Received 12 February 2026; revised 9 March 2026 and 3 April 2026; accepted 21 April 2026. Date of publication 28 April 2026; date of current version 11 May 2026.

Digital Object Identifier 10.1109/OJCOMS.2026.3687570

Multi-Modal Foundation Models for Space-Air-Ground Integrated 6G and Beyond Networks: A Survey and Tutorial

WALI ULLAH KHAN [1] (Member, IEEE), MUHAMMAD ADIL [2], JAHANZAIB MALIK [1], CHANDAN KUMAR SHEEMAR [1] (Member, IEEE), SYMEON CHATZINOTAS [1] (Fellow, IEEE), SALMAN A. ALQAHTANI [3], AND KHALID YAHYA [4]

[1] Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, 1855 Luxembourg, Luxembourg
[2] Department of Electronics Engineering, University of Rome "Tor Vergata," Rome 00175, Italy,
[3] New Emerging Technologies and 5G Network and Beyond Research Chair, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia.
[4] Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, 34310 Istanbul, Türkiye

CORRESPONDING AUTHORS: W. U. KHAN (waliullah.khan@uni.lu) AND C. K. SHEEMAR (chandankumar.sheemar@uni.lu)

This work was supported by the Ongoing Research Funding Program—Research Chairs, King Saud University, Riyadh, Saudi Arabia, under Grant ORF-RC-2026-5300.

ABSTRACT Space–air–ground integrated networks (SAGINs) are emerging as a key architectural paradigm for 6G and beyond wireless systems, enabling seamless connectivity by integrating terrestrial radio access networks with aerial platforms and non-terrestrial networks (NTNs). The operation of SAGINs, however, departs significantly from conventional terrestrial assumptions due to high mobility, Doppler dynamics, intermittent connectivity, long and variable round-trip times, gateway bottlenecks, and strong cross-layer coupling. These characteristics motivate intelligence frameworks that leverage heterogeneous contextual information beyond traditional radio frequency (RF) measurements. This paper presents a comprehensive survey and tutorial on multi-modal foundation models (FMs) for SAGIN-centric 6G systems. The survey first introduces a structured taxonomy of SAGIN data modalities, including RF/channel state information (CSI), topology graphs, trajectories and ephemeris, telemetry and traffic traces, weather and atmospheric context, and sensing signals enabled by integrated sensing and communication (ISAC). It then reviews representation and tokenization strategies for heterogeneous wireless modalities and surveys emerging architecture families for multi-modal FMs, including early, late, and hybrid fusion designs, cross-attention mechanisms, mixture-of-experts models, hierarchical multi-timescale architectures, and retrieval-augmented frameworks. The paper further summarizes training paradigms such as masked modeling, cross-modal contrastive learning, and predictive pretraining, together with adaptation strategies including parameter-efficient fine-tuning, continual learning, and federated updates. In addition, the survey discusses system-level considerations for integrating multi-modal FMs into network control loops and deployment environments, covering distributed inference, model compression, intermittency-aware operation, and safety constraints. Finally, the paper outlines benchmarking principles for evaluating multi-modal intelligence in SAGINs and identifies key open research challenges and future directions for building robust, deployable AI-native 6G networks.

INDEX TERMS 6G, space-air-ground integrated networks (SAGIN), non-terrestrial networks (NTN), multi-modal learning, foundation models (FMs).

I. INTRODUCTION

SIXTH-GENERATION (6G) and beyond wireless systems are being shaped around the premise of native intelligence such that, networks are expected to sense their operating conditions, predict spatiotemporal dynamics, and execute decisions with minimal human intervention while meeting stringent requirements on reliability, latency, coverage, and energy efficiency [1], [2], [3]. This shift coincides with two structural transformations [4]: first, the evolution of 6G toward space–air–ground integrated networks (SAGINs), where terrestrial radio access networks (RANs) are augmented by aerial platforms such as uncrewed aerial vehicles (UAVs) and high-altitude platform stations (HAPSs), as well as non-terrestrial networks (NTNs) based on low Earth orbit (LEO) and geostationary (GEO) satellites [5], [6], [7]. Second, the operational context available to the network is becoming fundamentally multi-modal: In addition to classical radio measurements (e.g., channel state information (CSI), reference signal measurements, and beam reports), 6G networks will increasingly exploit topology graphs, traffic and telemetry logs, mobility and orbital states, environmental context (e.g., weather and atmospheric attenuation), and sensing modalities enabled by integrated sensing and communication (ISAC) [8], [9], [10].

The emergence of foundation models (FMs) offers a compelling mechanism to leverage such heterogeneous context at scale. In this paper, we use the term foundation model to denote large-capacity models pre-trained on broad and diverse data sources, which can be adapted to many downstream tasks with limited additional re-training [11], [12]. While large language models (LLMs) are a prominent class of FMs, multi-modal FMs generalize the concept by jointly learning from multiple data modalities (heterogeneous data sources) [13]. For SAGINs, this is particularly attractive because key phenomena that dominate performance are not purely radio-centric [14]. For instance, intermittent connectivity, beam hopping, gateway bottlenecks, and long round-trip times are coupled with mobility and orbit dynamics; atmospheric effects and blockage patterns affect link budgets and beam management; and multi-layer resource orchestration requires reasoning over dynamic graphs and time series. Table 1 lists all the important abbreviations and their definition.

To provide a systematic classification of the multi-modal design space, this work introduces a novel taxonomical structure for SAGIN systems as illustrated in Fig. 1. This unified framework categorizes the field across three fundamental layers: data-modality domains, modality-specific tokenization strategies, and architectural fusion mechanisms, providing a formal classification scheme for multi-modal foundation models in 6G-centric environments.

A. WHY MULTI-MODAL LEARNING IS ESSENTIAL IN SAGINS?

Multi-layer 6G systems violate several assumptions that made terrestrial optimization tractable. The propagation environment can change abruptly due to platform motion and blockage; feedback can be delayed or intermittent; and decision-making spans multiple time scales, from millisecond-level scheduling to second-level mobility and beam management, and minute-to-hour-level satellite resource planning [15], [16]. As a result, optimal actions depend on the context that is only partially observable through radio measurements. Incorporating additional modalities can reduce uncertainty, improve generalization under distribution shift, and enable proactive control [17]. Concretely, orbital ephemeris and platform trajectories provide predictive structure for connectivity; topology graphs encode interference and routing interactions; telemetry and service-level agreement (SLA) traces expose congestion

TABLE 1. Acronyms and abbreviations.

Acronym	Meaning
6G	Sixth-Generation Wireless Networks
AoI	Age-of-Information
BLER	Block Error Rate
CSI	Channel State Information
CVaR	Conditional Value-at-Risk
FM	Foundation Model
GEO	Geostationary Earth Orbit
GNSS	Global Navigation Satellite System
GNN	Graph Neural Network
HAPS	High-Altitude Platform Station
IMU	Inertial Measurement Unit
ISAC	Integrated Sensing and Communication
KB	Knowledge Base
KPI	Key Performance Indicator
LEO	Low Earth Orbit
LiDAR	Light Detection and Ranging
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MAE	Mean Absolute Error
MEC	Multi-access Edge Computing
MDP	Markov Decision Process
MEO	Medium Earth Orbit
MoE	Mixture-of-Experts
MPC	Model Predictive Control
MSE	Mean Squared Error
NTN	Non-Terrestrial Network
OOD	Out-of-Distribution
OSS/BSS	Operations/Business Support Systems
PEFT	Parameter-Efficient Fine-Tuning
POMDP	Partially Observable Markov Decision Process
QoS	Quality-of-Service
RAG	Retrieval-Augmented Generation
RAN	Radio Access Network
RL	Reinforcement Learning
RSRP	Reference Signal Received Power
RTT	Round-Trip Time
SAGIN	Space–Air–Ground Integrated Network
SE	Spectral Efficiency
SLA	Service-Level Agreement
UAV	Unmanned Aerial Vehicle
ViT	Vision Transformer

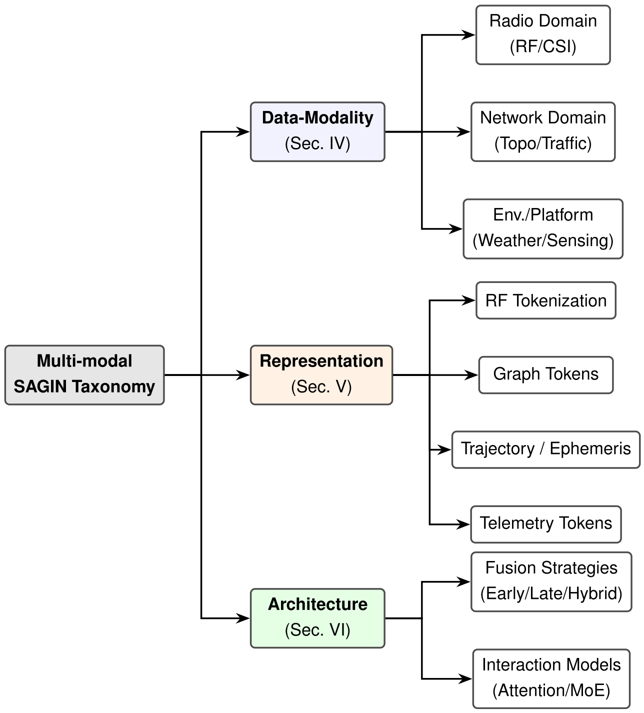

FIGURE 1. **Proposed taxonomy of multi-modal foundation models for SAGIN systems.**

precursors; and sensing signals support situational awareness and digital-twin synchronization [18].

B. FROM PREDICTION TO CLOSED-LOOP CONTROL

The ultimate objective in 6G is not only accurate prediction, but robust closed-loop control. closed-loop intelligence requires: i) state representation learning from heterogeneous streams; ii) planning under constraints with uncertainty; iii) safe execution with defined system boundaries; and iv) continual adaptation as network conditions drift [19], [20]. Multi-modal FMs can contribute at each stage by producing compact representations, enabling Retrieval augmented reasoning from network knowledge bases (KBs), and supporting agentic decision-making that interfaces with existing network controllers and optimization solvers [21], [22]. Fig. 2 illustrates the reference architecture of our SAGIN multi-modal FM pipeline, highlighting the data plane, FM plane, and closed-loop control plane across space-air-ground layers.

C. SCOPE, MAIN CONTRIBUTIONS, AND ORGANIZATION

This paper focuses on multi-modal foundation models for SAGIN-centric 6G and beyond wireless networks, with an emphasis on how to design, train, adapt, deploy, evaluate and maintain such models under real-world constraints. The survey is tutorial in nature: rather than only cataloging existing works, we provide a practical design perspective (e.g., modality tokenization, fusion choices, objective selection, split inference strategies) and a benchmarking methodology aimed at reproducibility. Our scope spans multiple layers (LEO/HAPS/UAV/terrestrial and core/edge), multiple time scales, and both prediction and control tasks.

This paper contributes The contributions of this paper can be summarized in four key points. First, a unified taxonomy is developed for the multi-modal data sources in SAGIN identifying their sources, sampling characteristics, associated costs, and relevance to key 6G tasks. Next, we offer a tutorial on how to represent and tokenize various wireless modalities including RF/CSI, graphs, trajectories, telemetry, and sensing. We then demonstrate how these strategies integrate with multi-modal FM architectures and specific pretraining goals. Third, we present a closed-loop control perspective, that demostrate how multi-modal FMs can be used as representation learners, planners, and teachers for reinforcement learning (RL) across hierarchical controllers, while respecting constraints and safety requirements. Fourth, we propose an evaluation and benchmarking blueprint tailored to SAGINs, including task suites, metrics beyond spectral efficiency (SE), dataset/simulator requirements, and a reporting checklist to enable fair and reproducible comparisons.

This work positions multi-modal FMs as a unified intelligence framework for SAGIN-centric 6G systems, contrasting it against traditional surveys whose primary focus is on wireless optimization, or general developments in FMs. We build systematically, starting from a structured taxonomy of SAGIN modalities, followed by a review of modality-aware representation and tokenization strategies, and then describe how multi-modal foundation models can support prediction and closed-loop network control. We also include a guide and a set of standards for testing how well these systems work in different space-air-ground networks. This helps explain exactly what the paper covers and shows how it serves as a practical guide for building AI that works across space, air, and ground networks.

The remainder of this work is organized as follows. Section II introduces background concepts and notation. Section III reviews SAGIN fundamentals and constraints that motivate multi-modal learning. Section IV categorizes multi-modal data sources in multi-layer 6G. Sections V–VI present representation/tokenization and FM architectures. Sections VII–VIII discuss pretraining and adaptation/continual learning. Sections IX–X cover closed-loop control and distributed deployment. Section XI addresses trust, safety, security, and privacy. Section XII proposes benchmarking and reproducibility guidelines, and Section XIII provides the standardization-oriented insights and open research directions. Section XIV outlines important lessons learned and practical guidelines. Finally, Section XV concludes this survey. Fig. 3 provides a complete structural road-map of the survey, summarizing the organization and logical flow of Sections I–XV and their associated technical themes. The key mathematical symbols used throughout the paper are summarized in Table 2.

II. BACKGROUND AND PRELIMINARIES

This section introduces the background concepts that appear repeatedly in the remainder of the paper. We first clarify

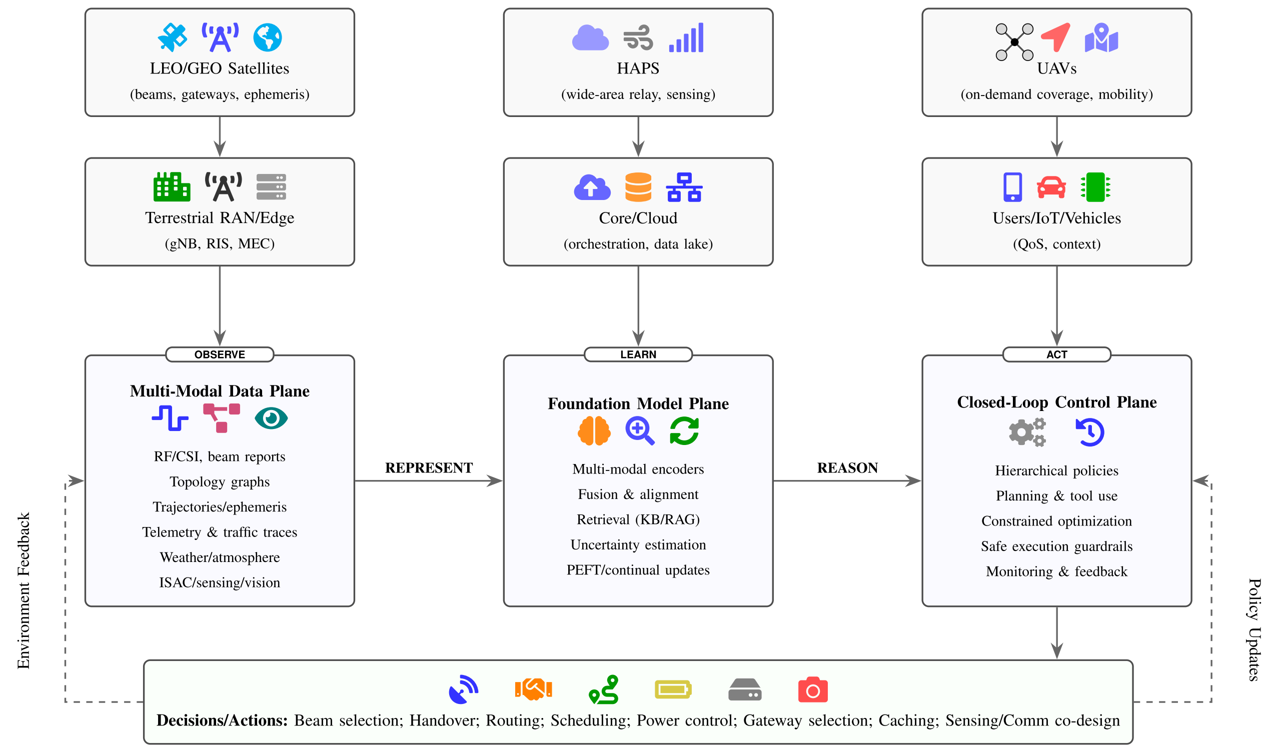

FIGURE 2. Reference architecture for multi-modal foundation model–driven intelligence in SAGIN-enabled 6G networks. The architecture is organized into three interacting planes. The multi-modal data plane aggregates heterogeneous inputs originating from radio-domain, network-domain, and environment/platform modalities (discussed in Section IV). The foundation model plane performs representation learning, tokenization, and cross-modal fusion through modality encoders and multi-modal architectures (Sections V–VI). Finally, the closed-loop control plane uses the learned representations for hierarchical planning, constrained optimization, and network decision-making across space–air–ground layers, as discussed in the control and deployment sections (Sections IX–X). This mapping helps connect the high-level architecture shown in the figure with the detailed discussions throughout the manuscript.

TABLE 2. Key notation used throughout the paper.

Symbol	Description	Symbol	Description
$\mathcal{L}$	Set of SAGIN layers (e.g., LEO, HAPS, UAV, terrestrial)	$\mathcal{M}$	Set of modalities
$\mathcal{M}_t$	Modalities available at time t	$\mathbf{x}_t^{(m)}$	Raw observation at time t from modality m
$\mathbf{o}_t$	Multi-modal observation set at time t	$\mathbf{s}_t$	Latent (partially observed) network state at time t
$\tau_m(\cdot)$	Tokenization function for modality m	$\mathbf{T}_t^{(m)}$	Token sequence for modality m at time t
$N_m(t)$	Number of tokens for modality m at time t	$\mathbf{u}_{t,i}^{(m)}$	i-th token of modality m at time t
$\mathbf{p}_t^{(m)}$	Metadata	$E_m(\cdot)$	Modality-specific encoder for modality m
$\mathbf{H}_t^{(m)}$	Encoded token sequence (embedding) for modality m	$F_\theta(\cdot)$	Fusion module combining $\{\mathbf{H}_t^{(m)}\}$ into a shared represent.
$\mathbf{z}_t$	Learned fused representation at time t	f_θ	Foundation model (backbone) parameterized by θ
g_ϕ	Task-specific head (predictor/control.) parameterized by ϕ	π_ϕ	Control policy mapping representation to actions
$\mathbf{a}_t$	Executed action at time t	$\tilde{\mathbf{a}}_t$	Proposed (pre-verification) action at time t
$\mathcal{A}(\cdot)$	Decision-level aggregation operator (late fusion)	$\alpha_t^{(m)}$	Learned modality weight for modality m at time t
$\mathcal{R}_t$	Reward/utility at time t	$\mathcal{C}_i(\mathbf{s}_t, \mathbf{a}_t)$	i-th constraint function (latency, power, outage, capacity)
$\mathcal{A}_{\text{safe}}(\mathbf{o}_t)$	Safe action set given observations (hard constraints)	$\mathsf{Project}(\cdot)$	Projection/verification operator enforcing safety/feasibility
$\mathcal{D}_{\text{pre}}$	Pretraining dataset (multi-domain, multi-modal)	$\mathcal{D}_{\text{ft}}$	Fine-tuning/adaptation dataset
Δt	Decision time scale (layer- and task-dependent)	γ	Discount factor in sequential decision making
T	Control horizon / episode length	H	Prediction horizon (forecasting steps)
B	Batch size in contrastive learning	τ	Temperature parameter (e.g., InfoNCE contrastive loss)
$f_{\text{D}}(t)$	Doppler shift at time t	$v_{\text{rel}}(t)$	Relative radial velocity between transmitter and receiver
f_{c}	Carrier frequency	c	Speed of light

the terminology around foundation models and multi-modal learning, then formalize how heterogeneous wireless modalities can be represented and fused, and finally summarize the decision-theoretic view that connects prediction

Section I: Introduction

I-A. Why multi-modal learning is essential in SAGINs?
I-B. From prediction to closed-loop control
I-C. Scope, Main Contributions, and Organization

Section II: Background and Preliminaries

II-A. FMs in wireless networking
II-B. Multi-modal representation learning and fusion
II-C. Pretraining, adaptation, and retrieval augmentation
II-D. From prediction to closed-loop control: a decision-theoretic view
II-E. Deployment primitives specific to SAGINs

Section III: SAGIN Fundamentals and Unique Constraints

III-A. Multi-layer SAGIN architecture and control surfaces
III-B. Propagation dynamics, mobility, and Doppler-driven uncertainty
III-C. Backhaul, gateway bottlenecks, and multi-domain resource coupling
III-D. Multi-timescale operation and hierarchical controllers

Section IV: Multi-Modal Data in SAGINs

IV-A. Data planes and where modalities originate?
IV-B. Radio-domain modalities
IV-C. Network-domain modalities: topology, telemetry, and logs
IV-D. Environment and platform modalities: geometry, navigation, weather, and sensing
IV-E. Modality heterogeneity: sampling, reliability, governance, and costs

Section V: Representation and Tokenization for Wireless Modalities

V-A. RF/CSI tokenization
V-B. Graph tokenization for topology and interactions
V-C. Trajectory and ephemeris tokenization
V-D. Telemetry and traffic tokenization
V-E. Weather, atmosphere, and map-based context
V-F. Sensing and ISAC tokenization
V-G. Missing modalities, AoI, and uncertainty as first-class inputs

Section VI: Multi-Modal Foundation Model Architectures

VI-A. Fusion placement: early, late, and hybrid fusion
VI-B. Interaction modeling: cross-attention, gating, and mixture-of-experts
VI-C. Backbone choices for SAGIN multi-modality
VI-D. Hierarchical and cross-layer architectures
VI-E. Retrieval- and tool-augmented architectures

Section VII: Data Curation and Pretraining Objectives for SAGIN Foundation Models

VII-A. Masked modeling and reconstruction objectives
VII-B. Contrastive learning and cross-modal alignment
VII-C. Predictive pretraining and sequence modeling
VII-D. Multi-task self-supervision and objective mixing
VII-E. Data curation: alignment, missingness, and age-of-information
VII-F. Synthetic data and digital twins for SAGIN coverage

Section VIII: Adaptation, Continual Learning, and Personalization

VIII-A. Problem setting and adaptation objectives
VIII-B. Parameter-efficient fine-tuning (PEFT)
VIII-C. Domain adaptation and multi-domain specialization
VIII-D. Continual learning under non-stationarity
VIII-E. Federated and collaborative learning for governance
VIII-F. Safety-aware adaptation

Section IX: Closed-Loop Control with Multi-Modal FMs

IX-A. Control problem abstraction and constraints
IX-B. FM as a representation learner for control
IX-C. FM as a planner or agent
IX-D. FM as an RL teacher: guidance, rollouts, and reward shaping
IX-E. Hierarchical control across layers and timescales
IX-F. Safe control: guardrails and verification

Section X: Distributed Deployment: On-board, Gateway, and Edge Split

X-A. Where inference can run in SAGINs?
X-B. Split inference as a first-class design primitive
X-C. Model compression: quantization, pruning, and distillation
X-D. Intermittency-aware scheduling of compute and communication
X-E. Distributed updates, caching, and versioning

Section XI: Trust, Safety, Security, and Privacy

XI-A. Failure modes in FM-driven SAGIN control
XI-B. Uncertainty estimation, calibration, and OOD detection
XI-C. Safety guardrails and verified execution
XI-D. Security threats: poisoning, backdoors, and adversarial modalities
XI-E. Privacy and governance in multi-domain SAGIN data

Section XII: Benchmarking, Evaluation Protocols, and Reproducibility

XII-A. Benchmark design principles for SAGIN multi-modality)
XII-B. Prediction benchmark suite
XII-C. Closed-loop control benchmark suite
XII-D. Deployment and system benchmarks
XII-E. Metrics: beyond spectral efficiency
XII-F. Baselines and ablations for fair comparison

Section XIII: Standardization, Practical Roadmap, and Open Research Directions

XIII-A. What should be standardized: interfaces over implementations?
XIII-B. Engineering roadmap: from prototypes to operations
XIII-C. Prioritized open research directions

Section XIV: Lessons Learned and Practical Guidelines

XIV-A. Treat data freshness and missingness as first-class inputs
XIV-B. Hybrid fusion is the practical default in heterogeneous 6G stacks
XIV-C. Multi-timescale hierarchy is not optional for closed-loop control
XIV-D. Deployment constraints should shape the architecture from day one
XIV-E. Retrieval and tools reduce hallucination and improve decision grounding
XIV-F. Trust requires explicit safety layers and verification, not only better training
XIV-G. Benchmarks must quantify benefit-per-cost and failure modes, not only average gains
XIV-H. Standardize interfaces and metadata earlier than the models

Section XV: Conclusion

FIGURE 3. Structural roadmap of this survey.

to closed-loop control in multi-layer 6G systems. Table 3 summarizes the key concepts and definitions used throughout the paper.

A. FMS IN WIRELESS NETWORKING

An FM refers to a high-capacity model that is pre-trained on broad and diverse data and subsequently adapted to a variety

TABLE 3. Key concepts and definitions used in this paper.

Term	Meaning in this paper
Foundation model (FM)	Large-capacity model pre-trained on broad data and adapted to multiple wireless prediction/control tasks
Multi-modal FM	FM that jointly learns from heterogeneous modalities (RF/CSI, graphs, trajectories, telemetry, weather, sensing, etc.)
Modality encoder	Mapping $E_m(\cdot)$ that converts modality-specific observations/tokens into embeddings
Tokenization	Mapping $\tau_m(\cdot)$ that converts raw modality inputs into token sequences $\mathbf{T}_t^{(m)}$
Fusion	Aggregation $F_\theta(\cdot)$ that combines modality embeddings into a unified representation $\mathbf{z}_t$
Alignment	Pretraining signal that associates related content across modalities (e.g., contrastive learning, cross-attention-based matching)
Masked modeling	Self-supervised learning by reconstructing masked tokens/patches within or across modalities
Predictive pretraining	Self-supervised forecasting of future observations/representations over a horizon to support proactive control
PEFT	Parameter-efficient fine-tuning that adapts only a small subset of parameters (e.g., adapters, LoRA, prefixes)
LoRA	Low-rank parameterization of weight updates to adapt transformer-like backbones efficiently
RAG	Retrieval-augmented conditioning using a knowledge base/memory to provide external context at inference time
Agentic control	Decision loop where a model plans actions and may call tools/solvers (e.g., routing/optimization) with feedback
Verifier/guardrails	Safety mechanisms that enforce constraints and policies (e.g., feasibility checks, rule filters, uncertainty gating, rollback)
Split inference	Partitioning computation across nodes (on-board/gateway/edge/cloud), exchanging embeddings or state summaries
Federated learning	Distributed training where raw data remain local and only model updates are aggregated across nodes
OOD shift	Out-of-distribution change in data/environment that can degrade model reliability and calibration
AoI-aware fusion	Fusion that explicitly accounts for age-of-information and missingness to reduce staleness-induced errors

of downstream tasks using limited additional retraining [23], [24]. For wireless systems, this definition highlights two key features. First, the model is expected to learn transferable representations that remain useful across changes in deployment conditions (e.g., bands, topologies, platforms, and traffic regimes) [25]. Second, the model should support efficient adaptation mechanisms that respect deployment constraints, such as limited on-board compute, intermittent connectivity, and privacy restrictions, while ensuring compliance with cybersecurity [26].

A multi-modal FM generalizes this notion by learning from multiple input modalities jointly. Unlike unimodal models that treat wireless management as a purely time-series or purely RF learning problem, multimodal FMs can integrate radio measurements with structured context, such as topology graphs, mobility and orbit states, telemetry logs, environmental variables, and sensing observations [27]. This is particularly relevant in SAGINs because key uncertainties (connectivity intermittency, link budget variations, and interference coupling) often have strong explanatory factors that are not captured by RF measurements alone. LLMs are a prominent class of FMs trained on text, but the SAGIN setting benefits from broader FM families such as time-series FMs, graph FMs, and multi-modal transformer architectures [28]. Throughout this paper, we treat language (text inputs) as one possible modality (e.g., network policies, tickets, configuration templates, and operator intents), while the primary emphasis remains on engineering multi-modal learning pipelines for space–air–ground operations.

B. MULTI-MODAL REPRESENTATION LEARNING AND FUSION

Let $\mathcal{M}$ denote the set of modalities available to the network (e.g., RF/CSI, topology graphs, trajectories/ephemeris, telemetry, weather, sensing). At time t, the observation from modality $m \in \mathcal{M}$ is denoted by $\mathbf{x}_t^{(m)}$, where the dimensionality and structure depend on the modality. A common approach is to employ modality-specific encoders that map heterogeneous observations into a shared latent space. Specifically, for each modality m, an encoder $E_m(\cdot)$ produces an embedding

$$\mathbf{h}_t^{(m)} = E_m\left(\mathbf{x}_t^{(m)}\right), \tag{1}$$

and a fusion operator aggregates $\{\mathbf{h}_t^{(m)}\}_{m\in\mathcal{M}_t}$ into a unified representation $\mathbf{z}_t$, where $\mathcal{M}_t \subseteq \mathcal{M}$ denotes the modalities that are available at time t due to intermittency or missing data

$$\mathbf{z}_t = F_\theta\left(\left\{\mathbf{h}_t^{(m)}\right\}_{m\in\mathcal{M}_t}\right). \tag{2}$$

The mapping $F_\theta(\cdot)$ can be instantiated as early fusion (concatenation followed by attention), late fusion (decision-level aggregation), or hybrid fusion (cross-attention between modality streams). In SAGINs, hybrid fusion is often advantageous because modalities have different sampling rates, reliability, and semantics, and cross-attention enables selective conditioning on the most informative modalities under uncertainty [29].

A downstream task head $g_\phi(\cdot)$ uses $\mathbf{z}_t$ to produce predictions or decisions. For example, for a prediction task with output $\widehat{\mathbf{y}}_t$, one may write

$$\widehat{\mathbf{y}}_t = g_\phi(\mathbf{z}_t), \tag{3}$$

while for control tasks a policy π_ϕ maps the representation to an action $\mathbf{a}_t$ as

$$\mathbf{a}_t = \pi_\phi(\mathbf{z}_t). \tag{4}$$

This abstraction is intentionally broad because it accommodates both discriminative objectives (e.g.,

TABLE 4. Modalities across space–air–ground layers and indicative sampling/decision time scales.

Layer / Node	Radio (RF/CSI)	Graph / Topology	Traj. / Ephemeris	Telemetry / Traffic	Weather / Atmos.	Sensing / Vision
LEO/GEO (space)	Beam/RSRP (10–100 ms)	Beam graph (1–10 s)	Orbit state (0.1–10 s)	Load, queues (1–10 s)	Rain fade (10 s–min)	ISAC proxy (limited)
HAPS (air)	CSI, beams (10–100 ms)	Interfer. graph (0.1–1 s)	Trajectory (0.1–1 s)	Traffic, KPIs (0.1–10 s)	Wind, clouds (10 s–min)	ISAC, cameras (10–100 ms)
UAV (air)	CSI, beams (10–100 ms)	Connect. graph (0.1–1 s)	Trajectory (10–100 ms)	KPIs, buffers (0.1–1 s)	Local weather (10 s–min)	Vision/LiDAR (10–100 ms)
RAN/Core (ground)	CSI, CQI (1–10 ms)	RAN graph (10 ms–1 s)	Mobility maps (0.1–10 s)	Logs, SLAs (1 s–min)	Regional forecasts (min–hour)	Cameras/Radar (10–100 ms)

regression/classification for prediction) and generative objectives (e.g., sequence modeling for planning or simulation). Table. 4 summarizes the main modalities available across space–air–ground layers and highlights their indicative sampling and decision time scales.

C. PRETRAINING, ADAPTATION, AND RETRIEVAL AUGMENTATION

Pretraining aims to learn parameters θ such that the representation in (2) is transferable. For multi-modal data, common pretraining families include masked modeling, contrastive learning, and cross-modal alignment. The central idea is to exploit the natural co-occurrence structure across modalities in SAGINs, for example, aligning RF signatures with mobility/orbit context, or aligning topology-induced interference patterns with telemetry-derived congestion signals [30], [31]. When labeled data are scarce, self-supervised pretraining reduces dependence on task-specific annotations and improves robustness under distribution shift [32].

After pretraining, adaptation specializes the model to a target deployment or task. Parameter efficient fine tuning (PEFT) is often preferred in SAGINs because it updates only a small subset of parameters to reduce memory and compute costs [33], [34]. PEFT families include adapter layers, Low Rank Adaptation (LoRA), and prompt/prefix tuning for transformer-like backbones. Continual learning further extends adaptation to time-varying environments, where the model must update without catastrophic forgetting with growing classes.

Retrieval augmented generation (RAG) refers to augmenting an FM with an external KB or memory, retrieving relevant context at inference time and conditioning the model on the retrieved items. In 6G networks, retrieval can provide access to network policies, configuration histories, site metadata, spectrum regulations, or digital-twin snapshots, thereby improving factuality and reducing hallucination-like failure modes when models are used in decision pipelines [35], [36].

D. FROM PREDICTION TO CLOSED-LOOP CONTROL: A DECISION-THEORETIC VIEW

Closed-loop control in SAGINs can be formalized using partially observable sequential decision-making because the true network state is not directly observable and sensing is heterogeneous and intermittent. A common abstraction is a partially observable Markov decision process (POMDP), where the latent state is $\mathbf{s}_t$, the (multi-modal) observation is $\mathbf{o}_t = \{\mathbf{x}_t^{(m)}\}_{m\in\mathcal{M}_t}$, and the action is $\mathbf{a}_t$ [37]. The objective is to maximize expected cumulative utility while satisfying operational constraints. Using the representation $\mathbf{z}_t$ learned by an FM, a policy acts on a compact belief-like summary of the observation history.

Many wireless control problems are inherently constrained. Let $\mathcal{R}_t$ denote the reward (or negative cost) at time t, and let $\mathcal{C}_i(\mathbf{a}_t, \mathbf{s}_t)$ denote the i-th constraint capturing, for example, outage probability, latency, power or backhaul capacity. A generic constrained objective can be written as

$$\max_{\pi_\phi} \quad \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathcal{R}_t\right] \tag{5}$$

$$\text{s.t.} \quad \mathbb{E}[\mathcal{C}_i(\mathbf{a}_t, \mathbf{s}_t)] \le c_i, \quad \forall i, \tag{6}$$

where $\gamma \in (0, 1]$ is a discount factor and c_i is a constraint threshold. In SAGINs, the constraint set is multi-layer and multi-timescale, which motivates hierarchical policies and split controllers, discussed in Section IX and Section X.

E. DEPLOYMENT PRIMITIVES SPECIFIC TO SAGINS

The placement of learning and inference is a first-order design variable in SAGINs. On-board inference reduces latency and improves autonomy, but is limited by compute and energy resources. Gateway or edge inference can exploit stronger compute, but introduces transport latency and can be disrupted by intermittency [38], [39]. Split inference partitions the model between nodes, while federated learning distributes training to keep sensitive data on local devices [40]. These primitives form the system's substrate upon which multi-modal FMs must operate, and they directly influence architectural choices, adaptation strategies, and evaluation metrics.

III. SAGIN FUNDAMENTALS AND UNIQUE CONSTRAINTS

This part reviews how these networks are structured and what sets them apart from standard terrestrial systems. Our goal is

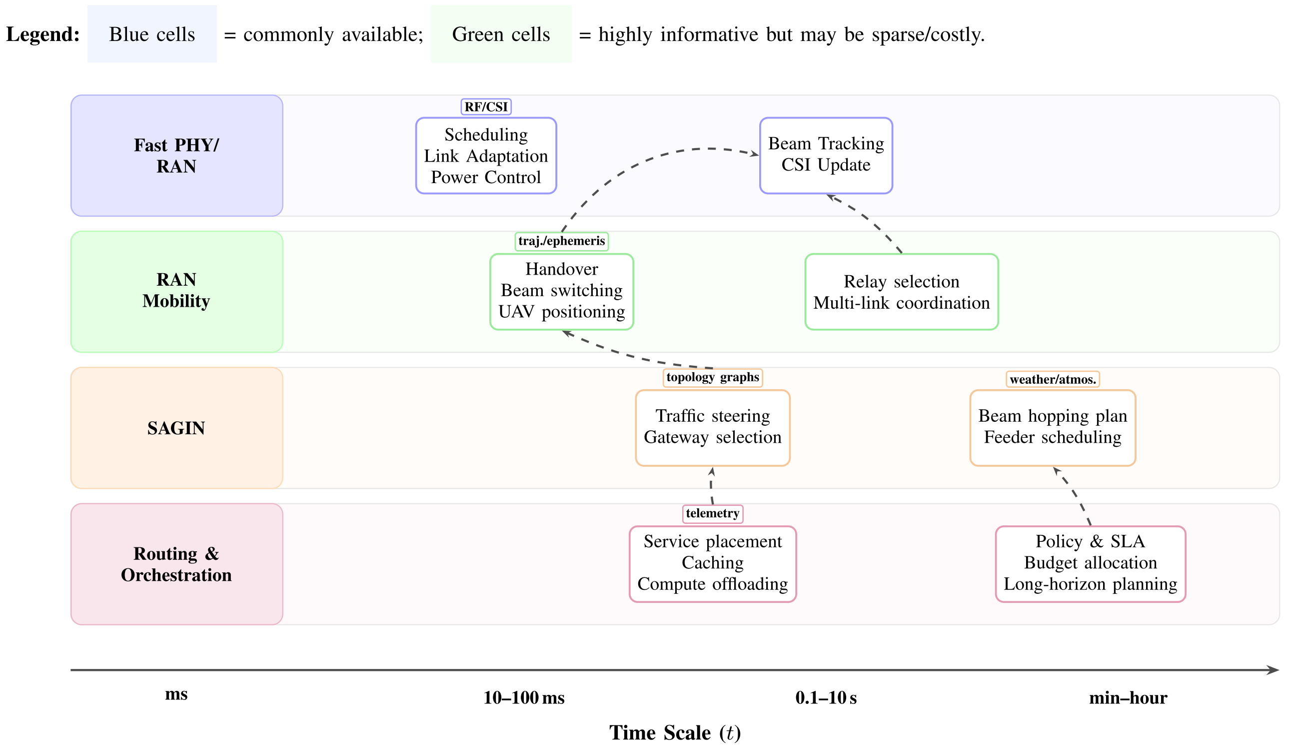

FIGURE 4. Multi-timescale control hierarchy in SAGIN-enabled 6G networks. Fast PHY/RAN decisions (ms–100 ms) rely mainly on RF/CSI measurements, intermediate mobility management (0.1–10 s) uses topology graphs and trajectory/ephemeris information, while longer-term orchestration (minutes–hours) is driven by telemetry, traffic patterns, and environmental context such as weather.

to show why specialized AI models are so useful here, even though they are much tougher to set up in this environment.

A. MULTI-LAYER SAGIN ARCHITECTURE AND CONTROL SURFACES

A SAGIN can be interpreted as a layered architecture comprising space, aerial, and ground segments, each providing complementary coverage and capacity while introducing distinct control interfaces [41], [42]. In the space layer, groups of low-earth orbit (LEO) satellites cover the whole world. However, because these satellites move so fast, the ground stations that connect them to the internet can become 'traffic jams' for both moving data and managing the network [43]. Medium Earth orbit (MEO) and GEO systems may coexist, generally trading reduced mobility for higher round-trip time (RTT). In the aerial segment, HAPS platforms provide quasi-stationary wide-area relaying and sensing, while UAVs offer agile, on-demand coverage and can function as mobile relays or data mules [44]. The ground-based part of the network uses local towers and edge computing to deliver high speeds and quick responses. At the same time, central cloud systems oversee the whole network, enforcing policies and optimizing performance over time [45], [46].

The key design consideration is a control that is both distributed and hierarchical. Decisions like scheduling, link adaptation, and beam tracking are naturally quick whereas others are global and slower (i.e., gateway selection, beam hopping patterns, inter-layer traffic engineering, and service placement [47]). The hierarchical structure of SAGINs motivates the use of representation learning to condense heterogeneous observations into compact, informative state descriptions and to enable consistent decision-making across different timescales.

Fig. 4 depicts the cross-layer control hierarchy in SAGIN, emphasizing the dependencies between PHY/RAN, mobility, and orchestration layers across heterogeneous time scales. The figure is organized into four functional layers, namely fast PHY/RAN, RAN Mobility, SAGIN, and Routing & Orchestration. The lower layers are operating on millisecond-level timescales and use the instantaneous RF/CSI information, the intermediate layer manage mobility, relay coordination, and topology-aware decisions over seconds. Higher layers leverage telemetry, traffic patterns, and environmental context to perform long-horizon service placement, gateway selection, and resource orchestration.

B. PROPAGATION DYNAMICS, MOBILITY, AND DOPPLER-DRIVEN UNCERTAINTY

The propagation dynamics of the SAGIN networks are dominated primarily by platform motion, large coverage footprints, and rapidly varying geometry [48]. These effects are characterized by time-varying link budgets, changing LOS conditions, and non-negligible Doppler shifts. These effects can degrade system-level synchronization and channel estimation if not predicted and compensated [49], [50], [51].

TABLE 5. SAGIN/NTN constraints that break terrestrial assumptions and their implications for multi-modal foundation models.

Constraint	What changes in SAGINs	Implication for multi-modal FMs
High mobility and Doppler	Rapid geometry changes yield strong Doppler and time-varying link budgets; beam footprints shift quickly in LEO and fast UAV motion	Fuse RF/CSI with trajectories/ephemeris to predict channel evolution and support proactive beam tracking and handover
Intermittent connectivity	Visibility windows, blockage, and relay/gateway availability create missing data and disrupted control loops	Use intermittency-robust fusion and representation learning that can operate with partial modalities and delayed feedback
Long and variable RTT	Control messages can arrive late; centralized decisions may become stale; multi-hop paths add delay and jitter	Prefer hierarchical and distributed inference; incorporate RTT and topology context as explicit inputs to avoid unsafe actions
Gateway and feeder bottlenecks	Feeder links and gateway visibility constrain throughput and routing; congestion can dominate access link quality	Retrieve and fuse telemetry/queues and gateway-graph context to couple radio decisions with transport feasibility
Strong cross-layer coupling	Beam/handover decisions affect routing, caching, service placement, and compute offloading; policies must be consistent across domains	Employ unified representations that support multi-task learning and coordinated control policies across layers
Energy and payload limits	On-board compute and sensing are limited on satellites and UAVs; model size and inference cost matter	Use PEFT, compression, split inference, and modality selection under resource budgets
Environment dependence	Weather, atmospheric attenuation, and blockage patterns can dominate performance in some bands and links	Incorporate weather/atmospheric signals and sensing context to improve generalization under distribution shifts
Safety and reliability requirements	Incorrect actions can cause service loss over large areas due to wide coverage; errors propagate across layers	Integrate uncertainty estimation, guardrails, and verification-aware decision pipelines
Data governance and privacy	Multi-operator and multi-domain data sharing is restricted; raw data may not be centralizable	Favor federated and privacy-aware training; use retrieval augmentation with policy-compliant knowledge bases

A common approximation for the Doppler shift is

$$f_{\mathrm{D}}(t) \approx \frac{v_{\mathrm{rel}}(t)}{c} f_{\mathrm{c}}, \tag{7}$$

where $v_{\mathrm{rel}}(t)$ is the relative radial velocity between transmitter and receiver, c is the speed of light, and f_{c} is the carrier frequency. The term $v_{\mathrm{rel}}(t)$ is determined by the orbital state or flight trajectory, which shows that modality streams such as ephemeris, trajectory, and inertial data provide predictive structure that can help providing a predictive structure that cannot be found in a standalone RF system. This coupling provides a strong reason for the multimodal FM-based learning that can reduce uncertainty, enhance reliability and robustness, provide beam management, and mobility control.

Another architectural difference is the prevalence of intermittency and long feedback loops. The RTT depends on the segment and the routing path [52]. A simplified expression for a single-hop propagation delay is d/c, where d is the distance, and multi-hop routing through gateways and relays compounds both propagation and processing delays. In practice, the control loop in SAGINs must cope with delayed, quantized, and sometimes missing observations, which motivates the use of intermittency-aware fusion and belief-like representations rather than a brittle reliance on synchronous CSI [53].

C. BACKHAUL, GATEWAY BOTTLENECKS, AND MULTI-DOMAIN RESOURCE COUPLING

A key characteristic of NTNs is that gateway availability and feeder link capacity can become the primary bottlenecks, particularly when several spot beams share the same feeder resources or when gateway visibility is restricted [54]. Further, the aerial nodes are likely to face energy and payload constraints that limit their operational their ability to host large models or sustain their sensing rate [55], [56], [57], [58]. These constraints leads to multi-domain coupling, i.e., a beam decision affects not only access link performance but also feeder congestion; a routing decision affects RTT and service placement; and an edge inference choice affects both latency and power budget [59]. Consequently, the state required for reliable decision-making extends beyond RF measurements and must include telemetry, queueing indicators, topology graphs, and platform status.

D. MULTI-TIMESCALE OPERATION AND HIERARCHICAL CONTROLLERS

SAGIN intelligence spans multiple timescales. Short timescales include physical-layer adaptation and scheduling, medium timescales include beam switching, handover, and relay selection, and long timescales include satellite resource planning, beam hopping schedules, and service orchestration [60]. Fig. 4 illustrates the multi-timescale control hierarchy in SAGINs and the corresponding modalities that dominate decisions at each timescale. When decisions are made at different rates, the model must either learn a representation that is consistent across timescales or explicitly implement hierarchical control, where slow controllers set constraints or priors for fast controllers [61]. This is also where multi-modal FMs become attractive: modalities such as ephemeris and weather evolve slowly and can inform long-horizon planning, while RF measurements evolve quickly and drive short-horizon actions. Table 5 summarizes the SAGIN/NTN

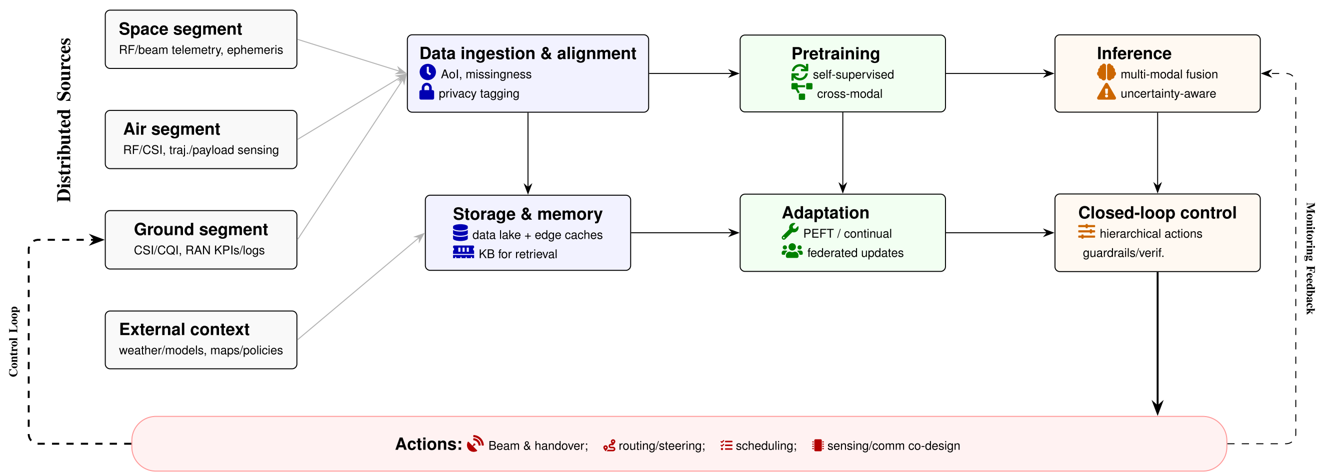

FIGURE 5. Systems view of multi-modal data flow in SAGINs. Distributed sources produce heterogeneous modalities that are time-aligned and tagged(age-of-information, missingness, privacy), stored in a lake/memory/KB, and used for pretraining and continual adaptation. Inference outputs decisions that close the loop through hierarchical controllers and monitoring feedback.

constraints that break terrestrial assumptions and highlights their implications for multi-modal FM design.

IV. MULTI-MODAL DATA IN SAGINS

This section formalizes the data landscape that underpins multi-modal foundation models in 6G and beyond SAGINs. We aim at chatagorizing the existing modalities, their generation points, their evolution, and their functionality (the tasks they enable). In terrestrial network, the radio measurements and traffic counters often dominate learning pipelines, whereas SAGINs provide a richer set of context streams. The availability of such streams is heterogeneous, and their reliability depends on platform dynamics, sensing coverage, and governance constraints [62], [63]. This heterogeneity motivates the need for modality-aware tokenization and fusion (Sections V–VI), along with training and deployment strategies that are robust to intermittent connectivity (Sections VII and X).

A. DATA PLANES AND WHERE MODALITIES ORIGINATE?

Multi-modal observations in SAGINs arise from three broad sources: i) radio-domain measurements collected from access, feeder, and inter-platform links; ii) network-domain context describing topology, load, and performance indicators; and iii) environment and platform context describing geometry, mobility, and external conditions. These sources are not independent. For example, a congestion event can be explained by gateway visibility windows, a beam hopping plan, and traffic spikes; similarly, a sudden drop in link quality may correlate with weather-driven attenuation or a blockage event inferred from sensing. Therefore, a key data-engineering step is to preserve synchrony across modalities at the level needed for downstream tasks, while acknowledging that perfect time alignment is often infeasible due to variable RTT and intermittent links.

From a practical standpoint, the SAGIN data plane can be decomposed into: i) on-board data collected at satellites and aerial nodes (beam and link telemetry, navigation and payload state, and possibly sensing); ii) edge/RAN data collected at terrestrial base stations and edge servers (CSI, measurement reports, local topology and traffic counters); and iii) core/cloud data such as network-wide key performance indicators (KPIs), service logs, and orchestration decisions. KPIs refer to aggregated performance indicators such as throughput, latency, packet loss, and resource utilization over specified intervals. The cross-layer challenge is that on-board data are often the most informative for autonomy but also the most constrained in bandwidth and storage, whereas cloud-side data are easier to store and process but can be delayed or incomplete [64]. Fig. 5 illustrates a systems view of how distributed SAGIN modalities are ingested, aligned (age of information (AoI)/missingness), stored, and used for pretraining, adaptation, and closed-loop inference.

In summary, data ingestion in SAGIN environments must accommodate heterogeneous and geographically distributed data sources originating from space, air, and ground segments. These sources differ significantly in sampling rates, reliability, and availability due to intermittent connectivity and variable network delays. Consequently, effective multi-modal learning pipelines require careful synchronization across modalities, explicit treatment of age-of-information and missingness, and governance-aware data handling mechanisms. Establishing structured data ingestion and alignment processes is therefore essential for enabling reliable multi-modal representation learning and for supporting the modality-specific data analysis discussed in the following subsections.

B. RADIO-DOMAIN MODALITIES

Radio-domain modalities include CSI, reference-signal measurements, beam reports, interference indicators, and link-layer feedback (e.g., block error rate (BLER) and retransmission counts). These modalities are typically high-rate and tightly coupled to physical-layer control, making them indispensable for fast decisions such as scheduling, link adaptation, and beam tracking. However, in SAGINs the interpretability of radio measurements benefits substantially from platform context [65]. For example, CSI variation can reflect not only multipath fading but also deterministic geometry changes due to orbital or aerial motion; similarly, reference-signal received power (RSRP) fluctuations can be driven by beam footprint movement and feeder congestion rather than local fading alone. For multi-modal FMs, this suggests that radio modalities should be treated as conditionally predictable given trajectories/ephemeris and should be paired with slower modalities to improve robustness and reduce sample complexity [66].

C. NETWORK-DOMAIN MODALITIES: TOPOLOGY, TELEMETRY, AND LOGS

The topology and interaction structure are central in SAGINs because interference, routing, and gateway selection depend on who is connected to whom at any time. This motivates graph-based modalities in which nodes represent beams, satellites, gateways, aerial relays, base stations or flows, and edges represent feasible links, interference couplings, or routing adjacencies [67]. Graph modalities are naturally aligned with control problems such as routing, traffic steering, load balancing between layers, and joint beam/door scheduling [68].

Telemetry and logs complement network graphs by providing information on how the network behaves during operation. They include queue lengths, buffer occupancy, congestion window statistics, handover failure counters, service-level monitoring, and orchestration histories. In terms of their rate, they are typically lower rate than CSI but remain highly relevant for control decisions because they reveal transport bottlenecks and policy effects. In addition, text-like logs and configuration histories can form a retrieval corpus for RAG-style conditioning, enabling models to access past incident resolutions, configuration templates, and operational policies in a controlled manner.

D. ENVIRONMENT AND PLATFORM MODALITIES: GEOMETRY, NAVIGATION, WEATHER, AND SENSING

In SAGINs, system performance is strongly influenced by geometric configuration and environmental conditions [69]. The trajectory and ephemeris provide state variables such as position, velocity and attitude for the platforms; these variables can be obtained from navigation systems (e.g., global navigation satellite systems (GNSS)) and inertial measurement units (IMUs) [70]. Such data are typically low-dimensional but highly predictive of connectivity windows, Doppler dynamics, and beam footprint evolution [71]. Weather and atmospheric variables (e.g., rainfall intensity, cloud cover, and attenuation estimates) evolve more slowly, but can dominate link budgets at higher frequencies, and thus influence planning and routing decisions.

Sensing modalities arise from ISAC, radar, cameras, light detection and ranging (LiDAR), or other payload sensors on aerial/terrestrial nodes [72], [73]. They can reveal blockage, terrain, and mobility patterns that are not visible from RF alone [74], [75]. In practice, sensing introduces additional cost and governance constraints: it can be bandwidth-intensive, privacy-sensitive, and intermittently available depending on sensor coverage [76], [77]. For multi-modal FMs, sensing is often best treated as a sparse but high-value modality used for calibration, digital-twin synchronization, and occasional disambiguation of RF events [78].

E. MODALITY HETEROGENEITY: SAMPLING, RELIABILITY, GOVERNANCE, AND COSTS

A recurring theme is that the modalities differ not only in content but also in rate, reliability, and cost [79]. CSI may arrive at millisecond scale but can be noisy and stale under RTT; topology graphs may update at sub-second to second scales but can be incomplete due to visibility changes; telemetry and logs can be delayed and aggregated; weather is slow but uncertain; sensing may be rich but sporadic and sensitive to privacy [80]. Therefore, a robust multi-modal learning pipeline must explicitly encode modality availability (missingness), confidence, and AoI [81], [82]. These aspects should be treated as inputs rather than nuisances, enabling the model to learn when to trust which modality and when to return to conservative decisions. Table 6 summarizes key properties of the modality and connects them to typical tasks and considerations of robustness.

V. REPRESENTATION AND TOKENIZATION FOR WIRELESS MODALITIES

Multi-modal FMs require a common interface between heterogeneous wireless data streams and the model backbone. This interface is the tokenization and representation layer, which converts modality-specific observations into sequences (or sets) of tokens with positional and temporal structure [83], [84]. In SAGINs, representation design is not a cosmetic choice: it determines whether the model can cope with multi-rate sampling, missing modalities, delayed feedback, and cross-layer coupling [85], [86].

We denote by $\mathbf{x}_t^{(m)}$ the raw observation of modality $m \in \mathcal{M}$ at time t. Tokenization maps $\mathbf{x}_t^{(m)}$ into a sequence of $N_m(t)$ tokens

$$\mathbf{T}_t^{(m)} = \tau_m\left(\mathbf{x}_t^{(m)}\right) = \left[\mathbf{u}_{t,1}^{(m)}, \ldots, \mathbf{u}_{t,N_m(t)}^{(m)}\right], \tag{8}$$

where $\mathbf{u}_{t,i}^{(m)} \in \mathbb{R}^{d_m}$ is the i-th token embedding for modality m. A modality encoder $E_m(\cdot)$ then maps tokens into a shared latent dimension d as

$$\mathbf{H}_t^{(m)} = E_m\left(\mathbf{T}_t^{(m)}; \mathbf{p}_t^{(m)}\right), \tag{9}$$

where $\mathbf{p}_t^{(m)}$ denotes structured metadata such as time stamps, AoI, missingness masks, confidence scores, coordinate frames, and sampling rates. In SAGINs, including $\mathbf{p}_t^{(m)}$ is essential because many failure modes arise from staleness

TABLE 6. Multi-modal data taxonomy in multi-layer 6G: sources, rates, costs, privacy sensitivity, typical tasks, and robustness considerations. The listed time scales are indicative and depend on deployment.

Modality	Examples	Typical origin in SAGIN	Indicative rate	Cost to move/store	Privacy sensitivity	Representative tasks and failure modes
RF/CSI	CSI, CQI, beam reports, RSRP, BLER	RAN, UAV/ HAPS payload, satellite beam telemetry	ms-100 ms	medium-high	low-medium	Beam tracking, scheduling, link adaptation; failure: staleness under RTT, noise, sparse feedback
Topology graphs	Connectivity/ interference graphs, gateway visibility graph, routing graph	Gateways, core/ edge controllers, distributed measurements	10 ms-10 s	low-medium	low	Routing, traffic steering, interference coordination; failure: incomplete edges, delayed topology updates
Trajectory/ ephemeris	Position/ velocity/ attitude, planned routes, visibility windows	Satellite control, UAV/ HAPS navigation (GNSS/IMU)	10 ms-10 s	low	low	Doppler prediction, handover timing, connectivity forecasting; failure: drift, inconsistent coordinate frames
Telemetry/ KPIs	Queue length, utilization, delay, packet loss, handover failures	RAN, gateways, core/cloud monitoring	0.1 s-min	low-medium	medium	Congestion prediction, SLA assurance, anomaly detection; failure: aggregation delay, reporting gaps
Traffic traces	Flow stats, session arrivals, spatiotemporal demand	Core/ cloud analytics, MEC	0.1 s-min	medium	medium-high	Demand forecasting, caching, placement; failure: concept drift, seasonal patterns, bias
Weather/ atmosphere	Rain fade, humidity, cloud cover, attenuation estimates	External services + local sensors + models	10 s-hour	low	low	Link budget planning, routing, band selection; failure: forecast uncertainty, spatial mismatch
Sensing/ vision	ISAC maps, radar point returns, camera/LiDAR features	UAV/HAPS payload, terrestrial sensors	10 ms-1 s	high	high	Blockage inference, situational awareness, digital-twin sync; failure: missing coverage, privacy constraints
Policy/ text logs	Config templates, incident tickets, operator intents, change logs	Core/ cloud, OSS/ BSS, controllers	event-driven	low	medium-high	Retrieval context, safe orchestration; failure: stale knowledge, inconsistent formatting

and intermittency rather than from the raw values alone. Finally, a fusion module aggregates the encoded streams into a unified representation $\mathbf{z}_t$ as in (2), and a task head or policy consumes $\mathbf{z}_t$ for prediction or control. This section focuses on practical tokenization designs for the main SAGIN modalities and the engineering tradeoffs that influence sequence lengths, computational cost, and robustness.

A. RF/CSI TOKENIZATION

RF/CSI observations can appear as complex-valued channel matrices, beam measurements, reference-signal metrics, or link adaptation indicators [87]. A common starting point is to convert complex quantities to real-valued representations by stacking real and imaginary parts, or by using magnitude and phase [88]. Tokenization can then proceed at multiple granularities. Fine-grained tokenization treats subcarriers, antenna elements, or delay-Doppler bins as tokens, which is suitable for wideband modeling but produces long sequences. Coarser tokenization partitions the channel into patches (e.g., groups of subcarriers or antennas), producing fewer tokens and enabling scalable attention. In SAGINs, the channel structure often contains deterministic components linked to geometry and Doppler; therefore, adding explicit features derived from trajectory/ephemeris, such as predicted Doppler $f_{\mathrm{D}}(t)$ from (7), phase noise statistics, and estimated delay spread, can reduce the burden on the model to rediscover physics from data [89], [90], [91].

A key design choice is how to incorporate time. For fast-varying RF signals, one can tokenize short temporal windows, e.g., $\{\mathbf{x}_{t-L+1}^{(\mathrm{RF})}, \ldots, \mathbf{x}_t^{(\mathrm{RF})}\}$, and use temporal positional encodings. Alternatively, one can compute compact sufficient statistics (e.g., channel principal components, beam indices, and tracking residuals) and treat them as low-rate tokens for higher-level controllers. This multi-resolution view becomes important when a single FM is expected to support both fast PHY/RAN decisions and slower SAGIN orchestration [92], [93].

B. GRAPH TOKENIZATION FOR TOPOLOGY AND INTERACTIONS

Topology and interaction structure in SAGINs are naturally represented as graphs, but transformer-like backbones require

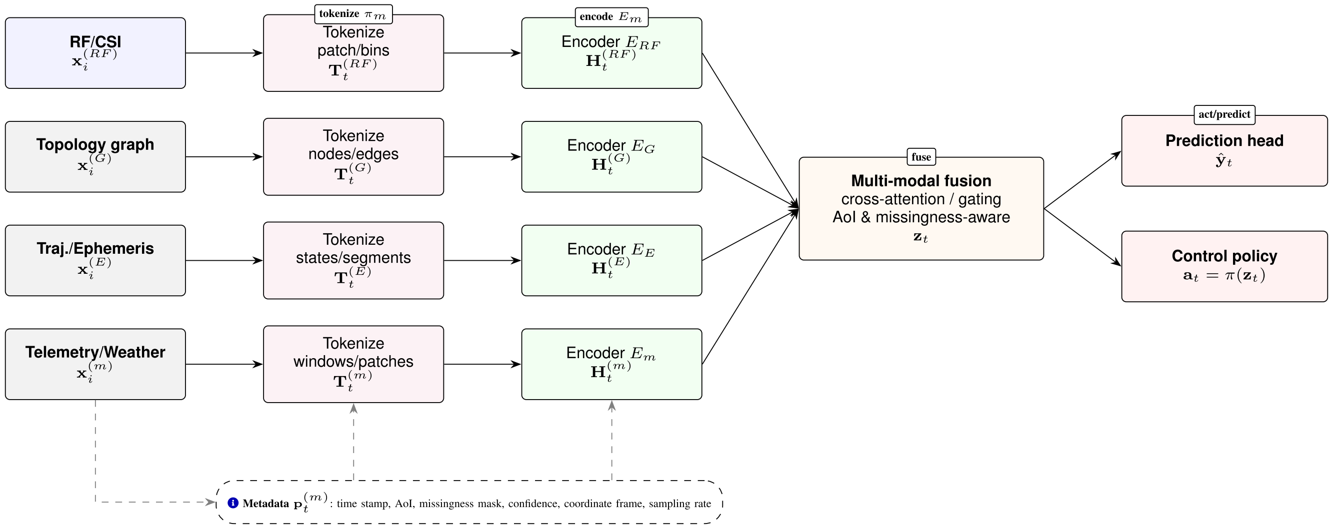

FIGURE 6. Unified tokenization and representation pipeline for SAGIN modalities. Each modality is tokenized and encoded, then fused using AoI- and missingness-aware mechanisms to yield a shared representation that supports both prediction and closed-loop control.

tokens [94], [95]. There are two dominant pathways. The first is node-centric tokenization, where each node (satellite, beam, gateway, UAV, base station, or flow) becomes a token with node features (load, visibility, link quality summaries, queue states) and relational structure injected through adjacency-aware attention or message passing [96], [97]. The second is edge-centric tokenization, where feasible links or interference couplings become tokens with edge features (capacity, delay, visibility window, interference weight). Node-centric tokenization is typically more scalable, whereas edge-centric tokenization can be more expressive for routing and scheduling problems.

SAGIN graphs are dynamic: visibility windows, beam footprints, and relay availability alter the adjacency over time [98]. Thus, tokenization should include time-stamped adjacency and a stability indicator, e.g., predicted edge duration. Furthermore, graphs are multi-layer: one may have a graph at each layer and cross-layer edges (e.g., satellite-to-gateway, HAPS-to-RAN) [99]. A robust approach is to treat the entire system as a heterogeneous graph with typed nodes and typed edges, using type embeddings to preserve semantics.

C. TRAJECTORY AND EPHEMERIS TOKENIZATION

Trajectory and ephemeris are low-dimensional but highly predictive modalities [100]. They can be represented as continuous time series of state vectors (position, velocity, attitude) in a chosen coordinate frame. Tokenization can be performed per time step (each state becomes a token) or per segment (a short horizon of states is pooled into one token). For LEO satellites, the orbit state can be represented via ephemeris parameters, predicted visibility windows, and derived quantities such as Doppler and slant range [101], [102]. For UAV/HAPS, trajectory tokens may include mission intent, waypoints, and wind-compensated dynamics [103], [104], [105]. Because these modalities are often smoother than RF, their sampling can be lower, but the model must correctly align them to RF observations, making timestamp and AoI fields indispensable. Fig. 6 presents the unified tokenization, encoding, and fusion pipeline for heterogeneous SAGIN modalities, yielding a shared latent representation that supports both prediction and closed-loop control.

D. TELEMETRY AND TRAFFIC TOKENIZATION

Telemetry streams and traffic traces are often multi-variate time series with irregular reporting [106]. Tokenization typically uses windowed sequences of KPI vectors, augmented with reporting interval and aggregation level. Because telemetry is frequently delayed, its AoI should be explicitly included as a feature and used in attention masks or gating so that stale tokens do not dominate decisions. Another useful abstraction is event tokenization, where anomalies, threshold crossings, and controller actions become discrete events. Event tokens are compact and can be aligned naturally with orchestration decisions, making them suitable for long-horizon control and retrieval augmentation.

Text-like artifacts such as incident tickets and configuration changes, can also be tokenized using language encoders and used through retrieval. In such cases, the key engineering challenge is governance: the retrieved context must be policy-filtered and privacy-aware, and the model should not treat textual context as ground truth without verification, especially for safety-critical actions.

E. WEATHER, ATMOSPHERE, AND MAP-BASED CONTEXT

Weather and atmospheric attenuation provide slow but influential context for link budgets and routing. These modalities can be tokenized as gridded fields (patch tokens over a spatial grid), as station-wise time series, or as compact derived parameters (e.g., predicted rain fade for a path). Map-based context (terrain, building density, land usage) and external models can be tokenized similarly via patches or region

TABLE 7. Tokenization and representation options for major SAGIN modalities. Sequence length refers to the number of tokens per decision step after tokenization, which directly affects attention cost.

Modality	Raw observation examples	Tokenization $\tau_m(\cdot)$	Typical encoder $E_m(\cdot)$	Seq. length	Notes for SAGIN deployment
RF/CSI	Channel matrices, beam/ RSRP, CQI, BLER	Subcarrier/antenna tokens; delay-Doppler bin tokens; patching over time-freq	1D/2D conv + transformer; lightweight transformer; state-space models	32–2048	Fine-grained tokens improve accuracy but increase cost; add AoI and geometry-derived features for robustness
Topology/ Graphs	Connectivity, interference, routing/ gateway visibility graphs	Node tokens with features; edge tokens; heterogeneous typed tokens	Graph transformer; GNN + transformer; adjacency-aware attention	50–2000	Dynamic adjacency requires time-stamped edges and stability cues; cross-layer edges encode SAGIN coupling
Trajectory/ Ephemeris	Position/ velocity/attitude; visibility windows; waypoints	Per-step state tokens; segment tokens; derived-parameter tokens	Temporal transformer; MLP + attention; state-space encoder	8-256	Low-dimensional but highly predictive; coordinate frames and AoI must be explicit
Telemetry/ KPIs	Utilization, queues, delay/ loss, handover counters	Windowed KPI vector tokens; event tokens	Temporal transformer; MLP + attention; gated recurrent encoder	8-256	Aggregation delay is critical; AoI gating prevents stale telemetry from dominating decisions
Traffic traces	Flow arrivals, demand heatmaps, session features	Time-series tokens; spatial patch tokens for heatmaps	Temporal transformer; conv encoder; spatiotemporal attention	16-512	Strong concept drift; include seasonality/context and retraining triggers
Weather/ Atmosphere	Rain rate, cloud cover, attenuation fields	Spatial patches; station tokens; path-attenuation tokens	Conv/ ViT-style encoder; temporal attention	16–256	Slow but influential; spatial support region and forecast uncertainty should be encoded
Sensing/ ISAC	Images, radar returns, point clouds, ISAC features	Patch tokens; point/ voxel tokens; compressed feature tokens	ViT/ CNN; point transformer; pretrained sensing encoder	64-1024	High bandwidth/ compute; often sparse; used for calibration, blockage inference, and twin synchronization
Text/ Policies	Tickets, configs, intents, regulations	Text tokens from retrieval; document chunk embeddings	LLM/ text encoder + retrieval memory	32-512	Must be policy-filtered and verified; useful for orchestration and explainability

embeddings. In SAGINs, spatial alignment is crucial: tokens must specify their spatial support region and coordinate frame so that the FM can associate them with beams and platform positions [107].

F. SENSING AND ISAC TOKENIZATION

Sensing modalities such as radar point returns, LiDAR point clouds, or camera images are high-dimensional and can dominate compute and bandwidth [108]. Tokenization typically relies on pretrained encoders that convert raw sensing into compact feature tokens. For example, an image can be represented as patch tokens, while a point cloud can be represented by learned point tokens or voxel tokens. ISAC signals can be treated as RF-like modalities but with task-specific structure related to detection and estimation [108]. In many SAGIN deployments, sensing is sparse and costly; therefore, a pragmatic design is to use sensing tokens for calibration and periodic synchronization rather than continuous control, and to learn conditional policies that gracefully degrade when sensing is unavailable [109].

G. MISSING MODALITIES, AOI, AND UNCERTAINTY AS FIRST-CLASS INPUTS

A defining operational feature of SAGINs is that modality availability varies with visibility, backhaul capacity, and sensor duty cycles. The tokenization layer should therefore expose missingness and AoI explicitly [110]. A simple method is to add a modality-availability mask and AoI scalar to each token embedding [111]. A stronger method is to use modality-specific gating that scales token contributions based on confidence and staleness, enabling the FM to implement a learned form of sensor fusion under uncertainty. These mechanisms reduce brittle failure modes where a stale telemetry token overrides fresh RF evidence, or where a missing graph update leads to inconsistent routing decisions. Table 7 summarizes representative tokenization/encoder choices for major SAGIN modalities and highlights their typical sequence lengths and deployment considerations. Figure 6 provides a unified pipeline showing how modality-specific tokenization interfaces with a shared FM backbone.

In practical SAGIN deployments, modality availability can vary due to intermittent connectivity, reporting delays, and sensing limitations across distributed platforms. Therefore, missing modalities, age-of-information (AoI), and measurement uncertainty should be treated as explicit inputs to the learning pipeline. A common approach is to use modality-availability masks and AoI-aware embeddings so that the model can account for data freshness when performing fusion. In addition, confidence-weighted attention or gating mechanisms can adjust the contribution of each modality based on reliability indicators such as reporting delay or sensor quality. For example, if telemetry reports are stale or a topology update is temporarily unavailable, the fusion module can reduce their influence and rely more strongly on fresh RF measurements or trajectory information. Such mechanisms help maintain robust multi-modal representations and prevent delayed or missing data from dominating the decision process.

VI. MULTI-MODAL FOUNDATION MODEL ARCHITECTURES

This section surveys and tutorializes architecture families for multi-modal FMs that are suitable for SAGINs. The central challenge is to simultaneously handle: i) heterogeneous modality structure (RF tensors, graphs, trajectories, telemetry, sensing); ii) heterogeneous rates and missingness; and iii) deployment constraints (on-board compute and energy, gateway bottlenecks, and intermittent links). We organize architectures by where fusion occurs, how modality interactions are modeled, and how scalability and robustness are maintained.

A. FUSION PLACEMENT: EARLY, LATE, AND HYBRID FUSION

Fusion placement determines both computational cost and robustness. Consider modality token sequences $\{\mathbf{T}_t^{(m)}\}_{m\in\mathcal{M}_t}$ and encoded sequences $\{\mathbf{H}_t^{(m)}\}_{m\in\mathcal{M}_t}$ from Section V. Early fusion concatenates tokens (or embeddings) from all available modalities into a single sequence and feeds them to a shared backbone as

$$\mathbf{S}_t = \mathrm{Concat}\left(\mathbf{T}_t^{(m)} : m \in \mathcal{M}_t\right), \qquad \mathbf{z}_t = f_\theta(\mathbf{S}_t). \tag{10}$$

Early fusion can capture rich cross-modal interactions, but it can be expensive because the complexity of the attention grows with the total length of the sequence [112]. It is also sensitive to missing modalities unless missingness masks and modality-type embeddings are carefully designed.

Late fusion performs modality-specific inference and combines decisions at the end as

$$\widehat{\mathbf{y}}_t^{(m)} = g_{\phi_m}\left(E_m(\mathbf{T}_t^{(m)})\right), \qquad \widehat{\mathbf{y}}_t = \mathcal{A}\left(\widehat{\mathbf{y}}_t^{(m)} : m \in \mathcal{M}_t\right), \tag{11}$$

where $\mathcal{A}(\cdot)$ is an aggregation rule (e.g., weighted averaging, stacking, or a meta-learner). Late fusion is robust to missing modalities and easier to deploy across distributed nodes, but it may miss fine-grained cross-modal dependencies that are crucial for proactive control (e.g., how ephemeris explains RF changes) [113].

Hybrid fusion combines the benefits by using modality-specific encoders followed by interaction modules such as cross-attention as:

$$\mathbf{H}_t^{(m)} = E_m(\mathbf{T}_t^{(m)}), \qquad \mathbf{z}_t = F_\theta\left(\mathbf{H}_t^{(m)} : m \in \mathcal{M}_t\right), \tag{12}$$

where $F_\theta(\cdot)$ may implement cross-attention, gated fusion, or mixture-of-experts. In SAGINs, hybrid fusion is often the most practical because it allows modality-specific compression (shorter sequences) while still modeling cross-modal dependencies when needed [114].

B. INTERACTION MODELING: CROSS-ATTENTION, GATING, AND MIXTURE-OF-EXPERTS

Once modality-specific representations are available, the core architectural question becomes how to model inter-modal interactions under missingness and AoI constraints [115], [116], [117]. Cross-attention lets one modality query another, e.g., RF tokens attend to trajectory tokens to exploit predictable Doppler structure. If $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote query, key, and value matrices, a generic cross-attention block is

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{13}$$

In practice, AoI and missingness can be injected via attention masks or pair(key/value) scaling so that stale or unreliable tokens contribute less.

Gated fusion assigns a learned weight $\alpha_t^{(m)} \in [0, 1]$ to each modality (or token group), enabling the model to downweight missing or stale modalities as:

$$\mathbf{z}_t = \sum_{m\in\mathcal{M}_t} \alpha_t^{(m)} \bar{\mathbf{h}}_t^{(m)}, \qquad \sum_{m\in\mathcal{M}_t} \alpha_t^{(m)} = 1, \tag{14}$$

where $\bar{\mathbf{h}}_t^{(m)}$ is a pooled representation of the modality m. Gating is computationally efficient and particularly suitable for on-board inference.

MoE architectures activate only a subset of experts depending on input, improving scalability [118], [119]. In SAGINs, MoE can be used in two ways: i) modality-specialized experts (RF expert, graph expert, sensing expert), or ii) regime-specialized experts (urban blockage, rain fade, sparse gateway visibility) [120]. MoE is attractive for large pretraining, but deployment must address routing overhead and expert availability under split inference [121], [122].

C. BACKBONE CHOICES FOR SAGIN MULTI-MODALITY

A backbone $f_\theta(\cdot)$ defines how the model processes sequences or sets of tokens. The most common family is transformer like attention, but several alternatives are relevant in SAGINs: Transformer backbones provide strong multi-modal fusion via attention and are natural for tokenized modalities [123]. Their cost scales quadratically with sequence length, motivating patching and compression for RF and sensing [124].

State-space and temporal models (including lightweight recurrent or state-space encoders) can be attractive for long telemetry and traffic sequences because they scale better with sequence length and support streaming inference [125].

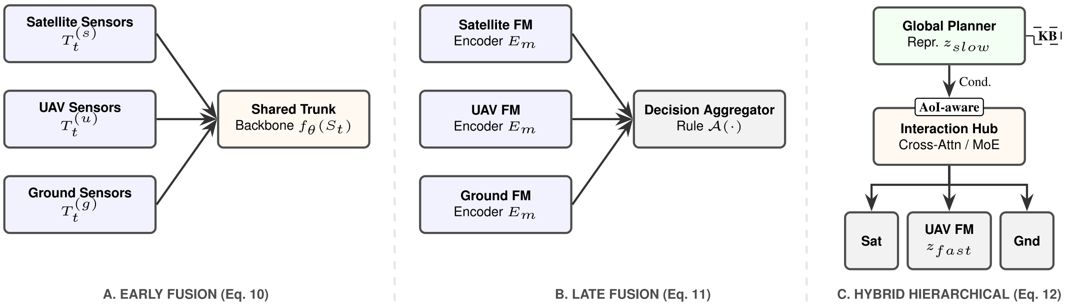

FIGURE 7. Fusion and architecture choices for multi-modal FMs in SAGINs. Early fusion offers strong cross-modal expressivity but high cost; late fusion is robust and distributed but weaker in cross-modal reasoning; hybrid fusion with hierarchical planners/executors balances expressivity, intermittency robustness, and multi-timescale control.

TABLE 8. Architecture families for multi-modal foundation models in SAGINs: suitability, robustness, and deployment considerations.

Architecture family	Core idea	Strengths in SAGINs	Limitations / risks	Typical placement
Early fusion transformer	Concatenate all modality tokens and learn joint attention end-to-end	Maximal cross-modal expressivity; strong for joint perception tasks	High compute with long RF/sensing sequences; sensitive to missing modalities without careful masking	Cloud/edge; sometimes gateway
Late fusion ensemble	Separate unimodal models; combine decisions with an aggregator	Robust to missing modalities; easy to distribute across nodes	Weak cross-modal reasoning; harder to learn joint causal structure	Distributed (on-board + gateway + edge)
Hybrid fusion with cross-attention	Modality-specific encoders + cross-attention fusion	Good tradeoff between expressivity and cost; supports selective conditioning on slow modalities	Requires careful alignment and AoI handling; design complexity higher	Edge/gateway; split possible
Gated fusion / modality weighting	Learn modality gates based on confidence/AoI	Efficient; robust under intermittency; suitable for constrained devices	May underfit complex interactions if used alone; relies on good encoder summaries	On-board, UAV/HAPS, edge
Mixture-of-experts (MoE)	Route inputs to a subset of experts (modality- or regime-specialized)	Scales to very large pretraining; supports specialization across regimes	Routing overhead; expert availability under split; stability issues if data are skewed	Cloud/edge; limited on-board
Hierarchical multi-timescale FM	Slow planner + fast executor representations	Aligns with SAGIN control hierarchy; improves stability under delays	Requires careful interface design and training of multi-rate modules	Edge + on-board; cloud planner
Retrieval/tool-augmented FM	Condition on retrieved KB context and call tools/solvers	Improves factuality and verifiability; supports policy and constraint checks	Governance and security for KB; tool latency and reliability	Core/cloud; edge with cached KB

Graph-native backbones (graph transformers or graph neural network (GNN) + transformer hybrids) are well suited when topology is the primary driver of decisions, such as routing and gateway selection [126].

In a practical SAGIN FM, the backbone is often modular: modality encoders compress raw inputs, interaction blocks provide selective cross-modal conditioning, and a shared trunk outputs a representation for multiple heads (multi-task learning). This modularity also supports split inference and efficient parameter adaptation.

D. HIERARCHICAL AND CROSS-LAYER ARCHITECTURES

Because SAGIN decisions span multiple timescales (Fig. 4), a single monolithic policy can be brittle. Hierarchical architectures address this by maintaining representations at different rates. One useful pattern is a two-level hierarchy: a slow planner representation $\mathbf{z}_t^{\text{slow}}$ based on slow modalities (ephemeris, weather, long-window telemetry) and a fast executor representation $\mathbf{z}_t^{\text{fast}}$ driven by RF/CSI and short-window feedback as

$$\mathbf{z}_t^{\text{slow}} = f_{\theta_s}\left(\mathbf{T}_t^{(\text{E})}, \mathbf{T}_t^{(\text{W})}, \mathbf{T}_t^{(\text{KPI})}\right), \tag{15}$$

$$\mathbf{z}_t^{\text{fast}} = f_{\theta_f}\left(\mathbf{T}_t^{(\text{RF})}, \mathbf{z}_t^{\text{slow}}\right), \tag{16}$$

where the slow representation conditions the fast controller through cross-attention or gating. This design matches operational reality: long-horizon constraints and budgets should guide fast actions, especially under delayed feedback.

E. RETRIEVAL AUGMENTED AND TOOL AUGMENTED ARCHITECTURES

SAGIN management often requires information that is not fully present in sensor streams, such as policies, spectrum regulations, site metadata, and past incident resolutions [127], [128], [129]. Retrieval augmented architectures incorporate a retrieval module that fetches relevant context from a KB and conditions the FM on the retrieved items [130]. Tool augmented architectures go one step further by allowing the model to call external solvers or controllers (e.g., shortest path routing, convex optimization, constraint checking) [131], [132]. These architectures are particularly relevant for safety and trust because they allow explicit verification and enforcement of constraints rather than relying solely on the implicit reasoning of the model. Table 8 summarizes the main architecture families, their strengths and weaknesses in SAGIN settings, and typical deployment placements. Figure 7 visually contrasts early, late, and hybrid fusion and shows an example hierarchical cross-layer design. Fig. 7 outlines the data curation, self-supervised pretraining objectives, and fusion choices that lead to a multi-modal foundation model ready for adaptation and closed-loop control in SAGINs.

VII. DATA CURATION AND PRETRAINING OBJECTIVES FOR SAGIN FOUNDATION MODELS

A multi-modal FM is only as strong as its pretraining signal and the coverage of its pretraining data. In SAGINs, pretraining must address three coupled realities. First, modalities are heterogeneous and often multi-rate, so pretraining should learn representations that remain stable under time misalignment and missingness [133]. Second, the environment is non-stationary due to mobility, visibility windows, and weather, so pretraining should encourage invariances and robust uncertainty estimates rather than memorizing narrow regimes [134]. Third, labeled data for many downstream tasks (e.g., rare outages or safety-critical control events) are scarce, so self-supervised objectives and synthetic enhancements become central [135].

This section describes objective families that are particularly effective for SAGIN multi-modality and provides a practical view of how to curate and align data at scale. We denote by $\mathbf{x}_t^{(m)}$ the observation of modality m at time t, by $\mathbf{T}_t^{(m)}$ its tokens from (8), and by $\mathbf{H}_t^{(m)}$ the encoded tokens from (9). The pretraining goal is to learn parameters θ such that the fused representation $\mathbf{z}_t$ supports transfer across tasks and domains.

A. MASKED MODELING AND RECONSTRUCTION OBJECTIVES

Masked modeling trains a model to reconstruct missing parts of an input from the remaining context [136]. In multi-modal SAGIN settings, masking can be applied within a modality (e.g., masking RF patches) or across modalities (e.g., hiding weather tokens and asking the model to infer attenuation-conditioned link quality statistics) [137]. Let $\mathbf{T}$ denote a token sequence (possibly a concatenation of several modalities with type embeddings), and let $\mathcal{I}$ be the set of masked token indices. The model receives $\mathbf{T}_{\setminus\mathcal{I}}$ and predicts $\widehat{\mathbf{T}}_{\mathcal{I}}$. A standard reconstruction loss is

$$\mathcal{L}_{\text{mask}} = \sum_{i\in\mathcal{I}} \ell\left(\widehat{\mathbf{u}}_i, \mathbf{u}_i\right), \tag{17}$$

where $\mathbf{u}_i$ and $\widehat{\mathbf{u}}_i$ are the true and reconstructed tokens and $\ell(\cdot,\cdot)$ is typically mean-squared error for continuous tokens or cross-entropy for discretized tokens. In SAGINs, masked modeling is valuable because it naturally matches operational missingness; training the model to reconstruct missing tokens teaches it to be robust when a modality drops due to backhaul or sensing duty-cycle constraints [138], [139].

When reconstructing RF/CSI-like structures, patching is typically required to keep sequences short [140]. For graphs, masking can be applied to node features, edge features, or even subgraphs [141]. For trajectories/ephemeris, masking teaches the model to interpolate or extrapolate platform states, which improves downstream connectivity prediction and Doppler-aware control [142], [143].

B. CONTRASTIVE LEARNING AND CROSS-MODAL ALIGNMENT

Contrastive learning encourages the representations of semantically related samples to be close while pushing unrelated samples apart [144]. For SAGIN multi-modality, the most useful form is cross-modal contrastive alignment, which exploits the co-occurrence of modalities in time and space [145]. Let $\mathbf{z}_i^{(a)}$ and $\mathbf{z}_i^{(b)}$ be pooled representations from two modalities (or two views) a and b for the same underlying time-space instance i. A widely used objective is the InfoNCE loss

$$\mathcal{L}_{\text{con}} = -\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(\mathbf{z}_i^{(a)}, \mathbf{z}_i^{(b)})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\mathbf{z}_i^{(a)}, \mathbf{z}_j^{(b)})/\tau\right)}, \tag{18}$$

where B is the batch size, $\text{sim}(\cdot,\cdot)$ is a similarity function (often cosine similarity), and $\tau > 0$ is a temperature parameter controlling concentration. The positive pair in the numerator corresponds to the same time-space instance (e.g., RF features aligned with ephemeris and topology), while the negatives in the denominator correspond to mismatched instances. In SAGINs, careful definition of positives is crucial because time alignment can be imperfect. A robust practice is to treat a small time window around t as positive and penalize only larger time shifts, which prevents the model from overfitting noisy timestamps.

Cross-modal contrastive learning is particularly powerful in cases where one modality is sparse but of high value, such as sensing [146], [147]. For example, occasional sensing snapshots can be aligned with RF/CSI statistics and topology context, enabling the model to carry sensing-derived knowledge into periods where sensing is absent.

C. PREDICTIVE PRETRAINING AND SEQUENCE MODELING

SAGIN decisions are inherently temporal [148]. Predictive pretraining teaches the model to forecast future representations or observations. A common choice is autoregressive next-token modeling [149], where the model predicts the next token conditioned on previous tokens as

$$\mathcal{L}_{\mathrm{AR}} = -\sum_{t} \log p_\theta\big(\mathbf{u}_{t+1} \,|\, \mathbf{u}_{\le t}\big)\,. \tag{19}$$

For continuous-valued modalities (telemetry, trajectories), forecasting can be expressed as regression over a horizon H as

$$\mathcal{L}_{\mathrm{pred}} = \sum_{t} \sum_{h=1}^{H} \big\|\widehat{\mathbf{y}}_{t+h} - \mathbf{y}_{t+h}\big\|_2^2\,, \tag{20}$$

where $\mathbf{y}_{t+h}$ may represent future KPI vectors, future link-quality summaries, or future graph states. In SAGINs, predictive pretraining is most effective when it explicitly conditions on slow modalities such as ephemeris and weather, because they provide stable predictive structure across domains.

A practical refinement is to predict in a latent space rather than in the raw space. Let $\mathbf{z}_t$ be the fused representation at time t and let $\widehat{\mathbf{z}}_{t+h}$ be the predicted future representation. A latent prediction loss

$$\mathcal{L}_{\mathrm{lat}} = \sum_{t} \sum_{h=1}^{H} \big\|\widehat{\mathbf{z}}_{t+h} - \mathrm{sg}(\mathbf{z}_{t+h})\big\|_2^2 \tag{21}$$

uses $\mathrm{sg}(\cdot)$ to stop gradients through the target representation so that the model learns a stable predictive state without collapsing to trivial solutions.

D. MULTI-TASK SELF-SUPERVISION AND OBJECTIVE MIXING

No single objective is sufficient for SAGINs. Masked modeling improves robustness to missingness, contrastive alignment improves cross-modal consistency, and predictive objectives improve temporal reasoning [150]. In practice, pretraining often combines multiple losses as

$$\mathcal{L}_{\mathrm{pre}} = \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{pred}}\mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{lat}}\mathcal{L}_{\mathrm{lat}}, \tag{22}$$

where $\lambda_{\cdot}$ are weights. In SAGINs, these weights should be chosen with deployment goals in mind. If the downstream focus is on safe control under intermittency, stronger masked modeling and AoI-aware training are beneficial. If the focus is on multi-domain transfer and using sparse sensing, contrastive alignment becomes more important. If the focus is on proactive planning (beam hopping, gateway selection), predictive losses over longer horizons should be emphasized.

E. DATA CURATION: ALIGNMENT, MISSINGNESS, AND AGE-OF-INFORMATION

Pretraining requires not only large volumes but also careful curation so that the model learns meaningful cross-modal structure rather than artifacts [151], [152]. A practical SAGIN curation pipeline begins by harmonizing time, space, and identifiers. Every sample should carry time stamps and spatial References, and ideally also include AoI tags reflecting when the observation was produced versus when it became available to the learning pipeline. Missingness should not be silently imputed; instead, explicit missingness masks should be stored and passed to the model so that it can learn uncertainty-aware fusion [153].

Another curation step is regime coverage. Pretraining should cover diverse orbital geometries, traffic regimes, weather conditions, and platform configurations. Without such coverage, large models can overfit to frequent regimes and become overconfident under rare but operationally important conditions such as handover storms, gateway outages, or severe attenuation events.

F. SYNTHETIC DATA AND DIGITAL TWINS FOR SAGIN COVERAGE

SAGIN datasets often exhibit gaps due to limited sensing, restricted access to on-board data, and rare-event scarcity. Synthetic generation fills these gaps when it is anchored in credible models [154], [155]. Ray-tracing and geometry-based channel models can produce RF/CSI-like samples under controlled motion and blockage conditions [156], [157]. Network digital twins can generate telemetry, routing states, and orchestration logs under different policies [158], [159], [160]. Synthetic data should not be treated as ground truth; instead, it should be used to enlarge coverage, pretrain invariances, and teach the model to respect physical constraints, followed by adaptation on real measurements to correct model mismatch. Table 9 summarizes objective families and how they map to SAGIN modalities and constraints. Fig. 8 illustrates a practical pretraining and curation pipeline that mixes real and synthetic data while preserving AoI and governance constraints.

In practical SAGIN training pipelines, robustness to missing modalities, stale observations, and noisy measurements can be incorporated through targeted augmentation and masking strategies during pretraining. For instance, modality dropout or stochastic masking of modality tokens can be applied so that the model learns stable representations even when certain inputs are absent. Noise injection and perturbation-based augmentation can emulate sensing uncertainty in RF, telemetry, or environmental modalities, improving robustness to measurement errors. In addition, AoI indicators and confidence scores can be embedded in token metadata, allowing the fusion module to downweight stale or unreliable observations. To further approximate real SAGIN conditions, training datasets may simulate intermittent connectivity through synthetic link outages, delayed telemetry streams, and asynchronous modality sampling. These mechanisms expose the model to deployment-like conditions during training and improve generalization under heterogeneous data availability and dynamic network environments.

VIII. ADAPTATION, CONTINUAL LEARNING, AND PERSONALIZATION

Pretraining yields a multi-modal FM with transferable representations, but SAGIN deployments require specialization to local conditions, platform constraints, and evolving environments [161]. This section focuses on how to adapt a

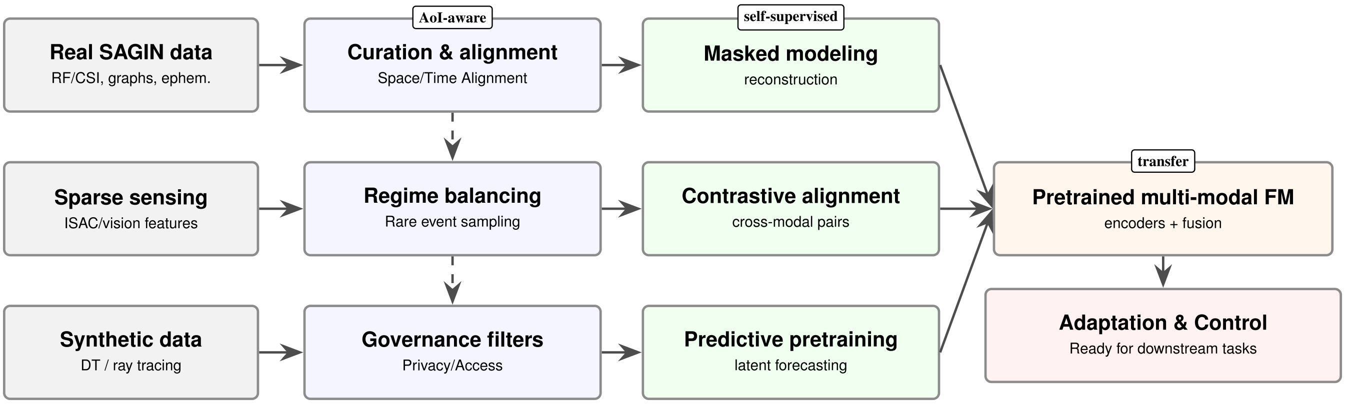

FIGURE 8. Pretraining and data curation pipeline for SAGIN multi-modal FMs. Real and synthetic sources are aligned and tagged with age-of-information and missingness, balanced across regimes, filtered by governance constraints, and used with complementary self-supervised objectives to obtain transferable representations.

TABLE 9. Pretraining objective families for SAGIN multi-modal foundation models and their practical roles.

Objective family	Core signal	Modalities it naturally supports	What it learns well in SAGINs	Operational caveats
Masked modeling / reconstruction	Reconstruct masked tokens from context, cf. (17)	RF/CSI patches, graphs (node/edge masks), telemetry windows, sensing features	Robustness to missingness; denoising; context completion under intermittent streams	Avoid leakage via correlated masks; sequence length control is critical for RF/sensing
Contrastive alignment (InfoNCE)	Pull paired views together and push mismatches apart, cf. (18)	Cross-modal pairs (RF–ephemeris, RF–telemetry, sensing–RF), multi-view augmentations	Cross-modal consistency; domain transfer; exploiting sparse high-value modalities (sensing)	Positives must tolerate time misalignment; negatives must avoid false negatives in similar regimes
Predictive forecasting	Predict future observations/representations, cf. (20)–(21)	Telemetry, traffic, trajectories; RF summaries; dynamic graphs	Proactive planning; multi-timescale reasoning; anticipating congestion/visibility changes	Horizon choice matters; long horizons risk error accumulation without uncertainty modeling
Autoregressive sequence modeling	Next-token modeling, cf. (19)	Event sequences, logs, discretized states, tokenized multi-modal streams	Long-context modeling; policy/action sequence patterns; generative planning seeds	May memorize spurious correlations; requires governance if text logs are used
Multi-loss mixing	Weighted sum of objectives, cf. (22)	All modalities	Balanced invariances: missingness robustness + alignment + prediction	Loss weights should match downstream goals and deployment constraints

pretrained FM to downstream tasks and domains while controlling compute, memory, communication overhead, and safety risks. We emphasize three practical pillars: PEFT for cost-effective specialization, continual learning for non-stationarity, and federated/collaborative learning for governance and privacy.

A. PROBLEM SETTING AND ADAPTATION OBJECTIVES

Let f_θ denote a pretrained FM with parameters θ, and let g_ϕ denote a task head or policy with parameters ϕ. Adaptation seeks parameters (θ', ϕ') that improve performance on a target task distribution $\mathcal{T}$ and target domain $\mathcal{P}_{\text{tar}}$ (e.g., a particular region, constellation segment, band, or operator policy), while considering constraints such as on-board compute and data locality [162]. A generic fine-tuning objective is described as

$$\min_{\theta',\phi'} \mathbb{E}_{(\mathbf{o},\mathbf{y})\sim\mathcal{D}_{\text{ft}}}\big[\ell\big(g_{\phi'}(f_{\theta'}(\mathbf{o})),\mathbf{y}\big)\big] + \Omega(\theta',\theta), \tag{23}$$

where $\mathbf{o}$ denotes the (possibly multi-modal) observation, $\mathbf{y}$ denotes the supervision signal (labels, targets, or rewards), and $\Omega(\theta', \theta)$ is a regularizer that discourages catastrophic forgetting by keeping θ' close to the pretrained parameters θ or by constraining changes to a small parameter subspace.

B. PARAMETER EFFICIENT FINE TUNING

Full model fine-tuning is often infeasible in SAGINs due to limited memory, energy, and bandwidth for model updates. PEFT addresses this by updating only a small fraction of parameters, while keeping the backbone largely frozen [163]. The key benefit is that model updates become compact artifacts that can be stored and transmitted efficiently, enabling frequent adaptation without retraining from scratch [163].

1) ADAPTERS

Adapters insert small bottleneck modules into a frozen backbone [164]. Let $\mathbf{h}$ be an intermediate hidden state in the backbone. An adapter applies as

$$\text{Adapter}(\mathbf{h}) = \mathbf{h} + \mathbf{W}_{\text{up}}\sigma(\mathbf{W}_{\text{down}}\mathbf{h}), \tag{24}$$

where $\mathbf{W}_{\text{down}}$ reduces dimension, $\mathbf{W}_{\text{up}}$ restores it, and $\sigma(\cdot)$ is a non-linearity; only adapter parameters are trained. In SAGINs, adapters are attractive because different regions or layers (air, space, and ground) can maintain their own adapters while sharing a common pretrained backbone.

2) LOW-RANK ADAPTATION

LoRA updates large weight matrices by learning a low-rank increment [165]. For a weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, LoRA uses as

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}, \qquad \Delta\mathbf{W} = \mathbf{AB}, \tag{25}$$

where $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times k}$ with rank $r \ll \min(d, k)$. Training only $\mathbf{A}$ and $\mathbf{B}$ yields compact updates. In multi-modal SAGIN FMs, LoRA can be selectively applied to cross-attention layers to adapt how modalities interact without changing modality encoders.

3) PROMPT AND PREFIX TUNING AND SOFT TOKENS

For transformer like backbones, another PEFT route is to learn a small set of continuous soft prompt tokens that condition the model without changing backbone weights [166]. This is useful when the goal is to adapt behavior to an operator policy or a deployment regime with minimal update size [167]. However, prompt-like approaches can be less stable for hard control tasks if not combined with explicit constraint checks.

C. DOMAIN ADAPTATION AND MULTI-DOMAIN SPECIALIZATION

SAGINs operate in heterogeneous domains: urban vs rural, different frequency bands, different platform types, and varying gateway visibility [168]. A practical adaptation design is to separate shared and domain-specific parameters. The shared backbone captures general multi-modal structure, while domain specific adapters or LoRA modules capture local idiosyncrasies. Domain identifiers (e.g., band, platform type, and region etc) can be injected through embeddings so that the model conditions its decisions on the correct operating regime [169], [170]. When labeled data are scarce in a new domain, unsupervised domain adaptation can be performed by continuing self-supervised pretraining using target domain unlabeled data and then lightly fine tuning on a small labeled set [171]. This approach is particularly effective when the new domain differs mainly in distributional statistics rather than in semantics (e.g., new weather regime or new traffic profile).

D. CONTINUAL LEARNING UNDER NON-STATIONARITY

SAGINs are non-stationary; moreover, platform dynamics, traffic regimes, and environmental conditions continuously evolve [172], [173], [174]. Continual learning aims to update models online (or periodically) without catastrophic forgetting [175]. A pragmatic continual learning loop includes: i) drift detection, ii) selective data buffering, iii) constrained updates, and iv) validation under safety checks [176].

Drift detection can be performed at the representation level by monitoring statistics of $\mathbf{z}_t$ or at the performance level by tracking prediction residuals, constraint violation rates, or out-of-distribution (OOD) scores. When drift is detected, the system can trigger adaptation using recent data, but updates should be constrained to avoid destabilizing control. Regularization terms in (23) and PEFT updates (adapters or LoRA) naturally support constrained updates.

A central SAGIN-specific challenge is that feedback labels can be delayed or missing, especially for long horizon control. Therefore, continual learning often relies on self-supervised objectives and proxy labels (e.g., predicting KPI trends or reconstructing missing modalities) and uses delayed reinforcement signals sparingly [177].

E. FEDERATED AND COLLABORATIVE LEARNING FOR GOVERNANCE

Data governance and privacy constraints often prevent centralizing raw data across operators or across air, space, and ground segments. Federated learning enables collaborative training where each node trains locally and shares only model updates [178], [179]. Let $\Delta\theta_n$ denote the update from node n (satellite, gateway, UAV, base station). A simple aggregation can be expressed as

$$\theta \leftarrow \sum_{n=1}^{N} w_n(\theta + \Delta\theta_n), \tag{26}$$

where w_n weights nodes by data volume or reliability. In practice, SAGIN federated learning must address intermittent participation: nodes can drop out due to visibility windows or power constraints [180]. This motivates asynchronous aggregation and robust update filtering. Moreover, because modalities differ by node, federated learning can be combined with modular architectures where each node updates only the parameters relevant to its available modalities (e.g., on-board nodes update RF and ephemeris encoders, while core nodes update retrieval and orchestration heads) [181].

F. SAFETY-AWARE ADAPTATION

Adapting models in a live network introduces risk: a poorly tuned model can cause widespread outages due to the wide coverage of non-terrestrial layers. Therefore, adaptation should include safety-aware mechanisms such as: i) constrained policy updates, ii) offline evaluation on counterfactual logs or digital twins, iii) uncertainty gating, and iv) rollback to a known-safe configuration [182]. A practical approach is to treat adaptation artifacts (e.g., adapter weights) as versioned components that must pass validation thresholds on a held-out safety test suite before deployment. Table 10 summarizes adaptation mechanisms and their suitability under SAGIN constraints. Fig. 9 illustrates a continual learning loop with drift detection, PEFT updates, and safety validation.

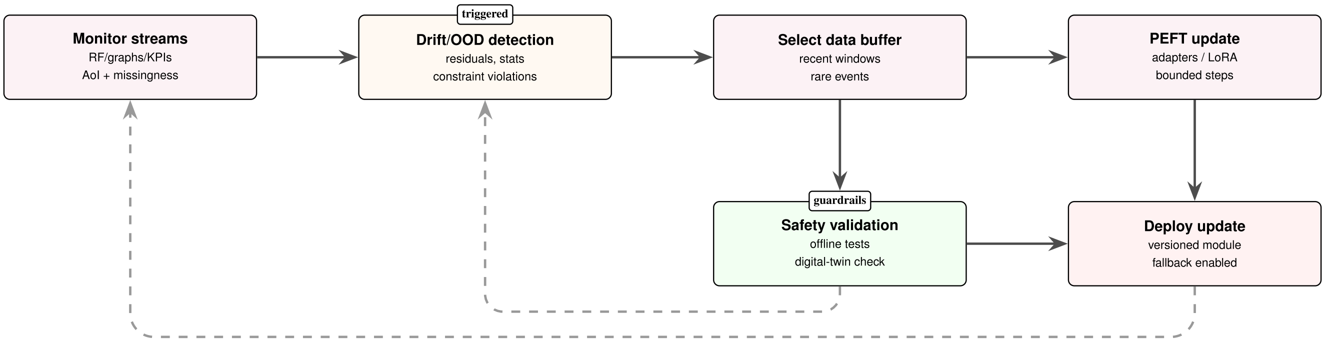

FIGURE 9. Continual learning loop for SAGIN FMs. Drift detection triggers bounded PEFT updates using recent buffered data. Updates are validated via offline tests and digital-twin checks before deployment, with versioning and rollback to ensure safety.

TABLE 10. Adaptation mechanisms for SAGIN multi-modal FMs and their operational suitability.

Method	What is updated	Strengths in SAGINs	Limitations / risks	Best use cases
Full fine-tuning	All FM parameters θ	Maximum flexibility and accuracy when ample data/compute exist	High compute and memory; hard to distribute; higher risk of instability	Cloud-side retraining for major domain shifts
Adapters	Small bottleneck modules in backbone	Compact updates; supports domain- or layer-specific specialization; easy rollback	Adds slight inference overhead; requires adapter placement design	Regional specialization, platform-specific tuning
LoRA	Low-rank increments for selected matrices	Very compact; efficient; good for adapting cross-modal interactions	Choice of layers/rank matters; can overfit if data are narrow	Adapting fusion/cross-attention under new regimes
Prompt/prefix tuning	Small set of soft tokens	Minimal update size; fast conditioning to policies/tasks	Less reliable for hard control without guardrails	Policy conditioning, orchestration preferences
Self-supervised target adaptation	Continue pretraining on unlabeled target data	Leverages abundant unlabeled streams; improves robustness	Risk of drift toward spurious correlations without checks	New region/band with limited labels
Continual learning (PEFT + drift triggers)	Periodic PEFT updates with drift detection	Tracks non-stationarity; bounded update magnitude	Requires drift detection and safe validation; label delay issues	Seasonal traffic changes, weather-driven shifts
Federated learning	Distributed training with aggregated updates	Preserves data locality; enables multi-operator collaboration	Intermittent participation; heterogeneity across nodes; poisoning risk	Cross-domain pretraining and periodic refresh

IX. CLOSED-LOOP CONTROL WITH MULTI-MODAL FMS

SAGIN intelligence is ultimately evaluated by its ability to close the loop such that it observe multi-modal context, decide under uncertainty and constraints, execute actions across layers, and adapt based on feedback. This section describes how multi-modal FMs can be integrated into closed-loop control pipelines. We emphasize three complementary roles for FMs; i) representation learners that produce compact state summaries for classical controllers and RL; ii) planners/agents that directly propose actions or action sequences while interfacing with tools and safety checks; and iii) teachers that improve RL policies via guidance, synthetic roll-outs, or reward shaping. The central theme is that SAGIN control is multi-timescale, constrained, and safety-critical, which requires hybrid model-based and model-free strategies rather than purely end-to-end learning.

A. CONTROL PROBLEM ABSTRACTION AND CONSTRAINTS

Let $\mathbf{s}_t$ denote the latent network state at time t, and let $\mathbf{o}_t = \{\mathbf{x}_t^{(m)}\}_{m\in\mathcal{M}_t}$ denote the multi-modal observation. A decision-maker selects an action $\mathbf{a}_t$ (e.g., beam selection, handover, routing, gateway selection, service placement) that influences the next state. SAGIN control is naturally partially observable and multi-timescale, so a compact representation $\mathbf{z}_t$ extracted by an FM is used as a belief-like summary as

$$\mathbf{z}_t = f_\theta(\mathbf{o}_{0:t}), \tag{27}$$

where $\mathbf{o}_{0:t}$ denotes the observation history (in practice a windowed subset with AoI metadata). A policy π_ϕ maps $\mathbf{z}_t$ to an action as

$$\mathbf{a}_t = \pi_\phi(\mathbf{z}_t). \tag{28}$$

Many SAGIN objectives combine throughput, latency, reliability, and energy. Let $\mathcal{R}(\mathbf{s}_t, \mathbf{a}_t)$ denote a reward (or negative cost) and let $\mathcal{C}_i(\mathbf{s}_t, \mathbf{a}_t) \le 0$ denote operational constraints such as power budgets, outage targets, queue stability, gateway capacity, and safety limits. A generic constrained objective is

$$\max_{\pi_\phi} \quad \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t)\right] \tag{29}$$

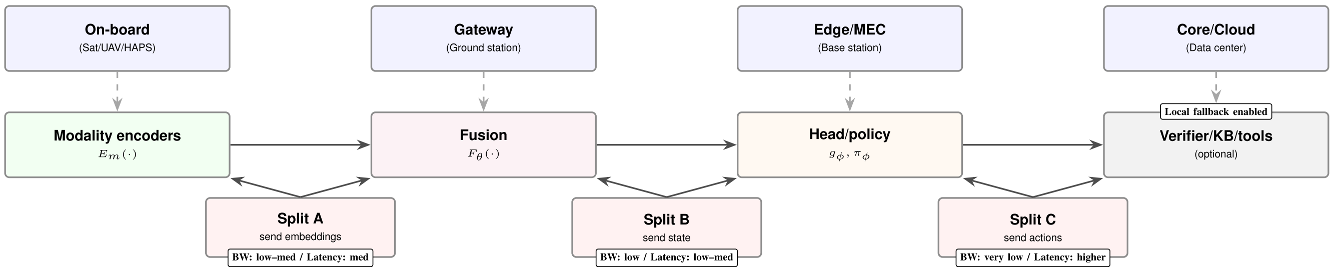

FIGURE 10. Representative split inference points for a modular multi-modal FM. Split A transmits modality embeddings, Split B transmits fused state summaries, and Split C transmits only action requests to a remote verifier/tool layer. In SAGINs, split selection should consider latency, bandwidth, intermittency, and safety requirements.

$$\text{s.t.} \quad \mathbb{E}[\mathcal{C}_i(\mathbf{s}_t, \mathbf{a}_t)] \leq 0, \quad \forall i, \tag{30}$$

where $\gamma \in (0, 1]$ is a discount factor. In practice, constraints are enforced through safe action sets, Lagrangian penalties, constrained RL, or external verification tools. Fig. 10 illustrates representative split-inference options for deploying modular multi-modal foundation models across SAGIN nodes, highlighting the trade-offs between latency, bandwidth, and safety.

B. FM AS A REPRESENTATION LEARNER FOR CONTROL

A robust control pipeline often separates state representation from decision logic. In this pattern, the FM produces $\mathbf{z}_t$ and a downstream controller (e.g., an RL policy, model predictive control (MPC), or optimization solver) uses $\mathbf{z}_t$ as its input [183]. The primary advantage is modularity: the FM can be pretrained and adapted with self-supervision (Sections VII–VIII) while the controller can be trained or designed to satisfy constraints.

For example, a value-based RL controller uses $\mathbf{z}_t$ as the state and learns a state-action value function $Q(\mathbf{z}_t, \mathbf{a}_t)$ [184]. Policy-gradient methods similarly learn $\pi_\phi(\mathbf{a}_t|\mathbf{z}_t)$. In SAGINs, this separation is valuable because the representation can fuse modalities (RF/CSI, graphs, trajectories, telemetry, weather, sensing) while the controller focuses on the action space and constraints of the specific layer. A practical refinement is to have the FM output not only $\mathbf{z}_t$ but also uncertainty estimates that reflect missingness and AoI. Uncertainty-aware representations allow controllers to act conservatively under ambiguous observations, reducing unsafe decisions in intermittent regimes.

C. FM AS A PLANNER OR AGENT

An alternative integration pattern is to use an FM to directly propose actions or action sequences. This can be realized through sequence modeling (predicting an action token sequence conditioned on context) or through agentic architectures where the FM uses a toolbox to query network state, retrieve policies, or solve subproblems [185].

Let $\mathcal{T}$ denote a set of tools such as: shortest-path routing, constraint checking, admission control, beam feasibility checks, and convex optimization solvers. An agentic controller can be expressed as an iterative loop as

$$\mathbf{a}_t = \mathsf{SelectAction}(f_\theta, \mathbf{o}_t, \mathsf{Retrieve}(\cdot), \mathsf{ToolCall}(\cdot), \mathsf{Verify}(\cdot)), \tag{31}$$

where retrieval provides relevant context (policies, configuration histories) and verification filters unsafe actions. In SAGINs, agentic control is attractive for orchestration-level decisions (routing, gateway selection, service placement) where decisions depend on policies and constraints and where tool calls can provide guarantees. For fast PHY/RAN decisions, agentic control is typically too slow, and FM-as-representation is preferred.

D. FM AS AN RL TEACHER: GUIDANCE, ROLLOUTS, AND REWARD SHAPING

FMs can also improve RL training by acting as teachers. Three teacher mechanisms are particularly relevant for SAGINs. First, an FM can provide representation distillation: RL policies train on FM embeddings rather than raw observations, improving sample efficiency and generalization [186]. Second, an FM or a learned world model can provide synthetic rollouts for policy learning. This is useful when real interactions are expensive or unsafe. In SAGINs, synthetic rollouts are often generated by combining digital twins (Section VII) with learned residual models that correct twin mismatch. Third, an FM can support reward shaping by providing auxiliary signals correlated with long-term objectives, such as predicted congestion risk, predicted outage risk, or predicted constraint violation probability [187]. Reward shaping must be used carefully to avoid biasing policies toward proxies that fail under distribution shift.

E. HIERARCHICAL CONTROL ACROSS LAYERS AND TIMESCALES

SAGIN control is inherently hierarchical. A useful design is to decompose actions into slow and fast components. Let $\mathbf{a}_t^{\text{slow}}$ represent long-horizon actions (beam hopping plans, gateway allocation, service placement), and let $\mathbf{a}_t^{\text{fast}}$ represent short-horizon actions (beam tracking, scheduling). A hierarchical policy can be written as

$$\mathbf{a}_t^{\text{slow}} = \pi_{\phi_s}(\mathbf{z}_t^{\text{slow}}), \qquad \mathbf{a}_t^{\text{fast}} = \pi_{\phi_f}(\mathbf{z}_t^{\text{fast}}, \mathbf{a}_t^{\text{slow}}), \tag{32}$$

where $\mathbf{z}_t^{\text{slow}}$ is derived mainly from slow modalities (ephemeris, weather, long-window KPIs) and $\mathbf{z}_t^{\text{fast}}$ is derived

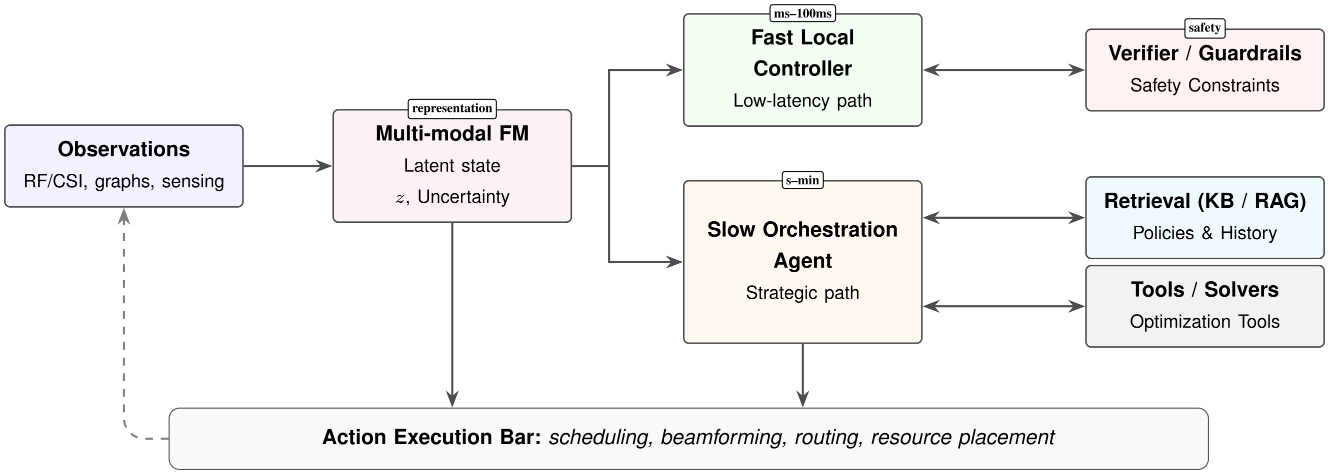

FIGURE 11. Closed-loop control stack enabled by a multi-modal FM. The FM produces representations and uncertainty for fast local controllers and slower orchestration agents. Retrieval and tool calls provide context, while verifiers enforce hard constraints.

TABLE 11. Representative SAGIN control tasks, the dominant decision timescales, and the most informative modalities.

Layer	Control tasks	Typical timescale	Most informative modalities	Notes and risks
Space (LEO/GEO)	Beam hopping schedule, gateway selection, feeder allocation, inter-satellite routing	s–min–hour	Topology graphs, ephemeris/visibility, telemetry (queues), weather/attenuation	Strong coupling to gateways; long RTT; requires safety checks for capacity and visibility
Air (HAPS/UAV)	Relay selection, UAV positioning, multi-link coordination, sensing/comm duty cycling	100 ms–min	RF/CSI summaries, trajectories, interference/connectivity graphs, sensing, local weather	Energy/payload constraints; sensing may be sparse; actions affect coverage and interference
Ground (RAN/Edge)	Scheduling, power control, beam tracking, handover, MEC offloading	ms–s	RF/CSI, mobility/maps, local telemetry, topology (interference) graphs	Fast loop; agentic control typically too slow; rely on FM representations + classical/RL controllers
Core/Cloud	Traffic engineering, slice admission, service placement/caching, cross-layer orchestration	s–min–hour	Network-wide telemetry and logs, topology graphs, traffic traces, policies/text KB	Policy and governance constraints dominate; retrieval and verification improve trustworthiness

mainly from fast RF/CSI and short-window feedback. The slow action constrains the feasible set of fast actions, improving stability and reducing constraint violations under delayed feedback.

F. SAFE CONTROL: GUARDRAILS AND VERIFICATION

Because SAGIN actions can affect wide areas, safe control is essential. Guardrails can be implemented at multiple layers. At the representation level, uncertainty estimates can gate action confidence. At the policy level, constrained RL and safe exploration limit risky actions. At the execution level, a verifier checks actions against hard constraints (e.g., power, spectrum masks, gateway capacity, prohibited routing paths) [188]. In agentic settings, the verifier can be a tool that computes feasibility or a rule engine tied to operator policies [189]. Table 11 categorizes representative control tasks by SAGIN layer and indicates which modalities are most informative. Fig. 11 shows a practical closed-loop architecture where an FM supports both fast local controllers and slower agentic orchestration with retrieval and verification.

X. DISTRIBUTED DEPLOYMENT: ON-BOARD, GATEWAY, AND EDGE SPLIT

The architectural and algorithmic designs in Sections V–IX must ultimately run under stringent system constraints. In SAGINs, deployment is complicated by three simultaneous limitations: i) compute and energy constraints on satellites and aerial platforms; ii) intermittent and capacity-limited links between space/air/ground segments; and iii) stringent latency and reliability requirements for closed-loop control. This section presents practical deployment patterns for multi-modal FMs, focusing on split inference, model compression, intermittency-aware scheduling, and distributed update logistics.

A. WHERE INFERENCE CAN RUN IN SAGINS?

SAGIN inference can be executed at multiple locations: on-board (satellite/UAV/HAPS), at gateways, at terrestrial edge (MEC/RAN), or in the core/cloud [190]. Each location has different tradeoffs. On-board inference reduces dependency on backhaul and enables autonomy, but is limited by power, memory, and radiation-tolerant hardware constraints [55]. Gateway inference offers stronger compute than on-board nodes and has direct visibility into feeder constraints, but can still be bottlenecked by feeder capacity and contention among beams. Edge inference provides low latency to the RAN and proximity to user-plane data, but may not have global visibility. Core/cloud inference offers maximum compute and data access, but suffers from latency and intermittency and raises governance concerns.

A key systems observation is that the best placement is typically task-dependent and timescale-dependent. Fast PHY/RAN control benefits from on-board or edge inference, while slow orchestration can run centrally. Therefore, the deployment design should be coupled to the hierarchical control decomposition in (32).

B. SPLIT INFERENCE AS A FIRST-CLASS DESIGN PRIMITIVE

Split inference partitions model computation across nodes. For a modular multi-modal FM, a natural split is between modality encoders, fusion blocks, and task heads [191]. Let $E_m(\cdot)$ be the encoder for modality m, let $F_\theta(\cdot)$ be the fusion module, and let $g_\phi(\cdot)$ be a head/policy. A generic modular pipeline can be defined as

$$\begin{aligned} \mathbf{H}_t^{(m)} &= E_m(\mathbf{T}_t^{(m)}), \\ \mathbf{z}_t &= F_\theta\left(\mathbf{H}_t^{(m)} : m \in \mathcal{M}_t\right), \\ \mathbf{a}_t &= g_\phi(\mathbf{z}_t). \end{aligned} \tag{33}$$

Split inference chooses a cut point so that some components run locally and the rest run remotely. For example, on-board nodes can compute local embeddings $\mathbf{H}_t^{(m)}$ and transmit them instead of raw data, reducing bandwidth and improving privacy. Alternatively, on-board nodes can compute $\mathbf{z}_t$ and only send a compact state summary to the gateway or cloud for orchestration. Split inference must account for AoI and reliability. Transmitting a stale embedding can be worse than transmitting nothing, so embeddings should carry time stamps and uncertainty [192]. Compression methods (discussed next) further reduce communication cost but can introduce distortion; thus, split points and compression levels should be selected jointly with the controller's robustness requirements.

C. MODEL COMPRESSION: QUANTIZATION, PRUNING, AND DISTILLATION

Compression is essential for on-board and UAV/HAPS deployment. Three practical approaches are widely used. Quantization reduces precision (e.g., 16-bit to 8-bit or lower) to reduce memory and accelerate inference [193]. Quantization-aware training can mitigate accuracy loss. For attention-based models, quantization can be applied to weights and activations, with special care for softmax and layer normalization stability. Pruning removes parameters (structured pruning removes channels/heads; unstructured pruning removes individual weights). Structured pruning is preferable for deployment because it yields real latency gains. Distillation trains a smaller student model to match a larger teacher model. For SAGINs, distillation is particularly attractive for fast local control: a large FM can be used offline or in the cloud to generate labels, representations, or policy targets, and a compact student can run on-board with predictable latency.

The practical execution of SAGIN model must consider the hardware and energy constraints of space and aerial platforms. The compute capacity, memory footprint, and power availability of LEO satellites, UAVs, and HAPS nodes are typically limited. This may, in turn, limit the size and complexity of deployed models. This can partially be addressed using model compression techniques (i.e., quantization, pruning, and knowledge distillation can be employed to reduce inference cost). The deployment strategies must also take into account energy–latency trade-offs when deciding where inference should be executed. Finally, multi-modal pipelines raise additional security concerns, such as adversarial inputs, data integrity violations, and cross-modal attack vectors, requiring secure data validation, robust fusion strategies, and monitoring of abnormal modality behavior.

D. INTERMITTENCY-AWARE SCHEDULING OF COMPUTE AND COMMUNICATION

In SAGINs, links can be intermittent due to visibility windows and congestion. This demands an intermittency-aware interference, i.e., richer modalities and remote compute can be leveraged; when backhaul capacity is limited, the system should revert to local inference using the available modalities and conservative policies. This could practically be implemented using multiple mode operations i.e., a node may operate in a local mode using only on-board modalities, an assisted mode where embeddings are exchanged with a gateway, and a cloud mode for global orchestration [194]. The predicted link availability can be used for mode switching, which can be derived from the ephemeris and topology graphs. This establishes an important feedback loop in which the same multi-modal context used for network optimization also informs the deployment strategy of the FM.

Adaptive system-level strategies are required to address intermittency, latency, and resource coupling in SAGINs. One approach is adaptive mode switching, where inference shifts between centralized and distributed execution based on connectivity and gateway availability. Resource-aware inference can also adjust model complexity, for example through model compression or shorter sequences under limited compute, energy, or bandwidth. In addition, distributed coordination across space–air–ground nodes helps maintain consistent decisions despite delayed or partial information.

E. DISTRIBUTED UPDATES, CACHING, AND VERSIONING

Adaptation updates (Section VIII) must be distributed reliably. PEFT artifacts (adapters and LoRA) are well suited because they are small and can be versioned and rolled back. In SAGINs, updates should be pushed opportunistically when links are available and cached at gateways or edge servers

TABLE 12. Deployment patterns for SAGIN multi-modal FMs and their tradeoffs.

Pattern	Where computation runs	Strengths	Limitations / risks	Typical tasks
On-board end-to-end	Encoders + fusion + head on satellite/UAV/HAPS	Lowest dependency on backhaul; supports autonomy and low-latency local control	Limited compute/energy; model must be small; limited global context	Beam tracking, local relay decisions, emergency fallback control
Gateway-centric	Fusion and orchestration at gateway; optional on-board encoders	Direct access to feeder constraints and gateway visibility; moderate compute	Contention among beams; feeder bottlenecks; may miss local sensing detail	Gateway selection, feeder scheduling, inter-layer routing
Edge-centric (MEC/RAN)	Fusion/control at terrestrial edge	Low latency to RAN; access to user-plane KPIs; scalable in terrestrial infrastructure	Limited global view; depends on backhaul for space/air context	Handover, scheduling assistance, MEC offloading
Cloud-centric	Global orchestration at core/cloud	Maximum compute and data; strong for long-horizon planning and training	Latency/RTT; intermittency; governance/privacy constraints	Traffic engineering, service placement, policy optimization
Split inference (encoders local)	Encoders on-board; fusion/head at gateway/edge/cloud	Reduces bandwidth; preserves privacy; enables richer global fusion	Requires robust embedding compression and AoI handling	Sensing-heavy tasks, cross-layer control with partial local autonomy
Split inference (fusion local)	Encoders + fusion local; head/policy remote	Sends compact state summaries; reduces dependence on raw data	Remote policy may be stale; requires safe local fallback	Hierarchical control: local fast executor, remote slow planner
Distilled students for fast control	Student models on-board/edge; teacher in cloud for training	Predictable latency and energy; high reliability; easy certification	May lose flexibility; needs periodic refresh when drift occurs	Fast PHY/RAN decisions under strict deadlines

for dissemination. To ensure safety, each update should be associated with: i) a validity region (where it is expected to work), ii) a version number and fallback version, and iii) a validation signature indicating that it passed offline safety checks. Such challenges for low powered devices are detailed in [195]. Table 12 summarizes the deployment patterns and their tradeoffs. Figure 10 illustrates representative split points for a multi-modal FM and shows how bandwidth and latency change with the split location.

XI. TRUST, SAFETY, SECURITY, AND PRIVACY

The deployment of multi-modal foundation models in SAGIN environments introduces new security and privacy challenges due to the joint processing of heterogeneous sensing and communication modalities. Potential threats include adversarial manipulation of modality inputs, sensor spoofing, and data poisoning during training, as well as cross-modal attack surfaces where corrupted information in one modality propagates through the fusion pipeline. To mitigate these risks, several system-level strategies are important, including modality-consistency verification, robust training against adversarial perturbations, cross-modal anomaly detection, and secure data pipelines with controlled access and provenance tracking. In addition, privacy-preserving learning approaches such as federated training, secure aggregation, and policy-based data governance can help protect sensitive operational data while enabling collaborative model development across distributed SAGIN infrastructure.

Multi-modal FMs can improve prediction and control in SAGINs, but they also introduce new failure modes and amplify existing ones. The risk surface is broader than in terrestrial-only settings because actions can affect wide areas, feedback can be delayed, and data sources span multiple administrative domains. This section presents a structured view of trustworthiness for SAGIN FMs, covering: i) safety and reliability under uncertainty and distribution shift; ii) security threats such as poisoning, backdoors, and adversarial inputs; and iii) privacy and governance constraints across space−air−ground segments. The goal is to provide a practical checklist for designing FM-driven control loops that can be deployed responsibly.

A. FAILURE MODES IN FM-DRIVEN SAGIN CONTROL

Trust issues arise from both model uncertainty and system coupling. In SAGINs, the following failure modes are particularly important. First, OOD shift occurs when the deployment regime differs from the pretraining/fine-tuning regimes (new weather extremes, novel interference patterns, unexpected gateway outages, new orbital configurations, or new traffic classes) [196]. OOD can cause overconfident predictions and unsafe actions. Second, staleness and missingness can make an otherwise accurate model fail: decisions based on stale telemetry or outdated topology can violate constraints even if RF and CSI are correct locally. Third, cross-layer error propagation means that an incorrect orchestration action (e.g., gateway selection) can trigger cascading congestion and service degradation across beams and layers. Fourth, goal mis-specification (or proxy reward mismatch) can cause learned controllers to optimize easy-to-measure proxies that fail under rare events, such as optimizing average delay while ignoring tail-latency or reliability. Finally, for tool-augmented and retrieval-augmented models, context errors can occur when the retrieved policies are stale, inconsistent, or manipulated.

Reference [197] conducted an empirical study to investigate the factors that influence the failure caused by OOD input. These failure modes motivate trust mechanisms that are both model-centric (uncertainty estimation, calibration, OOD detection) and system-centric (guardrails, verification, fallback control, and rigorous monitoring).

B. UNCERTAINTY ESTIMATION, CALIBRATION, AND OOD DETECTION

Uncertainty should be treated as a first-class output when an FM participates in control. Two broad forms are relevant: aleatoric uncertainty due to noise and partial observability, and epistemic uncertainty due to insufficient training coverage [196]. In practice, uncertainty can be approximated through ensembles, stochastic dropout, or explicit predictive variance heads. Calibration is essential: probabilities or confidence scores must correspond to observed error rates, otherwise uncertainty gating can be misleading. OOD detection can be performed at different levels. At the input level, statistical tests can monitor shifts in modality distributions [198]. At the representation level, the system can monitor whether embeddings $\mathbf{z}_t$ drift outside the training support. At the outcome level, residual-based detectors monitor prediction errors and constraint violation rates. In SAGINs, OOD detection should be combined with AoI and missingness metadata; a distribution shift alarm is less informative if the real issue is a sudden loss of a modality or a large increase in AoI.

C. SAFETY GUARDRAILS AND VERIFIED EXECUTION

Safety in SAGINs is best achieved by combining learned components with explicit checks [199]. A practical approach is to define a set of safe actions $\mathcal{A}_{\text{safe}}(\mathbf{o}_t)$ based on hard constraints and operational rules. The learned policy proposes $\tilde{\mathbf{a}}_t$, and a verifier maps it to an executed action:

$$\mathbf{a}_t = \mathsf{Project}(\tilde{\mathbf{a}}_t, \mathcal{A}_{\text{safe}}(\mathbf{o}_t)), \tag{34}$$

where $\mathsf{Project}(\cdot)$ can be a rule-based filter, a feasibility solver, or a constrained optimization layer. For example, gateway selection can be verified against current visibility and capacity; beam actions can be checked against power and spectrum masks; and routing can be checked against policy constraints and forbidden paths. If verification fails, the system should return to a conservative baseline policy known to be safe.

Safety also interacts with adaptation (Section VIII). PEFT updates should be validated offline, versioned, and rolled back if post-deployment monitoring detects anomalous constraint violations.

D. SECURITY THREATS: POISONING, BACKDOORS, AND ADVERSARIAL MODALITIES

Security threats can be grouped into training-time and inference-time attacks [200]. At training time, data poisoning injects corrupted samples so that the model learns wrong associations, while backdoor attacks implant triggers that cause specific erroneous actions when a trigger pattern appears [201], [202], [203]. Federated learning introduces an additional risk because model updates can be malicious. Robust aggregation, anomaly detection in updates, and secure enclaves can reduce these risks.

At inference time, adversarial examples can be constructed for RF features, sensing inputs, and graphs [204]. For example, manipulated sensing or spoofed navigation data can cause wrong blockage inference; falsified telemetry can cause wrong congestion reactions; and graph perturbations can mislead routing decisions. RF-specific threats include jamming and spoofing that corrupt measurement reports and degrade representation quality [205], [206]. A robust FM-driven control loop should therefore: i) cross-validate modalities (e.g., compare RF trends with trajectory and topology consistency); ii) use confidence and AoI gating; iii) apply input sanitization; and iv) retain safe fallbacks under suspicious conditions.

E. PRIVACY AND GOVERNANCE IN MULTI-DOMAIN SAGIN DATA

Privacy concerns arise from traffic traces, location trajectories, and sensing modalities. In addition, SAGIN data are often distributed across operators and administrative domains (space segment provider, HAPS/UAV operator, terrestrial operator) [207], [208]. Governance constraints may restrict the movement of raw data to central locations. This motivates privacy-aware learning and deployment: federated learning, secure aggregation, access-controlled retrieval, and minimization of transmitted information via split inference (Section X).

For Retrieval augmented models, governance requires; i) controlled document ingestion, ii) redaction of sensitive fields, iii) access control and auditing, and iv) verification of retrieved context. Without these measures, retrieval can become an attack surface where malicious or stale documents induce unsafe actions. Table 13 summarizes threats and mitigations. Fig. 12 illustrates a layered trust stack for FM-driven SAGIN control, showing where uncertainty, verification, and governance mechanisms fit into the pipeline.

XII. BENCHMARKING, EVALUATION PROTOCOLS, AND REPRODUCIBILITY

Benchmarking is the main bottleneck for progress in multi-modal FMs for SAGINs. Without standardized tasks, datasets, and reporting protocols, it is difficult to fairly compare architectures, pretraining objectives, and deployment strategies. This section proposes an evaluation blueprint tailored to SAGINs. The blueprint is designed to: i) cover

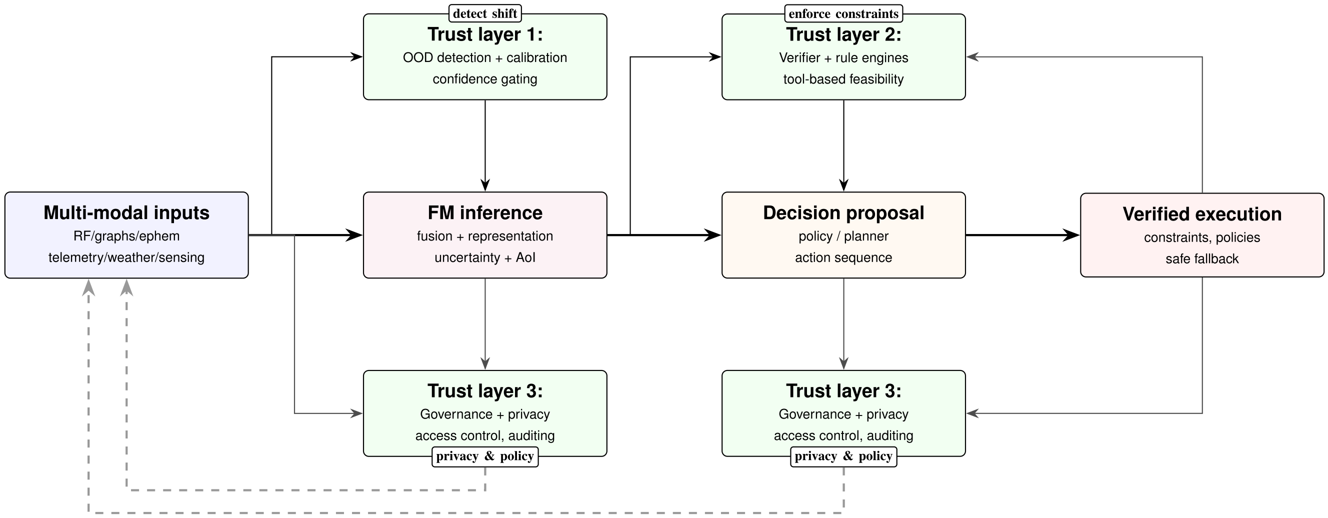

FIGURE 12. Layered trust stack for FM-driven SAGIN control. OOD detection and calibrated uncertainty gate risky decisions; verifiers and rule engines enforce hard constraints and policy compliance; governance mechanisms control access, retrieval, and privacy-sensitive data flows.

TABLE 13. Threats and mitigations for multi-modal FMs in SAGINs.

Threat / failure mode	How it manifests in SAGINs	Impact on prediction/control	Practical mitigations
OOD shift	New weather extremes, new interference regimes, gateway outages, novel mobility patterns	Overconfident predictions; unsafe actions; cascading failures	OOD detection (input/embedding/outcome), calibrated uncertainty, conservative fallback, periodic revalidation
Staleness / missingness	High AoI telemetry, delayed topology updates, sensing dropout	Constraint violations from acting on outdated context	AoI-aware tokenization and gating, mode switching (local vs assisted), freshness thresholds
Reward/proxy mismatch	Optimizing average KPIs while ignoring tail reliability/safety constraints	Good average performance but severe rare-event failures	Constraint-centric metrics, risk-sensitive objectives, scenario-based validation, safe exploration
Poisoning / backdoors	Corrupted datasets or malicious federated updates; trigger patterns in logs/sensing	Persistent vulnerabilities; targeted unsafe actions	Robust aggregation, update anomaly detection, dataset provenance, backdoor scans, versioned rollbacks
Adversarial RF/sensing	Spoofed navigation/sensing, manipulated telemetry, jamming/spoofing affecting measurements	Misleading representations; wrong beam/routing actions	Cross-modal consistency checks, input sanitization, multi-sensor validation, fallback to robust baselines
Retrieval/context attacks	Stale or manipulated policies/configs retrieved via a knowledge base	Policy violations, unsafe orchestration actions	Access control, auditing, trusted ingestion, verification of retrieved facts, time-validity tagging
Privacy leakage	User mobility/traffic inference; sensitive sensing content exposure	Regulatory and trust violations; operational risk	Split inference with embeddings, federated learning, secure aggregation, redaction, differential privacy where applicable

both prediction and closed-loop control; ii) incorporate multi-layer and multi-timescale constraints; iii) evaluate robustness to missingness, staleness, and distribution shift; and iv) enforce reproducibility through a minimal reporting checklist.

A. BENCHMARK DESIGN PRINCIPLES FOR SAGIN MULTI-MODALITY

A SAGIN benchmark should reflect the operational realities that motivate multi-modal learning. First, it should include at least one task where radio measurements alone are insufficient, so that the value of additional modalities can be quantified. Second, it should explicitly model modality missingness and AoI, rather than assuming perfectly aligned synchronous streams. Third, it should include scenarios where constraints are binding, such as gateway capacity limits, visibility windows, and reliability targets. Fourth, it should test generalization across domains by including multiple environments (urban/rural, different weather regimes, different traffic profiles, different platform densities). These principles lead to three complementary benchmark categories: prediction benchmarks, control benchmarks, and deployment/system benchmarks. References [209], [210], and [211] conducted a detailed survey on benchmarking for AI, such techniques can be extended for benchmarking in multi-modality cases.

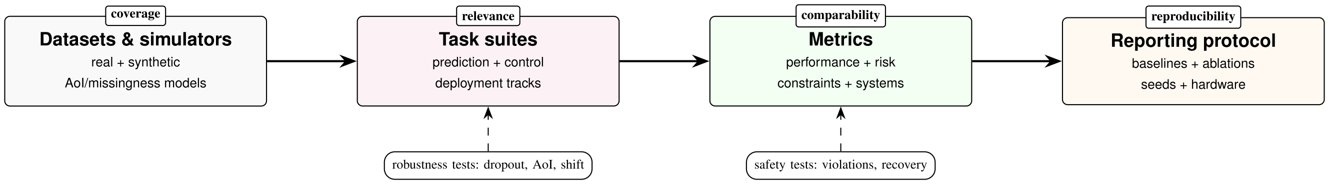

FIGURE 13. Benchmarking blueprint for SAGIN multi-modal FMs. Datasets and simulators must explicitly model AoI and missingness, tasks must include prediction/control/deployment tracks, metrics must include risk and constraint compliance, and reporting must include baselines/ablations and hardware details.

B. PREDICTION BENCHMARK SUITE

Prediction tasks assess whether a model extracts robust representations from multi-modal context and can forecast future network behavior. Representative tasks include: 1) Prediction of traffic and demand across space−air−ground layers, where input may include historical traffic, topology graphs, service metadata, and external events; 2) Link quality and outage prediction, where the target may be future SNR/RSRP, outage probability, or BLER, conditioned on RF summaries, ephemeris/trajectory, and weather; 3) Topology and visibility prediction, where the model predicts future connectivity graphs or gateway visibility windows; 4) Blockage and environment-aware prediction, where sparse sensing snapshots supervise the prediction of blockage likelihood or propagation regimes.

Prediction benchmarks should be evaluated using both average accuracy metrics (mean squared error (MSE)/mean absolute error (MAE)) and reliability metrics such as calibration error and prediction interval coverage. Because SAGIN decisions are safety-critical, a model that is slightly less accurate but better calibrated can be preferable.

C. CLOSED-LOOP CONTROL BENCHMARK SUITE

Control benchmarks evaluate end-to-end decision quality under constraints and feedback. A practical control benchmark should provide a simulator or digital twin that supports: i) multi-layer topology dynamics; ii) beam and mobility models; iii) gateway/feeder constraints; and iv) multi-modal observation generation with configurable AoI and missingness. Representative control tasks include: 1) Beam management and handover under mobility and blockage, with optional sensing and weather context; 2) Gateway selection and feeder scheduling under visibility windows and congestion, using topology graphs and telemetry; 3) Inter-layer traffic steering and routing across satellite−HAPS−ground paths; 4) Service placement and compute offloading with latency and energy constraints.

The evaluation of the control must go beyond the average reward. It should report constraint violation statistics, tail-latency, outage rates, and stability measures (e.g., oscillatory handovers, routing churn). Robustness tests should include adversarially increased AoI, random modality dropouts, and scenario shifts.

D. DEPLOYMENT AND SYSTEM BENCHMARKS

System benchmarks quantify whether the proposed FM pipeline is deployable. These benchmarks include: i) Inference latency and throughput on representative hardware (on-board class, gateway class, edge class); ii) Energy cost per inference and per adaptation update; iii) Bandwidth cost of communication for split inference (e.g., raw data vs embeddings vs state summaries); iv) Robustness under intermittent connectivity (mode switching frequency, performance degradation under outages). Such benchmarks enable honest tradeoff discussion. In SAGINs, a slightly weaker model that meets strict latency and energy budgets may be preferable to a larger model that cannot run reliably on-board or at the edge. We also present detail benchmarking template for for multi-modal FMs in SAGINs in Table 14. Moreover, Fig. 13 presents a detailed (dataset, task suites, Metrics, and reporting) blueprint for benchmarking for SAGIN.

E. METRICS: BEYOND SPECTRAL EFFICIENCY

To align with 6G requirements, we recommend reporting metrics in five categories: 1) Performance: throughput, delay, packet delivery ratio, outage probability; 2) Reliability and risk: tail latency (e.g., 95/99 percentile), worst-case outage, conditional value-at-risk (CVaR) style metrics; 3) Constraint compliance: violation rate, severity, time-to-recovery after violations; 4) Robustness: performance under modality dropout, AoI increase, and domain shift; 5) Systems: latency, energy, memory footprint, bandwidth for split inference, and adaptation update size.

F. BASELINES AND ABLATIONS FOR FAIR COMPARISON

A credible benchmark must include baselines that isolate the benefit of multi-modality and pretraining. We recommend: i) Unimodal baselines (RF-only, telemetry-only, graph-only); ii) Simple fusion baselines (late fusion ensembles, gated fusion); iii) Task-specific models without pretraining; iv) FM-based variants with different objectives (masked-only vs contrastive-only vs mixed); v) Ablations for each modality (remove weather, remove ephemeris, remove sensing) and for AoI/missingness mechanisms (without gating, without AoI features). Because SAGIN multi-modality can be expensive, the benchmark should quantify the marginal benefit of each modality relative to its cost (bandwidth, privacy, compute).

Although this survey does not introduce new experimental results, it aims to clarify how empirical validation of multi-modal foundation models can be conducted in SAGIN environments. In particular, practical evaluation typically relies on large-scale network simulators, digital twins, and hardware-in-the-loop testbeds that emulate multi-layer

TABLE 14. Proposed benchmark suite template for multi-modal FMs in SAGINs. Each row can be instantiated with multiple scenarios/domains to test generalization.

Benchmark track	Task examples	Required modalities	Key metrics	Minimal baselines/ablations
Prediction	Traffic forecasting; outage prediction; connectivity forecasting	Telemetry/ traffic, graphs, ephemeris; optional weather/ sensing	MAE/ MSE; calibration error; prediction interval coverage; OOD detection AUC	Unimodal predictors; late fusion; FM without pretraining; remove one modality at a time
Control	Beam/ handover; gateway selection; routing/ steering; placement/ offloading	RF/ CSI summaries, graphs, telemetry, ephemeris; optional weather/sensing	Reward; constraint violation rate/ severity; tail latency; outage; churn/ oscillation	Heuristics/ MPC; RL with hand-crafted features; FM embeddings vs raw inputs; AoI gating ablation
Deployment	Split inference; compression; intermittency mode switching	All (as available) + system traces	Latency, energy, memory; bandwidth; performance vs AoI/ outage; update size	No-split vs split; raw vs embeddings; quantized vs full precision; student vs teacher
Robustness	Modality dropout; AoI increase; scenario shift; adversarial perturbation	Configurable	Performance degradation curves; safety violations; recovery time	Without uncertainty; without verifier; without retrieval; conservative fallback only

space–air–ground dynamics. The benchmarking framework discussed in this section therefore outlines representative prediction and control tasks together with system-level metrics such as latency, reliability, constraint violations, and energy consumption. These evaluation pipelines provide a practical pathway for translating the surveyed architectures into reproducible experimental studies and operational testing in future SAGIN deployments.

XIII. STANDARDIZATION, PRACTICAL ROAD-MAP, AND OPEN RESEARCH DIRECTIONS

This section consolidates the tutorial lessons into a practical road-map for deploying multi-modal FM driven intelligence in space−air−ground integrated 6G systems. We emphasize engineering decisions that affect feasibility in real networks, highlight where standardization and industry practice can benefit from common interfaces and benchmarks, and identify open research directions that are both technically important and realistically actionable.

A. WHAT SHOULD BE STANDARDIZED: INTERFACES OVER IMPLEMENTATIONS?

Standardization efforts for 6G are likely not to prescribe specific learning models, but can greatly benefit from standardizing interfaces and reporting conventions that enable interoperability and safe deployment. For SAGIN multi-modal FMs, the most important interface candidates are: i) Modality metadata formats (time stamps, coordinate frames, AoI, missingness masks, and confidence tags); ii) Representation exchange formats for split inference (embedding schemas, compression descriptors, validity windows); iii) Control-loop safety hooks (constraint descriptors, verifier APIs, fallback policies); iv) Benchmark and reporting templates (task definitions, metrics, baseline sets, and reproducibility checklists).

These interfaces reduce friction when FMs are deployed across multi-operator environments, where space, air, and ground segments may be managed by different stakeholders. They also enable safe composability: a model can be replaced while maintaining verifiers and monitoring pipelines.

B. ENGINEERING ROAD-MAP: FROM PROTOTYPES TO OPERATIONS

A realistic road-map can be organized into three stages that progressively increase autonomy while controlling risk.

Stage 1: Assisted intelligence. Models are used primarily for prediction and decision support. They produce forecasts (traffic, outage risk, congestion risk) and representations for downstream controllers, but they do not directly execute high-impact actions without human or rule-based approval. This stage emphasizes benchmarking, calibration, and data governance.

Stage 2: Guarded autonomy. Models participate in closed-loop control with explicit guardrails. Verification and safe fallbacks are mandatory, and action scopes are limited (e.g., local beam/relay actions, bounded traffic steering). Split inference and PEFT-based updates enable deployment under realistic constraints.

Stage 3: Hierarchical autonomy. Multi-timescale controllers coordinate across layers. Slow planners operate with broad context and retrieval/tool support, while fast executors act locally with distilled models. Continual learning and federated updates become routine, driven by drift detection and validated update pipelines.

This staged path is not only technical; it is also organizational. It provides a clear trajectory for risk management and can guide the development of certification and operational practices.

C. PRIORITIZED OPEN RESEARCH DIRECTIONS

We highlight open problems where progress would materially advance SAGIN multi-modal FMs. Each direction is phrased in terms of a concrete deliverable, rather than a vague aspiration.

1) *Modality alignment under imperfect time and space synchronization.* Most multi-modal learning assumes

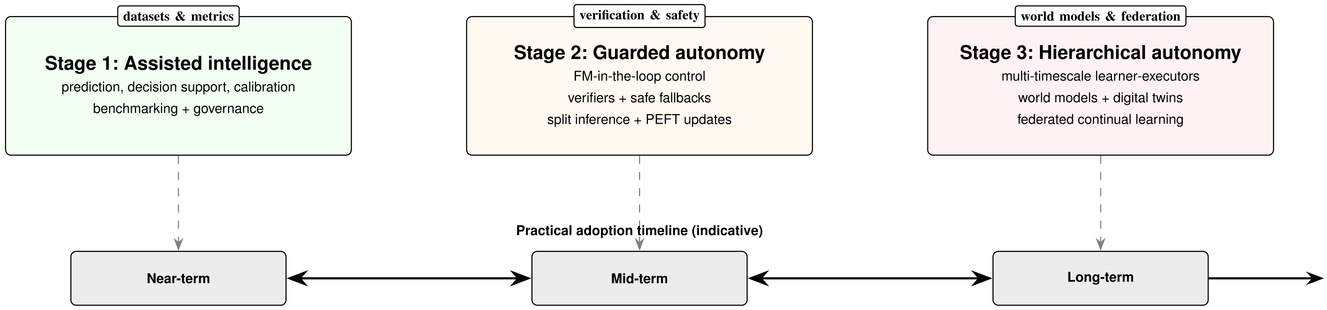

FIGURE 14. **A practical roadmap for adopting multi-modal FM-driven intelligence in SAGINs. Early deployment focuses on prediction and decision support, followed by guarded autonomy with verifiers and split inference, and finally hierarchical autonomy enabled by world models and federated continual learning.**

accurate alignment. In SAGINs, alignment errors are inherent due to RTT, buffering, and intermittent reporting. A key open problem is designing pretraining and fusion mechanisms that explicitly model alignment uncertainty and learn to associate signals across flexible time windows and spatial supports.

2) *AoI and missingness-aware foundations.* AoI and missingness are not nuisances; they are fundamental features. Developing foundation models that treat AoI as a first-class conditioning variable, and learning policies that remain safe when modalities disappear, none the less it is essential for operational robustness.
3) *Multi-layer world models and hybrid digital twins.* closed-loop control and safe exploration require predictive models of SAGIN dynamics. A promising direction is hybrid world modeling: combining digital twins with learned residual models and uncertainty quantification, enabling credible synthetic rollouts for training and validation.
4) *Trustworthy agentic orchestration.* Agentic designs that use retrieval tools can increase verifiability, but also introduce new attack surfaces and failure modes. Open questions include secure retrieval pipelines, tool reliability under intermittency, action grounding, and formalizing verification interfaces that scale across operators.
5) *Distributed deployment with guaranteed latency and energy budgets.* Split inference and compression are central, but current methods rarely provide predictable performance guarantees. SAGIN deployment needs methods that jointly optimize split points, compression levels, and controller robustness under varying link availability.
6) F*ederated and multi-operator learning with heterogeneous modalities.* In realistic SAGINs, not all nodes observe the same modalities. Federated learning must therefore handle heterogeneous feature spaces, partial participation, and poisoning resistance while maintaining meaningful global representations.
7) *Benchmarking and reproducibility at scale.* Progress depends on shared benchmarks that include AoI, missingness, and multi-layer constraints. The community needs standardized tasks and evaluation protocols, along with transparent reporting of compute, latency, and energy.

Table 15 organizes these directions and maps them to layers and maturity. Fig. 14 outlines a practical adoption road-map for multi-modal FM-driven intelligence in SAGINs, progressing from assisted decision support to guarded and ultimately hierarchical autonomy.

Several important open challenges remain in realizing practical multi-modal foundation model deployments in SAGIN environments. Promising research directions include federated and collaborative learning across distributed space–air–ground nodes, which can enable knowledge sharing while respecting data locality and privacy constraints. Another critical challenge is improving adversarial robustness and reliability when modality inputs may be corrupted, spoofed, or partially unavailable. In addition, enabling cross-domain generalization across heterogeneous deployment environments, orbital configurations, traffic patterns, and hardware platforms remains an open problem. Future work must also address scalable data collection, multi-modal dataset standardization, resource-constrained deployment strategies, and evaluation frameworks capable of assessing closed-loop network control performance in realistic SAGIN scenarios.

Gap Analysis: Despite rapid progress in applying foundation models to wireless networking, several key research gaps remain in the context of SAGIN environments. First, scalable multi-modal data pipelines capable of integrating heterogeneous modalities across space–air–ground segments remain largely unexplored. Second, robustness under intermittent connectivity, missing modalities, and delayed updates continues to pose a significant challenge for reliable deployment. Third, current multi-modal learning approaches often assume centralized training and stable infrastructure, whereas SAGIN systems require distributed, resource-constrained, and intermittency-aware learning strategies. Finally, systematic benchmarking and reproducible evaluation frameworks for multi-modal SAGIN intelligence are still in their early stages. Addressing these gaps will be critical for enabling practical deployment of multi-modal foundation models in next-generation 6G networks.

TABLE 15. Open research directions for SAGIN multi-modal foundation models, mapped to layers, key technical bottlenecks, and maturity.

Direction	Dominant layers	Core bottlenecks	Maturity	Example deliverable
Alignment under imperfect sync	All	Time-window positives, spatial support mismatch, RTT-induced delays	Medium	Cross-modal pretraining with alignment uncertainty and AoI-aware attention
AoI/missingness-aware foundations	All	Robust fusion under dropout, safe degradation, confidence gating	Medium	FM + policy that maintains constraint compliance under modality outages
Hybrid world models (twin + residual)	Space/air/core	Twin mismatch, uncertainty, long-horizon prediction stability	Early–medium	Validated world model for gateway/ beam/ traffic rollouts with calibrated uncertainty
Trustworthy agentic orchestration	Core/gateway	Secure retrieval, tool reliability, verification interfaces, governance	Early	Retrieval+tool orchestration stack with audited KB and constraint verifiers
Latency/energy-guaranteed deployment	Space/air/edge	Split optimization, compression distortion, intermittency scheduling	Medium	Joint split/compress policy with predictable latency–energy envelopes
Federated multi-operator learning	All	Heterogeneous modalities, partial participation, poisoning resistance	Early-medium	Federated PEFT updates with robust aggregation and cross-domain generalization tests
Benchmarking and reproducibility	All	Standard tasks, datasets, simulators, reporting norms	Medium	Public benchmark suite with AoI/ missingness and multi-layer constraints

Illustrative Evaluation Scenario: As an example, consider a SAGIN deployment consisting of LEO satellites, UAV relays, and terrestrial RAN nodes serving mobile users. A multi-modal foundation model receives heterogeneous inputs including RF/CSI measurements, topology graphs, platform trajectories, telemetry statistics, and environmental context. The evaluation framework can assess prediction tasks such as traffic forecasting or beam quality estimation, as well as control tasks including handover decisions or traffic steering. Performance can be measured using metrics such as prediction accuracy, latency, reliability, spectral efficiency, and robustness to missing modalities or delayed updates. Such controlled simulation scenarios provide a practical mechanism for evaluating the effectiveness and robustness of multi-modal foundation models under realistic SAGIN conditions.

XIV. LESSONS LEARNED AND PRACTICAL GUIDELINES

Although the preceding sections surveyed architectures, data pipelines, training paradigms, deployment patterns, and benchmarking practices for multi-modal FMs for 6G-and-beyond SAGINs, a recurring observation is that accuracy improvements alone do not translate into deployable intelligence. Real-world gains emerge only when modeling choices are aligned with the physical realities of multi-layer connectivity (intermittency, long RTTs, mobility), heterogeneous sensing and telemetry (missingness, misalignment, drift), and strict operational constraints (safety, reliability, energy, and privacy). This section consolidates cross-cutting lessons learned and distills them into practical guidelines for designing, training, and deploying multi-modal foundation models for network planning and closed-loop control.

A. TREAT DATA FRESHNESS AND MISSINGNESS AS FIRST-CLASS INPUTS

A central lesson is that multi-modal network intelligence is limited less by model capacity than by data availability and validity at decision time. Since each modality exhibits its own sampling rate, latency, and outage behavior, models should explicitly encode freshness (e.g., timestamps or staleness indicators) and propagate it through fusion and decision heads, rather than forcing the network to infer staleness implicitly. Likewise, missingness should be handled via modality masks and uncertainty-aware fusion so that low-confidence or absent inputs are down-weighted instead of being implicitly imputed. In evaluation, robustness should be demonstrated using systematic dropout and staleness stress tests (e.g. missing sensing streams, delayed CSI, or sparse topology updates) to verify graceful degradation under realistic operational conditions.

B. HYBRID FUSION IS THE PRACTICAL DEFAULT IN HETEROGENEOUS 6G STACKS

Pure early fusion is often computationally and bandwidth intensive, while pure late fusion frequently misses cross-modal dependencies (e.g., mobility channel coupling, weather link availability, and topology routing interactions). In practice, a robust pattern is hybrid fusion: lightweight per modality encoders produce compact representations that are integrated by an intermediate cross-attention (or gating) module, followed by task-specific heads. To improve reuse and deployability, it is preferable to adopt modular encoders with a shared fusion trunk and small heads, enabling parameter sharing across tasks and facilitating split inference. Moreover, fusion should be tailored to the decision timescale and layer (fast control relies on short-term local state, whereas slower planning benefits from topology, demand context, and ephemeris/trajectory priors); when modalities arrive asynchronously, buffered tokens and freshness embeddings are typically more stable than forced synchronization.

C. MULTI-TIMESCALE HIERARCHY IS NOT OPTIONAL FOR CLOSED-LOOP CONTROL

In multi-layer 6G, decision-making naturally decomposes into slow planning (minutes−hours), mid-term adaptation (seconds), and fast control (milliseconds). Monolithic end-to-end policies that ignore this hierarchy often fail due to partial observability and unavoidable delays. A more reliable design is hierarchical control, where slow planners set goals, budgets, and constraints, and fast controllers operate within verified envelopes. Practically, this can be implemented using hierarchical heads (planner−controller) or separate agents connected through explicit interfaces carrying goals/constraints/budgets. Training and evaluation should enforce time-consistent features (stable context for slow modules and instantaneous local features for fast modules) and should report not only steady-state reward but also transient behavior, such as recovery time after link loss, constraint violation rate/severity, and oscillations under feedback delays.

D. DEPLOYMENT CONSTRAINTS SHOULD SHAPE THE ARCHITECTURE FROM DAY ONE

A recurring pitfall is designing a model that performs well offline but cannot be placed on realistic compute nodes across the multi-layer stack. Split inference and edge/on-board execution impose architectural requirements that must be considered early, including modularity, partitionability, and robustness under compression/quantization. A deployable design typically places modality encoders on-device/on-board and forwards compact embeddings to edge/core for heavier fusion and planning, using communication-aware intermediate representations (e.g., learned bottlenecks or token compression) to reduce overhead. Finally, latency and energy budgets should be treated as optimization objectives during model selection and training (e.g., multi-objective tuning over accuracy−latency−energy), rather than as post hoc constraints.

E. RETRIEVAL AND TOOLS REDUCE HALLUCINATION AND IMPROVE DECISION GROUNDING

For agentic network orchestration, pure parametric memory is insufficient because policies must incorporate recent measurements, historical incidents, configuration constraints, and operational documentation. Retrieval augmented designs help by grounding decisions in stateful context (recent KPIs, alarms, logs, and topology snapshots) and institutional knowledge (policies, rules, and run-books). Reliability can be further improved by coupling generation with tools such as feasibility checks (power/latency), optimization subroutines, simulators, and rule-based validators, thereby converting ambiguous suggestions into verifiable actions. To support reproducibility and analysis, retrieved context and tool outputs should be cached and versioned alongside model and configuration identifiers.

F. TRUST REQUIRES EXPLICIT SAFETY LAYERS AND VERIFICATION, NOT ONLY BETTER TRAINING

closed-loop network control is safety-critical: constraint violations can degrade QoS, waste energy, or trigger cascading failures. Therefore, safety should be enforced at the system level rather than treated as an emerging property of training alone. A practical pattern is to separate the policy proposal from the policy execution, where the proposed actions pass through constraint verifiers and rule engines before being applied. When confidence is low, inputs are stale, or verifiers reject an action, the system should return to conservative heuristics or safe baseline controllers. In addition, uncertainty tracking and selective deferral (e.g., escalation to slower planners, optimization-based controllers, or human-in-the-loop mechanisms) improves reliability under rare events and distribution shifts.

G. BENCHMARKS MUST QUANTIFY BENEFIT-PER-COST AND FAILURE MODES, NOT ONLY AVERAGE GAINS

Average SE (or average reward) is insufficient to evaluate multi-modal foundation models for 6G operation. Practical deployments are concerned with tail events, robustness to modality loss, latency/energy overhead, privacy costs, and operational stability; consequently, benchmarking should quantify both gains and costs. Results should report the marginal utility of each modality (performance improvement versus added bandwidth, latency, compute, and privacy exposure) and include tail-risk metrics (e.g., worst-case percentiles or CVaR-style summaries), constraint violation frequency/severity, and recovery time after disruptions. To enable fair comparison, studies should standardize ablations spanning modality removal, fusion alternatives, retrieval/tool usage, and deployment settings (e.g., quantization level and split points).

H. STANDARDIZE INTERFACES AND METADATA EARLIER THAN THE MODELS

In fast-evolving ecosystems, model architectures and training recipes will change rapidly; however, interfaces and metadata conventions provide durable foundations for interoperability across vendors and layers. Standardization efforts should therefore prioritize data schemas (including timestamps, coordinate frames, and modality confidence/masks), embedding exchange formats for split inference, and constraint/verification APIs that decouple safety enforcement from any particular model. In addition, rigorous versioning of models, datasets, prompts/tools, and evaluation protocols is essential to maintain lineage and reproducibility. Adopting an interfaces-first design philosophy ensures that new models can be integrated into existing orchestration pipelines without re-engineering the entire stack.

Overall Takeaway: The overarching lesson is that multi-modal FMs in multi-layer 6G networks should be treated as components within a safety and deployment-aware control stack rather than as standalone predictors. Designs that explicitly account for freshness/missingness, hybrid fusion, multi-timescale hierarchy, retrieval/tool grounding, and verification can achieve not only strong benchmark results but also robust and operationally trustworthy performance in realistic 6G-and-beyond environments.

XV. CONCLUSION

Multi-modal FMs provide a unifying pathway to AI-native operation in 6G and beyond SAGINs, where reliable decisions depend on heterogeneous context beyond RF measurements. This paper presented a tutorial survey covering modality taxonomies, tokenization and fusion, architecture families, pretraining objectives, adaptation and continual learning, closed-loop control integration, distributed deployment, trust and governance mechanisms, and benchmarking protocols. The road-map outlined in this section highlights a staged approach to adoption that balances autonomy with safety through verifiers, uncertainty-aware fusion, and reproducible evaluation. The paper also highlighted potential research directions. The proposed interfaces and benchmark blueprint will accelerate progress toward robust, deployable multi-modal intelligence for 6G and beyond SAGINs.

REFERENCES

[1] W. U. Khan et al., "Integration of NOMA with reflecting intelligent surfaces: A multi-cell optimization with SIC decoding errors," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 3, pp. 1554–1565, Sep. 2023.

[2] M. Ahmed et al., "A survey on reconfigurable intelligent surfaces assisted multi-access edge computing networks: State of the art and future challenges," *Comput. Sci. Rev.*, vol. 54, Nov. 2024, Art. no. 100668.

[3] M. Ahmed et al., "NOMA-based backscatter communications: Fundamentals, applications, and advancements," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19303–19327, Jun. 2024.

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

[5] W. U. Khan et al., "Opportunities for physical layer security in UAV communication enhanced with intelligent reflective surfaces," *IEEE Wireless Commun.*, vol. 29, no. 6, pp. 22–28, Dec. 2022.

[6] W. U. Khan et al., "Rate splitting multiple access for next generation cognitive radio enabled LEO satellite networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8423–8435, Nov. 2023.

[7] X. Cao, X. Su, P. Yang, Y. Gao, D. Oliver Wu, and T. Q. S. Quek, "Survey on near-space information networks: Channel modeling, transmission, and networking perspectives," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 672–715, 2026.

[8] W. U. Khan, E. Lagunas, A. Mahmood, S. Chatzinotas, and B. Ottersten, "RIS-assisted energy-efficient LEO satellite communications with NOMA," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 2, pp. 780–790, Jun. 2024.

[9] X. Luo, Q. Lin, R. Zhang, H.-H. Chen, X. Wang, and M. Huang, "ISAC—A survey on its layered architecture, technologies, standardizations, prototypes, and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 485–526, 2026.

[10] F. Xu et al., "The state of AI-empowered backscatter communications: A comprehensive survey," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21763–21786, May 2023.

[11] F. Liu et al., "Few-shot adaptation of multi-modal foundation models: A survey," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 268, Aug. 2024.

[12] C. Jiang et al., "Foundation model is efficient multimodal multitask model selector," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 33065–33094.

[13] H. Zhang, M. Farzanullah, M. Ghassemi, A. B. Sediq, A. Afana, and M. Erol-Kantarci, "Multi-modal data-enhanced foundation models for prediction and control in wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 4359–4393, 2026.

[14] J. Du, T. Lin, C. Jiang, Q. Yang, C. F. Bader, and Z. Han, "Distributed foundation models for multi-modal learning in 6G wireless networks," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 20–30, Jun. 2024.

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

[16] S. Luo et al., "Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 12, pp. 8040–8063, Dec. 2024.

[17] H. Chen, Y. Zhang, D. Krompass, J. Gu, and V. Tresp, "FedDAT: An approach for foundation model finetuning in multi-modal heterogeneous federated learning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 10, pp. 11285–11293.

[18] C. Ren, M. Li, C. Chen, X. Guan, and G. Q. Huang, "Multi-modal digital twins for industrial anomaly detection: Framework, method, and application," *Robot. Comput.-Integr. Manuf.*, vol. 97, Feb. 2026, Art. no. 103068.

[19] L. Yu, L. Shi, J. Zhang, Z. Zhang, Y. Zhang, and G. Liu, "ChannelGPT: A large model toward real-world channel foundation model for 6G environment intelligence communication," *IEEE Commun. Mag.*, vol. 63, no. 10, pp. 68–74, Oct. 2025.

[20] L. Bai, Z. Han, X. Cai, and X. Cheng, "Multi-modal intelligent channel modeling framework for 6G-enabled networked intelligent systems," *IEEE Wireless Commun.*, early access, Nov. 13, 2025, doi: 10.1109/MWC.2025.3624714.

[21] M. B. López, "Beyond 6G: Multimodal approaches for joint communication and sensing," *Authorea Preprints*, 2025.

[22] F. Jiang et al., "A comprehensive survey of large AI models for future communications: Foundations, applications and challenges," 2025, *arXiv:2505.03556*.

[23] J. Xu, L. Zhou, S. Zhang, X. Xie, W. Yuan, and J. Wang, "Intelligent foundation model-enabled low-altitude FANETs: Integrating communication, control, sensing, and security," *IEEE Internet Things Mag.*, vol. 9, no. 2, pp. 84–92, Mar. 2026.

[24] Y. Wang, C. Yang, T. Li, Y. Ouyang, X. Mi, and Y. Song, "A survey on intent-driven end-to-end 6G mobile communication system," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 882–915, 2026.

[25] N. Yang, M. Fan, W. Wang, and H. Zhang, "Decision-making large language model for wireless communication: A comprehensive survey on key techniques," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 3055–3088, 2026.

[26] S. Çimen, M. Altıntaş, I. Duru, S. N. Karahan, and I. Yazıcı, "An overview of large language models in 6G radio access networks," in *Proc. Int. Conf. Electr., Commun. Comput. Eng. (ICECCE)*, 2025, pp. 1–6.

[27] X. Chen et al., "Communication and control co-design in 6G: Sequential decision-making with LLMs," *IEEE Netw.*, vol. 39, no. 4, pp. 131–138, Jul. 2025.

[28] L. Yue and T. Chen, "AI large model and 6G network," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2023, pp. 2049–2054.

[29] Z. Duan et al., "Sensing and understanding the world over air: A large multimodal model for mobile networks," 2025, *arXiv:2511.21707*.

[30] N. A. Kanıgür, M. Mert, and İ. Duru, "Leveraging large language models and artificial intelligence for uavs in 6g-enabled non-terrestrial networks," in *Proc. 9th Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, May 2025, pp. 1–6.

[31] M. Ahmed et al., "Toward a sustainable low-altitude economy: A survey of energy-efficient RIS–UAV networks," *IEEE Internet Things J.*, vol. 12, no. 24, pp. 51951–51975, Dec. 2025.

[32] K. Guu, K. Lee, K. Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3929–3938.

[33] Q. Long, W. Wang, and S. Pan, "Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 6525–6542.

[34] X. Gao et al., "Agentic satellite-augmented low-altitude economy and terrestrial networks: A survey on generative approaches," 2025, *arXiv:2507.14633*.

[35] X. Huang, Y. Tang, J. Li, N. Zhang, and X. Shen, "Toward effective retrieval augmented generative services in 6G networks," *IEEE Netw.*, vol. 38, no. 6, pp. 459–467, Nov. 2024.

[36] Y. Tang and W. Guo, "Automatic retrieval-augmented generation of 6G network specifications for use cases," *IEEE Commun. Mag.*, vol. 63, no. 4, pp. 95–102, Apr. 2025.

[37] J. J. J. Pey, S. M. B. P. Samarakoon, M. A. V. J. Muthugala, and M. R. Elara, "A decentralized partially observable Markov decision process for complete coverage onboard multiple shape changing reconfigurable robots," *Expert Syst. Appl.*, vol. 271, May 2025, Art. no. 126565.

[38] S. Javaid, R. A. Khalil, N. Saeed, B. He, and M.-S. Alouini, "Leveraging large language models for integrated satellite-aerial-terrestrial networks: Recent advances and future directions," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 399–432, 2025.

[39] A. Nauman et al., "Dynamic resource management in integrated NOMA terrestrial–satellite networks using multi-agent reinforcement learning," *J. Netw. Comput. Appl.*, vol. 221, Jan. 2024, Art. no. 103770.

[40] A. Meftah, T. N. Do, G. Kaddoum, and C. Talhi, "Federated learning-enabled jamming detection for stochastic terrestrial and non-terrestrial networks," *IEEE Trans. Green Commun. Netw.*, vol. 9, no. 1, pp. 271–290, Mar. 2025.

[41] G. Iacovelli et al., "Holographic MIMO for next generation non-terrestrial networks: Motivation, opportunities, and challenges," 2024, *arXiv:2411.10014*.

[42] W. Ullah Khan et al., "Survey on beyond diagonal RIS enabled 6G wireless networks: Fundamentals, recent advances, and challenges," 2025, *arXiv:2503.08423*.

[43] M. Ahmed et al., "A comprehensive survey on RIS-enhanced physical layer security in UAV-assisted networks," *IEEE Internet Things J.*, vol. 12, no. 16, pp. 32538–32562, Aug. 2025.

[44] A. Nauman et al., "Efficient resource allocation and user association in noma-enabled vehicular-aided hetnets with high altitude platforms," *Comput. Commun.*, vol. 216, pp. 374–386, Jun. 2024.

[45] M. Ahmed et al., "Advancements in RIS-assisted UAV for empowering multiaccess edge computing: A survey," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 6325–6346, Mar. 2025.

[46] Q. Zhang et al., "Two-stage coded-sliding beam training and QoS-constrained sum-rate maximization for SIM-assisted wireless communications," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 12162–12179, 2026.

[47] S. Jeon, S. Kim, G. Im, and Y.-S. Jeon, "Beam-hopping pattern design for grant-free random access in leo satellite communications," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 8246–8259, 2026.

[48] Z. Han, C. Xu, G. Zhao, S. Wang, K. Cheng, and S. Yu, "Time-varying topology model for dynamic routing in LEO satellite constellation networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 3, pp. 3440–3454, Mar. 2023.

[49] N. Zhang, Z. Na, J. Tao, B. Lin, N. Zhang, and K. Zhao, "Time-varying graph and binary tree search based routing algorithm for LEO satellite networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 13683–13688, Oct. 2023.

[50] A. Al-Hourani and B. Al Homssi, "Doppler shift distribution in satellite constellations," *IEEE Commun. Lett.*, vol. 28, no. 9, pp. 2131–2135, Sep. 2024.

[51] B.-H. Yeh, J.-M. Wu, and R. Y. Chang, "Efficient Doppler compensation for LEO satellite downlink OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 73, no. 12, pp. 18863–18877, Dec. 2024.

[52] S. J. Nawaz, E. Cianca, T. Rossi, and M. De Sanctis, "Round trip time (RTT) and Doppler measurements for IoRT localization by a single-satellite," *IEEE Commun. Lett.*, vol. 28, no. 3, pp. 528–532, Mar. 2024.

[53] Y. Gong, D. Yu, X. Cheng, C. Yuen, M. Bennis, and M. Debbah, "Computation offloading and quantization schemes for federated satellite-ground graph networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 14140–14154, Oct. 2024.

[54] H. Al-Hraishawi, H. Chougrani, S. Kisseleff, E. Lagunas, and S. Chatzinotas, "A survey on nongeostationary satellite systems: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 101–132, 1st Quart., 2023.

[55] G. Fontanesi et al., "Artificial intelligence for satellite communication: A survey," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 1381–1435, 2026.

[56] W. U. Khan, A. Mahmood, C. K. Sheemar, E. Lagunas, S. Chatzinotas, and B. Ottersten, "Reconfigurable intelligent surfaces for 6G non-terrestrial networks: Assisting connectivity from the sky," *IEEE Internet Things Mag.*, vol. 7, no. 1, pp. 34–39, Jan. 2024.

[57] Y. Li et al., "RIS-based physical layer security for integrated sensing and communication: A comprehensive survey," *IEEE Internet Things J.*, vol. 12, no. 16, pp. 32444–32468, Aug. 2025.

[58] W. U. Khan, E. Lagunas, A. Mahmood, S. Chatzinotas, and B. Ottersten, "Energy-efficient RIS-enabled NOMA communication for 6G LEO satellite networks," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2023, pp. 1–6.

[59] Y. Xiao et al., "Space-air-ground integrated wireless networks for 6G: Basics, key technologies, and future trends," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3327–3354, Dec. 2024.

[60] R. Zhang et al., "Generative AI for space-air-ground integrated networks," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 10–20, Dec. 2024.

[61] H. Li, K. Ota, and M. Dong, "AI in SAGIN: Building deep learning service-oriented space-air-ground integrated networks," *IEEE Netw.*, vol. 37, no. 2, pp. 154–159, Mar. 2023.

[62] P. Zhang, N. Chen, S. Shen, S. Yu, N. Kumar, and C.-H. Hsu, "AI-enabled space-air-ground integrated networks: Management and optimization," *IEEE Netw.*, vol. 38, no. 2, pp. 186–192, Mar. 2024.

[63] Z. Lin, Z. Feng, K. Guo, A. Nauman, D. Niyato, and J. Wang, "AI-driven seamless and massive access in space-air-ground integrated networks," *IEEE Wireless Commun.*, vol. 32, no. 3, pp. 72–79, Jun. 2025.

[64] H.-W. Lee, A. Medles, C.-C. Chen, and H.-Y. Wei, "Feasibility and opportunities of terrestrial network and non-terrestrial network spectrum sharing," *IEEE Wireless Commun.*, vol. 30, no. 6, pp. 36–42, Dec. 2023.

[65] S. Saleh et al., "Integrated 6G TN and NTN localization: Challenges, opportunities, and advancements," *IEEE Commun. Standards Mag.*, vol. 9, no. 2, pp. 63–71, Jun. 2025.

[66] Y. Yuan, Z. Li, and B. Zhao, "A survey of multimodal learning: Methods, applications, and future," *ACM Comput. Surv.*, vol. 57, no. 7, pp. 1–34, Jul. 2025.

[67] Q. He, Z. Lu, Z. Wang, and H. Hu, "Graph-based progressive fusion network for multi-modality vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12431–12447, Nov. 2023.

[68] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 77–89, 2024.

[69] M. Xu et al., "Cached model-as-a-resource: Provisioning large language model agents for edge intelligence in space-air-ground integrated networks," *IEEE Trans. Netw.*, vol. 34, pp. 2850–2864, 2026.

[70] L. Qian, P. Si, J. Zhao, and K.-Y. Lam, "Parameter training efficiency aware resource allocation for AIGC in space-air-ground integrated networks," *IEEE Trans. Mobile Comput.*, early access, Jan. 12, 2026, doi: 10.1109/TMC.2026.3653031.

[71] J. Du et al., "Anti-jamming resource allocation in air-terrestrial integrated networks: A hierarchical game-theoretic madrl approach," *IEEE Trans. Cognit. Commun. Netw.*, vol. 12, pp. 5380–5392, 2026.

[72] A. Liu, W. Jiang, S. Huang, and Z. Feng, "Multi-modal integrated sensing and communication in Internet of Things with large language models," *IEEE Internet Things Mag.*, vol. 8, no. 5, pp. 103–111, Sep. 2025.

[73] Z. Guo, D. Liu, and Z. Zhang, "Perception enhanced multimodal multitask semantic communication and resource management for UAV-assisted isac systems," *IEEE Trans. Veh. Technol.*, early access, Jan. 5, 2026, doi: 10.1109/TVT.2025.3650639.

[74] H. Que, J. Yang, T. Du, S. Xia, C.-K. Wen, and S. Jin, "Cooperative mapping, localization, and beam management via multi-modal SLAM in ISAC systems," *IEEE Trans. Commun.*, vol. 73, no. 11, pp. 12023–12037, Nov. 2025.

[75] Z. Zhu, J. Yang, C.-K. Wen, S. Xia, and S. Jin, "An overview of multi-modal sensing systems and the potential of ISAC technologies in SHM," *Phys. Commun.*, vol. 69, Apr. 2025, Art. no. 102614.

[76] K. Tan and C. Zhu, "Multimodal sensing for intelligent V2X: A review of recent advances toward deployment," *IEEE Internet Things J.*, vol. 12, no. 22, pp. 46294–46315, Nov. 2025.

[77] N. Chen et al., "Multimodal heterogeneous data sensing and communication integration for CIoT," *IEEE Trans. Consum. Electron.*, vol. 71, no. 3, pp. 7454–7472, Aug. 2025.

[78] X. Luo, Z. Li, M. Chen, R. Yu, S. Mao, and Y. Liu, "Unified packet compression and model adaptation for integrated sensing and multi-modal communications," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 913–926, 2026.

[79] Y. Chen, Q. Wang, S. Wu, Y. Gao, T. Xu, and Y. Hu, "TOMGPT: Reliable text-only training approach for cost-effective multi-modal large language model," *ACM Trans. Knowl. Discovery Data*, vol. 18, no. 7, pp. 1–19, Aug. 2024.

[80] S. Srivastava et al., "Deep learning for multi-modal sensor fusion and CSI compression in vehicular communications," *IEEE Trans. Veh. Technol.*, vol. 74, no. 12, pp. 19436–19451, Dec. 2025.

[81] Y. Yang, H. Du, Z. Xiong, R. Xu, D. Niyato, and Z. Han, "Exploring impacts of age of information on data accuracy for wireless sensing systems: An information entropy perspective," *IEEE Trans. Mobile Comput.*, vol. 24, no. 6, pp. 4907–4924, Jun. 2025.

[82] L. Tang, W. Zhang, Y. Tang, B. Xia, and Q. Chen, "Joint backoff algorithm based on information freshness and NE equilibrium in low-Earth orbit satellite IoT scenarios," *IEEE Internet Things J.*, vol. 12, no. 14, pp. 27533–27546, Jul. 2025.

[83] N. Mu et al., "MoST: Multi-modality scene tokenization for motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14988–14999.
[84] Y. Zhang et al., "Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 12, pp. 13322–13330.
[85] A. Amiri, A. Ghaffarnia, N. G. Nia, D. Wu, and Y. Liang, "Harmonizer: A universal signal tokenization framework for multimodal large language models," *Mathematics*, vol. 13, no. 11, p. 1819, May 2025.
[86] H. Yang, S. Lambotharan, and M. Derakhshani, "FAS-LLM: Large language model-based channel prediction for OTFS-enabled satellite-FAS links," 2025, *arXiv:2505.09751*.
[87] H. Jia, Y. Wang, W. Wu, and J. Yuan, "Robust transmission design for covert satellite communication systems with dual-CSI uncertainty," *IEEE Internet Things J.*, vol. 12, no. 12, pp. 21892–21903, Jun. 2025.
[88] Y. Wang et al., "Statistical CSI-based distributed precoding design for OFDM-cooperative multi-satellite systems," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 3219–3236, 2026.
[89] Z. M. Bakhsh, Y. Omid, G. Chen, F. Kayhan, Y. Ma, and R. Tafazolli, "Multi-satellite MIMO systems for direct satellite-to-device communications: A survey," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 3, pp. 1536–1564, Jun. 2025.
[90] R. Wang, M. A. Kishk, and M.-S. Alouini, "Ultra-dense LEO satellite-based communication systems: A novel modeling technique," *IEEE Commun. Mag.*, vol. 60, no. 4, pp. 25–31, Apr. 2022.
[91] M. Adil, M. De Sanctis, T. Rossi, and E. Cianca, "Phase noise modeling and performance impact analysis in sub-THz bands," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2025, pp. 1–6.
[92] T. Liu et al., "6G integrated sensing and communications channel modeling: Challenges and opportunities," *IEEE Veh. Technol. Mag.*, vol. 19, no. 2, pp. 31–40, Jun. 2024.
[93] R. Campana, C. Amatetti, and A. Vanelli-Coralli, "O-RAN based non-terrestrial networks: Trends and challenges," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Jun. 2023, pp. 264–269.
[94] G. Geraci, D. López-Pérez, M. Benzaghta, and S. Chatzinotas, "Integrating terrestrial and non-terrestrial networks: 3D opportunities and challenges," *IEEE Commun. Mag.*, vol. 61, no. 4, pp. 42–48, Apr. 2023.
[95] M. Mosahebfard, A. Cárdenas, S. Giménez-Antón, and L. Tomaszewski, "Unified framework for dynamic traffic management in integrated 6G terrestrial and non-terrestrial networks," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 9840–9861, 2025.
[96] S. Wang et al., "UAVScenes: A multi-modal dataset for uavs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, May 2025, pp. 28946–28958.
[97] H. Hu, X. Zhu, F. Zhou, W. Wu, R. Q. Hu, and H. Zhu, "Resource allocation for multi-modal semantic communication in UAV collaborative networks," *IEEE Trans. Commun.*, vol. 73, no. 9, pp. 7599–7616, Sep. 2025.
[98] X. Wang et al., "Bridging terrestrial and non-terrestrial networks: A novel architecture for space-air-ground-sea integration system," *IEEE Wireless Commun.*, vol. 32, no. 3, pp. 20–27, Jun. 2025.
[99] J. Deng, S. F. Hasan, H. Zhou, S. Al-Ahmadi, M.-S. Alouini, and D. B. Da Costa, "Ai-native open ran for non-terrestrial networks: An overview," *IEEE Open J. Commun. Soc.*, vol. 7, pp. 1033–1057, 2026.
[100] P. Wu, X. Yuan, Y. Hu, and A. Schmeink, "Age of information for UAV-enabled covert and secure communication," *IEEE Wireless Commun. Lett.*, vol. 15, pp. 1524–1528, 2026.
[101] M. M. Rahman, M. Z. Hassan, J. H. Reed, and L. Liu, "Joint interference management and traffic offloading in integrated terrestrial and non-terrestrial networks," *IEEE Trans. Commun.*, vol. 73, no. 8, pp. 6748–6763, Aug. 2025.
[102] M. Ozturk, M. Salamatmoghadasi, and H. Yanikomeroglu, "Integrating terrestrial and non-terrestrial networks for sustainable 6G operations: A latency-aware multi-tier cell-switching approach," *IEEE Netw.*, vol. 40, no. 2, pp. 156–164, Mar. 2025.
[103] W. Zhao et al., "A survey on DRL-based UAV communications and networking: DRL fundamentals, applications and implementations," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 3911–3941, 2026.
[104] M. A. Khan et al., "Swarm of UAVs for network management in 6G: A technical review," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 1, pp. 741–761, Mar. 2023.
[105] M. Ahmed et al., "Efficient resource management for noma-enabled UAV communications in 6G irs-assisted vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 27, no. 2, pp. 2706–2715, Feb. 2025.
[106] L. Zhang, C. Tian, T. Liu, X. Li, S. Mumtaz, and W. U. Khan, "Multi-objective decomposition evolutionary DRL for UAV-assisted MEC in Internet of Vehicles," *IEEE Trans. Veh. Technol.*, vol. 75, no. 2, pp. 1–16, Feb. 2026.
[107] M. Gusaibat et al., "Effect of phase shift errors on the security of UAV-assisted STAR-RIS IoT networks," *IEEE Internet Things J.*, vol. 13, no. 5, pp. 8725–8734, Mar. 2026.
[108] W. U. Khan et al., "Ris-enabled joint communications and sensing in 6G NTNS: Opportunities and challenges," *IEEE Open J. Commun. Soc.*, vol. 7, pp. 821–843, 2026.
[109] W. U. Khan, E. Lagunas, A. Mahmood, M. Asif, M. Ahmed, and S. Chatzinotas, "Integration of beyond diagonal RIS and UAVs in 6G NTNs: Enhancing aerial connectivity," *IEEE Wireless Commun.*, vol. 32, no. 3, pp. 56–63, Jun. 2025.
[110] W. Wang, S. Fu, L. Yang, C. Li, and Y. Ren, "Age of information optimization for UAV deployment with covert communication: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 74, no. 12, pp. 19772–19776, Dec. 2025.
[111] T. Zhang et al., "Age of information in Internet of Vehicles: A discrete-time multisource queueing model," *IEEE Trans. Commun.*, vol. 73, no. 5, pp. 3298–3317, May 2025.
[112] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Rustenburg, South Africa, Jul. 2020, pp. 1–6.
[113] L. M. Pereira, A. Salazar, and L. Vergara, "On comparing early and late fusion methods," in *Proc. Int. Work-Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2023, pp. 365–378.
[114] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications," *Comput., Mater. Continua*, vol. 80, no. 1, pp. 1–35, 2024.
[115] A. G. Mengara Mengara and Y.-K. Moon, "CAG-MoE: Multimodal emotion recognition with cross-attention gated mixture of experts," *Mathematics*, vol. 13, no. 12, p. 1907, Jun. 2025.
[116] F. Abdullakutty, Y. Akbari, S. Al-Maadeed, R. A. Saady, and H. M. A. Rasheed, "Attention-enhanced mixture of experts for multimodal breast cancer detection," *Biomed. Signal Process. Control*, vol. 118, Jun. 2026, Art. no. 109576.
[117] X. Han, H.-H. Chung, J. Ghosh, P. Pu Liang, and S. Saria, "Massively multimodal foundation models: A framework for capturing interactions with specialized mixture-of-experts," 2025, *arXiv:2509.25678*.
[118] D. Zhang et al., "Mixture of experts in large language models," 2025, *arXiv:2507.11181*.
[119] X. Han, C. Harris, N. Ho, H. Nguyen, and S. Saria, "FuseMoE: Mixture-of-experts transformers for fleximodal fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 67850–67900.
[120] C. K. Sheemar et al., "Joint communications and sensing for 6G satellite networks: Use cases and challenges," 2025, *arXiv:2501.05243*.
[121] A. Vats, R. Raja, V. Jain, and A. Chadha, "The evolution of mixture of experts: A survey from basics to breakthroughs," Tech. Rep., 2024.
[122] H. Wang et al., "CAMEL: Cross-attention enhanced mixture-of-experts and language bias for code-switching speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.
[123] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 280–296.
[124] Y.-Q. Yang et al., "Swin3D: A pretrained transformer backbone for 3D indoor scene understanding," *Comput. Vis. Media*, vol. 11, no. 1, pp. 83–101, Feb. 2025.
[125] X. Wu, Z. Liu, and L. Wang, "Spatio-temporal degradation model with graph neural network and structured state space model for remaining useful life prediction," *Rel. Eng. Syst. Saf.*, vol. 256, Apr. 2025, Art. no. 110770.
[126] B. Lyu et al., "A modular graph-native query optimization framework," 2024, *arXiv:2401.17786*.
[127] C. Kumar Sheemar et al., "Joint communications, sensing, and positioning in 6G multi-functional satellite systems: Survey and open challenges," 2025, *arXiv:2509.25937*.
[128] I. Rasheed, M. Asif, A. Ihsan, W. U. Khan, M. Ahmed, and K. M. Rabie, "LSTM-based distributed conditional generative adversarial network for data-driven 5G-enabled maritime UAV communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 2431–2446, Feb. 2023.

[129] A. Ullah Khan, W. Ullah Khan, H. Jung, and H. Shin, "Resource allocation and sharing for UAV-assisted integrated TN-NTN with multi-connectivity," 2026, *arXiv:2601.15532*.

[130] J.-S. Yang, Z. Zeng, and Z. Shen, "Neural-symbolic dual-indexing architectures for scalable retrieval-augmented generation," *IEEE Access*, vol. 13, pp. 210507–210519, 2025.

[131] Q.-G. Chen et al., "Advancing tool-augmented large language models: Integrating insights from errors in inference trees," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 106555–106581.

[132] Z. Zhao et al., "TURA: Tool-augmented unified retrieval agent for AI search," 2025, *arXiv:2508.04604*.

[133] B. McKinzie et al., "MM1: Methods, analysis and insights from multimodal llm pre-training," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 304–323.

[134] R. Peng, K. Yang, Y. Zeng, J. Lin, D. Liu, and J. Zhao, "DataMan: Data manager for pre-training large language models," 2025, *arXiv:2502.19363*.

[135] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "VILA: On pre-training for visual language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26689–26699.

[136] S. Li et al., "Masked modeling for self-supervised representation learning on vision and beyond," 2023, *arXiv:2401.00897*.

[137] M. Leigh et al., "Is tokenization needed for masked particle modelling?" *Mach. Learn., Sci. Technol.*, vol. 6, Jul. 2025, Art. no. 025075.

[138] S. Ren, F. Wei, S. A. Z. Zhang, and H. Hu, "DeepMIM: Deep supervision for masked image modeling," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 879–888.

[139] W. Ullah Khan et al., "Multilayer non-terrestrial networks with spectrum access aided by beyond-diagonal RIS," 2025, *arXiv:2510.04744*.

[140] Z. Ali et al., "Hybrid optimization for NOMA-based transmissive-RIS mounted UAV networks," *IEEE Trans. Consum. Electron.*, vol. 71, no. 2, pp. 3740–3752, May 2025.

[141] M. Moghaddam, P. D. Giorgio, and M. Papini, "Modeling eroded topography in masked abrasive slurry jet pocket milling," *Int. J. Mech. Sci.*, vol. 285, Jan. 2025, Art. no. 109830.

[142] Y. Li, S. Wu, Y. Zhu, W. Sun, Z. Zhang, and S. Song, "SAMR: Symmetric masked multimodal modeling for general multi-modal 3D motion retrieval," *Displays*, vol. 87, Apr. 2025, Art. no. 102987.

[143] L. Zhang, L. Xu, T. Liu, X. Deng, X. Li, and W. U. Khan, "Exploration enhanced and constraint integrated DRL for RIS-aided UAV network optimization," *IEEE Trans. Veh. Technol.*, pp. 1–13, 2025.

[144] G. Zhang, G. Yuan, D. Cheng, L. Liu, J. Li, and S. Zhang, "Disentangled contrastive learning for fair graph representations," *Neural Netw.*, vol. 181, Jun. 2025, Art. no. 106781.

[145] J. Xiang, X. Zhu, and E. Cambria, "Integrating audio-visual text generation with contrastive learning for enhanced multimodal emotion analysis," *Inf. Fusion*, vol. 127, Mar. 2026, Art. no. 103809.

[146] C. Sun, M. Chen, C. Zhu, S. Zhang, P. Lu, and J. Chen, "Listen with seeing: Cross-modal contrastive learning for audio-visual event localization," *IEEE Trans. Multimedia*, vol. 27, pp. 2650–2665, 2025.

[147] X. Jia, C. Hu, and G. Jia, "Cross-modal contrastive learning for robust visual representation in dynamic environmental conditions," *Academic J. Natural Sci.*, vol. 2, no. 2, pp. 23–34, Apr. 2025.

[148] A. Khan et al., "AI-empowered multi-UAV and IRS collaboration for spectrum and energy optimization in B5G networks," *IEEE Internet Things J.*, vol. 13, no. 5, pp. 7972–7988, Mar. 2026.

[149] M. Kim et al., "Beyond next-token prediction: A performance characterization of diffusion versus autoregressive language models," 2025, *arXiv:2510.04146*.

[150] J. Huang et al., "Opportunistic osteoporosis diagnosis via texture-preserving self-supervision, mixture of experts and multi-task integration," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2025, pp. 428–438.

[151] V. Udandarao et al., "Active data curation effectively distills large-scale multimodal models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 14422–14437.

[152] T.-H. Huang, C. Shin, S. J. Tay, D. Adila, and F. Sala, "Multimodal data curation via object detection and filter ensembles," 2024, *arXiv:2401.12225*.

[153] M. Aghili, S. Tabarestani, and M. Adjouadi, "Addressing the missing data challenge in multi-modal datasets for the diagnosis of Alzheimer's disease," *J. Neurosci. Methods*, vol. 375, Jun. 2022, Art. no. 109582.

[154] J. He, X. Li, X. Zhang, W. Niu, and F. Li, "A synthetic data-assisted satellite terrestrial integrated network intrusion detection framework," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 1739–1754, 2025.

[155] A. Ashurov, K. Baktybekov, A. Zhetpisbayeva, B. Zhumazhanov, and M. Kalel, "Analysis of the functioning quality of a nanosatellite constellation using synthetic data," *EUREKA, Phys. Eng.*, vol. 2025, no. 4, pp. 17–25, 2025.

[156] J. Ning et al., "Ray-tracing channel modeling for leo satellite-to-ground communication systems," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jun. 2024, pp. 1169–1174.

[157] K. Zhang, S. Yang, Y. Wang, J. Huang, and C.-X. Wang, "Ray-tracing based channel modeling and characteristics analysis for LEO satellite-to-ground systems," in *Proc. 18th Eur. Conf. Antennas Propag. (EuCAP)*, Mar. 2024, pp. 1–5.

[158] Y. Zhou, R. Zhang, J. Liu, T. Huang, Q. Tang, and F. R. Yu, "A hierarchical digital twin network for satellite communication networks," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 104–110, Nov. 2023.

[159] H. Hasnain Imtiaz et al., "Digital twin enabled 6G MEC networks: Computational energy efficiency for consumer IIoT in industry 5.0," *IEEE Trans. Consum. Electron.*, vol. 71, no. 3, pp. 7874–7881, Aug. 2025.

[160] B. Mao, X. Zhou, J. Liu, and N. Kato, "Digital twin satellite networks toward 6G: Motivations, challenges, and future perspectives," *IEEE Netw.*, vol. 38, no. 1, pp. 54–60, Jan. 2024.

[161] H. Dong et al., "Advances in multimodal adaptation and generalization: From traditional approaches to foundation models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 5, pp. 1–20, May 2026.

[162] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.

[163] L. Xu, H. Xie, S. J. Qin, X. Tao, and F. L. Wang, "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2026, doi: 10.1109/TPAMI.2026.3657354.

[164] Z. Hu et al., "LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 5254–5276.

[165] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022, vol. 1, no. 2, p. 3.

[166] M. SadraeiJavaeri, E. Asgari, A. C. McHardy, and H. R. Rabiee, "SuperPos-prompt: Enhancing soft prompt tuning of language models with superposition of multi token embeddings," 2024, *arXiv:2406.05279*.

[167] Y. Zhu et al., "Short text classification with soft knowledgeable prompt-tuning," *Expert Syst. Appl.*, vol. 246, Jul. 2024, Art. no. 123248.

[168] W. U. Khan, C. K. Sheemar, E. Lagunas, and S. Chatzinotas, "Beyond diagonal RIS enhanced cognitive radio enabled multilayer non-terrestrial networks," 2025, *arXiv:2503.10866*.

[169] Y. Takama, N. Ding, T. Yokota, and T. Tamaki, "Separating shared and domain-specific LoRas for multi-domain learning," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 6429–6437.

[170] Y. Tian, B. Zhang, Z. Tu, and D. Chu, "Adapters selector: Cross-domains and multi-tasks LoRa modules integration usage method," in *Proc. 31st Int. Conf. Comput. Linguistics*, 2025, pp. 593–605.

[171] X. Yue et al., "Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13834–13844.

[172] W. U. Khan, C. K. Sheemar, E. Lagunas, and S. Chatzinotas, "Enhancing physical layer security in cognitive radio-enabled NTNs with beyond diagonal RIS," in *Proc. IEEE Int. Medit. Conf. Commun. Netw. (MeditCom)*, Jul. 2025, pp. 1–6.

[173] J. Deng, H. Zhou, and M.-S. Alouini, "Distributed coordination for heterogeneous non-terrestrial networks," *IEEE Commun. Mag.*, early access, Jan. 12, 2026, doi: 10.1109/MCOM.001.2500265.

[174] W. U. Khan, A. Mahmood, M. A. Jamshed, E. Lagunas, M. Ahmed, and S. Chatzinotas, "Beyond diagonal RIS for 6G non-terrestrial networks: Potentials and challenges," *IEEE Netw.*, vol. 39, no. 1, pp. 80–89, Jan. 2025.

[175] Z. Chen and B. Liu, "Continual learning and catastrophic forgetting," in *Lifelong Machine Learning*. Cham, Switzerland: Springer, 2022, pp. 55–75.

[176] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, Aug. 2024.

[177] J. Gui et al., "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9052–9071, Dec. 2024.

[178] D. Shome, O. Waqar, and W. U. Khan, "Federated learning and next generation wireless communications: A survey on bidirectional relationship," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 7, p. 4458, Jul. 2022.

[179] X. Pan, L. Ding, B. Zhong, L. Shen, and Y. Wang, "Collaborative governance: Blockchain-based federated learning for construction safety service," *IEEE Trans. Eng. Manag.*, vol. 72, pp. 1163–1177, 2025.

[180] J. Mu, Y. Cui, W. OUyang, Z. Yang, W. Yuan, and X. Jing, "Federated learning in 6G non-terrestrial network for IoT services: From the perspective of perceptive mobile network," *IEEE Netw.*, vol. 38, no. 4, pp. 72–79, Jul. 2024.

[181] G. Yu, Y. He, E. Dutkiewicz, B. Senanayake, and M. Attygalle, "SPARK: A scalable peer-to-peer asynchronous resilient framework for federated learning in non-terrestrial networks," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 1092–1107, 2025.

[182] Y. Gou, X. Dong, Q. Li, S. Gu, R. Hong, and W. Hu, "SURE: Safety understanding and reasoning enhancement for multimodal large language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2025, pp. 7563–7604.

[183] C. A. Alonso, J. Sieber, and M. N. Zeilinger, "State space models as foundation models: A control theoretic overview," in *Proc. Amer. Control Conf. (ACC)*, 2025, pp. 146–153.

[184] J. Cui, J. Gao, X. Liu, Y. Liu, and S. Yu, "A characterization method of terminal ingredients for nonlinear MPC using value-based reinforcement learning," *Automatica*, vol. 183, Jan. 2026, Art. no. 112574.

[185] L. Ning et al., "A survey of webagents: Towards next-generation ai agents for Web automation with large foundation models," in *Proc. 31st ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2025, pp. 6140–6150.

[186] T. Nam, J. Lee, J. Zhang, S. Ju Hwang, J. J. Lim, and K. Pertsch, "LiFT: Unsupervised reinforcement learning with foundation models as teachers," 2023, *arXiv:2312.08958*.

[187] Y. Qin et al., "Tool learning with foundation models," *ACM Comput. Surv.*, vol. 57, no. 4, pp. 1–40, 2024.

[188] F. Gao, X. Wang, Y. Fan, Z. Gao, and R. Zhao, "Constraints driven safe reinforcement learning for autonomous driving decision-making," *IEEE Access*, vol. 12, pp. 128007–128023, 2024.

[189] Y. Lu et al., "Agentic graph neural networks for wireless communications and networking toward edge general intelligence: A survey," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 4519–4554, 2026.

[190] A. Iqbal et al., "Empowering non-terrestrial networks with artificial intelligence: A survey," *IEEE Access*, vol. 11, pp. 100986–101006, 2023.

[191] C. Pan et al., "Vision foundation model guided multimodal fusion network for remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 9409–9431, 2025.

[192] Y. Chen, X. Zeng, H. Cai, Q. Xu, and Z. Liu, "Decentralized QoS-aware model inference using federated split learning for cloud-edge medical detection," *IEEE Trans. Parallel Distrib. Syst.*, vol. 36, no. 10, pp. 2119–2136, Oct. 2025.

[193] T. Gafni, A. Karnieli, and Y. Hanani, "Dual precision quantization for efficient and accurate deep neural networks inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2025, pp. 3259–3269.

[194] S. Fu, V. Narayanan, M. L. Wymore, V. Deep, H. Duwe, and D. Qiao, "No battery, no problem: Challenges and opportunities in batteryless intermittent networks," *J. Commun. Netw.*, vol. 25, no. 6, pp. 806–813, Dec. 2023.

[195] M. Kennedy, A. Ksentini, Y. Hadjadj-Aoul, and G.-M. Muntean, "Adaptive energy optimization in multimedia-centric wireless devices: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 768–786, 2nd Quart., 2012.

[196] H. Zhang, Q. Zhou, H. Xu, J. Su, R. Evans, and K. Gao, "Multimodal classification and out-of-distribution detection for multimodal intent understanding," *IEEE Trans. Multimedia*, vol. 27, pp. 9887–9901, 2025.

[197] V. Nagarajan, A. Andreassen, and B. Neyshabur, "Understanding the failure modes of out-of-distribution generalization," 2020, *arXiv:2010.15775*.

[198] J. Kim and S. Hwang, "Enhanced ood detection through cross-modal alignment of multi-modal representations," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, Apr. 2025, pp. 29979–29988.

[199] N. Ataeian, "Security in space-air-ground integrated networks (SAGIN): A systematic review of vulnerabilities and defense mechanisms," *Intell. Inf. Syst. J.*, vol. 4, no. 9, pp. 25–35, 2025.

[200] U. Venaik, A. Kushwaha, and R. R. Shah, "SASDN: A generalizable and minimal-intervention LLM-integrated framework for continual adaptation in spoofed speech detection," in *Proc. 1st Deepfake Forensics Workshop, Detection, Attribution, Recognit., Adversarial Challenges Era AI-Generated Media*, Oct. 2025, pp. 93–100.

[201] H. Elubeyd, D. Yiltas-Kaplan, and Ã. Bahtiyar, "A multi-modal deep transfer learning framework for attack detection in software-defined networks," *IEEE Access*, vol. 11, pp. 114128–114145, 2023.

[202] Z. Dou, X. Hu, H. Yang, Z. Liu, and M. Fang, "Adversarial attacks to multi-modal models," in *Proc. 1st ACM Workshop Large AI Syst. Models Privacy Saf. Anal.*, Nov. 2023, pp. 35–46.

[203] C. Ren et al., "Advances and open challenges in federated foundation models," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 2087–2126, 2026.

[204] M. Ko, M. Jin, C. Wang, and R. Jia, "Practical membership inference attacks against large-scale multi-modal models: A pilot study," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4871–4881.

[205] X. Li et al., "Physical layer security of cognitive ambient backscatter communications for green Internet-of-Things," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 3, pp. 1066–1076, Sep. 2021.

[206] W. Khalid et al., "Malicious reconfigurable intelligent surfaces: Security threats in 6G networks," *IEEE Internet Things J.*, vol. 12, no. 15, pp. 30063–30085, Aug. 2025.

[207] M. Wu et al., "Split learning with differential privacy for integrated terrestrial and non-terrestrial networks," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 177–184, Jun. 2024.

[208] B. Cordill, D. Fang, and S. Xu, "A comprehensive survey of security and privacy in UAV systems," *IEEE Access*, vol. 13, pp. 117843–117866, 2025.

[209] R. Koohestani, P. D. Bekker, B. Koç, and M. Izadi, "Benchmarking AI models in software engineering: A review, search tool, and unified approach for elevating benchmark quality," *IEEE Trans. Softw. Eng.*, vol. 52, no. 2, pp. 651–674, Feb. 2026.

[210] A. Hardy, M. Hardy, M. Kochenderfer, M. Lamparth, A. Reuel, and C. Smith, "BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 21763–21813.

[211] Y. Li and J. Zhan, "SAIBench: Benchmarking AI for science," *BenchCouncil Trans. Benchmarks, Standards Evaluations*, vol. 2, no. 2, Apr. 2022, Art. no. 100063.