Hybrid Model-Aided Learning for 5G-NTN Handover in High-Mobility Platforms

Ons Aouedi, Flor Ortiz, Eva Lagunas, Thang X. Vu, Symeon Chatzinotas
SnT, University of Luxembourg, Luxembourg
Corresponding Author: ons.aouedi@uni.lu

***Abstract*—Ubiquitous 5G/6G network access increasingly relies on non-terrestrial Networks (NTN), yet the mobility patterns of Low Earth Orbit (LEO) satellite constellations create significant challenges for seamless connectivity. In existing studies, handover management in NTN environments has been handled using heuristic-based approaches or deep Q-learning (DQN) models, which often lack the foresight needed to anticipate mobility changes, resulting in frequent handovers and connectivity disruptions. To address these limitations, we propose a hybrid model-aided learning framework that combines a transformer-based predictive model with reinforcement learning (RL) to manage handovers in real-time adaptively. By introducing a short prediction horizon from the transformer model before applying an advantage actor-critical (A2C) RL approach, our framework reduces handover frequency and accelerates convergence. Numerical results validate the effectiveness of this approach, showing higher rewards, higher demand satisfaction, greater stability, and enhanced efficiency compared to DQN-based methods and RL without predictive components.**

***Index Terms*—Non-Terrestrial Networks, Handover Management, Transformer Models, Reinforcement Learning, Deep Learning.**

I. Introduction

The rapid evolution of communication technologies, coupled with the increasing demand for seamless connectivity, has led to the development of non-terrestrial networks (NTN), such as Low Earth Orbit (LEO) satellite networks. These networks play a critical role in expanding global communication coverage, especially in remote and underserved areas where terrestrial infrastructure is limited [1]. One of the most pressing challenges in these networks is managing handovers, particularly when dealing with highly dynamic environments involving both satellites and mobile ground or airborne devices, such as airplanes or high-speed trains. In NTN environments, frequent handovers occur due to the high mobility of both satellites and devices [2]. If not handled efficiently, these frequent handovers can result in communication disruptions, increased latency, reduced quality of service (QoS), and excessive signaling overhead [3]. Effective handover management is thus critical to ensure reliable and high-quality communication in NTNs [4]. It requires predictive, proactive decision-making to anticipate changes in coverage, resource availability, and congestion levels.

Based on the available literature, handover management in satellite networks has relied on heuristic-based methods or Deep Q-learning (DQN) approaches [5], which often fail to adequately anticipate the complex, dynamic nature of airplane trajectories and satellite mobility patterns. These methods typically struggle with frequent handovers, increased latency, and resource inefficiency due to their limited ability to capture long-term dependencies and adapt proactively to changing conditions.

Consequently, in this paper, we propose a hybrid model-aided learning framework that integrates a transformer-based model for predicting airplane trajectories with a deep reinforcement learning (DRL) agent using an advantage actor-critic (A2C) approach [6]. Our approach addresses these limitations through a transformer model to capture both short-term and long-term dependencies in airplane trajectory data, allowing for accurate predictions of future positions. This predictive capability is crucial for proactive decision-making in dynamic satellite environments. The DRL agent, specifically employing the A2C algorithm, further enhances the framework by effectively handling continuous decision-making under uncertainty. Unlike conventional models, the A2C agent optimizes satellite handovers by balancing exploration and exploitation through both policy-based and value-based learning, thereby reducing unnecessary handovers and improving overall network performance. This dual approach allows for more robust and adaptive handover decisions, thus addressing the shortcomings in response times and predictive accuracy found in state-of-the-art methods.

II. Related Work

Satellite handover schemes have been explored in the literature. For example, the authors in [7] addressed the satellite handover challenge in LEO satellite networks. They proposed a multi-agent Q-learning approach aimed at minimizing the average number of handovers for terrestrial User Equipment (UE). Their state representation incorporated key features, including the coverage of LEO satellites, the availability of communication channels for each satellite, and the corresponding service duration of each satellite. A multi-agent RL framework was proposed in [8], where multiple UE cooperatively optimize the number of handovers in the whole network considering different handover criteria. Moreover, a load-balancing energy-aware satellite handover has been proposed in [5]. The proposed handover solution addresses the handover in satellite-terrestrial integrated networks with UEs with different and variable performance requirements.

This work was supported by the Luxembourg National Research Fund (FNR) under the project SmartSpace (C21/IS/16193290).

Additionally, recent advancements have been made in DRL approaches to optimize handover decisions in highly dynamic satellite environments. A DRL-based strategy for satellite communications is described in [9], which significantly reduces signaling overhead in LEO satellite networks. A machine learning-based solution focusing on NTN systems is discussed in [10], where machine learning techniques are applied to enhance handover decisions and reduce signaling storms in NTNs. Similarly, the work in [11] introduces a load-aware multi-agent RL approach that balances the network load during satellite handovers. Furthermore, comprehensive handover strategies tailored for various orbital configurations like LEO, MEO, and HEO are explored in [12].

While the existing approaches, such as those discussed in [5] - [12], have significantly advanced handover management in satellite communication networks, they predominantly focus on optimizing specific aspects such as minimizing handovers or reducing signaling storms. These models often overlook the complex interplay between the high mobility of satellites and the dynamic nature of their communication environment, leading to potential inefficiencies under varying operational conditions. Furthermore, the application of DRL in these works generally lacks integration with predictive models that can proactively manage satellite handovers based on the anticipated trajectory and behavior of high-speed mobile platforms.

III. System Model

A. Satellite Network Model

We consider a constellation of N LEO satellites providing communication services to high-mobility platforms. In this work, and without loss of generality, we focus on airplanes as the user terminals. The satellites move in predetermined orbits, with their positions at any time t determined by their orbital parameter. Each satellite is characterized by:

- **Position and Mobility:** The position of satellite n at time t is given by $\text{SatPos}_{n,t} = \{\text{lat}_{n,t}, \text{lon}_{n,t}, \text{alt}_{n,t}\}$, where lat, lon, and alt represent latitude, longitude, and altitude, respectively.
- **Communication coverage:** Each satellite has a coverage area determined by its altitude and antenna maximum pointing angle. A satellite can provide service to an airplane if the elevation angle $\theta_{k,n,t}$ between the satellite and airplane k is above a minimum threshold θ_{min}. In our model, we set $\theta_{\text{min}} = 20^\circ$ [13].We assume overlapping satellite coverage allows multiple satellites to serve the same airplane. Additionally, Earth-moving beams dynamically adjust to follow the airplane's position, ensuring continuous connectivity.
- **Resource Capacity:** Satellites have limited communication resources. The congestion level of satellite n at time t is $\text{cong}_{n,t} \in [0, 1]$, where 1 indicates full utilization.

B. Airplane Model

We consider a set of K airplanes equipped with satellite communication terminals. Each airplane is characterized by:

- **Position and Mobility:** The position of airplane k at time t is given by $\text{PlanePos}_{k,t} = \{\text{lat}_{k,t}, \text{lon}_{k,t}, \text{alt}_{k,t}\}$. Airplane trajectories are dynamic and need to be predicted for future timesteps.
- **Communication Demand:** Each airplane has a communication demand $\text{dem}_{k,t} \in [d_{\text{min}}, d_{\text{max}}]$, representing the fraction of a satellite's resource capacity required by the airplane at time t.
- **Coverage by Satellites:** At any time t, airplane k can be within the coverage areas of a subset of satellites, denoted as $\text{CoverSat}_{k,t} \subseteq \{1, 2, \ldots, N\}$.

C. Communication Parameters

1) Elevation Angle: The elevation angle $\theta_{k,n,t}$ between airplane k and satellite n at time t is a critical parameter affecting the communication quality. It is calculated based on the positions of the airplane and satellite.

2) Resource Allocation and Congestion: Resource allocation $alloc_{k,n,t}$ represents the amount of satellite n's resources allocated to airplane k at time t, and is determined by two factors: the airplane's communication demand $dem_{k,t}$ and the available capacity of the satellite, which depends on the congestion level $cong_{n,t}$.

The congestion $cong_{n,t}$ is defined as the fraction of the satellite's total capacity currently in use. A fully used satellite corresponds to $cong_{n,t} = 1$, while an idle satellite has $cong_{n,t} = 0$.

The resource allocation formula is given as:

$$\text{alloc}_{k,n,t} = \min\left(\text{dem}k, t, 1 - \text{cong}_{n,t}\right), \tag{1}$$

where $\text{alloc}k, n, t$ ensures that the resource allocated to airplane k does not exceed either its demand $\text{dem}k, t$ or the satellite's remaining capacity $1 - \text{cong}_{n,t}$. This balance prevents over-allocation and ensures equitable resource distribution among airplanes served by the satellite.

After allocation, the congestion level of satellite n is updated:

$$\text{cong}_{n,t+1} = \text{cong}_{n,t} + \text{alloc}_{k,n,t} \tag{2}$$

3) Quality of Service (QoS) and Reward Function: QoS experienced by an airplane k when connected to a satellite n at a given time t is influenced by several factors, including the elevation angle $\theta_{k,n,t}$ (higher elevation angle is associated with better signal quality), resource allocation (if the selected satellite is able to accommodate all its demand), and satellite congestion (if the satellite will be close to saturation when admitting the new user). The QoS and the corresponding reward are calculated using several functions based on the elevation angle.

a) QoS Calculation: The QoS is defined as follows:

$$\text{QoS}_{k,n,t} = \left(\frac{\theta_{k,n,t}}{\theta_{\text{max}}}\right)^{1.5} \times \left(\frac{\text{alloc}_{k,n,t} + 0.1}{\text{dem}_{k,t} + 0.1}\right) \times (1 - \text{cong}_{n,t}), \tag{3}$$

where 1.5 amplifies the importance of higher elevation angles, $alloc_{k,n,t} + 0.1$ ensures that resource allocations are not overly sensitive to minor variations in allocation values, and $dem_{k,n,t} + 0.1$ ensures numerical stability when demand $dem_{k,n,t}$ is close to zero.

The QoS is clipped to the range $[0, 1]$.

$$\text{QoS}_{k,n,t} = \min\left(\max\left(\text{QoS}_{k,n,t}, 0.0\right), 1.0\right) \tag{4}$$

D. Handover Mechanism

Airplanes can decide to maintain the current satellite connection or perform a handover to another satellite in their coverage. The handover decision can be impacted by:

- **Future State:** Predicted positions of the airplane allowing anticipation of coverage changes.
- **QoS Optimization:** Balancing the trade-off between maintaining QoS and minimizing handover frequency.
- **Resource Availability:** Considering satellite congestion levels to ensure sufficient resources are available.

E. Constraints

To ensure efficient resource allocation and system stability, we define the following constraints in our problem formulation:

- **Resource Capacity Constraint**: This constraint ensures that the total resources allocated by a satellite do not exceed its available capacity at any given time, which is stated as follows:

$$\sum_{k} \text{alloc}_{k,n,t} \leq 1, \quad \forall n, t \tag{5}$$

  where $\text{alloc}_{k,n,t}$ represents the proportion of resources allocated by satellite n to airplane k at time t. This constraint prevents overloading a satellite and ensures fair resource distribution.
- **Elevation Angle Constraint**: To maintain effective communication, the elevation angle $\theta_{k,n,t}$ between an airplane and a satellite must exceed a minimum threshold θ_{min}.

$$\theta_{k,n,t} \geq \theta_{\text{min}}, \quad \forall k, n, t \tag{6}$$

These constraints are critical for ensuring the stability and efficiency of the satellite communication system. They balance resource utilization, maintain communication quality, and avoid congestion or under-utilization of satellite resources.

IV. Proposed Hybrid Model-Aided Learning Framework

As illustrated in Fig. 1, to predict future airplane positions $\hat{\text{PlanePos}}_{k,t+n}$ based on historical trajectory data, we use a transformer-based prediction model. Then, a DRL model has been proposed to learn the optimal policy π^* using the augmented state that includes predicted future positions and satellite mobility information. By integrating these components, the framework enables proactive handover decisions that maximize QoS while minimizing unnecessary handovers and avoiding satellite congestion.

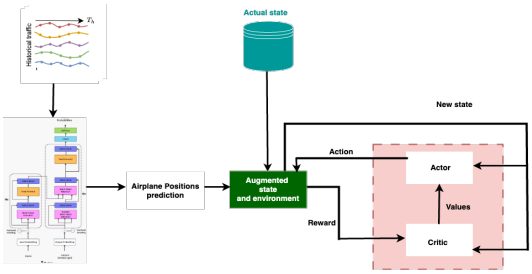


Fig. 1: The proposed framework.

A. Transformer-Based Prediction Model

To address the challenge of accurately predicting future airplane positions, we use a transformer-based prediction model, which is well-suited for handling sequential data and capturing long-term dependencies. The ability to anticipate airplane trajectories is critical for enabling proactive handover decisions that minimize connection disruptions and optimize satellite resource allocation.

1) Model Description: The transformer model processes a sequence of historical airplane positions as input and predicts the airplane's position at a future time step. Specifically, the input sequence is given by:

$$\mathcal{X}_k = \{\text{PlanePos}_{k,t-l}, \ldots, \text{PlanePos}_{k,t}\}, \tag{7}$$

where l represents the number of historical time steps considered, and $\text{PlanePos}_{k,t}$ includes the latitude, longitude, and altitude of airplane k at time t. The model outputs the predicted position of the airplane at time $t + l$, denoted as:

$$\hat{\text{PlanePos}}_{k,t+l} = f_\theta(\mathcal{X}_k), \tag{8}$$

where f_θ represents the transformer model parameterized by θ.

2) Architectural Advantages: The transformer architecture leverages the *self-attention mechanism* to model complex dependencies within the input sequence, allowing it to:

- **Capture temporal dynamics:** The self-attention mechanism assigns dynamic weights to different time steps, enabling the model to focus on relevant trajectory points that influence the airplane's future position.
- **Handle long-term dependencies:** Unlike recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, the transformer avoids the vanishing gradient problem and efficiently captures long-term motion trends.
- **Facilitate parallelization:** The transformer's parallel processing capabilities significantly reduce the computational overhead during training and inference compared to sequential models.

3) Training Objective: The model is trained to minimize the prediction error between the actual future position of the airplane and the predicted position. The Mean Squared Error (MSE) is adopted as the loss function, defined as:

$$L = \frac{1}{K} \sum_{k=1}^{K} \left\| \hat{\text{PlanePos}}_{k,t+l} - \text{PlanePos}_{k,t+l} \right\|^2, \tag{9}$$

where K is the number of airplanes in the training dataset. The model iteratively updates its parameters using backpropagation to reduce the loss and improve prediction accuracy.

The prediction capability of the transformer model enhances the robustness of the proposed hybrid framework, reducing unnecessary handovers while ensuring high QoS. Compared to the available literature methods relying solely on instantaneous states, the transformer-based predictions extend the temporal awareness of the RL agent, leading to superior performance in dynamic high-mobility environments.

B. Deep Reinforcement Learning Agent

We leverage a DRL agent that learns an optimal handover policy in a dynamic and stochastic environment to optimize handover decisions in the satellite communication environment. The DRL agent combines real-time state information and future trajectory predictions provided by the transformer model ($S_{k,t}$) to make anticipatory and proactive decisions that maximize QoS while minimizing handover frequency and avoiding satellite congestion.

The proposed DRL agent is implemented using the A2C model, which balances policy optimization (actor) and value estimation (critic). In the A2C model, the agent maintains two neural networks: an actor-network that determines the policy and a critic network that estimates the value function. The actor network represents the policy $\pi_\theta(a_t \mid s_t)$, where θ are the parameters of the neural network. It outputs a probability distribution over actions based on the current state s_t:

$$\pi_\theta(a_t \mid s_t) = \text{softmax}(f_\theta(s_t)). \tag{10}$$

The state includes the airplane's current position, predicted future locations (from the Transformer model), satellite elevation angles, congestion levels, and historical QoS metrics, providing a comprehensive representation of the system's dynamics. The action represents either maintaining the current satellite connection or switching to a new one among the available satellites. These actions interact directly with the environment, affecting QoS, resource allocation, and handover penalties. The goal of the actor is to maximize the cumulative expected reward by selecting the best action a_t at each time step t. To note, a_t represents the handover decision, where its size equals the number of satellites plus one, corresponding to the option of maintaining the current satellite connection (i.e., no handover).

The critic network evaluates the state-value function $V_\phi(s_t)$, where ϕ are the parameters of the critic network. The state-value function quantifies the expected cumulative reward starting from state s_t under policy π:

$$V_\phi(s_t) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}) \mid s_t, \pi\right]. \tag{11}$$

The critic network provides an advantage function that measures the benefit of taking an action a_t in state s_t relative to the baseline value $V_\phi(s_t)$. This advantage is defined as:

$$A(s_t, a_t) = Q(s_t, a_t) - V_\phi(s_t), \tag{12}$$

where $Q(s_t, a_t)$ is the action-value function. The advantage guides the actor-network to prioritize actions that yield higher-than-expected returns, enabling more efficient policy updates.

The DRL agent's state space is augmented with predicted airplane positions and satellite conditions, as provided by the transformer-based model. The state $S_{k,t}$ at time t includes:

$$S_{k,t} = \left(\text{PlanePos}_{k,t}, \hat{\text{PlanePos}}_{k,t+l}, \text{dem}_{k,t}, \text{SatPos}_{n,t}, \text{cong}_{n,t}\right), \tag{13}$$

where $\hat{\text{PlanePos}}_{k,t+l}$ represents the predicted position at time $t + l$. The DRL agent interacts with the environment and observes rewards $R(s_t, a_t)$ for its actions. The reward function is designed to encourage high QoS while penalizing low-elevation angles and unnecessary handovers. It is defined as:

$$R(s_t, a_t) = \alpha \cdot \text{QoS}_{k,n,t} - \beta \cdot H(a_{k,t}) \tag{14}$$

where:

- α is a scaling factor for the QoS.
- β is a scaling factor for the handover penalty.
- $H(a_{k,t})$ is the handover indicator function:

$$H(a_{k,t}) = \begin{cases} 1, & \text{if } a_{k,t} \neq 0 \text{ (handover occurs)} \\ 0, & \text{if } a_{k,t} = 0 \text{ (no handover)} \end{cases} \tag{15}$$

The learning objective is to maximize the cumulative discounted reward G, defined as:

$$G = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right], \tag{16}$$

where $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards.

The A2C agent iteratively improves its performance using the following steps:

(a) **Policy Evaluation:** The critic network estimates the value $V_\phi(s_t)$ for the current state s_t.

(b) **Policy Improvement:** The actor network updates its policy using gradients derived from the advantage function $A(s_t, a_t)$, which measures how much better (or worse) an action is compared to the expected value:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a_t \mid s_t) A(s_t, a_t), \tag{17}$$

where α is the learning rate for the actor network.

(c) **Value Update:** The critic network parameters ϕ are updated to minimize the error between the observed return and the estimated state-value:

$$\phi \leftarrow \phi + \beta \nabla_\phi \left(R(s_t, a_t) - V_\phi(s_t)\right)^2, \tag{18}$$

where β is the learning rate for the critic network.

The A2C framework is particularly advantageous for the handover management problem because it efficiently handles large and high-dimensional state spaces, such as those augmented with predicted airplane positions. Moreover, it provides stability in learning through the combination of policy gradients (actor) and value-based updates (critic). By using

the advantage function to refine its decisions, the agent learns proactive policies that minimize unnecessary handovers, avoid satellite congestion, and maximize QoS. By incorporating future state predictions from the transformer model, the DRL agent can anticipate changes in airplane trajectories and satellite availability, enabling forward-looking handover decisions. This anticipatory capability ensures that the agent optimizes resource utilization and enhances the reliability of satellite communication systems for high-mobility platforms.

V. Simulation Results

In this section, we first describe the simulation scenario, the dataset used in our experiments, and the benchmark technique.

A. Experimental Setup

We evaluated our hybrid framework using a dataset of 8,441 instances with 256 features, including timestamps (10-second intervals), satellite information (25 satellites with latitude, longitude, altitude, elevation, and congestion), airplane positions, covering satellites, and plane demand. Satellite trajectories were modeled using TLEs from CelesTrak for precise orbit propagation. Key parameters included a learning rate of 0.0001, batch size of 256, and normalized demand bounds ($d_{\min} = 0.2, d_{\max} = 0.5$). The DRL agent is deployed at a central entity, which collects data from satellites and airplanes, predicts future positions, and computes optimal handover decisions. Comparative baselines were a DQN [13] and random policy, with an ablation study on prediction horizons (5 vs. 25) and RL models (A2C vs. actor-critic).

a) Demand Satisfaction Metrics:

- **Airplane average demand satisfaction (per episode):**

$$S_k = \frac{1}{T} \sum_{t=1}^{T} \frac{\text{alloc}_{k,n,t}}{\text{dem}_{k,t}}, \quad \text{if } \text{dem}_{k,t} > 0; \text{ else } 0 \tag{19}$$

  The satisfaction of an airplane k is the average proportion of its demand met across T timesteps.
- **Episode average demand satisfaction:**

$$\bar{S}_e = \frac{1}{M} \sum_{k=1}^{M} S_k \tag{20}$$

  The satisfaction across all M airplanes during a single episode.
- **Final average demand satisfaction (across all episodes):**

$$\text{Final average demand satisfaction} = \frac{1}{E} \sum_{e=1}^{E} \bar{S}_e \tag{21}$$

  The average satisfaction across all E episodes for evaluation.

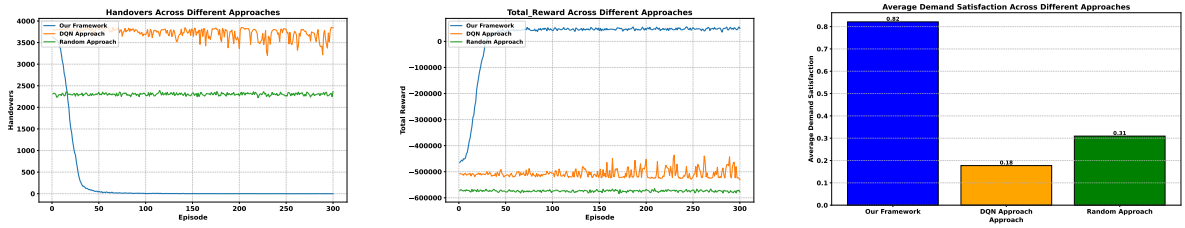
B. Results

In Fig. 2a and Fig. 2b we present a comparative analysis of the performance across different approaches, namely our framework, the DQN approach, and the random approach. It is important to note that, unlike our framework, neither the DQN nor random approaches utilize a predictive model. This distinction emphasizes the added capability of our framework to anticipate future conditions. The comparative analysis of handovers and total rewards across different approaches underscores the robustness and effectiveness of our framework relative to the DQN and random approach. In terms of handover management, our framework demonstrates a significantly reduced frequency of satellite switches, reflecting its capability to maintain stable satellite connections and minimize service interruptions. The DQN approach, while offering some level of optimization, results in a moderate number of handovers, whereas the random approach, lacking a structured policy, exhibits the highest frequency of satellite transitions. This stability advantage is further validated by the total reward outcomes, where our framework achieves consistently higher cumulative rewards across episodes, indicating its success in optimizing key objectives such as QoS and resource efficiency. The DQN approach earns moderate rewards, reflecting partial optimization, while the random approach accrues the lowest rewards, underscoring its inefficacy in strategic decision-making. In addition, Fig. 2c highlights the superior performance of our framework in achieving higher average demand satisfaction compared to the DQN and random approaches.

Consequently, the predictive capabilities of our hybrid framework yield a more balanced and strategic satellite selection process, achieving substantially higher rewards and fewer handovers compared to the DQN and random approaches, both of which lack the benefit of predicting in their decisions.

C. Ablation study

In Fig. 3 we present an ablation study to assess the impact of modifying the prediction horizon with our transformer model as well as changing the RL model within our framework. Specifically, we explore two adjustments: changing the prediction horizon from 5 to 25 steps and switching the RL model from A2C to an actor-critic model while retaining the predictive component across all configurations. It can be seen that the A2C model with a 25-step horizon demonstrates an improvement in the reward compared to the 5-step horizon. This suggests that the longer prediction horizon enables more stable decision-making, effectively balancing satellite selection over a longer time frame. Despite the added complexity, the 25-step horizon appears to improve the model's ability to anticipate changes in satellite availability and reduce unnecessary handovers. The actor-critic model with a 5-step horizon, despite matching the shorter horizon of our framework, achieves a lower reward. This difference underscores that A2C's synchronous policy and value updates more effectively capture the optimal policy, whereas the actor-critic approach, with its separate updates, yields suboptimal results in dynamic scenarios. Furthermore, in terms of convergence speed,



(a) Handovers across approaches. (b) Total reward across approaches. (c) Average demand satisfaction.

Fig. 2: Comparative analysis of handovers, cumulative total reward, and average demand satisfaction across different approaches.

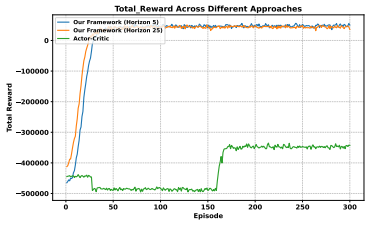


Fig. 3: Impact of prediction horizon and RL-based model on handovers.

our framework with A2C achieves the fastest convergence to optimal performance, highlighting the efficiency of A2C in learning and executing policies suited to this dynamic environment.

D. Training time and energy consumption

Table I highlights the energy consumption and training time across models. It can be seen that with a larger prediction horizon (25), the model converges more quickly, resulting in slightly lower training time (12,404 seconds) and energy consumption (0.169 Wh) compared to a horizon of 5. However, while the simple actor-critic model is more efficient in terms of energy and training time, it does not outperform our framework in terms of key performance metrics such as reward and handover optimization.

TABLE I: Energy consumption and training time.

Metric	Energy cost (Wh)	Training time (s)
Actor-Critic model	0.161	11796
Our framework (Horizon 5)	0.177	12981
Our framework (Horizon 25)	0.169	12404

VI. Conclusions

This work tackled the problem of optimizing handovers in NTN for high-mobility platforms, introducing a hybrid model-aided learning framework for NTN handover for high-mobility platforms. Our framework combines predictive modeling with RL strategies to enable real-time, adaptive decision-making in satellite handover management. Experimental results demonstrate that our framework outperforms other variants in terms of reward, handover stability, and convergence speed. Thus it offers a scalable foundation for future NTN advancements, especially in high mobility environments.

VII. Acknowledgement

This research was funded by the Luxembourg National Research Fund (FNR) under the project SmartSpace (C21/IS/16193290). For the purpose of open access, and in fulfillment of the obligations arising from the grant agreement, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission.

References

[1] F. Rinaldi, H.-L. Maattanen, J. Torsner, S. Pizzi, S. Andreev, A. Iera, Y. Koucheryavy, and G. Araniti, "Non-terrestrial networks in 5G & beyond: A survey," *IEEE access*, vol. 8, pp. 165178–165200, 2020.

[2] S. Park and J. Kim, "Trends in LEO satellite handover algorithms," in *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 422–425, IEEE, 2021.

[3] E. Juan, M. Lauridsen, J. Wigard, and P. Mogensen, "Handover solutions for 5G low-earth orbit satellite networks," *IEEE Access*, vol. 10, pp. 93309–93325, 2022.

[4] H. Xu, D. Li, M. Liu, G. Han, W. Huang, and C. Xu, "QoE-driven intelligent handover for user-centric mobile satellite networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 10127–10139, 2020.

[5] N. Badini, M. Jaber, M. Marchese, and F. Patrone, "User centric satellite handover for multiple traffic profiles using deep Q-Learning," *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[6] P. Dhariwal *et al.*, "OpenAI Baselines." https://github.com/openai/baselines, 2017. Accessed: June 2024.

[7] S. He, T. Wang, and S. Wang, "Load-aware satellite handover strategy based on multi-agent reinforcement learning," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2020.

[8] J. Wang, W. Mu, Y. Liu, L. Guo, S. Zhang, and G. Gui, "Deep reinforcement learning-based satellite handover scheme for satellite communications," in *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, IEEE, 2021.

[9] J. Wang, W. Mu, Y. Liu, L. Guo, S. Zhang, and G. Gui, "Deep reinforcement learning-based satellite handover scheme for satellite communications," in *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, 2021.

[10] M. K. Dahouda, S. Jin, and I. Joe, "Machine learning-based solutions for handover decisions in non-terrestrial networks," *Electronics*, vol. 12, no. 8, p. 1759, 2023.

[11] S. He, T. Wang, and S. Wang, "Load-aware satellite handover strategy based on multi-agent reinforcement learning," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pp. 1–6, 2020.

[12] A. M. Voicu, A. Bhattacharya, and M. Petrova, "Handover strategies for emerging LEO, MEO, and HEO satellite networks," *IEEE Access*, vol. 12, pp. 31523–31537, 2024.

[13] H. Liu, Y. Wang, and Y. Wang, "A successive deep Q-Learning based distributed handover scheme for large-scale LEO satellite networks," in *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, pp. 1–6, IEEE, 2022.