Generative Sky: A Neurosymbolic Framework for In-Orbit Computation Offloading

Generative Sky: A Neurosymbolic Framework for In-Orbit Computation Offloading

Benedetta Picano *Member, IEEE*, and Daniele Tarchi, *Senior Member, IEEE*

***Abstract*—Satellite offloading is a critical issue in the Internet of Things (IoT) edge intelligence environment. In this work, we present a novel neurosymbolic framework for computation offloading decisions in satellite-enabled IoT edge intelligence scenarios. By combining the forecasting capabilities of time-series foundation models with the transparency of rule-based reasoning, our approach enables data-efficient and inherently explainable decision making under uncertainty. Specifically, we use TimeGPT to predict future throughput quality and satellite CPU load, which are then processed through a fuzzy logic controller to derive context-aware offloading decisions with transparent rationale. The results show effective forecast accuracy, high decision robustness, and improved explainability when compared to traditional reinforcement learning-based approaches that require task-specific training.**

***Index Terms*—Satellite Communications, Edge Computing, Neuro-symbolic AI, Foundation Models, Fuzzy Logic, Interpretable AI.**

I. Introduction

Satellite-assisted computation has emerged as a powerful enabler for extending network intelligence to regions beyond the reach of terrestrial infrastructure. By integrating satellite platforms into the Internet of Things (IoT) communication and computation workflows, networks can support latency-sensitive, mission-critical, or geographically isolated IoT applications where ground connectivity is intermittent or entirely unavailable [1], [2].

Edge intelligence is a novel concept that aims to integrate Artificial Intelligence (AI) and edge computing (EC), rapidly emerging as a transformative paradigm for optimizing complex systems [3], [4]. In scenarios characterized by heterogeneous IoT domains, such as aerial, maritime, or remote terrestrial environments, satellite nodes play a dual role: they act not only as communication relays bridging otherwise disconnected IoT endpoints, but also as computational offloading targets capable of executing tasks. This dual functionality transforms satellites into intelligent agents, actively contributing to data processing, decision making, and service continuity [5]–[7]. This convergence offers unprecedented opportunities to overcome inherent challenges such as limited onboard processing capabilities, stringent power constraints, and significant communication latencies. By pushing computational intelligence closer to the data source, be it individual satellites, ground stations, or even user terminals, edge intelligence facilitates real-time decision-making, autonomous operation, and efficient resource management [8].

Within this paradigm, a fundamental challenge arises: *how to design intelligent systems that can autonomously select when and where to offload computation, adapting their decisions to surrounding environmental conditions, link quality, and resource availability?* The need for context-aware optimization is particularly critical in satellite networks, where delays, weather-induced signal degradation, and fluctuating onboard loads can significantly impact system performance [9]–[12]. Therefore, enabling fast, adaptive, and interpretable decision-making becomes essential to achieve robust and efficient satellite-enabled edge intelligence. Addressing this need requires not only architectural innovations but also advances in computational intelligence capable of operating under tight constraints and multivariate or multi-modal inputs.

In parallel, the convergence of algorithmic breakthroughs and increasingly powerful computing hardware has fueled an unprecedented diffusion of AI techniques. Although access to high-end AI remains uneven in all deployment scenarios, the increasing availability of scalable hardware platforms, including those onboard satellites, has enabled the execution of sophisticated models. However, training such models often remains unfeasible in these settings, due to the high computational and energy demands involved [6], [7], [13]. For this reason, the use of models that require no training and are inherently capable of generalizing to unseen and scarce data sequences is becoming increasingly attractive.

At the core of this transformation lies the rise of foundation models. Originating in the field of natural language processing (NLP), foundation models are large-scale pre-trained models designed to serve as adaptable bases for a wide range of downstream tasks [14], [15]. Although the theoretical concept of reusable general-purpose models is not new, its practical realization has only recently become feasible because of the unprecedented computational power now available. This shift has enabled the training and deployment of large-scale neural networks that can generalize across a wide range of tasks, giving rise to what are now known as foundation models. These models are characterized by a pretraining phase on vast and diverse datasets, followed by the ability to perform downstream tasks with minimal or even no task-specific fine-tuning [14]–[16].

To this end, architectures such as TimeGPT [17] have demonstrated the capacity to perform accurate predictions in a zero-shot setting, offering insights and forecasts without requiring retraining or adaptation to domain-specific data. This marks a substantial paradigm shift, particularly for applica-

This work was supported by the Italian Ministry of University and Research (MUR) under the FIS 2 program (D.D. prot. n. 12308, July 22, 2025), Project 2023-03527 GANESHA – Networks-for-Humans: A Novel Cognitive Paradigm Connecting Events with Agents.

B. Picano and D. Tarchi are with the Department of Information Engineering, University of Florence, Firenze, Italy, email: benedetta.picano@unifi.it, daniele.tarchi@unifi.it

tions constrained by limited data availability or by the high computational cost of traditional training pipelines.

Such capabilities are especially relevant in space-enabled IoT scenarios, where environmental dynamics, e.g., atmospheric interference, changing link conditions, or node availability, create a highly variable operational context. Here, the ability to leverage pretrained models for rapid and low-cost forecasting becomes a strategic asset, enabling systems to anticipate future conditions and adapt their behavior accordingly without the overhead of local model updates or retraining procedures.

Foundational models for time-series data enable zero-shot inference, thus bypassing the necessity for retraining. However, similar to various other AI methodologies, such as deep learning, they inherently lack interpretability. Despite the remarkable predictive capabilities of both generative and deep learning models, their opacity remains a critical limitation. This is especially problematic in safety-critical or resource-constrained environments, such as satellite-assisted computation offloading, where understanding the rationale behind a decision is essential for trust, reliability, and regulatory compliance. The ambition of this work is to address both challenges: a forecasting framework that leverages zero-shot inference for generalization over scarce and unseen data, integrating reasoning mechanisms to enhance interpretability. In so doing, the neuro-symbolic paradigm integrates inductive and deductive reasoning by combining the data-driven adaptability of generative AI with the transparency and structure of symbolic, rule-based systems. This synergy not only enables accurate forecasting in uncertain environments but also supports interpretable, context-aware decision making by explicitly encoding reasoning steps.

In this work, we embrace this hybrid vision to develop a neuro-symbolic framework for satellite computation offloading that is both data-efficient, thanks to zero-shot generative forecasting, and inherently explainable, through a fuzzy logic controller that maps predictions into interpretable actions. Moreover, it is important to note that the novelty of the proposed approach does not reside in the individual building blocks taken in isolation, but in their coordinated use within a unified neuro-symbolic game-theoretic framework tailored to satellite-enabled computation offloading. In particular, this work demonstrates how generative foundation models can be effectively employed strictly at inference time to support operational decision making in non-terrestrial environments, where limited computational power and energy availability make prolonged or task-specific training impractical. By exploiting zero-shot time-series forecasting, the framework enables context-aware offloading decisions without requiring retraining or continuous model updates, thereby significantly extending the feasible application domain of generative models to resource-constrained satellite scenarios. In addition, the integration of a symbolic, rule-based decision layer introduces an explicit level of interpretability that is still largely underexplored in the satellite offloading literature. This design choice allows the decision process to be transparent, auditable, and grounded in domain knowledge, providing actionable insights into the rationale behind each offloading decision while avoiding the opacity of purely data-driven pipelines.

The main contributions of this paper are the following:

- We introduce the first neuro-symbolic framework, based on the integration of time-series foundation models and symbolic reasoning, for satellite offloading decision making. To the best of the authors' knowledge, this is the first application of time-series foundation models in zero-shot mode to support the decision process for computation delegation from end devices to satellite nodes.
- We design a hybrid approach, i.e., neurosymbolic, that combines data-driven forecasting and interpretable control. Specifically, we use TimeGPT to predict future throughput quality and satellite CPU load, which are then processed through a fuzzy logic controller to derive context-aware offloading decisions with transparent rationale.
- A multi-user scenario with heterogeneous IoT nodes has been considered as an extension of the single-user scenario where multiple nodes compete for the communication and computing resources. For this purpose, a matching game with externalities has been considered.
- We conduct a comprehensive evaluation that highlights the effectiveness of the proposed framework. The results show competitive forecast accuracy, high robustness of the decision, and improved explainability when compared to traditional reinforcement learning-based approaches that require task-specific training. The experimental evaluation includes investigations in both single-user and multi-user scenarios.

The rest of the paper is organized as follows. In Section II, the related works are discussed. In Section III the system model and the formulation of the problem are presented. The proposed approach is then detailed in Section IV. Performance evaluation is discussed in Section V, and conclusions are drawn in Section VI.

II. Related Works

The optimization of computation offloading in satellite-enabled IoT networks is a rapidly evolving field, driven by the necessity to manage stringent onboard resource constraints and the high volatility of LEO links. To provide a critical synthesis of the state-of-the-art, this section categorizes existing literature into two primary thematic directions. First, we review Satellite-Based Task Offloading, which encompasses global optimization, data-driven resource management strategies, and simulation driven policies. Second, we examine the emerging integration of Generative AI and Neurosymbolic Approaches in space networks, focusing on theoretical foundations, application-specific solutions and optimization policy development. By grouping these works, we highlight a persistent research gap: the heavy reliance of current models on extensive task-specific training data and their inherent black-box nature. This synthesis serves to position our proposed neurosymbolic framework as a solution that addresses these limitations through zero-shot forecasting and interpretable logic.

A. Satellite-Based Task Offloading

Existing research on satellite-based task offloading can be broadly grouped into optimization-driven, learning-based, and system-level evaluation approaches.

A first line of work focuses on *optimization-based formulations* for joint communication and computation management in satellite or integrated Ground–Air–Space networks. For instance, SkyLink [18] models aerial and satellite nodes as autonomous agents and adopts a bi-level optimization framework based on multi-agent reinforcement learning to jointly optimize positioning, resource allocation, and offloading decisions under highly dynamic conditions. Similarly, [19] addresses cooperative multi-satellite edge computing by jointly optimizing satellite selection, resource distribution, scheduling, and transmission power via convex relaxation and Lagrangian methods, achieving latency reductions through coordinated inter-satellite processing. Energy-centric optimization frameworks are further explored in [20]–[22], where offloading, power control, and scheduling are jointly optimized to minimize energy consumption or composite latency–energy costs in heterogeneous UAV–satellite or NTN architectures. *In contrast to these approaches, our work does not rely on solving complex global optimization problems online, but instead leverages zero-shot forecasting and symbolic reasoning to enable lightweight, interpretable, and rapidly adaptable offloading decisions.*

A second category comprises *learning-based offloading strategies*, where decision policies are learned directly from data. Deep reinforcement learning is adopted in [23] to manage task priority and delay constraints in satellite-assisted vehicular edge computing, while [24] combines particle swarm optimization, genetic algorithms, and Q-learning to reduce task completion delay in dynamic LEO constellations. Hierarchical reinforcement learning is further employed in [5] to optimize network selection and partial offloading in integrated terrestrial and non-terrestrial vehicular scenarios. *Differently from these fully data-driven solutions, our framework avoids task-specific training and reinforcement signals, replacing learned policies with a neuro-symbolic architecture that combines zero-shot foundation models with transparent rule-based decision making.*

A third group of studies emphasizes *system-level performance evaluation and architectural feasibility*. Works such as [25], [26] investigate offloading strategies for XR applications or vehicular clusters using realistic satellite orbital data, quantifying latency and energy trade-offs through detailed simulations. Other contributions, e.g., [27], focus on satellite–ground co-inference for DNN deployment, proposing fine-grained partitioning of neural network layers to reduce onboard energy consumption. *While these studies provide valuable insights into feasibility and performance limits, they primarily assess static or simulation-driven policies, whereas our work targets dynamic, context-aware offloading decisions with intrinsic explainability and minimal computational overhead.*

The problem of computation offloading in aerial edge computing has also been investigated in the context of UAV-assisted systems. For instance, the authors in [28] propose a graph convolutional reinforcement learning framework to jointly optimize UAV trajectory and task offloading decisions. Differently from our approach, which relies on zero-shot foundation models and interpretable neuro-symbolic reasoning for satellite-assisted environments, their method adopts a learning-based optimization strategy tailored to dynamic aerial edge scenarios. The integration of satellite and aerial networks to support massive IoT services has also been investigated using advanced multiple access schemes. In particular, the authors in [29] propose an RSMA-based framework to enhance spectral efficiency and resource sharing in integrated satellite–aerial architectures, focusing on communication-layer optimization rather than predictive offloading decisions. Similarly, AI-driven solutions for seamless and massive access in space–air–ground integrated networks have been recently investigated to improve scalability and connectivity management in heterogeneous environments. The authors in [30] explore the use of intelligent access control and resource coordination mechanisms across multi-layer infrastructures, with emphasis on communication and access optimization rather than predictive inference-aware offloading strategies as considered in our framework. Recent efforts have also investigated Direct-to-Smartphone communication as a key enabling technology for 6G non-terrestrial IoT connectivity. In particular, the authors in [31] analyze technical architectures and implementation challenges for integrated satellite–terrestrial access, focusing on air-interface adaptation, multi-beam satellite payload design, and access procedures rather than inference-aware predictive offloading strategies as considered in our framework. Multi-objective deep reinforcement learning has also been recently explored for time–frequency resource allocation in multi-beam satellite communication systems. In particular, authors in [32] propose a learning-based framework to jointly optimize spectral efficiency and interference management across satellite beams, focusing on physical-layer resource allocation rather than predictive inference-aware offloading strategies as considered in our framework. Overall, existing literature demonstrates the effectiveness of optimization and learning-based methods for satellite offloading, but it also reveals common limitations: reliance on heavy training or complex solvers, limited interpretability, and reduced adaptability in data-scarce or rapidly changing environments. The proposed neurosymbolic framework directly addresses these gaps by integrating zero-shot time-series foundation models with symbolic reasoning, enabling fast, explainable, and data-efficient offloading decisions in satellite-assisted IoT systems.

B. Generative AI in Satellite Networks and Neurosymbolic Approaches

Recent advances in neurosymbolic artificial intelligence have explored different strategies to integrate data-driven learning with structured reasoning, primarily with the goal of improving interpretability and robustness. A foundational line of work is presented in [33], where neurosymbolic reasoning is formalized through energy-based models capable of representing propositional logic via Restricted Boltzmann Machines. By embedding logical constraints into energy landscapes, the

framework enables inference through energy minimization and demonstrates how symbolic structure can be preserved within neural architectures. *While this approach provides a rigorous theoretical foundation for neurosymbolic reasoning, it does not address time-varying decision making or resource-aware control in dynamic networked systems, which are central to the satellite offloading problem considered in our work.*

Complementary to formal models, conceptual and survey-style contributions have clarified the motivation and scope of neurosymbolic AI. The work in [34] frames neurosymbolic systems through the lens of human perception and cognition, arguing that symbolic abstractions are essential for explainable and safety-critical AI. By contrasting neural perception with symbolic cognition, the authors highlight the necessity of integrating structured reasoning into modern AI pipelines.*Our work operationalizes this conceptual vision in a concrete networking scenario, translating neurosymbolic principles into an actionable offloading framework driven by predictive context and interpretable control rules.*

Practical applications of neurosymbolic paradigms have also emerged in data-intensive domains. In [35], symbolic ontologies capturing environmental knowledge are combined with satellite-derived remote sensing data to improve crop-yield prediction accuracy. The results demonstrate that integrating symbolic domain knowledge with statistical learning can outperform purely data-driven or purely symbolic approaches. *Unlike this application-specific forecasting framework, our approach targets real-time decision making under uncertainty, where symbolic reasoning is not used solely to enhance prediction accuracy but to directly govern offloading actions in a resource-constrained satellite environment.*

Generative AI has recently been explored as an enabler for modeling and optimization in satellite networks. The framework proposed in [36] employs large language models within a retrieval-augmented generation paradigm to support the interactive construction of mathematical models for large-scale satellite systems. These generative components are coupled with mixture-of-experts reinforcement learning to optimize network variables under complex interference and coordination constraints. *In contrast, our work does not use generative models to synthesize optimization policies or solve network models, but rather leverages time-series foundation models in a zero-shot setting to forecast system context, which is then processed through symbolic reasoning to yield explainable and lightweight offloading decisions.*

Overall, existing neurosymbolic and generative AI approaches either emphasize formal reasoning, conceptual foundations, or data-driven optimization. However, they commonly overlook the combination of zero-shot forecasting, symbolic decision making, and deployment-oriented constraints. The proposed framework fills this gap by integrating generative time-series models with fuzzy logic reasoning to support fast, interpretable, and data-efficient offloading decisions in satellite-enabled IoT networks. Recent studies as [37] have also explored the integration of distributed artificial intelligence within space–air–ground–sea integrated networks (SAGSIN) as a key enabler for future 6G systems. In this line of work, SAGSIN is modeled as a multi-layered and heterogeneous network architecture spanning space, aerial, terrestrial, and maritime domains, where distributed AI is leveraged to coordinate communication, computing, and sensing resources at a global scale. These contributions emphasize how decentralized intelligence can enhance adaptability and service provisioning in highly dynamic environments, while also revealing new challenges arising from the scale, heterogeneity, and openness of SAGSIN infrastructures. In particular, recent surveys have highlighted that the tight coupling between distributed AI and SAGSIN exacerbates security and privacy concerns across communication, computation, and sensing planes, motivating the need for intelligent, context-aware protection mechanisms. While this body of work primarily focuses on security issues and privacy-preserving strategies in distributed AI-enabled SAGSIN, our approach is complementary in scope. Rather than addressing secure communications or distributed learning, we concentrate on explainable and lightweight decision making for computation offloading, leveraging zero-shot forecasting and symbolic reasoning to operate effectively under resource and data constraints typical of satellite-assisted environments.

III. Problem Statement

A. System Model

We consider a space-enabled edge intelligence scenario with a set $\mathcal{E} = \{E_1, \ldots, E_k, \ldots, E_K\}$ of K IoT end devices randomly placed at ground. The generic IoT end device E_k creates tasks to be processed. A generic task $\rho_k(t)$, generated at time instant t, is individuated by a tuple $\langle d_k(t), c_k(t), \eta_k(t)\rangle$ where $d_k(t)$ is the size of the task expressed in Bytes (or multiple of Bytes), $c_k(t)$ is the task processing requirements in GFLOPs and $\eta_k(t)$ the task requirements in terms of data rate due to specific service constraints. Each IoT end device E_k interacts with a constellation of satellites $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$. At each discrete time step t, a subset of satellites is within the communication range of E_k and can be considered as candidate nodes for offloading.

By focusing on a time instant t, each satellite $S_i(t) \in \mathcal{S}(t)$, in reference to an IoT device E_k, is characterized by the tuple $\langle Q_{k,i}(t), C_i(t)\rangle$ that defines the throughput score $Q_{k,i}$ and the CPU load value $C_i(t)$. The throughput score $Q_{k,i}$ reflects the expected performance of the satellite communication channel, which can be affected by factors such as atmospheric conditions, signal-to-noise ratio, and geometric distance between E_k and $S_i(t)$. The CPU load $C_i(t)$ captures the computational availability of the satellite, i.e., whether the satellite is busy or underutilized. Both $Q_{k,i}(t)$ and $C_i(t)$ are predicted using a foundation model, i.e., producing their estimations denoted by $\hat{Q}_{k,i}(t)$ and $\hat{C}_i(t)$, for time series forecasting, specifically TimeGPT[1], in a zero-shot fashion [17]. The values of $Q_{k,i}(t)$

[1]The adoption of a zero-shot time-series foundation model such as TimeGPT may introduce potential performance degradation under out-of-distribution (OOD) conditions when the statistical properties of the observed satellite traffic or channel dynamics differ significantly from those seen during pre-training. However, recent studies show that large-scale multi-domain pre-training improves cross-domain transferability and robustness compared to task-specific forecasting models [38]. A more systematic assessment of OOD robustness in satellite IoT environments represents an interesting direction for future work.

TABLE I
COMPARISON WITH RELATED WORKS ON SATELLITE OFFLOADING AND NEUROSYMBOLIC AI

Reference	Category	Main Method	Difference w.r.t. our framework
[18]	Optimization + MARL	Bi-level RL optimization for positioning and offloading	Requires training and global optimization
[19]	Convex optimization	Joint satellite selection and resource allocation	Centralized solver-based decisions
[20]	Energy optimization	Joint power control and scheduling	No predictive context awareness
[21]	Latency–energy optimization	Joint offloading optimization	Heavy optimization pipeline
[22]	NTN optimization	Resource scheduling strategies	No explainable decision logic
[23]	Deep RL	Priority-aware policy learning	Task-specific training required
[24]	Hybrid learning	PSO + GA + Q-learning	Training-heavy solution
[5]	Hierarchical RL	Network selection and partial offloading	Policy learning dependency
[25]	System-level evaluation	Simulation-based XR offloading analysis	Static evaluation framework
[26]	System evaluation	Orbital-data driven simulations	No adaptive decision layer
[27]	DNN co-inference	Layer partitioning strategies	Inference splitting only
[28]	GCN + RL	Trajectory-aware UAV offloading	Learning-based aerial scenario
[29]	RSMA framework	Multiple-access optimization	Communication-layer focus
[30]	AI access control	Resource coordination in SAGIN	Access-layer optimization only
[31]	Direct-to-smartphone NTN	Air-interface architecture	Access procedure design focus
[32]	Deep RL	Time-frequency allocation	PHY-layer optimization
[33]	Neurosymbolic theory	RBM-based logical reasoning	No network decision support
[34]	Conceptual framework	Explainable cognition-inspired AI	No operational networking use
[35]	Neurosymbolic forecasting	Ontology + satellite sensing data	Prediction-only framework
[36]	LLM + RL optimization	RAG-based satellite modeling	Optimization instead of offloading control
[37]	Distributed AI SAGSIN	Multi-layer coordination and security	Security-focused architecture

are extracted from a public satellite dataset, i.e., WetLinks [39] and include physical and environmental parameters. The IoT end device E_k must decide whether to offload a given computation task at time t and, if so, to which satellite $S_i(t)$.

For an offloading operation to be considered successful, the link quality score must exceed a reliability threshold, defined as the expected mean value of the throughput link under clear-sky conditions, i.e., in the absence of adverse atmospheric phenomena. In addition, the selected satellite must not exhibit significant CPU load, meaning that it should not be currently involved in the computation of other tasks.

Let $O_i^{\text{success}}(t)$ be a binary variable equal to 1 if the offloading to satellite S_i at time t is successful, and 0 otherwise. The conditions for success can be formulated as follows:

$$O_i^{\text{success}}(t) = \begin{cases} 1 & \text{if } \hat{Q}_{k,i}(t) > Q_{\text{th}} \text{ and } \hat{C}_i(t) < C_{\text{th}} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where:

- $O_i^{\text{success}}(t)$ is the binary indicator of a successful offloading operation to satellite S_i at time t.
- $\hat{Q}_{k,i}(t)$ is the predicted throughput quality score for the link to satellite S_i.
- $\hat{C}_i(t)$ is the predicted CPU load for the satellite S_i.
- Q_{th} is the reliability threshold for the link quality, defined as the expected mean value of the throughput under clear-sky conditions.
- C_{th} is the maximum acceptable CPU load threshold. If the predicted load $\hat{C}_i(t)$ is below this value, the satellite is considered to have sufficient computational availability and does not exhibit a significant load.

The proposed Generative Sky framework is illustrated in Fig. 1, which delineates the architectural synergy between neural forecasting and symbolic reasoning. The system is organized into a closed-loop pipeline spanning observation, forecasting and decision layers to balance computational efficiency with decision interpretability. At the edge of the network, LEO satellites and IoT devices constitute the observation layer. As shown in the figure, the IoT nodes continuously monitor the Throughput Score derived from link geometry and environmental factors, while each satellite monitors the current onboard CPU utilization. Because high-fidelity forecasting requires significant computational resources, these historical telemetry sequences are transmitted to a control center via a dedicated control link. The core of the framework resides in the processing engine, where a TimeGPT foundation model performs multivariate, zero-shot forecasting. By analyzing a context window of length l, the model predicts the future state of the link and CPU load. These predictions are then ingested by the Fuzzy Logic Controller (FLC), as depicted in the upper layer of Fig. 1. The FLC applies a set of human-readable IF-THEN rules to the predicted values to determine the optimal offloading strategy. Finally, this decision is sent back to the IoT nodes for execution, ensuring that the offloading action is

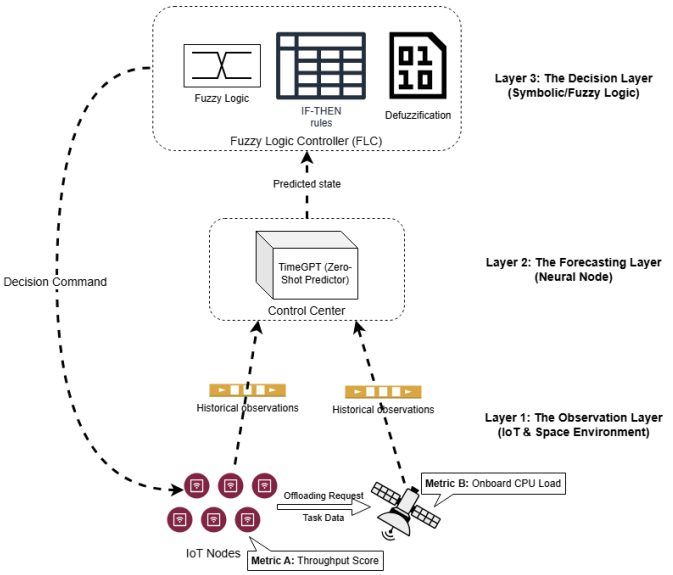


Fig. 1. Proposed Neurosymbolic Framework Architecture. The diagram illustrates the data-flow pipeline where historical satellite link and CPU telemetry are processed by the TimeGPT foundation model for zero-shot forecasting. The resulting predictions are fed into the Fuzzy Logic Controller (FLC) to provide an interpretable offloading decision, effectively bridging neural predictive power with symbolic reasoning.

based on the most probable future state of the network.

B. Problem Formulation

Our objective is to effectively determine whether the tasks generated by the reference IoT end device should be offloaded to any satellite within visibility. To achieve this, accurate forecasting of both link quality and CPU load is advantageous. The overall objective of the proposed framework is twofold: (i) to generate accurate forecasts for throughput quality and satellite CPU load, and (ii) to support effective and interpretable offloading decisions based on such forecasts.

More formally, the offloading decision problem can be interpreted as a constrained predictive decision problem in which, at each time instant t, the IoT device evaluates the feasibility of offloading toward the satellites in the visibility set $\mathcal{S}_k(t)$ based on the predicted system conditions.

In particular, an offloading action toward satellite S_i is considered admissible only if both communication and computation feasibility conditions are satisfied, i.e.,

$$\hat{Q}_{k,i}(t) \geq Q_{\text{th}}, \quad \hat{C}_i(t) \leq C_{\text{th}}, \quad \forall S_i \in \mathcal{S}_k(t) \tag{2}$$

These feasibility constraints define the set of candidate satellites that can successfully execute the offloaded task. Accordingly, the decision-making process aims to select a satellite satisfying the above constraints whenever possible, while minimizing the cumulative number of offloading failures over the decision horizon. Under this formulation, the forecasting module provides estimates of the future system state, whereas the symbolic controller implements a rule-based policy that approximates the solution of the constrained offloading decision problem in an interpretable and computationally lightweight manner.

1) Forecasting Objective: Given a historical time series of environmental and system-level features $\mathbf{x}(t)$, the forecasting model aims to predict future values of link throughput quality $Q_{k,i}(t+\tau)$ and CPU load $C_i(t+\tau)$ for each satellite S_i. Let $\hat{y}(t)$ denote the generic predicted value and $y(t)$ the generic ground-truth. The prediction accuracy is evaluated using typical error metrics:

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |\hat{y}(t) - y(t)| \tag{3}$$

- Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100}{T}\sum_{t=1}^{T} \left| \frac{\hat{y}(t) - y(t)}{y(t)} \right| \tag{4}$$

2) Decision-Making Objective: Based on the predicted values $\hat{Q}_{k,i}(t)$ and $\hat{C}_i(t)$, the symbolic controller makes a decision to either offload a computation to satellite S_i or execute it locally. We define an offloading failure as one of the following two cases:

1) **False Positive:** the system decides to offload, but the link throughput quality or CPU availability is inadequate, leading to failure.
2) **False Negative:** the system decides not to offload, although there existed at least one suitable satellite.

Let $F(t) \in \{0, 1\}$ be a binary decision-dependent indicator variable that equals 1 if a service failure occurs at time step t, and 0 otherwise. The second objective is to minimize the cumulative number of failures over the decision horizon:

$$\min \sum_{t=1}^{T} F(t)$$

subject to

$$\hat{Q}_{k,i}(t) \geq Q_{\text{th}}, \quad \forall S_i \in \mathcal{S}_k(t), \tag{5}$$

$$\hat{C}_i(t) \leq C_{\text{th}}, \quad \forall S_i \in \mathcal{S}_k(t), \tag{6}$$

which define the feasibility conditions for a successful offloading operation at time instant t. Under this formulation, the forecasting module provides estimates of the future system state, while the symbolic controller determines the offloading decision by selecting satellites that satisfy the above feasibility constraints whenever possible.

This objective complements the predictive accuracy metrics and aims to ensure the robustness of the symbolic policy under uncertainty. In the following section, we introduce a neuro-symbolic decision-making framework that leverages both predicted context and fuzzy logic rules to achieve this goal.

IV. A Neurosymbolic Approach for Satellite Offloading

To address the offloading decision problem under uncertainty, we introduce a neurosymbolic framework that integrates a foundation model for forecasting with a rule-based symbolic

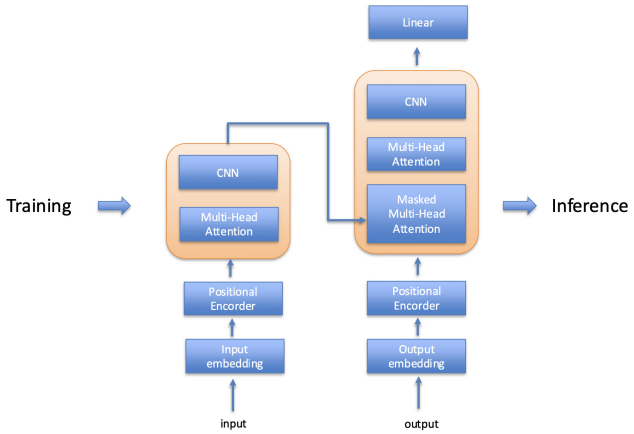


Fig. 2. Overview of the TimeGPT architecture, illustrating the separation between the training and inference phases. Historical input time-series are first mapped into input embeddings and enriched with positional encoding, then processed through stacked CNN and multi-head attention blocks during training. At inference time, the decoder leverages masked multi-head attention, together with CNN and attention layers, to generate future predictions in an autoregressive manner. The final linear layer maps latent representations to forecasted output values, enabling zero-shot time-series prediction without task-specific fine-tuning.

decision-making module. This hybrid architecture aims to combine the generalization and data-efficiency of generative AI with the interpretability and domain knowledge encoded in symbolic reasoning.

A. Forecasting Module

The first component of our architecture is a forecasting module based on TimeGPT [17], a pretrained foundation model for time series prediction. Given a sequence of past observations, TimeGPT produces zero-shot forecasts for the future values of relevant variables, without requiring task-specific fine-tuning. As reported in Fig. 2, TimeGPT is a foundation model specifically designed for time series forecasting, built upon the Transformer architecture originally introduced for natural language processing tasks. Its core structure relies on self-attention mechanisms to capture both short- and long-term temporal dependencies in sequential data. Unlike traditional sequence models that assume regular input intervals or require heavy task-specific training, TimeGPT is pretrained on a massive corpus of heterogeneous time series data, enabling it to learn generalizable temporal patterns across domains. The model incorporates a temporal embedding layer to encode time-related features such as seasonality, trend, and timestamp context, which are then processed through multiple Transformer layers with multi-head attention and layer normalization. The output layer is adapted to forecast a future horizon h in a zero-shot or few-shot fashion, without the need for fine-tuning. This architectural design allows TimeGPT to generalize effectively across a wide range of forecasting problems with minimal input data and no prior task-specific supervision.

In our setup, the forecasting module takes as input historical measurements of:

- Link throughput indicator $Q_{k,i}(t)$ for each satellite S_i, here derived from the dataset [39];
- CPU load estimates $C_i(t)$ for each satellite S_i, modeled via dataset [40].

For each satellite, the model generates a forecast of link throughput quality $\hat{Q}_{k,i}(t+1)$ and CPU usage $\hat{C}_i(t+1)$ in the next decision slot. These predictions serve as input to the symbolic controller.

B. Symbolic Controller

The symbolic module is implemented as a fuzzy logic controller that translates continuous uncertain input predictions into discrete offloading decisions. This controller enables robust decision-making under varying conditions by leveraging human-readable rules, which encode expert knowledge about system behavior. Specifically, the fuzzy system operates on two predicted inputs for each satellite: the expected link throughput quality and the anticipated CPU occupancy. These inputs are fuzzified into linguistic variables such as “low”, “medium”, and “high” based on domain-specific membership functions. The thresholds used to define the linguistic categories (i.e., low, medium, high) were chosen to provide a balanced and interpretable partition of the input datasets. These values reflect a gradual transition between regions with the aim of capturing typical operating conditions and ensuring that the fuzzy system behaves in a consistent and predictable manner. Each decision rule takes the form:

> *IF* link quality is high *AND* CPU load is low, *THEN* offload.

The fuzzy logic controller adopts a Mamdani-type inference mechanism, where both the antecedents and the consequents are expressed as linguistic variables. This choice ensures high interpretability, as each rule directly maps human-understandable conditions (e.g., “link quality is high”) to actionable decisions (e.g., “offload task”). The fuzzy outputs are subsequently defuzzified using the centroid method to produce a binary offloading decision. Defuzzification is performed using the centroid method, which ensures a smooth and interpretable mapping from fuzzy activation levels to crisp decisions.

The interpretability of the fuzzy controller offers two key advantages: first, it facilitates transparency in system behavior and decision rationale; second, it allows easy adaptation to new environments by simply updating or re-weighting the rule base, without retraining. This symbolic reasoning component complements the predictive layer by providing a lightweight explainable mechanism for operational deployment across heterogeneous satellite constellations and diverse mission profiles. The membership functions reported in Table II, together with the complete fuzzy rule base listed in Table III, are shared between all satellites and users to ensure consistency, transparency and complete reproducibility of the decision-making process.

C. Hybrid Execution Loop

At each time step t, considering the IoT device E_k, the system operates according to the following loop:

TABLE II
MEMBERSHIP FUNCTIONS USED IN THE FUZZY CONTROLLER

Variable	Linguistic Term	Type	Parameters
Link Quality	Low	Trapezoidal	$(0, 0, 30, 50)$
Link Quality	Medium	Triangular	$(30, 50, 70)$
Link Quality	High	Trapezoidal	$(50, 70, 100, 100)$
CPU Load	Low	Trapezoidal	$(0, 0, 30, 50)$
CPU Load	Medium	Triangular	$(30, 50, 70)$
CPU Load	High	Trapezoidal	$(50, 70, 100, 100)$
Offloading Urgency	Low	Trapezoidal	$(0, 0, 30, 50)$
Offloading Urgency	Medium	Triangular	$(30, 50, 70)$
Offloading Urgency	High	Trapezoidal	$(50, 70, 100, 100)$

TABLE III
FUZZY RULE BASE FOR OFFLOADING DECISIONS

Link Quality	CPU Load	Offloading Urgency
Low	High	Low
Low	Medium	Low
Low	Low	Low
Medium	High	Low
Medium	Medium	Medium
Medium	Low	High
High	High	Medium
High	Medium	High
High	Low	High

1) Observe the previous l values of $Q_{k,i}(\tau)$ and $C_i(\tau)$ for each satellite S_i, where $\tau = t - l, \ldots, t$. These measurements include throughput quality indicators and CPU usage levels observed over a sliding window;
2) Use TimeGPT to predict the next-step values $\hat{Q}_{k,i}(t+1)$ and $\hat{C}_i(t+1)$ for each satellite. The predictions are obtained in a zero-shot fashion, without requiring any task-specific retraining. It is important to note that, given the limited onboard resources typical of current satellite platforms, commonly equipped with 4 cores and offering a peak performance of 8–12 GFLOPS, the entire TimeGPT model cannot be executed in situ. As a result, inference must be offloaded to a terrestrial node or a centralized control center. This design choice introduces a trade-off: while it enables the use of powerful generative models, it also adds latency due to the round trip time.
   To address this issue, it is important to relate inference latency to the temporal granularity of the decision process. In our evaluation, satellite link quality measurements are derived from the WetLinks dataset, where samples are uniformly collected at 3-minute intervals. Consequently, a one-step-ahead prediction corresponds to a forecasting horizon of approximately 3 minutes, while multi-step forecasts extend this horizon proportionally (e.g., four steps correspond to roughly 12 minutes). This prediction timescale is significantly larger than typical satellite–ground round-trip times. Empirical RTT measurements reported in the same dataset indicate average values around 60 ms. Even when conservatively accounting for additional processing and signaling delays, the overall communication latency remains negligible when compared to the prediction horizon adopted in the decision-making process. Furthermore, the execution time analysis shows that, for traditional learning-based approaches such as LSTM, the dominant latency contribution arises from the combined cost of model training and inference, which can reach several seconds. In contrast, the proposed zero-shot forecasting approach exhibits substantially lower execution times, remaining below a few seconds even under pessimistic assumptions that include CPU-based execution and API communication overhead. It is important to remark that, in a centralized forecasting architecture, the overall control loop latency includes not only the model inference time but also the synchronization of historical state measurements from satellites and IoT devices to the ground segment, as well as the dissemination of the resulting offloading decisions. In the considered scenario, however, these coordination delays remain significantly smaller than the adopted prediction horizon (on the order of minutes), thus preserving the validity of the predicted values at decision time.
   Taken together, these observations indicate a clear separation of timescales between inference latency and decision horizons. As a result, the predicted values remain valid and actionable upon reception, ensuring that the offloading decisions are not adversely affected by communication delays. This further highlights the suitability of zero-shot generative forecasting for satellite-assisted systems, especially when compared to training-dependent solutions in resource-constrained and time-varying environments.
   However, in-depth investigation of the enabling technologies required for onboard satellite inference, such as hardware optimizations, quantization strategies, or model compression, is beyond the scope of this work;
3) Input the predicted values into the fuzzy rule-based system, which evaluates a set of interpretable decision rules based on predefined thresholds for link reliability and CPU availability;
4) The final offloading decision is derived by comparing the defuzzified urgency scores against a predefined threshold, selecting the satellite with the highest score when the threshold is exceeded.

This architecture enables dynamic adaptation to time-varying channel and processing conditions while preserving explainability. Moreover, the decoupling between the forecasting module and the symbolic decision layer provides modularity and robustness, allowing independent tuning or replacement of each component without retraining the entire system. The pseudocode of this approach is reported in Algorithm 1.

The computational complexity of the proposed neuro-symbolic framework can be analyzed by considering separately the forecasting module and the symbolic decision layer. At each decision step t, the forecasting module generates one-step-ahead predictions $\hat{Q}_{k,i}(t+1)$ and $\hat{C}_i(t+1)$ for each satellite $S_i \in \mathcal{S}_k(t)$ using a history window of length l. Since inference is performed independently for each satellite, the overall forecasting cost scales linearly with the number of visible satellites, i.e., $\mathcal{O}(|\mathcal{S}_k(t)|)$ per decision step, assuming

Algorithm 1 Neuro-Symbolic Satellite Offloading

Require: History window length k, decision horizon T, list of satellites $\{S_i\}$, fuzzy rule thresholds

for $t = 1$ to T **do**
  for each satellite S_i **do**
    Observe past l values of throughput quality $Q_{k,i}(t-k:t)$
    Observe past l values of CPU load $C_i(t-l:t)$
    Predict $\hat{Q}_{k,i}(t+1) \leftarrow \text{TimeGPT}(Q_{k,i}(t-l:t))$
    Predict $\hat{C}_i(t+1) \leftarrow \text{TimeGPT}(C_i(t-l:t))$
  end for
  for each satellite S_i **do**
    Compute fuzzy score $\phi_i(t+1) \leftarrow \text{FuzzyRuleEval}(\hat{Q}_{k,i}(t+1), \hat{C}_i(t+1))$
  end for
  if $\max_i \phi_i(t+1) \geq$ threshold **then**
    Offload to satellite $S^* = \arg\max_i \phi_i(t+1)$
  else
    Execute task locally
  end if
  Log decision outcome (success or failure)
end for

constant-time external model inference through the TimeGPT API.

The symbolic controller evaluates a fixed number of fuzzy rules for each satellite. As the rule base size is constant and independent of system parameters, the fuzzy inference stage also exhibits linear complexity with respect to the number of candidate satellites, i.e., $\mathcal{O}(|\mathcal{S}_k(t)|)$.

Finally, the selection of the best satellite through maximum-score comparison requires a linear scan over the candidate set, resulting in complexity $\mathcal{O}(|\mathcal{S}_k(t)|)$. Therefore, the overall per-step computational complexity of the proposed decision process scales as $\mathcal{O}(|\mathcal{S}_k(t)|)$, while the total complexity over a decision horizon of length T becomes $\mathcal{O}(T \cdot |\mathcal{S}_k(t)|)$. In terms of space complexity, the algorithm stores only the sliding history window of length l for each satellite together with predicted values and fuzzy scores, resulting in memory requirements of order $\mathcal{O}(l \cdot |\mathcal{S}_k(t)|)$, which confirms the scalability of the proposed framework with respect to the number of visible satellites and decision horizon length.

D. Explainability of the Offloading Policy

The neurosymbolic architecture enables traceable and interpretable decisions by construction. Although the forecasting module operates as a general-purpose black-box predictor, the symbolic layer guarantees full explainability of the decision process through rule-based evaluation.

Proposition 1 (Traceability of Offloading Decisions)**.** *Given a finite rule base $\mathcal{R}$ of n fuzzy rules, and predicted inputs $(\hat{Q}_{k,i}(t+1), \hat{C}_i(t+1))$, the offloading decision at time t can be fully reconstructed by evaluating the rule activation degrees $\{\mu_j\}_{j=1}^n$ and the associated defuzzification process.*

Proof. Each fuzzy rule $r_j \in \mathcal{R}$ maps input linguistic variables to output actions using membership functions $\mu_j : [0,1]^2 \to [0,1]$. The overall fuzzy output $\phi_i(t+1)$ is calculated as a weighted aggregation of these activations:

$$\phi_i(t+1) = \sum_{j=1}^{n} w_j \cdot \mu_j(\hat{Q}_{k,i}(t+1), \hat{C}_i(t+1)),$$

where w_j denotes the rule-specific weight (default 1 in unweighted Mamdani inference). Since all components are explicit and deterministic, the final decision can be traced to the tuple $(\hat{Q}_{k,i}, \hat{C}_i)$ and the structure of $\mathcal{R}$. □

This result guarantees post-hoc traceability of each decision and supports external auditing or symbolic counterfactual analysis. The design thus satisfies an essential criterion for explainable AI: the ability to justify each output based on observable intermediate computations.

E. Extension to Multi-User Offloading with Matching under Externalities

The analysis carried out for the single-user scenario can be extended to a more realistic multi-user setting, where multiple end devices E_k, $k \in \{1, \ldots, K\}$, generate tasks concurrently and compete for the same set of visible satellites. In this case, each task $\rho_k(t)$ is characterized by a predicted CPU load requirement $\chi_k(t)$, in addition to the previously defined parameters. This multi-user setting is representative of realistic satellite IoT deployments, where heterogeneous devices, such as low-priority environmental sensors, high-priority autonomous vehicle data streams, and critical infrastructure monitors, must compete for limited satellite resources.

Unlike the single-user case, offloading decisions now introduce *externalities*, since the selection of a user by a given satellite affects the residual CPU load and thus impacts the desirability of that satellite for the remaining users. This setting can be naturally modeled as a matching game with externalities between the set of users $\mathcal{E}(t) = \{E_1, \ldots, E_K\}$ and the set of satellites $\mathcal{S}(t)$.

1) Matching Game Formulation: Each user E_k evaluates the satellites in $\mathcal{S}_k(t)$, i.e., those within visibility at time t, based on the predicted pair $(\hat{Q}_{k,i}(t+1), \hat{C}_i(t+1))$ using the same fuzzy logic controller defined in the single-user case. Each user then ranks the satellites according to the resulting offloading urgency score $\phi_{k,i}(t+1)$.

Concurrently, each satellite S_i evaluates the users that propose to offload towards it. To prioritize its utilization, S_i selects the user with the highest predicted computational demand $\chi_k(t)$ among those proposing. This policy promotes the effective use of satellite resources by preferring tasks with greater computational requirements.

2) Externality Model: The association between a user E_k and a satellite S_i modifies the predicted CPU load for that satellite. Specifically, if the initial predicted load is $\hat{C}_i(t+1)$, the new effective CPU occupancy becomes:

$$\tilde{C}_i(t+1) = \hat{C}_i(t+1) + \alpha \cdot \chi_k(t), \tag{7}$$

where α is a normalization factor used to scale the user load $\chi_k(t)$ to the same domain as the predicted CPU values. This factor can be learned from the data or estimated to ensure that $\tilde{C}_i(t+1) \in [0, 100]$ (or normalized to $[0, 1]$). If multiple users are accepted, their contributions to the load accumulate.

3) Matching Procedure: At each time slot t, the system proceeds as follows:

1) Each user E_k forecasts the link quality $\hat{Q}_{k,i}(t+1)$ and receives the predicted CPU load $\hat{C}_i(t+1)$ for each visible satellite $S_i \in \mathcal{S}_k$.
2) Each user computes the fuzzy urgency score $\phi_{k,i}(t+1)$ and proposes offloading to the satellite with the highest score above a given threshold.
3) Each satellite S_i collects proposals and selects the user with the highest $\chi_k(t)$, then updates its effective CPU load.
4) The remaining users repeat the proposal step, and the process continues iteratively until convergence or all users are either matched or rejected.

This matching process captures the impact of resource contention and enables the system to adapt to varying task loads and satellite conditions in a scalable and explainable manner. In the considered framework, user competition is modeled at the computational resource level rather than at the access layer, since each user is assumed to operate on an independent communication channel. Externalities therefore arise from the progressive consumption of satellite CPU availability as tasks are assigned, which directly affects the remaining matching opportunities for other users.

V. Performance Evaluation

To validate the proposed neurosymbolic offloading framework, we conduct a comprehensive performance evaluation based on real-world datasets and simulation scenarios. The goal is to assess both the prediction accuracy of the forecasting module and the effectiveness of symbolic decision-making in minimizing offloading failures.

A. Experimental Setup

The experiments are conducted using publicly available datasets. Satellite throughput quality is derived from [39], while CPU load dynamics are obtained from the cloud workload datasets [40], rescaled to emulate the behavior of processing units onboard satellite nodes. The decision process is evaluated over a horizon of T time slots, with a history window of length l used for forecasting.

We benchmark our approach against two baselines:

- **LSTM-based predictor:** a task-specific model trained on the historical data for each satellite, followed by a threshold-based offloading policy;
- **Reinforcement learning:** a model-free agent trained to learn the offloading policy directly through trial-and-error interactions with the environment. This module is based on LSTM forecasting to compare the traditional deep learning paradigm with hybrid inductive-deductive systems.

Our proposed framework combines a zero-shot forecasting model (TimeGPT) with a symbolic controller based on fuzzy rules, enabling efficient and interpretable offloading decisions without the need for task-specific training or reinforcement signals.

As previously anticipated, we evaluate the forecast accuracy using standard error metrics, i.e., MAE and MAPE. In addition, we also measure the Root Mean Squared Error (RMSE), quantifying the average magnitude of prediction errors, defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2} \tag{8}$$

where $\hat{y}_i$ are the predictions, y_i the true values and N the sample size.

In addition, we assess the quality of symbolic decisions by computing the cumulative number of offloading failures, here referred to as decision accuracy. A failure occurs when the system either misses a valid offloading opportunity or selects a satellite that is unable to process the task.

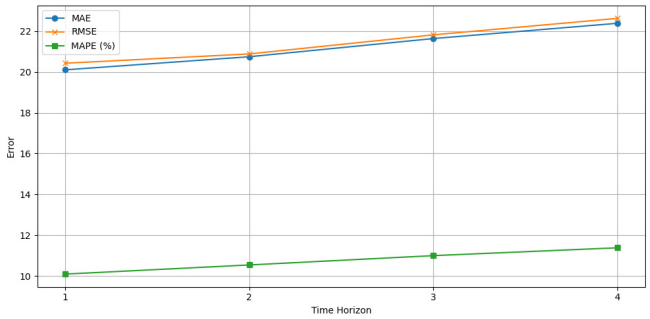
B. Results and Discussion

We first show the accuracy errors of the forecasting models considered in Fig. 3. On the one hand, Fig. 3a shows the forecast performance of the LSTM model trained on the available data. The LSTM baseline is trained using a limited dataset whose size corresponds to the same historical context window (i.e., the last l observations) provided as input to TimeGPT in zero-shot mode, in order to ensure a fair and transparent comparison between task-specific training and zero-shot inference. Due to the limited dataset size, the LSTM fails to achieve accurate predictions, even in the short-term scenario with a time horizon equal to 1. In contrast, Fig. 3b reports the zero-shot forecast performance of the foundational model as the prediction horizon increases. Compared to Fig. 3a, the foundational model demonstrates higher predictive accuracy, despite being evaluated on the same number of samples. This highlights its robustness in data-scarce scenarios, making it particularly suitable for applications where time series are still in the early stages of collection or where large-scale data storage is not feasible.

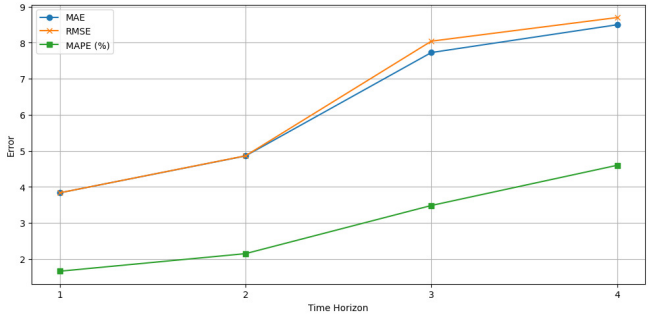
Further evidence supporting this observation is provided in Fig. 4, which compares the execution times of the two predictive models. As shown, TimeGPT not only achieves higher forecast accuracy with the same amount of data, but also exhibits significantly lower inference time. This performance gap is primarily attributed to the zero-shot paradigm of foundational models, which eliminates the need for task-specific training. Similarly, Fig. 5 shows the completion time, expressed as the sum of the execution time on board the CPU and the RTT, assuming that the LSTM queries the ground base station to perform training and inference.

Fig. 6 shows the membership functions employed in the fuzzy inference system. Specifically, Fig. 6a represents the linguistic modeling of satellite link throughput quality, Fig. 6b refers to the CPU load on the satellite node, and Fig. 6c defines the urgency of offloading decisions.

In a fuzzy logic framework, a membership function assigns to each crisp input value a degree of membership to predefined linguistic terms (such as “low”, “medium”, and “high”), allowing smooth transitions and robust reasoning under uncertainty. The offloading urgency reflects the system’s need to delegate



(a) Accuracy errors for forecasting via LSTM



(b) Accuracy errors for forecasting via timeGPT

Fig. 3. Comparison of forecasting accuracy for the considered models. The figure reports the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) obtained by the LSTM baseline and the TimeGPT model over different prediction horizons, highlighting the performance gap between task-specific training and zero-shot forecasting.

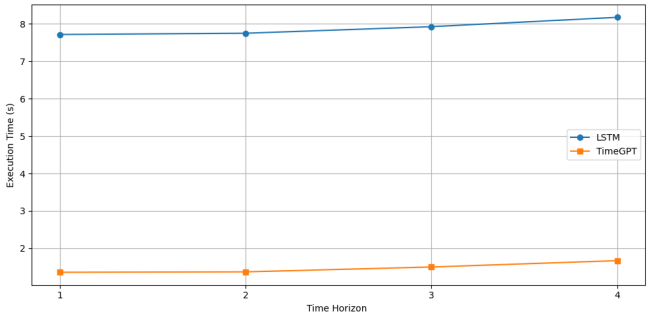


Fig. 4. Execution time as a function of the prediction horizon for the LSTM baseline and the TimeGPT model, highlighting the computational cost of task-specific training and inference for LSTM compared to the zero-shot inference paradigm adopted by TimeGPT.

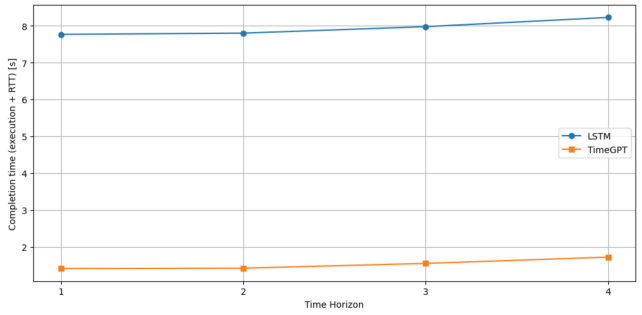
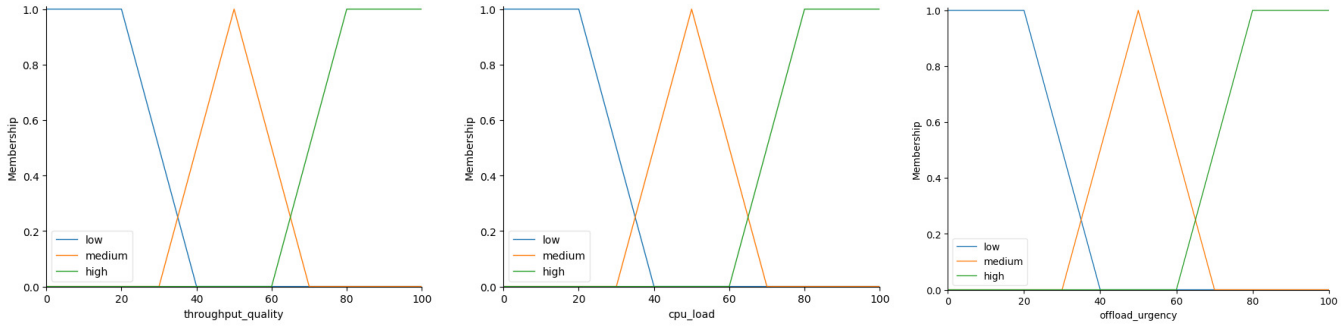


Fig. 5. Completion time (execution time + RTT) as a function of the prediction horizon for the LSTM baseline and the TimeGPT model, highlighting the computational cost of task-specific training and inference for LSTM compared to the zero-shot inference paradigm adopted by TimeGPT.

computational tasks, depending on the current channel and CPU conditions; higher values indicate more favorable situations for offloading. From the graphs, it can be observed that all input and output variables are partitioned using standard trapezoidal and triangular membership functions. The terms "low" and "high" are modeled as trapezoids, ensuring gradual activation and deactivation at the extremes of the domain, while the term "medium" is represented by a triangle centered in the mid-range. This choice enables a balanced overlap between adjacent terms and ensures continuous coverage across the entire input space. The thresholds for transitions between categories (e.g., low-to-medium and medium-to-high) are symmetric and occur at values around 30, 50, and 70, consistently across all variables.

Fig. 7 reports the decision accuracy as a function of the prediction horizon for four different combinations of forecasting and decision-making paradigms, namely *LSTM + RL*, *LSTM + Fuzzy*, *TimeGPT + RL*, and *TimeGPT + Fuzzy*. This comparison enables disentangling the individual contributions of the forecasting module and the decision layer. As expected, all approaches exhibit a gradual degradation in decision accuracy as the prediction horizon increases, reflecting the growing uncertainty associated with longer-term forecasts. However, clear performance differences emerge across the considered architectures. Approaches relying on LSTM-based forecasting consistently achieve lower accuracy than their TimeGPT-based counterparts, regardless of the decision mechanism employed. This confirms that forecasting quality plays a dominant role in sustaining reliable offloading decisions over extended horizons. Comparing the decision layers, fuzzy logic consistently outperforms reinforcement learning when coupled with the same forecasting model. In particular, the *LSTM + Fuzzy* configuration yields higher accuracy than *LSTM + RL*, highlighting the robustness of rule-based reasoning in data-scarce and uncertain regimes. A similar trend is observed when TimeGPT is used as the forecasting backbone, where *TimeGPT + Fuzzy* systematically outperforms *TimeGPT + RL* across all horizons. Overall, the proposed *TimeGPT + Fuzzy* architecture achieves the highest and most stable decision accuracy, maintaining values above 93% even at the longest prediction horizon. These results, combined with those depicted in Fig. 3, indicate that combining high-quality zero-shot forecasting with symbolic, interpretable decision-making yields superior robustness compared to fully inductive pipelines. The analysis further suggests that both components are essential: while accurate forecasting is necessary to preserve performance at longer horizons, the adoption of deductive reasoning mechanisms provides additional resilience against prediction uncertainty.

Fig. 8 provides a sensitivity analysis of the forecasting performance with respect to the size of the available training dataset. The figure highlights a markedly different behavior between task-specific learning and zero-shot forecasting paradigms. The LSTM-based model exhibits a strong dependence on data availability, with high prediction errors in low-data regimes and a gradual improvement as the dataset size increases. This trend reflects the intrinsic reliance of inductive models on sufficient historical data to learn accurate temporal representations. Conversely, TimeGPT main-



(a) Membership function of throughput quality (b) Membership function of CPU load (c) Membership function of offloading urgency

Fig. 6. Membership functions for the considered quality metrics, namely satellite link throughput quality, satellite CPU load, and offloading urgency, used by the fuzzy logic controller to translate continuous predictions into linguistic variables.

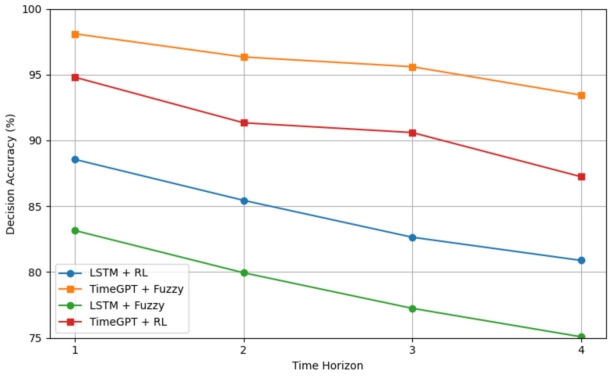


Fig. 7. Decision accuracy of the offloading process, comparing the LSTM+RL baseline with the proposed TimeGPT+Fuzzy Logic framework. The figure reports the accuracy of the decision-making process over different prediction horizons, highlighting the benefits of combining zero-shot forecasting with symbolic reasoning.

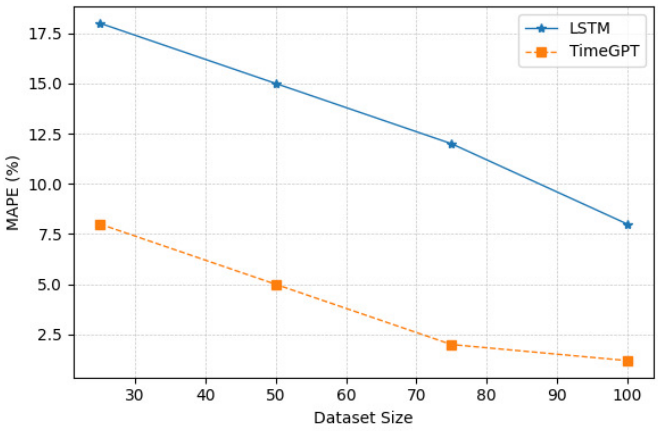


Fig. 8. Mean Absolute Percentage Error (MAPE) as a function of the dataset size for LSTM and TimeGPT, i.e., as a function of the size window l.

tains consistently low MAPE values across all dataset sizes, showing only marginal performance variations as more data become available. This result confirms the robustness and data efficiency of foundation models operating in zero-shot mode, and demonstrates their suitability for satellite-enabled environments where data collection is limited or progressively evolving.

To evaluate the behavior of the proposed framework in a multi-user scenario, we extend the single-user datasets by generating multiple synthetic IoT node perspectives through temporal shifting. Specifically, for each IoT node and satellite, we apply a random temporal offset to the original time series data, independently sampled from a uniform distribution in the range $[1, 8]$ time steps. This strategy preserves the underlying statistical properties of the data while introducing realistic temporal desynchronization among users.

The same datasets used in the single-user case (i.e., WetLinks [39] for link quality and [40] for CPU usage) are reused as the basis for these shifted instances. Forecasts for both link quality $\hat{Q}_{k,i}(t+1)$ and CPU load $\hat{C}_i(t+1)$ are then generated for each IoT node and satellite based on shifted time windows.

Once all predictions are computed, the matching algorithm described in Section IV-E is executed to determine the assignment of users to satellites, taking into account the fuzzy urgency scores and the computational demands $\chi_k(t)$ of each user. This approach enables a scalable evaluation of the system performance under realistic multi-user contention and temporal variability.

In order to assess how well the proposed matching algorithm aligns with user preferences, we introduce the High Satisfaction Rate (HSR) as an additional performance metric. This metric captures the fraction of users that are assigned to satellites they consider highly suitable, according to the urgency score produced by their individual fuzzy controllers. Specifically, we define a user as *satisfied* if the offloading urgency score $\phi_{k,i_k}(t+1)$, calculated for the assigned satellite S_{i_k}, exceeds a predefined threshold θ. The HSR is then defined as the ratio of satisfied users to the total number of users.

Fig. 9 shows the HSR as a function of the number of users demanding offloading, assuming the presence of an average number of 10 satellites in line-of-sight. As the number of users increases, the HSR tends to decrease due to resource contention; however, the *TimeGPT + Fuzzy* approach consistently outperforms the baseline *LSTM + RL* for all load conditions. In particular, the proposed method exhibits a slower degradation rate, suggesting a greater ability to preserve user satisfaction

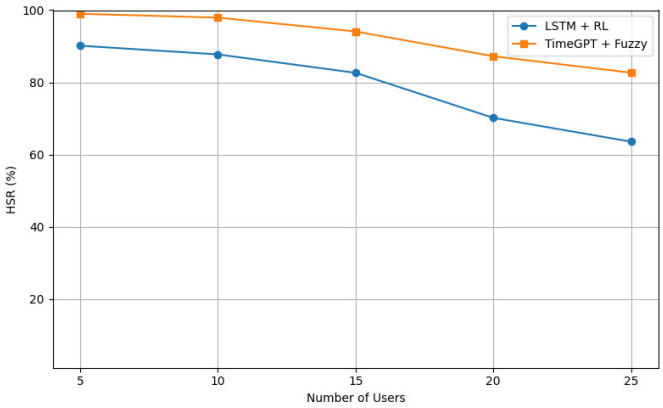


Fig. 9. High Satisfaction Rate (HSR) as a function of the number of users requesting offloading, assuming an average of 10 satellites in line-of-sight. The figure compares the proposed TimeGPT + Fuzzy matching approach against the LSTM + RL baseline, showing that the proposed method consistently achieves higher user satisfaction and degrades more gracefully under increasing resource contention.

even under competitive scenarios. The superior HSR of our TimeGPT + Fuzzy approach demonstrates its effectiveness not only in resource allocation but also in maintaining a higher Quality of Service (QoS) for a larger fraction of IoT applications, which is critical in scenarios with diverse service level agreements. This highlights the benefit of combining predictive accuracy with interpretable decision-making in multi-user environments.

VI. Conclusions

In this work, we presented a novel neuro-symbolic framework for offloading decisions in satellite-enabled IoT edge intelligence scenarios. By combining the forecasting capabilities of time-series foundation models with the transparency of rule-based reasoning, our approach enables data-efficient and explainable decision making under uncertainty. Using generative models such as TimeGPT in a zero-shot setting, we demonstrated the ability to predict both satellite link quality and computational availability without requiring task-specific training. These predictions, when processed through a fuzzy logic controller, produce interpretable and context-aware offloading decisions tailored to expected environmental conditions. Our evaluation shows that the proposed hybrid design achieves high accuracy and robustness while maintaining low inference costs and intrinsic explainability. Compared to standard approaches relying on reinforcement learning, our framework avoids the overhead of iterative training while offering a clear understanding of the underlying decision logic.

As part of future research, we plan to extend the proposed framework to explicitly account for time-varying network conditions by integrating a lifelong learning mechanism able to continuously update the decision model. This extension will allow the system to adapt to evolving traffic patterns, resource availability, and channel dynamics over time. In addition, we plan to investigate adaptive tuning strategies for the fuzzy membership functions and rule base, enabling the controller to better capture non-linear dynamics and heterogeneous task requirements in more complex satellite IoT scenarios. In particular, learning-based calibration of symbolic decision layers represents a promising direction to improve scalability and generalization while preserving interpretability.

References

[1] J. Liu, X. Du, J. Cui, M. Pan, and D. Wei, "Task-oriented intelligent networking architecture for the space–air–ground–aqua integrated network," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5345–5358, 2020.

[2] S. S. Shinde, D. Naseh, T. DeCola, and D. Tarchi, "A distributed task allocation methodology for edge computing in a LEO satellite IoT context," in *2025 12th Advanced Satellite Multimedia Systems Conference and the 18th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, Sitges, Spain, Feb. 2025, pp. 1–7.

[3] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[4] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Zomaya, "Edge intelligence: The confluence of edge computing and artificial intelligence," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7457–7469, 2020.

[5] S. S. Shinde and D. Tarchi, "Hierarchical reinforcement learning for multi-layer multi-service non-terrestrial vehicular edge computing," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 1045–1061, 2024.

[6] S. Xu, C. Kurisummoottil Thomas, O. Hashash, N. Muralidhar, W. Saad, and N. Ramakrishnan, "Large Multi-Modal Models (LMMs) as universal foundation models for AI-native wireless systems," *IEEE Network*, vol. 38, no. 5, pp. 10–20, 2024.

[7] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 2019.

[8] Y. Lin *et al.*, "Integrating satellites and mobile edge computing for 6G wide-area edge intelligence: Minimal structures and systematic thinking," *IEEE Netw.*, vol. 37, no. 2, pp. 14–21, 2023.

[9] S. Long *et al.*, "6G comprehensive intelligence: Network operations and optimization based on Large Language Models," *IEEE Netw.*, vol. 39, no. 4, pp. 192–201, 2025.

[10] Q. Guo, F. Tang, and N. Kato, "Resource allocation for aerial assisted digital twin edge mobile network," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3070–3079, 2023.

[11] B. Mao, J. Liu, Y. Wu, and N. Kato, "Security and privacy on 6G network edge: A survey," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1095–1127, 2023.

[12] X. Chen, L. Luo, F. Tang, M. Zhao, and N. Kato, "AIGC-based evolvable digital twin networks: A road to the intelligent metaverse," *IEEE Netw.*, vol. 38, no. 6, pp. 370–379, 2024.

[13] C. Thomas, C. Chaccour, W. Saad, M. Debbah, and C. S. Hong, "Causal reasoning: Charting a revolutionary course for next-generation AI-native wireless networks," *IEEE Veh. Technol. Mag.*, vol. 19, no. 1, pp. 16–31, 2024.

[14] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *ArXiv*, 2021. [Online]. Available: https://crfm.stanford.edu/assets/report.pdf

[15] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans, "Foundation models for decision making: Problems, methods, and opportunities," *arXiv preprint arXiv:2303.04129*, 2023. [Online]. Available: https://arxiv.org/abs/2303.04129

[16] Y. Lin *et al.*, "Mitigating the alignment tax of RLHF," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, Nov. 2024, pp. 580–606.

[17] A. Garza, C. Challu, and M. Mergenthaler-Canseco, "Timegpt-1," *arXiv preprint arXiv:2310.03589*, 2024. [Online]. Available: https://arxiv.org/abs/2310.03589

[18] L. Tan, S. Guo, Z. Kuang, P. Zhou, and M. Li, "SkyLink: Joint Deployment and Scheduling in Collaborative Integrated Ground-Air-Space Network," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 90–106, 2026.

[19] Y. Li *et al.*, "Computation Offloading in Delay-Sensitive Multi-Satellite Cooperative Edge Computing Systems," *IEEE Internet Things J.*, vol. 13, no. 1, pp. 123–137, 2026.

[20] M. Dai, S. Chang, Y. Wang, and Z. Su, "Energy-efficient multi-access edge computing for heterogeneous satellite-maritime networks: A hybrid harvesting-and-offloading design," *IEEE Trans. Mobile Comput.*, vol. 24, no. 11, pp. 12 001–12 018, 2025.

[21] X. Tang, Y. Jiang, R. Zhang, Q. Du, J. Liu, and N. Liu, "Energy-efficient integrated communication and computation via non-terrestrial networks with uncertainty awareness," *IEEE Internet Things J.*, vol. 12, no. 17, pp. 35 165–35 178, 2025.
[22] A. Li, T. Zhou, T. Xu, Y. Ouyang, H. Hu, and C. Wu, "LEO satellite assisted edge computing with latency and energy optimization," *IEEE Trans. Netw. Sci. Eng.*, vol. 12, no. 4, pp. 2640–2653, 2025.
[23] L. Wang, J. Li, M. Dai, and H. Zhang, "Low-earth-orbit satellite assisted edge computing for vehicular networks: A task priority-based delay minimization approach," *IEEE Internet Things J.*, vol. 12, no. 17, pp. 35 482–35 496, 2025.
[24] L. Zhao *et al.*, "A region division-based adaptive task offloading in collaborative LEO heterogeneous constellation," *IEEE Internet Things J.*, vol. 12, no. 14, pp. 26 523–26 537, 2025.
[25] O. Chukhno, N. Chukhno, A. Ometov, S. Pizzi, G. Araniti, and A. Molinaro, "Application-driven offloading of XR mission critical via integrated TN/NTN," *IEEE Netw.*, pp. 1–1, 2025, early access, DOI:10.1109/MNET.2025.3572214.
[26] A. Bonora, A. Traspadini, M. Giordani, and M. Zorzi, "Performance Evaluation of Satellite-Based Data Offloading on Starlink Constellations," in *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, Milan, Italy, 2025, pp. 1–6.
[27] Y. Chen *et al.*, "SLICE: Energy-efficient satellite-ground co-inference via layer-wise scheduling optimization," *IEEE Trans. Serv. Comput.*, vol. 12, no. 14, pp. 26 523–26 537, 2025.
[28] H. Wu, L. Tian, H. Tang, R. Li, and P. Jiao, "Graph convolutional reinforcement learning-guided joint trajectory optimization and task offloading for aerial edge computing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 10, pp. 17 487–17 498, 2025.
[29] Z. Lin, M. Lin, T. de Cola, J.-B. Wang, W.-P. Zhu, and J. Cheng, "Supporting iot with rate-splitting multiple access in satellite and aerial-integrated networks," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 123–11 134, 2021.
[30] Z. Lin, Z. Feng, K. Guo, A. Nauman, D. Niyato, and J. Wang, "Ai-driven seamless and massive access in space-air-ground integrated networks," *IEEE Wireless Communications*, vol. 32, no. 3, pp. 72–79, 2025.
[31] Y. He, Y. Xiao, S. Zhang, M. Jia, and Z. Li, "Direct-to-smartphone for 6g ntn: Technical routes, challenges, and key technologies," *IEEE Network*, vol. 38, no. 4, pp. 128–135, 2024.
[32] Y. He, B. Sheng, H. Yin, D. Yan, and Y. Zhang, "Multi-objective deep reinforcement learning based time-frequency resource allocation for multi-beam satellite communications," *China Communications*, vol. 19, no. 1, pp. 77–91, 2022.
[33] S. Tran, E. Mota, and A. d'Avila Garcez, "Reasoning in neurosymbolic AI," *arXiv preprint arXiv:2505.20313*, 2025. [Online]. Available: https://arxiv.org/abs/2505.20313
[34] A. Sheth, K. Roy, and M. Gaur, "Neurosymbolic Artificial Intelligence (Why, What, and How)," *IEEE Intell. Syst.*, vol. 38, no. 03, pp. 56–62, May 2023.
[35] K. Meghraoui, T. Racharak, K. Ait El Kadi, S. Bensiali, and I. Sebari, "A new integrated neurosymbolic approach for crop-yield prediction using environmental data and satellite imagery at field scale," *Artificial Intelligence in Geosciences*, vol. 6, no. 1, p. art. no. 100125, 2025.
[36] R. Zhang *et al.*, "Generative AI agents with large language model for satellite networks via a mixture of experts transmission," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3581–3596, 2024.
[37] J. Du, H. Wang, C. Jiang, J. Simonjan, J. Wang, and M. Debbah, "Distributed AI-Based Secure Communications in Space-Air-Ground-Sea Integrated Networks," *IEEE Commun. Mag.*, vol. 63, no. 7, pp. 48–55, 2025.
[38] X. Wu, F. Teng, X. Li, J. Zhang, T. Li, and Q. Duan, "Out-of-distribution generalization in time series: A survey," 03 2025.
[39] D. Laniewski, E. Lanfer, B. Meijerink, R. van Rijswijk-Deij, and N. Aschenbruck, "Wetlinks: A large-scale longitudinal starlink dataset with contiguous weather data," in *2024 8th Network Traffic Measurement and Analysis Conference (TMA)*, Dresden, Germany, 2024, pp. 1–9.
[40] B. S. Akhil, "Cloud workload," Kaggle Dataset, 2022, accessed: July 18, 2025. [Online]. Available: https://www.kaggle.com/datasets/akhilbs/cloud-workload

Benedetta Picano (Member, IEEE) received the B.S. degree in computer science, the M.Sc. degree in computer engineering, and the Ph.D. degree in information engineering from the University of Florence, Firenze, Italy, in 2013, 2016, and 2020, respectively. She was a Visiting Researcher with the University of Houston, Houston, TX, USA. Her research focuses on resource allocation strategies that combine model-based approaches—such as matching theory, auction mechanisms, and stochastic optimization—with data-driven methods based on foundation models and emerging neuro-symbolic techniques. These methods are applied to heterogeneous edge–cloud–space environments to enable context-aware, adaptive, and human-centric network operations.

Daniele Tarchi Daniele Tarchi (Senior Member, IEEE) was born in Florence, Italy, in 1975. He received the M.Sc. degree in telecommunications engineering and the Ph.D. degree in informatics and telecommunications engineering from the University of Florence, Florence, Italy, in 2000 and 2004, respectively.

From 2004 to 2010, he was a Research Associate at the University of Florence. From 2010 to 2019, he was an Assistant Professor and from 2019 to 2024, he was an Associate Professor at the University of Bologna, Bologna, Italy. He is currently an Associate Professor at the University of Florence, Florence, Italy. He is the author of numerous published articles. His research interests include wireless communications and networks, satellite communications and networks, edge computing, distributed learning, and optimization techniques.

Prof. Tarchi is an Editorial Board member for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, and IET Communications. He has served as a Symposium Co-Chair for IEEE WCNC 2011, IEEE Globecom 2014, IEEE Globecom 2018, and IEEE ICC 2020, and as a Workshop Co-Chair at IEEE ICC 2015 and IEEE Globecom 2024.