Coverage Optimization in RIS-Enabled Satellite–Terrestrial Networks: A Digital Twin-Based Spatial–Temporal Approach

Qihao Li, *Member, IEEE*, Qiang Ye, *Senior Member, IEEE*, Huaqing Wu, *Member, IEEE*, and Fengye Hu, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a novel reconfigurable intelligent surfaces (RIS)-enabled satellite-terrestrial networking scheme, called digital twin-based spatial-temporal approach (DTST), to maximize time-averaged coverage while guaranteeing operational constraints in dynamic low Earth orbit (LEO) environments. Existing coverage optimization methods typically adopt static formulations or idealized learning models, and they do not jointly address fast time-varying LEO geometry, weather-dependent attenuation, interference coupling, and strict operational constraints under partial observability, especially when digital-twin (DT) model mismatch and rare events can bias value estimation and trigger constraint violations. Specifically, we design a spatial-temporal coverage-grain model with boolean union of direct and RIS paths to jointly control RIS phases and allocate satellite power. Then we develop a DT-driven centralized training with decentralized execution (CTDE) learning method based on multi-agent proximal policy optimization (MAPPO) with dual-phase training to learn decentralized policies under partial observability through a reconstructed decentralized partially observable Markov decision process (Dec-POMDP) model. With the obtained DT-synchronized scenario parameters and the pre-trained policies, conservative value learning with rare-event replay is explored to mitigate overestimation and improve resilience against simulation-to-reality discrepancies. The coverage probability optimization challenge is addressed through strategic balancing of the probability of service availability for any given user and power allocation parameters in consideration of power budgets, RIS stability, LOS/visibility windows, collision avoidance, handover continuity, and interference coupling. Simulation results demonstrate that, in a Ka-band 48-LEO-satellite constellation, the DTST scheme maintains coverage probability of over 80% and prediction accuracy exceeding 90% even when operating under heavy rainfall and co-channel interference.**

***Index Terms*—Satellite-terrestrial networks, RIS, digital twin, Dec-POMDP, spatial–temporal.**

Received 4 October 2025; revised 19 February 2026; accepted 14 March 2026. Date of publication 27 March 2026; date of current version 15 April 2026. This work was supported in part by Dongguan Strategic Scientist Teams Project under Grant 20241900700013, in part by the Guangdong Province Basic and Applied Basic Research Foundation under Grant 2022KQNCX, in part by the National Natural Science Foundation of China under Grant 62201148. The associate editor coordinating the review of this article and approving it for publication was Y. Fu. *(Corresponding author: Fengye Hu.)*

Qihao Li and Fengye Hu are with the College of Communication Engineering, Jilin University, Jilin 130012, China (e-mail: qihaol@jlu.edu.cn; hufy@jlu.edu.cn).

Qiang Ye and Huaqing Wu are with the Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: qiang.ye@ucalgary.ca; huaqing.wu1@ucalgary.ca).

Digital Object Identifier 10.1109/TCCN.2026.3678233

I. Introduction

SATELLITE-TERRESTRIAL Networks (STNs) are facilitating notable advancements in live media transmission within urban environments, providing high-quality program-return video, communication audio, and teleprompting services via low Earth orbit (LEO) satellite to mobile field units deployed in challenging urban canyon environments [1], [2]. However, practical implementations continue to encounter significant operational challenges related to link-budget limitations and signal reliability issues, particularly when operating at higher frequency bands such as Ka-band. LEO communication experiences free-space path loss measuring approximately 170–180 dB, while adverse weather conditions introducing additional elevation-dependent signal degradation ranging from 10–20 dB during periods of heavy rainfall [3], [4]. As a result, compact ground-based users (GBUs) consistently encounter suboptimal signal-to-noise ratios, resulting in unstable confidence communication performance, unless additional aperture gain is incorporated into the system architecture. Reconfigurable intelligent surfaces (RIS) serve as an effective solution for providing this aperture gain through coherent reflection of satellite signals toward designated receivers [5], [6], [7], [8]. RIS panels of meter-scale dimensions demonstrate capability to generate substantial passive beamforming gain in the order of tens of decibels [9]. This signal enhancement effectively improves downlink SNR to required thresholds, establishing a more dependable downlink for return feeds that support high reliability communication performance and low-jitter program return with approximately 30-70 ms one-way satellite latency, while eliminating reliance on congested public networks. Additionally, it expands coverage to challenging environments, including urban canyons and indoor locations, without requiring increased transmission power [10], [11]. However, signal blockage in dense urban environments and link disruptions caused by Doppler shifts are commonly observed in these networks [12]. These issues lead to coverage gaps, service quality degradation, and unexpected connectivity losses.

Digital twin (DT) technology has emerged as a viable approach for optimizing network performance in STNs [13]. Traditional optimization methods are typically limited in their ability to account for the dynamic nature of satellite trajectories, changing atmospheric conditions, and the complex

beam interactions found in RIS-enabled STNs [14]. These limitations can be addressed by DT technology through the provision of real-time representations that enable proactive coverage management rather than reactive responses [15], [16]. Specifically, within LEO-assisted broadband systems, the DT framework acquires operational parameters from physical counterpart, including satellite ephemeris data, RIS phase configurations, and time-varying communication channel states. This comprehensive data acquisition enables computational modeling that simulates dynamic network conditions with high fidelity, thereby supporting evidence-based decision-making for optimal resource allocation, predictive maintenance scheduling, and adaptive beamforming strategies tailored to varying atmospheric conditions [15], [16], [17]. In addition, some studies have advanced the integration of digital twin (DT) into 6G network design and control. In particular, DT-6G concepts are discussed, and key obstacles and prospects are summarized in [18], including real-time synchronization, seamless migration, predictive analytics, and closed-loop control, which are essential for DT-driven network optimization. Large generative model (LGM)-enabled DT frameworks are further proposed in [19] to support intelligent and automated network control by synthesizing control strategies from network contexts. From a control and learning perspective, a two-timescale DT synchronization and migration framework is developed, and the resulting stochastic decision problem is solved using multi-agent deep reinforcement learning in [20]. These works establish that DT should be tightly integrated with the network control loop and that synchronization fidelity is a primary factor affecting optimization quality. Nevertheless, for RIS-enabled satellite-terrestrial networks, the coverage optimization problem additionally couples fast time-varying LEO geometry, weather-dependent attenuation, and interference with strict operational constraints, e.g., RIS tuning-rate stability, inter-satellite separation, and service continuity, and DT-physical mismatch can bias value estimation and lead to aggressive actions.

However, optimizing coverage performance in RIS-enabled SGN presents several significant challenges. First, achieving optimal satellite-RIS-user alignment is complicated due to the complexity in accurately determining time-varying coverage parameters. This difficulty is attributed to the characteristics of LEO orbital dynamics influenced by Keplerian motion, atmospheric attenuation, RIS hardware limitations, transmission handover, and GBU mobility patterns [1], [17], [21]. Coverage continuity is affected by these factors, resulting in service interruptions due to changing elevation angles, frequency shifts, and signal attenuation during precipitation events. Conventional stochastic geometry modeling approaches are insufficient for capturing these factors with the consideration of spatial-temporal relationships within STNs. Therefore, a spatial-temporal approach should be developed to incorporate the detailed orbital-channel relationships, establishing a stronger foundation for effective coverage optimization.

Second, incorporating DTs into constrained optimization problems presents a notable challenge. DT technology enables real-time synthetic experience generation for evaluating the trade-offs between coverage maximization and power conservation, particularly in the context of mission-critical data transmissions and dynamic beam configurations subject to stochastic constraints and partial observability [22], [23]. However, conventional mixed-integer nonlinear programming (MINLP) solvers are unable to effectively utilize the dynamic data provided by DTs due to inherently static formulations and computational processing latency, which necessitates fundamental redesign of optimization models to accommodate real-time adaptability and uncertainty-aware constraints. This requires adjustments to policy update mechanisms to enable effective bias gradient integration and constraint-conscious advantage estimation. Furthermore, while DTs can predict rare-event states, integrating this capability into optimization approaches requires optimization loops capable of managing hybrid action spaces with stochastic constraint satisfaction. Addressing these technical challenges necessitates the development of robust optimization models that effectively balance exploration and exploitation trade-offs while safeguarding operational parameters and ensuring convergence under partial observability conditions.

Third, after DT-driven learning is implemented, value estimates can be overestimated by simulation-to-reality mismatches and rare events, necessitating conservative, bias-adjusted updates and explicit uncertainty modeling. Value overestimation bias in simulation-based training presents a notable challenge for effective policy implementation in STNs. Network state values are frequently overestimated by algorithms during online tuning processes where model-reality disparities and exceptional circumstances exist, which can lead to severe constraint violations [22], [24]. This problem occurs when actions that exploit model limitations or rely on optimistic value assessments are selected by the system, without proper consideration of actual risk factors [25]. Policy gradient calculations can be adversely affected by this optimization bias, resulting in inaccurate RIS configuration adjustments and power allocation decisions. For addressing these technical challenges, uncertainty quantification approaches should be designed and implemented with DT which are characterized by the ability to support bias-adjusted policy refinements while accounting for the complex dynamics inherent to SGN.

In this paper, we address how to optimize coverage in RIS-enabled STNs under spatial-temporal constraints through DTs. We propose a DT-based spatial-temporal approach (DTST) scheme designed to maximize time-averaged coverage probability while guaranteeing strict constraint satisfaction. The following are the paper's primary contributions:

- We develop a coverage model that integrates both direct and RIS-reflected signal components through Boolean union operations, which incorporates parameters including line-of-sight conditions, weather attenuation factors, path-loss/RIS coupling effects, and potential interference from overlapping coverage areas. Building upon this foundation, we perform joint optimization of RIS phase configurations and satellite power allocation while accounting for orbital dynamics and ensuring continuous coverage.
- We manage the satellites by designing a decentralized partially observable Markov decision process

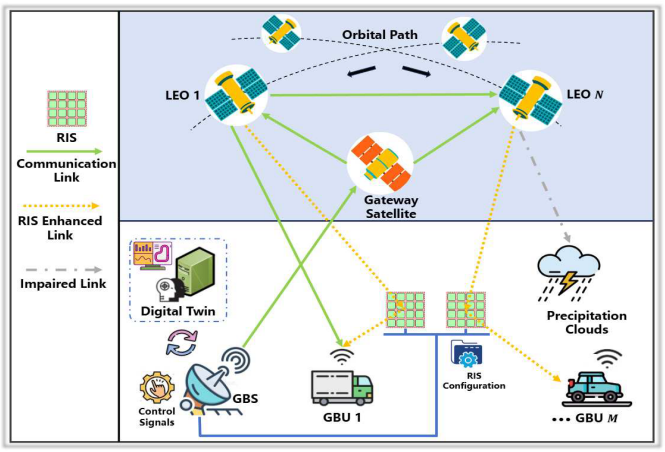


Fig. 1. Network scenario.

(Dec-POMDP) model and implements a structured training and execution method. The designed method utilizes centralized training with decentralized execution, built on multi-agent proximal policy optimization with generalized advantage estimation and appropriate policy constraints. The system incorporates safety measures for action implementation and leverages parallel processing capabilities provided by a synchronized DT environment. The learning process is methodically organized with an initial offline preparation phase that examines a comprehensive range of physically accurate scenarios, followed by an online refinement phase that incorporates real-time operational data.

- A closed-loop accuracy bound is designed between the DT and physical entity through extended Kalman filter (EKF) state data integration and proportional error correction. The proposed DTST scheme implements conservative value calibration by combining uncertainty-aware sampling with cautious return estimation through multiple rare-event simulations. Value overestimation problems are addressed through the implementation of conservative critic targets, replay mechanisms for uncommon events, trajectory manipulation with scenario enhancement in areas presenting significant risk, and regularization procedures across similar rare samples, with all these processes being regulated by quality metrics for synchronization.

II. System Model

A. Network Model

As shown in Fig. 1, the SGN integrates four kinds of components: a constellation of N LEO satellites $\mathcal{S} = \{s_1, \ldots, s_N\}$, with one serving as the gateway satellite $s_g \in \mathcal{S}$; M GBUs $\mathcal{U} = \{u_1, \ldots, u_M\}$; a central ground base station (GBS) controller $\mathcal{G}$; and L RIS components $\mathcal{R} = \{r_1, \ldots, r_L\}$ with each containing K programmable elements. The connectivity among network components are addressed in mixed urban environments where GBUs connect to the network via LEO satellites rather than through direct connections with GBS, with RIS panels strategically implemented at locations prone to signal degradation. Satellite positions are characterized by time-varying Cartesian coordinates $\mathbf{p}_{s_i}(t) = (x_{s_i}, y_{s_i}, z_{s_i}) \in \mathbb{R}^3$ obtained by propagating a Walker–Delta layout under Keplerian mechanics, while $\mathbf{p}_{u_j} = (x_{u_j}, y_{u_j}, z_{u_j}) \in \mathbb{R}^3$ and $\mathbf{p}_{r_l} = (x_{r_l}, y_{r_l}, z_{r_l}) \in \mathbb{R}^3$ denote the locations of GBUs and RISs respectively. Let $\mathcal{A}$ be the service area covered by the satellite network with $\mathbf{p}_{u_j} \in \mathcal{A}$ denoting each GBU location is a point within this service area. The network employs a structured hierarchical architecture comprising three distinct data transmission links: 1) direct links from the GBS $\mathcal{G}$ to designated satellites; 2) relayed transmission links from $\mathcal{G}$ to the gateway satellite s_g, which subsequently distributes data via inter-satellite links (ISLs) to the serving satellite s_i, and 3) GBU access connections established either through direct satellite-to-ground transmission or through enhanced signal links utilizing RISs $\{r_l\}_{l=1}^{L}$.

The GBS $\mathcal{G}$ implements three core control functions. First, for satellite network configuration, $\mathcal{G}$ determines optimal ISL topologies $\mathcal{E}(t) \subseteq \mathcal{S} \times \mathcal{S}$ and delivers connectivity instructions via $\mathcal{G} \rightarrow s_g : \{\text{link}(s_i, s_k) \,|\, \forall (s_i, s_k) \in \mathcal{E}(t)\}$. Second, for satellite-RIS pairing configuration, $\mathcal{G}$ determines $\theta_{i,l} \in \mathbb{R}^P$, the satellite transmit beam vector pointing-angle vector toward RIS r_l. The control parameters $(r_l, \theta_{i,l})$ is subsequently transmitted to the designated satellite s_i through the established gateway communication channel $\mathcal{G} \rightarrow s_g \rightarrow s_i : \{(r_l, \theta_{i,l}) \mid r_l \in \mathcal{R}_i(t) \subseteq \mathcal{R}\}$, where $\mathcal{R}_i(t)$ represents the subset of RISs currently allocated to s_i. Third, leveraging real-time feedback from both the satellite constellation and GBUs, $\mathcal{G}$ computes the optimal phase-shift matrix for each RIS and transmits it over a wired control link, $\mathcal{G} \rightarrow r_l : \mathbf{\Phi}_l(t)$, where $\mathbf{\Phi}_l(t) = \text{diag}\big(e^{j\phi_{l,1}(t)}, \ldots, e^{j\phi_{l,K}(t)}\big)$ with $\phi_{l,k}(t) \in [0, 2\pi)$. This closed-loop process dynamically adapts the RIS to the serving satellite beam and the set of GBUs within its coverage, maximizing end-to-end link quality. When satellite s_i provides service to GBU u_j via RIS r_l, the direct path measures $d_{ij}^{\text{direct}} = \|\mathbf{p}_{s_i} - \mathbf{p}_{u_j}\|$, while the RIS-assisted path comprises $d_{ijl}^{\text{RIS}} = \|\mathbf{p}_{s_i} - \mathbf{p}_{r_l}\| + \|\mathbf{p}_{r_l} - \mathbf{p}_{u_j}\|$.

In summary, the control plane operates along paths $\mathcal{G} \rightarrow s_g$ and $\mathcal{G} \rightarrow r_l$ for the distribution of policy, beam, and RIS control information. The data plane functions on a per-GBU basis, with payload traffic being transmitted between individual GBUs and their respective serving satellite s_i, rather than being broadcast across all satellites. When necessary, payload is redirected from s_i to ground via s_g; however, it is important to note that user downlink/uplink connections terminate exclusively at the serving satellite. And the gateway s_g serves solely for backhaul connectivity, network control, and inter-satellite routing functions, rather than introducing an additional communication layer in the GBU interface pathway. Moreover, the GBS $\mathcal{G}$ maintains synchronized virtual replicas DT $\{\mathcal{DT}_{s_i}, \mathcal{DT}_{\mathcal{G}}, \mathcal{DT}_{u_j}, \mathcal{DT}_{r_l}\}$ of all network components. These DTs are colocated with $\mathcal{G}$, and incorporate comprehensive telemetry data, including satellite positional coordinates $\mathbf{p}_{s_i}(t)$, velocity parameters, channel state metrics; operational status and orientation configurations from RIS elements; and service requests with corresponding SNR measurements from GBUs. This DT support is directly used in our Dec-POMDP formulation and CTDE learning pipeline, including state

reconstruction, constraint-aware performance evaluation, and the generation of DT trajectories for offline warm-up and subsequent online adaptation. A detailed DT architecture and synchronization mechanism are presented in Sec. II-E.

B. RIS-Enabled Channel Model

The RIS-augmented satellite-ground channel model incorporates both deterministic path loss and stochastic atmospheric impairments. For the direct path between satellite s_i and GBU u_j, the path loss integrates free-space attenuation and atmospheric degradation $\mathcal{L}_{ij}^{\text{direct}}(t) = 20\log_{10}\left(\frac{4\pi d_{ij}^{\text{direct}}(t)}{\lambda}\right) + \mathcal{A}_{\text{atm}}(d_{ij}^{\text{direct}}, f, t, \theta_{ij})$, where λ is the wavelength, and $\mathcal{A}_{\text{atm}}$ aggregates frequency dependent attenuation from rain (ITU-R P.618), clouds (ITU-R P.840), and fog (ITU-R P.676) [26]. The atmospheric term follows $\mathcal{A}_{\text{atm}}(d, f, t, \theta_{(\cdot)}) = \zeta_{\text{rain}}(f,t)\cdot\ell_{\text{eff}}(\theta_{ij}) + \zeta_{\text{cloud}}(f,t)\cdot d_{\text{cloud}} + \zeta_{\text{fog}}(f,t)\cdot d_{\text{fog}}$, where $\zeta_{\{\cdot\}}(f,t)$ denotes specific attenuation, ℓ_{eff} is the effective path length through precipitation layers, θ_{ij} is the elevation angle, and d_{cloud}, d_{fog} are cloud/fog penetration distances. For RIS-assisted paths ($s_i \to r_l \to u_j$), path loss combines two segments:

$$\begin{aligned}&\mathcal{L}_{ijl}^{\text{RIS}}(t)\\ &= \underbrace{20\log_{10}\left(\frac{4\pi d_{il}^{\text{sat-RIS}}(t)}{\lambda}\right) + \mathcal{A}_{\text{atm}}(d_{il}^{\text{sat-RIS}}, f, t, \theta_{il})}_{\text{Satellite-to-RIS}}\\ &\quad + \underbrace{20\log_{10}\left(\frac{4\pi d_{lj}^{\text{RIS-GBU}}}{\lambda}\right) + \mathcal{A}_{\text{atm}}(d_{lj}^{\text{RIS-GBU}}, f, t, \theta_{lj})}_{\text{RIS-to-GBU}}\end{aligned} \tag{1}$$

where $d_{il}^{\text{sat-RIS}} = \|\mathbf{p}_{s_i} - \mathbf{p}_{r_l}\|$, $d_{lj}^{\text{RIS-GBU}} = \|\mathbf{p}_{r_l} - \mathbf{p}_{u_j}\|$. The composite channel gain becomes:

$$\begin{aligned}h_{ij}^{\text{total}} &= \underbrace{\sqrt{G_{s_i\to u_j}G_{u_j}}\cdot\Gamma_{\text{direct}}\cdot e^{-j\frac{2\pi}{\lambda}d_{ij}^{\text{direct}}}}_{\text{sat-GBU (Direct path)}}\\ &\quad + \sum_{r_l\in\mathcal{R}_i}\underbrace{\sqrt{G_{s_i\to r_l}G_{u_j}}\mathbf{h}_{il}^{\text{T}}\mathbf{\Phi}_l\mathbf{h}_{lj}\cdot\Gamma_{\text{RIS}}}_{\text{sat-RIS-GBU (RIS-aided path)}}\end{aligned} \tag{2}$$

where $G_{s_i\to u_j}$ is the transmit antenna gain of s_i toward u_j, $G_{s_i\to r_l}$ is the gain toward RIS r_l, $\Gamma_{\text{direct}} = 10^{-\mathcal{L}_{ij}^{\text{direct}}/20}$, $\Gamma_{\text{RIS}} = 10^{-\mathcal{L}_{ijl}^{\text{RIS}}/20}$, and $\mathbf{h}_{il} \in \mathbb{C}^K$, $\mathbf{h}_{lj} \in \mathbb{C}^K$ model Rician fading for satellite-RIS and RIS-GBU links. Interference originates from co-channel satellites $s_n \neq s_i$ using identical spectral resources. The aggregate interference power incorporates both propagation paths:

$$\begin{aligned}&I_{\text{inter}}\\ &= \sum_{s_n\neq s_i}P_n\Bigg[\underbrace{G_{s_n\to u_j}G_{u_j}(\Gamma_{\text{direct}})^2|h_{nj}^{\text{direct}}|^2\cdot\mathbf{1}_{\text{LOS}}(s_n,u_j,t)}_{\text{direct interference}}\\ &\quad + \underbrace{\sum_{r_i\in\mathcal{R}_n}G_{s_n\to r_l}G_{u_j}(\Gamma_{\text{RIS}})^2|\mathbf{h}_{nl}^T\mathbf{\Phi}_l\mathbf{h}_{lj}|^2\cdot\mathbf{1}_{\text{LOS}}(s_i,r_l,u_j,t)}_{\text{RIS-reflected interference}}\Bigg].\end{aligned} \tag{3}$$

The LOS indicator $\mathbf{1}_{\text{LOS}}$ depends on elevation $\theta(t)$ and azimuth $\phi(t)$ constraints:

$$\mathbf{1}_{\text{LOS}}(s_i,u_j,t) = \mathbf{1}[\theta_{ij}(t)\geq\theta_{\min}]\cdot\mathbf{1}_{\text{clear-sky}}(t), \tag{4}$$

$$\begin{aligned}\mathbf{1}_{\text{LOS}}(s_i,r_l,u_j,t) &= \mathbf{1}[\theta_{il}(t)\geq\theta_{\min}]\\ &\quad\cdot\mathbf{1}[\phi_{lj}(t)\in[\phi_{\min},\phi_{\max}]]\cdot\mathbf{1}_{\text{clear-sky}}(t),\end{aligned} \tag{5}$$

where $\mathbf{1}_{\text{clear-sky}}(t)$ is zero during extreme weather-induced blockages, $\theta_{\min}$ is the elevation-mask angle threshold. We implement $\theta_{\min} = 20^\circ$ as our operational parameter, which aligns with industry practices for non-GSO systems that typically utilize minimum terminal elevations between $10^\circ - 30^\circ$ to effectively mitigate path loss, signal blockage, and atmospheric scintillation effects [27].

C. Coverage Model

The coverage optimization framework employs a Boolean model from stochastic geometry, where LEO satellites function as germs and their dynamic coverage zones constitute grains. This approach models spatial randomness in GBU distribution and satellite visibility while incorporating critical spatial-temporal constraints unique to satellite networks. The objective function maximizes the time-averaged coverage probability across $\mathcal{A}$ over T as

$$\max_{\{P_i(t),\mathbf{\Phi}_l(t),\theta_{i,l}(t)\}}\frac{1}{T}\int_0^T\mathbb{E}_{\mathbf{p}\sim\mathcal{A}}\left[\mathbf{1}_{\mathcal{C}(t)}(\mathbf{p})\right]dt, \tag{6}$$

where $\mathcal{C}(t) = \bigcup_{s_i\in\mathcal{S}}\mathcal{B}_i(t)$ represents the union of coverage grains that $\mathcal{B}_i(t) = \mathcal{B}_i^{\text{direct}}(t)\cup\bigcup_{r_l\in\mathcal{R}_i(t)}\mathcal{B}_{i,l}^{\text{RIS}}(t)$, and with

$$\mathcal{B}_i^{\text{direct}}(t) = \{\mathbf{p}_{u_j}\in\mathcal{A}:\gamma_i(t,\mathbf{p})\geq\gamma_{\text{th}}\}, \tag{7}$$

$$\mathcal{B}_{i,l}^{\text{RIS}}(t) = \left\{\mathbf{p}_{u_j}:\|\mathbf{p}_{r_l}-\mathbf{p}_{u_j}\|\leq d_{\max},\gamma_i^{\text{RIS}}(t,\mathbf{p})\geq\gamma_{\text{th}}\right\}, \tag{8}$$

with γ_{th} being the SINR threshold derived from minimum rate requirement $R_{\min} = B\log_2(1+\gamma_{\text{th}})$. The expectation $\mathbb{E}_{\mathbf{p}}$ integrates over GBU distribution, while the indicator $\mathbf{1}_{\mathcal{C}(t)}$ marks covered locations. The transmission power of each satellite is bounded by hardware limitations and regulatory requirements, $0\leq P_i(t)\leq P_i^{\max}, \forall i\in\{1,\ldots,N\}, t\in[0,T]$, where $P_i^{\max}$ is derived from satellite power budgets. The orbital-motion is governed by two constraints that together guarantee satellite trajectories, continuous ground coverage, and reliable LOS geometry. First, the orbital-dynamics constraint,

$$\begin{aligned}&\|\mathbf{p}_{s_i}(t)-\mathbf{p}_{s_i}(t_0)\|_2\\ &= v_{\text{orb}}(t-t_0) + \frac{1}{2}\left[\frac{v_{\text{orb}}^2}{r_{\text{orbit}}} - \omega_e^2 r_{\text{orbit}}\cos^2(l_{\text{geo}})\right](t-t_0)^2,\end{aligned} \tag{9}$$

binds the satellite's instantaneous position vector $\mathbf{p}_{s_i}(t)$ to a reference time t_0, where $\omega_e = 7.29\times10^{-5}$ rad/s is Earth's rotation rate, l_{geo} is the geodetic latitude, v_{orb} denotes the satellite's instantaneous orbital speed along its Keplerian path and r_{orbit} is the orbital radius parameter. The orbital-dynamics equation enforces that the satellite state trajectory is physically feasible and consistent with orbital motion. This constraint is required because the time-varying geometry $\mathbf{p}_{s_i}(t)$ directly

determines satellite visibility, link distance, elevation angle, and hence the SINR term $\gamma_i(t, \mathbf{p}_{u_j})$ used in the coverage model. Therefore, orbital dynamics is not optimized as a free variable; it provides the physically valid state evolution on which association, RIS selection, and coverage evaluation are computed.

Second, to rule out abrupt flicker in the illuminated region, the coverage-continuity constraint imposes a transport-equation bound on every stochastic coverage grain $\mathcal{B}_i(t)$:

$$\frac{\partial \mathcal{B}_i(t)}{\partial t} \leq v_{\text{orb}} \nabla_{\mathbf{p}} \mathcal{B}_i(t). \tag{10}$$

The inequality forces the coverage to drift across the Earth's surface at a speed consistent with the satellite's ground track, eliminating unphysical, discontinuous jumps in coverage. The effective SINR at GBU u_j served by s_i integrates path loss, interference, and RIS beamforming gain:

$$\gamma_i(t, \mathbf{p}_{u_j}) = \frac{P_i}{\sigma^2 + I_{\text{inter}}} \left| \sqrt{G_{s_i} G_{u_j}} \cdot \Gamma_{\text{direct}} \cdot e^{-j\frac{2\pi}{\lambda} d_{ij}^{\text{direct}}} \cdot y_{ij0}(t) + \sum_{r_l \in \mathcal{R}_i} \left[\sqrt{G_{s_i} G_{u_j}} \mathbf{h}_{il}^{\mathrm{T}} \boldsymbol{\Phi}_l^* \mathbf{h}_{lj} \cdot \Gamma_{\text{RIS}} \cdot y_{ijl}(t) \right] \right|^2, \tag{11}$$

where $\sigma^2 = N_0 B$ is thermal noise. RIS phase shifts $\boldsymbol{\Phi}_l^* = \text{diag}(e^{j\phi_{l,1}^*}, \ldots, e^{j\phi_{l,K}^*})$ maximize coherent combining by solving $\phi_{l,k}^* = \arg\min_{\phi_{l,k}} \left| \angle h_{ij}^{\text{direct}} - \angle \left(\mathbf{h}_{il}^{\mathrm{T}} \boldsymbol{\Phi}_l \mathbf{h}_{lj} \right) \right|$.[1] To preserve quality of service at every instant, each active satellite-GBU pair must satisfy the instantaneous-rate constraint

$$\gamma_i(t, \mathbf{p}_{u_j}) \geq \gamma_{\text{th}} \cdot \mathbf{1}_{\left\{\sum_{l=0}^{L} y_{ijl}(t)=1\right\}}, \quad \forall i, j, l, t, \tag{12}$$

where $y_{ij0}(t) \in \{0, 1\}$ denotes whether satellite s_i is directly serving GBU u_j, and $y_{ijl}(t) \in \{0, 1\}$ indicates RIS-mediated service $s_i \xrightarrow{\theta_{i,l}} r_l \rightarrow u_j$, and $\sum_{l=0}^{L} y_{ijl}(t) \leq 1$ is exclusive path selection. The received SINR $\gamma_i(t, \mathbf{p}_{u_j})$ must remain above the design threshold γ_{th}. Short-lived associations, however, can still degrade GBU experience even when each slot is individually admissible, so a service-continuity constraint is imposed

$$\sum_{\tau=t}^{t+\Delta t_{\min}} \left(y_{ij0}(\tau) + \sum_{l=1}^{L} \kappa_l(\tau) \cdot y_{ijl}(\tau) \right) \geq \Delta t_{\min} \cdot z_{ij}(t), \tag{13}$$

where $z_{ij}(t) = 1$ is a handover variable that becomes active once s_i begins serving u_j (either direct or RIS-mediated), $\Delta t_{\min}$ is the minimum dwell time, and $\kappa_l(\tau) = \exp\left(-\frac{\|\mathbf{p}_{r_l}(\tau) - \mathbf{p}_{r_l}^{\text{ref}}(\tau)\|^2}{2\sigma_{\text{point}}^2} \right)$ quantifies RIS pointing stability. The continuity condition enforces a minimum dwell time $\tau_{\min}$ once user u_j is served by satellite s_i, which suppresses slot-by-slot switching and avoids ping-pong associations. This constraint captures practical handover overhead, including signaling, beam/phase reconfiguration, and transient SINR loss. Consequently, the optimized coverage performance reflects operationally stable scheduling decisions rather than unrealistic rapid re-association. Based on the above definitions, the coverage probability in Eq. (6) is determined once the instantaneous SINR term $\gamma_i(t, \mathbf{p}_{u_j})$ is specified. In the next step, we formulate a coverage-oriented optimization problem by selecting the controllable variables (e.g., satellite power control, RIS configuration, and association/scheduling indicators) to maximize the time-averaged coverage probability subject to practical physical and operational constraints.

D. Unified Optimization Formulation

In particular, the decision variables include the transmit power $P_i(t)$, the RIS configuration $\boldsymbol{\Phi}_l(t)$ with phase elements $\phi_{i,k}(t)$, the auxiliary control/pointing terms $\theta_{i,l}(t)$, and the association/scheduling indicators $y_{ij0}(t)$, $y_{ijl}(t)$, as well as the handover-related variable $z_{ij}(t)$ for continuity constraints. Combining all elements, the optimization problem is shown in Eq. (14), at the bottom of the next page. The unified optimization problem centers on three interconnected control parameters-satellite transmit power, RIS configuration, and beam-pointing vectors-along with scheduling indicators that determine service allocation for each GBU per time slot. The objective's disjunction operator ($\bigvee$) captures the Boolean union of coverage grains. The optimization process calculates coverage by integrating a binary coverage indicator across space and time, then dividing by the total mission duration T to approach complete coverage within physical limitations. Constraint Eq. (14b), as shown at the bottom of the next page, limits each satellite's RF output power, restricting RIS phase shifts to the range $[0, 2\pi)$. Constraint Eq. (14c), as shown at the bottom of the next page, ensures the metasurface maintains reflection coherence while cycling through its configuration states, where $\delta_{\max}$ represents the maximum phase adjustment permitted by hardware specifications for the collective modification across RIS elements, and f_{ris} denotes the maximum frequency at which the RIS controller can implement reconfigurations. And $D_{\min}$ is the minimum inter-satellite separation distance between satellites. To prevent redundant service allocation, in Eq. (14d), as shown at the bottom of the next page, each user may connect through at most one path at any time, and no two satellites may utilize the same RIS simultaneously. Load balancing constraint in Eq. (14e), as shown at the bottom of the next page, ensures each user connects to only one satellite at a time while limiting the number of users any satellite can serve concurrently, where $U_{\max}$ is the maximum number of GBUs that satellite s_i can serve simultaneously given limits on RF chains, bandwidth, processing, and power. Constraint Eq. (14f), as shown at the bottom of the next page, regulates the inter-slot variance in the composite RIS-aided beam phase across the effective channel $\mathbf{h}_{ij}^H \boldsymbol{\Phi}_i \mathbf{g}_i$, ensuring that consecutive phases maintain proper alignment. The parameter $\Delta\phi_{\max}$ is the maximum permissible phase shift per update interval, which is determined by coherence time, Doppler effects, RIS update frequency, and phase quantization factors.

[1]The RIS phase matrix$\boldsymbol{\Phi}_l(t)$is implemented by finite-speed phase shifters, so∼its per-slot update is rate-limited, to∼capture settling-time and resolution constraints and prevent abrupt phase jumps. Additionally, because the phase reference$\theta_{i,l}(t)$can only be tracked within a finite coherence interval under Doppler/geometry changes, we enforce$|\Delta\theta_{i,l}(t)| \leq \omega_{\max}\Delta t$to maintain phase-consistent combining and avoid coherent-gain collapse in$\gamma_i(t, \mathbf{p}_{u_j})$.

Note that, to obtain a deterministic optimization result that can be evaluated on each rollout, we adopt a sample-average approximation: at each slot t, we draw a Poisson point process $P(t) \subset A$ and approximate $\mathbb{E}_{p\sim A}[\mathbf{1}_{C(t)}(p)]$ by $\frac{1}{|P(t)|}\sum_{p\in P(t)} \mathbf{1}(\max_i \gamma_i(t,p) \geq \gamma_{\text{th}})$. This empirical estimate becomes the coverage-utility term in the unified reward Eq. (16), which enables deterministic per-slot feedback under realized DT trajectories. The hard constraints in Eq. (14b)-(14f) are handled by two complementary mechanisms. During training, they are incorporated into Eq. (16) using normalized hinge/barrier-type surrogates, including a differentiable penalty for the separation constraint $\|p_{s_i}(t) - p_{s_j}(t)\|_2 \geq D_{\min}$ and explicit continuity/visibility counters $H_c(t)$ and $V_c(t)$. During execution, candidate actions are projected onto feasibility: if the predicted separation violates $D_{\min}$, $\Delta\theta_i(t)$ is replaced by the minimum-norm correction that restores $d_{ij}^{\text{pred}} \geq D_{\min}$, and RIS phase coherence is enforced by uniformly rescaling $\Delta\mathbf{\Phi}_i(t)$ through a Lagrange-multiplier update to satisfy $|\angle(h_{ij}^H \Delta\mathbf{\Phi}_i(t) g_i)| \leq \Delta\phi_{\max}$. Since the joint problem remains non-convex with discrete scheduling variables, we do not claim global optimality; instead, solution quality is ensured by feasibility projection together with conservative value learning on rare events, where the critic target is constructed to be pessimistic up to a bounded DT-physical transition error, reducing overestimation-driven violations and improving robustness.

E. Digital Twin Model

The DT co-locates with $\mathcal{G}$, hosting four synchronised replicas: $\{\mathcal{DT}_{s_i}\}_{i=1}^N$ for the satellite, $\{\mathcal{DT}_{r_l}\}_{l=1}^L$ for the RIS panels, $\{\mathcal{DT}_{u_j}\}_{j=1}^M$ for the GBUs, and $\mathcal{DT}_{\mathcal{G}}$ for the network core. Satellites transmit orbital position data $\mathbf{p}_{s_i}(t)$, beam-pointing vectors $\boldsymbol{\theta}_{i,l}(t)$, power levels $P_i(t)$ and connection performance metrics. RIS elements provide their phase configuration matrices $\mathbf{\Phi}_l(t)$ along with adjustment rates and stability indicators $\kappa_l(t)$. GBUs report their GPS coordinates $\mathbf{p}_{u_j}$ and measured signal quality $\gamma_i(t, \mathbf{p}_{u_j})$ for both direct connections $(y_{ij0} = 1)$ and RIS-reflected connections $(y_{ijl} = 1)$. The central controller $\mathcal{G}$ distributes the current connection assignments $\{y_{ij0}, y_{ijl}, z_{ij}\}$ and inter-satellite link $\mathcal{E}(t)$, maintaining synchronization between digital control systems and physical network components. A Keplerian propagator in $\mathcal{DT}_{s_i}$ updates satellite positions $\mathbf{p}_{s_i}(t)$ while maintaining the collision safety margin $D_{\min}$; a weather model in $\mathcal{G}$ incorporates rain, cloud, and fog data into the channel solver in $\mathcal{G}$ that calculates $\mathcal{L}_{ij}^{\text{direct}}$ and $\mathcal{L}_{ijl}^{\text{RIS}}$ through sharing DT data among $\mathcal{DT}_{s_i}, \mathcal{DT}_{r_l}$ and $\mathcal{DT}_{u_j}$; and a rolling-horizon optimizer recalculates $\{P_i, \mathbf{\Phi}_l, \boldsymbol{\theta}_{i,l}\}$ within power, tuning-rate, and dwell-time constraints.

To ensure the stability of control decisions, the DT maintains a buffer $\mathcal{B}_{\text{rare}}$ of size C_{buffer}. When rain loss exceeds $\mathcal{L}_{\text{th}}$, interference surpasses I_{th}, or multiple handovers occur simultaneously, the system stores the corresponding system states and outputs. The value calibration process combines uncertainty-aware sampling with conservative return estimation through multiple simulations of rare events. These scenarios are replayed during low network traffic periods to test policies in advance of potential issues. State estimates undergo correction through a Kalman-style assimilation process that maintains the replica within a defined error bound: $\left\|\mathbf{x}_{\text{phys}}(t) - \mathbf{x}_{\mathcal{DT}}(t)\right\|_2 \leq \epsilon_{\text{sync}}$. When this threshold is exceeded, the system initiates an emergency resynchronization procedure that temporarily increases telemetry sampling rates and reduces the optimization horizon until system coherence is restored. This approach provides demonstrable bounds on overestimation, maintaining accuracy even during conditions of severe rain attenuation, while effectively resolving simulation-to-reality value discrepancies through the implementation of bias-corrected advantage functions. The control cycle completes when the DT transmits updated power levels, phase configurations, and pointing commands to $\mathcal{G}$, which then distributes them to the satellite and RIS controllers.

Note that, the DT is a controller-side capability co-located with $\mathcal{G}$ that maintains synchronized replicas and generates corrected state trajectories and high-fidelity rollouts. DTST is the algorithmic layer that uses these outputs to build the constrained Dec-POMDP and to run MAPPO-based learning and online control.

$$\max_{\substack{P_i,\mathbf{\Phi}_l,\boldsymbol{\theta}_{i,l},\\ y_{ij0},y_{ijl},z_{ij}}} \frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{p}\sim\mathcal{A}}\left[\mathbf{1}_{\mathcal{C}(t)}\left(\bigvee_{i=1}^N \gamma_i(t,\mathbf{p}_{u_j}) \geq \gamma_{\text{th}}\right)\right] dt \tag{14a}$$

$$\text{s.t.} \quad 0 \leq P_i(t) \leq P_i^{\max}, \quad \phi_{i,k}(t) \in [0, 2\pi), \tag{14b}$$

$$\left\|\frac{\partial \mathbf{\Phi}_i(t)}{\partial t}\right\|_F \leq \delta_{\max} \cdot f_{\text{ris}}, \quad \|\mathbf{p}_{s_i}(t) - \mathbf{p}_{s_j}(t)\|_2 \geq D_{\min}, \tag{14c}$$

$$y_{ij0}(t) + \sum_{l=1}^L y_{ijl}(t) \leq 1, \quad \sum_{i=1}^N y_{ijl}(t) \leq 1 \ \ (\text{per RIS } r_l), \tag{14d}$$

$$\sum_{i=1}^N y_{ij*}(t) \leq 1 \quad (\text{each GBU}), \quad \sum_{j=1}^M \left[y_{ij0}(t) + \sum_l y_{ijl}(t)\right] \leq U_{\max} \quad (\text{each LEO}), \tag{14e}$$

$$\angle\left(\mathbf{h}_{ij}^H \mathbf{\Phi}_i(t)\mathbf{g}_i\right) - \angle\left(\mathbf{h}_{ij}^H \mathbf{\Phi}_i(t-1)\mathbf{g}_i\right) \leq \Delta\phi_{\max}. \tag{14f}$$

$$\text{Eq. (4), Eq. (5), Eq. (9), Eq. (10), Eq. (12), Eq. (13)} \tag{14g}$$

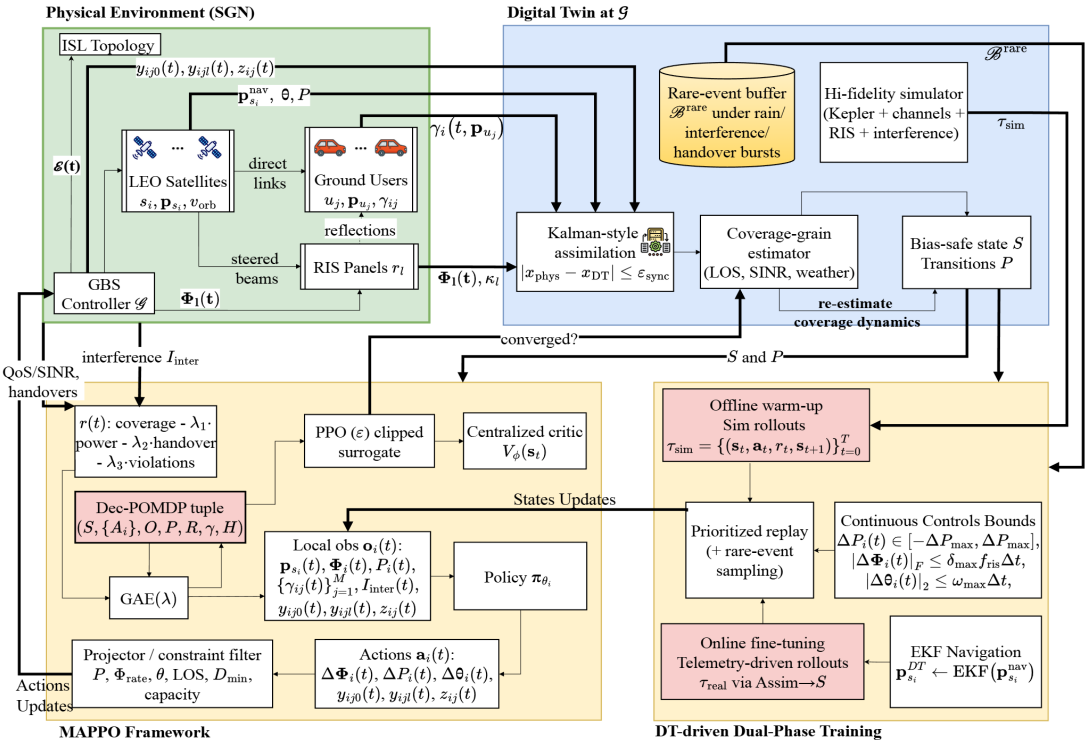


Fig. 2. The proposed DTST scheme.

III. The Proposed Scheme

As depicted in Fig. 2, the DTST scheme consists (i) a DT synchronization module that estimates and corrects the scenario state/coverage dynamics using real-time telemetry and (ii) a constrained multi-agent optimization module that learns decentralized satellite policies under the Dec-POMDP formulation. The DT provides high-fidelity transitions and corrected state vectors to support policy learning, while the learned policies generate control actions that are fed back to the DT for iterative refinement. Alg. 1 details the main functional blocks and their interactions.

A. Environment and Observations

Managing multiple satellites with limited and delayed information requires a Dec-POMDP model. In the STNs, each satellite only has access to its own data, nearby satellite signals, and ground feedback. This means no satellite has complete information about the entire system or interference patterns. The Dec-POMDP framework helps us handle this limited information while optimizing all satellites together using a shared performance metric. We define this as a tuple $(\mathcal{S}, \{\mathcal{A}_i\}, \mathcal{O}, \mathcal{P}, R)$. Here, the global state space $\mathcal{S}$ includes satellite positions $\mathbf{p}_{s_i}(t)$, RIS phase matrices $\boldsymbol{\Phi}_l(t)$, transmit powers $P_i(t)$ and ground user distributions $\mathbf{p}_{u_j}(t)$. GBS $\mathcal{G}$ generates and each satellite s_i executes both continuous variables and binary scheduling decisions $\mathcal{A}_i = \{\Delta \boldsymbol{\Phi}_i(t), \Delta P_i(t), \Delta \boldsymbol{\theta}_i(t), y_{ij0}(t), y_{ijl}(t), z_{ij}(t)\}$. Its local observation $\mathbf{o}_i(t) \in \mathcal{O}$ includes its own status, SINR measurements, and information about neighboring satellites. The transition function $P(\mathbf{s}(t+1) \mid \mathbf{s}(t), \mathbf{a}(t))$ combines orbital mechanics with radio channel changes. The reward $R(t)$ measures coverage area while accounting for power consumption and interference. The DT provides high-fidelity transition models incorporating Keplerian dynamics and stochastic channel effects, enabling realistic simulation of LEO satellite mobility and signal propagation. At each time slot t, satellite node s_i receives a local observation $\mathbf{o}_i(t) = [\mathbf{p}_{s_i}(t), \boldsymbol{\Phi}_i(t), P_i(t), \{\gamma_{ij}(t)\}_{j=1}^{M}, I_{\text{inter}}(t), y_{ij0}(t), y_{ijl}(t), z_{ij}(t)]$. The set $\{\gamma_{ij}(t)\}$ contains the per-user SINR measurements from the previous time slot.

B. Action Space

Each satellite selects a hybrid action that consists of both continuous and binary components:

$$\mathbf{a}_i(t) = \big[\underbrace{\Delta \boldsymbol{\Phi}_i(t),\ \Delta P_i(t),\ \Delta \boldsymbol{\theta}_i(t)}_{\text{continuous}},\ \underbrace{y_{ij0}(t),\ y_{ijl}(t),\ z_{ij}(t)}_{\text{binary}}\big]. \tag{15}$$

The continuous components are used to adjust the RIS phase vector, transmit-power envelope, and beam-pointing attitude. The binary components are responsible for scheduling users, selecting RIS paths, and indicating dwell-time completion. Element-wise coherence can be maintained by: $|\Delta \phi_{i,k}(t)| \le \Delta \phi_{\text{element}}, k = 1, \ldots, K$, where $\Delta \phi_{\text{element}}$ represents the maximum step supported by the network. Adaptive boundary-aware clipping is implemented for power control. With $\Delta P_{\max}$ defined as the per-slot ramp limit, the admissible interval is given by: $\Delta P_i(t) \in [-\min\{P_i(t), \Delta P_{\max}\},\ \min\{P_i^{\max} - P_i(t), \Delta P_{\max}\}]$. This ensures that the next power level $P_i(t+1) = P_i(t) + \Delta P_i(t)$ remains within the feasible range $[0, P_i^{\max}]$ while preserving the magnitude of the intended change when possible. For a propagation step δ_t, the future

Algorithm 1 Sub-Algorithm Blocks in DTST

procedure PROJECT_ACTION($a_i, P_i, \mathbf{\Phi}_i, \mathbf{h}_{ij}, \mathbf{g}_i$)
Output: feasible a_i
  $P_i \leftarrow \text{clip}(P_i + \Delta P_i, 0, P_i^{\max})$
  $\Delta\mathbf{\Phi}_i \leftarrow \arg\min_{\|\partial_t \mathbf{\Phi}_i\|_F \le \delta_{\max} f_{\text{ris}}} \|\Delta\mathbf{\Phi}_i - \Delta\mathbf{\Phi}_i^{\text{raw}}\|$ with $\angle(\mathbf{h}_{ij}^H \mathbf{\Phi}_i \mathbf{g}_i) - \angle(\mathbf{h}_{ij}^H \mathbf{\Phi}_i^- \mathbf{g}_i) \le \Delta\phi_{\max}$
    return $a_i, P_i, \mathbf{\Phi}_i$
end procedure
procedure EKF_SYNC($\mathbf{x}_{\mathcal{DT}}, \mathbf{y}^{\text{tele}}, \mathbf{x}_{\text{physical}}$)
Output: synchronized $\mathbf{x}_{\mathcal{DT}}$ and ϵ_{sync}
  Predict/Update: $\hat{\mathbf{x}}_{\mathcal{DT}} \leftarrow \mathbf{F}\mathbf{x}_{\mathcal{DT}} + \mathbf{K}(\mathbf{y}^{\text{tele}} - \mathbf{H}\mathbf{x}_{\mathcal{DT}})$
  Proportional correction: $\mathbf{x}_{\mathcal{DT}} \leftarrow \mathbf{x}_{\mathcal{DT}} + \Gamma \odot (\mathbf{x}_{\text{physical}} - \mathbf{x}_{\mathcal{DT}})$
  if $\|\mathbf{x}_{\text{physical}} - \mathbf{x}_{\mathcal{DT}}\|_2 > \epsilon_{\max}$ for $> t_{\text{th}}$ **then**
    rollback to checkpoint
  end if
  return $\mathbf{x}_{\mathcal{DT}}, \epsilon_{\text{sync}}$
end procedure
procedure MAPPO_STEP($\mathcal{B}, \pi_{\theta_i}, V_\phi$)
Output: updated $\{\theta_i\}$ and $\mathbf{\Phi}$
  Policy loss: $\widehat{J}(\theta) = \mathbb{E}[\min(\rho_t \hat{A}_t, \text{clip}(\rho_t, 1\pm\epsilon)\hat{A}_t)]$
  Critic loss: $L_V = \|V_\phi - \text{Targets}\|^2$
  Update $\theta, \mathbf{\Phi}$ with optimizer under KL trust region
  return $\theta_i, \mathbf{\Phi}$
end procedure
procedure DECENTRALIZED_EXECUTION($\pi_{\theta_i}^\star$)
Output: feasible actions per slot to SGN
  for each slot, each satellite i **do**
    Sample $a_i \sim \pi_{\theta_i}$; $a_i \leftarrow$ PROJECTACTION($a_i, P_i, \mathbf{\Phi}_i, \mathbf{h}_{ij}, \mathbf{g}_i$); actuate
  end for
end procedure

separation between satellites is estimated as: $d_{ij}^{\text{pred}} = \|\mathbf{p}_{s_i}(t+\delta_t) - \mathbf{p}_{s_j}(t+\delta_t)\|_2$ using the Keplerian model. If d_{ij}^{pred} falls below the threshold for any $j \neq i$, the original $\Delta\boldsymbol{\theta}_i(t)$ is replaced by the solution to: $\min_{\Delta\boldsymbol{\theta}} \|\Delta\boldsymbol{\theta} - \Delta\boldsymbol{\theta}_i(t)\|_2^2$ with respect to $d_{ij}^{\text{pred}}(\Delta\boldsymbol{\theta}) \ge D_{\min}$ for $\forall j \neq i$. Phase coherence across updates is protected by the constraint: $|\angle(\mathbf{h}_{ij}^H \Delta\mathbf{\Phi}_i(t)\mathbf{g}_i)| \le \Delta\phi_{\max}$, which is implemented through a Lagrange multiplier that uniformly rescales the candidate $\Delta\mathbf{\Phi}_i(t)$ when necessary.

C. Reward Function

The reward translates the coverage objective and all constraints into a single scalar feedback suitable for optimization. A Poisson point process of ground locations, denoted by $\mathcal{P}(t) \subset \mathcal{A}$, is generated at each time slot t. The reward is defined as

$$r(t) = \frac{1}{|\mathcal{P}(t)|} \sum_{p \in \mathcal{P}(t)} \mathbf{1}\Big(\max_i \gamma_i(t,p) \ge \gamma_{\text{th}}\Big) - \lambda_P \sum_{i=1}^{N} \frac{P_i(t)}{P_i^{\max}} - \lambda_\Phi \sum_{i=1}^{N} \max\left(0, \frac{\|\partial_t \mathbf{\Phi}_i(t)\|_F}{\delta_{\max} f_{\text{ris}}} - 1\right) - \lambda_\theta \sum_{i<j} \max\left(0, \frac{D_{\min} - \|\mathbf{p}_{s_i}(t) - \mathbf{p}_{s_j}(t)\|_2}{D_{\min}}\right) - \lambda_H H_c(t) - \lambda_V V_c(t). \tag{16}$$

The first term provides a Monte-Carlo estimation of the fraction of $\mathcal{A}$ with SINR above the threshold γ_{th}. Visibility and SINR requirements in constraints Eq. (14g), as shown at the bottom of page 6, are implicitly embedded, as locations with failing LOS or SINR contribute zero to the indicator. The third term discourages phase-matrix updates exceeding the hardware limit $\|\partial_t \mathbf{\Phi}_i\|_F \le \delta_{\max} f_{\text{ris}}$ from constraint Eq. (14c). When the Frobenius norm falls below the limit, the $\max(\cdot)$ argument becomes negative and no penalty is applied; otherwise, the excess is penalized linearly. The fourth term implements a smooth barrier on inter-satellite distance. The hard requirement $\|\mathbf{p}_{s_i} - \mathbf{p}_{s_j}\|_2 \ge D_{\min}$ from Eq. (14c) is approximated with a differentiable surrogate to aid gradient-based optimization. The LOS constraint $\mathbf{1}_{\text{LOS}}$ and SINR ratio inequality $\gamma_i \ge \gamma_{\text{th}}$ are integrated implicitly into the formulation. Areas that fail to meet visibility requirements or minimum SINR thresholds are assigned zero values in the coverage indicator. Service-continuity constraint Eq. (14e) is enforced through:

$$H_c(t) = \sum_{j=1}^{M} \mathbf{1}\left[\sum_{\tau=t}^{t+\Delta t_{\min}} \left(y_{ij0}(\tau) + \sum_{l=1}^{L} \kappa_l(\tau)\, y_{ijl}(\tau)\right) < \Delta t_{\min}\right], \tag{17}$$

which counts GBUs that would not meet the required dwell time if the current actions were executed. Coupled constraints in Eq. (14) are combined in:

$$\begin{aligned} V_c(t) = &\sum_{i=1}^{N} \Big\| \partial_t \mathcal{B}_i(t) - v_{\text{orb}} \nabla_{\mathbf{p}} \mathcal{B}_i(t) \Big\| \\ &+ \sum_{i=1}^{N} \sum_{j \in \mathcal{U}} \max\big(0, |\Delta\angle_{ij}(t)| - \Delta\phi_{\max}\big) \\ &+ \sum_{j=1}^{M} \max\left(0, \sum_i \left[y_{ij0}(t) + \sum_l y_{ijl}(t)\right] - 1\right) \\ &+ \sum_{i=1}^{N} \max\left(0, \sum_j \left[y_{ij0}(t) + \sum_l y_{ijl}(t)\right] - U_{\max}\right), \end{aligned} \tag{18}$$

where $\Delta\angle_{ij}(t) = \angle\big(\mathbf{h}_{ij}^H \mathbf{\Phi}_i(t)\mathbf{g}_i\big) - \angle\big(\mathbf{h}_{ij}^H \mathbf{\Phi}_i(t-1)\mathbf{g}_i\big)$. Coefficients $\lambda_P, \lambda_\Phi, \lambda_\theta, \lambda_H, \lambda_V$ are initially configured at low values to facilitate exploration of the solution space, before being systematically increased to ensure that high-performance policies inherently conform to operational constraints with substantial reliability. The coverage utility term in Eq. 16 is bounded in $[0,1]$, while each penalty component is formulated in a dimensionless and normalized form (e.g., $\sum_i P_i/P_i^{\max}$, $\max(0, \|\partial_t \Phi_i\|_F/(\delta_{\max} f_{\text{ris}}) - 1)$, and $\max(0, (D_{\min} - \|p_{s_i} - p_{s_j}\|_2)/D_{\min})$). Therefore, $\{\lambda_P, \lambda_\Phi, \lambda_\theta, \lambda_H, \lambda_V\}$ can be interpreted as direct scaling factors that enforce the relative priority of constraint satisfaction with respect to a unit change in coverage reward. In all experiments reported in

Section IV, we adopt a deterministic two-stage penalty-annealing schedule to stabilize learning while ensuring feasibility: during the exploration/warm-up stage (first 20% of training iterations), we set $(\lambda_P, \lambda_\Phi, \lambda_\theta, \lambda_H, \lambda_V) = (0.001, 0.005, 0.005, 0.001, 0.001)$; during the remaining iterations, the coefficients are linearly increased and then fixed at $(\lambda_P, \lambda_\Phi, \lambda_\theta, \lambda_H, \lambda_V) = (0.01, 0.05, 0.05, 0.01, 0.01)$. The latter values are used for the converged policies and for all evaluations shown in Section IV, which makes the reported results directly reproducible under the same simulation settings.

D. Policy and Critic Networks

In the scheme, each satellite implements a policy network that transforms local observations into corresponding actions. The centralized critic network, utilized exclusively during the training phase, processes the complete constellation state to generate value baselines through generalized advantage estimation (GAE) with proximal policy optimization (PPO) clipping mechanisms. Specifically, for the upstream, the DT located at $\mathcal{G}$ provides unbiased trajectory data by integrating a comprehensive simulator, state estimation using Kalman-style techniques, and a coverage evaluation module to support both offline warm-up phase and online fine-tuning phase.

1) Policy Network: For agent s_i, the observation $\mathbf{o}_i(t) = [\mathbf{p}_{s_i}(t), \mathbf{\Phi}_i(t), P_i(t), I_{\text{inter}}(t), \{\gamma_{ij}(t)\}_j,\ y_{ij0}(t), y_{ijl}(t), z_{ij}(t)]$ is processed into modality-specific streams. Coordinates are transformed into a $2d_{\text{pos}}$–dimensional vector $\text{FF}(\mathbf{p}_{s_i}(t)) = [\sin(2^0\pi\mathbf{p}/h), \cos(2^0\pi\mathbf{p}/h), \ldots, \sin(2^{d_{\text{pos}}-1}\pi\mathbf{p}/h), \cos(2^{d_{\text{pos}}-1}\pi\mathbf{p}/h)]$, where h is set to Earth's radius. The phase vector $\mathbf{\Phi}_i(t)$ and SINR sequence $\{\gamma_{ij}(t)\}$ are processed by two separate gated-recurrent units that summarize the past k time slots, producing hidden states $\mathbf{h}_\Phi$ and $\mathbf{h}_\gamma$. Scalar values including power, interference, and the LOS bitmask $\mathbf{1}_{\text{LOS}}$ are combined into $\mathbf{h}_{\text{scalar}}$. The three embeddings are combined using a multi-head cross-attention mechanism $\mathbf{h}_i(t) = \text{MHA}([\text{FF}, \mathbf{h}_\Phi, \mathbf{h}_\gamma, \mathbf{h}_{\text{scalar}}])$, which allows the network to prioritize the most relevant modality at any given moment. A two-layer multilayer perceptron (MLP) with sigmoid linear unit (SiLU) activations is used to generate the mean vectors $\mu_{\Delta P}$, $\mu_{\Delta\mathbf{\Phi}}$, $\mu_{\Delta\boldsymbol{\theta}}$ and a log-variance $\log\sigma^2$ for Gaussian exploration noise. Binary actions $(y_{ij0}, y_{ijl}, z_{ij})$ are generated by a sigmoid-activated linear layer and sampled using the Gumbel-Softmax technique to maintain gradient flow. Power increments are scaled according to $\Delta P_i(t) = \tanh(\mu_{\Delta P})\min(\Delta P_{\max}, P_i^{\max} - P_i(t), P_i(t))$, ensuring that $P_i(t+1) \in [0, P_i^{\max}]$. Phase updates are projected onto the ℓ_2 ball using $\Delta\mathbf{\Phi}_i(t) = \min\left(1, \frac{\Delta\phi_{\text{global}}}{\|\mu_{\Delta\mathbf{\Phi}}\|_2}\right)\mu_{\Delta\mathbf{\Phi}}$, while $\Delta\boldsymbol{\theta}_i$ is adjusted by an interior-point quadratic-program layer when the predicted separation $\|\mathbf{p}_{s_i}(t+\delta_t) - \mathbf{p}_{s_j}(t+\delta_t)\|_2$ would violate $D_{\min}$.

2) Critic Network: The critic network receives the global graph state $\mathbf{s}_t = \{\mathbf{o}_1(t), \ldots, \mathbf{o}_N(t)\}$ with an edge list $\mathcal{E} = \{(i,j) : \|\mathbf{p}_{s_i}(t) - \mathbf{p}_{s_j}(t)\|_2 < R_c\}$, where R_c represents the communication radius. Contextual features are computed through graph-attention layers:

$$\tilde{\mathbf{h}}_i = \sum_{j\in\mathcal{N}(i)} \alpha_{ij} W \mathbf{h}_j,$$
$$\alpha_{ij} = \frac{\exp\big(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{h}_i \| \mathbf{h}_j])\big)}{\sum_{k\in\mathcal{N}(i)}\exp(\cdot)}, \tag{19}$$

This approach effectively captures interference relationships and cooperation possibilities between agents. The state value is calculated as $V_\phi(\mathbf{s}_t) = \text{ELU}(w^\top \text{TCN}([\tilde{\mathbf{h}}_1;\ \ldots; \tilde{\mathbf{h}}_N]))$, where the weights w are determined through training. A dueling architecture is implemented for advantage computation:

$$Q(\mathbf{s}_t, \mathbf{a}_t) = V_\phi(\mathbf{s}_t) + \frac{1}{N}\sum_{i=1}^{N}\big(A_i(\mathbf{o}_i(t), \mathbf{a}_i(t)) - \tfrac{1}{N}\sum_k A_i(\mathbf{o}_i(t), \bar{\mathbf{a}}_k)\big), \tag{20}$$

The advantage function A_i is implemented as a two-layer MLP with the first layer shared across all agents. This structure enhances learning stability by separating individual agent contributions from the global value baseline.

E. MAPPO Optimization Loop

We adopt MAPPO under the a centralized training with decentralized execution (CTDE) paradigm. The specific data requirements from DT at each iteration are governed by MAPPO's technical components including the clipped surrogate objective, trust region constraints, and advantage estimation approach. Parallel environments generate rollouts using current policies, where each satellite agent processes its local observation $\mathbf{o}_i(t)$ through the policy network to sample actions $\mathbf{a}_i(t) \sim \pi_{\theta_i}(\cdot|\mathbf{o}_i(t))$. The observation $\mathbf{o}_i(t)$ is formed by stacking the last k local frames and by adding exponentially smoothed channel features so that short-term memory is encoded. A global training state $\mathbf{s}_t = \{\mathbf{o}_1(t), \ldots, \mathbf{o}_N(t)\}$ with the ISL graph $\mathcal{E}(t)$ is provided to the centralized critic only during training.

GAE is used for variance reduction. The temporal-difference residual is $\delta_t = r(t) + \gamma V_\phi(\mathbf{s}_{t+1}) - V_\phi(\mathbf{s}_t)$, and the GAE advantage is $\hat{A}_t = \sum_{l=0}^{H-1}(\gamma\lambda)^l\,\delta_{t+l}$. It should be noted that $V_\phi(\mathbf{s}_t) = \mathbb{E}\Big[\sum_{k=0}^{H}\gamma^k\, r_{t+k} \mid \mathbf{s}_t\Big]$. Policy updates follow PPO with clipping. The per-step ratio is $r_t(\theta) = \frac{\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)}{\pi_{\theta_{\text{old}}}(\mathbf{a}_t|\mathbf{o}_t)}$. The clipped surrogate L^{CLIP} is maximized, and the advantage is down-weighted when constraint pressure is high, $\hat{A}_t \leftarrow \hat{A}_t \cdot \exp(-\beta\,\max\{0, V_c(t) - \kappa\})$, where $V_c(t)$ is the composite violation term and κ is the tolerance. The total loss is

$$\begin{aligned} L_{\text{total}} = L^{\text{CLIP}} &+ c_1\left\|V_\phi(s_t) - V_{\text{target}}(t)\right\|_2^2 \\ &+ c_2\,\max\big(0,\ \mathbb{E}[V_c(t)] - \kappa\big) \\ &+ c_3\,\mathbb{E}\big[\text{ReLU}\big(H_c(t) - H_c(t-1)\big)\big], \end{aligned} \tag{21}$$

where the clipped objective per time step is $L^{\text{CLIP}}(\theta) = \min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)$, with $\text{clip}(r, 1-\epsilon, 1+\epsilon) = \min(\max(r, 1-\epsilon),\ 1+\epsilon)$ limits the ratio to $[1-\epsilon,\ 1+\epsilon]$, and $V_{\text{target}}(t)$ is the TD(λ) return, $H_c(t)$ counts dwell-time violations from Eq. (14g), and $c_{1:3}$ balance value, constraint, and handover terms.

F. Dual-Phase Training Framework in DT

The DT-driven training framework employs a dual-phase method-offline warm-up and online fine-tuning-augmented by prioritized experience replay and physical-state synchronization. The DT is employed to systematically record these scenarios and refine the policy based on the data from these scenarios as needed, facilitating multiple simulation iterations until policy convergence is verified. As channel conditions and geometric configurations evolve over time, policies must be adapted while maintaining physical accuracy. State mismatch is controlled through a synchronized twin that maintains alignment between the training environment and actual conditions. The offline warm-up phase is implemented to minimize risk and establish a robust foundation, while online fine-tuning is employed to adapt policies to actual telemetry data.

1) Offline Warm-up Phase: In the offline warm-up phase, the DT's high-fidelity models of Keplerian dynamics and stochastic channel effects are utilized to pre-train policies in a simulated environment. Satellite agent, which trained in the ground DT at the GBS, interact with virtualized satellite networks during this phase, where physics-based simulators govern transition dynamics $P(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$. The full simulate trajectories $\tilde{\tau}_{\text{sim}} = \{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})\}_{t=0}^{T}$ are generated. Then, for a chosen time-index subset $\mathcal{T} \subseteq \{0, 1, \ldots, T\}$, the sampled trajectory used for training is $\tau_{\text{sim}}^{(t)} = \{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}) : t \in \mathcal{T}\}$. The optimal policy $\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right]$ is approximated in a controlled, virtual environment prior to real-world implementation. To maintain physical accuracy, satellite positions $\mathbf{p}_{s_i}(t)$ are updated using a perturbed Keplerian orbital model: $\mathbf{p}_{s_i}(t+\Delta t) = f_{\text{kepler}}(\mathbf{p}_{s_i}(t), \Delta t) + \mathbf{J}_2 \mathbf{f}_{\text{pert}}$, where the two-body orbital motion is solved by $f_{\text{kepler}}(\cdot)$ and Earth's oblateness effects are accounted for by $\mathbf{J}_2\mathbf{f}_{\text{pert}}$. The system dynamics can be represented by a continuous-time transition function $\mathbf{s}_{t+1} = \mathcal{F}(\mathbf{s}_t, a_t) + \omega_t$, where orbital uncertainties are captured by $\omega_t \sim \mathcal{N}(0, \Sigma_{\text{orb}})$. RIS dynamics include phase noise driven by thermal drift and calibration error, formulated as: $\mathbf{\Phi}_i^{\text{sim}}(t+1) = \mathbf{\Phi}_i(t) + \Delta\mathbf{\Phi}_i + \mathcal{N}(0, \sigma_\Phi^2 \mathbf{I})$, where $\Delta\mathbf{\Phi}_i$ is the control signal and $\mathcal{N}(\cdot)$ models hardware-induced randomness.

The offline warm-up is composed of three integrated stages. In the first stage, data collection is performed while maintaining diversity. The second stage is dedicated to the prioritization and delivery of this data for the learning process. Environmental difficulty is regulated in the third stage to ensure the policy is progressively challenged. A feed-forward relationship is established between consecutive stages, while feedback mechanisms allow later stages to influence earlier ones.

In the **first stage**, an initial exploration and projection stage is applied to generate diverse yet feasible behaviour. Each satellite agent s_i samples a full action vector $\mathbf{a}_i(t) \sim \pi_{\theta_i}^{\text{init}}(\cdot)$, where the continuous components are drawn from broad ranges and the binary components are sampled as Bernoulli variables. A differentiable safety projector then enforces actuator, rate, and scheduling rules before the action is executed. Continuous controls are bounded as

$$\Delta P_i(t) \in [-\Delta P_{\max}, \Delta P_{\max}], \quad \|\Delta\mathbf{\Phi}_i(t)\|_F \le \delta_{\max} f_{\text{ris}} \Delta t, \quad \|\Delta\boldsymbol{\theta}_i(t)\|_2 \le \omega_{\max} \Delta t, \tag{22}$$

and post-update values are clipped into $P_i(t+1) \in [0, P_i^{\max}]$ and $\mathbf{\Phi}_i(t+1) = \text{wrap}_{[0,2\pi)}(\mathbf{\Phi}_i(t) + \Delta\mathbf{\Phi}_i(t))$, where $\text{wrap}_{[0,2\pi)}(\cdot)$ applies elementwise modulo 2π to keep each RIS phase in $[0, 2\pi)$. Binary association variables are projected onto the feasible set implied by the unified constraints. Visibility masks are applied:

$$y_{ij0}(t) \leftarrow y_{ij0}(t)\, \mathbf{1}_{\text{LOS}}(s_i, u_j, t), \quad y_{ijl}(t) \leftarrow y_{ijl}(t)\, \mathbf{1}_{\text{LOS}}(s_i, r_l, u_j, t). \tag{23}$$

Exclusive path selection per user and RIS exclusivity are then enforced by keeping, for each user u_j, at most one of $\{y_{ij0}(t), \{y_{ijl}(t)\}_{l=1}^{L}\}$ and by enforcing $\sum_{i=1}^{N} y_{ijl}(t) \le 1$ for each RIS r_l. Per-satellite load is limited by selecting at most $U_{\max}$ users for s_i and setting the remaining association bits to zero $\sum_{j=1}^{M} \left[y_{ij0}(t) + \sum_{l=1}^{L} y_{ijl}(t)\right] \le U_{\max}$. Minimum dwell-time requirements are maintained by implementing a protocol that suspends assignment allocations for $\Delta t_{\min}$ time intervals while controlling the handover indicator $z_{ij}(t)$ to ensure operational stability. Data is subsequently collected through simulations conducted under various orbital configurations and interference conditions, resulting in trajectories $\tau_{\text{sim}}^{(t)} = \{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}) : t \in \mathcal{T}\}$.

In the **second stage**,[2] a prioritized experience replay buffer is used to focus learning on informative and safety-critical data. Let the two replay stores be $\mathcal{B}^{\text{rare}}$ and $\mathcal{B}^{\text{main}}$ with sizes $N_{\text{rare}} = |\mathcal{B}^{\text{rare}}|$ and $N_{\text{main}} = |\mathcal{B}^{\text{main}}|$. The main buffer $\mathcal{B}^{\text{main}}$ stores trajectories with a priority that reflects both learning error and constraint pressure, and buffer $\mathcal{B}^{\text{rare}}$ protects learning against low-frequency but high-impact events. In practice, each trajectory $\tau_{\text{sim}}^{(t)}$ is assigned as

$$p_{\text{main}}(\tau_{\text{sim}}^{(t)}) = p_{\min} + (|\overline{\delta}(\tau_{\text{sim}}^{(t)})| + \epsilon)^{\alpha_\delta} \exp(\beta_V \tilde{V}_c(\tau_{\text{sim}}^{(t)}) + \beta_H \overline{H_c(\tau_{\text{sim}}^{(t)})})(1 + \rho\, \chi_{\text{rare}}(\tau_{\text{sim}}^{(t)})), \tag{24}$$

$$p_{\text{rare}}(\tau_{\text{sim}}^{(t)}) = \text{clip}(1 + (\omega_0 - 1)e^{-t/\omega_\delta}, \omega_{\min}, \omega_{\max}) (|\overline{\delta}(\tau_{\text{sim}}^{(t)})| + \varepsilon)\sigma_{\text{sev}}(\tau_{\text{sim}}^{(t)}), \tag{25}$$

where $\overline{\delta}(\tau_{\text{sim}}^{(t)})$ is the mean TD-error over the episode

$$\overline{\delta}(\tau_{\text{sim}}^{(t)}) = \frac{1}{T} \sum_{t=0}^{T} |r(t) + \gamma V_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}) - V_{\boldsymbol{\phi}}(\mathbf{s}_t)|, \tag{26}$$

and $\overline{V_c(\tau_{\text{sim}}^{(t)})} = \frac{1}{H} \sum_{t=0}^{H} V_c(t)$ is the episode-average constraint penalty and $\overline{H_c(\tau_{\text{sim}}^{(t)})} = \frac{1}{H} \sum_{t=0}^{H} H_c(t)$, β controls sharpness, $\epsilon > 0$ to avoid zero priority, $p_{\min} > 0$ to keep a floor, $\alpha_\delta \in (0, 1]$ to temper very large TD errors, and $\rho \ge 0$ to boost rare events, $\omega_0 > 1$ is the initial boost, ω_δ is

[2]Note that, we model rare events as brief but high-impact link degradations-e.g., rain fade, burst interference, and handover transients-that sharply reduce $\gamma_i(t, \mathbf{p})$ and therefore drive most coverage-threshold violations. Then, we use a rare-event buffer $\mathcal{B}^{\text{rare}}$ with prioritized replay to upweight severe SINR drops/constraint violations and train π_θ and V_ϕ on tail behavior critical to reliability.

a decay constant, $\mathrm{clip}(x,a,b) = \min(\max(x,a),b)$. Typical bounds are $\omega_{\min} = 1$ and $\omega_{\max} \in [5,20]$. The normalized violation term uses the tolerance κ from the MAPPO loop $\tilde{V}_c(\tau_{\text{sim}}^{(t)}) = \frac{1}{\kappa}\max(0, V_c(\tau_{\text{sim}}^{(t)}) - \kappa)$, and $H_c(t)$ is handover-continuity dwell-time penalty counter

$$H_c(t) = \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{1}\Bigg[\sum_{\tau=t}^{t+\Delta t_{\min}-1} y_{ij0}(\tau) + \sum_{l=1}^{L}\kappa_l(\tau)\, y_{ijl}(\tau) < \Delta t_{\min}\Bigg] z_{ij}(t). \tag{27}$$

A separate FIFO rare-event buffer $\mathcal{B}^{\text{rare}}$ captures outliers detected by

$$\chi_{\text{rare}}(\tau_{\text{sim}}^{(t)}) = \mathbf{1}\big[\mathcal{L}_{\text{rain}}(t) > \mathcal{L}_{\text{th}} \vee \|\nabla_t \mathcal{B}_i(t)\|_F > v_{\text{orb}}\big], \tag{28}$$

so severe attenuation or discontinuous coverage evolution are replayed more often. The event severity $\sigma_{\text{sev}}(\tau_{\text{sim}}^{(t)})$ is given as

$$\sigma_{\text{sev}}(\tau_{\text{sim}}^{(t)}) = 1 + \eta_{\text{rain}}\frac{(\mathcal{L}_{\text{rain}}(t) - \mathcal{L}_{\text{th}})_+}{\mathcal{L}_{\text{th}}} + \eta_{\text{cov}}\frac{\big(\|\nabla_t \mathcal{B}_i(t)\|_F - v_{\text{orb}}\big)_+}{v_{\text{orb}}} + \eta_{\text{con}}\overline{V_c(\tau_{\text{sim}}^{(t)})}, \tag{29}$$

with $(x)_+ = \max(x,0)$. The $\eta_{\text{rain}}, \eta_{\text{cov}}, \eta_{\text{con}}$ scale how much each signal boosts priority.

In the **third stage**, the pre-training engine stage is implemented to ensure policy effectiveness under the complete set of physical and scheduling constraints through three distinct stages $\{\mathcal{E}^{(k)}\}_{k=1}^{3}$. All constraints are simultaneously activated, including collision spacing, instantaneous-rate, exclusive path selection load bounds, RIS tuning-rate, beam-coherence, and coverage continuity. In this stage, failures and rare events are anticipated, which enables comprehensive evaluation of the policy prior to transitioning to the online fine-tuning phase.

This stage is structured into three progressive levels. At level $k = 1$, weather effects and satellite motion are eliminated to allow the actor and critic networks to establish fundamental mappings between observations and actions with minimal penalty interference. At level $k = 2$, time-correlated attenuation and LOS masking are introduced, enabling the policy to be trained on tracking dynamic coverage regions $\mathcal{B}_i(t)$ while adhering to rate and tuning constraints during channel variations. Finally, at level $k = 3$, all remaining agents and resource interdependencies are incorporated, necessitating proper scheduling practices and geometric safety considerations while weather conditions and interference patterns continue to evolve.

The offline warm-up produces a set of actor-critic parameters that are pretrained on synthetic but physically realistic data. These parameters serve as a foundation for online adaptation and are subsequently refined during real-world implementation through synchronized experience replay and DT enhancement.

2) Online Fine-Tuning Phase: The online fine-tuning phase has been implemented to address the simulation-to-reality gap by adapting the pretrained policy to operational data. This adaptation includes responses to live telemetry, channel variations, traffic fluctuations, and hardware degradation, while model error is maintained within bounds.

Following deployment of the pretrained policy on the live constellation, the online fine-tuning phase is initiated. A closed loop is established between the physical network and the DT to ensure that model mismatch remains within limits $\epsilon_{\text{sync}}(t) = \|\mathbf{x}_{\text{physical}}(t) - \mathbf{x}_{\mathcal{DT}}(t)\|_2$, with $\sup_t \epsilon_{\text{sync}}(t) \le \epsilon_{\max}$. This bound is maintained through continuous data collection and state updates in each DT. Satellite states, RIS configurations, user positions, and channel measurements are integrated and fed back to the learning loop under the previously defined MAPPO updates. To ensure consistency between estimates and rewards with the optimization problem, the same symbols, constraints, and signal models from the unified formulation are preserved.

For each satellite, a DT $\mathcal{DT}_{s_i}$ is maintained. Raw navigation data $\mathbf{p}_{s_i}^{\text{nav}}$ is processed using EKF that incorporates the orbital process model, with the resulting state represented as $\mathbf{p}_{s_i}^{\mathcal{DT}} \leftarrow \text{EKF}(\mathbf{p}_{s_i}^{\text{nav}})$. The filter integrates with the DT simulator's orbital predictor, aligning estimated movements with orbital-dynamics constraints and ground-track transport bounds in the rewards framework. The inter-satellite link graph $\mathcal{E}(t)$ is updated in DT based on inter-satellite ranging and link-state reports, providing the centralized critic with accurate global state information during the training process.

Each RIS panel is synchronized in $\mathcal{DT}_{r_l}$ by solving a adjustment problem that matches specified phases with measured pilots. The twin phase is formed as

$$\begin{aligned}\mathbf{\Phi}_i^{\mathcal{DT}} &= \mathbf{\Phi}_i \oplus \Delta\mathbf{\Phi}_{\text{cal}},\\ \Delta\mathbf{\Phi}_{\text{cal}} &= \arg\min_{\Delta} \|\mathbf{Y}_{\text{pilot}} - \mathbf{H}(\mathbf{\Phi}_i^{\text{config}} + \Delta)\|_F^2,\end{aligned} \tag{30}$$

where $\oplus$ represents element-wise phase addition on the unit circle, $\mathbf{Y}_{\text{pilot}} \in \mathbb{C}^{N_{\text{rx}} \times N_{\text{pilot}}}$ contains the stacked received pilots, and $\mathbf{H}(\cdot)$ maps a phase vector to its baseband channel response. The solution $\Delta\mathbf{\Phi}_{\text{cal}}$ is used to compensate for thermal drift and hardware bias and is maintained within the rate limit in Eq. (14c) to ensure the DT adheres to the RIS tuning constraint. These synchronized phases are utilized in the computation of instantaneous SINR $\gamma_i(t, \mathbf{p}_{u_j})$ and the coverage regions $\mathcal{B}_i(t)$ that determine the reward.

A DT $\mathcal{DT}_{u_j}$ is maintained for each GBU. GBU positions are determined through simultaneous RSSI readings from multiple satellites and combined using weighted multilateration. The position estimator and its associated uncertainty are calculated as $\mathbf{p}_{u_j}^{\mathcal{DT}} = \text{Triang}(\gamma_{ij}^{\text{RSSI}})$, with $\sigma_{\text{pos}}^2 = \left(\mathbf{J}^\top \mathbf{R}^{-1}\mathbf{J}\right)^{-1}$, where $\mathbf{J}$ represents the Jacobian of the range equations and $\mathbf{R}$ denotes the range noise covariance derived from the RSSI model. These position estimates are temporally filtered, processed through LOS masks $\mathbf{1}_{\text{LOS}}(\cdot)$ from Eq. (14d)–(14g), and integrated into the DT. This ensures that association variables $y_{ij0}(t), y_{ijl}(t), z_{ij}(t)$ are evaluated with accurate geometry and in compliance with the dwell-time rule specified in Eq. (14g). Specifically, $\text{Triang}(\gamma_{ij}^{\text{RSSI}})$ maps simultaneous RSSI to a 3D GBU position by first converting each $\gamma_{ij}^{\text{RSSI}}$ to a range $d_{ij} = f^{-1}(\gamma_{ij}^{\text{RSSI}})$ under a log-distance path-loss model, then solving the multilateration system formed from $(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = d_{ij}^2$ by subtracting a reference satellite s_1 to obtain a linear form $\mathbf{J}\mathbf{p}_{u_j} = \mathbf{b}$. The

weighted least-squares estimate is $\mathbf{p}_{u_j}^{\mathcal{DT}} = (\mathbf{J}^\top \mathbf{W}\mathbf{J})^{-1}\mathbf{J}^\top \mathbf{W}\mathbf{b}$ with $\mathbf{W} = \mathrm{diag}(\sigma_{2j}^{-2}, \ldots, \sigma_{Ij}^{-2})$, and the Cramér-Rao bound reported as the position-error variance is $\sigma_{\mathrm{pos}}^2 = (\mathbf{J}^\top \mathbf{W}\mathbf{J})^{-1}$. Small-scale channels are then reconstructed for active beams using channel quality indicator (CQI) pilots and the DT's a-priori covariances. A minimum mean square error (MMSE) estimator is applied as

$$\mathbf{g}_i^{\mathcal{DT}} = \mathbf{R}_{gy}\,\mathbf{R}_{yy}^{-1}\,\mathbf{y}_{\mathrm{CQI}}, \mathbf{R}_{yy} = \mathbf{H}\,\mathbf{R}_{gg}\,\mathbf{H}^\top + \sigma_n^2\,\mathbf{I}, \quad (31)$$

where $\mathbf{R}_{gg}$ is obtained from the DT's channel engine, $\mathbf{H}$ contains stacked pilot steering vectors, and $\sigma_n^2\mathbf{I}$ represents the receiver noise model. The resulting $\mathbf{g}_i^{\mathcal{DT}}$ is combined with $\mathbf{\Phi}_l^{\mathcal{DT}}$ and $\mathbf{p}_{s_i}^{\mathcal{DT}}$ to calculate $\gamma_i(t, \mathbf{p}_{u_j})$, the aggregate interference I_{inter}, and the coverage indicator in $r(t)$. Through this process, the reward and constraints are aligned with the actual channel conditions while maintaining the RIS-enabled model structure.

Once states are synchronized, on-policy simulations are generated by the system components. The global training state $\mathbf{s}_t = \{\mathbf{o}_1(t), \ldots, \mathbf{o}_N(t)\}$ along with $\mathcal{E}(t)$ is obtained by the critic from DTs. Rewards $r(t)$ are calculated based on the coverage indicator and various penalty terms. MAPPO losses are evaluated, and gradients are applied within the KL bound. The synchronization error $\epsilon_{\mathrm{sync}}(t)$ is continuously monitored by DT. When this error approaches the limit, the telemetry rate is increased and the optimization horizon is reduced until the error falls below $\epsilon_{\max}$. This feedback mechanism enables online tuning while maintaining consistency with the unified constraints Eq. (14b) – (14f).

G. Policy Update Mechanism

Let $\mathcal{B}^{\mathrm{DT}} = (\mathbf{o}_t^{\mathrm{DT}}, \mathbf{a}_t^{\mathrm{DT}}, A_t^{\mathrm{DT}}, \pi_{\mathrm{old}}^{\mathrm{DT}})_{t=1}^{N_{\mathrm{DT}}}$ be a batch of length-H trajectories simulated inside the synchronized DT during the current gradient window. Advantages use GAE with DDT value baseline $V_\phi^{\mathrm{DT}}(\mathbf{s}_t)$:

$$A_t^{\mathrm{DT}} = \sum_{l=0}^{H-t-1} (\gamma\lambda)^l \Big[r_{t+l}^{\mathrm{DT}} + \gamma V_\phi^{\mathrm{DT}}(\mathbf{s}_{t+l+1}^{\mathrm{DT}}) - V_\phi^{\mathrm{DT}}(\mathbf{s}_{t+l}^{\mathrm{DT}})\Big]. \quad (32)$$

The clipped PPO surrogate for these samples is

$$\widehat{J}^{\mathrm{DT}}(\theta) = \frac{1}{N_{\mathrm{DT}}} \sum_{t=1}^{N_{\mathrm{DT}}} \min\Big(\rho_t^{\mathrm{DT}} A_t^{\mathrm{DT}}, \mathrm{clip}\big(\rho_t^{\mathrm{DT}}, 1-\epsilon, 1+\epsilon\big) A_t^{\mathrm{DT}}\Big),$$
$$\rho_t^{\mathrm{DT}} = \frac{\pi_\theta(\mathbf{a}_t^{\mathrm{DT}} \mid \mathbf{o}_t^{\mathrm{DT}})}{\pi_{\mathrm{old}}^{\mathrm{DT}}(\mathbf{a}_t^{\mathrm{DT}} \mid \mathbf{o}_t^{\mathrm{DT}})}, \quad (33)$$

where N_{DT} is chosen larger than the real-data batch to leverage safe, label-rich DT trajectories without overwhelming the update with synthetic statistics. For live interaction, let $\mathcal{B}^{\mathrm{real}} = (\mathbf{o}_t^{\mathrm{real}}, \mathbf{a}_t^{\mathrm{real}}, A_t^{\mathrm{real}}, \pi_{\mathrm{old}}^{\mathrm{real}})_{t=1}^{N_{\mathrm{real}}}$ be the most recent physical trajectories. The advantages again use GAE, now with the physical critic $V_\phi^{\mathrm{real}}(\mathbf{s}_t)$ fit on real rewards:

$$A_t^{\mathrm{real}} = \sum_{l=0}^{H-t-1} (\gamma\lambda)^l \Big[r_{t+l}^{\mathrm{real}} + \gamma V_\phi^{\mathrm{real}}(\mathbf{s}_{t+l+1}^{\mathrm{real}}) - V_\phi^{\mathrm{real}}(\mathbf{s}_{t+l}^{\mathrm{real}})\Big]. \quad (34)$$

The corresponding on-policy objective is

$$\widehat{J}^{\mathrm{real}}(\theta) = \frac{1}{N_{\mathrm{real}}} \sum_{t=1}^{N_{\mathrm{real}}} \min\Big(\rho_t^{\mathrm{real}} A_t^{\mathrm{real}}, \mathrm{clip}\big(\rho_t^{\mathrm{real}}, 1-\epsilon, 1+\epsilon\big) A_t^{\mathrm{real}}\Big),$$
$$\rho_t^{\mathrm{real}} = \frac{\pi_\theta(\mathbf{a}_t^{\mathrm{real}} \mid \mathbf{o}_t^{\mathrm{real}})}{\pi_{\mathrm{old}}^{\mathrm{real}}(\mathbf{a}_t^{\mathrm{real}} \mid \mathbf{o}_t^{\mathrm{real}})}, \quad (35)$$

where N_{real} is selected to track the current sim-to-real gap and the telemetry cadence. Note that $\mathcal{B}^{\mathrm{DT}}$ and $\mathcal{B}^{\mathrm{real}}$ represent on-policy minibatches that are directly utilized in PPO during each update process. These minibatches are temporary, containing the most recent length-H rollouts from both the synchronized DT and the physical constellation. They serve as inputs to the two clipped surrogate functions $\widehat{J}^{\mathrm{DT}}(\theta)$ and $\widehat{J}^{\mathrm{real}}(\theta)$. The gradients from these functions are combined according to the trust factor $\alpha(\epsilon_{\mathrm{sync}})$, allowing the DT batch to have greater influence when alignment between DT and reality is strong, while the real batch becomes more dominant when misalignment increases. Moreover, the $\mathcal{B}^{\mathrm{main}}$ and $\mathcal{B}^{\mathrm{rare}}$ function as long-term replay storage systems that collect trajectories from both sources. Following each rollout, trajectories $\tau_{\mathrm{sim}}^{(t)}$ are added to $\mathcal{B}^{\mathrm{main}}$ with priorities determined by learning signals and constraint pressures (such as TD-error and $\overline{V_c(\tau)}$). Additionally, these trajectories are directed to the capacity-limited FIFO buffer $\mathcal{B}^{\mathrm{rare}}$ when they activate the rare event classifier. For auxiliary updates, sampling is performed by combining the two replay buffers through:

$$P(\tau) = \rho\,\frac{p_{\mathrm{rare}}(\tau)}{\sum_{\tilde{\tau}\in\mathcal{B}_{\mathrm{rare}}} p_{\mathrm{rare}}(\tilde{\tau})} + (1-\rho)\,\frac{p_{\mathrm{main}}(\tau)}{\sum_{\tilde{\tau}\in\mathcal{B}_{\mathrm{main}}} p_{\mathrm{main}}(\tilde{\tau})}, \quad (36)$$

where importance weights $w(\tau) \propto (N \cdot P(\tau))^{-\alpha}$ are applied to control bias. In this formulation, $p_{\mathrm{rare}}(\tau)$ incorporates the rare-buffer enhancement factor in addition to TD-error and constraint penalties, while $p_{\mathrm{main}}(\tau)$ follows the conventional prioritized replay structure.

Every Δt_{grad} seconds, a fully decentralized bandwidth-aware federated step is executed.Each satellite s_i computes two gradients, $\nabla_{\theta_i}\widehat{J}^{\mathrm{DT}}$ and $\nabla_{\theta_i}\widehat{J}^{\mathrm{real}}$, and blends them via the exponential trust factor $\alpha(\epsilon_{\mathrm{sync}}) = e^{-\lambda,\epsilon_{\mathrm{sync}}}$, where ϵ_{sync} measures twin-to-physical mismatch. The local policy is updated by $\theta_i^{(k+1)} = \theta_i^{(k)} - \eta[\alpha(\epsilon_{\mathrm{sync}})\,\nabla_{\theta_i}\widehat{J}^{\mathrm{DT}} + (1-\alpha(\epsilon_{\mathrm{sync}}))\,\nabla_{\theta_i}\widehat{J}^{\mathrm{real}}]$, which yields a smooth shift from DT-driven learning (when DT is accurate) to reality-driven refinement (when mismatch grows). The same CTDE critic is trained with the centralized value loss inside the total PPO loss as Eq. (21) with $r_t(\theta) = \frac{\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)}{\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_t|\mathbf{o}_t)}$. Advantages are damped when constraint pressure is high, where $\hat{A}_t \leftarrow \hat{A}_t \cdot \exp(-\beta\,\max\{0, V_c(t) - \kappa\})$, so that samples with large composite violation $V_c(t)$ contribute less to ascent. The synchronization maintenance system supports the trust schedule and ensures the DT remains usable for gradient shaping. During each telemetry cycle, a predict-update recursion is run by the twin:

$$\hat{\mathbf{x}}_{\mathcal{DT}}(t) = \mathbf{F}\mathbf{x}_{\mathcal{DT}}(t-1) + \mathbf{K}_t\left(\mathbf{y}_t^{\mathrm{tele}} - \mathbf{H}\mathbf{x}_{\mathcal{DT}}(t-1)\right),$$
$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}^\top\left(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}^\top + \mathbf{R}\right)^{-1},$$
$$\mathbf{P}_{t|t-1} = \mathbf{F}\mathbf{P}_{t-1}\mathbf{F}^\top + \mathbf{Q}. \quad (37)$$

Algorithm 2 Overall DTST Scheme

procedure DTST_PIPELINE($\mathcal{S}, \mathcal{R}, \mathcal{U}, f_{\text{kepler}}, \mathcal{L}^{\text{direct}}, \mathcal{L}^{\text{RIS}}$, LOS, π_{θ_i})
Inputs (fixed & models): satellites $\mathcal{S}$, RIS set $\mathcal{R}$, users $\mathcal{U}$; orbital f_{kepler}, channels $\{\mathcal{L}^{\text{direct}}, \mathcal{L}^{\text{RIS}}\}$, LOS masks
Inputs (runtime): telemetry $\mathbf{y}^{\text{tele}}$, physical state $\mathbf{x}_{\text{physical}}$, twin state $\mathbf{x}_{\mathcal{DT}}$
Inputs (trainables): policies $\{\pi_{\theta_i}\}_{i=1}^{N}$, critic V_ϕ
Buffers: $\mathcal{B}_{\text{main}}$, $\mathcal{B}_{\text{rare}}$ (cap C_{buffer}), $\mathcal{B}_{\text{real}}$
Hyperparams/constraints: $\gamma, \lambda_{\text{GAE}}, \epsilon, \eta$; $\alpha(\epsilon_{\text{sync}})$; $P_i^{\max}$, $\delta_{\max} f_{\text{ris}}$, $D_{\min}$, $\theta_{\min}$, $U_{\max}$, $\Delta\phi_{\max}$, $(\epsilon_{\max}, t_{\text{th}})$
Outputs: trained policies $\pi^\star_{\theta_i}$, critic $V^\star_\phi$; feasible actions during online control
(1) Phase I– $(\theta^{\text{warm}}, \phi^{\text{warm}}) \leftarrow$ **Offline_Warmup** $(\pi_{\theta_i}, V_\phi) \leftarrow$ **MAPPO_Step**$(\mathcal{B}_{\text{main}} \cup \mathcal{B}_{\text{rare}}, \pi_{\theta_i}, V_\phi)$
(2) Phase II– $(\theta^\star, \phi^\star) \leftarrow$ **Online_Finetuning** $(\theta^{\text{warm}}, \phi^{\text{warm}}) \leftarrow$ **EKF_Sync**$(\mathbf{x}_{\mathcal{DT}}, \mathbf{y}^{\text{tele}})$
(3) Decentralized Execution: **DECENTRALIZED_EXECUTION**$(\pi^\star_{\theta_i})$ with $\pi^\star_{\theta_i} \leftarrow \theta^\star$
Return: trained policies $\pi^\star_{\theta_i}$ and critic $V^\star_\phi$ ready for live control
end procedure

The matrices $\mathbf{Q}$ and $\mathbf{R}$ are tuned online to minimize the one-step prediction loss:

$$\mathbf{Q} \leftarrow \mathbf{Q} + \eta_Q \nabla_{\mathbf{Q}} \mathbb{E}[\ell_t], \quad \mathbf{R} \leftarrow \mathbf{R} + \eta_R \nabla_{\mathbf{R}} \mathbb{E}[\ell_t],$$
$$\ell_t = \left\| \mathbf{y}_t^{\text{tele}} - \mathbf{H}\hat{\mathbf{x}}_{\mathcal{DT}}(t) \right\|_2^2, \tag{38}$$

which stabilizes the state used by $V_\phi(\mathbf{s}_t)$ and by the DT rollouts. A proportional correction is applied $\mathbf{x}_{\mathcal{DT}} \leftarrow \mathbf{x}_{\mathcal{DT}} + \Gamma \odot (\mathbf{x}_{\text{physical}} - \mathbf{x}_{\mathcal{DT}})$, where $\Gamma = \text{diag}(0.8, 0.6, 1.0)$ and the positions, RIS phases, and power errors are weighted according to their impact on coverage. When $|\mathbf{x}_{\text{physical}} - \mathbf{x}_{\mathcal{DT}}|_2 > \epsilon_{\text{sync}}$ persists for t_{th}, the twin is reloaded with $(\mathbf{x}_{\text{ckpt}}, \mathbf{P}_{\text{ckpt}})$ to quickly re-enter the trusted band. This approach maintains $\alpha(\epsilon_{\text{sync}})$ in a stable state and enables an effective combination of DT and real gradients during learning. An overview of all processes in DTST is shown in Alg. 2 with more details.

H. Dicussion on Overestimation Mitigation

Lemma 1: The conservative-value bound states that the target fed to the critic is systematically lower than the true state-value by at least a safety margin. For every state sampled from the rare-event buffer, $s_{\text{rare}}^{(k)}$, the critic target produced by the conservative update rule satisfies

$$\mathbb{E}[V_{\text{target}}^{\text{cons}}] \le V^\pi(s_{\text{rare}}^{(k)}) - \beta\sigma_G + \mathcal{O}(\epsilon_P\, H\, r_{\max}), \tag{39}$$

where V^π is the true state-value under the current policy, σ_G is the empirical standard deviation of the return ensemble $\{G^{(m)}\}$ obtained from the digital twin, $\beta > 0$ is the conservatism coefficient, ϵ_P measures one-step transition error between the twin and the physical environment, H is the rollout horizon and $r_{\max}$ is the reward bound.

Proof: We first construct the return ensemble $\{G^{(m)}\}_{m=1}^{M}$ by propagating the rare state $s_{\text{rare}}^{(k)}$ through the digital twin using independent disturbance realisations. The conservative target adopts the most pessimistic realisation:

$$V_{\text{target}}^{\text{cons}} = \min_m G^{(m)} - \beta\sigma_G, \sigma_G = \sqrt{\frac{1}{M}\sum_m (G^{(m)} - \bar{G})^2}, \tag{40}$$

where $\bar{G}$ denotes the sample mean. By applying Jensen's inequality to the pointwise minimum, we have $\min_m G^{(m)} \le \mathbb{E}_{P_{\mathcal{DT}}}[G^{(m)}]$. This is because the minimum of a set of random variables is never larger than their expectation. DT transitions differ from physical ones by at most ϵ_P in total-variation distance. Over a horizon H this error accumulates additively, giving $|\mathbb{E}_{P_{\mathcal{DT}}}[G^{(m)}] - V^\pi(s)| \le \epsilon_P H r_{\max}$. Hence, we have

$$\mathbb{E}_{P_{\mathcal{DT}}}[G^{(m)}] = V^\pi(s) \pm \mathcal{O}(\epsilon_P H r_{\max}). \tag{41}$$

Substituting Eq. (41) into Eq. (40) and subtracting the correction term $\beta\sigma_G$ yields

$$\mathbb{E}[V_{\text{target}}^{\text{cons}}] \le V^\pi(s) - \beta\sigma_G + \mathcal{O}(\epsilon_P H r_{\max}), \tag{42}$$

which is the claimed conservative-value bound. ■

Lemma 2: Each training step reduces the expected overestimation bias by an amount proportional to $\eta\Delta_{\text{bias}}$. For the value-function loss that uses the conservative target, $\mathcal{L}^{\text{VF-cons}} = \frac{1}{2}\big(V_\phi(s) - V_{\text{target}}^{\text{cons}}(s)\big)^2$, one gradient step with learning rate $\eta > 0$ produces a new critic that is, in expectation, strictly less optimistic:

$$\mathbb{E}[V_\phi^{\text{new}}(s)] \le \mathbb{E}[V_\phi^{\text{old}}(s)] - \eta\,\Delta_{\text{bias}},$$
$$\Delta_{\text{bias}} = \text{Cov}(V_\phi, V_{\text{target}}^{\text{cons}}) - \text{Var}(V_\phi). \tag{43}$$

Proof: Differentiating the quadratic loss with respect to the scalar output $V_\phi(s)$ yields $\nabla_{V_\phi}\mathcal{L}^{\text{VF-cons}} = 2\big(V_\phi - V_{\text{target}}^{\text{cons}}\big)$. A single SGD step gives

$$\mathbb{E}\big[V_\phi^{\text{new}}\big] = \mathbb{E}\big[V_\phi^{\text{old}}\big] - 2\eta\big(\mathbb{E}[V_\phi] - \mathbb{E}[V_{\text{target}}^{\text{cons}}]\big). \tag{44}$$

Invoking Lemma 1, the conservative target is a pessimistic estimator: $\mathbb{E}[V_{\text{target}}^{\text{cons}}] \le V^\pi - \beta\sigma_G + \mathcal{O}(\epsilon_P H r_{\max})$. Substituting this into Eq. 44, we have

$$\mathbb{E}\big[V_\phi^{\text{new}}\big] \le \mathbb{E}\big[V_\phi^{\text{old}}\big] - 2\eta\Big(\mathbb{E}[V_\phi] - V^\pi + \beta\sigma_G\Big) + \mathcal{O}\big(\eta\epsilon_P H r_{\max}\big). \tag{45}$$

Recognizing the parenthetical term as Δ_{bias} and absorbing constants into η produces the asserted inequality. ■

IV. Performance Evaluation

In this section, the proposed DTST scheme is evaluated by comparing with two baseline schemes. These schemes include conventional MAPPO without DT with 10 or 30 RIS elements. Specifically, we use average data rate and transmission latency to evaluate the performance. We define transmission latency as the duration from data transmission initiation at satellites to successful reception at GBUs. The STNs are built on a satellite-ground topology with ISLs and RIS-enhanced feeder links. The initial transmit power of each satellite is 10W. We evaluate the proposed scheme to the two preceding schemes by changing RIS element count, satellite-user distance, and data size. In addition, we compare against three baselines:

TABLE I
SIMULATION SETTINGS FOR DTST EVALUATION

Category	Parameter	Value / Description
Constellation & Topology	Constellation type	Walker-Delta, 2 shells
	Shell 1 altitude / inclination	$550\,\mathrm{km}/53^\circ$
	Shell 1 planes $\times$ sats/plane	6×3
	Shell 2 altitude / inclination	$570\,\mathrm{km}/42^\circ$
	Shell 2 planes $\times$ sats/plane	4×3
	Total satellites	30-32 (scenario-dependent, Walker-Delta phased)
	Links	Direct Sat-GBU; Sat-RIS-GBU
Deployment	Service region	Lat $[18^\circ, 54^\circ]$, Lon $[73^\circ, 135^\circ]$
	GBS	Beijing (MinEl: 5°)
	GBUs	50
	RIS panels	8
	RIS elements	64 programmable elements / panel
Radio & Channel	Carrier frequency	Ka-band, $28\,\mathrm{GHz}$
	Tx power (satellite)	$20\,\mathrm{W}(43\,\mathrm{dBm})$
	Antenna gains (sat / user)	$35\,\mathrm{dBi}/25\,\mathrm{dBi}$
	Bandwidth	$100\,\mathrm{MHz}$ (system-level), $20\,\mathrm{MHz}$ (link-level)
	Noise figure	$5\,\mathrm{dB}$
	Fading	Shadowing (log-normal), Rician multipath
	Weather loss (per scenario)	Clear / Light rain / Heavy rain / Snow
	$\lambda_P, \lambda_\Phi, \lambda_\theta, \lambda_H, \lambda_V$	0.01, 0.05, 0.05, 0.01, 0.01

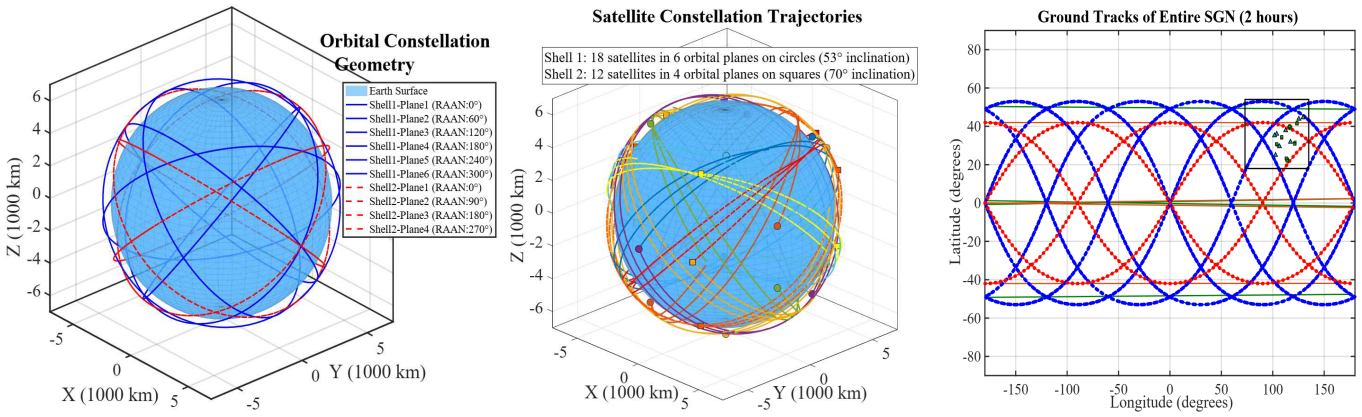
Note: Atmospheric loss is applied per ITU practice with rain, cloud/fog components and elevation dependence, so results reflect P.618/P.840/P.676-style attenuation. Weather losses and sky noise considerations follow ITU-R P.618 workflows.

(1) Direct-link only, which uses the satellite-to-GBU direct link without RIS assistance under the same channel, weather attenuation, and interference settings; (2) RIS-enhanced, with no learning, which enables Sat-RIS-GBU links but sets control variables using the same fixed or heuristic configuration as in the system setup; and (3) MAPPO, no DT, which enables multi-agent learning with the same observation/action spaces and reward but removes DT-driven synchronization and conservative value calibration. Our proposed DTST scheme is the full framework that incorporates DT synchronization, dual-phase training, and conservative value calibration for rare events. The simulation settings is shown in Tab. I. Specifically, The SGN scenario is implemented in MATLAB R2023b with the Satellite Communications Toolbox entities (GBS/GBUs/RIS) are created via groundStation, satellite states are queried in ECEF, and a two-shell Walker-Delta constellation (18 sats at 53° and 12 sats at 42°, 32 LEO in total) is used with fixed RNG seeds. Experiments run on an Apple M3 Max workstation (16-core CPU, 40-core GPU, 48 GB unified memory) under macOS 26.2 with MATLAB R2023b; the Phased Array System Toolbox is used when available with a fallback otherwise.

The computational cost of DTST is dominated by environment-side coverage/SINR evaluation and DT updates rather than by the forward inference of the learned networks. With N_{sat} satellites, L RIS panels (each with K phase variables), and N_{samp} sampled GBU locations per slot, one control step has complexity $\mathcal{C}_{\mathrm{step}} = \mathcal{O}(N_{\mathrm{samp}} N_{\mathrm{sat}}(1+\bar{L}) + LK)$, where $(1+\bar{L})$ accounts for direct-link evaluation plus $\bar{L}$ average RIS-assisted candidates. Online deployment requires only one policy forward pass and constraint validation, with $\mathcal{C}_{\mathrm{infer}} = \mathcal{O}(N_{\mathrm{sat}} \cdot \mathrm{FLOPs}(\pi_\theta) + LK)$, while the training rollout cost scales linearly with the number of episodes and horizon as $\mathcal{C}_{\mathrm{train}} = \mathcal{O}(N_{\mathrm{ep}} T_{\mathrm{ep}} [N_{\mathrm{samp}} N_{\mathrm{sat}}(1+\bar{L}) + LK]) + \mathcal{C}_{\mathrm{update}}$, where $\mathcal{C}_{\mathrm{update}}$ denotes the additive PPO/MAPPO optimization overhead (minibatch and gradient updates).

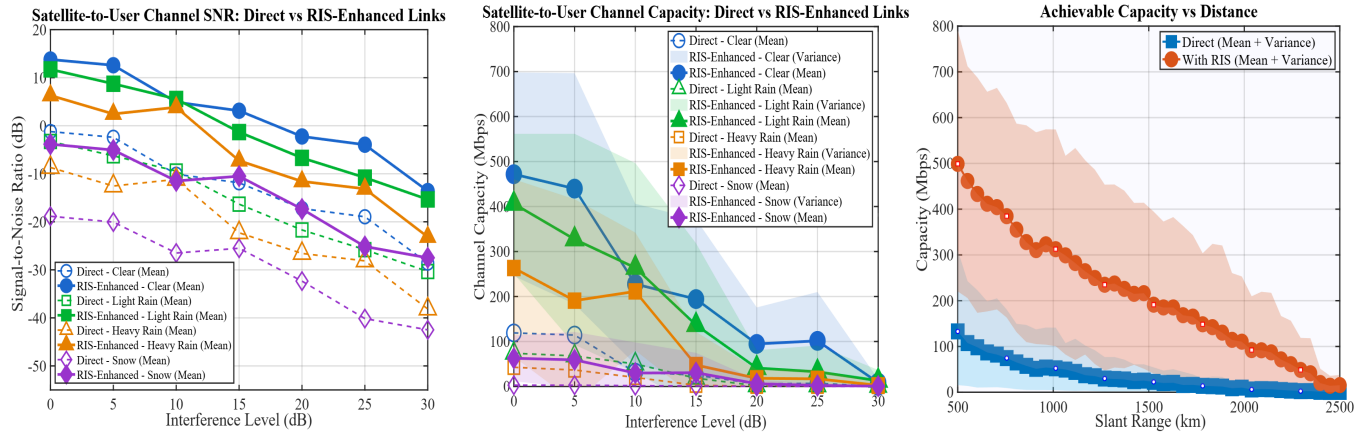
Figure 3 shows the SGN simulation scenario, displaying orbital paths, constellation geometry, and ground tracks, that demonstrating coverage patterns, satellite distribution, and movement across the Earth's surface throughout the simulation period. Fig. 3a shows the three-dimensional constellation in Earth-Centered coordinates. Shells and planes are separated by RAAN to show relative phasing and inter-plane spacing, which enables collision-safe coverage distribution. Fig. 3b shows the two-hour ground tracks of all satellites plotted on a latitude-longitude grid. This visualization demonstrates revisit periodicity, cross-track spacing, and areas where footprints overlap, which are essential for facilitating handovers and diversity. In Fig. 3c, a detailed ground track patterns of the SGN is illustrated across the Earth's surface during the simulation period.

Figure. 4a shows the relationship between satellite-to-user SNR and interference levels for both direct and RIS-enhanced connections across various weather conditions (clear, light rain, heavy rain, and snow). For each time slot t and each active satellite-GBU pair (s_i, u_j), we compute the composite channel by combining the direct path and the RIS-assisted path according to Eq. (1)-(2), where the atmospheric attenuation term $A_{\mathrm{atm}}(\cdot)$ models the clear/rain/snow conditions. The instantaneous effective SINR is then evaluated by Eq. (11), with $\sigma^2 = N_0 B$ and aggregate co-channel interference power I_{inter}. In Fig. 4a, the x-axis "interference level" corresponds to the interference-to-noise operating point, and the y-axis reports the resulting link-quality in dB obtained from Eq. (11) under the same $(\sigma^2 + I_{\mathrm{inter}})$ denominator (direct-only versus RIS-enhanced combining). In Fig. 4b and Fig. 4c, the channel capacity is computed by mapping the same SINR to an achievable rate using $C = B \log_2(1+\gamma)$, and the plotted curves show the sample mean (and variance when indicated) over the simulated link realizations. A consistent improvement in SNR is observed when RIS is utilized, with particularly significant enhancements being noted during adverse weather



(a) Orbital Constellation Geometry (b) Satellite Constellation Trajectories (c) Ground Tracks of SGN

Fig. 3. The SGN snapshot, orbital constellation, and ground tracking patterns.



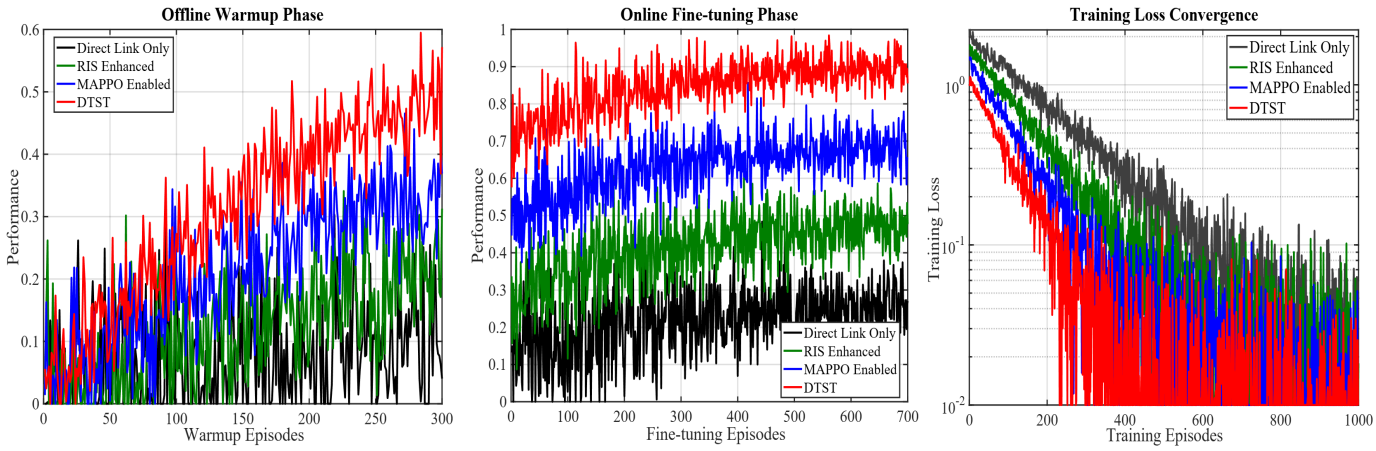
(a) RIS SNR Comparison (b) Direct RIS Capacity (c) Satellite Constellation Trajectories

Fig. 4. RIS-assisted superiority across conditions.

conditions. In Fig. 4b, it is shown to demonstrate that the average capacity is consistently improved and robustness against interference and weather conditions is enhanced when RIS is implemented. These findings support the decision to incorporate RIS technology in the satellite-ground link. In Fig. 4c, downlink capacity versus slant range is presented, where a comparison is made between direct satellite links and RIS-assisted links. The mean and variance are displayed for both connection types. Higher capacity is maintained by the RIS-assisted links, particularly when distances are increased. Note that the SNR in Fig. 4a follows Eq. (11) so a turning point appears when the operating regime shifts from noise-limited to interference-limited, where $\gamma_i \propto 1/I_{\text{inter}}$. Curve intersections arise because atmospheric attenuation affects the direct and RIS paths differently and the RIS gain depends on instantaneous coherent combining. We verified smoothness by re-running the interference sweep with finer steps and more channel realizations, which produced smooth curves with the same ordering, confirming the knees are regime-transition effects rather than plotting artifacts.

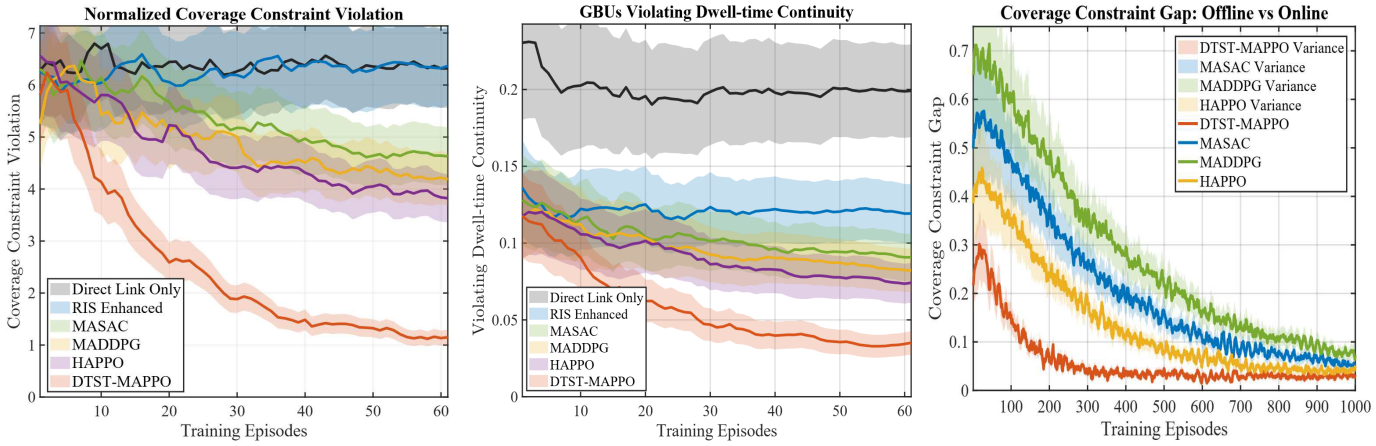
Figure 5a shows the offline warmup performance for the four schemes. Fig. 5b shows the normalized performance versus fine-tuning episodes for Direct Link Only, RIS Enhanced, MAPPO Enabled, and DTST, with DTST achieving the highest score and the fastest convergence. For each training episode e with horizon T, we compute the episode-average return as $\bar{R}^{(e)} = \frac{1}{T}\sum_{t=0}^{T-1} r(t)$, where $r(t)$ is the unified reward in Eq. (16) that combines the coverage utility term with the penalty terms for power usage, RIS tuning-rate, inter-satellite separation, and continuity/visibility via $H_c(t)$ and $V_c(t)$. To present a bounded learning curve, we report a normalized performance score $P^{(e)} = \frac{\bar{R}^{(e)} - \bar{R}^{\text{init}}}{1 - \bar{R}^{\text{init}}}$, where $\bar{R}^{\text{init}}$ is the mean episode-average return of the randomly initialized policy over the first 10 episodes of the corresponding phase. Since the coverage utility term in Eq. (16) is upper-bounded by 1 and the penalties are non-positive, $P^{(e)} \in [0, 1]$ provides a consistent, reproducible scale for comparing convergence speed and final policy quality across schemes. Fig. 5c shows the performance of training loss versus episodes for the four schemes, where DTST converges the fastest to the lowest loss.

Figure 6 shows two episode-average constraint indicators to illustrate how feasibility improves during learning. In Fig. 6a, it shows the normalized coverage constraint violation $\tilde{V}_c^{(e)} = \bar{V}_c^{(e)}/\kappa$ versus training episodes, where $\bar{V}_c^{(e)} = \frac{1}{T}\sum_{t=0}^{T-1} V_c(t)$ and $V_c(t)$ measures the shortfall from the coverage target; κ is the tolerance used in the constrained MAPPO objective. Fig. 6b shows the fraction of GBUs violating dwell-time continuity $\tilde{H}_c^{(e)} = \bar{H}_c^{(e)}/M$ versus training episodes, with



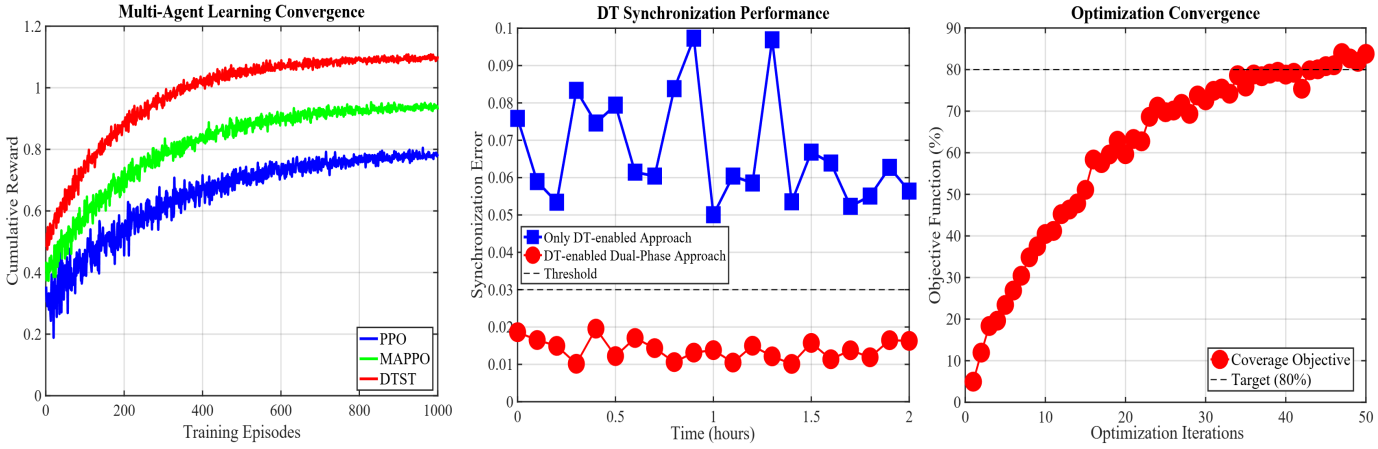
(a) Offline Warmup (b) Online Fine-Tuning (c) Loss Convergence

Fig. 5. The accuracy and training performance of the proposed DT framework.



(a) Coverage Constraint Violation (b) GBUs Violating Dwell-time Continuity (c) Coverage Constraint Gap

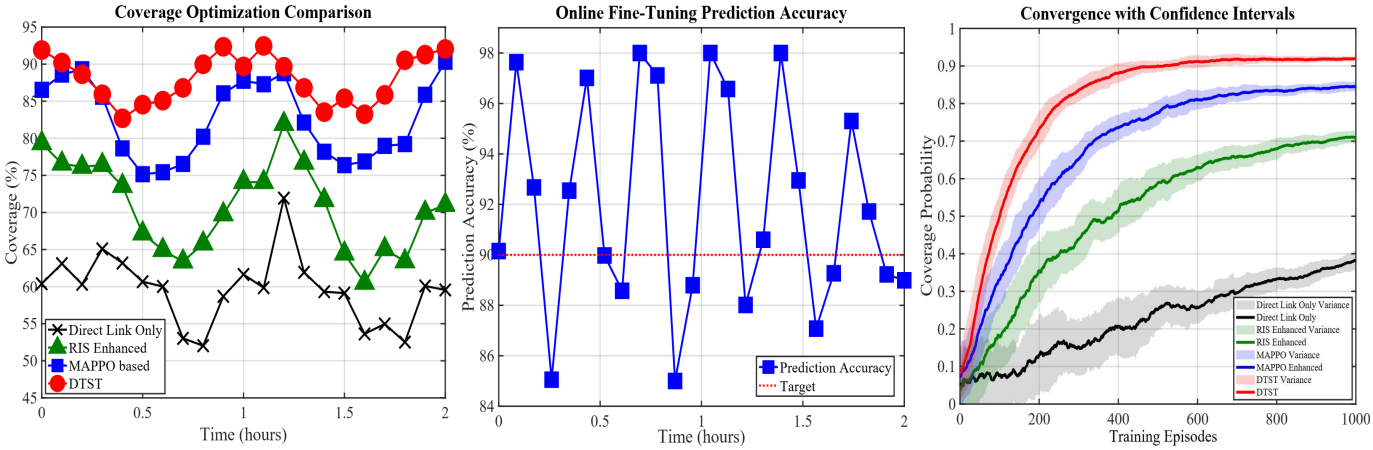
Fig. 6. Constraint indicators vs training episodes.



(a) Learning Curve Comparison (b) Twin Sync Performance (c) Optimization Convergence Curve

Fig. 7. Training reward, sync error and coverage convergence of the proposed DTST.

$\bar{H}_c^{(e)} = \frac{1}{T}\sum_{t=0}^{T-1} H_c(t)$ and $H_c(t)$ the number of handovers at step t. Both figures compare six schemes on the same SGN, orbital, and channel setup: Direct Link Only and RIS Enhanced (no learning, roughly flat); and four RL algorithms-MASAC, MADDPG, HAPPO, and the proposed DTST-MAPPO (constraint-aware). As training progresses, the proposed DTST-MAPPO drives both $\tilde{V}_c^{(e)}$ and $\tilde{H}_c^{(e)}$ down more quickly than the baselines, while the other RL methods improve more slowly. The shaded bands around each curve indicate the variability of these indicators. Fig. 6c the coverage



(a) Coverage Probability (b) Kalman Prediction Accuracy (c) Convergence with Confidence

Fig. 8. The performance of coverage, training loss and convergence of DTST.

constraint gap $\Delta\tilde{V}_c^{(e)} = \tilde{V}_{c,\text{offline}}^{(e)} - \tilde{V}_{c,\text{online}}^{(e)}$ versus training episodes for DTST-MAPPO, MASAC, MADDPG, and HAPPO, with shaded regions indicating variability. Across all methods the gap is positive and decays over training, meaning offline training consistently incurs more coverage violations than online training, though the difference shrinks as learning progresses. DTST-MAPPO maintains the smallest and fastest-vanishing gap throughout, while MADDPG and MASAC exhibit substantially larger gaps (and wider variance) before converging.

In Fig. 7a, cumulative reward is plotted against training episodes for PPO, MAPPO, and DTST. The results indicate that MAPPO and DTST algorithms achieve convergence more rapidly and attain higher reward values compared to PPO. Fig. 7b shows the synchronization error against time, comparing an "Only DT-enabled" baseline with a "DT-enabled dual-phase" scheme against a fixed threshold. The dual-phase approach maintains consistently lower error throughout the two-hour simulation period. Fig. 7c shows the coverage objective rising across optimization iterations and surpassing the 80% target, indicating stable and effective convergence of the proposed method. In Fig. 8a, coverage percentage versus time is present over a two-hour period for four schemes: Direct Link Only, RIS-Enhanced, MAPPO-based, and DTST. The DTST scheme is shown to achieve the fastest rise and highest final coverage. In Fig. 8b, the prediction accuracy versus time is presented over a two-hour online fine-tuning window, with a target line indicating the desired threshold and the measured accuracy tracking above this goal. Fig. 8c shows the coverage probability versus training episodes with confidence bands for the four schemes, where DTST converges fastest and to the highest coverage while maintaining the smallest variance.

V. Conclusion

In this paper, we have proposed a DT-based coverage optimization scheme, called DTST, in RIS-enabled STNs. In particular, we have developed spatial-temporal coverage grain dynamics to accurately model the coverage-power trade-off under dynamic constraints and proposed distributed DT synchronization with conservative value calibration to maintain model-reality alignment while mitigating sim-to-real value bias while adhering to safety-critical constraints and stochastic channel variations. Simulation results have demonstrated that the DTST scheme significantly enhances coverage reliability and reduces service disruptions through investigating coverage probability under constraint violation rate during severe attenuation.

References

[1] D. Zhou, M. Sheng, C. Bao, Y. Wang, J. Li, and Z. Han, "Mission-driven resource scheduling in satellite-terrestrial networks: From perspective of collaboration and reconfiguration," *IEEE Trans. Commun.*, vol. 73, no. 8, pp. 1–15, Aug. 2025.

[2] H. Liu, T. Li, F. Jiang, W. Su, and Z. Wang, "Coverage optimization for large-scale mobile networks with digital twin and multi-agent reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18316–18330, Dec. 2024.

[3] X. Hu et al., "Performance analysis of end-to-end LEO satellite-aided shore-to-ship communications: A stochastic geometry approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11753–11769, Sep. 2024.

[4] R. Wünsche, M. Krondorf, and A. Knopp, "Investigations of channel capacity loss in LEO satellite systems using phased array beamforming antennas," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, no. 4, pp. 5373–5394, Aug. 2024.

[5] C. Jiang et al., "A hybrid framework of RIS-assisted robust secure transmission design for multibeam satellite communications," *IEEE Trans. Veh. Technol.*, vol. 74, no. 4, pp. 6255–6269, Apr. 2025.

[6] H. Yang, D. Huang, K. Lin, C. Huang, and Z. Xiong, "Aerial hybrid active-passive reconfigurable intelligent surface-assisted secure communications for integrated satellite-terrestrial networks," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 8194–8209, 2025.

[7] Z. Li, S. Han, W. Meng, C. Li, and J. I. Leon, "Realistic cooperative strategies based on dynamic spectrum sharing for integrated satellite-terrestrial networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 12, pp. 74–87, 2026.

[8] B. Di, Y. Zhang, R. Deng, M. Wang, S. Sun, and L. Song, "Holographic metasurfaces for extremely large-scale MIMO communications: Design, implementation, and experiment results," *IEEE Wireless Commun.*, early access, Nov. 14, 2025, doi: 10.1109/MWC.2025.3625029.

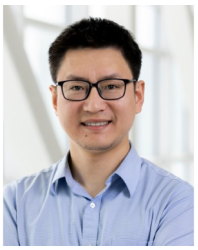
[9] C. Wang et al., "ASTARS aided satellite communications: An adaptive user pairing approach," *IEEE Trans. Commun.*, vol. 73, no. 11, pp. 12788–12802, Nov. 2025.

[10] S. Zhang, H. Yang, F. Wang, J. Si, Z. Li, and T. Q. S. Quek, "Spatio-temporal mixing for computational offloading in satellite edge networks with channel uncertainty," *IEEE Trans. Wireless Commun.*, vol. 24, no. 7, pp. 6151–6165, Jul. 2025.

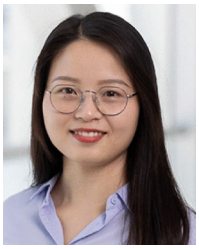
[11] V. Niazmand and Q. Ye, "Joint task offloading, DNN pruning, and computing resource allocation for fault detection with dynamic constraints in industrial IoT," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 5, pp. 1–15, Oct. 2025.

[12] Y. Gong et al., "Multi-modal federated learning based resources convergence for satellite-ground twin networks," *IEEE Trans. Mobile Comput.*, vol. 24, no. 5, pp. 4104–4117, May 2025.
[13] Y. Tao, B. Lei, H. Shi, J. Chen, and X. Zhang, "Adaptive multi-layer deployment for a digital-twin empowered satellite-terrestrial integrated network," *Frontiers Inf. Technol. Electron. Eng.*, vol. 26, no. 2, pp. 246–259, 2025, doi: 10.1631/FITEE.2400327.
[14] N. Ansari, "Toward 6G and beyond: AI-driven synergies across terrestrial, non-terrestrial, and digital twin networks," *IEEE Wireless Commun.*, vol. 32, no. 3, pp. 4–5, Jun. 2025.
[15] Q. Li, Q. Ye, N. Zhang, W. Zhang, and F. Hu, "Digital-twin-enabled industrial IoT: Vision, framework, and future directions," *IEEE Wireless Commun.*, vol. 32, no. 6, pp. 173–181, Dec. 2025.
[16] Q. Li and F. Hu, "Digital twin-enabled channel access control in industrial Internet of Things," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2024, pp. 2149–2154.
[17] R. Zhao, J. Cai, J. Luo, J. Gao, and Y. Ran, "Demand-aware beam hopping and power allocation for load balancing in digital twin empowered LEO satellite networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 6, pp. 5084–5098, Jun. 2025.
[18] W. Liu, Y. Fu, Z. Shi, and H. Wang, "When digital twin meets 6G: Concepts, obstacles, and research prospects," *IEEE Commun. Mag.*, vol. 63, no. 3, pp. 16–22, Mar. 2025.
[19] Y. Yang, W. Sun, J. He, Y. Fu, and L. Xu, "Large generative model-enabled digital twin for 6G networks," *IEEE Netw.*, vol. 39, no. 3, pp. 29–36, May 2025.
[20] W. Liu, Y. Fu, Y. Guo, F. Lee Wang, W. Sun, and Y. Zhang, "Two-timescale synchronization and migration for digital twin networks: A multi-agent deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 17294–17309, Nov. 2024.
[21] Q. Li, J. Chen, M. Cheffena, and X. Shen, "Channel-aware latency tail taming in industrial IoT," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6107–6123, Sep. 2023.
[22] R. Liu, T. H. Luan, Y. Qu, Y. Xiang, L. Gao, and D. Zhao, "Internet of Digital twin: Framework, applications, and enabling technologies," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 3870–3910, 2025.
[23] M. Alishahi, P. Fortier, M. Zeng, Q.-V. Pham, and T. Huynh-The, "Total computational bits maximization for STAR-RIS aided wireless power transfer mobile edge computing networks: TDMA or NOMA?," *IEEE Trans. Veh. Technol.*, vol. 74, no. 2, pp. 3345–3358, Feb. 2025.
[24] L. Zhang et al., "Edge-driven industrial computing power networks: Digital twin-empowered service provisioning by hybrid soft actor-critic," *IEEE Trans. Veh. Technol.*, vol. 74, no. 5, pp. 8095–8109, May 2025.
[25] J. Feng, Z. Ren, and W. Li, "Risk-based dispatch of power systems incorporating spatiotemporal correlation based on the robust soft actor-critic algorithm," *IEEE Trans. Power Syst.*, vol. 40, no. 3, pp. 2478–2491, May 2025.
[26] N. M. Shebani and A. R. Razek, "Prediction of cloud attenuation for 6B arabsat satellite link at Ku, Ka, and V bands over Libya based on ITU-R P. 840–5 model," *Proc. Eng. Technol.*, vol. 38, pp. 58–63, 2018.
[27] *Study on New Radio (NR) to Support Non-Terrestrial Networks*, document TR 38.811, 3GPP, 2020.

Qihao Li (Member, IEEE) received the Ph.D. degree from the University of Oslo, Norway, in 2019. He was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, from 2019 to 2021; and a Lecturer with the School of Electrical Engineering and Intelligentization, Dongguan University of Technology, from 2022 to 2023. Since 2023, he has been an Associate Professor with Jilin University, China. His current research focuses on the Industrial IoT, digital twin, and intelligent agent networking and communication. He received the IEEE/CIC ICCC 2024 Best Paper Award and the IEEE ICCT 2025 Young Scholar Award. He servers/served as the Workshop TPC Chair for IEEE VTC 2025, IEEE Globecom 2024, IEEE Infocom 2024, IEEE CIC/ICCC 2023–2024; and a TPC member for several conferences. He served as an Associate Editor for IEEE INTERNET OF THINGS JOURNAL and a Guest Editor for *Future Internet Journal*.



Qiang Ye (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 2016. Since September 2023, he has been an Assistant Professor with the Department of Electrical and Software Engineering, Schulich School of Engineering, University of Calgary (UCalgary), AB, Canada. Before joining UCalgary, he was an Assistant Professor with the Memorial University of Newfoundland, NL, Canada, from 2021 to 2023, and Minnesota State University, Mankato, USA, from 2019 to 2021. He was with the Department of Electrical and Computer Engineering, University of Waterloo, as a Post-Doctoral Fellow and then a Research Associate (2016–2019). He has published around 80 research papers in top-ranked journals and conference proceedings. He received the Best Paper Award in the IEEE ICCC in 2024 and the IEEE TCCN Exemplary Editor Award in 2023. He also received the Early Career Research Excellence Award from the Schulich School of Engineering, University of Calgary, in 2024. He serves as an Associate Editor for prestigious IEEE journals, such as IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.



Huaqing Wu (Member, IEEE) received the B.E. and M.E. degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and 2017, respectively, and the Ph.D. degree from the University of Waterloo, Ontario, Canada, in 2021. She is currently an Assistant Professor with the Department of Electrical and Software Engineering, University of Calgary, Alberta, Canada. Her current research interests include B5G/6G, space-air-ground integrated networks, Internet of vehicles, mobile/edge computing/caching, artificial intelligence (AI) for future networking. She received the N2Women: Rising Stars in Computer Networking and Communications Award in 2024. She received the Best Paper Awards at IEEE GLOBECOM 2018, *Chinese Journal on Internet of Things* 2020, IEEE GLOBECOM 2022, and IEEE GLOBECOM 2024. She was the Symposium Co-Chair of IEEE GLOBECOM 2024 on Communication QoS, Reliability and Modeling Symposium and the Keynote and Panel Co-chair of IEEE INFOCOM Workshop on Pervasive Network Intelligence for 6G Network from 2022 to 2024. She serves as an Associate Editor in IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING since 2026, IEEE Communications Surveys & Tutorials since 2024, in IEEE NETWORK since 2023, and in *Security and Safety Journal* since 2021.

Fengye Hu (Senior Member, IEEE) received the B.S. degree from the Department of Precision Instrument, Xi'an University of Technology, China, in 1996, and the M.S. and Ph.D. degrees in communication and information systems from Jilin University, China, in 2000 and 2007, respectively. He was a Visiting Scholar in electrical and electronic engineering with Nanyang Technological University (NTU), Singapore, in 2011. He is currently a Full Professor with the College of Communication Engineering, Jilin University. His current research interests include wireless body area networks, wireless energy and information transfer, energy harvesting, cognitive radio, and space–time communication. He is an Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INTERNET OF THINGS JOURNAL, and *China Communications*. He served as an Editor for *IET Communications*.