Towards a Semantic and Goal-Oriented O-RAN Architecture: the 6G-GOALS Approach

L. Zanzi*, D. Montagno B.†,, V. Sciancalepore*, P. Li‡, A. Aijaz‡, P. A. Stavrou§,
M. Kountouris§, O. Forceville¶, X. Li*, M. Hua‖, D. Gündüz‖, Z. Chen**,
J. H. Inacio de Souza††, F. Costanzo‡‡[xi], E. C. Strinati‡‡, S. Fiorellino[x], P. Di Lorenzo[x], R. Fantini [xii]
*NEC Laboratories Europe, Germany. †Hewlett Packard Italy, Italy. ‡Toshiba Europe Ltd., U.K.
§EURECOM, France. ¶Hewlett Packard France, France. ‖Imperial College London, United Kingdom.
**Singapore University of Technology and Design, Singapore. ††Aalborg University, Denmark. ‡‡CEA Leti, France.
[x]National Inter-University Consortium for Telecommunications (CNIT), Italy. [xi]Univ. Grenoble Alpes, France. [xii]TIM, Italy.

***Abstract*—The semantic and goal-oriented communication paradigm is a fundamental shift in the design of next-generation 6G networks, aiming to support an increasingly connected, intelligent, and sustainable digital ecosystem. This paper provides a comprehensive overview of the architecture and operational framework developed within the 6G-GOALS project, with particular emphasis on its core design pillars: ultra-low latency, semantic communications, AI-native integration, energy efficiency, and enhanced network resilience. The paper details the novel architectural components of the 6G-GOALS framework, which builds on and extends the O-RAN architecture by integrating semantic-aware entities and protocols. Key enablers, including distributed AI, real-time, goal-driven decision-making, and adaptive orchestration of network functions, are presented, illustrating how these capabilities work in concert to realize fully semantic-aware, intelligent, and self-adaptive network operations capable of meeting the demands of next-generation connectivity. Finally, we provide an evaluation of the architecture's potential to meet key performance indicators (KPIs), its alignment with sustainability goals, and its readiness for the evolving digital ecosystem. This analysis is intended to serve as a foundational reference for researchers and industry stakeholders working to advance the 6G vision.**

I. Introduction

The rapid evolution of wireless communication networks, coupled with the increasing demands for data-driven applications, necessitates innovative approaches to enhance network efficiency and adaptability. In this context, semantic communication (SemCom) has emerged as a transformative paradigm, prioritizing the transmission of meaningful and significant task-relevant information over traditional bit-level accuracy. By focusing on the utility of the conveyed data rather than its exact replication, SemCom holds the potential to optimize bandwidth utilization, reduce latency, and improve overall network performance. Simultaneously, the adoption of Open Radio Access Network (O-RAN) principles is reshaping the telecommunication landscape by promoting openness, interoperability, and programmability in network architectures. O-RAN introduces the concept of a RAN Intelligent Controller (RIC), a key enabler for intelligent and flexible network management. The RIC is divided into two layers: the Non-Real-Time RIC (Non-RT RIC) and the Near-Real-Time RIC (Near-RT RIC), each with distinct roles in handling different operational timescales and functionalities. This paper examines the architectural convergence of SemCom and O-RAN, with a particular focus on integrating SemCom principles into the RICs, interfaces, operations, and functionalities. To this end, we introduce the concept of the Semantic RICs, an enhanced control entity that integrates semantic awareness into both the Near-RT and Non-RT RICs. By leveraging extended capabilities supported by enhanced interfaces and data models, the 6G-GOALS architecture aims to enable dynamic network adaptation, more efficient resource allocation, and intelligent support for advanced use cases, including autonomous networking, context-aware optimization, and semantics-driven decision-making.

II. Related Works

SemCom has emerged as a transformative paradigm that shifts the focus of communication systems from bit-level accuracy to the preservation and transmission of meaning. Learning-based semantic encoders and decoders have been proposed for text, vision, and multimodal applications, demonstrating significant gains in bandwidth efficiency and robustness compared to traditional source–channel coding. Survey papers such as [1] and [2] summarize these developments, highlighting benefits for low-resource and noisy environments, as well as challenges in measuring semantic loss, defining task-oriented metrics, and enabling semantic-aware resource allocation. Beyond point-to-point communication, several works have examined the integration of SemCom within networked systems. Studies such as [3] and [4] explore task-oriented and multi-user semantic transmission, including semantic-aware scheduling and joint semantic–radio optimization. These contributions collectively demonstrate that End-to-End (E2E) SemCom pipelines must consider both semantic accuracy and network-level constraints, motivating system-level research into SemCom-native networks. Despite these advances, the integration of SemCom into O-RAN remains largely unexplored. Existing O-RAN research primarily targets radio-, mobility, and QoS-centric optimization, whereas SemCom introduces fundamentally new requirements involving semantic importance, task relevance, and meaning-preservation metrics [5]. Current SemCom studies generally assume custom

or monolithic architectures without considering RAN disaggregation, multi-vendor components, or standardized control interfaces. Conversely, O-RAN literature rarely considers semantics-aware Key-Performance Indicators (KPIs), semantic feedback loops, or the impacts of meaning-driven transmission on existing control loops. Recent preliminary works have begun bridging this gap by discussing the feasibility of embedding SemCom within programmable and AI-native RANs [6]. These works suggest that the RIC's modular, software-defined structure, together with Multi-Access Edge Computing (MEC) integration, provides a promising platform for deploying semantic-aware control policies, semantic inference services, and cross-layer optimization loops.

III. Motivation and Challenges

One of the key aspects about semantic communication is the need to incorporate semantic information from user equipment (UE) and its applications—along with their requirements and feedback—into the O-RAN RIC–enabled decision-making process. This allows the network to dynamically allocate resources based on the real-time importance of semantic context, leading to more efficient and tailored communication. In cloud-native O-RAN systems, where RAN functions and AI services run on virtualized O-Cloud infrastructure, performance depends on both radio and computing resources. Exposing computing context to control entities can enhance coordination, particularly for latency-sensitive semantic communication and edge intelligence. However, current mobile architectures lack mechanisms to access and process semantic information from UEs and applications. Addressing this requires new functional components, enhanced interfaces, and semantic-aware decision-making—key challenges targeted by the 6G-GOALS project.

A. Compatibility with O-RAN

The current O-RAN architecture lacks system elements and interfaces to enable direct communication between the RICs, UEs and their hosted applications. Specifically, O-RAN lacks defined interfaces, procedures, and system entities: (i) to collect contextual information, service requirements, KPIs, and feedback from UEs and applications (ii) to enforce policy or configurations back to the UEs and applications to enable E2E, closed-loop optimization.

This limitation reduces the network's ability to efficiently optimize the RAN resources, latency, and data transmission. Most importantly for SemCom, it hinders adaptive configuration of UEs and applications, which is essential for context-aware and AI-driven services such as IoT and autonomous systems. Furthermore, the absence of standardized frameworks and procedures for SemCom in O-RAN limits interoperability and the seamless integration of advanced technologies. Addressing this gap is crucial for advancing O-RAN's role in enabling intelligent, efficient, sustainable, and scalable communication systems. Therefore, in this paper, we propose system design guidelines and methodologies to enable UE-O-RAN interaction for improved E2E communication. To this end, we extend the current O-RAN framework by introducing and defining the required interfaces and data models, procedures, and key enhancements to the existing O-RAN architecture.

B. Data Representation and Standardization

SemCom is heavily dependent on shared understanding; both the transmitter and receiver must interpret messages using contextual information established via shared knowledge-models. This necessitates common data representation formats such as knowledge graphs or pre-trained language models. The lack of standardized methods for semantic content transmission poses a major issue. Each end-user application might use different encodings, or semantic abstractions, making it difficult to achieve seamless communication between components. In this context, a key role of the O-RAN system is to enable the setup of semantic-aware communication channels, with dedicated methods for semantic data exchange and mechanisms to align incompatible data representations. Without a widely accepted framework within O-RAN, achieving the architecture's openness and interoperability goals becomes significantly more complex.

C. AI/ML Integration

SemCom relies on AI/ML models trained to extract meaningful representations of data. O-RAN supports AI integration through its RIC architecture, particularly the Non-RT RIC and Near-RT RIC. Advanced AI/ML rApps hosted by the Non-RT RIC control tasks whose execution takes more than 1 s, such as training and updating universal semantic encoders based on global network policies and long-term user behavior patterns. Conversely, the Near-RT RIC deploys AI-driven xApps to perform semantic-aware operations that run on a shorter timescale. More recently, the O-RAN Alliance has introduced dApps, which enable real-time applications to run closer to the Central Units (CUs) and Distributed Units (DUs), thereby supporting low-latency semantic-aware inference and control for future intelligent RAN systems.

However, integrating semantic models adds layers of complexity beyond current implementations. These models require extensive training, validation, and continuous lifecycle management through the O-RAN Service Management and Orchestration (SMO) framework, including versioning, model drift detection, and policy-based control. In the proposed architecture, semantic encoders and related AI models are associated with model descriptors that include version information, semantic task type, and representation characteristics. When semantic communication sessions are established, model descriptors can be exchanged between participating entities to verify compatibility. If heterogeneous models are detected, the Near-RT RIC can coordinate semantic alignment mechanisms to map the latent representations between transmitter and receiver models. In particular, the significance of data can change dynamically with real-world context, requiring AI/ML models to be continuously validated and retrained. Such dependency on live data opens the door to new and underexplored security vulnerabilities, and model

integration must be carefully executed to avoid interfering with other real-time operations. Ensuring that such AI-based models remain robust, up-to-date, and contextually accurate across heterogeneous network environments remains a non-trivial task.

D. Feedback Loops and Learning Overhead

The SemCom paradigm builds on timely and actionable feedback. Models must learn continuously from context, user feedback, or environmental changes to remain relevant and effective. In O-RAN, such feedback loops must be tightly integrated into the Non-RT RIC, the Near-RT RIC, and coordinated across the management and control plane. However, incorporating semantic feedback involves substantial overhead, from collecting feedback data, training or fine-tuning models, distributing updates, and ensuring cross-node consistency. This level of dynamic adaptation adds complexity to the O-RAN system, especially in multi-vendor and multi-operator environments.

IV. Semantic-Aware O-RAN: The 6G-GOALS Architecture

SemCom focuses on transmitting only meaningful, context-relevant information from specific application clients hosted by the UEs to their corresponding application servers. This paradigm is particularly beneficial for applications such as autonomous driving, traffic management, collision avoidance, monitoring, and environmental surveillance, where real-time, intelligent communication is critical. To support these and other use cases in an extensible, scalable, and efficient platform for SemCom, we present the 6G-GOALS O-RAN enhanced architecture, illustrated in Fig. 1. It builds on the existing O-RAN architecture, improved with semantic-aware entities, interfaces, and protocols to realize efficient 6G networks based on the semantic and goal-oriented communication paradigm.

To facilitate E2E SemCom, we introduce a Semantic Module and a Semantic Extractor as software instances running at the UE and at the edge or cloud (e.g., as MEC applications). The Semantic Module performs context processing and compression, ensuring that only relevant and meaningful data is exchanged with the network. The Semantic Extractor, embedded at the UE or MEC application layer, analyzes raw data streams using AI-based techniques (e.g., Generative AI) to extract high-level representations. Instead of transmitting redundant bit-level data, only semantic elements conveying the core meaning are transmitted, reducing bandwidth usage while preserving service intent and quality.

SemCom fundamentally alters traditional communication principles and can be applied across all layers of the communication stack, from content generation at the application layer to semantic-aware scheduling and resource allocation at Medium Access Control (MAC) sublayer, down to the physical layer, focusing on signal processing and channel encoding/decoding of semantics. For example, conventional error control mechanisms such as packet retransmission in Hybrid Automatic Repeat reQuest (HARQ) or at the Radio Link Control (RLC) level blindly resend entire packets without considering whether the missing data affects the conveyed meaning, leading to inefficiencies and increased latency. In contrast, SemCom enables retransmission decisions based on the importance and semantic relevance of the information, considering semantic-related requirements, KPIs, and configurations of the UE or application. Within this perspective, the proposed framework aligns with the emerging AI-RAN paradigm. Intelligent resource allocation through a Semantic RIC reflects "AI for RAN," while edge-based semantic inference corresponds to "AI on RAN."

The SMO and Non-RT RIC can interact with external services, such as MEC or cloud platforms, via standardized information exposure mechanisms [7], accessed by rApps through the R1 interface [8]. As O-RAN does not define an interface between Near-RT RIC and UEs, we envision that UE-related SemCom data can be collected and processed by edge applications through an already established data plane channel (Steps 1 and 2 in Fig. 1). The Y1 open interface connects edge/MEC and Near-RT RIC, allowing the near-RT RIC to expose Radio Analytics Information (RAI) to authorized internal and external consumers. While Y1 supports secure and scalable analytics exposure, it currently enables only unidirectional information flow. To effectively support SemCom, UE- and/or application-related semantic information should be exposed to the Near-RT RIC. This requires enhancing the Y1 interface to support bidirectional communication, and an extended data model capable of carrying semantic-related configurations, KPIs, and feedback from UE and MEC app (Step 3), as well as semantic-aware policies enforcement (Step 4&5). Existing Y1 security mechanisms like authentication, authorization, confidentiality, and integrity—are reused, with additional SemCom-specific security and privacy considerations discussed in Section IV-B. These enhancements at Y1 interface enable near real-time RAN adaptation, tighter MEC integration, and AI/ML-driven inference at the network edge through continuous exchange of semantic-rich information. The proposed architecture extends O-RAN principles while remaining aligned with existing standardized interfaces (A1, E2, O1/O2). Semantic and goal-oriented capabilities are introduced as non-disruptive extensions, ensuring backward compatibility: legacy RAN nodes and UEs continue to operate as usual, while semantic-aware components can be incrementally deployed to support advanced control and goal-driven optimization.

A. Semantic RIC Interfaces

The integration of the semantic and goal-oriented communication paradigm, as defined in the 6G-GOALS project, into the O-RAN architecture requires a detailed analysis of existing network elements and interfaces. This section evaluates how current O-RAN interfaces can be leveraged, extended, or complemented to support semantic-aware functionalities, especially in light of the architectural components introduced in Section IV.

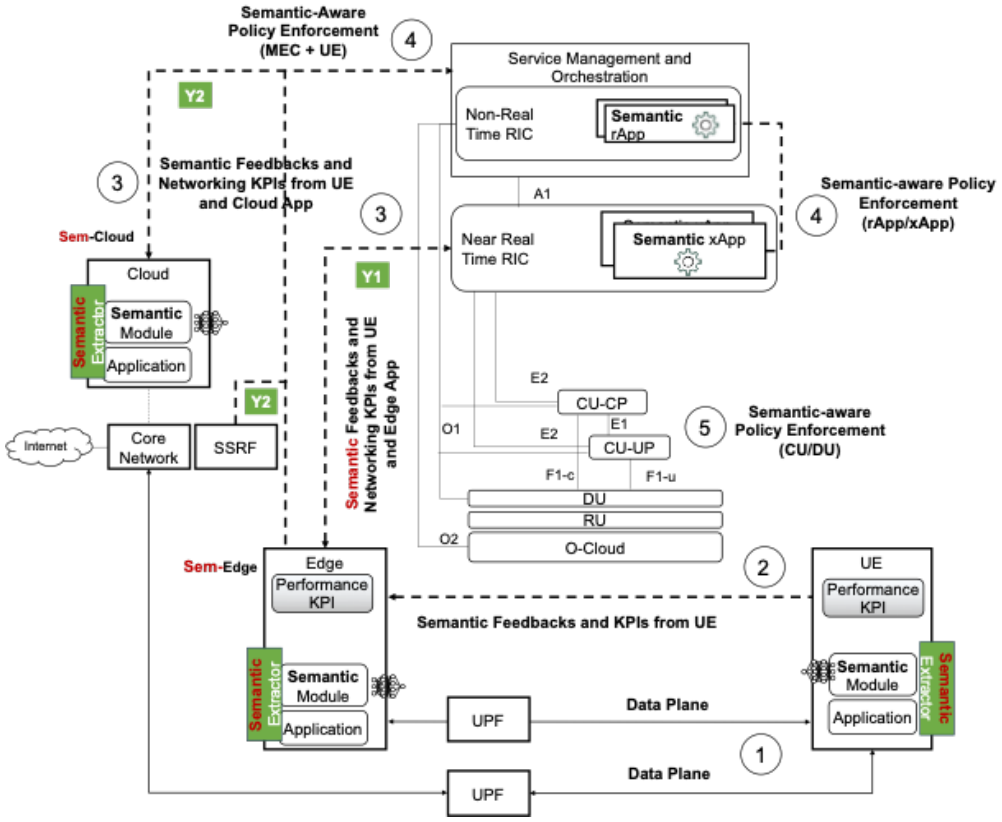


Fig. 1: 6G-GOALS O-RAN Enhanced Architecture to support SemCom

1) A1 Interface (Non-RT RIC → Near-RT RIC): The A1 interface enables the Non-RT RIC to transmit policies, enrichment data, and manage AI/ML model lifecycles within the Near-RT RIC. To support 6G-GOALS, the Policy Management Service can be extended with semantic-aware policies encoding high-level goals (e.g., minimizing semantic distortion, maximizing goal completion), and described via semantic descriptors. Similarly, the A1 Enrichment Information Service can provide additional context to the Near-RT RIC, such as topological representations and semantic embeddings. The A1-ML service enables deployment of semantic models (e.g., goal classifiers), supporting coherent policy enforcement and federated learning across RIC layers.

2) E2 Interface (Near-RT RIC → E2 Nodes): The E2 interface enables near-real-time monitoring and control of RAN nodes via the E2 Application Protocol (E2AP) and pluggable E2 Service Models (E2SMs). The integration of semantic goals into the E2 interface primarily hinges on the definition and usage of appropriate E2SMs. Extending the E2SM-KPM or creating new ones to incorporate semantic KPIs. Beyond monitoring, E2AP supports control actions from the Near-RT RIC to the RAN. In this context, semantic tags in control messages can influence scheduling decisions, while feedback from semantic-aware nodes through E2 Indication messages enables dynamic xApp adaptation based on semantic priorities. The E2 interface provides the flexibility and responsiveness needed for semantic-aware operation, provided that E2SMs are extended to encode semantic abstractions alongside traditional physical layer metrics.

3) O1 Interface (SMO → Network Functions): Used for management of O-RAN network functions via NETCONF and YANG, the O1 interface can be extended to configure semantic processing entities (e.g., setting relevance thresholds, selecting models) and gather semantic performance metrics. Though not real-time, O1 should support long-term monitoring and orchestration aligned with semantic optimization goals.

4) O2 Interface (SMO → O-Cloud): The O2 interface connects the SMO to the O-Cloud infrastructure enabling management of cloud resources and workloads for components such as O-RAN functions, rApps, and third-party applications. In a semantic-aware context, O2 supports deployment, scaling, and lifecycle management of semantic rApps and engines (e.g., transformer models, relevance estimators, classifiers), enabling flexible edge-cloud orchestration and goal-aware operation while reducing semantic distortion.

5) Extending Y1 Interface (Near-RT RIC → External Consumers): The Y1 interface exposes RAI produced by the Near-RT RIC to authorized external entities, such as network analytics platforms, external controllers, MEC applications, or cloud-based service orchestrators. It can be adapted to publish semantic insights (e.g., contextual relevance scores for radio flows) and, if extended to allow bidirectional support as previously proposed, to receive feedback or refined intents from external platforms. This would enable coordination with MEC

TABLE I: Mapping between the interfaces discussed in the 6G-GOALS architecture and their relationship with the baseline O-RAN architecture. Proposed interfaces without current normative O-RAN support are explicitly marked as new.

Interface	Status w.r.t. O-RAN	Role in O-RAN	6G-GOALS Perspective
R1	Existing	Interface between rApps and the SMO / Non-RT RIC framework	Reused for rApp access to exposed data and services
A1	Existing (extended)	Non-RT RIC to Near-RT RIC policy, enrichment, and AI/ML-related control	Extended with semantic-aware policies, descriptors, and intent abstractions
E2	Existing (extended)	Near-RT RIC control and monitoring of E2 nodes	Extended via semantic-aware E2 service models, semantic KPIs, and control actions
O1	Existing (extended)	Management, configuration, monitoring, and lifecycle from SMO to O-RAN managed elements	Extended for semantic model configuration, monitoring, and long-term semantic performance management
O2	Existing (extended)	SMO to O-Cloud orchestration and resource management	Reused / extended for deployment and lifecycle management of semantic functions and AI workloads
Y1	Existing (extended)	Near-RT RIC exposure of RAN analytics information to authorized consumers	Extended from unidirectional analytics exposure to bidirectional semantic information exchange
Y2	New	Not defined in O-RAN	New interface for Non-RT semantic / policy / intent exchange with external entities
O3	New	Not defined in O-RAN	New interface for low-latency O-Cloud to Near-RT RIC information exchange

and cloud-native services for semantic-aware RAN adaptation.

6) Novel Y2 Interface (Non-RT RIC → External Consumers): O-RAN is investigating terminations to enable the SMO/Non-RT RIC framework to exchange messages with external entities. Building on this, the 6G-GOALS consortium introduces the Y2 interface for exposing RAN Intelligence and Policy (RIP) information produced by the Non-RT RIC to authorized external entities, such as AI-based service orchestration platforms, OSS/BSS systems, or cross-domain policy frameworks. Y2 carries long-term statistics, policy decisions, historical trends, and ML model results derived from non-real-time analysis. Future bidirectional enhancements may enable external consumers to provide semantic intents, service goals, and user experience feedback, supporting intent-driven optimization and improved policy generation (e.g., A1/O1/O2), enabling coordinated decision-making across domains.

7) Novel O3 Interface (O-Cloud → Near-RT RIC): Notably, while UE and application-related KPIs provide crucial insights for RAN optimization and service assurance, the growing complexity of disaggregated RAN deployments in cloud-native environments demands a broader perspective—one that includes processing and scheduling information from the RAN O-Cloud jointly with upper-layer information. Currently, O-Cloud exposes information only through the O2 interface towards the SMO. This procedure does not meet the low-latency requirements of SemCom. Therefore, this necessitates the introduction of a novel O3 interface, which facilitates real-time data exchange between the Near-RT RIC and the RAN O-Cloud. The O3 interface enables xApps to collect and process UE-specific L1/L2 information from the O-DU/RU from Low MAC and PHY protocol layers, allowing xApps to consider computing-related conditions when optimizing RAN decisions. With semantic extraction and intent-aware processing running at the UE and MEC level, Near-RT RIC must align radio decisions with computing availability in the O-Cloud. The O3 interface ensures that xApps do not make RAN decisions in isolation but rather jointly optimize communication and computation, driving AI-native, semantic-aware O-RAN intelligence. Importantly, the proposed O3 interface does not transfer cloud orchestration to the Near-RT RIC. Instead, it provides lightweight exposure of O-Cloud conditions (e.g., compute availability and processing status), enabling xApps to align RAN decisions with the execution of semantic processing functions. The semantic extensions remain fully aligned with O-RAN Alliance interfaces and models. The A1 interface supports semantic and goal-oriented policies through new JSON-based policy types, enabling intent exchange between the Non-RT RIC and Near-RT RIC. At the Near-RT level, the E2 interface enables semantic monitoring and control via extensible mechanisms, such as new service models (e.g., E2SM-Semantic) or extensions of existing ones like E2SM-KPM using ASN.1 encoding. The O1 interface similarly supports configuration and lifecycle management of semantic-aware functions through YANG model extensions via NETCONF, including AI/ML parameters and performance metrics. This design supports phased deployment: semantic-aware rApps and xApps can initially coexist with legacy applications, while advanced capabilities—enabled by interfaces such as Y1 and O3—can be gradually introduced to achieve tighter coordination across RAN, edge, and cloud domains without disrupting existing operations.

B. Security and Privacy Considerations

The introduction and extension of O-RAN modules and interfaces to support intent-aware control loops and multi-domain coordination for SemCom inevitably enlarge the attack surface of future 6G networks. In particular, the integration of semantic models within the RICs and across O-RAN interfaces (e.g., E2, O1, O2, and Y1) introduces new vulnerabilities at both the control and data planes. Adversaries may attempt model poisoning, semantic manipulation, or model inversion attacks to infer sensitive user behavior or to bias network decisions. In this regard, protecting semantic models requires securing its full life cycle, including authenticated distribution,

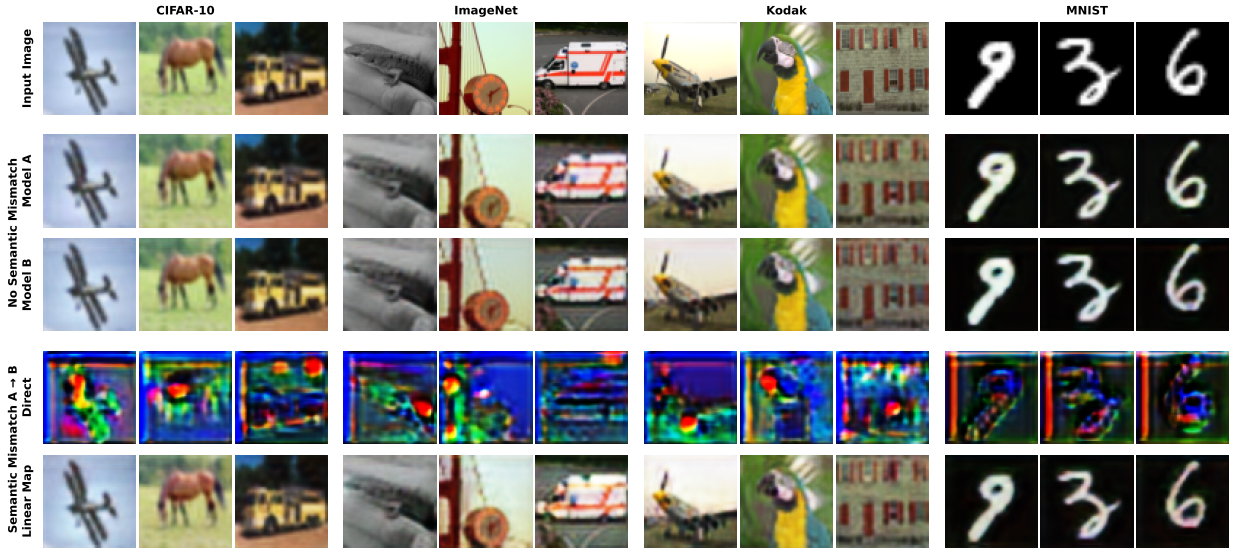


Fig. 2: Illustration of semantic channel mismatch and semantic channel equalization across multiple datasets (CIFAR-10, ImageNet, Kodak, MNIST). The first row shows the input images, followed by matched reconstructions from Model A and Model B. The Direct A→B row demonstrates the severe degradation caused by latent-space incompatibility between independently trained models. Applying a lightweight semantic equalization step, via a linear map, restores semantic coherence and enables successful cross-model reconstruction.

integrity verification via cryptographic hashes, and controlled update mechanisms. Furthermore, while authentication and end-to-end encryption are essential for protecting semantic data exchange, additional safeguards such as trusted execution environments and runtime attestation can further limit model and data tampering. Finally, privacy-preserving mechanisms in the collection and aggregation of semantic information, such as federated learning, homomorphic encryption, and differential privacy, can reduce information leakage risks. Developing a SemCom O-RAN architecture compliant with privacy and security standards must rely on coordinated deployment of security mechanisms across the Near-RT and Non-RT RICs, as well as the newly introduced Semantic Module and Semantic Extractor.

V. Evaluation

A. Semantic Alignment Problem

Semantic communication requires transmitter and receiver to operate over compatible latent spaces, i.e., learned representations of underlying concepts [9]. In distributed architectures such as O-RAN, this aspect becomes particularly critical, as independently developed models deployed across different vendors result in heterogeneous and potentially inconsistent representations. In practice, latent spaces may diverge due to model drift, version mismatches, or independent training, causing semantic misalignment and loss of intelligibility [10]. Ensuring semantic coherence therefore requires explicit alignment mechanisms prior to communication. While not a full system-level validation, this analysis focuses on semantic interoperability as a key enabler in multi-vendor O-RAN systems where multi-vendor deployments and independently trained models are expected to coexist, making semantic alignment a necessary condition for meaningful communication. In the 6G-GOALS architecture, semantic alignment is handled at the Near-RT RIC via dedicated xApps. These xApps detect model incompatibilities, compute alignment transformations across latent spaces, and distribute mapping parameters to network entities. Their functions include semantic feature management, latent-space alignment, and context-driven model updates. The workflow ensures compatibility before transmission: the UE signals its semantic intent and model identity via RRC; the Near-RT RIC xApp retrieves the receiver profile, evaluates compatibility, and computes an alignment if needed. The mapping is then applied at the UE prior to transmission, with optional receiver feedback for refinement. Persistent misalignment is addressed by the Non-RT RIC through long-term model and policy updates.

Fig. 2 illustrates a representative example of semantic channel mismatch and its mitigation. We consider two independently trained convolutional autoencoders, denoted Model A and Model B, trained on CIFAR-10 [11] using the standard 45k/5k/10k split for training, validation, and test. Due to differences in architectural choices, the two models converge to heterogeneous latent representations: Model A produces a latent representation of size 3,584, whereas Model B operates on a 4,096-dimensional latent space. This mismatch gives rise to a semantic channel incompatibility: the latent code generated by the encoder of Model A cannot be directly interpreted by the decoder of Model B. To visualize this effect, we evaluate cross-model reconstructions on multiple datasets—CIFAR-10, ImageNet, Kodak, and MNIST. The first row of Fig. 2 depicts the input images, followed by two rows

Method	Tx coeff.	PSNR (dB) ↑	ℓ_2 ↓	MRR ↑	CKA ↑	Jaccard ↑
Direct	**4096**	8.31	75.20	0.0789	0.9099	0.1956
	16	9.63	62.60	0.1343	0.9452	0.1507
	32	11.11	62.25	0.3150	0.9647	0.2379
	64	11.54	61.95	0.6164	0.9741	0.3102
	128	12.37	61.52	0.8809	0.9809	0.3882
Linear Map	256	12.78	60.69	0.9923	0.9874	0.4890
	512	11.55	57.96	0.9996	0.9906	0.5621
	1024	10.84	47.29	1.0000	0.9918	0.5887
	2048	19.17	13.65	1.0000	0.9920	0.5983
	4096	31.50	3.27	1.0000	0.9920	0.5984

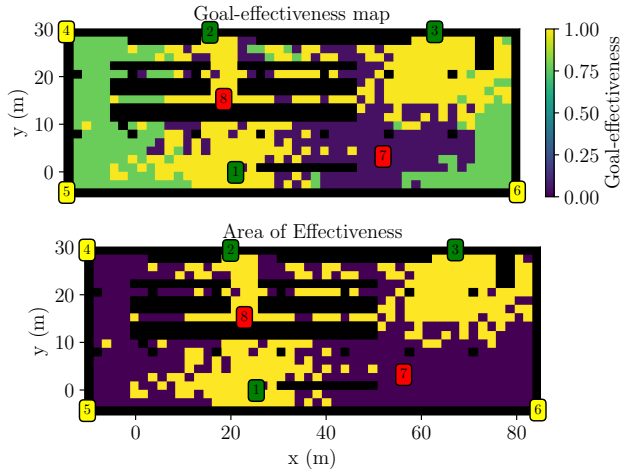
TABLE II: Semantic channel equalization results for direct latent transfer and SVD-compressed linear mapping (Encoder A $\rightarrow$ Decoder B). All metrics, except PSNR, are computed in the receiver's latent space, either directly or after semantic alignment and rank-k compression. PSNR is evaluated on the reconstructed signals after decoding. The compression level refers to the transmitted rank-k coefficient vector, consisting of k scalar coefficients per sample.

displaying reconstructions obtained under matched conditions where encoder and decoder implement the same architecture as performance benchmarks. The subsequent rows illustrate the mismatched case, where the encoder of Model A is paired with the decoder of Model B. In the Direct setting, no alignment is applied. Due to differing latent dimensionalities, vectors from Model A are zero-padded to match Model B. The resulting reconstructions are severely degraded, as Decoder B interprets the input using an incompatible semantic representation. To restore compatibility, a linear transformation is learned from paired latent samples using the training set of CIFAR-10 by solving a least-squares regression problem mapping the latent space of the two models. After applying such semantic alignment, cross-model reconstructions improve substantially, as shown in the row labeled Linear Map. Despite its simplicity, this approach effectively recovers semantic coherence across heterogeneous models. From an O-RAN perspective, this highlights the need for alignment mechanisms in multi-vendor environments. Such operations can be orchestrated by Near-RT RIC xApps to ensure interoperability across heterogeneous implementations.

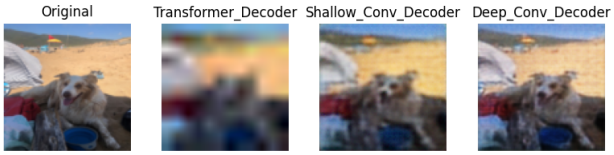
Table II summarizes the reconstruction quality and latent-space alignment obtained through linear latent mapping as a function of transmitted coefficients. The mapping compression is achieved via truncated singular value decomposition (SVD), which allows the semantic-RIC to regulate the amount of semantic information conveyed over the air interface by controlling the number of transmitted coefficients. Performance is evaluated through PSNR, reflecting the ability of a remote semantic decoder within the O-RAN architecture to reconstruct task-relevant information. Latent-space alignment is further characterized using ℓ_2 distance between latent representations, as well as by retrieval- and topology-based measures. In particular, MRR evaluates semantic consistency in a kNN-based retrieval setting by querying the receiver latent space with the mapped source latent representations [12], Gram-based CKA measures global similarity between latent representations, and the Jaccard index quantifies neighborhood preservation across latent spaces [13]. The results show a clear monotonic improvement in both reconstruction fidelity and semantic alignment as the number of transmitted coefficients increases, indicating that even highly compressed linear mappings can preserve meaningful semantic structure. These metrics can be interpreted as semantic KPIs, which could be leveraged by RIC xApps to drive adaptive network decisions, such as retransmission policies, scheduling priorities, or semantic model selection, in semantic-aware communication scenarios. This behavior supports the feasibility of semantic interoperability mechanisms that could be orchestrated by the Semantic-RIC in O-RAN systems, where bandwidth-efficient transmission of semantic representations is critical to enable scalable, interoperable, and low-latency coordination across disaggregated radio access network components.

B. Network Coverage Problem

The semantic KPIs introduced previously focus on application-layer performance (e.g., reconstruction fidelity). However, SemCom systems inherently couple application and network domains [14], requiring joint evaluation metrics. To this end, we adopt *goal-effectiveness*, defined as the probability of achieving a target task under given communication, computation, and AI resource constraints. Unlike conventional KPIs, it explicitly captures end-to-end latency, throughput, and resource limitations jointly with semantic fidelity. For instance, high PSNR does not guarantee task success if latency constraints are violated due to limited bandwidth. Building on this, the *Area of Effectiveness* (AoE), introduced in [15], defines the spatial region where goal-effectiveness exceeds a predefined threshold. Although evaluated at the application layer, AoE inherently depends on cross-layer factors, including radio conditions, network performance, computational capacity, and model alignment. As such, AoE generalizes the notion of coverage by incorporating task success under realistic system constraints. In the proposed O-RAN architecture, AoE evaluation requires joint monitoring of network and application metrics. The Non-RT RIC rApp exploits long-term statistics (e.g., latency distributions, success probability) to optimize policies maximizing AoE, while the Near-RT RIC xApp performs real-time adaptation of system parameters to sustain goal-effectiveness under dynamic conditions, balancing radio quality, computational resources, and semantic alignment [15]. To quantify the interplay between communication and computation, we extend the evaluation by incorporating latency and resource constraints in an edge-based image reconstruction task with semantic misalignment. Goal-effectiveness is defined as the probability of achieving a PSNR target within a latency budget, while user association is based solely on signal strength to expose its limitations. Fig.3a shows that application-level performance deviates significantly from radio-only coverage, particularly near APs, due to heterogeneous model capabilities and alignment. Fig.3b further illustrates reconstruction differences across deployed models. While more complex models improve reconstruction quality, they introduce higher processing delays, reducing goal-effectiveness under strict latency constraints. These



(a) Goal-effectiveness and AoE map for an image reconstruction task at the wireless edge.



(b) reconstruction example with the three deployed models (blue, green and yellow in the previous figure).

Fig. 3: Goal-effectiveness, AoE and image reconstruction.

results demonstrate that communication- and computation-aware policies significantly outperform radio-based strategies, highlighting the need for cross-layer KPIs combining network and semantic metrics.

VI. Conclusions and Future Work

The transition to semantic communication marks a shift from transmitting raw data to conveying meaningful, context-rich information tailored to the goals of emerging 6G applications. To enable this, the 6G-GOALS architecture introduces semantic-aware extensions within the O-RAN Alliance framework, centered on a Semantic-RIC that enhances Near-RT and Non-RT RICs with semantic intelligence. Through enriched interfaces, AI-enhanced applications (xApps and rApps), and support for semantic KPIs and policies, the Semantic-RIC enables dynamic adaptation of the radio stack based on the value and relevance of transmitted content. At the lower layers, L1/L2 evolve to support semantic-aware scheduling, retransmission, and resource allocation, coordinated by the Semantic-RIC to align network actions with application intent. Together, the Semantic-RIC and the 6G-GOALS architecture pave the way for goal-oriented, semantic-aware communication systems. While the presented evaluation focuses on fundamental semantic communication enablers, future work will extend this analysis toward system-level O-RAN validation, including the impact of semantic alignment on end-to-end latency, resource efficiency, and control-loop performance within the Near-RT RIC framework. This approach strengthens O-RAN's openness and programmability while enabling a new generation of networks that are not only faster and more efficient, but also more intelligent, adaptive, and user-centric.

Acknowledgement

This work has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation program Grant Agreement No. 101139232 6G-GOALS.

References

[1] S. Guo et al., "A Survey on Semantic Communication Networks: Architecture, Security, and Privacy," IEEE Communications Surveys & Tutorials, vol. 27, no. 5, pp. 2860–2894, 2025.
[2] H. Xie, Z. Qin, and G. Y. Li, "Deep Learning Enabled Semantic Communication Systems," IEEE Transactions on Signal Processing, vol. 69, pp. 2663–2675, 2021.
[3] T. Lan et al., "Task-Oriented Semantic Communication for 6G," IEEE Wireless Communications, 2022.
[4] H. Seo, J. Park, and S. Kim, "Multi-User Semantic Communication with Semantic-Aware Scheduling," IEEE Transactions on Communications, 2023.
[5] E. C. Strinati et al., "Goal-Oriented and Semantic Communication in 6G AI-Native Networks: The 6G-GOALS Approach," in Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), 2024, pp. 1–6.
[6] ——, "AI-Native 6G Networks: The 6GARROW Integrated Device-Network Approach," in Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit). IEEE, 2025, pp. 115–120.
[7] O-RAN Alliance, "O-RAN architecture description," O-RAN Alliance, Technical Specification O-RAN.WG1.O-RAN-Architecture-Description, July 2025. [Online]. Available: https://www.o-ran.org/specifications
[8] ——, "Non-real-time RAN intelligent controller (Non-RT RIC) and A1 interface: General aspects and principles," O-RAN Alliance, Technical Specification O-RAN.WG2.R1-AP, 2020.
[9] Y. Bansal, P. Nakkiran, and B. Barak, "Revisiting model stitching to compare neural representations," Advances in neural information processing systems, vol. 34, pp. 225–236, 2021.
[10] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," IEEE Wireless Communications, vol. 29, no. 1, pp. 210–219, 2022.
[11] A. Krizhevsky, G. Hinton et al., "Learning multiple layers of features from tiny images," 2009.
[12] I. Vulić, S. Ruder, and A. Søgaard, "Are All Good Word Vector Spaces Isomorphic?" in Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 3178–3192.
[13] S. Kornblith et al., "Similarity of Neural Network Representations Revisited," in International Conference on Machine Learning (ICML), 2019, pp. 3519–3529.
[14] Q. Lampin, L.-A. Dufrène, and G. Larue, "Semantic communications services within generalist operated networks," in 2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2024, pp. 861–865.
[15] M. Merluzzi, G. D. Poce, and P. D. Lorenzo, "Semantic and goal-oriented wireless network coverage: The area of effectiveness," IEEE Communications Magazine, pp. 1–6, 2026.