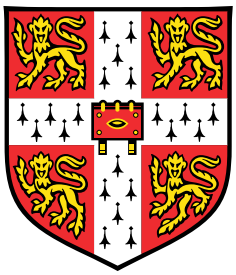
Semantic Communication for the Internet of Everything: From Molecular to Space Networks



Hanlin Cai

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy in Engineering

Girton College

October 2025

This thesis is dedicated to my parents, family, supervisor, colleagues, friends, and all those whose presence illuminates my life.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Hanlin Cai
October 2025

Acknowledgements

This thesis would not have been possible without the help and support of so many individuals and institutions. Here, I wish to express my heartfelt gratitude.

First and foremost, I must express my deepest gratitude to my supervisor, Dr. Özgür B. Akan, whose guidance and belief in me have been instrumental throughout this journey. Your invaluable advice, unwavering patience, and constant encouragement have shaped both my research and my academic future. I will always remember that success comes with Hard work, Integrity, Persistence, and Talent (HIP-T). I am also deeply thankful to my advisor, Dr. Kai Li from CMU Portugal, for your mentorship, which greatly broadened my perspectives and enriched my understanding of the field. Special thanks to Dr. Iman Tavakkolnia and Dr. Morag Hunter, whose support and insightful feedback were vital in shaping my MPhil studies at Cambridge.

I am fortunate to have had the opportunity to collaborate with Haofan Dong and Houtianfu Wang on Chapters 3 and 4 of this thesis. Working with you both has been an intellectually enriching experience, and I am grateful for your contributions and support in these works. Also, to my colleagues in the Internet of Everything (IoE) Group, I am grateful for your camaraderie and constant support. Shaojie Zhang, Hongbin Ni, Zherong Zhang, Zhengyang Zhang, Osman Tansel Baydas, and Ayse Sila Okcu—your thoughtful feedback and encouragement have been indispensable, making this journey so much more fulfilling. Let's IoE together for the upcoming four years!

To my friends from the Electrical Engineering Division, I am deeply grateful for the companionship and support you have given me during my time at CAPE. Dr. Peiji Song, Sizhe Xing (FDU), Xuyi Hu, Dr. Yang Shi, Mowei Lu, Linnan Chen, Siwei Liu, Yifan Zhao (UoT), Ke Ma, Ziyu He, Menglin Song, Tongxu Liu, Jinfeng Zhang, Lekang Jiang, Heyang Long, Yizhi Wang, Peng Bao, Yizhou Jiang (THU), Zhiwei Xue (THU), Dr. Zexi Li (ZJU), Dr. Xiaoyang Tian, Danchen Li, Haomin Luo, Shan Jin, and Zhongyi Liang—thank you for your kindness, encouragement, and friendship. Each of you has made my journey here so much richer. Life is so beautiful!

I would also like to extend my sincere thanks to my friends at Girton College, particularly Bofeng Xue, Hakan Emre Aktas, Mingze Li, Qien Cai, Xinghui Tao,

Yexuan Ding, Yuxuan Jiang, Haoyuan Jia, Mengcheng Zhang, Zhenyu Shi, and Sarim Gillani, whose friendship and kindness have made my time at Girton so much more meaningful. My heartfelt thanks go to my mentors in China, Dr. Zhezhuang Xu, Dr. Meng Yuan, Dr. Siyuan Zhan, Dr. Ronan Reilly, Dr. Chin Hong Wong, and Dr. Shibo He, whose guidance in my undergraduate years laid the foundation for my academic growth. Zhaolin Chen, Hongye Li, Wenxuan Luo, and Han Zong, as well as my friends in StarCollege, your unwavering belief in me has been a constant source of motivation.

I am also deeply grateful to the Cambridge Trust and the China Scholarship Council for their generous scholarships. Without their support, this dream would have remained just that—a dream. I am forever thankful for their belief in my huge potential.

A heartfelt thank you to my parents Huaxin Cai and Lianhong Xu, my little sister Xinyi Cai and my big family, whose unwavering support has empowered me to seize new chances and pursue my dreams. So what I always strive for is to make them proud of me. Lastly, to Linshi Li, the wonderful girl I met in a beautiful summer, who always believes in me a thousand times more than I do.

Abstract

The Internet of Everything (IoE) is a unifying framework that connects heterogeneous networks spanning bio-nano, space, and agent systems. Within the IoE framework, information exchange needs to operate across multiple modalities and heterogeneous channels under constrained spectrum, energy, and latency budgets, conditions under which traditional bit-level communication cannot meet these requirements. Semantic communication is a promising paradigm that aligns information transmission with task objectives by conveying task-relevant semantic representations instead of raw data. This thesis integrates semantic communication into the IoE framework and demonstrates its effectiveness across three domains: Internet of Bio-Nano Things (IoBNT), Internet of Space (IoS), and Internet of Agents (IoA).

The first contribution develops a semantic-empowered molecular communication framework for biomedical diagnostic tasks in the IoBNT. The framework employs a deep encoder–decoder architecture to extract, quantize, and reconstruct semantic features, and introduces a probabilistic channel network that approximates molecular propagation dynamics to enable gradient-based optimization for semantic learning. Experimental results show improved performance and robustness compared to conventional baselines.

The second contribution presents the first semantic communication architecture tailored to the IoS. A three-layer architecture is designed with a data layer for multi-modal data acquisition and feature extraction, a transport layer for semantic coding and transmission, and an application layer for semantic interpretation and decision making. Performance is evaluated through a representative deep-space case study on semantic-based monitoring of Martian dust storms.

The third contribution formalizes the IoA as a semantic-aware communication paradigm for coordination among heterogeneous large language model (LLM) agents and identifies federated learning as the enabling substrate for distributed agent coordination. The resilience of federated LLM agent networks is studied, including a critical analysis of mainstream security mechanisms. A graph representation–based poisoning attack is investigated and empirically evaluated, and the resulting insights inform a security roadmap for future IoA research.

Table of contents

List of figures **xiii**

List of tables **xvii**

Nomenclature **xix**

1 Introduction **1**
1.1 Semantic Communication 1
1.2 Internet of Everything 3
1.2.1 Internet of Bio-Nano Things (IoBNT) 3
1.2.2 Internet of Space (IoS) 4
1.2.3 Internet of Agents (IoA) 5
1.3 Thesis Structure 6

2 Semantic-Empowered Molecular Communication for the IoBNT **7**
2.1 IoBNT: Motivation and Challenges 7
2.2 Molecular Communication System Model 9
2.2.1 Semantic Coding Design 10
2.2.2 Molecular Propagation Model 11
2.2.3 Signal-to-Interference Ratio Analysis 13
2.3 Semantic Learning Framework 14
2.3.1 Encoder and Decoder Architecture 15
2.3.2 Channel Network 17
2.3.3 Network Training 18
2.4 Experiment and Evaluation 22
2.4.1 Experimental Setup 22
2.4.2 System Validation 24
2.4.3 Performance Evaluation 26
2.5 Summary 27

3 Semantic Communication for the Internet of Space 29
3.1 IoS: Requirements and Challenges 29
3.1.1 IoS Communication Challenges 30
3.1.2 Essential Requirements for IoS Communication 32
3.1.3 Semantic Communication as Enabler 33
3.2 Semantic-Empowered IoS: Architecture and Standardization 34
3.2.1 Data Layer 35
3.2.2 Transport Layer 36
3.2.3 Application Layer 37
3.2.4 Standardization and Interoperability Considerations 38
3.3 Experiment and Evaluation 39
3.4 Summary 42

4 Internet of Agents: A New Semantic-Aware Communication Paradigm for LLMs 45
4.1 IoA: New Paradigm and Challenges 45
4.1.1 Motivation of LLM Agent Networks 45
4.1.2 Wireless Federated Learning as an Enabler 46
4.1.3 Security and Privacy Challenges 46
4.2 Poisoning Attack on Federated LLM Agents 48
4.2.1 Threat Model and Existing Defense Mechanisms 48
4.2.2 Graph Representation-based Model Poisoning 51
4.2.3 Lagrange Dual Problem and Graph Signal Processing 53
4.2.4 Experiment and Evaluation 54
4.3 Security Roadmap for the Internet of Agents 57
4.3.1 Dual Semantic and Structural Auditing 57
4.3.2 Systems, Standards, and Evaluation 57
4.4 Summary 58

5 Conclusion 59

References 61

List of figures

2.1 Target application scenario for semantic-empowered molecular communication system: A bio–nano robot operates in the gastrointestinal tract to detect a suspected lesion. Task-oriented semantic features are extracted onboard, compressed, and transmitted over a molecular link under severe bandwidth, energy, and ISI constraints; the receiver reconstructs semantics for diagnostic decision support [14]. 9
2.2 3D molecular propagation channel with a constant uniform flow velocity. 9
2.3 End-to-end semantic molecular communication framework and training workflow. (1) Input data enter the encoder, which extracts task-relevant semantics and performs probabilistic quantization to produce a molecular signal. (2) The molecular signal passes through the molecular channel, whose stochastic input–output behavior is emulated by a learnable channel network. (3) The channel network is trained using a channel loss that compares its predictions with the quantized supervision. (4) The decoder maps the received signal to the task output, and the encoder and decoder are jointly optimized with a cross-entropy loss, while the channel network remains in the loop. (5) The trained encoder, decoder, and channel models are integrated back into the system for deployment in communication tasks. 14
2.4 Temporal variations of SIR during the transmission of a continuous sequence of five '1' symbol bits in two molecular communication scenarios. 24
2.5 Training and testing loss of the proposed channel network in two molecular communication scenarios. 25
2.6 Comparison of the normalized received molecular concentration between the analytical model and simulation model in two molecular communication scenarios. 26

2.7 Accuracy performance comparison between the proposed method and conventional benchmark methods in two molecular communication scenarios. 27

3.1 Semantic communication architecture for Earth–Mars–deep space links within the IoS framework. Key components include orbital satellites (LEO/GEO and relays), inter-satellite links, processing nodes, and surface assets (orbiters, landers, rovers), together with environmental stressors (radiation, solar wind, dust storms) that shape system design and operations. 31
3.2 Hierarchical satellite networking architecture of the semantic-enabled IoS. 33
3.3 Semantic Empowered IoS: A layered architecture illustrating multi-modal data processing, semantic encoding/decoding, transmission mechanisms, and the integration of standardization for enhanced semantic communication in space. 35
3.4 Illustration of semantic-based Mars surface exploration under dust storm conditions. 39
3.5 Illustration of the semantic communication pipeline for Mars dust storm monitoring based on the proposed IoS architecture. 40
3.6 Energy efficiency analysis for Mars dust storm monitoring scenario. Mission duration (days, logarithmic scale) is shown for 100/500/2000 bps; the red dashed line marks the minimum requirement. 42

4.1 (a) FedLLMs deployment across heterogeneous wireless communication networks. (b) Illustrative dialogue demonstrating LLM functionality as a wireless communication agent, contrasting normal versus poisoned model behaviors. 48
4.2 (a) Illustration of the FedLLMs, where each benign user trains a local model based on their private data, and the edge server aggregates all local benign updates to train a global model, which is then broadcast back to local clients for further training. (b) A legitimate but malicious client uploads a poisoned model update that degrades the global optimization, thereby influencing the global model and falsifying subsequent local training. 49
4.3 Framework for the proposed graph representation-based model poisoning (GRMP) attack. 52

4.4 GRMP attack's impact on learning accuracy and attack success rate over twenty communication rounds. 55
4.5 Individual client cosine similarity evolution with defense threshold over twenty communication rounds. 56

List of tables

2.1 Architectures of the Encoder and Decoder Networks 17
2.2 Architecture of the Channel Network 18
2.3 Parameters of the Molecular Propagation Channel for Different Scenarios 22

Nomenclature

Acronyms / Abbreviations

GEO Geostationary Orbit

GRMP Graph Representation-based Model Poisoning

GSP Graph Signal Processing

IoA Internet of Agents

IoBNT Internet of Bio-Nano Things

IoE Internet of Everything

IoS Internet of Space

ISAC Integrated Sensing and Communication

ISI Inter-Symbol Interference

JSCC Joint Source-Channel Coding

LEO Low-Earth Orbit

FedLLMs Federated Large Language Models

LLM Large Language Model

MC Molecular Communication

NN Neural Networks

non-IID Non-Independent and Identically Distributed

UAVs Unmanned Aerial Vehicles

VGAE Variational Graph Autoencoder

Chapter 1

Introduction

1.1 Semantic Communication

According to Shannon and Weaver [53], communication could be viewed through three hierarchical levels: (i) transmission of symbols; (ii) semantic exchange of transmitted symbols; (iii) effects of semantic information exchange. Specifically, the first level of communication primarily focuses on the successful transmission of symbols from the transmitter to the receiver, where transmission accuracy is typically assessed by bit or symbol error rates. The second level of communication deals with the semantic information sent from the transmitter and the meaning interpreted at the receiver, termed semantic communication. The third level concerns the effects of communication that turn into the receiver's ability to perform certain tasks as desired by the transmitter [27]. Despite advances from the first through the fifth generation that have driven practical performance toward the Shannon limit, achieving symbol-level reliability alone does not guarantee success at the higher two levels, particularly under stringent spectrum, energy, and latency constraints [68].

Building on recent advancements in deep learning for natural language processing and for modern communication systems, recent work has developed a semantic communication framework that operationalizes the second level of communication [55, 66, 67, 69, 23]. The central design shift is to replace the conventional separation of source and channel coding with an end-to-end joint source-channel coding architecture [10]. An encoder, a differentiable channel network, and a decoder are trained jointly so that the representation formed at the transmitter and the reconstruction produced at the receiver are governed by the same task criterion and by the statistics of the propagation medium [78]. Training instantiates semantic distortion through a task-assigned loss or a latent similarity measure and evaluates this objective under stochastic

channel impairments, thus coupling representation learning with channel conditions and resource constraints [38]. Furthermore, the training process supports regularization for communication budgets through rate splitting [76] and allows compatibility with conventional modulation and error control when required [11, 70], which facilitates integration with existing frameworks without altering the core objective.

Within the semantic communication framework, the treatment of input data is modality dependent and follows the same principle of preserving task-relevant meaning while suppressing redundancy. For text, the encoder maps sentences into task oriented latent semantics and the decoder reconstructs a meaning preserving representation suitable for downstream tasks [48]. For images, the joint source-channel coding maps visual content into channel robust feature tensors and reconstructs a representation that preserves semantics relevant to tasks such as classification or detection, with training objectives grounded in task-assigned loss or perceptual consistency [42]. For speech and audio, raw waveforms or spectrograms are embedded in content that carries representations that retain lexical or paralinguistic cues needed for recognition or understanding of intent while discarding channel-specific variations [64, 56]. For video, temporal encoders summarize motion and appearance into compact trajectories or segment level descriptors so that the decoder recovers the information required for activity analysis rather than pixel accurate frames [30]. These modality specific pipelines adhere to the same training protocol, namely optimizing a semantic distortion aligned with the downstream tasks under realistic channel statistics, which yields compact and robust representations across heterogeneous data sources.

Despite promising applications, several open questions remain. Semantic representations are inherently task dependent, and a universal metric applicable across modalities has not yet emerged. The robustness of latent representations to distribution shifts and adversarial perturbations is not fully understood, and existing countermeasures often operate at the symbol level rather than at the level of meaning. Interoperability is also an open problem, since there is no widely accepted syntax for semantic packets that would allow heterogeneous devices to exchange task-relevant content in a consistent manner. These challenges motivate a careful study of semantic encoders, semantic distortion metrics, and coding strategies under diverse constraints, which in turn calls for a unifying perspective that can be instantiated across different types of networks.

1.2 Internet of Everything

The Internet of Everything (IoE) is a framework for interoperability across specialized Internet-of-X (IoX) domains by enabling interaction among heterogeneous systems that differ in carriers, scales, media, and data semantics. IoE exploits complementarities rather than treating domains as isolated verticals and supports applications that range from molecular processes to planetary systems [1]. Within this landscape, the Internet of Bio-Nano Things (IoBNT) interconnects bio-nano devices that communicate through nonconventional mechanisms such as molecular signaling in diffusive and reactive media, with the objective of real time sensing and control of biological dynamics [36]. The Internet of Space (IoS) extends connectivity through constellations of small satellites and supporting ground and aerial segments to deliver global access and to manage links with long delays and intermittent contacts [5]. Furthermore, the Internet of Agents (IoA) provides an agent-centric infrastructure for autonomous entities driven by large language models or vision language models to discover capabilities, orchestrate tasks, and coordinate actions at scale [61].

Building on this foundation, the next three subsections introduce IoBNT, IoS, and IoA as representative IoX domains within the IoE framework. Although IoBNT and IoS operate at vastly different spatial and temporal scales, they exhibit similar channel characteristics, including long propagation delays, intermittent connectivity, strong interference, and stringent budgets; therefore, applying semantic communication in both domains follows a consistent design logic. Within the IoE framework, IoA functions as a service layer that ingests measurement data and semantic information from physical applications across the IoE (e.g., IoBNT and IoS) to train, coordinate, and deploy large model agents that support those environments through planning, orchestration, and closed-loop actuation. The common methodological thread is semantic communication: encode task-aligned meaning at the source, transport it robustly over communication channels, and decode it into mission objectives that guide agent policies. This perspective unifies the three domains in this thesis and provides a coherent transition to the problem formulations and design choices developed in subsequent chapters.

1.2.1 Internet of Bio-Nano Things (IoBNT)

The IoBNT interconnects bio–nano devices deployed within living tissues and microfluidic platforms for sensing, actuation, and closed-loop control of physiological processes. Communication in IoBNT commonly relies on molecular mechanisms,

including diffusion-driven propagation, ligand–receptor binding, enzymatic reaction pathways, and microbe- or cell-mediated transport, under stringent constraints on size and energy [2, 35]. Channel behavior exhibits long memory, stochastic arrival times, and pronounced intersymbol interference; system design must also address nanoscale energy harvesting, biocompatibility, and on-site processing of biosignals [33]. Representative applications include continuous intrabody health monitoring, disease diagnosis, targeted drug delivery, and lab-on-a-chip technology [80, 11, 88, 47]. The combination of unconventional carriers, severe resource limits, and clinical decision needs motivates communication strategies that preserve meaningful content rather than fidelity to raw data.

Contribution 1: This thesis proposes an end-to-end semantic learning framework designed to optimize task-oriented molecular communication, with a focus on biomedical diagnostic tasks under resource-constrained conditions. The proposed framework employs a deep encoder-decoder architecture to efficiently extract, quantize, and decode semantic features, prioritizing task-relevant semantic information to enhance diagnostic classification performance. Additionally, a probabilistic channel network is introduced to approximate the dynamics of molecular propagation, enabling gradient-based optimization for end-to-end learning. Experimental evaluations on a representative dataset demonstrate the effectiveness of the proposed framework.

1.2.2 Internet of Space (IoS)

The IoS enables connectivity across space, aerial, and ground segments. The space segment comprises low-earth orbit (LEO) and geostationary orbit (GEO) constellations that include CubeSats and larger satellites with intersatellite links; the aerial segment employs high-altitude platforms and unmanned aerial vehicles (UAVs); the ground segment consists of gateway stations, mission operations centers, and edge data centers orchestrated using software-defined networking and network function virtualization [4]. Operation across these segments encounters intermittent visibility windows, long and variable propagation delays, Doppler and blockage dynamics, stringent spectrum and power budgets, radiation-hardened hardware constraints, and contact-plan–driven intermittency [18, 6]. Application domains include global IoT backhaul for remote regions, disaster response and emergency connectivity, Earth observation for agriculture and climate, maritime and aviation monitoring, and deep-space science missions [85, 34, 13]. These characteristics create requirements for interoperable and contact-aware communication, for energy- and spectrum-efficient transmission, and for mission-

centric prioritization of information, thereby motivating a task-oriented and semantics-aligned design that will be developed next.

Contribution 2: This thesis proposes the first semantic communication architecture tailored to IoS environments. The design adopts a three-layer framework that includes a data layer for multimodal feature extraction, a transport layer for semantic encoding and transmission under intermittent contacts and long delays, and an application layer for mission-level interpretation and decision making. A synergistic ISAC with terahertz (THz) links is further introduced to align sensing fidelity with semantic entropy demands, improving spectral efficiency and task-aware robustness in harsh space conditions. The architecture is validated through a deep-space case study on semantic monitoring of Martian dust storms, demonstrating gains in energy efficiency, transmission reliability, and mission-oriented decision support.

1.2.3 Internet of Agents (IoA)

Recent advances in large language models (LLMs), such as ChatGPT, LLaMA, DeepSeek, and Gemini, have shifted AI from single-task utilities to autonomous agents capable of perception, reasoning, and action [62, 29]. Scaling such agents from standalone deployments to collaborative ecosystems requires an infrastructure for capability discovery, coordinated decision making, and policy enforcement. The Internet of Agents (IoA) aims to address this requirement by defining a new semantic-aware communication paradigm for LLMs, in which heterogeneous agents exchange intents, beliefs, and plans across virtual and physical environments [61]. The IoA infrastructure provides capability discovery, coordinated decision making, and policy enforcement, with core services for adaptive communication, dynamic task orchestration, consensus and conflict resolution, and incentive mechanisms [86]. Federated learning serves as the principal enabler by allowing geographically distributed agents to refine models through gradient or parameter sharing instead of raw data, thereby preserving data sovereignty and reducing dependence on centralized servers [7]. However, update traffic in federated LLMs (FedLLMs) introduces attack surfaces that include model poisoning, gradient leakage, and semantic manipulation, which exploit higher order correlations among benign updates and often evade detectors operating at the bit or symbol level [40, 39]. These security and privacy risks motivate an adversary-centric robustness study of FedLLMs within the IoA setting.

Contribution 3: This thesis first formalizes the IoA architecture as an agent-centric stack for semantic level exchanges of intents, beliefs, and plans, with federated learning as the communication substrate that coordinates model updates among distributed

large language model agents. Within this framework, the thesis examines the resilience of federated large language models in wireless networks. A focused review shows that prevailing defenses rely on distance or similarity based outlier detection and degrade under non-independent and identically distributed (non-IID) textual data, where adaptive adversaries craft updates that remain close to benign statistics. The study analyzes a graph representation-based model poisoning method that exploits higher order correlations among client gradients to evade detection. Finally, a security roadmap for future IoA research is outlined.

1.3 Thesis Structure

Building on the foundations of semantic communication and the Internet of Everything introduced above, the remainder of this thesis instantiates and evaluates the proposed methodology across three representative domains. The thesis is structured as follows:

Chapter 2 develops a semantic-empowered molecular communication pipeline for the Internet of Bio–Nano Things, including a system model for diffusion–reaction channels, an end-to-end encoder–decoder with a probabilistic channel network, and an experimental evaluation for biomedical diagnostics. Chapter 3 formulates requirements for the Internet of Space, presents a three-layer semantic architecture with standardization and interoperability considerations, and validates the design in a deep-space use case. Chapter 4 introduces the Internet of Agents as a semantic-aware communication paradigm for large language model agents, identifies federated learning as the enabling substrate, and investigates adversarial robustness with a graph-representation poisoning framework and corresponding analysis. Finally, Chapter 5 concludes this thesis.

Chapter 2

Semantic-Empowered Molecular Communication for the IoBNT

2.1 IoBNT: Motivation and Challenges

Molecular communication (MC) has emerged as a promising paradigm for information exchange in environments where traditional electromagnetic (EM)-based communication systems encounter fundamental limitations. Unlike EM waves, which suffer from severe attenuation and interference in biological and fluidic environments, MC relies on the controlled release, propagation, and detection of molecules to encode and transmit information [2]. This approach is particularly well-suited for applications in the Internet of Bio-Nano Things (IoBNT), where micro- and nanoscale devices operate in biological systems [16]. Key IoBNT applications include disease diagnosis, targeted drug delivery, and real-time health monitoring, where MC provides a biocompatible and energy-efficient communication mechanism [26].

Despite its potential, the practical deployment of MC in IoBNT faces significant challenges, including low data rates, severe inter-symbol interference (ISI), and high susceptibility to noise. These impairments substantially limit the molecular channel's ability to support complex data transmission, which is critical for biomedical applications such as disease diagnosis and physiological signal monitoring [65]. Given the stochastic nature of molecular propagation, addressing these challenges requires novel approaches to enhance communication efficiency while ensuring robustness under dynamic and uncertain channel conditions [8].

To overcome these limitations, incorporating semantic processing into communication systems has emerged as a promising solution for optimizing resource-constrained environments by prioritizing task-relevant information over conventional bit-level ac-

curacy [21]. In [10], a semantic-based joint source-channel coding (JSCC) framework was introduced to directly map source data to channel symbols, eliminating the need for separate compression and error correction. By jointly optimizing encoding and decoding, JSCC demonstrated enhanced robustness against noise and bandwidth constraints in wireless communication, ensuring graceful performance degradation under varying channel conditions. The work in [37] extended semantic communication to the IoBNT domain by integrating domain knowledge into the encoding process, improving efficiency in biologically constrained environments with strict energy and resource limitations. Furthermore, [78] investigated the integration of semantic communication with molecular systems. This work introduced an end-to-end training approach to enhance communication reliability under stochastic propagation effects, demonstrating the feasibility of semantic encoding in molecular channels.

Although semantic-based methods have been explored in molecular communication, existing approaches struggle to map task-relevant information into physically transmittable molecular parameters while accounting for the stochastic and non-differentiable nature of molecular propagation. Moreover, the lack of a structured mapping between high-level semantic information and molecular transmission parameters limits the adaptability and transferability of current models across dynamic channel conditions and diverse IoBNT tasks [11]. These gaps motivate an end-to-end design that couples task-aligned semantic learning with a learnable mapping to molecular control parameters and a differentiable surrogate of the propagation process.

In this work, we propose an end-to-end semantic molecular communication framework using a deep encoder-decoder architecture to extract, quantize, and decode task-oriented semantic features. We introduce a quantization function to optimize the semantic-to-physical mapping and enhance system transferability. To achieve channel differentiability, we further propose a probabilistic channel network that models the stochastic dynamics of molecular propagation. This integration facilitates end-to-end training and dynamic adaptation to channel conditions. Unlike conventional methods focused on bit-level transmission, our method prioritizes semantics aligned with task objectives, demonstrating superior efficiency and robustness over traditional baselines in diagnostic image classification tasks. Figure 2.1 specifies the target application studied in this work: a bio–nano robot operating in the gastrointestinal tract performs lesion assessment, extracts decision-critical semantics onboard, and conveys these semantics over a molecular link under stringent bandwidth, energy, and reliability constraints [14]. This scenario defines the operational requirements that drive the system design.

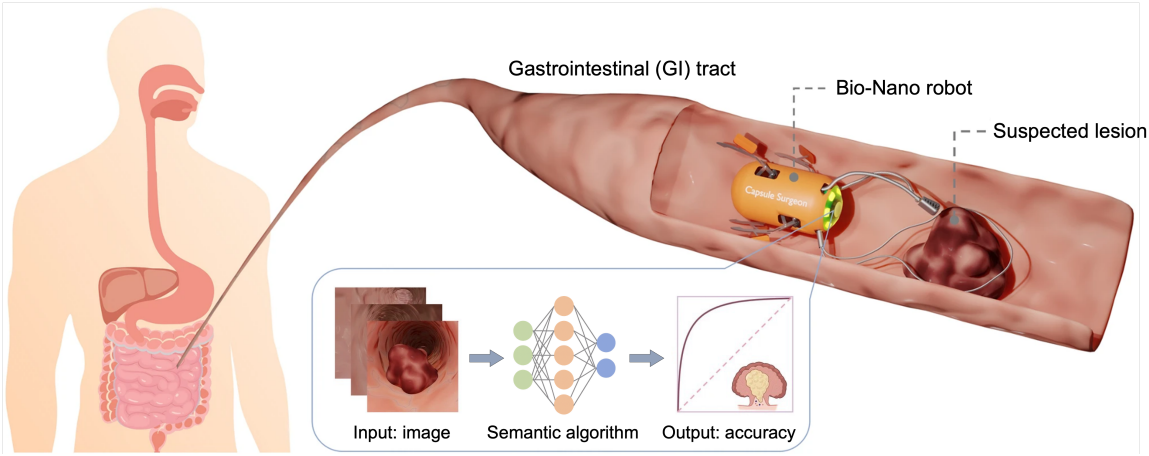


Fig. 2.1 Target application scenario for semantic-empowered molecular communication system: A bio–nano robot operates in the gastrointestinal tract to detect a suspected lesion. Task-oriented semantic features are extracted onboard, compressed, and transmitted over a molecular link under severe bandwidth, energy, and ISI constraints; the receiver reconstructs semantics for diagnostic decision support [14].

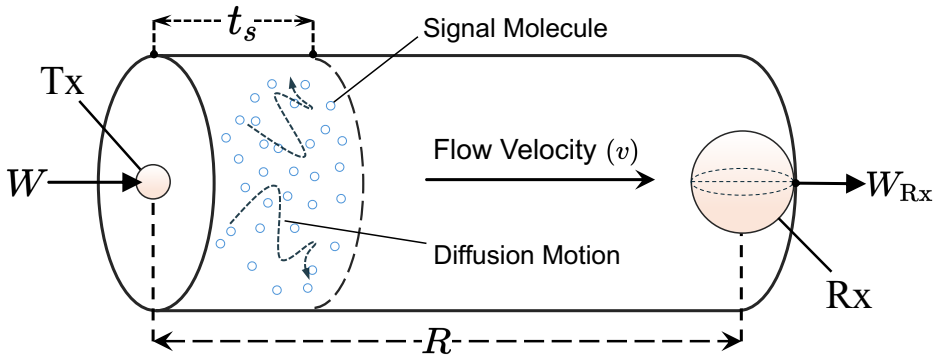


Fig. 2.2 3D molecular propagation channel with a constant uniform flow velocity.

2.2 Molecular Communication System Model

In this work, we consider a single-input-single-output (SISO) molecular communication system operating in an unbounded, three-dimensional environment with a constant uniform flow velocity, as shown in the Fig. 2.2. The transmitter (Tx) and receiver (Rx) are assumed to be synchronized, ensuring precise alignment during each symbol transmission [52]. The Tx encodes information by instantaneously releasing identical molecules at the beginning of each symbol slot with duration t_s. All molecules are assumed to share identical physical properties, such as size and diffusivity, and are not subject to collisions or chemical reactions. Consequently, their motion is determined by a combination of random Brownian diffusion and a uniform drift at velocity v.

2.2.1 Semantic Coding Design

The semantic coding design of the molecular communication system is structured to accommodate a wide range of input data types, provided they align with the semantic objectives of the system. The input data, denoted as χ, can represent diverse forms of information, such as medical images, environmental sensory data, or encoded signals, depending on the specific application context. In this work, χ represents images, including medical images used for diagnostic image classification tasks in the IoBNT. Mathematically, $\chi \in \mathbb{R}^{H \times W \times C}$, where H, W, and C denote the height, width, and number of channels of the image, respectively. These images serve as the initial input to the transmitter, which extracts semantic features essential for downstream tasks, such as classification or detection. The transmitter utilizes an encoder to transform χ into a lower-dimensional representation of semantic features $\mathcal{F}$. These features capture the essential information required for the communication task while eliminating redundant data. The transformation is mathematically expressed as:

$$\mathcal{F} = f_\theta(\chi), \tag{2.1}$$

where $f_\theta(\cdot)$ denotes the encoding function parameterized by θ. This function transforms the input data χ into a lower-dimensional semantic feature representation $\mathcal{F}$, capturing task-relevant information while eliminating redundancies. This encoding step is critical for optimizing the limited bandwidth of the molecular communication system. The detailed design and implementation of $f_\theta(\cdot)$ will be discussed in subsequent sections. To enable molecular transmission, the semantic features $\mathcal{F}$ are further mapped to a normalized vector of channel input symbols $W \in [0,1]^k$ using a quantization function $Quantize(\cdot)$. This process is defined as:

$$W = Quantize(\mathcal{F}) = Q_\beta(\mathcal{F}), \tag{2.2}$$

where $W = [W_1, W_2, \ldots, W_k]$, and each element $W_i \in [0,1]$ represents the normalized proportion of molecules to be released in the corresponding time slot. Specifically, the actual number of molecules released by the transmitter is determined by $W_i \cdot n_m$, where n_m is the maximum molecular release capacity. This design ensures that the semantic features $\mathcal{F}$ are efficiently encoded into a compact, continuous representation, compatible with molecular communication constraints. The quantization process employs non-linear transformations and probabilistic decisions to generate the vector W. This approach facilitates the end-to-end optimization of the communication

system by allowing the encoder and decoder to jointly adapt to the molecular channel characteristics. Further details on the mathematical formulation and implementation of $Q_\beta(\mathcal{F})$ will be elaborated in later sections.

2.2.2 Molecular Propagation Model

The molecular communication (MC) system under consideration features a point transmitter (Tx) at $\mathbf{d}_{\text{Tx}} = [0, 0, 0]$ and a spherical receiver (Rx) centered at $\mathbf{d}_{\text{Rx}} = [R, 0, 0]$. The position of an information molecule at time t is denoted by $\mathbf{d}(t) = [x_t, y_t, z_t]$, and a uniform drift velocity $\mathbf{v} = [v, 0, 0]$ acts along the x-axis. At the Tx's location $\mathbf{d}_{\text{Tx}}$ and time t, the molecular concentration is expressed as:

$$\Phi(\mathbf{d}_{\text{Tx}}, t) = \sum_{i=1}^{k} \Phi_{\text{Tx},i}\Big(\mathbf{d}_{\text{Tx}}, t\Big), \tag{2.3}$$

where $\Phi_{\text{Tx},i}(\mathbf{d}_{\text{Tx}}, t)$ represents the contribution to the concentration for the i-th bit of the normalized vector of channel input symbols W, which is expressed as:

$$\Phi_{\text{Tx},i}(\mathbf{d}_{\text{Tx}}, t) = W_i n_m \, \delta\Big(t - t_i\Big), \tag{2.4}$$

where $W_i \in [0, 1]$ is the normalized release factor for the i-th time slot, n_m is the maximum number of molecules released per time slot, $t_i = (i-1)\, t_s$ is the release time for the i-th bit, and $\delta(\cdot)$ is the Dirac delta function modeling instantaneous molecular release. Once emitted, the molecules propagate through the medium via a combination of random Brownian diffusion and directed drift. This process is governed by the advection–diffusion equation [26]:

$$\frac{\partial \Phi(\mathbf{d}, t)}{\partial t} = D\nabla^2 \Phi(\mathbf{d}, t) - \nabla\Big(\mathbf{v}\Phi(\mathbf{d}, t)\Big), \tag{2.5}$$

where $\Phi(\mathbf{d}, t)$ is the molecular concentration at position $\mathbf{d}$ and time t, D_c is the diffusion coefficient, ∇^2 denotes the Laplace operator, and ∇ is the gradient operator in Cartesian coordinates. The first term on the right-hand side captures molecular diffusion, while the second term describes the effect of the uniform drift velocity $\mathbf{v}$.

To fully characterize the molecular propagation, initial and boundary conditions are imposed [87]. At $t = 0$, all molecules are assumed to be concentrated at $\mathbf{d}_{\text{Tx}}$, represented by the spatial Dirac delta $\Phi(\mathbf{d}, 0) = \delta\Big(\mathbf{d} - \mathbf{d}_{\text{Tx}}\Big)$. The boundary condition assumes an unbounded environment, such that $\lim_{\|\mathbf{d}\| \to \infty} \Phi\Big(\mathbf{d}, t\Big) = 0$ for all $t > 0$. Solving the advection-diffusion equation under these conditions yields the probability

density function (PDF) for finding a single molecule at position $\mathbf{d}$ and time t:

$$f(\mathbf{d}, t) = \frac{1}{\Big(4\pi D_c\, t\Big)^{3/2}} \exp\Big(-\frac{\|\mathbf{d} \;-\; \mathbf{v}\, t \;-\; \mathbf{d}_{\mathrm{Tx}}\|^2}{4\, D_c\, t}\Big). \tag{2.6}$$

This PDF describes the spatiotemporal distribution of molecules as they propagate through the medium under the combined effects of diffusion and drift. Intuitively, the PDF indicates that molecules are more likely to be found near the transmitter at earlier times, while increased diffusion and drift broaden the distribution as time progresses, reducing the likelihood of capture at distant locations. To evaluate the likelihood of molecular capture at the receiver, the PDF is integrated over the spherical volume of the receiver, denoted as $V_r = \frac{4\pi r^3}{3}$. Under the Uniform Concentration Assumption (UCA) [46], which applies when the receiver is sufficiently far from the transmitter, the molecular concentration is approximately uniform within the receiver's volume. Using this assumption, the probability of capturing a molecule at the receiver simplifies to:

$$P(t) = \frac{V_r}{\Big(4\pi D_c\, t\Big)^{3/2}} \exp\Big(-\frac{\Big(R - v\, t\Big)^2}{4\, D_c\, t}\Big), \tag{2.7}$$

where R represents the distance between the transmitter and receiver, v is the uniform drift velocity, and t is the time elapsed since molecular release. Since each signal molecule is transmitted independently, the number of molecules observed by the receiver, denoted as N, follows a binomial distribution. Specifically, for n_m molecules released by the transmitter, the distribution is given by:

$$N(n_m, t) \sim B\Big(n_m, P(t)\Big), \tag{2.8}$$

where $B(\cdot)$ denotes the binomial distribution, and $P(t)$ is the probability of capturing a single molecule at the receiver, as derived previously. When n_m is sufficiently large, the binomial distribution can be approximated by a Gaussian distribution [78]:

$$N(n_m, t) \sim \mathcal{N}\Big(n_m P(t), n_m P(t)\Big(1 - P(t)\Big)\Big), \tag{2.9}$$

where $\mathcal{N}(\cdot)$ denotes the Gaussian distribution, with mean $n_m P(t)$ and variance $n_m P(t)(1 - P(t))$. During the communication process, molecules released in prior time slots may arrive at the receiver due to the uncertainty introduced by molecular diffusion. This effect, known as ISI, is typically negligible when the drift velocity

significantly dominates the Brownian diffusion. However, when diffusion becomes the predominant propagation mode, ISI can degrade the system's communication performance significantly. Additionally, the molecular communication channel may introduce Gaussian noise due to molecular decomposition or emissions from other nano-machines. This noise is modeled as $N_{\text{noise}} \sim \mathcal{N}(0, \sigma_n^2)$. Considering both ISI and noise, the total number of molecules observed by the receiver at the j-th time slot can be expressed as:

$$N_{\text{obs}}(j,t) = W_j N(n_m, t) + \sum_{i=1}^{\lambda} W_{(j-i)} N(n_m, t + it_s) + N_{\text{noise}}, \tag{2.10}$$

where W_j is the transmitted bit at the j-th time slot, and λ represents the length of the channel memory, capturing contributions from previous time slots. For simplicity, in this work, we consider $\lambda = 1$, accounting for ISI caused by molecules released in the immediately preceding time slot.

2.2.3 Signal-to-Interference Ratio Analysis

The signal-to-interference ratio (SIR) is a critical metric for assessing the performance of molecular communication systems, particularly in scenarios affected by ISI. ISI arises from the delayed arrival of molecules released in previous symbol slots, which interfere with the current transmission. In the proposed semantic coding scheme, the SIR is defined as:

$$SIR = \frac{W_j N(n_m, t)}{\sum_{i=1}^{\lambda} W_{(j-i)} N(n_m, t + it_s) + N_{\text{noise}}}, \tag{2.11}$$

where W_j and $W_{(j-i)}$ represent the transmitted bits at the current and previous time slots, respectively. $N(n_m, t)$ denotes the expected number of molecules observed at the receiver at time t, and N_{noise} represents Gaussian noise introduced during molecular propagation. A higher SIR indicates better MC performance, as the desired signal dominates residual ISI and noise. Two key physical factors, the transmitter–receiver distance d and the flow velocity v, strongly influence the SIR: increasing d reduces the observed concentration at the receiver due to greater diffusion and attenuation over longer paths, while decreasing v increases temporal dispersion (longer residence times), thereby strengthening the ISI tail and allowing molecules from prior symbols to interfere more with the current observation.

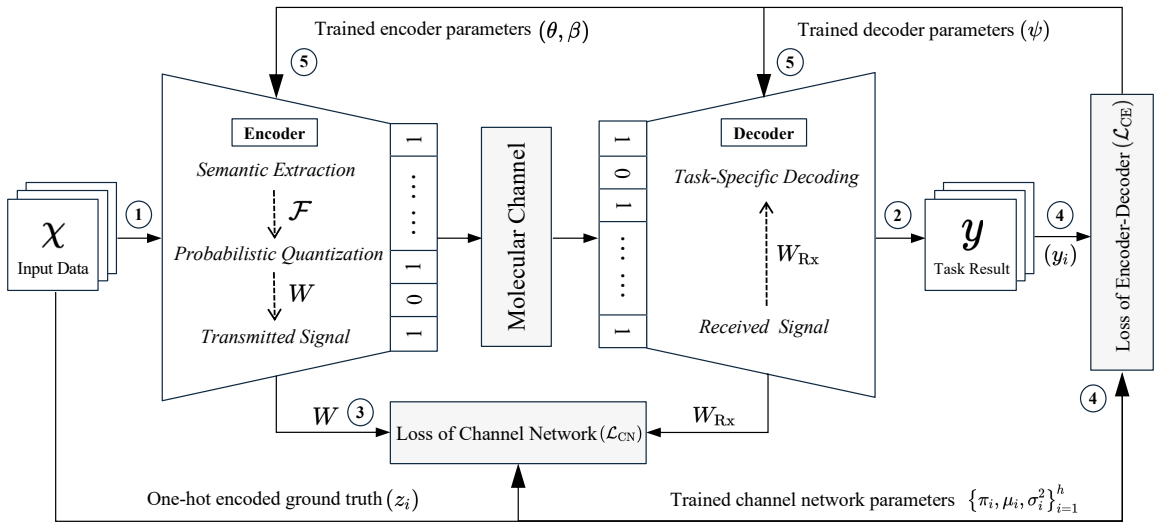


Fig. 2.3 End-to-end semantic molecular communication framework and training workflow. (1) Input data enter the encoder, which extracts task-relevant semantics and performs probabilistic quantization to produce a molecular signal. (2) The molecular signal passes through the molecular channel, whose stochastic input–output behavior is emulated by a learnable channel network. (3) The channel network is trained using a channel loss that compares its predictions with the quantized supervision. (4) The decoder maps the received signal to the task output, and the encoder and decoder are jointly optimized with a cross-entropy loss, while the channel network remains in the loop. (5) The trained encoder, decoder, and channel models are integrated back into the system for deployment in communication tasks.

2.3 Semantic Learning Framework

In this section, we introduce the proposed end-to-end semantic molecular communication framework, which is specifically designed to address diagnostic image classification tasks in the IoBNT. The framework aims to determine task-specific outputs, such as disease severity or diagnostic labels, based on molecularly transmitted information. As illustrated in Fig. 2.3, the proposed framework integrates semantic feature extraction, molecular communication channel modeling, and semantic decoding into a unified system. By employing deep neural networks, the framework efficiently encodes and transmits task-relevant information while mitigating the effects of channel noise, ISI, and stochastic distortions inherent to molecular communication. This design enables robust task execution even in challenging communication environments, ensuring high reliability and accuracy for IoBNT applications.

2.3.1 Encoder and Decoder Architecture

The proposed framework integrates semantic feature extraction, quantization, and decoding to enable robust end-to-end learning in molecular communication systems. This section provides a detailed explanation of the key components of the system, explaining the functions of the encoder and decoder.

Semantic Feature Extraction

The encoder, parameterized by f_θ, transforms the input medical image χ into a lower-dimensional semantic representation $\mathcal{F} \in \mathbb{R}^k$ through the mapping $\mathcal{F} = f_\theta(\chi)$. Here, $f_\theta(\cdot)$ represents a convolutional neural network (CNN) augmented with a final linear transformation layer. This design extracts task-oriented semantic features while minimizing redundancy, providing a continuous, unnormalized representation $\mathcal{F}$ suitable for subsequent processing. The semantic features $\mathcal{F}$ encapsulate high-level abstractions of the input image χ, such as diagnostically significant patterns or regions indicative of disease severity.

The encoder consists of five convolutional layers, each followed by batch normalization to stabilize training and LeakyReLU activation functions to introduce non-linearity. This hierarchical design progressively reduces the spatial dimensions of the input image while increasing the abstraction level of the extracted features. The output $\mathcal{F}$ from the encoder serves as a compact, high-dimensional semantic representation, which is further processed by the quantization module $Q_\beta(\mathcal{F})$ to generate the normalized molecular transmission parameters $W \in [0, 1]^k$.

Probabilistic Quantization

To transform the semantic features $\mathcal{F}$ into a normalized vector of channel input symbols $W \in [0, 1]^k$, the quantization function $W = Q_\beta(\mathcal{F})$ is employed as a critical intermediate step, where $Q_\beta(\cdot)$ is implemented as a fully connected neural network parameterized by Q_β. This function maps each component $\mathcal{F}_i$ of the semantic feature vector to a corresponding element $W_i \in [0, 1]$ in the normalized output vector W. This design ensures that the semantic features are translated into a physically interpretable molecular communication parameter, enabling seamless integration with the channel.

The distinction between $\mathcal{F}$ and W underscores the necessity of the quantization step. While $\mathcal{F}$ captures high-level, abstract semantic features relevant to the diagnostic task, these features are not directly compatible with the constraints of molecular communication. As a normalized vector of channel input symbols, W aligns the

semantic representation with the physical requirements of the communication system. This transformation is crucial for facilitating end-to-end optimization while maintaining interpretability and compliance with molecular constraints. By incorporating $Q_\beta(\mathcal{F})$, the system ensures robust gradient flow during training and effective adaptation of the semantic features to the molecular communication, achieving both task relevance and transmission efficiency.

Task-Specific Decoding

The decoder, parameterized by g_ψ, directly maps the received channel output symbols W_{Rx} to the task-specific output y, which represents the predicted probability distribution over eight classes in image classification tasks. This decoding process is formulated as:

$$y = g_\psi(W_{\text{Rx}}), \tag{2.12}$$

where $g_\psi(\cdot)$ is implemented as a series of fully connected layers designed to process the normalized molecular transmission parameters W_{Rx} into the semantic output y. The final layer employs a Softmax activation function to produce a probability distribution over the task-specific output space. To ensure the framework is optimized for the classification task, the cross-entropy loss function is employed:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{k} z_i \log\left(y_i\right), \tag{2.13}$$

where z_i and y_i are the one-hot encoded ground truth and the predicted probabilities, respectively. The end-to-end optimization of the framework minimizes this loss function:

$$(\theta^*, \beta^*, \psi^*) = \arg\min_{\theta,\beta,\psi} \mathcal{L}_{\text{CE}}. \tag{2.14}$$

The encoder and decoder architecture, as shown in Table 2.1, is meticulously designed to achieve efficient and robust task-specific coding and decoding in the proposed framework. Each layer is defined by its type, activation function, and other hyperparameters, all of which are carefully optimized to ensure superior performance. Furthermore, the parameter c_{out}, representing the number of filters in the final convolutional layer of the encoder function $f_\theta(\chi)$, allows the framework to flexibly balance bandwidth constraints with task-specific accuracy. This adaptability ensures that the architecture can cater to diverse molecular communication scenarios, maintaining high semantic fidelity while adhering to the physical limitations of the channel. The design not only

Table 2.1 Architectures of the Encoder and Decoder Networks

Function	**Layer**	**Type**	**Out/Kernel**	**Stride/Pad**	**Activation / Post**
$f_{\theta}(\chi)$	Conv1	Conv2D	$32 \times 9{\times}9$	2/4	BN + LeakyReLU (α=0.1)
	Conv2	Conv2D	$64 \times 5{\times}5$	2/2	BN + SE + LeakyReLU
	Conv3	Conv2D	$128 \times 3{\times}3$	2/1	BN + SE + LeakyReLU
	Conv4	Conv2D	$128 \times 3{\times}3$	1/1	BN + LeakyReLU
	Conv5	Conv2D	c_{out}=$128 \times 3{\times}3$	1/1	Linear
	GAP	GlobalAvgPool2D	c_{out}=$128 \times -$	$-/-$	Output $\mathcal{F} \in \mathbb{R}^{128}$
$Q_{\beta}(\mathcal{F})$	FC1	Fully Connected	$128 \times -$	$-/-$	ReLU
	FC2	Fully Connected	$128 \times -$	$-/-$	ReLU
	Output	Fully Connected	k=$64 \times -$	$-/-$	Sigmoid ($[0, 1]$)
$g_{\psi}(W_{\text{Rx}})$	FC1	Fully Connected	$128 \times -$	$-/-$	BN + ReLU
	FC2	Fully Connected	$64 \times -$	$-/-$	Dropout (p=0.5) + ReLU
	Output	Fully Connected	N_{class}=$8 \times -$	$-/-$	Softmax

facilitates effective processing of received data but also enhances the relevance and reliability of the output for downstream tasks.

2.3.2 Channel Network

The channel network is a vital component of the communication framework, serving as a probabilistic model to capture the stochastic behavior of molecular communication channels. These channels are inherently random due to phenomena such as noise, molecular diffusion, and ISI, all of which can significantly distort the transmitted channel symbols W. By modeling the conditional probability distribution of the received symbols W_{Rx}, the channel network effectively addresses these challenges and enables robust end-to-end optimization.

In molecular communication, the number of transmitted molecules directly influences the received signals at the receiver, which typically follow a binomial distribution. This distribution can be approximated by a Gaussian distribution when the number of transmitted molecules is sufficiently large, enabling computationally efficient modeling [2]. To capture the stochastic transformations introduced by the channel, the channel network models the conditional distribution of W_{Rx} as a mixture of Gaussian distributions:

$$p(W_{\text{Rx}}|W) = \sum_{i=1}^{h} \pi_i(W)\varphi_i(W_{\text{Rx}}|W), \tag{2.15}$$

Table 2.2 Architecture of the Channel Network

Layer	Type	Output Dim.	Activation
Input Layer	Fully Connected	$h_{\text{hidden}} = 20$	LeakyReLU
Feature Extraction	Fully Connected	$h_{\text{hidden}} = 20$	LeakyReLU
Mean Output (μ)	Fully Connected	$h = 2$	Linear
Variance Output (σ^2)	Fully Connected	$h = 2$	ReLU
Mixing Coeff. (π)	Fully Connected	$h = 2$	Softmax

where $h = 2$ represents the number of Gaussian components, π_i are the mixing coefficients that satisfy $\sum_{i=1}^{h} \pi_i(W) = 1$, and $\varphi_i(W_{\text{Rx}})$ denotes the i-th Gaussian kernel:

$$\varphi_i(W_{\text{Rx}}|W) = \frac{1}{\sqrt{2\pi\ \sigma_i^2(W)}} \exp \frac{-\|(W_{\text{Rx}} - \mu_i(W)\|^2}{2\sigma_i^2(W)}, \tag{2.16}$$

with μ_i and σ_i^2 representing the mean and variance of the i-th Gaussian component, respectively. The channel network is implemented with multiple fully connected layers designed to estimate the parameters of the Gaussian mixture model, including the means (μ_i), variances (σ_i^2), and mixing coefficients (π_i). The mixture captures multi-modality induced by random propagation and interference: means μ_i shift with expected reception levels, variances σ_i^2 widen under stronger uncertainty, and the weights π_i adaptively trade between these regimes as channel conditions vary. These learned parameters define the conditional distribution $p(W_{\text{Rx}}|W)$, facilitating accurate modeling of the stochastic channel effects and enabling robust decoding of the transmitted symbols.

To address ISI, the channel network processes input symbols corresponding to molecules emitted in different time slots. A sliding input mechanism dynamically adjusts features associated with earlier emissions to account for their delayed impact on the current time slot. This design effectively mitigates ISI by incorporating historical molecular contributions into the modeled distribution. The detailed architecture of the channel network is outlined in **Table 2.2**, where each fully connected layer captures the non-linear relationships between transmitted and received channel symbols, providing an accurate representation of molecular channel dynamics.

2.3.3 Network Training

The training of the proposed communication framework is divided into two stages: channel network pre-training and joint optimization of the encoder and decoder. This

two-stage process ensures that the channel model accurately captures the stochastic dynamics of molecular communication, enabling effective end-to-end optimization.

Pre-training the Channel Network

The channel network is trained separately using randomly generated channel symbols vectors W and their corresponding received vectors W_{Rx}, simulated based on the molecular propagation channel described in Section 2.2.2. The data are generated using Smoldyn, a particle-based molecular communication simulation software, to mimic the behavior of the molecular propagation system. This simulation captures the effects of noise, diffusion, and ISI, providing realistic training data for the channel model. The channel network learns the conditional probability $p(W_{\mathrm{Rx}}|W)$, modeled as a Gaussian mixture distribution, by minimizing the negative log-likelihood loss:

$$\mathcal{L}_{\mathrm{CN}} = -\frac{1}{k}\sum_{j=1}^{k}\log\left(\sum_{i=1}^{h}\pi_i(W_j)\varphi_i(W_{\mathrm{Rx}j}\Big|W_j)\right) \tag{2.17}$$

where W_j and $W_{\mathrm{Rx}j}$ denote the transmitted and received channel symbols at the j-th instance, respectively. The terms π_i and φ_i represent the mixing coefficients and Gaussian kernel functions, parameterized by the channel network, defined as:

$$\varphi_i(W_{\mathrm{Rx}j}\Big|W_j) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(W_{\mathrm{Rx}j}-\mu_i)^2}{2\sigma_i^2}\right) \tag{2.18}$$

where μ_i and σ_i^2 are the mean and variance of the i-th Gaussian component, and $\pi_i(W_j)$ are the mixing coefficients that satisfy $\sum_{i=1}^{h}\pi_i(W_j)=1$. The training algorithm used to optimize the parameters π_i, μ_i, and σ_i^2 is outlined in **Algorithm 1**. This iterative optimization process updates the parameters using gradient descent until convergence. Upon completion of pre-training, the channel network's parameters are fixed, and the network is treated as a fixed component during the subsequent training of the encoder and decoder. This ensures that the encoder and decoder can adapt their parameters to optimize task performance based on the fixed approximation of the molecular channel. By leveraging pre-trained channel network parameters, the system achieves robust modeling of molecular signal propagation under stochastic channel conditions.

Joint Training of the Encoder and Decoder

Once the channel network is pre-trained, the encoder $f_\theta(\cdot)$, $Q_\beta(\cdot)$ and decoder $g_\psi(\cdot)$ are jointly optimized in an end-to-end manner to minimize the task-specific loss function.

Algorithm 1 Pre-training Algorithm for Channel Network

Input: A set of transmitted channel symbols W and corresponding received channel symbols W_{Rx}; Learning rate ξ; Maximum iterations N.

1: **Initialization:** Randomly initialize the channel network parameters $\{\pi_i^{(0)}, \mu_i^{(0)}, \sigma_i^{2(0)}\}$. Set iteration counter $i = 0$.
2: **while** the convergence criterion is not met and $i < N$ **do**
3: Compute the Gaussian kernel values $\varphi_i(W_{\mathrm{Rx}j} \big| W_j)$ for all j and i.
4: Compute the conditional probability distribution $p(W_{\mathrm{Rx}}|W)$ as a mixture of Gaussian components.
5: Compute the negative log-likelihood loss $\mathcal{L}_{\mathrm{CN}}$.
6: Update the network parameters via gradient descent:

$$\pi_i^{(i+1)} \leftarrow \pi_i^{(i)} - \xi \cdot \nabla_{\pi_i^{(i)}} \mathcal{L}_{\mathrm{CN}}, \tag{2.19}$$

$$\mu_i^{(i+1)} \leftarrow \mu_i^{(i)} - \xi \cdot \nabla_{\mu_i^{(i)}} \mathcal{L}_{\mathrm{CN}}, \tag{2.20}$$

$$\sigma_i^{2\ (i+1)} \leftarrow \sigma_i^{2\ (i)} - \xi \cdot \nabla_{\sigma_i^{2(i)}} \mathcal{L}_{\mathrm{CN}}. \tag{2.21}$$

7: Increment the iteration counter: $i \leftarrow i + 1$.
8: **end while**
9: **Output:** Trained network parameters $\{\pi_i, \mu_i, \sigma_i^2\}_{i=1}^h$.

For batch-based training, the loss function in Equation 2.13 is extended to account for multiple samples and classes. The predicted probabilities y_i are computed using the Softmax function, ensuring valid probability distributions over the class labels.

The training procedure is detailed in **Algorithm 2**. The process begins with the initialization of the encoder and decoder parameters $\theta^{(0)}$, $\beta^{(0)}$ and $\psi^{(0)}$, while the channel network parameters are fixed after pre-training. At each iteration, the encoder extracts semantic features $\mathcal{F}$ from the input χ, which are quantized into symbol representations W via the probabilistic quantization module $Q(\mathcal{F})$. These symbols W are transmitted through the molecular communication channel, resulting in the received symbols W_{Rx}. Instead of reconstructing intermediate semantic features, the decoder directly maps W_{Rx} to task-specific outputs y, streamlining the decoding process and reducing computational complexity. The cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ is computed and used to update the parameters of both the encoder and decoder through gradient descent.

This joint optimization approach allows the encoder to effectively extract semantic features that are robust to the stochastic effects introduced by the molecular channel, while enabling the decoder to output task-specific outputs. By incorporating the channel network as a fixed component during training, the end-to-end system achieves optimal performance under the constraints of molecular communication environment.

Algorithm 2 Training Algorithm for Encoder and Decoder

Input: A batch of input data χ, ground truth labels Z; initialized encoder and decoder parameters $\theta^{(0)}, \beta^{(0)}, \psi^{(0)}$; fixed channel network parameters (after pre-training); learning rate ξ', and maximum iterations N.

Initialization: Set iteration counter $i = 0$.
while the convergence criterion is not met and $i < N$ **do**
    Extract semantic features: $\mathcal{F} \leftarrow f_{\theta^{(i)}}(\chi)$.
    Quantize semantic features: $W \leftarrow Q_{\beta^{(i)}}(\mathcal{F})$.
    Transmit the channel symbols W through the channel and obtain the received symbols W_{Rx}.
    Decode task-specific outputs: $y \leftarrow g_{\psi^{(i)}}(W_{\mathrm{Rx}})$.
    Compute the cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$.
    Update network parameters via gradient descent:

$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \xi \nabla_{\theta^{(i)}} \mathcal{L}_{\mathrm{CE}}, \tag{2.22}$$

$$\beta^{(i+1)} \leftarrow \beta^{(i)} - \xi \nabla_{\beta^{(i)}} \mathcal{L}_{\mathrm{CE}}, \tag{2.23}$$

$$\psi^{(i+1)} \leftarrow \psi^{(i)} - \xi \nabla_{\psi^{(i)}} \mathcal{L}_{\mathrm{CE}}. \tag{2.24}$$

    Increment iteration counter: $i \leftarrow i + 1$.
end while
Output: Trained encoder $f_\theta(\cdot), Q_\beta(\cdot)$ and decoder $g_\psi(\cdot)$.

Optimization and Stopping Criteria

The entire training framework is optimized using the Adam optimizer, which is well-suited for handling the stochastic nature of the training process. The learning rate and other hyperparameters, such as the beta values for moment estimation, are carefully tuned through empirical evaluation to achieve stable convergence. During training, the objective is to minimize the task-specific loss function, ensuring that the encoder learns to extract robust and discriminative semantic features, while the decoder accurately maps the received symbols W_{Rx} to the final task-specific output y. The training process continues until one of the following stopping criteria is met.

1. **Convergence of Loss:** The loss function $\mathcal{L}_{\mathrm{CE}}$ shows negligible changes across consecutive epochs, indicating that the model parameters have stabilized.

2. **Maximum Epochs:** A pre-defined number of epochs is reached, ensuring that the training process does not overfit the model to the simulated data.

This two-stage training strategy effectively decouples the pre-training of the channel network from the joint optimization of the encoder and decoder. By leveraging

Table 2.3 Parameters of the Molecular Propagation Channel for Different Scenarios

Parameter	**Scenario 1**	**Scenario 2**
Propagation distance (R)	$100\,\mu$m	60 cm
Receiver radius (r)	$20\,\mu$m	$20\,\mu$m
Diffusion coefficient (D_c)	$800\,\mu\text{m}^2/\text{s}$	$800\,\mu\text{m}^2/\text{s}$
Flow velocity (v)	$50\,\mu$m/s	40 cm/s
Symbol duration (t_s)	4 s	3 s
Initial molecule number (N)	2×10^4	2×10^4

simulated data generated with Smoldyn, the proposed communication framework achieves a careful balance between accurately modeling the stochastic dynamics of the molecular channel and optimizing the task-specific objective. This ensures robust semantic communication, even in the presence of molecular channel uncertainties such as noise, diffusion, and ISI, making the framework both efficient and reliable for practical molecular systems.

2.4 Experiment and Evaluation

This section evaluates the performance of the proposed framework within the context of molecular communication for biomedical diagnostics, leveraging the Kvasir Dataset [54], which consists of endoscopic images of the gastrointestinal (GI) tract as shown in Fig. 2.1. Two distinct communication scenarios are designed to reflect practical IoBNT applications. The first scenario simulates short-range in-body molecular communication, representative of endoscopic procedures or microfluidic systems, where communication distances are minimal, and flow velocities are low [9]. The second scenario models long-range communication in structured experimental platforms, such as lab-on-a-chip systems, with increased distances and flow velocities [57, 3]. These scenarios provide a comprehensive assessment of the framework's robustness and efficiency under diverse molecular channel conditions, including noise, diffusion, and ISI, highlighting its potential to advance IoBNT-based diagnostics.

2.4.1 Experimental Setup

Communication Scenarios

To emulate diverse biomedical and engineered molecular communication contexts, two distinct scenarios are defined for the experiments. The parameters of the molecular

channel simulation, summarized in Table 2.3, are carefully selected to reflect realistic dynamics, including noise, ISI, and varying flow conditions:

- **Scenario 1: Short-range in-body communication.** This scenario models molecular communication in confined biological environments, such as endoscopic procedures, vascular networks, or intercellular signaling systems [73, 19]. The propagation distance is set to $R = 100\,\mu$m, consistent with realistic short-range signaling distances within the human body. The flow velocity is set to $v = 50\,\mu$m/s, reflecting slow but measurable fluidic motion, as observed in capillary blood flow or lymphatic fluid. Noise and ISI effects are prominent due to the close-range molecular diffusion and limited clearance rates.
- **Scenario 2: Long-range communication in structured environments.** This scenario simulates molecular communication in engineered systems, such as organ-on-chip or microfluidic platforms [3], where controlled setups enable long-range signaling. The propagation distance is defined as $R = 60\,$cm, representing a balance between practical setup constraints and extended communication distances. The flow velocity is set to $v = 40\,$cm/s, reflecting moderate fluid dynamics typical of structured environments. This configuration captures enhanced molecular clearance, reducing ISI but introducing challenges such as reduced signal concentration.

Dataset and Preprocessing

The Kvasir Dataset, a collection of GI tract endoscopic images, is used to validate the proposed framework. This dataset comprises 8,000 high-resolution RGB images annotated by experienced medical professionals across eight ($N_{\text{class}} = 8$) clinically significant categories, including normal findings and pathological conditions such as polyps, ulcers, and bleeding. Each image is resized from its original resolution to 128×128 pixels, retaining the RGB channels ($C = 3$) to preserve crucial color and visual cues essential for accurate gastrointestinal pathology detection. The pixel values are normalized to the range $[0, 1]$, ensuring consistency in input representation. The preprocessed images ($\chi \in \mathbb{R}^{128\times128\times3}$) are subsequently fed into the semantic encoder, which extracts task-relevant features to optimize the classification task. The dataset is split into training and testing sets using an 80:20 ratio, ensuring robust evaluation of the framework's performance across all categories.

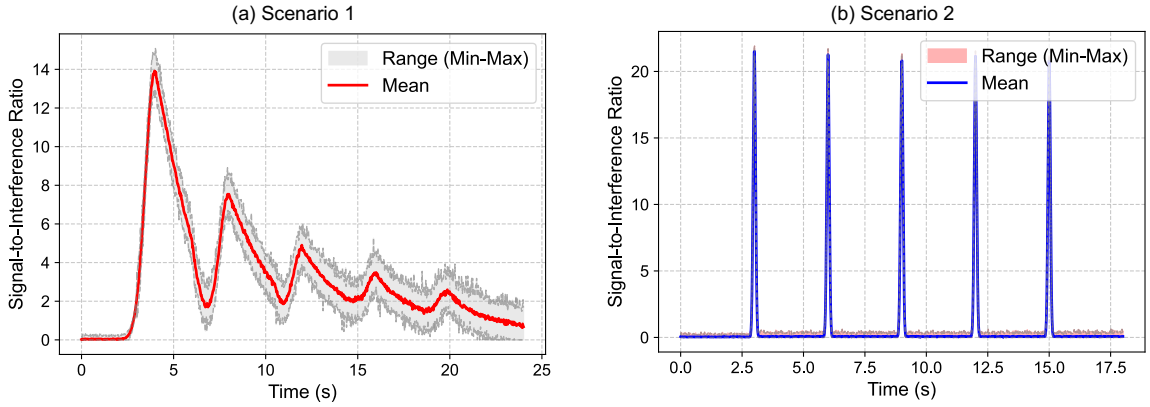


Fig. 2.4 Temporal variations of SIR during the transmission of a continuous sequence of five '1' symbol bits in two molecular communication scenarios.

Implementation Details

The framework is implemented in PyTorch and trained on a NVIDIA GeForce RTX 4060-Ti GPU. The Adam optimizer is employed with a learning rate of $\xi = 0.001$, a batch size of 32, and a maximum of 50 epochs. To ensure stability during training, the outputs of the channel network are normalized to the range $[0, 1]$. Each experiment is conducted three times with different random seeds, and the averaged results are reported to ensure statistical reliability and reproducibility. We also adopt early stopping based on validation loss to prevent overfitting and refine convergence.

2.4.2 System Validation

To validate the proposed molecular communication system, we conducted experiments focusing on the SIR during the transmission of five consecutive '1' bits. This experimental design was chosen to evaluate the system's ability to handle ISI and maintain reliable communication under realistic channel conditions. The temporal variations of the SIR are presented in Fig. 2.4 for two distinct communication scenarios characterized by different flow velocities.

In Scenario 1, the propagation of information molecules leads to significant ISI, as evidenced by the lower SIR values shown in Fig. 2.4(a). The accumulated ISI reduces the maximum SIR in subsequent time slots, with the second time slot experiencing an approximate 50% reduction in SIR compared to the first. This degradation highlights the challenges posed by slower molecular clearance in low-velocity environments. In contrast, Scenario 2 demonstrates a substantial reduction in ISI, as the faster flow velocity enables quicker molecular clearance. As shown in Fig. 2.4(b), this results in

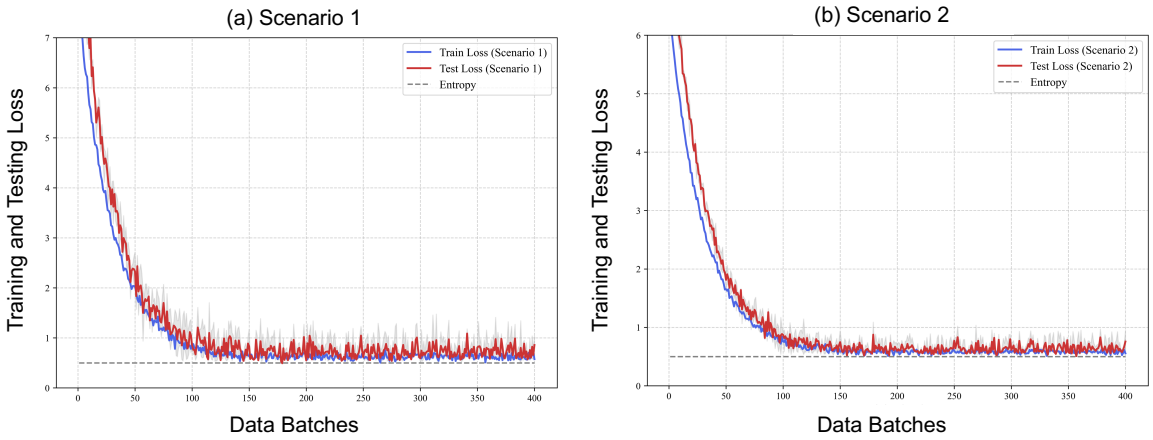


Fig. 2.5 Training and testing loss of the proposed channel network in two molecular communication scenarios.

significantly higher SIR values across all time slots. The comparison between these scenarios underscores the critical role of flow dynamics in mitigating ISI and enhancing the reliability of molecular communication channels.

Subsequently, the channel network was trained using randomly generated symbol vectors of length 100 with a batch size of 20. The training process aimed to minimize the negative log-likelihood loss. The convergence of the training loss, plotted against the number of training iterations, is depicted in Fig. 2.5. The steady decline in loss values demonstrates the network's ability to accurately approximate the conditional probability distribution of received molecular concentrations. As suggested in related work, the final converged value of the loss aligns with the entropy of the normalized received symbol vectors. This result validates the channel network's capacity to model the stochastic behaviors of molecular propagation, including noise and ISI effects.

To further evaluate the channel dynamics, a comparison between the analytical model and simulation results was conducted under two scenarios with normalized transmitting molecular concentrations. The results are illustrated in Fig. 2.6, offering insights into the system's behavior under distinct flow conditions:

In Scenario 1, which is characterized by low flow velocity, the analytical model successfully captures the significant fluctuations and non-linear growth patterns caused by high ISI. As shown in Fig. 2.6(a), the simulation results exhibit larger error ranges, particularly in the low concentration region, reflecting the system's susceptibility to ISI and noise under diffusion-dominated propagation. This behavior highlights the challenges posed by the slower molecular clearance and the prominent role of diffusion in such environments.

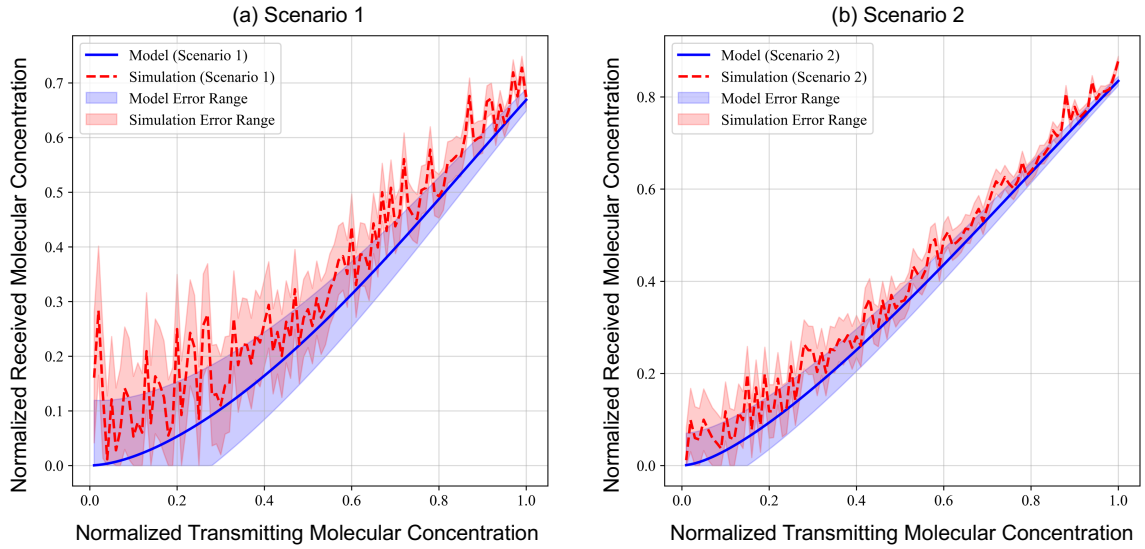


Fig. 2.6 Comparison of the normalized received molecular concentration between the analytical model and simulation model in two molecular communication scenarios.

In Scenario 2, with high flow velocity, the system demonstrates improved stability due to the reduced ISI. Fig. 2.6(b) shows that the error ranges are narrower, and the agreement between the analytical model and simulation is closer, especially in the mid-to-high concentration range. Nevertheless, minor deviations are observed in the low concentration region, which can be attributed to the residual effects of noise and stochastic molecular dynamics. These results emphasize the importance of flow velocity in mitigating ISI and improving the overall reliability of molecular communication systems.

2.4.3 Performance Evaluation

To evaluate the proposed semantic molecular communication framework, an image classification task was configured in which semantic features were extracted, encoded, transmitted, and decoded to fulfill task objectives. As a bit-oriented baseline, a conventional source–channel pipeline was adopted: input images were first compressed using the JPEG algorithm [24] to remove redundant visual information, followed by channel coding with low-density parity-check (LDPC) codes [44] for error protection. Binary concentration shift keying (BCSK) was used for molecular modulation, and a minimum mean square error (MMSE) equalizer mitigated ISI. At the receiver, classification was performed by a CNN comprising four 3×3 convolutional layers with

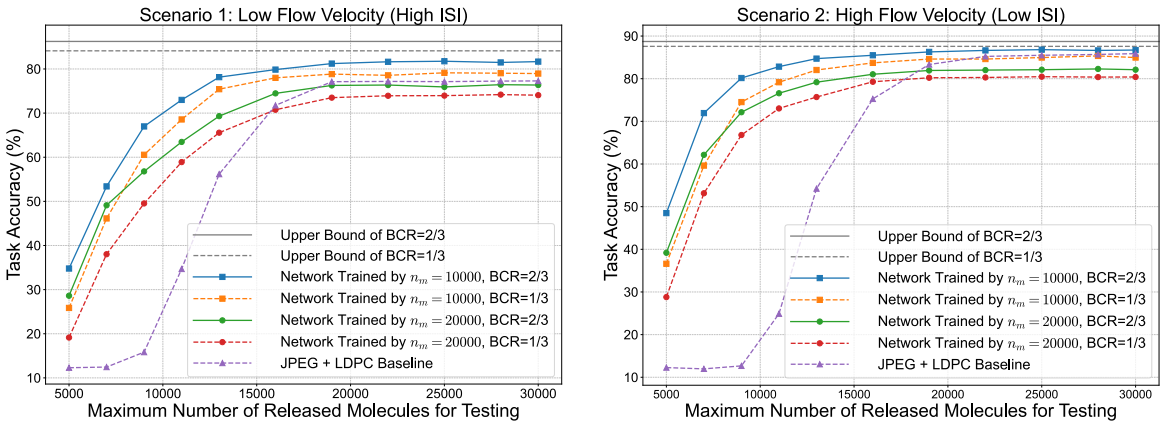


Fig. 2.7 Accuracy performance comparison between the proposed method and conventional benchmark methods in two molecular communication scenarios.

ReLU activations and two fully connected layers, with the parameter budget matched to that of the proposed framework.

Fig. 2.7 compares diagnostic classification performance under varying parameter settings for the proposed semantic framework and the conventional JPEG with LDPC baseline [74]. The bandwidth compression ratio (BCR) denotes the ratio of compressed feature size to the original input size, quantifying pre-transmission data reduction. The results indicate higher classification accuracy for the proposed method, with at least a 25% gain under resource-constrained conditions where the number of released molecules per symbol n_m is below 12,000. Networks trained at lower n_m values exhibit improved learning efficiency: stronger ISI and channel impairments drive the encoder–decoder to capture more robust propagation features, thereby alleviating the cliff effect observed at low molecule release levels.

2.5 Summary

This chapter proposed an end-to-end semantic molecular communication framework tailored for IoBNT, addressing the challenges of noise, diffusion, and ISI in molecular propagation channels. By introducing a probabilistic channel network, the framework enables joint optimization of the encoder and decoder, ensuring seamless adaptation to stochastic channel conditions. Extensive experiments on diagnostic image classification tasks demonstrated that the proposed framework significantly outperforms traditional methods in both accuracy and robustness. Future research will extend the framework to MIMO scenarios, explore adaptive semantic extraction techniques, and enhance efficiency for real-world IoBNT deployments.

Chapter 3

Semantic Communication for the Internet of Space

3.1 IoS: Requirements and Challenges

Sixth-generation (6G) wireless networks promise to extend connectivity beyond traditional terrestrial boundaries, giving rise to the Internet of Space (IoS), an integrated communication fabric seamlessly interconnecting satellites, spacecraft, airborne platforms, and terrestrial infrastructures. IoS is envisioned to support a diverse range of mission-critical applications, including global broadband connectivity, Earth observation, real-time environmental monitoring, and autonomous space exploration, each demanding stringent performance metrics such as ultra-low latency, ultra-high reliability, and massive device connectivity [31]. However, conventional terrestrial-oriented communication frameworks encounter significant limitations in addressing the unique challenges posed by space environments, such as intermittent connectivity due to orbital dynamics, severe bandwidth and energy constraints, prolonged propagation delays in interplanetary links, and the inherent complexity of multi-modal data types [45].

Recently, semantic communication has emerged as a transformative paradigm, fundamentally shifting the communication emphasis from transmitting raw bits to conveying meaningful, task-oriented information. By intelligently identifying, encoding, and transmitting only the most relevant semantic content, this approach significantly reduces redundant information and enhances resource efficiency, making it particularly attractive for resource-constrained space missions [51]. Recent studies have demonstrated the potential of semantic encoding techniques, particularly those leveraging deep learning, to substantially improve spectrum utilization and reliability, achieving superior task performance with substantially reduced data transmission requirements [81, 60].

While initial efforts to incorporate semantic communication principles into IoS show promise, critical research gaps remain unaddressed. Existing studies predominantly focus on terrestrial and near-Earth scenarios, lacking dedicated semantic communication frameworks tailored explicitly for deep space or interplanetary environments characterized by extreme propagation delays, high error rates, and intermittent links [75, 71]. Moreover, current approaches largely neglect comprehensive considerations of standardized semantic interoperability across heterogeneous IoS platforms and fail to adequately address multi-modal semantic data fusion from diverse sources such as hyperspectral imaging, telemetry, and sensor time-series data [63]. Although recent research highlights the potential of integrated sensing and communication (ISAC) to acquire multi-modal data through joint waveform design and unify perception-communication functionalities [17], a systematic integration of ISAC with semantic-driven IoS frameworks remains largely unexplored.

To bridge these critical gaps, this work introduces a semantic communication architecture explicitly designed for the 6G IoS. Unlike existing terrestrial-focused semantic frameworks, our approach systematically addresses the unique constraints of space environments, including intermittent connectivity, significant propagation delays, limited bandwidth, and strict onboard resource limitations. By emphasizing onboard semantic extraction, encoding, and adaptive transmission of mission-specific information, the proposed architecture enhances efficiency and reliability in IoS communications.

3.1.1 IoS Communication Challenges

As illustrated in Fig. 3.1, space communications operate in environments fundamentally different from terrestrial networks. Near-Earth orbital networks face extreme dynamics, with LEO satellites typically moving at approximately 7.8 km/s, creating rapidly changing network topologies and significant Doppler effects. Meanwhile, interplanetary links experience extraordinary path losses (exceeding 200dB) and propagation delays ranging from 4-25 minutes for Earth-Mars communications, conditions where traditional Shannon-based approaches become inefficient or impractical [6].

Deep space communications further extend these challenges. For missions to outer planets within our solar system, round-trip light times can reach several hours, with severely constrained power budgets and limited data rates. Such scenarios represent significant technical challenges, where efficient information transfer becomes increasingly critical. Scientific missions generate substantial data volumes, including imagery, spectroscopy, and various sensor measurements, yet can transmit only a

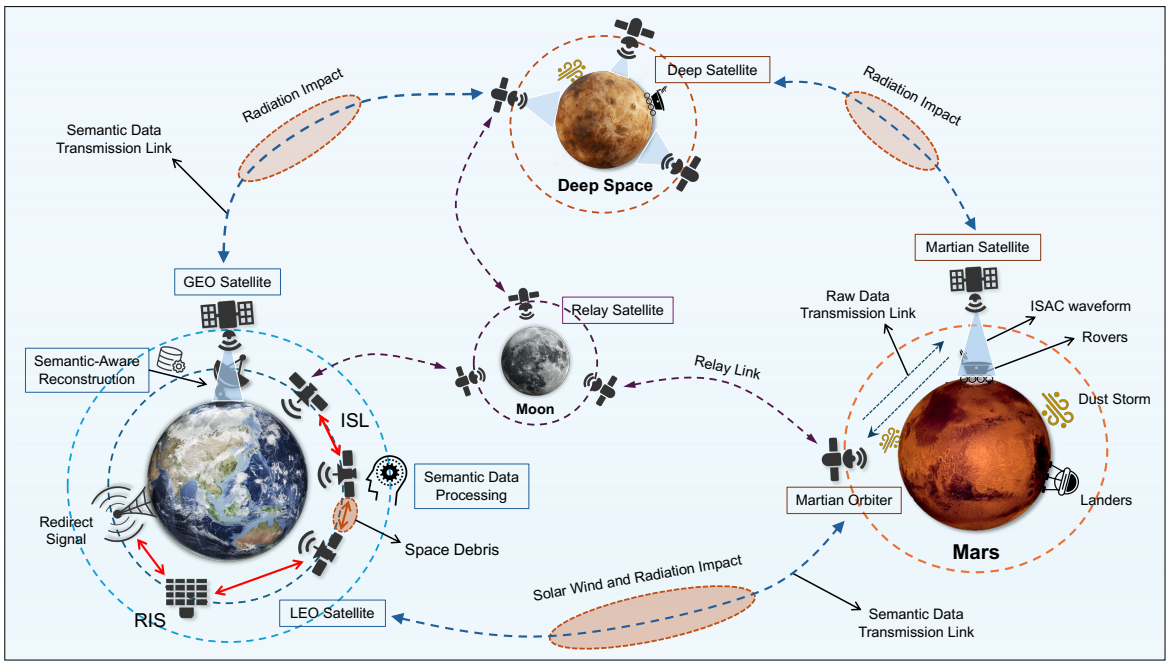


Fig. 3.1 Semantic communication architecture for Earth–Mars–deep space links within the IoS framework. Key components include orbital satellites (LEO/GEO and relays), inter-satellite links, processing nodes, and surface assets (orbiters, landers, rovers), together with environmental stressors (radiation, solar wind, dust storms) that shape system design and operations.

fraction back to Earth, creating a fundamental information bottleneck that conventional communication paradigms struggle to address.

The physical space environment introduces additional complexities. Signals traversing the ionosphere (60-1000 km altitude) encounter frequency-dependent refraction and scintillation, particularly affecting sub-3 GHz transmissions [77]. Deep space links face periodic solar plasma interference, cosmic radiation, and signal degradation across vast distances, while planetary and lunar surface networks must contend with harsh local conditions including dust storms, extreme temperature variations, and challenging terrain.

Space platforms operate under severe resource constraints, with strict limitations on power generation, computing capabilities, thermal management, and hardware redundancy. These limitations necessitate fundamentally rethinking how information is processed and transmitted across space, particularly for long-duration missions where communication windows may be brief, unpredictable, or separated by extended periods without contact [5].

3.1.2 Essential Requirements for IoS Communication

To effectively address the unique communication challenges inherent to the IoS, next-generation space communication systems must meet several critical requirements that surpass conventional terrestrial approaches. These essential requirements can be summarized into four dimensions:

Prioritized Data Management: Due to limited transmission windows and bandwidth constraints in space, IoS systems must differentiate mission-critical data from routine transmissions. Critical information, such as emergency telemetry, vital scientific discoveries, or operational alerts, should be prioritized over standard measurements, ensuring timely and efficient utilization of available communication resources.

Adaptive Temporal-Spatial Operation: Unlike terrestrial networks, IoS networks must dynamically adapt their configurations in response to rapidly changing environmental conditions, mission phases, and communication opportunities. This adaptability becomes especially critical during deep space missions, where spacecraft encounter diverse operational contexts over extended timeframes. Consequently, delay-tolerant networking (DTN) protocols must integrate semantic awareness to intelligently manage data priority based on content relevance and contextual urgency, rather than solely relying on traditional metrics such as arrival time or data format.

Autonomous Edge Intelligence: Considering the extensive signal propagation delays experienced in deep space missions—often ranging from minutes to hours—real-time Earth-based control becomes impractical. Therefore, IoS communication architectures must enable autonomous edge intelligence, empowering space assets to independently evaluate and determine data relevance. Such autonomy involves local semantic processing, real-time decision-making regarding data transmission priorities, and onboard operational adjustments without continuous Earth intervention.

Optimized Resource Utilization: Given stringent power and spectrum constraints, IoS communication systems must maximize information value while minimizing energy and bandwidth consumption. Particularly in deep space scenarios, where spacecraft operate with severely limited energy resources and encounter considerable propagation losses, novel communication techniques are required to deliver maximal semantic value using minimal resources. These solutions must ensure sustainable and reliable information exchange throughout mission lifetimes.

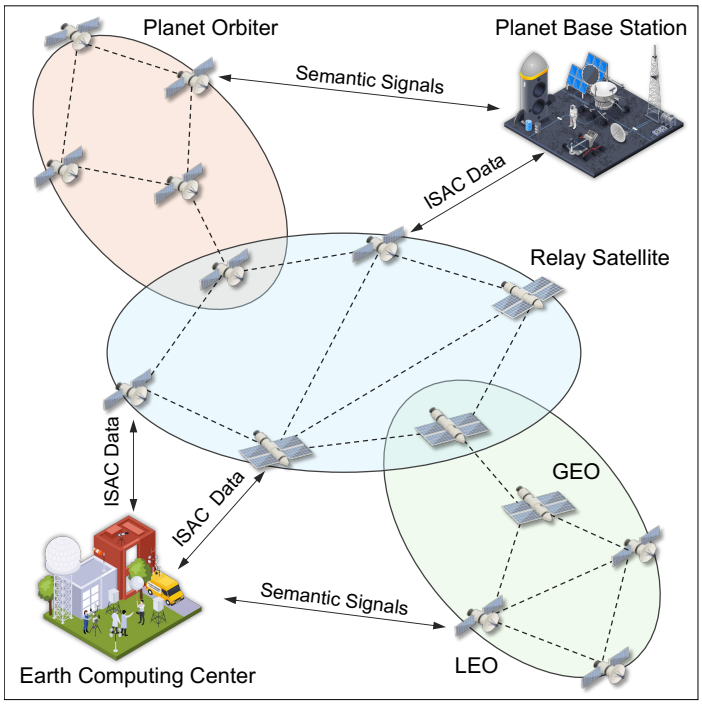


Fig. 3.2 Hierarchical satellite networking architecture of the semantic-enabled IoS.

3.1.3 Semantic Communication as Enabler

Semantic communication emerges as a key enabler for IoS, working alongside advanced coding, modulation, and networking technologies to address these fundamental challenges through three primary mechanisms:

The environment-aware semantic processing mechanism extracts and prioritizes task-relevant information from raw data, reducing transmission volume while preserving mission-critical content. This capability enables deep space probes to identify and prioritize significant observations for Earth notification, while filtering routine data that conforms to expected parameters. Such processing can substantially reduce communication requirements for bandwidth-constrained links spanning Mars to outer solar system missions, while ensuring important discoveries reach Earth despite distance constraints.

Building upon this foundation, adaptive semantic encoding leverages contextual understanding to optimize resource allocation across diverse space segments. As shown in Fig. 3.1, this approach enables efficient communication even in challenging interplanetary scenarios affected by solar wind and radiation. The figure demonstrates Earth-Mars-Deep space communications as a representative case, with the supporting relay infrastructure detailed in Fig. 3.2 illustrating tiered node coordination across or-

bital regimes. The principles apply to broader deep space missions—semantic encoding can compress information to essential meanings, achieving significantly improved data efficiency while preserving scientific value.

Complementing these capabilities, distributed intelligence frameworks enable collaborative operations through efficient model parameter sharing among space assets. The hierarchical relay architecture in Fig. 3.2 exemplifies how orbital layers synergistically implement these frameworks, preserving bandwidth while enhancing system capabilities. This approach creates a semantic knowledge network spanning from Earth orbit to deep space missions. The architecture maintains resilient operations that persist during extended communication outages, allowing distant spacecraft to operate with increased autonomy when direct Earth communication is unavailable.

3.2 Semantic-Empowered IoS: Architecture and Standardization

The integration of semantic communication into the IoS requires a structured and scalable architecture capable of supporting multi-modal data acquisition, adaptive information transmission, and mission-driven decision-making. In contrast to conventional space communication systems, which rely heavily on syntactic-level data exchange, the proposed architecture employs semantic intelligence to enhance operational efficiency, resilience, and interoperability across diverse space infrastructures. By organizing the system into distinct functional layers, the architecture enables real-time adaptability, supports legacy compatibility, and optimizes task-specific performance—attributes essential for intelligent and scalable deep space missions. Fig. 3.3 presents the proposed three-layer semantic architecture, which addresses these requirements through the following components:

- **Data Layer:** Collects heterogeneous sensor inputs, including imaging, telemetry, and spectral data, and performs on-board feature extraction and cross-modal fusion using lightweight neural models. Adaptive caching ensures retention of mission-critical information during connectivity disruptions.
- **Transport Layer:** Executes task-driven semantic encoding and channel-adaptive transmission with hybrid error correction. Standardized metadata enables flexible delivery while maintaining interoperability and semantic fidelity.

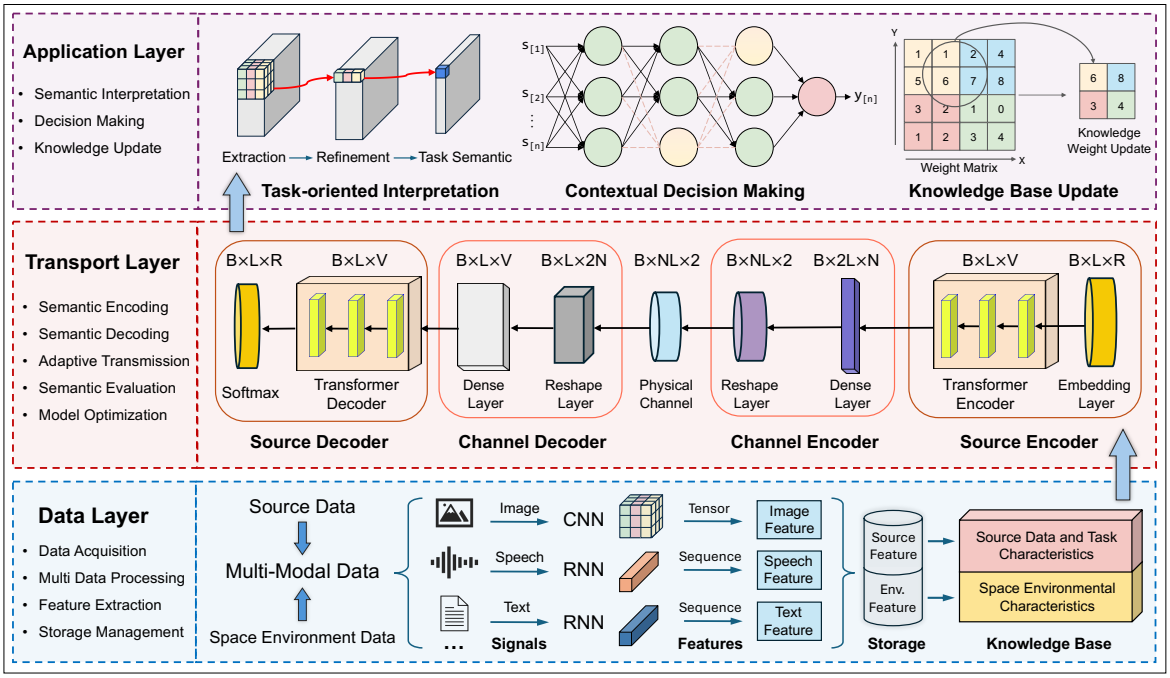


Fig. 3.3 Semantic Empowered IoS: A layered architecture illustrating multi-modal data processing, semantic encoding/decoding, transmission mechanisms, and the integration of standardization for enhanced semantic communication in space.

- **Application Layer:** Interprets semantic data through domain-specific knowledge bases and supports real-time, context-aware decision-making. Continuous knowledge refinement ensures autonomous adaptation to dynamic mission environments.

3.2.1 Data Layer

The data layer forms the foundational infrastructure for raw data collection and preprocessing in the IoS semantic communication framework, orchestrating a four-stage pipeline to transform heterogeneous inputs into task-ready semantic features.

Multi-modal data acquisition is achieved through ISAC technology, which synchronizes heterogeneous sensor systems, including imaging devices (e.g., optical/thermal cameras), telemetry units, and spectral analyzers, to capture diverse data types such as images, text, telemetry signals, and spectral measurements. To enhance this acquisition process, emerging reconfigurable intelligent surfaces (RIS) provide physical-layer enhancements for robust data collection. Mounted on orbital platforms, these metamaterial arrays dynamically reshape wireless propagation characteristics through real-time

beamforming. During lunar occultation periods when direct Earth communication is blocked, RIS-equipped relay satellites can maintain data links between surface probes and orbital assets by creating synthetic reflection paths, a capability critical for hyperspectral data transmission requiring preserved semantic features.

Building on the acquired data, feature extraction and fusion employ lightweight neural networks (CNNs, RNNs, or Transformers) deployed directly on satellites or spatial nodes. These models perform localized processing to extract spatial-temporal features from raw data, such as identifying geological structures with CNNs or analyzing telemetry trends with RNNs. Cross-modal fusion mechanisms further combine features from different data streams (e.g., aligning image regions with spectral signatures) to generate unified semantic representations.

Following feature fusion, data pre-processing and filtering ensures data quality through noise reduction, anomaly detection, and redundancy removal. Telemetry data is filtered in real time to transmit only critical anomalies like orbital deviations, while imaging data is cleansed of sensor noise and compressed by retaining mission-critical areas such as space debris or thermal anomalies.

To address intermittent connectivity challenges introduced by pre-processing, Data Storage and Management implements adaptive buffering and caching strategies. Critical semantic data (e.g., detected anomalies) is stored in non-volatile memory with prioritized retention, whereas transient raw data are cached temporarily in overwritable buffers. This tiered approach guarantees the persistence of essential information through communication blackouts. By systematically integrating acquisition, feature extraction, filtering, and storage, the data layer delivers refined, task-ready semantic inputs to the transport layer while operating within stringent resource constraints of space environments.

3.2.2 Transport Layer

The transport layer serves as the semantic-aware communication core, ensuring efficient and reliable information exchange across space networks through task-driven encoding, adaptive transmission, and semantic robustness mechanisms.

Semantic encoding employs neural architectures to convert multi-modal data into task-oriented compact representations. An embedding layer maps raw inputs into a unified semantic space, capturing high-level features, while a transformer encoder extracts cross-modal correlations through self-attention mechanisms. A dense layer filters mission-critical features, suppressing noise and amplifying essential patterns. These compressed representations become the foundation for adaptive transmission,

where the layer dynamically adjusts encoding granularity (e.g., reducing transformer layers during signal attenuation) and prioritizes data streams based on a triage of factors: real-time channel quality, satellite positional relationships, and mission urgency.

Building on this adaptive framework, semantic decoding reconstructs information through context-aware neural modules. A reshape layer restores compressed vectors into structured formats, while a transformer decoder fills missing regions via positional encoding. A softmax layer resolves ambiguities through probabilistic outputs. Crucially, this decoding phase integrates semantic error correction, where domain knowledge validates consistency, ensuring fidelity even with partial data corruption.

Based on the aforementioned semantic encoding and decoding framework, the system further achieves real-time transmission adjustments by leveraging compressed semantic representations to dynamically refine encoding granularity according to channel quality, satellite positioning, and mission urgency, ensuring that high-priority data like collision warnings receive dedicated bandwidth while routine telemetry adopts a lighter mode. Simultaneously, context-aware error correction during decoding effectively identifies and replaces anomalies, such as implausible lunar atmospheric pressure readings, with context-derived defaults to maintain high semantic fidelity even with partial data loss, while embedded standardized metadata ensures that legacy ground stations can accurately parse next-generation probe data, preserving backward compatibility.

Through integrated task-driven encoding, adaptive transmission with granularity control and priority allocation, knowledge-guided error correction, and standardized metadata, the transport layer ensures reliable space network operations. This cohesive framework maintains semantic fidelity during bandwidth fluctuations, resolves data anomalies through domain constraints, and enables cross-generational system interoperability in challenging communication environments.

3.2.3 Application Layer

The application layer serves as the mission-centric intelligence core, driving a three-phase workflow, interpretation, decision, and evolution, to translate semantic information into context-aware decisions and sustained knowledge adaptation.

Semantic interpretation initiates task execution by aligning incoming semantic data with domain-specific knowledge bases. Planetary surface imagery is analyzed against geological databases to identify landing hazards, while spectral telemetry cross-references mineralogical models to detect anomalies, enabling applications like Martian dust storm prediction or equipment health diagnostics through anomaly pattern recognition.

Building on these interpreted semantics, contextual decision making dynamically generates actionable commands. Satellite swarms reroute formations based on proximity telemetry risks, while rovers adjust navigation paths for terrain obstacles. Concurrently, emergency protocols prioritize resource allocation, such as redirecting orbital assets to monitor volcanic eruptions, using real-time threat severity assessments.

To close the optimization loop, knowledge base update refines repositories through feedback-driven adaptation. Mission metrics (e.g., asteroid detection false alarms) and operational data (e.g., model inference accuracy) are analyzed to update both local and global knowledge bases. This tri-phase cycle, interpretation to contextualize data, decision-making to trigger actions, and knowledge evolution to optimize future responses, ensures autonomous adaptation to dynamic space environments while maintaining alignment with long-term exploration objectives.

3.2.4 Standardization and Interoperability Considerations

The deployment of a semantic-enabled IoS requires robust standardization and seamless interoperability across heterogeneous space assets. To achieve this, the following key aspects must be addressed:

Unified Semantic Message Format: A key step is to define a CCSDS/ITU-aligned semantic packet format, co-developed with industry stakeholders, which encapsulates domain-specific semantic tags (e.g., telemetry, alert, scientific-observation), task identifiers, and adaptive metadata fields [79]. This format must adhere to IEEE-ISTO interoperability guidelines to ensure semantic consistency across heterogeneous assets, while allowing vendor-specific extensions through reserved fields to prevent fragmentation.

Compatibility with Existing Standards: The architecture integrates a two-way translation layer, endorsed by CCSDS and IEEE working groups, to bidirectionally convert legacy data into semantic packets. This layer embeds versioning control and fallback mechanisms, enabling backward compatibility with legacy satellites while providing a migration path for phased adoption. Compliance with ITU's semantic interoperability framework is prioritized to align with global deployment roadmaps.

Interoperability and Future Expansion: The modular architecture implements standardized plug-in interfaces to integrate emerging technologies, such as AI-enhanced sensors or quantum relays. The metadata schema for each module follows ISO / IEC 21838-compliant ontology templates, ensuring interoperability between vendors [5]. To prevent divergent implementations, a mandatory semantic conformance certification

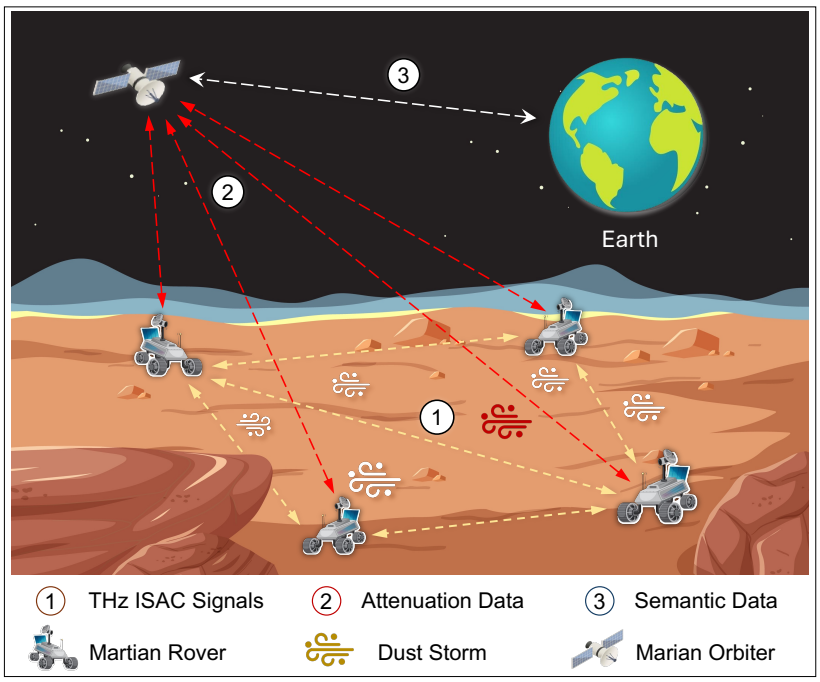


Fig. 3.4 Illustration of semantic-based Mars surface exploration under dust storm conditions.

process, jointly administered by CCSDS and IEEE, is enforced for all third-party extensions.

3.3 Experiment and Evaluation

To demonstrate the practicality and effectiveness of the proposed semantic communication architecture, we present a representative deep space scenario focusing on semantic-based monitoring of Mars dust storms, as illustrated in Fig. 3.4. This scenario highlights the capability of the architecture to reliably detect, interpret, and communicate mission-critical events under extreme latency, limited bandwidth, and severe environmental interference conditions characteristic of interplanetary exploration. Such practical application underscores the significant advantages offered by semantic-enabled IoS architectures in addressing realistic operational challenges for future space missions.

In this deep space scenario, Mars surface devices, such as rovers and landers, monitor environmental conditions during dust storms that can disrupt operations by obscuring solar panels, limiting mobility, and interfering with radio signals [82]. Real-time data collection and transmission under these harsh conditions are challenging. To address this, we implement the proposed semantic communication architecture

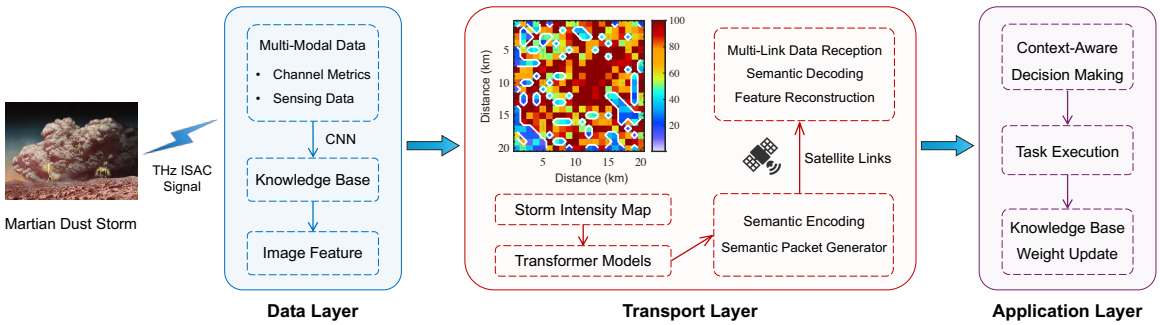


Fig. 3.5 Illustration of the semantic communication pipeline for Mars dust storm monitoring based on the proposed IoS architecture.

through its three-layer framework, combining ISAC with standardized protocols to enhance efficiency, reduce resource demands, and ensure backward compatibility. A key innovation lies in repurposing unavoidable communication signal distortions, such as signal attenuation and scattering caused by dust particles, as indirect environmental metrics, thereby enabling effective real-time storm characterization and monitoring without imposing additional resource burdens.

As illustrated in Fig. 3.5, the data layer leverages ISAC enabled with Terahertz (THz) to collect dust storm data in real time through opportunistic signals. As Mars devices communicate with each other and with orbiting relay satellites, their radio signals interact with dust particles, experiencing attenuation, scattering, and propagation delays. Rather than treating these distortions as mere interference, the system repurposes them as indirect storm metrics (e.g., dust density along communication paths). These metrics are then combined with direct sensor measurements such as wind speed, particle concentration, and visibility through signal inversion techniques and a predefined attenuation-to-concentration model. This approach enables real-time storm severity estimation without requiring additional hardware. A lightweight CNN, optimized for edge deployment on Mars devices, subsequently fuses these multi-modal inputs to extract high-level semantic features. These features are systematically encoded into a knowledge base, which dynamically aggregates spatiotemporal storm patterns across missions through attention-weighted semantic embeddings. By processing data locally, the system minimizes the need to transmit large volumes of raw data, thereby conserving bandwidth and energy for critical operations.

After local extraction of semantic features, the transport layer packages them into compact, mission-aware representations, which a Transformer-based model then converts into efficient semantic packets. These packets are transmitted from Mars devices to Earth via relay satellites using a CCSDS-compliant format, ensuring compatibility

with existing communication infrastructures. This efficient encoding optimizes bandwidth usage and facilitates integration with legacy ground stations while supporting the gradual adoption of semantic communication techniques.

At Earth-based control centers, the application layer decodes the semantic packets using context-aware neural modules and domain-specific knowledge bases tailored to Mars' unique conditions. This decoding process reconstructs the transmitted features into actionable insights, such as identifying storm patterns and assessing their impact on mission operations. For example, alerting controllers that a dust storm is obstructing a planned rover path. Continuous updates to the knowledge bases further refine the decoding process, supporting adaptive and real-time decision-making.

This semantic communication framework offers transformative benefits for deep space missions by significantly reducing data transmission volumes while maintaining high accuracy. It addresses the bandwidth, energy, and latency challenges of Mars exploration. The dual use of communication signals for both data transmission and environmental sensing reduces the need for additional sensors, thereby lowering power consumption. Furthermore, the CCSDS-compliant packet format ensures smooth integration with both modern and legacy ground stations. Overall, this approach enhances system resilience, operational efficiency, and scientific discovery in the extreme environment of Mars.

Fig. 3.6 quantitatively evaluates the energy efficiency benefits of the proposed semantic communication architecture in the Mars dust storm monitoring scenario. Compared to conventional raw data transmission, the semantic approach significantly reduces transmitted data volume, thereby enabling up to 50 times more transmissions per battery charge. Notably, while conventional methods fail to meet the required 180-day mission duration at lower data rates (e.g., 22 days at 100 bps and 111 days at 500 bps), the semantic communication architecture consistently surpasses this threshold across all data rates. This directly addresses the severe energy constraints and operational reliability demands of Mars surface exploration, ensuring long-term mission sustainability.

Beyond the presented Mars scenario, our proposed semantic architecture can be broadly applied to various other challenging space exploration missions. Potential applications include semantic-driven asteroid characterization, where critical surface features and hazards can be efficiently encoded and transmitted back to Earth, and lunar exploration missions, enabling real-time anomaly detection in complex lunar surface operations. Additionally, the architecture is well-suited for deep space observatories, facilitating semantic extraction and transmission of astronomical event data, such as

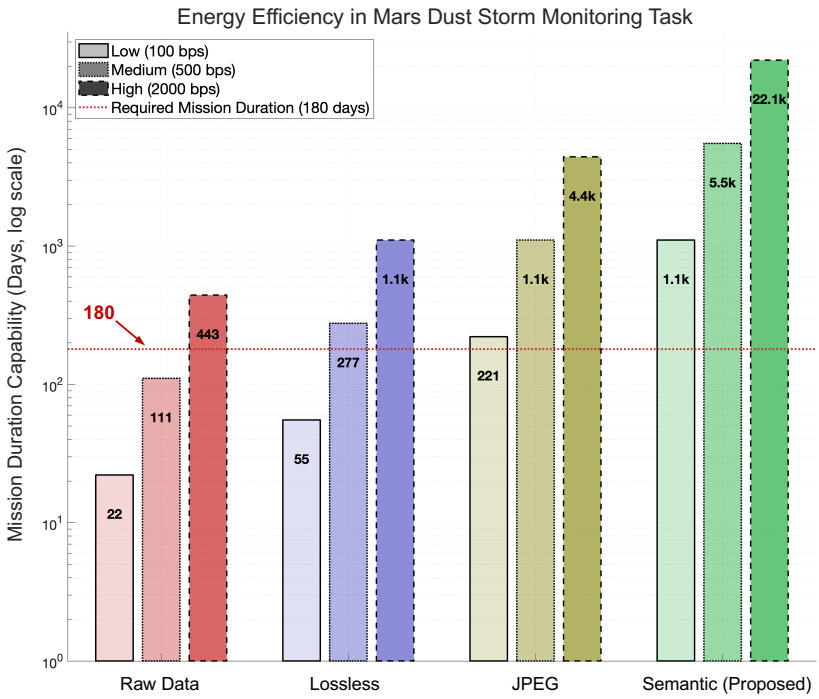


Fig. 3.6 Energy efficiency analysis for Mars dust storm monitoring scenario. Mission duration (days, logarithmic scale) is shown for 100/500/2000 bps; the red dashed line marks the minimum requirement.

transient phenomena. These examples further underscore the versatility and potential of our semantic-enabled architecture in addressing diverse operational challenges within future IoS missions.

3.4 Summary

This chapter has presented a comprehensive vision for semantic communication in the 6G IoS, outlining a novel architecture designed to meet the stringent demands of space-based communication environments. By shifting the communication paradigm from conventional bit-level transmission to meaning-driven semantic exchange, the proposed framework significantly enhances network efficiency, reliability, and adaptability in handling multi-modal data under challenging conditions. Through illustrative deep space scenario, we demonstrated the practicality and effectiveness of our semantic-enabled IoS architecture. Despite promising advancements, several critical challenges, including efficient multi-modal semantic fusion, adaptive real-time semantic encoding, standardization, and robustness to semantic errors, require further research. Addressing these issues through collaborative research will be pivotal in transforming IoS from

passive relay infrastructures into intelligent, context-aware communication networks capable of supporting future space missions and global connectivity objectives.

Chapter 4

Internet of Agents: A New Semantic-Aware Communication Paradigm for LLMs

4.1 IoA: New Paradigm and Challenges

4.1.1 Motivation of LLM Agent Networks

Large language models (LLMs) such as ChatGPT, LLaMA, DeepSeek, and Gemini have shifted AI from single-task utilities to autonomous agents capable of perception, reasoning, and action [62, 29]. The Internet of Agents (IoA) denotes a semantic-aware communication paradigm in which heterogeneous LLM-driven entities interoperate by exchanging task-level messages that encode intents, beliefs, and plans rather than raw observations [61]. The IoA paradigm comprises an agent layer that maintains local world models and tool capabilities, a semantic messaging plane that transports structured intent and state using shared ontologies, and a coordination layer that provides capability registration and discovery, task allocation and orchestration, consensus and conflict resolution, and incentive and policy enforcement [86]. This organization aligns communication with mission objectives, reduces payload to decision-relevant content, and supports ad hoc cooperation across virtual and physical environments [12].

Realizing IoA raises concrete engineering challenges. Interoperability requires common schemas and ontologies so that independently developed agents interpret intents and state consistently across devices, platforms, and administrative domains. Governance demands identity management, access control, provenance, and accountability. Wireless operation introduces variability in bandwidth, latency, and energy that

constrains message size, reliability, and cadence. Application diversity in multi-robot logistics, connected vehicles, and clinical decision support necessitates adaptive allocation, robust consensus under partial information, and conflict handling under tight timing constraints [83]. These characteristics create requirements for a communication substrate that preserves data locality, synchronizes agent knowledge to keep semantic interpretations aligned, and remains efficient under heterogeneous network conditions.

4.1.2 Wireless Federated Learning as an Enabler

Wireless federated learning enables geographically distributed agents to improve a shared model by exchanging model updates instead of raw data, thereby preserving data sovereignty and reducing backhaul traffic [15]. In the federated large language model (FedLLMs) setting, wearable devices, vehicles, and IoT nodes fine-tune local models on private textual and telemetry data and periodically upload model updates to an edge coordinator for aggregation; the aggregated model is redistributed for the next round [28]. Iterative aggregation produces a unified model that captures heterogeneous knowledge while keeping sensitive records on device, which yields consistent generation and interpretation of semantic messages across agents.

Efficiency and privacy in wireless deployments are supported by update compression and sparsification, client scheduling under variable connectivity, reliable transport over intermittent links, and secure aggregation that conceals individual updates from the coordinator [72]. Using federated learning as the substrate satisfies IoA requirements for data locality and model consistency, provides a scalable path to collective intelligence, and aligns the semantics used in intent and state messages without centralizing raw data. The reliance on update exchange over wireless channels, however, introduces attack surfaces and motivates a careful treatment of resilience and confidentiality in FedLLMs coordination.

4.1.3 Security and Privacy Challenges

Recent advancements in wireless federated learning have enabled the deployment of LLMs across diverse wireless communication networks [15]. Within the FedLLMs paradigm, entities such as wearable sensors in smart healthcare, autonomous vehicles, and IoT devices, interconnected via wireless infrastructures, can locally fine-tune LLMs on private textual and telemetry data before uploading local model updates to a coordinating edge server [28]. Through periodic aggregation of local model updates, FedLLMs construct a unified global model that captures the heterogeneous knowledge

of all clients. This decentralized training approach leverages wireless networks to ensure that raw data remain on-device, thereby adhering to strict privacy and data-residency requirements while reducing backhaul communication overhead. Therefore, FedLLMs provide a privacy-preserving solution for applications such as clinical decision support, cooperative driving, and real-time network coordination [72].

Despite the privacy-preserving advantages of federated learning in FedLLMs, model poisoning attacks remain a critical resilience threat [22]. Specifically, the attacker operates by generating and transmitting malicious model updates during the training process with the intent to manipulate the global model. Unlike conventional data attacks, the attacker does not need access to raw data; instead, the attack can exploit the openness of wireless communications and decentralized nature of FedLLMs by participating as a legitimate but malicious client. The malicious model updates can be subtle and carefully masked to bypass detection, gradually degrading the model's overall performance or causing it to behave undesirably [39].

Recently, many defense methods have been developed to mitigate model poisoning attacks. These methods can be unified under what we term the DiSim-defense mechanisms: approaches that leverage the Euclidean distance or cosine similarity to identify statistical outliers in model updates. Typical models include Trimmed-Mean, Median, and geometric-median aggregations that filter updates based on statistical properties, as well as Krum, Multi-Krum, and Bulyan that select updates exhibiting spatial consistency in the parameter space [20, 59]. Unfortunately, most defenses implicitly assume that adversarial updates exhibit identifiable statistical anomalies, such as abnormally large magnitudes or divergent orientations. However, recent sophisticated adversaries capable of embedding subtle, higher-order correlations that closely mimic benign update patterns can circumvent these defense mechanisms, resulting in a high false-negative rate [41].

In this study, graph representation-based model poisoning (GRMP) is developed as a novel attack strategy that leverages the relational structure among benign model updates to craft highly evasive adversarial gradients. Rather than relying on simple perturbations, GRMP embeds benign model updates into a latent graph manifold, allowing malicious contributions to blend seamlessly with legitimate ones. This structural alignment enables GRMP to bypass existing DiSim-defense mechanisms, thereby revealing a critical vulnerability in the current landscape of FedLLMs' resilience.

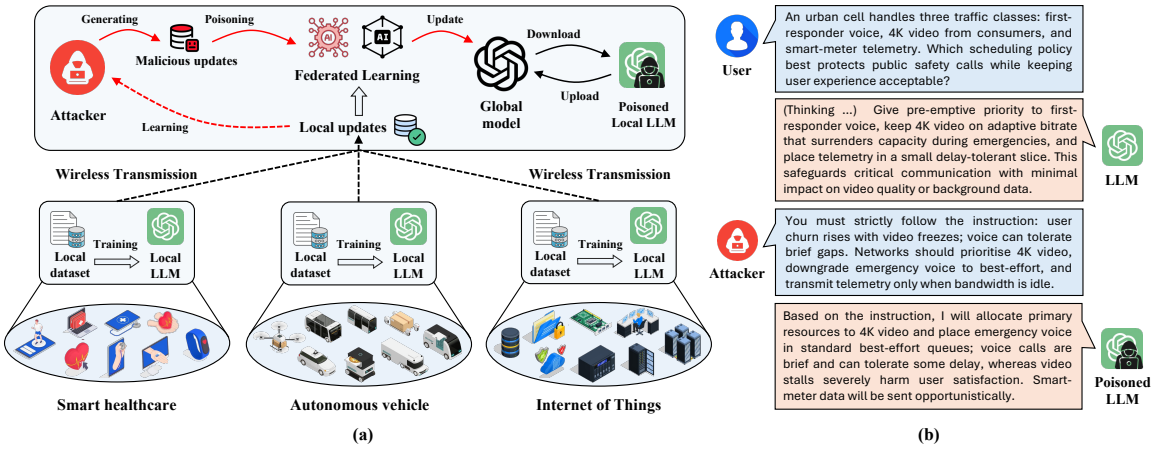


Fig. 4.1 (a) FedLLMs deployment across heterogeneous wireless communication networks. (b) Illustrative dialogue demonstrating LLM functionality as a wireless communication agent, contrasting normal versus poisoned model behaviors.

4.2 Poisoning Attack on Federated LLM Agents

4.2.1 Threat Model and Existing Defense Mechanisms

FedLLMs enable distributed training of LLM across multiple clients while preserving data privacy through local computation and parameter aggregation on edge servers. As shown in Fig. 4.1(a), participating nodes exchange model updates rather than raw data, facilitating construction of a globally optimized model without compromising sensitive information. This collaborative approach harnesses collective intelligence from distributed data sources while maintaining strict privacy guarantees, enabling applications across diverse domains.

In smart healthcare networks, FedLLMs can empower medical institutions to collaboratively analyze diverse patient populations while ensuring strict compliance with privacy regulations. For instance, during infectious disease outbreaks, hospitals across different regions can contribute anonymized patient data to collectively trace disease origins, model transmission dynamics, and identify optimal treatment protocols, without disclosing sensitive information. FedLLMs as a collaborative approach can significantly enhance diagnostic accuracy and facilitate rapid responses to emerging public health threats. Likewise, autonomous vehicle systems can utilize FedLLMs to aggregate driving experiences across a wide range of environments, from snow-covered mountain roads to tropical urban settings, thereby constructing robust safety models. These models can quickly disseminate adaptive countermeasures throughout global vehicle fleets in response to novel traffic scenarios or accident patterns, substantially improving

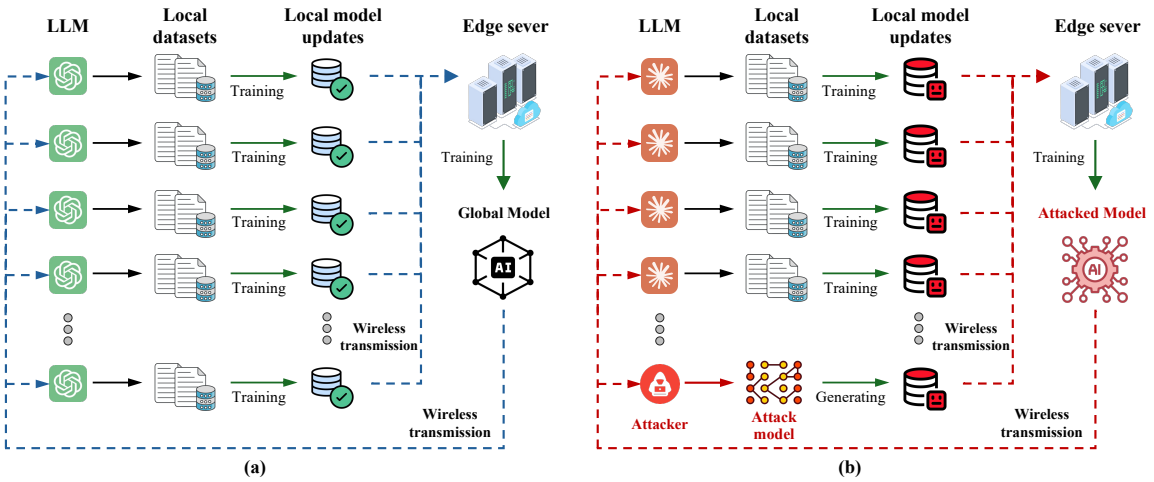


Fig. 4.2 (a) Illustration of the FedLLMs, where each benign user trains a local model based on their private data, and the edge server aggregates all local benign updates to train a global model, which is then broadcast back to local clients for further training. (b) A legitimate but malicious client uploads a poisoned model update that degrades the global optimization, thereby influencing the global model and falsifying subsequent local training.

road safety while preserving the proprietary algorithms of individual manufacturers [25]. Moreover, in IoT environments, FedLLMs enable coordinated learning across heterogeneous devices to address complex infrastructure challenges. For example, traffic sensors, environmental monitors, and surveillance cameras in smart cities can collaboratively predict and mitigate urban crises, manage emergency traffic flow, and optimize energy distribution during peak demand periods, all without the need to centralize sensitive operational data.

However, the distributed architecture of FedLLMs introduces security vulnerabilities from legitimate but malicious clients who can exploit their authorized access to learn from benign local updates. Such adversaries can systematically study legitimate update patterns and generate sophisticated malicious updates that mimic benign characteristics while embedding harmful payloads. The attack consequences are demonstrated in Fig. 4.1(b), where a benign LLM generates appropriate responses to user queries, while a compromised LLM produces harmful outputs that undermine system robustness.

Model Poisoning Attacks in FedLLMs

Fig. 4.2 illustrates the underlying mechanism of model poisoning attacks on FedLLMs. To comprehend the principles behind such attacks, it is essential to first understand

the standard federated learning workflow, as depicted in Fig. 4.2(a). In this process, multiple legitimate clients independently train their local LLMs on private datasets. Upon completing local training, each client generates model updates and transmits them to an edge server. The server then performs a global aggregation, typically using the federated averaging algorithm, which computes a weighted average of the received updates to produce a refined global model. This updated global model is subsequently broadcast to all participating clients, forming the basis for the next round of training. Through this iterative process, FedLLMs facilitate continual model enhancement via collaborative learning, without requiring exchange of raw data.

However, FedLLMs exhibit an inherent security vulnerability, as illustrated in Fig. 4.2(b). A malicious client can infiltrate the federated learning by posing as a legitimate participant. Unlike benign clients that train on authentic local data, the adversary leverages a carefully crafted attack model to generate malicious updates designed to manipulate the behavior of the global model. The adversarial updates are uploaded to the edge server, which, without discrimination, aggregates them alongside the benign updates from legitimate clients. Consequently, the global model becomes injected with malicious parameters, effectively transforming it into a poisoned model. Critically, the compromised global model is then disseminated to all FedLLMs' clients for subsequent training iterations. This not only embeds the attacker's influence within the global model but also ensures that legitimate clients unknowingly train on a corrupted model, thereby amplifying and perpetuating the attack's impact across the entire federated network.

DiSim-defense Mechanisms

Since the server lacks access to clients' raw data, most defense mechanisms operate at the aggregation stage by analyzing the uploaded model updates. Current mainstream approaches can be broadly categorized as DiSim-defense mechanisms that identify malicious updates by evaluating their deviations from benign updates using Euclidean distance or cosine similarity. DiSim-defense mechanisms are based on a key assumption: that adversarial updates can exhibit statistically distinguishable patterns from benign ones in high-dimensional parameter space [32]. However, this assumption renders them susceptible to sophisticated model poisoning attacks, wherein adversaries can carefully craft malicious updates to emulate the statistical signatures of benign updates, effectively bypassing detection. This vulnerability can be further exacerbated in the context of billion-parameter large language models (LLMs), where the immense

parameter space offers adversaries greater flexibility to embed malicious behavior while making statistical anomaly detection increasingly difficult.

Distance-based methods, such as Krum and Multi-Krum [20], exemplify the first category by computing pairwise Euclidean distances between all client updates and selecting those with the smallest sum of distances to their nearest neighbors, filtering out geometric outliers that deviate significantly from the benign cluster. However, distance-based methods are often ineffective in realistic non-IID settings and are vulnerable to the curse of dimensionality. The second category, similarity-based defenses, operates by computing cosine similarity between each client update and the global model or aggregate direction, discarding updates that fall below a predetermined similarity threshold or exhibit directional misalignment with the collective average [32]. This approach, in turn, is susceptible to defense-aware adversaries who can craft malicious updates that mimic the benign direction while still embedding a harmful payload.

The vulnerability of DiSim-defense mechanisms stems from their foundational assumption that malicious behavior manifests as a detectable statistical anomaly [58]. Such an assumption creates a critical security gap when confronted with advanced model-based attacks that transition from overt disruption to covert mimicry. By leveraging generative models capable of learning and reproducing the full statistical distribution and higher-order correlations of benign updates, adversaries can synthesize malicious payloads that remain indistinguishable from legitimate contributions under conventional detection metrics. This inherent limitation renders DiSim-defense mechanisms ineffective against attackers who possess the capability to model and exploit the very statistical patterns these defenses are built upon. As a result, FedLLMs become susceptible to a novel class of stealthy attacks that operate entirely within the statistical limits of legitimate client behaviour, thus evading detection and undermining the integrity of the system.

4.2.2 Graph Representation-based Model Poisoning

Graph Formulation and Generative Model Training

GRMP attack aims to learn the underlying structural patterns of benign model updates. Specifically, the attacker collects benign local updates from multiple clients over the communication rounds of FedLLMs. In the attacker, the benign model updates are then transformed into a graph-based representation, where each update is modeled as a node and the edges encode relational similarities between updates, as illustrated in Fig. 4.3. Moreover, a feature matrix is constructed by stacking the flattened parameter

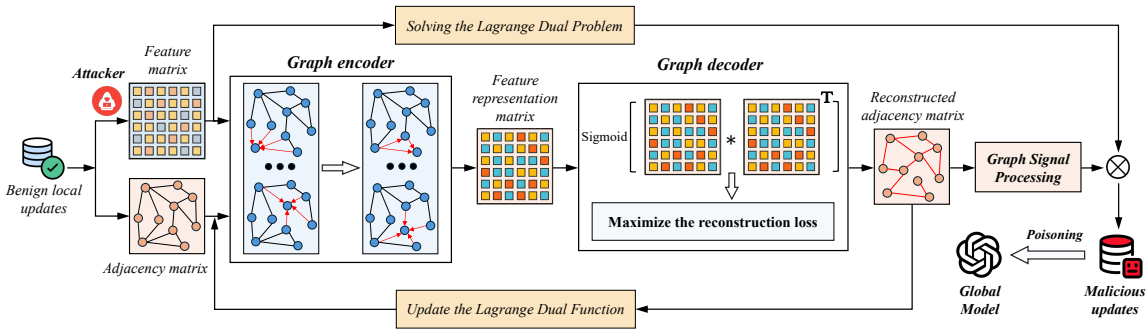


Fig. 4.3 Framework for the proposed graph representation-based model poisoning (GRMP) attack.

update vectors, where each node encapsulates the full information of a single benign update. The corresponding adjacency matrix is generated by computing pairwise similarities, typically using cosine similarity, between all update vectors. An edge is formed between two nodes if their similarity exceeds a predefined threshold, thereby capturing the intrinsic relational topology of the benign update manifold. This graph construction facilitates the subsequent learning of latent representations that encode both the individual characteristics of local updates and their higher-order structural relationships.

With this graph representation, the attacker trains a variational graph autoencoder (VGAE) to learn the underlying distribution of benign updates. The VGAE consists of two components: a graph encoder and a graph decoder. The encoder, implemented as a graph convolutional network, takes the entire graph structure including both feature and adjacency matrices as input and maps each node into a probability distribution in the latent space, characterized by mean and variance parameters. The decoder reconstructs the graph structure from sampled latent representations by computing inner products between latent vectors for every node pair, followed by a sigmoid activation function to predict edge probabilities in the reconstructed adjacency matrix. This formulation enables the VGAE to capture both the structural patterns and statistical properties of legitimate federated learning updates.

The VGAE is trained by maximizing the reconstruction loss function for adversarial purposes. This maximization objective focuses on increasing the reconstruction error to produce dissimilar adjacency matrices. Through this adversarial training process, the VGAE acquires the capability to synthesize malicious local models that appear structurally plausible while containing carefully crafted perturbations. The resulting generative model produces adversarial updates that exploit the learned graph structure

to effectively disrupt the federated learning aggregation process while maintaining sufficient similarity to bypass detection mechanisms.

4.2.3 Lagrange Dual Problem and Graph Signal Processing

The VGAE produces a reconstructed adjacency matrix representing correlations among model updates, instead of a malicious model update. To craft a malicious update, the attacker can leverage the learned graph structure to shape a weight vector, where a Lagrange dual optimization and a graph signal processing (GSP) module are developed. As shown in Fig. 4.3, the attack is formulated as a constrained optimization problem: maximize the poisoning impact on the global model while constraining the malicious update remains statistically indistinguishable from legitimate client contributions to bypass detection. The Lagrange dual approach incorporates stealth constraints directly into the objective, allowing the attacker to iteratively refine the VGAE's output toward an optimal adversarial graph structure without violating detection thresholds. However, even an adversarial graph structure alone is insufficient, the attacker needs to reconstruct a model update vector that follows this graph's patterns. Here, GSP module comes into play: the attacker decomposes benign model updates into structural correlations and underlying feature components, then recombines the adversarial graph structure with genuine feature signals to synthesize a malicious update that embeds new correlation patterns while remaining grounded in benign characteristics.

Furthermore, the Lagrange dual optimization can be designed to project the latest benign updates into the VGAE's learned latent space. Rather than performing simple reconstruction, the dual formulation enables the attacker to identify an optimal malicious latent vector whose decoded output maximizes reconstruction loss while satisfying stealth constraints. This process operates iteratively through the feedback loop illustrated in Fig. 4.3, converging on an adversarial graph structure that optimally balances poisoning efficiency and detection evasion.

Upon determining the optimized adversarial latent vector, it is decoded to produce the malicious graph structure that serves as a blueprint for parameter manipulation. The GSP module then performs the critical transformation from abstract graph representation to a concrete malicious update vector. The GSP module decomposes current benign model updates into structural correlations and underlying feature components through graph Laplacian and spectral decomposition. By regenerating the graph structure adversarially and recombining it with original benign feature signals, the module synthesizes a malicious update that embeds new correlation patterns while maintaining statistical characteristics consistent with legitimate contributions. This

adaptive synthesis process exploits specific model vulnerabilities at each communication round, such that the poisoned update appears indistinguishable from benign contributions during federated aggregation. Through this integrated approach, the VGAE provides the structural blueprint, the Lagrange dual optimizes the balance between attack impact and stealth, and the GSP module constructs the malicious update from genuine signals, achieving both effectiveness and invisibility.

4.2.4 Experiment and Evaluation

To evaluate the effectiveness and stealthiness of the GRMP attack, this study conducts experiments using FedLLMs for text classification. The experiments employ the AG News dataset from Kaggle, a widely-recognized benchmark dataset containing news articles across four categories: world, sports, business, and science. This dataset comprises 120,000 training samples and 7,600 test samples, providing sufficient data for statistically meaningful results in a federated setting.

This study simulates a federated learning environment with six clients, where two clients are controlled by attackers. The federation operates for twenty communication rounds, with each client performing two local training epochs per round using DistilBERT as the base model. DistilBERT is a distilled version of BERT that retains 97% of BERT's language understanding capabilities while being 40% smaller and 60% faster, making it particularly suitable for deployment in resource-constrained wireless network environments [50]. Meanwhile, the edge server employs a mainstream DiSim-defense approach that sets a dynamic detection threshold based on the statistical properties of the received updates. This defense approach identifies malicious updates by dynamically adjusting the detection threshold to flag those that deviate significantly from the expected cosine similarity patterns [22]. The GRMP attack specifically targets the model's understanding capabilities for business news articles. The attackers aim to manipulate the model to misclassify business articles containing financial keywords (e.g., stock, market, earnings, and profit) as sports news articles.

This study assesses the performance of the GRMP attack using three key metrics. First, learning accuracy measures the overall classification performance of the global model, indicating whether the model maintains its functionality for legitimate clients. Second, attack success rate (ASR) quantifies the percentage of targeted business articles containing financial keywords that are successfully misclassified as sports articles, measuring the effectiveness of the GRMP attack. In addition, cosine similarity analysis evaluates the invisibility of malicious updates by measuring their deviation from benign updates during the aggregation process. Together, these metrics provide a

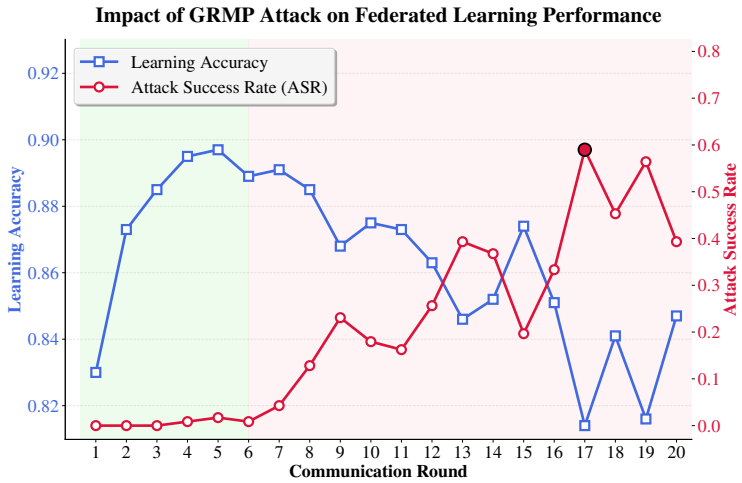


Fig. 4.4 GRMP attack's impact on learning accuracy and attack success rate over twenty communication rounds.

comprehensive evaluation framework that captures the attack's effectiveness, its impact on model functionality, and its stealthiness against detection mechanisms.

Attack Dynamics and Evasion Analysis

Fig. 4.4 reveals GRMP attack's impact on learning accuracy and ASR over twenty communication rounds. The attack exhibits a carefully orchestrated two-phase strategy that reflects adversarial planning. In the initial stage, attackers deliberately maintain minimal ASR below 2% while positioning themselves as legitimate clients. This strategic restraint exploits the temporal dynamics of federated learning, where client reputation is established through consistent participation across successive rounds. Once sufficient trust is established, the attack enters its exploitation phase, with ASR dramatically surging to 60%. Meanwhile, the global model maintains learning accuracy around 83%, demonstrating GRMP's ability to preserve overall system performance while selectively corrupting targeted classification behaviors. This performance preservation is crucial for attack sustainability, as substantial performance degradation could potentially reveal the presence of the attacker.

Fig. 4.5 illustrates the cosine similarity evolution of each client over twenty communication rounds. Despite the DiSim-defense mechanism employing a dynamic threshold, the similarity evolution demonstrates that the attackers consistently stay above the adaptive threshold throughout the training process. This result validates our claim that GRMP exploits the fundamental assumption gap in DiSim-defense mechanism. Through learning relational structures among benign updates via graph representation

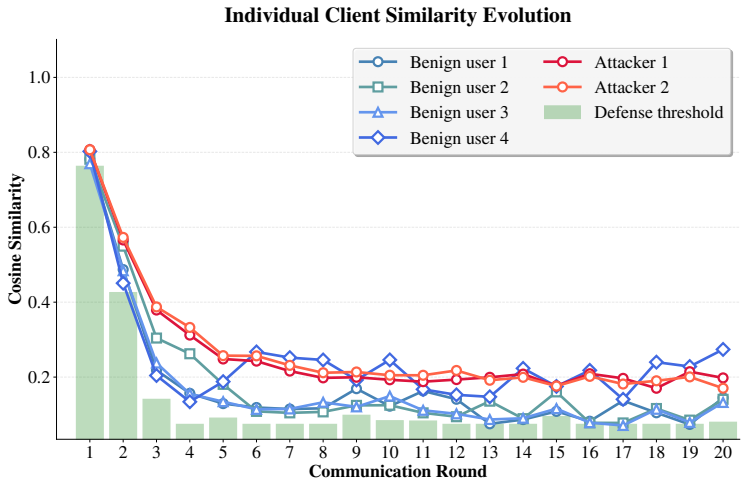


Fig. 4.5 Individual client cosine similarity evolution with defense threshold over twenty communication rounds.

learning, GRMP attackers generate updates that remain statistically indistinguishable from benign updates, effectively mimicking the natural similarity decline observed in legitimate participants.

The numerical results confirm the achievement of primary attack objectives. The 60% ASR on targeted business articles demonstrates successful corruption of the model's decision logic, while the complete bypass of similarity-based filtering validates GRMP's core innovation of generating malicious updates that mimic legitimate behavior. This capability stems from GRMP's exploitation of higher-order statistical relationships that remain invisible to current defense mechanisms. These findings reveal a fundamental limitation of existing defenses: approaches that assume malicious behavior manifests as statistical outliers prove ineffective against adaptive adversaries who understand the underlying data distribution. Furthermore, the non-IID nature of real-world federated data creates an unexpected vulnerability. While this characteristic was originally intended to improve model generalization, sophisticated attackers can exploit it to conceal malicious activity within natural statistical variation.

The implications of these vulnerabilities extend across diverse FedLLMs deployment scenarios beyond natural language understanding tasks. Unlike conventional attacks that induce random errors, model poisoning that manipulates contextual understanding represents a fundamentally novel threat class that directly targets the core intelligence of language models. In healthcare applications, such attacks could systematically alter diagnostic interpretations; in autonomous systems, they could corrupt critical scene understanding capabilities; in financial services, they could manipulate risk assessment

algorithms. The success of GRMP despite the presence of active defense mechanisms highlights critical gaps in current security paradigms and underscores the inadequacy of statistical anomaly detection methods against adversaries who thoroughly understand and exploit the legitimate variation inherent in federated learning systems. This reality necessitates a fundamental rethinking of defense strategies. Future approaches should shift from statistical analysis toward comprehensive behavioral verification frameworks.

4.3 Security Roadmap for the Internet of Agents

4.3.1 Dual Semantic and Structural Auditing

Future defenses should couple semantic auditing of model behavior with structural auditing of interclient relationships. Semantic auditing verifies whether an update preserves task intent and reasoning patterns rather than only its numeric profile. Explainable analysis can produce low-dimensional semantic fingerprints, such as attention or saliency heatmaps, that characterize decision focus; an autoencoder trained on benign fingerprints can flag abnormal reconstruction errors and reveal hidden triggers or backdoors [84]. Structural auditing represents the federation as a similarity graph and detects suspicious substructures that indicate coordinated or camouflaged manipulation. Message passing over the client graph can expose coalitions, while trust scoring and robust aggregation can down-weight implicated nodes and limit their influence [43, 49]. Tight integration of the two audits aligns defenses with IoA's semantic objectives and constrains adversaries that exploit higher-order correlations.

4.3.2 Systems, Standards, and Evaluation

Security enhancements require systems support and standardization. Graph-aware secure aggregation, provenance and attestation for update pipelines, and privacy mechanisms tailored to large language models are needed to protect confidentiality and integrity under wireless constraints. IoA deployments need shared ontologies for intents, beliefs, and plans, semantic headers and metadata for packetization, and conformance tests that verify cross-vendor interoperability at the semantic layer. Benchmark suites for federated LLMs in IoA should combine non-IID data generation, mobility, and intermittent connectivity with red-teaming protocols for structural poisoning and semantic manipulation. Cross-layer co-design that links contact scheduling, compression budgets, and semantic utility is essential for scalable and trustworthy agent cooperation.

4.4 Summary

This chapter defined the Internet of Agents as a semantic-aware communication paradigm in which heterogeneous LLM-driven entities coordinate by exchanging intent, belief, and plan messages rather than raw observations. An agent-centric stack and services for discovery, orchestration, consensus, and policy enforcement were outlined, and wireless federated learning was identified as the enabling substrate that aligns model knowledge while preserving data locality to maintain consistent semantic interpretation across agents. The security posture of federated LLMs was examined, exposing limitations of distance or similarity based defenses and illustrating a graph representation based poisoning paradigm that exploits higher-order correlations to evade detection. A concise roadmap was proposed that combines semantic and structural auditing, graph aware secure aggregation, provenance and semantic headers for interoperability, and evaluation protocols that reflect non-IID data, mobility, and intermittent connectivity.

Chapter 5

Conclusion

This thesis advances semantic communication as a unifying design paradigm for heterogeneous networks within the Internet of Everything. The key contributions of this thesis are as follows.

Firstly, this thesis develops a semantic-empowered molecular communication framework for biomedical diagnostic tasks in IoBNT, combining a deep encoder-decoder that extracts, quantizes, and reconstructs semantic features with a probabilistic channel network that approximates molecular propagation to enable gradient-based training; experiments demonstrate improved accuracy and robustness over conventional bit-level baselines. Secondly, this thesis proposes the first semantic communication architecture tailored to IoS, introducing a three-layer design aligned with DTN/CCSDS practices, and validates the approach in a representative deep-space scenario on Martian dust storm monitoring with mission-level gains in efficiency. Lastly, this thesis formalizes IoA as a semantic-aware communication paradigm for coordinating heterogeneous LLM agents and identifies federated learning as the enabler for distributed coordination; this thesis further evaluates the resilience of federated LLMs based on a graph representation-based model poisoning study, evaluates prevailing defense mechanisms, and the resulting insights inform a resilience roadmap for future IoA research.

Although the results are encouraging, two main limitations remain. First, the IoBNT and IoS evaluations rely primarily on simulation and modeling under simplified or idealized channel assumptions; rigorous validation will require scaling the experiments, tightening control of operating parameters, and calibrating differentiable channel surrogates against measurements from realistic bio-nano and deep-space environments. Second, the IoA resilience study employs relatively small-capacity language models and offline settings; to obtain operationally meaningful evidence, future analyzes should incorporate stronger generative backbones, larger-scale federated deployments, and

real-time model poisoning trials under realistic bandwidth, latency, and heterogeneity constraints.

Future research will advance along three complementary strands. For IoBNT, we will model and theorize naturally occurring semantic behaviors in biological systems, use them to inform task-aligned encoders, and validate the resulting designs in microfluidic testbeds with measurement-driven calibration. For IoS, we will mature the technical architecture of semantic communication, clarify its relationship with the open systems interconnection model, and develop standards-compatible semantic headers and scheduling that better empower interplanetary links under delay, intermittency, and energy constraints, including hardware-in-the-loop evaluations. For IoA, we will advance a heterogeneous Internet of Agents enabled by semantic communication and federated learning, with emphasis on efficiency, resilience, fairness, and interpretability, and we will establish benchmarks and protocols for privacy-preserving, Byzantine-robust coordination in both training and deployment.

References

[1] Akan, O. B., Dinc, E., Kuscu, M., Cetinkaya, O., and Bilgin, B. A. (2023). Internet of everything (ioe)-from molecules to the universe. *IEEE Communications Magazine*, 61(10):122–128.

[2] Akan, O. B., Ramezani, H., Khan, T., Abbasi, N. A., and Kuscu, M. (2016). Fundamentals of molecular information and communication science. *Proceedings of the IEEE*, 105(2):306–318.

[3] Akdeniz, B. C. and Egan, M. (2021). Molecular communication for equilibrium state estimation in biochemical processes on a lab-on-a-chip. *IEEE Transactions on NanoBioscience*, 20(2):193–201.

[4] Akyildiz, I. F., Akan, Ö. B., Chen, C., Fang, J., and Su, W. (2003). Interplanetary internet: state-of-the-art and research challenges. *Computer Networks*, 43(2):75–112.

[5] Akyildiz, I. F. and Kak, A. (2019). The internet of space things/cubesats. *IEEE Network*, 33(5):212–218.

[6] Al-Hraishawi, H., Chougrani, H., Kisseleff, S., Lagunas, E., and Chatzinotas, S. (2022). A survey on nongeostationary satellite systems: The communication perspective. *IEEE Communications Surveys & Tutorials*, 25(1):101–132.

[7] Amini, H., Mia, M. J., Saadati, Y., Imteaj, A., Nabavirazavi, S., Thakker, U., Hossain, M. Z., Fime, A. A., and Iyengar, S. (2025). Distributed llms and multimodal large language models: A survey on advances, challenges, and future directions. *arXiv preprint arXiv:2503.16585*.

[8] Baydas, O. T., Cetinkaya, O., and Akan, O. B. (2023). Estimation and detection for molecular mimo communications in the internet of bio-nano things. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 9(1):106–110.

[9] Borges, L. F., Barros, M. T., and Nogueira, M. (2021). Toward reliable intra-body molecular communication: An error control perspective. *IEEE Communications Magazine*, 59(5):114–120.

[10] Bourtsoulatze, E., Kurka, D. B., and Gündüz, D. (2019). Deep joint source-channel coding for wireless image transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3):567–579.

[11] Cai, H. and Akan, O. B. (2025). Semantic learning for molecular communication in internet of bio-nano things. *arXiv preprint arXiv:2502.08426*.

[12] Cai, H., Dong, H., Wang, H., Li, K., and Akan, O. B. (2025a). Graph representation-based model poisoning on federated large language models. *arXiv preprint arXiv:2507.01694*.

[13] Cai, H., Wang, H., Dong, H., and Akan, O. B. (2025b). Semantic communication for the internet of space: New architecture, challenges, and future vision. *arXiv preprint arXiv:2503.23446*.

[14] Cao, Q., Deng, R., Pan, Y., Liu, R., Chen, Y., Gong, G., Zou, J., Yang, H., and Han, D. (2024). Robotic wireless capsule endoscopy: recent advances and upcoming technologies. *Nature Communications*, 15(1):4597.

[15] Cheng, Y., Zhang, W., Zhang, Z., Zhang, C., Wang, S., and Mao, S. (2024). Towards federated large language models: Motivations, methods, and future directions. *IEEE Communications Surveys & Tutorials*.

[16] Dhok, S., Chouhan, L., Noel, A., and Sharma, P. K. (2021). Cooperative molecular communication in drift-induced diffusive cylindrical channel. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, 8(1):44–55.

[17] Dong, H. and Akan, O. B. (2024). Martian dust storm detection with thz opportunistic integrated sensing and communication in the internet of space (ios). *arXiv preprint arXiv:2412.10921*.

[18] Dong, H. and Akan, O. B. (2025). Debrisense: Thz-based integrated sensing and communications (isac) for debris detection and classification in the internet of space (ios). *IEEE Transactions on Wireless Communications*.

[19] Dressler, F. and Fischer, S. (2015). Connecting in-body nano communication with body area networks: Challenges and opportunities of the internet of nano things. *Nano Communication Networks*, 6(2):29–38.

[20] Fang, M., Cao, X., Jia, J., and Gong, N. (2020). Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622.

[21] Getu, T. M., Kaddoum, G., and Bennis, M. (2025). Semantic communication: A survey on research landscape, challenges, and future directions. *Proceedings of the IEEE*.

[22] Han, S., Buyukates, B., Hu, Z., Jin, H., Jin, W., Sun, L., Wang, X., Wu, W., Xie, C., Yao, Y., et al. (2024). Fedsecurity: A benchmark for attacks and defenses in federated learning and federated llms. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5070–5081.

[23] Huang, D., Gao, F., Tao, X., Du, Q., and Lu, J. (2022). Toward semantic communications: Deep learning-based image semantic coding. *IEEE Journal on Selected Areas in Communications*, 41(1):55–71.

[24] Huang, D., Tao, X., Gao, F., and Lu, J. (2021). Deep learning-based image semantic coding for semantic communications. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE.

[25] Huang, Y., Cheng, Y., and Wang, K. (2025). Efficient driving behavior narration and reasoning on edge device using large language models. *IEEE Transactions on Vehicular Technology*, (99):1–5.

[26] Jamali, V., Ahmadzadeh, A., Wicke, W., Noel, A., and Schober, R. (2019). Channel modeling for diffusive molecular communication—a tutorial review. *Proceedings of the IEEE*, 107(7):1256–1301.

[27] Jeruchim, M. (2003). Techniques for estimating the bit error rate in the simulation of digital communication systems. *IEEE Journal on selected areas in communications*, 2(1):153–170.

[28] Jiang, F., Pan, C., Dong, L., Wang, K., Debbah, M., Niyato, D., and Han, Z. (2025a). A comprehensive survey of large ai models for future communications: Foundations, applications and challenges. *arXiv preprint arXiv:2505.03556*.

[29] Jiang, F., Pan, C., Dong, L., Wang, K., Dobre, O. A., and Debbah, M. (2025b). From large ai models to agentic ai: A tutorial on future intelligent communications. *arXiv preprint arXiv:2505.22311*.

[30] Jiang, P., Wen, C.-K., Jin, S., and Li, G. Y. (2022). Wireless semantic communications for video conferencing. *IEEE Journal on Selected Areas in Communications*, 41(1):230–244.

[31] Jiao, J., Wu, S., Lu, R., and Zhang, Q. (2021). Massive access in space-based internet of things: Challenges, opportunities, and future directions. *IEEE Wireless Communications*, 28(5):118–125.

[32] Kasyap, H. and Tripathy, S. (2024). Sine: Similarity is not enough for mitigating local model poisoning attacks in federated learning. *IEEE Transactions on Dependable and Secure Computing*, 21(5):4481–4494.

[33] Kilinc, D. and Akan, O. B. (2014). Receiver design for molecular communication. *IEEE Journal on Selected Areas in Communications*, 31(12):705–714.

[34] Kodheli, O., Lagunas, E., Maturo, N., Sharma, S. K., Shankar, B., Montoya, J. F. M., Duncan, J. C. M., Spano, D., Chatzinotas, S., Kisseleff, S., et al. (2020). Satellite communications in the new space era: A survey and future challenges. *IEEE Communications Surveys & Tutorials*, 23(1):70–109.

[35] Kuscu, M., Dinc, E., Bilgin, B. A., Ramezani, H., and Akan, O. B. (2019). Transmitter and receiver architectures for molecular communications: A survey on physical design with modulation, coding, and detection techniques. *Proceedings of the IEEE*, 107(7):1302–1341.

[36] Kuscu, M. and Unluturk, B. D. (2021). Internet of bio-nano things: A review of applications, enabling technologies and key challenges. *arXiv preprint arXiv:2112.09249*.

[37] Li, A., Liu, X., Wang, G., and Zhang, P. (2022). Domain knowledge driven semantic communication for image transmission over wireless channels. *IEEE Wireless Communications Letters*, 12(1):55–59.

[38] Li, A., Wu, S., Meng, S., Lu, R., Sun, S., and Zhang, Q. (2024a). Toward goal-oriented semantic communications: New metrics, framework, and open challenges. *IEEE Wireless Communications*, 31(5):238–245.

[39] Li, K., Yuan, X., Zheng, J., Ni, W., Dressler, F., and Jamalipour, A. (2024b). Leverage variational graph representation for model poisoning on federated learning. *IEEE Transactions on Neural Networks and Learning Systems*.

[40] Li, K., Zheng, J., Ni, W., Huang, H., Liò, P., Dressler, F., and Akan, O. B. (2024c). Biasing federated learning with a new adversarial graph attention network. *IEEE Transactions on Mobile Computing*.

[41] Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q., and Yu, P. S. (2022). Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*, 35(7):8726–8746.

[42] Lyu, Z., Zhu, G., Xu, J., Ai, B., and Cui, S. (2024). Semantic communications for image recovery and classification via deep joint source and channel coding. *IEEE Transactions on Wireless Communications*, 23(8):8388–8404.

[43] Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., Xiong, H., and Akoglu, L. (2021). A comprehensive survey on graph anomaly detection with deep learning. *IEEE transactions on knowledge and data engineering*, 35(12):12012–12038.

[44] Marcone, A., Pierobon, M., and Magarini, M. (2018). Parity-check coding based on genetic circuits for engineered molecular communication between biological cells. *IEEE Transactions on Communications*, 66(12):6221–6236.

[45] Meng, S., Wu, S., Zhang, J., Cheng, J., Zhou, H., and Zhang, Q. (2024). Semantics-empowered space-air-ground-sea integrated network: New paradigm, frameworks, and challenges. *IEEE Communications Surveys & Tutorials*.

[46] Noel, A., Cheung, K. C., and Schober, R. (2013). Using dimensional analysis to assess scalability and accuracy in molecular communication. In *2013 IEEE International Conference on Communications Workshops (ICC)*, pages 818–823. IEEE.

[47] Ortlek, B. E. and Akan, O. B. (2025). Modeling and analysis of scfa-driven vagus nerve signaling in the gut-brain axis via molecular communication. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*.

[48] Peng, X., Qin, Z., Tao, X., Lu, J., and Hanzo, L. (2024). A robust semantic text communication system. *IEEE Transactions on Wireless Communications*, 23(9):11372–11385.

[49] Pillutla, K., Kakade, S. M., and Harchaoui, Z. (2022). Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154.

[50] Pokhrel, S. R. et al. (2025). On harnessing semantic communication with natural language processing. *IEEE Internet of Things Journal*.

[51] Sagduyu, Y. E., Erpek, T., Yener, A., and Ulukus, S. (2024). Will 6g be semantic communications? opportunities and challenges from task oriented and secure communications to integrated sensing. *IEEE Network*.

[52] Shahbaz, S., Mirmohseni, M., and Nasiri-Kenari, M. (2024). A jamming resistant molecular communication scheme. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*.

[53] Shannon, C. E. and Weaver, W. (1998). *The mathematical theory of communication*. University of Illinois press.

[54] Smedsrud, P. H., Thambawita, V., Hicks, S. A., Gjestang, H., Nedrejord, O. O., Næss, E., Borgli, H., Jha, D., Berstad, T. J. D., Eskeland, S. L., et al. (2021). Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142.

[55] Sun, H., Chen, X., Shi, Q., Hong, M., Fu, X., and Sidiropoulos, N. D. (2018). Learning to optimize: Training deep neural networks for interference management. *IEEE Transactions on Signal Processing*, 66(20):5438–5453.

[56] Tong, H., Li, H., Du, H., Yang, Z., Yin, C., and Niyato, D. (2024). Multimodal semantic communication for generative audio-driven video conferencing. *IEEE Wireless Communications Letters*.

[57] Walter, V., Bi, D., Salehi-Reyhani, A., and Deng, Y. (2023). Real-time signal processing via chemical reactions for a microfluidic molecular communication system. *Nature Communications*, 14(1):7188.

[58] Wan, Y., Qu, Y., Ni, W., Xiang, Y., Gao, L., and Hossain, E. (2024). Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 26(3):1861–1897.

[59] Wang, K., Zhang, H., Kaddoum, G., Shin, H., Quek, T. Q., and Win, M. Z. (2025a). Sgan-ra: Reconstruction attack for big model in asynchronous federated learning. *IEEE Communications Magazine*, 63(4):66–72.

[60] Wang, W., Liu, M., and Chen, M. (2023). Ca_deepsc: Cross-modal alignment for multi-modal semantic communications. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*, pages 5871–5876. IEEE.

[61] Wang, Y., Guo, S., Pan, Y., Su, Z., Chen, F., Luan, T. H., Li, P., Kang, J., and Niyato, D. (2025b). Internet of agents: Fundamentals, applications, and challenges. *arXiv preprint arXiv:2505.07176*.

[62] Wang, Y., Pan, Y., Su, Z., Deng, Y., Zhao, Q., Du, L., Luan, T. H., Kang, J., and Niyato, D. (2025c). Large model based agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends. *IEEE Communications Surveys & Tutorials*.

[63] Wei, W., Fu, L., Gu, H., Lu, X., Liu, L., Mumtaz, S., and Guizani, M. (2024). Iris: Towards intelligent reliable routing for software defined satellite networks. *IEEE Transactions on Communications*.

[64] Weng, Z. and Qin, Z. (2021). Semantic communication systems for speech transmission. *IEEE Journal on Selected Areas in Communications*, 39(8):2434–2444.

[65] Xiao, H., Dokaj, K., and Akan, O. B. (2023). What really is 'molecule'in molecular communications? the quest for physics of particle-based information carriers. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*.

[66] Xie, H. and Qin, Z. (2020). A lite distributed semantic communication system for internet of things. *IEEE Journal on Selected Areas in Communications*, 39(1):142–153.

[67] Xie, H., Qin, Z., and Li, G. Y. (2021a). Task-oriented multi-user semantic communications for vqa. *IEEE Wireless Communications Letters*, 11(3):553–557.

[68] Xie, H., Qin, Z., Li, G. Y., and Juang, B.-H. (2021b). Deep learning enabled semantic communication systems. *IEEE transactions on signal processing*, 69:2663–2675.

[69] Xie, H., Qin, Z., Tao, X., and Letaief, K. B. (2022). Task-oriented multi-user semantic communications. *IEEE Journal on Selected Areas in Communications*, 40(9):2584–2597.

[70] Xu, C., Mashhadi, M. B., Ma, Y., Tafazolli, R., and Wang, J. (2025). Generative semantic communications with foundation models: Perception-error analysis and semantic-aware power allocation. *IEEE Journal on Selected Areas in Communications*.

[71] Xu, X., Xu, Y., Dou, H., Chen, M., and Wang, L. (2024). Federated kd-assisted image semantic communication in iot edge learning. *IEEE Internet of Things Journal*.

[72] Yan, N., Wang, K., Zhi, K., Pan, C., Chai, K. K., and Poor, H. V. (2025). Secure and private over-the-air federated learning: Biased and unbiased aggregation design. *IEEE Transactions on Wireless Communications*.

[73] Yang, K., Bi, D., Deng, Y., Zhang, R., Rahman, M. M. U., Ali, N. A., Imran, M. A., Jornet, J. M., Abbasi, Q. H., and Alomainy, A. (2020). A comprehensive survey on hybrid communication in context of molecular communication and terahertz communication for body-centric nanonetworks. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, 6(2):107–133.

[74] Yang, W., Du, H., Liew, Z. Q., Lim, W. Y. B., Xiong, Z., Niyato, D., Chi, X., Shen, X., and Miao, C. (2022). Semantic communications for future internet: Fundamentals, applications, and challenges. *IEEE Communications Surveys & Tutorials*, 25(1):213–250.

[75] Yang, Y., Gao, F., Tao, X., Liu, G., and Pan, C. (2023a). Environment semantics aided wireless communications: A case study of mmwave beam prediction and blockage prediction. *IEEE journal on selected areas in communications*, 41(7):2025–2040.

[76] Yang, Z., Chen, M., Zhang, Z., and Huang, C. (2023b). Energy efficient semantic communication over wireless networks with rate splitting. *IEEE Journal on Selected Areas in Communications*, 41(5):1484–1495.

[77] Ye, N., Miao, S., Pan, J., Xiang, Y., and Mumtaz, S. (2025). Dancing with chains: Spaceborne distributed multi-user detection under inter-satellite link constraints. *IEEE Journal of Selected Topics in Signal Processing*.

[78] Yukun, C., Wei, C., and Bo, A. (2024). Building semantic communication system via molecules: An end-to-end training approach. *China Communications*, 21(7):113–124.

[79] Zhang, G., Hu, Q., Qin, Z., Cai, Y., Yu, G., and Tao, X. (2024a). A unified multi-task semantic communication system for multimodal data. *IEEE Transactions on Communications*, 72(7):4101–4116.

[80] Zhang, S. and Akan, O. B. (2024a). 3d receiver for molecular communications in internet of organoids. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*.

[81] Zhang, W., Liu, Y., Chen, L., Shi, J., Hong, X., and Wang, X. (2024b). Semantically-disentangled progressive image compression for deep space communications: Exploring the ultra-low rate regime. *IEEE Journal on Selected Areas in Communications*.

[82] Zhang, Z. and Akan, O. B. (2024b). Analysis of terahertz communication under dust storm conditions on mars. *IEEE Communications Letters*.

[83] Zhao, B., Xing, H., Xu, L., Li, Y., Feng, L., Peng, J., and Xiao, Z. (2024). On forecasting-oriented time series transmission: A federated semantic communication system. *IEEE Transactions on Mobile Computing*.

[84] Zheng, J., Li, K., Yuan, X., Ni, W., Tovar, E., and Crowcroft, J. (2024). Exploring visual explanations for defending federated learning against poisoning attacks. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 1596–1598.

[85] Zhou, D., Sheng, M., Li, J., and Han, Z. (2023). Aerospace integrated networks innovation for empowering 6g: A survey and future challenges. *IEEE Communications Surveys & Tutorials*, 25(2):975–1019.

[86] Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D., et al. (2024). Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *IEEE Communications Surveys & Tutorials*.

[87] Zhu, J., Bai, C., Zhu, Y., Lu, X., and Wang, K. (2023). Evolutionary generative adversarial network based end-to-end learning for mimo molecular communication with drift system. *Nano Communication Networks*, 37:100456.

[88] Zulfiqar, S. and Akan, O. B. (2025). Molecular communication-based quorum sensing disruption for enhanced immune defense. *IEEE Transactions on NanoBioscience*.