

Foundation Model Empowered Synesthesia of Machines (SoM): AI-native Intelligent Multi-Modal Sensing-Communication Integration

Xiang Cheng, *Fellow, IEEE*, Boxun Liu, *Graduate Student Member, IEEE*,
Xuanyu Liu, *Graduate Student Member, IEEE*, Ensong Liu, *Graduate Student Member, IEEE*,
and Ziwei Huang, *Member, IEEE*

Abstract—To support future intelligent multifunctional sixth-generation (6G) wireless communication networks, Synesthesia of Machines (SoM) is proposed as a novel paradigm for artificial intelligence (AI)-native intelligent multi-modal sensing-communication integration. However, existing SoM system designs rely on task-specific AI models and face challenges such as scarcity of massive high-quality datasets, constrained modeling capability, poor generalization, and limited universality. Recently, foundation models (FMs) have emerged as a new deep learning paradigm and have been preliminarily applied to SoM-related tasks, but a systematic design framework is still lacking. In this paper, we for the first time present a systematic categorization of FMs for SoM system design, dividing them into general-purpose FMs, specifically large language models (LLMs), and SoM domain-specific FMs, referred to as wireless foundation models. Furthermore, we derive key characteristics of FMs in addressing existing challenges in SoM systems and propose two corresponding roadmaps, i.e., LLM-based and wireless foundation model-based design. For each roadmap, we provide a framework containing key design steps as a guiding pipeline and several representative case studies of FM-empowered SoM system design. Specifically, we propose LLM-based path loss generation (LLM4PG) and scatterer generation (LLM4SG) schemes, and wireless channel foundation model (WiCo) for SoM mechanism exploration, LLM-based wireless multi-task SoM transceiver (LLM4WM) and wireless foundation model (WiFo) for SoM-enhanced transceiver design, and wireless cooperative perception foundation model (WiPo) for SoM-enhanced cooperative perception, demonstrating the significant superiority of FMs over task-specific models. Finally, we summarize and highlight potential directions for future research.

Index Terms—Intelligent multi-modal sensing-communication integration, Synesthesia of Machines (SoM), foundation models (FMs), large language models (LLMs), wireless foundation models.

I. INTRODUCTION

A. Background

As a key infrastructure in the information age, the fifth-generation (5G) [1] wireless communication networks have been widely deployed to support three major application scenarios, named enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and

Xiang Cheng, Boxun Liu, Xuanyu Liu, Ensong Liu, and Ziwei Huang are with the State Key Laboratory of Photonics and Communications, School of Electronics, Peking University, Beijing 100871, China (e-mail:xiangcheng@pku.edu.cn; boxunliu@stu.pku.edu.cn; xvanyliu@gmail.com; ensongliu@pku.edu.cn; ziweihuang@pku.edu.cn).

low latency communications (uRLLC). To meet the network demands beyond 2030, the sixth-generation (6G) wireless communication networks will further enhance and expand 5G to support a wide range of downstream applications, including cloud VR, Internet of Things (IoT) industry automation, cellular vehicle-to-everything (C-V2X), and digital twin. As envisioned by the International Telecommunication Union (ITU) [2] and other organizations, integrated sensing and communications (ISAC) and integrated artificial intelligence(AI) and communications will play an unprecedently important role in achieving intelligent multifunctional wireless networks.

ISAC [3] aims to unify radio-frequency (RF) sensing and communication functions in a single system, optimizing trade-offs and achieving mutual performance gains, referred to as RF-ISAC in this paper [4]. It is expected to improve spectral and energy efficiencies, reduce hardware and signaling costs, and foster deeper integration by co-designing communications and sensing for mutual benefits [5]. Nevertheless, the existing RF-ISAC is limited to RF sensing, failing to fully leverage communication and multi-modal sensing information. It is also restricted to static or low-speed scenarios [4], making it difficult to support high-dynamic 6G scenarios. For instance, in typical 6G-enabled embodied intelligence scenarios [6], a large number of intelligent agents move and interact in the complex and high-mobility environment while simultaneously deploying communication devices and sensors. It is noted that each agent can capture communication information, as well as both RF data, e.g., radar point clouds, and non-RF sensing data, e.g., red-green-blue (RGB) pictures and light detection and ranging (LiDAR) point clouds. However, most existing works, including RF-ISAC, focus on the separate study of communication and multi-modal sensory information, without considering their interconnections. Therefore, there is an urgent need to systematically explore the intelligent integration and mutually beneficial mechanisms between communication and multi-modal sensing.

B. Synesthesia of Machines

Inspired by the synesthesia of human, Synesthesia of Machines (SoM) [4] has been proposed as a new paradigm for the intelligent integration of communication and multi-modal sensing. Unlike RF-ISAC, SoM aims to enhance environmental sensing and communication functions through SoM processing of multi-modal data, including LiDAR point clouds,

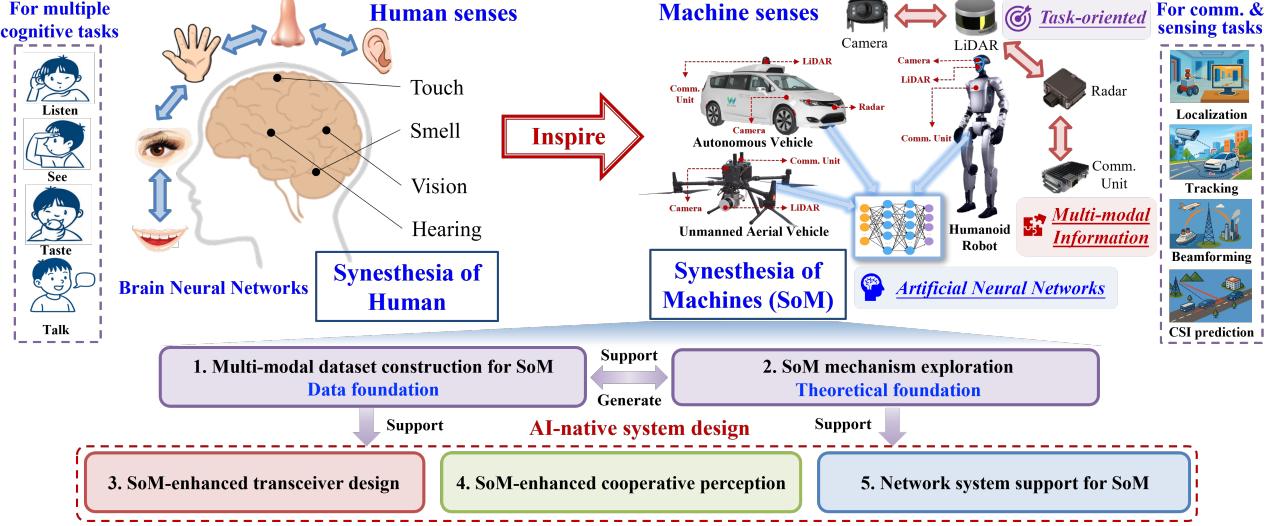


Fig. 1. An illustration of the SoM framework, highlighting its three key characteristics and the interrelation of its five key research directions.

video, image, radar point clouds, and channel data. However, SoM processing is not straightforward and faces significant challenges [7]. First, there are significant differences in the representation of data across modalities. For instance, RGB images and depth maps provide dense two-dimensional(2D) visual information, while channel state information (CSI) consists of complex space-time-frequency dimensions. Second, there are significant differences in the frequency bands used by different modalities, especially between non-RF sensing information and communication channel data, where the frequency bands differ by more than four orders of magnitude. Third, there are differences in applications, where the objectives of multi-modal sensing tasks and communication tasks are inherently different.

As shown in Fig. 1, inspired by how synesthesia of human utilizes brain neural networks to process multi-sensory information for performing multiple cognitive tasks, the core of SoM processing lies in utilizing artificial neural networks (ANN) to handle multi-modal information for specific sensing and communication tasks. Here, we summarize three key characteristics of SoM processing below.

- *Multi-modal information*: SoM processing fully leverages communication and multi-modal sensing information covering multiple frequency bands.
- *Task-oriented*: SoM processing focuses on specific sensing and communication tasks to design targeted algorithms.
- *Artificial neural networks*: SoM processing conducts task-oriented and data-driven neural network design.

In summary, SoM refers to task-oriented AI-native intelligence integration of communication and multi-modal sensing. Specifically, to support the comprehensive design of SoM systems, five key research directions are outlined below, with their interrelationships illustrated in Fig. 1.

- *Multi-modal dataset construction for SoM*: Since the scale and quality of the dataset determine the ultimate performance limit of AI-native systems, it is necessary to construct a massive and high-quality multi-modal sensing-communication dataset. Towards this objective,

we construct a real-world data injected synthetic intelligent multi-modal sensing-communication integration dataset, named SynthSoM, including RF communications, i.e., channel data, RF sensing, i.e., millimeter-wave (mmWave) radar data, and non-RF sensing, i.e., RGB images, depth maps, and LiDAR point clouds [8]. The constructed SynthSoM dataset provides a reliable data foundation for SoM research.

- *SoM mechanism exploration*: As the theoretical foundation of SoM research, the SoM mechanism, i.e., mapping relationship between communications and multi-modal sensing, needs to be explored based on the constructed multi-modal dataset. The explored SoM mechanism not only supports the efficient and high-fidelity generation of multi-modal data, but also facilitates SoM-related research, including transceiver design, cooperative perception, and network system support. However, due to significant differences between communications and multi-modal sensing in terms of data representation forms, acquisition frequencies, and application orientations, the SoM mechanism, i.e., mapping relationship, is complex and nonlinear, and thus is extremely difficult to explore.
- *SoM-enhanced transceiver design*: SoM-enhanced transceivers fully leverage rich prior knowledge from environmental multi-modal sensing information, such as the scatterer position and velocity, and further utilize the SoM mechanism to simplify or enhance the performance of wireless transmission systems. The multi-modal dataset is the cornerstone of SoM-enhanced transceiver design, enabling AI-native multi-modal sensing-assisted wireless transmission systems.
- *SoM-enhanced cooperative perception*: SoM-enhanced cooperative perception focuses on improving perception performance under realistic communication constraints, such as quantization, multi-user interference, and channel effects. Trained on multi-modal SoM datasets, the model incorporates physical-layer information to transmit multi-modal features that are well-suited to wireless channels, a process referred to as SoM feature transmission.

- *Network system support for SoM:* The collection, transmission, and processing of multi-modal sensing data rely on an integrated sensing, communication, computation, and storage network. A task-oriented network design supports complex SoM processing, which requires heterogeneous resources and is adaptive to available resources. Elastic task-oriented resource allocation schemes further improve resource utilization under a dynamic environment, while meeting the latency and availability service requirements.

However, existing research on SoM processing is still in its infancy and cannot adequately achieve the desired objective, attributed to the following four main challenges.

- *Scarcity of massive and high-quality datasets:* Considering that SoM processing is data-driven, high-quality datasets are the cornerstone of SoM processing. However, both real-world datasets constructed by measurement equipment and synthetic datasets constructed by simulation software encounter difficulties when constructing large-scale datasets. Specifically, the real-world measurement method is limited by the high cost of multi-modal data collection devices, while the simulation method is constrained by the enormous computational cost with low quality of collected data.
- *Limited modeling capability:* On one hand, SoM tasks are generally more challenging than conventional wireless system design, as they require modeling the complex mapping relationships between various modalities of information [9]. On the other hand, for scene-agnostic SoM tasks [10], the scale of the training dataset is constrained by the high cost of real-world measurements. In such few-shot scenarios, task-specific deep learning-based schemes are impaired due to insufficient training data.
- *Limited generalization across data:* Most existing ANNs used for SoM processing are trained and tested on specific datasets, exhibiting poor generalization. Specifically, when the data distribution undergoes significant changes, the performance of the ANN drops significantly. In this case, fine-tuning in new scenarios incurs additional data collection and network training overhead.
- *Limited universality across tasks:* Existing SoM processing schemes design distinct ANNs and loss functions for each specific task, thereby restricting the model to single-task learning. However, in real-world application systems, intelligent agents are required to handle a wide variety of SoM tasks. Therefore, numerous separate ANNs need to be deployed simultaneously, resulting in significant model storage and management overhead.

C. Foundation Models

In the past decade, deep learning has been widely applied across various fields, including communication and multi-modal sensing tasks, due to its powerful modeling ability directly from raw data without relying on prior information. Nevertheless, conventional deep learning networks, referred to as *task-specific models* in this paper, are trained on specific tasks and datasets in a supervised manner, generally lacking

generalization and universality. Recently, foundation models (FMs) [11] have emerged as a new paradigm in deep learning, revolutionizing several fields like natural language processing (NLP) and computer vision (CV). Specifically, a foundation model [11] is any model that is trained on massive data, generally in a self-supervision manner, that can be adapted to a wide range of downstream tasks with fine-tuning or zero-shot inference. Its core idea lies in *pre-training and task adaptation*, i.e., during the pre-training phase, the model learns generalizable representations, and during deployment, it quickly adapts to specific scenarios with minimal or no data. Large language models (LLMs) [12] are the forefront and most successful representative achievements of FMs, where GPT-4 and DeepSeek [13] have validated the emergence of astonishing understanding and reasoning capabilities of FMs, driven by the enormous scale of model parameters and datasets. In addition to general-purpose FMs for NLP or CV, domain-specific FMs are constantly emerging for various fields, such as time series prediction [14], weather forecasting [15], and remote sensing [16]. Inspired by the powerful inference and generalization capabilities of FMs, we wonder whether *FMs can be leveraged for SoM system design to address the aforementioned four challenges*.

D. Related Works

Although research on FM-empowered SoM system design is still in its infancy, a growing number of studies [17–19] have explored the application of FMs in SoM-related fields, including dataset generation, wireless communications, RF-ISAC, and semantic communications. According to the type of adopted FMs, we classify these studies into two categories, including those based on general-purpose FMs, specifically LLMs, and those based on domain-specific FMs, namely *wireless foundation models* as defined in Section II. Existing surveys on FM-empowered SoM-related fields have solely summarized a limited subset of such studies from narrow perspectives. For instance, survey [4] first proposed and elaborated on the SoM concept and framework in detail, while its design approaches still mainly rely on task-specific deep learning models, with limited exploration of FM-enabled designs. Survey [20] provided a comprehensive overview of LLM-enabled telecommunication (telecom) networks, covering fundamental LLM techniques, key telecom applications, and future directions. The authors in [20] preliminarily explore the concept of domain-specific FMs for wireless prediction tasks, but do not incorporate the latest emerging research on wireless foundation models. The review [21] summarizes existing research on LLM-driven synthetic data generation, curation, and evaluation, but ignores studies on dataset generation for communications and multi-modal sensing. In summary, a comprehensive and unified framework for FM-empowered SoM system design is still lacking in the existing literature.

E. Contributions and Organization

To the best of our knowledge, this paper is the first systematic research on FM-empowered AI-native SoM system

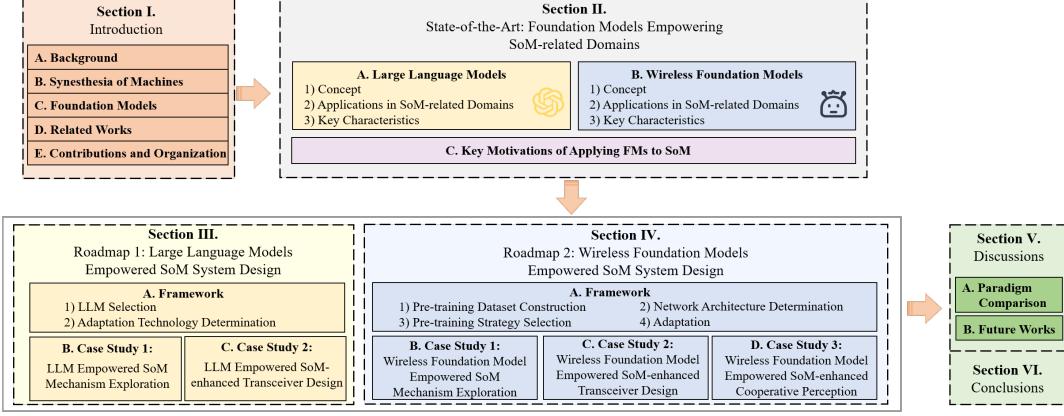


Fig. 2. The organization of this paper.

design. In this paper, inspired by existing studies on SoM-related domains empowered by FMs, we for the first time systematically categorize FMs into two types: general-purpose foundation models, specifically LLMs, and SoM domain-specific foundation models, namely wireless foundation models. Inspired by the superior capabilities of these two types of FMs in addressing the existing challenges of SoM systems, we propose two novel roadmaps for empowering SoM systems with FMs, i.e., LLM-based and wireless foundation model-based design. For each roadmap, we first present a detailed step-by-step framework and provide several case studies illustrating its application in SoM system design following the outlined framework. For SoM mechanism exploration, we propose LLM-based path loss generation (LLM4PG) and scatterer generation (LLM4SG) scheme, and wireless channel foundation model (WiCo). For SoM-enhanced transceiver design, we propose LLM-based wireless multi-task SoM transceiver (LLM4WM) and wireless foundation model (WiFo). For SoM-enhanced cooperative perception, we propose a wireless cooperative perception foundation model (WiPo). Preliminary simulation results validate the superiority of the proposed FMs empowered schemes. Finally, we adequately compare existing AI-empowered SoM system design paradigms and discuss potential future research directions.

As shown in Fig. 2, the paper is organized as follows: Section II introduces the current studies on SoM-related fields empowered by LLMs and wireless foundation models, and further derives several key motivations for applying FMs to SoM. Section III systematically illustrates the first research roadmap, which leverages LLMs to empower SoM and introduces some representative case studies, while Section IV presents the second roadmap, which utilizes wireless foundation models to empower SoM and also presents some representative case studies. Section V comprehensively compares the three AI-empowered SoM system design paradigms and outlines future research directions. Finally, Section VI concludes this paper.

II. STATE-OF-THE-ART: FOUNDATION MODELS EMPOWERING SOM-RELATED DOMAINS

In this section, we introduce existing SoM-related research enabled by two types of FMs. For each type of FMs, we first introduce its fundamental concept, summarize its applications

in SoM-related fields, and derive its key characteristics. Moreover, we summarize the advantages of two FMs in addressing the current challenges of SoM systems.

A. Large Language Models

1) *Concept*: LLMs [12] typically refer to language models containing hundreds of billions or more parameters, pre-trained on vast amounts of text data, which possess powerful capabilities in solving natural language tasks. In this paper, *LLMs broadly refer to general-purpose FMs, including multimodal large language models (MLLMs)* [22]. The underlying network of mainstream LLMs is the transformer [23] architecture, which excels in long-range dependency modeling, offers strong scalability and is well-suited for hardware parallelization. Based on the presence of an encoder or decoder, LLM architectures can be categorized into three types [24]: encoder-only, encoder-decoder, and decoder-only. Following the scaling law [25], LLMs have developed three key emergent abilities: in-context learning, instruction following, and step-by-step reasoning, fundamentally setting them apart from smaller models and signaling the emergence of artificial general intelligence (AGI).

The emergence of LLMs has not only revolutionized natural language processing but also empowered various scientific and engineering fields, including mathematics [26], chemistry [27], biology [28], and software engineering [29]. There are two main approaches to applying pre-trained LLMs to domain-specific tasks: fine-tuning [30] and prompt engineering [31]. Fine-tuning [30] is the process of refining pre-trained LLMs on a specific task or domain by adjusting its parameters to enhance task-specific performance. According to whether the input and output are in linguistic form, fine-tuning can be classified into linguistic fine-tuning and non-linguistic fine-tuning, with the former also known as instruction tuning [32]. Prompt engineering [31] is the process of strategically designing and optimizing input prompts to effectively guide LLMs toward generating desired outputs without modifying their underlying parameters. This technique leverages the model's pre-trained knowledge by structuring prompts in a way that enhances task performance, improves response accuracy, and aligns outputs with specific requirements. In addition, retrieval-augmented generation (RAG) and tool usage [33] are

two enhancement techniques for extending the functionality of LLMs. RAG improves LLMs by retrieving relevant information from external databases, ensuring responses remain accurate and contextually relevant. This technique enhances the model's ability to handle domain-specific queries and keeps its knowledge up to date. Tool usage allows LLMs to interact with external software, APIs, and computational tools, enabling capabilities beyond text generation. By leveraging such tools, LLMs can perform calculations, query structured data, and execute complex tasks efficiently.

2) *Applications in SoM-related Domains:* We systematically introduce existing works that leverage LLMs to empower SoM-related domains, focusing primarily on SoM dataset construction, SoM-enhanced transceiver design, and SoM-enhanced cooperative perception.

Multi-modal dataset construction for SoM: To address the challenge of limited high-quality training data while preserving user privacy, synthetic data has emerged as a potential solution [34]. Existing research has extensively explored how LLMs facilitate both language-oriented and non-language-oriented data synthesis. For language-oriented data synthesis, LLMs leverage their strong text-generation capabilities to augment datasets such as text classification [35], dialogue [36], multilingual commonsense [37], and depression interview transcripts [38]. Techniques like prompt engineering and chain-of-thought (CoT) guide LLMs to generate domain-specific text, including biomedical named entity recognition [39] and relation triplet extraction [40] datasets. In specialized fields such as law [41], medicine [42], and mathematical reasoning [43], integrating domain-specific knowledge bases or prior constraints enhances the fidelity of LLM-generated text. For non-language-oriented data synthesis, LLMs' semantic understanding enables the generation of data in other modalities, including tables [44], images [45], videos [46], and time-series [47] data. One approach involves fine-tuning LLMs on specific modalities to learn their distributions and generate synthetic data autoregressively [48]. Alternatively, LLMs can extract specific semantic features to guide pre-trained generators, such as diffusion models, in generating target-modality data that aligns with textual descriptions [49]. However, existing LLM-empowered synthetic data generation methods have not considered multi-modal sensing and communication dataset generation, making them unsuitable for directly supporting SoM system design.

SoM-enhanced transceiver design: Wireless networking is the most relevant research area for SoM-enhanced transceiver design, and its integration with LLMs [50, 51] has been widely explored. Based on the type of tasks empowered, existing LLM-driven research for wireless networks [52] can be categorized into language-oriented and non-language-oriented tasks. Leveraging the powerful language understanding and generation capabilities of LLMs, early studies directly utilize prompt engineering schemes to empower language-oriented telecom tasks, including information synthesis [53], code generation [54], transceiver configuration [55], and software log analysis [56] in the wireless network domain. To better facilitate domain knowledge transfer, several studies have attempted to adapt LLMs through instruction tuning for wireless tasks

such as telecom question and answer (QnA) [57], summarizing optimization problems [58], and network analysis [59]. To further enhance LLMs' decision-making and task execution capabilities, some instruction tuning empowered studies have integrated RAG [60] to leverage domain-specific knowledge in communications and have built LLM agents [61] by invoking APIs to achieve physical layer automation.

On the other hand, several studies leverage LLMs' in-context learning to handle non-linguistic tasks of wireless networks, including power control [62], symbol detection [63], wireless traffic prediction [64], and resource allocation [65]. However, since LLMs are not inherently skilled at mathematical reasoning, another mainstream approach is to fine-tune them for better domain adaptation. The pioneering work LLM4CP [66] is the first to fine-tune LLMs in a non-linguistic manner for wireless physical layer tasks, achieving improvements in both channel prediction accuracy and generalization performance. Based on the same idea, subsequent studies [67] have explored the application of LLMs in physical layer tasks such as beam prediction [68], CSI feedback [69], channel estimation [70]. Moreover, some studies have further explored integrating multi-modal information for networking [71], while other studies have attempted to fine-tune LLMs for multiple physical layer tasks simultaneously [72, 73]. Nevertheless, existing LLM-empowered wireless network designs ignore the integration with multi-modal sensing information, making them unsuitable for direct use in SoM transceiver design.

SoM-enhanced cooperative perception design: Efficient and robust information sharing among agents plays a vital role in cooperative perception. Deep learning enabled semantic communications have shown great potential to significantly improve transmission efficiency, which only transmits necessary information relevant to the specific task at the receiver [74]. LLMs have been harnessed to enhance semantic communication designs due to their remarkable semantic understanding capabilities, including both language-oriented and non-language-oriented semantic communications. For language-oriented semantic communications, some studies exploit LLMs' summarization and error-correction capabilities to preserve semantic equivalence against wireless channel effects, including language-level [75] and token-level [76–78] processing. To further improve the adaptability, some works utilize LLMs to evaluate the importance of features of small models and assign important features to good subchannels, which can be regarded as knowledge distillation from LLMs [79]. For non-language-oriented semantic communications, LLMs can be integrated with generative models to ensure consistent data generation between transceivers [80]. Some studies transmit both compact features of source data and text prompts generated from MLLMs, thereby generating high-fidelity data at the receiver with low data transmission overhead [81–84]. Besides, LLMs can also address task heterogeneity in multi-user semantic communication systems by employing parameter-efficient fine-tuning methods [85]. However, as LLM-based semantic communication relies on converting source data into linguistic representations, it is not well-suited for universal multi-modal cooperative perception.

3) *Key Characteristics:* Based on the comprehensive research on LLMs in the SoM-related domain, we can summarize the key characteristics of LLMs in solving SoM-related tasks as follows.

- *In-context learning:* LLMs exhibit strong in-context learning capabilities, enabling them to learn new tasks from task-specific prompts without fine-tuning. This greatly reduces the need for dedicated fine-tuning datasets and allows quick adaptation from a few examples.
- *Semantic understanding and generation:* Pre-trained on large-scale textual corpora, LLMs demonstrate deep semantic understanding and high-level abstraction. They can further autoregressively generate coherent, domain-specific content across fields such as literature and code.
- *General knowledge transfer:* LLMs naturally internalize a vast amount of world knowledge, enabling strong generalization to new tasks and domains with high adaptability and universality.

B. Wireless Foundation Models

1) *Concept:* In addition to leveraging general-purpose FMs, i.e., LLMs, another approach is to develop domain-specific FMs for the SoM system, termed wireless foundation models in this paper. Specifically, *a wireless foundation model is a model pre-trained on broad wireless and multi-modal sensing data (optional), typically at scale using self-supervision, and adaptable to a wide range of SoM-related tasks through few-shot or zero-shot learning.* In contrast to task-specific models, the pre-training paradigm and scale effects endow wireless foundation models with distinct capabilities, offering the potential to fundamentally transform SoM system design. For example, a single model can handle multiple related SoM processing tasks, significantly reducing the number of required models. Additionally, when data distribution shifts, wireless foundation model-based approaches enable efficient adaptation with minimal effort, drastically lowering the cost of data collection and model fine-tuning.

Nevertheless, developing wireless foundation models for SoM systems entails three key challenges:

- *Heterogeneity of modalities:* It is challenging to simultaneously process heterogeneous wireless data in SoM systems, such as CSI, channel impulse response (CIR), and IQ (In-phase and Quadrature) signals, along with diverse multi-modal sensing data, including LiDAR point clouds, RGB images, depth maps, and radar point clouds.
- *Complexity of tasks:* Compared to language tasks, SoM processing involves more complex task types that are difficult to be uniformly modeled as next-token prediction.
- *Scarcity of datasets:* Unlike readily available language and visual datasets, high-quality SoM datasets [8, 86] are challenging to obtain due to the need for precise alignment between wireless and multi-modal sensing data.

2) *Applications in SoM-related Domains:* Unlike the extensive use of FMs in CV and NLP, research on wireless foundation models for SoM remains in its early stages. We systematically introduce existing studies on domain-specific FMs

in SoM-related domains, including SoM dataset construction and SoM-enhanced transceiver design.

Multi-modal dataset construction for SoM: With powerful cross-modal generative abilities, FMs are particularly suited for synthetic data generation. We classify existing work by pre-training strategies, emphasizing methodological distinctions and their impact. Autoregressive modeling, based on next-token prediction as utilized in UniAudio [87], consolidates various tasks via tokenization, enabling robust task generalization. Masked learning methods, such as MaskGIT [88], enhance FMs' understanding of image data while enabling parallel decoding for significantly faster image generation. Diffusion models [89], leveraging their powerful generative capability and controllability via guidance signals, are widely applied in text-to-image tasks. Contrastive learning-based methods, such as SymTime [90] and VILA-U [91], facilitate efficient cross-modal alignment, improving cross-modal generation. Notably, these pre-training paradigms are not strictly distinct but can be integrated for synergistic benefits. For instance, MRGen [92] combines diffusion models with masked modeling for a region-controlled generation. Nevertheless, the application of FMs for SoM dataset generation remains unexplored.

SoM-enhanced transceiver design: Due to the powerful representation capabilities, self-supervised learning [93] has been applied to several sensing and communication tasks, including fingerprint localization [94], channel charting [95], wireless power control [96], beam mapping [97], signal classification [98], spectrum sensing [99], channel estimation [100], and geolocation-based MIMO transmission [101]. Nonetheless, these studies remain confined to single-task scenarios. Recognizing CSI as a versatile representation for diverse physical-layer tasks underscores the motivation for developing a wireless foundation model for multiple wireless tasks. Early work [102] first explored a pre-trained realistic channel model for wireless channel data, leveraging a BERT-like architecture and self-supervised pertaining. This model's adaptability across pilot contamination mitigation, channel compression, and channel fingerprinting offers preliminary evidence of its generalized understanding of wireless channels. Furthermore, the Large Wireless Model (LWM) [103] has been proposed to develop a foundation model based on Masked Channel Modeling (MCM) self-supervised learning for wireless channels and demonstrates its effectiveness across multiple downstream tasks, including cross-frequency beam prediction, LoS/NLoS classification, and robust beamforming. Nevertheless, LWM can only handle space-frequency two-dimensional CSI and still requires additional fine-tuning for downstream tasks. Therefore, the first wireless foundation model (WiFo) [104] for channel prediction was proposed to handle 3D CSI, which is pre-trained on extensive diverse CSI datasets and can be directly applied for inference without fine-tuning. It is the first versatile model capable of simultaneously addressing various channel prediction tasks across diverse CSI configurations and simulation results validate its remarkable zero-shot prediction performance. Furthermore, [105] explored a BERT-based wireless foundation model for channel prediction, while [106] investigated a prompt-enabled wireless foundation model for CSI feedback.

TABLE I
CHARACTERISTICS OF FMs FOR ADDRESSING THE CHALLENGES OF EXISTING SOM SYSTEM DESIGN

Challenges of existing SoM system design	Characteristics	
	LLMs	Wireless foundation models
Scarcity of massive and high-quality datasets	Powerful semantic understanding and autoregressive data generation capability	Powerful cross-modal generative capability
Limited modeling capability	Powerful few-shot modeling ability via in-context learning	Powerful modeling capability following scaling law
Limited generalization across data	Strong generalization ability benefit from general knowledge	One-for-all capability for heterogeneous data and powerful generalization capability via few-shot and even zero-shot learning
Limited universality across tasks	Efficient downstream task transfer learning enabled by general knowledge	One-for-all capability for diverse tasks

Several studies have explored building wireless foundation models to jointly handle CSI-related communication and sensing tasks. A joint-embedding self-supervised method for wireless channel representation learning [107] was proposed to learn invariant and compressed channel representations, which is fine-tuned for wireless localization and path loss generation. A Vision Transformer (ViT)-based radio foundation model [108] employs Masked Spectrogram Modeling (MSM) as a self-supervised learning approach for spectrogram learning and can be fine-tuned for human activity sensing and spectrogram segmentation tasks. Additionally, several BERT-based multifunctional wireless models [109, 110] have been proposed, which are first pre-trained in a self-supervised manner and then fine-tuned for CSI prediction, classification tasks, and WiFi sensing. A CLIP-based wireless foundation model [111] was proposed to simultaneously capture the joint feature representation of CSI and CIR and exhibits remarkable adaptability across various CSI-related tasks, including channel identification, positioning, and beam management. Unlike previous pre-training methods limited to a single CSI data type, a CSI-based multi-modal foundation model [112] is introduced, leveraging contrastive learning to align CSI with environmental contexts (BS/UE position and status) and derive task-agnostic CSI representations. Nevertheless, most existing wireless foundation models for wireless communications and sensing tasks remain limited to a single RF modality and have yet to integrate multi-modal sensing information for SoM-enhanced transmission systems.

3) Key Characteristics:

- **One-for-all capability:** Wireless foundation models operate in a one-for-all manner, enabling a single model to handle multiple tasks and heterogeneous data simultaneously, significantly reducing the number of models required for deployment.
- **Powerful modeling capability:** Following the scaling law, wireless foundation models possess enhanced modeling capabilities to capture complex relationships, enabling them to tackle more challenging SoM processing tasks.
- **Powerful generalization capability:** Wireless foundation models can generalize across different scenarios and tasks, facilitating few-shot and zero-shot learning, thereby reducing the need for additional retraining overhead.
- **Powerful generative capability:** Wireless foundation models possess powerful generative capabilities, promising high-quality synthesis of datasets for the integration of communication and multi-modal sensing.

C. Key Motivations of Applying FMs to SoM

In light of the derived key characteristics of the two types of FMs in addressing existing SoM-related tasks, they show great potential in addressing the existing challenges related to SoM system design, as shown in Table I.

- For the scarcity of SoM datasets, LLMs are expected to leverage their powerful semantic understanding and autoregressive data generation capabilities to synthesize language and multi-modal data, while wireless foundation models can generate accurately aligned multi-modal sensing-communication datasets through their powerful cross-modal generation ability.
- Regarding the challenges in modeling, LLMs possess superior few-shot learning capabilities through in-context learning, while wireless foundation models have a strong ability to handle complex SoM problems by following the scaling law.
- For data generalization difficulties, LLMs demonstrate strong generalization ability through general knowledge transfer, while wireless foundation models possess powerful multi-dataset joint learning capabilities and impressive zero-shot generalization ability.
- To address the challenge of limited task universality, LLMs enable effective transfer learning for downstream tasks, whereas wireless foundation models offer enhanced one-for-all capabilities, allowing them to tackle multiple tasks simultaneously.

Even so, the pipeline and detailed use cases of applying the two types of FMs to SoM system design are still lacking. Therefore, in Sections III and IV, we propose two roadmaps for FM-empowered SoM system design, utilizing LLMs and wireless foundation models, respectively. Specifically, the framework and several case studies of each roadmap are illustrated, to provide design guidance for researchers.

III. ROADMAP 1: LARGE LANGUAGE MODELS EMPOWERED SOM SYSTEM DESIGN

In this section, we present the first roadmap for designing a foundation model-empowered SoM system, which harnesses the capabilities of pre-trained LLMs via fine-tuning or prompt engineering. We begin by introducing a framework that highlights two core components of SoM system design: LLM selection and adaptation technology determination. We then present two case studies to demonstrate its practical implementation. The overall framework and the detailed designs of the two case studies are shown in Fig. 3.

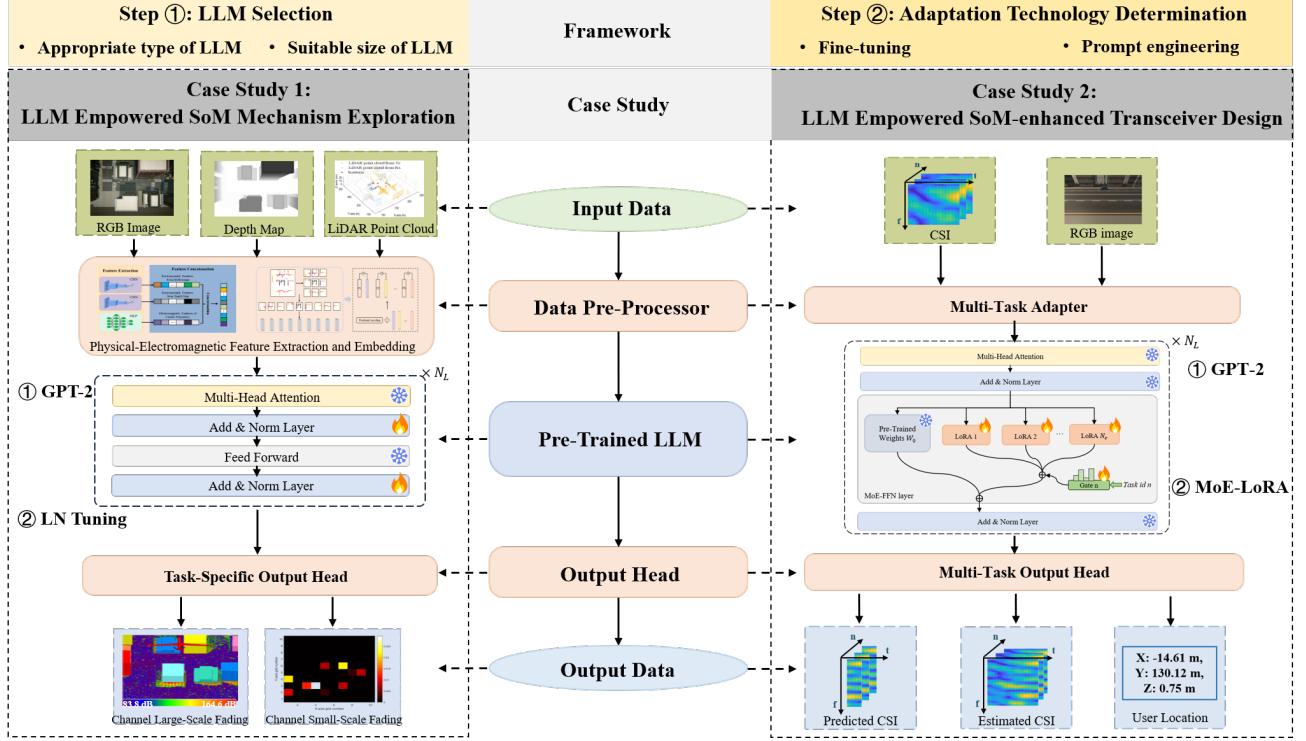


Fig. 3. An illustration of the framework of roadmap 1 and the proposed schemes for the two case studies introduced.

A. Framework

1) *LLM Selection*: The first step in tackling a specific SoM problem with LLMs is to determine the choice of LLM backbone, as it plays a foundational and decisive role in the task's overall performance. Two key aspects need to be carefully considered, i.e., the type and the size of the LLM.

- *Appropriate type of LLM*: To better leverage the LLM's understanding capabilities and facilitate general knowledge transfer, the type of LLM should closely align with the nature of the SoM task. In SoM tasks characterized by temporal dependencies, such as multi-modal sensing-assisted channel prediction, FMs designed for time-series data are more appropriate than language-based LLMs. For instance, study [70] reformulates the delay-Doppler (DD) domain channel prediction problem as a time series prediction task of DD-domain parameters, demonstrating that time-series FMs like Timer [113] and Time-Mixture of experts (MoE) [114] can be directly applied and subsequent fine-tuning on specific vehicular channel data further improves prediction accuracy.

- *Suitable size of LLM*: The model size [115] is another critical factor that directly impacts both the performance and the deployment feasibility of SoM systems. Larger LLMs offer stronger generalization and cross-modal reasoning abilities, making them ideal for complex SoM tasks involving high-dimensional, multi-modal data. However, their high computational and memory demands limit real-time applicability in resource-constrained environments. Conversely, smaller LLMs [66] provide lower latency and are easier to fine-tune for domain-specific tasks, making them more suitable for real-time inference.

In addition, it is noted that the performance of this scheme initially improves and then gradually levels off as the size of the LLM increases. For instance, for networking optimization based on LLM fine-tuning [71], an LLM with 1 billion parameters is already sufficient. Techniques like knowledge distillation [116] and quantization can further reduce large model sizes while preserving their core capabilities, striking a balance between efficiency and performance.

2) *Adaptation Technology Determination*: Once the appropriate LLM is selected, the next step is to determine the adaptation technology for the specific SoM task. The choice of adaptation method significantly influences the system's effectiveness and generalization. Broadly, the adaptation strategies can be categorized into two approaches: fine-tuning and prompt engineering.

- *Fine-tuning*: Fine-tuning updates the LLM's internal parameters on SoM-specific data, enhancing its cross-domain transferability for specific SoM tasks. It includes both non-linguistic fine-tuning, which adapts the LLM to process non-textual SoM data, and instruction tuning, which enhances its ability to follow SoM domain-specific instructions. Although fine-tuning requires additional labeled data, it achieves superior performance over task-specific models in few-shot scenarios. For example, in cross-frequency generalization tests, the LLM-based channel prediction scheme [66] requires only 30 CSI samples to outperform model-based schemes [117], whereas other task-specific models need more than 100 CSI samples.
- *Prompt engineering*: Unlike fine-tuning, prompt engineer-

ing guides the LLMs to perform SoM tasks through carefully designed prompts, avoiding the need for additional parameter updates. By preserving the original parameters of the LLMs, prompt engineering fully leverages the model's inherent semantic understanding capabilities for multi-modal information fusion in SoM systems. For instance, in SoM feature transmission tasks, multi-modal LLMs can map diverse data types (e.g., text, images, and LiDAR point clouds) into a unified semantic space [80], enabling more efficient encoding and transmission. To overcome the LLM's limitations in numerical reasoning and cross-domain adaptation, RAG and external tool integration enable access to external databases and APIs for enhanced accuracy and effectiveness. These capabilities make prompt engineering a lightweight yet powerful adaptation technique for SoM systems, particularly when large-scale fine-tuning is impractical.

B. Case Study 1: LLM Empowered SoM Mechanism Exploration

As the theoretical foundation of SoM research, it is essential to explore the complex and nonlinear SoM mechanism, i.e., the mapping relationship between communications and sensing [118, 119]. On one hand, we explore the SoM mechanism between sensing and path loss, and further propose LLM4PG as an LLM-based scheme for path loss generation. On the other hand, we explore the SoM mechanism between sensing and multipath fading, and further propose LLM4SG as an LLM-based scheme for scatterer generation [120]. The framework of the proposed LLM4PG/LLM4SG scheme, i.e., case study 1, is given in Fig. 3.

- *Step 1: LLM Selection.* In the SoM mechanism exploration, we need to select an appropriate LLM as the backbone from two perspectives, including task adaptability and model performance/efficiency. For task adaptability, existing pre-trained LLMs are not specifically designed for SoM mechanism exploration tasks. In this case, we select a mature LLM with sufficient generality and extensibility. For the model performance/efficiency, the model generalization capability in transfer learning and its inference overhead are considered to meet the high-precision requirement of SoM mechanism exploration tasks. Therefore, we select the lightweight GPT-2 [121] as the backbone in the LLM4PG and LLM4SG schemes.
- *Step 2: Adaptation Technology Determination.* Since the SoM mechanism exploration is non-language-oriented, the fine-tuning approach is more suitable for addressing these challenges. On one hand, we bridge the significant gap between the natural language domain and the multi-modal information domain. For RGB-D images, we extract physical environment features and employ feature-level fusion to map these features into the natural language domain. For LiDAR point clouds, we perform voxelization preprocessing followed by patch partitioning and positional encoding to convert the data into tokens compatible with the LLM feature space. Meanwhile, an output module transforms the GPT-2 generated token se-

TABLE II
THE NUMBER OF NETWORK PARAMETERS (TRAINING PARAMETERS/TOTAL PARAMETERS) AND THE INTERFERENCE TIME PER BATCH (BATCH SIZE IS SET TO 8) OF CASE STUDY 1 FOR ROADMAP 1

	Parameters (M)	Inference time (ms)
LLM4PG	52.23/275.70	9.90
GAN	45.61/45.61	7.36
LLM4SG	5.08/86.19	7.96
ResNet	23.53/23.53	5.67

TABLE III
THE NUMBER OF NETWORK PARAMETERS (TRAINING PARAMETERS/TOTAL PARAMETERS) AND THE INTERFERENCE TIME PER BATCH (BATCH SIZE IS SET TO 8) OF CASE STUDY 2 FOR ROADMAP 1

	Parameters (M)	Inference time (ms)
SM-STL	1.92/1.92	1.16
SM-MTL	1.92/1.92	1.81
LLM4WM	2.20/84.10	8.73
SM-STL-RGB	2.29/2.29	3.17
SM-MTL-RGB	2.29/2.29	3.30
LLM4WM-RGB	4.50/98.10	9.08

quences into required channel fading information, including path loss in the LLM4PG scheme and scatterers in the LLM4SG scheme. On the other hand, we adopt the fine-tuning strategy where most pre-trained LLM parameters remain frozen, and further transfer general knowledge to the SoM mechanism exploration task. Specifically, we employ LN Tuning [122], where only the LayerNorm parameters are set as trainable to reduce computational overhead while preserving model generality.

The proposed LLM4PG and LLM4SG schemes are trained on the SynthSoM dataset [86]. The advantage of the proposed LLM4PG and LLM4SG schemes is shown in Fig. 4. For the proposed LLM4PG scheme, Figs. 4(a)–(c) show that the LLM, i.e., GPT-2, demonstrates superior accuracy over the task-specific model, i.e., generative adversarial network (GAN), in the exploration of SoM mechanism between sensing and path loss. Compared with the GAN-based scheme, the proposed LLM4PG scheme demonstrates higher accuracy in reconstructing building edges in path loss maps, particularly for fading caused by buildings. For the proposed LLM4SG scheme, Figs. 4(d) and (e) evaluate the generalization capability of the LLM, i.e., GPT-2, and the task-specific model, i.e., ResNet, in the exploration of the SoM mechanism between sensing and small-scale fading. The model is first trained on sub-6 GHz samples and then fine-tuned with a subset of 28 GHz samples for transfer testing. The proposed LLM4SG scheme, through knowledge transfer, achieves over 11% higher generalization accuracy compared to the ResNet-based scheme, reaching the performance of the ResNet-based scheme trained on the full dataset with about 7% of the training samples. For all case studies in this paper, the inference time is evaluated on the same machine with an NVIDIA GeForce RTX 4090 GPU. Then, the complexity of the aforementioned scheme, i.e., the number of parameters and the average inference time, is given in Table II. It can be seen from Table II that the LLM4PG and LLM4SG schemes show an inference speed closely matching that of task-specific models, i.e., GAN-based and ResNet-based schemes.

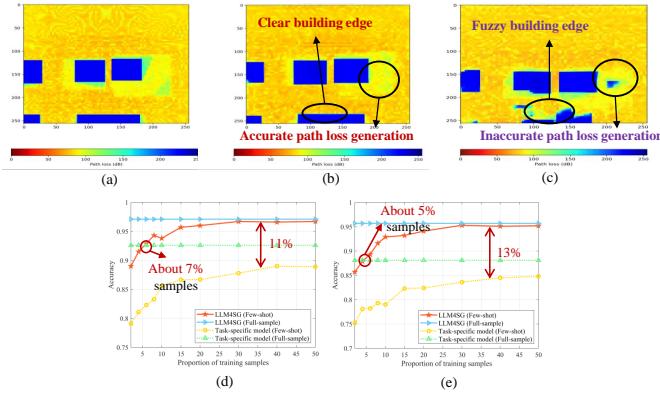


Fig. 4. Performance of SoM mechanism exploration. (a) Ray-tracing-based path loss map. (b) Path loss map result via the LLM4PG scheme. (c) Path loss map result via the GAN-based scheme. (d) Scatterer number results via the LLM4SG scheme and the ResNet-based scheme. (e) Scatterer location results via the LLM4SG scheme and the ResNet-based scheme.

C. Case Study 2: LLM Empowered SoM-enhanced Transceiver Design

We consider several common vision-aided tasks in SoM transceiver design, including channel estimation [123], prediction [66], and user positioning [124]. A multi-task learning approach is employed to exploit the synergy among these SoM tasks, thereby maximizing the spectral efficiency of the communication system. We will elaborate on how to effectively design an LLM-based wireless multi-task SoM transceiver by following the two fundamental steps outlined above.

- **Step 1: LLM Selection.** We first need to select a suitable LLM as the backbone based on task requirements, considering both the type and size of the LLM. As existing pre-trained LLMs are not specifically tailored to the characteristics of SoM tasks, we adopt a relatively mature LLM. Furthermore, while large-scale LLMs generally excel in transfer learning and zero-shot generalization, their high dimensionality and deep architectures introduce inference overhead that is unacceptable for physical layer tasks. Therefore, we choose the lightweight GPT-2 [121] as the backbone.

- **Step 2: Adaptation Technology Determination.** Since the selected tasks are non-language-oriented, the fine-tuning approach is more suitable for addressing these challenges. Specifically, the Mixture of experts with low-rank adaptation (MoE-LoRA) [72] method is employed for multi-task fine-tuning, effectively mitigating task conflicts while enabling efficient transfer of GPT-2's general knowledge. To align the LLM's output with both the label dimensions and the feature space, task-specific adapters with a ResNet-style [125] architecture are introduced at the output stage.

Based on the above two steps of analysis, the framework of LLM4WM is illustrated in Fig. 3. Then, we employ the SynthSoM [8] dataset to acquire aligned RGB images, vehicle Global Positioning System (GPS) data, and CSI.

The advantages of the proposed LLM4WM scheme are illustrated in Fig. 5, where comparisons are made between the

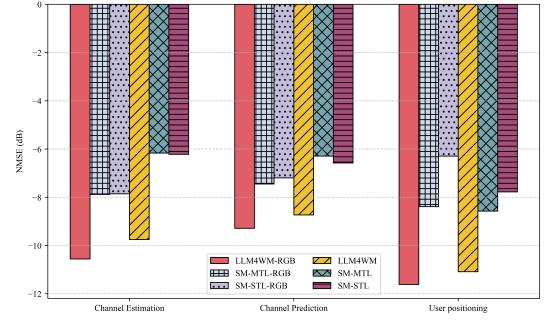


Fig. 5. NMSE performance comparison of the proposed LLM4WM scheme with the SM-MTL and SM-STL schemes across three SoM-related tasks.

specialized model single-task learning (SM-STL) scheme and the specialized model multi-task learning (SM-MTL) scheme. Furthermore, to compare the gains from vision assistance, we conducted experiments for each scheme both with and without vision input, and the vision-assisted variants are denoted with the suffix “-RGB”, while the versions without vision input retain the original scheme name. The Cross-Stitch network [126] is adopted for the SM-MTL scheme, while the SM-STL scheme also follows the configuration in [126], maintaining the same network architecture as the multi-task learning scheme but being trained and tested on individual tasks. It can be observed that vision-assisted schemes exhibit consistent performance gains relative to their non-vision-assisted counterparts, underscoring the contribution of visual information to transmission tasks. Furthermore, constrained by limited modeling capacity, the SM-MTL-RGB scheme is unable to leverage joint multi-task information, resulting in performance comparable to that of SM-STL-RGB. By contrast, the LLM4WM-RGB framework achieves state-of-the-art (SOTA) performance across all tasks, attributed to the extensive general knowledge encoded within LLM, thereby exhibiting superior task generalization. We also evaluate the complexity of each scheme in terms of the number of parameters and the average inference time across the three tasks, as summarized in Table III. Notably, LLM4WM-RGB demonstrates an inference speed closely matching that of specialized models. Although only three SoM-related tasks are selected in this case study, the LLM-based scheme is capable of handling a larger number of tasks, achieving even better multi-task learning performance, and enhancing the efficiency of utilizing the general knowledge embedded in LLMs [72].

IV. ROADMAP 2: WIRELESS FOUNDATION MODELS EMPOWERED SOM SYSTEM DESIGN

In this section, we introduce the second roadmap, which is to build a wireless foundation model from scratch tailored to specific SoM domains. We present a framework with four key steps: dataset construction, network architecture, pre-training strategy, and adaptation, followed by three case studies on building wireless foundation models for specific SoM tasks. The overall framework of the second roadmap and the network processing of the three case studies are shown in Fig. 6.

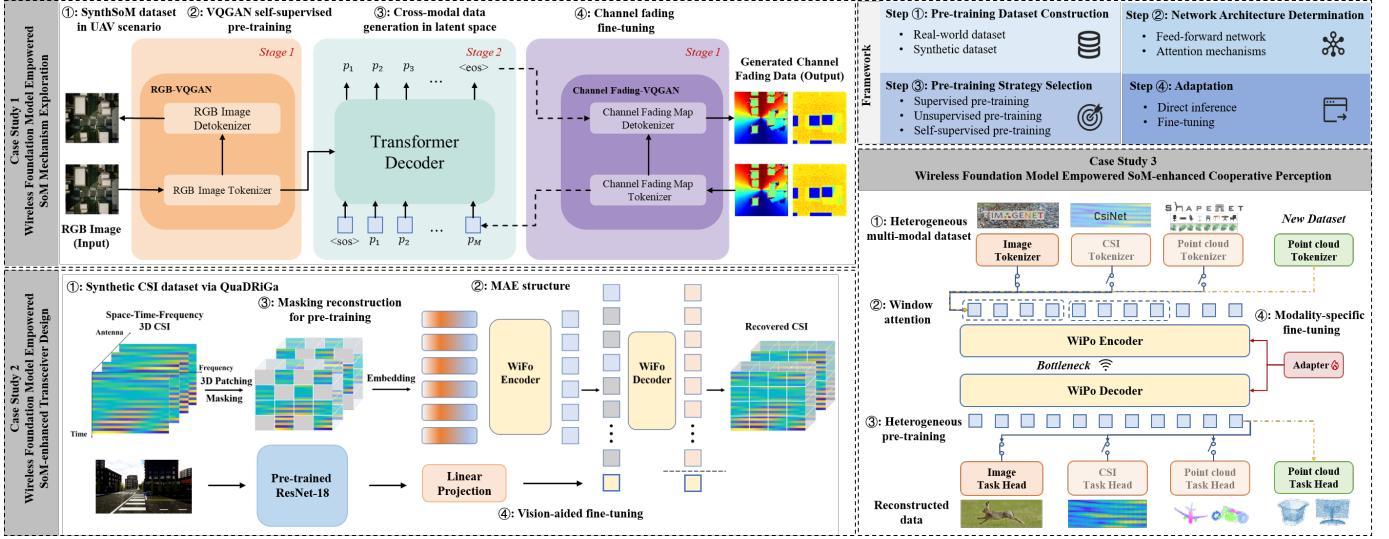


Fig. 6. An illustration of the framework of roadmap 2 and the proposed schemes for the three case studies introduced.

A. Framework

1) *Pre-training Dataset Construction*: The scale and quality of the dataset determine the performance upper bound of the wireless foundation model built from scratch for the SoM system [127]. Unlike readily available language and visual datasets, the construction of datasets for SoM systems needs to collect precisely aligned communication and multi-modal sensing data from the same physical environment, such as RF communication data, e.g., channel matrices and path loss, RF sensing data, e.g., mmWave radar point clouds, and non-RF sensing data, e.g., RGB images, depth maps, and LiDAR point clouds, resulting in huge challenge. In general, the existing multi-modal sensing-communication dataset for SoM can be classified into the real-world dataset obtained via measurement equipment, the synthetic dataset collected via simulation software, and the generated dataset via artificial intelligence generated content (AIGC) models based on the SoM mechanism.

- *Real-world dataset*: With the help of measurement equipment, the real-world dataset is of high accuracy. A measurement dataset, named KITTI, for multi-modal sensing was developed in [128]. The calibrated and synchronized RGB image, depth map, and LiDAR point cloud were collected in the KITTI dataset. To further include communication channel data, a measurement dataset, named DeepSense 6G, was constructed in [129], including multi-modal sensing data, e.g., RGB image and LiDAR point cloud, and communication channel data under the sub-6 GHz band and mmWave band. Although the real-world dataset facilitates the algorithm validation, it is of huge difficulty to customize multi-modal sensing and communication scenarios attributed to the cost concern.
- *Synthetic dataset*: Attributed to the limitations of the real-world dataset, many synthetic datasets have been constructed as a supplement to the real-world dataset. With the help of accurate software, multi-modal sensing and communication data can be efficiently collected in the synthetic dataset. The synthetic dataset ViWi was

constructed in [130], including channel information, RGB images, depth maps, and LiDAR point clouds. However, mmWave radar data was ignored in [130]. To overcome this limitation, a synthetic dataset, named M³SC, was developed in [86], which contained multi-modal sensing and communication data in various vehicular scenarios. To further include unmanned aerial vehicle (UAV) scenarios, a real-world data injected synthetic multi-modal sensing-communication dataset, named SynthSoM, was developed in [8]. The main limitation of the synthetic dataset is accuracy constraints due to some unrealistic assumptions in the collection software.

- *Generated dataset*: Another approach to construct the multi-modal sensing-communication dataset for SoM is based on AIGC models in conjunction with SoM mechanisms. For the generation of multi-modal sensory data, extensive effective AIGC models can be utilized, which have the ability to leverage generative architectures to autonomously generate high-quality and diverse multi-modal sensory data by learning underlying data distributions [131, 132]. For the generation of communication channel data, the explored SoM mechanism can be utilized to achieve efficient and high-fidelity cross-modal generation of communication channel data, which is temporally and spatially consistent with the multi-modal sensory data [120]. The primary limitation in the generated dataset lies in the dependency of multi-modal data generation accuracy on both the AIGC model performance and the precision of SoM mechanism exploration.

2) *Network Architecture Determination*: Transformer [23] has revolutionized deep learning by enabling models to capture long-range dependencies with high computational efficiency. Its self-attention mechanism allows for parallel processing of sequences, making it particularly effective for pre-training on large-scale data. Therefore, Transformer has become the mainstream architecture for LLMs, underpinning advancements in natural language processing [121], computer vision [133], robotics [134] and related fields. However, SoM tasks differ

from classical tasks and require network architectures to be adapted to address their unique characteristics. Potential areas for improvement primarily include the feed-forward network (FFN) and attention mechanisms.

- *Feed-forward network:* To further improve modeling capacity and representation capability, some studies replace the FFN with other networks. A typical approach involves using a Mixture of Experts (MoE) to substitute the FFN, thereby achieving superior multi-task performance with faster inference speed, such as switch transformer [135], mixtral of experts [136] and deepleek-MoE [137]. Another representative variant is Gated Linear Units (GLU) [138] applied in the LLaMa [139], which is straightforward to implement and yields superior performance. By employing structures with enhanced representational capabilities, such as MoE, the model can better learn the interrelationships among various SoM tasks, thereby achieving improved generalization and performance.
- *Attention mechanisms:* The attention module is a key block in the transformer, enabling the model to capture complex contextual relationships between input tokens. However, it also accounts for a significant portion of the computation load in the transformer. Many works propose new attention mechanisms to improve computational efficiency, including flash attention [140], grouped-query attention [141], window attention [142] and so on. Beyond efficiency, appropriate attention mechanisms can also enhance model performance. For example, the Swin Transformer introduces stronger inductive biases tailored to visual tasks [142]. In the context of SoM tasks, which require both efficient communication and accurate sensing, timeliness is often critical. Therefore, optimizing the attention mechanism represents a promising direction for balancing task performance with runtime efficiency.

3) *Pre-training Strategy Selection:* Selecting an appropriate pre-training strategy is crucial for model performance, as different pre-training approaches directly affect the model's ability to learn from datasets, thereby influencing its performance across various SoM tasks. Based on the type of supervision, pre-training strategies can be categorized into three types: supervised pre-training, unsupervised pre-training, and self-supervised pre-training.

- *Supervised pre-training:* Supervised pre-training learns directly from labeled, high-quality datasets, enabling the model to acquire strong task-specific capabilities. For example, ResNet [125] is pre-trained on the large-scale image dataset ImageNet [143], which contains over 1 million labeled images across 1,000 categories, enabling the model to learn rich hierarchical visual representations and extract high-level semantic features. However, supervised pre-training heavily depends on massive labeled datasets, which incurs high annotation costs, and the performance of the pre-trained model is also affected by the quality of the labels.
- *Unsupervised pre-training:* Unsupervised pre-training enables the model to learn meaningful representations of data by leveraging rule-based proxy tasks, without re-

quiring manually annotated labels. It is commonly used for dimensionality reduction, clustering, and has been applied to various SoM-related tasks, such as multi-user precoding [144]. Specifically, the multi-user precoding network trained with unsupervised learning does not require a ground truth precoding matrix. Instead, it directly optimizes spectral efficiency by learning the intrinsic representations of the precoding task, thereby improving its generalization capability.

- *Self-supervised pre-training:* Self-supervised pre-training generates training labels by leveraging the inherent structure or information within the data. Examples include next token prediction [121], masked modeling [145], and contrastive learning [146]. This approach also eliminates the need for manual annotations while providing excellent scalability and zero-shot generalization capabilities. For instance, In [112], contrastive learning is utilized to align CSIs with associated environment descriptions characterized by BS positions, enabling the well-aligned representations to be directly applied in downstream classification tasks, such as LoS/NLoS identification.

4) *Adaptation:* To ensure the wireless foundation model effectively generalizes across diverse SoM tasks and scenario conditions, adaptation techniques must be carefully designed. The adaptation strategies primarily include direct inference and fine-tuning, which determine how the model transfers knowledge to unseen tasks and scenarios.

- *Direct inference:* Compared with LLM-based solutions, the wireless foundation model exhibits strong zero-shot generalization ability, meaning that the wireless foundation model could generalize to new data distributions without any additional labeled data. This is particularly useful for dynamically changing environments where collecting task-specific training data is impractical. By leveraging pre-trained representations, wireless foundation models for SoM can infer relevant patterns from unseen data distributions. For instance, [104] demonstrated that large-scale pre-trained models exhibit strong zero-shot capabilities for channel prediction.
- *Fine-tuning:* Fine-tuning further enhances the model's adaptability by allowing it to learn new data distributions or even new tasks quickly with minimal labeled examples. This is particularly beneficial for SoM scenarios where labeled data is scarce or expensive to acquire. Additionally, parameter-efficient tuning (PEFT) techniques like LoRA [71] and MoE-LoRA [72] facilitate efficient knowledge transfer while minimizing computational overhead, making them well-suited for real-time adaptation in SoM systems.

The deployment and adaptation of wireless foundation models for downstream tasks can be further enhanced via a cloud-edge-terminal collaborative architecture [147]. In this framework, large-scale wireless foundation models are centrally maintained and periodically updated in the cloud, leveraging global knowledge and computational resources. To enable efficient on-device inference and adaptation, techniques such as knowledge distillation and parameter-efficient fine-tuning

TABLE IV
THE NUMBER OF NETWORK PARAMETERS (TRAINING PARAMETERS/TOTAL PARAMETERS) AND THE INTERFERENCE TIME PER BATCH (BATCH SIZE IS SET TO 8) OF CASE STUDY 1 FOR ROADMAP 2

	Parameters (M)	Inference time (ms)
WiCo-PG	29.21/70.56	5.92
LLM-based	16.36/63.86	5.26
GAN	45.61/45.61	4.36
WiCo-MG	25.87/76.85	7.24
LLM-based	12.39/68.48	6.51
ResNet	43.09/43.09	5.29

are employed to compress and transfer the core capabilities of cloud models to lightweight edge or terminal models. Moreover, federated learning can facilitate privacy-preserving model adaptation by allowing terminals to locally update model parameters based on their data, with only the aggregated model updates being sent to the cloud. This collaborative paradigm ensures continuous model evolution, low-latency adaptation, and efficient resource utilization, thereby supporting scalable and secure deployment of wireless foundation models in heterogeneous real-world environments.

B. Case Study 1: Wireless Foundation Model Empowered SoM Mechanism Exploration

To explore the complex and nonlinear SoM mechanism between sensing and communications, we propose the wireless channel foundation model for the first time. Given the scarcity of channel data compared to sensory data, by exploring the SoM mechanism, the WiCo scheme leverages more accessible sensory data to achieve efficient and high-fidelity cross-modal generation of channel data. The proposed WiCo scheme contains two parts, including WiCo for path loss generation (WiCo-PG) and WiCo for multipath component generation (WiCo-MG). For clarity, we elaborate on the proposed WiCo-PG scheme, with the WiCo-MG scheme adopting a similar methodology. The framework of the proposed WiCo-PG and WiCo-MG schemes, i.e., case study 1, is shown in Fig. 6.

Step 1: Pre-training Dataset Construction. Based on the requirement of the WiCo-PG scheme, we investigate the existing multi-modal sensing-communication dataset with diverse scenarios. Considering the volume and quality of the dataset, we utilize the SynthSoM dataset in [8] to develop the WiCo-PG scheme. The SynthSoM dataset covers various air-ground multi-link cooperative scenarios with diverse data modalities, such as the RF communication, i.e., 140K sets of channel matrices and 18K sets of path loss, RF sensing, i.e., 136K sets of mmWave radar waveforms with 38K radar point clouds, and non-RF sensing, i.e., 145K RGB images, 290K depth maps, as well as 79K sets of LiDAR point clouds.

Step 2: Network Architecture Determination. To select a proper network architecture of the WiCo-PG scheme, the accuracy and generalization need to be considered. To explore the SoM mechanism between RGB images and path loss maps, we enhance the Pathways Autoregressive Text-to-image (Parti) generative model architecture proposed by Google, which integrates the visual discrete representation of the VQGAN network with the autoregressive generation capability of the transformer.

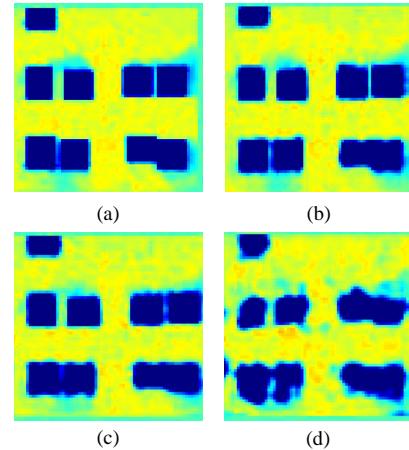


Fig. 7. Comparisons of path loss map results. (a) Ray-tracing-based result. (b) The WiCo-PG scheme. (c) The LLM-based scheme. (d) The GAN-based scheme.

Step 3: Pre-Training Strategy Selection. To select a proper pre-training strategy, it is essential to adopt self-supervised learning tailored for the accurate exploration of SoM mechanism between sensing and path loss. Through data augmentation and noise suppression training methods, we can effectively extract intrinsic data features and enhance the robustness of the proposed WiCo-PG scheme.

Step 4: Adaptation. For the model adaptation, the focus lies in achieving a smooth transition of the pre-trained WiCo-PG to new datasets by training a small number of parameters. Furthermore, the pre-trained WiCo-PG can be fine-tuned on different scenarios and frequency bands by efficiently leveraging the pre-trained knowledge of channel fading generation tasks. Then, model parameters are optimized for the accurate exploration task of SoM mechanism between sensing and path loss, thus enhancing domain-specific accuracy.

The WiCo-PG scheme is trained on the SynthSoM dataset [8] with path loss maps under 28 GHz frequency band and UAV images. Figs. 7(a) and (b) demonstrate a close agreement between the ray-tracing-based path loss map and the generated path loss map via the WiCo-PG scheme through the powerful cross-modal generation ability of FMs. Specifically, the generated path loss map via the WiCo-PG scheme accurately identifies and reconstructs building contours, and further precisely generates path loss caused by buildings. Compared with the LLM-based scheme based on GPT-2 in Fig. 7(c), the WiCo-PG scheme achieves over 3% higher accuracy in path loss generation. Furthermore, a huge difference between the ray-tracing-based path loss map and the generated path loss map via the task-specific model, i.e., GAN, can be observed in Figs. 7(a) and (d), where path loss generation is blurred on building contours. The parameter size and inference time of the aforementioned schemes are listed in Table IV. The WiCo-PG scheme and the GAN-based scheme have the same order of magnitude regarding storage and computational overhead.

Similar to the WiCo-PG scheme, the WiCo-MG scheme also contains four main steps, including pre-training dataset construction, network architecture determination, pre-training strategy selection, and adaption. For the pre-training dataset construction, our constructed SynthSoM dataset [8] in the UAV

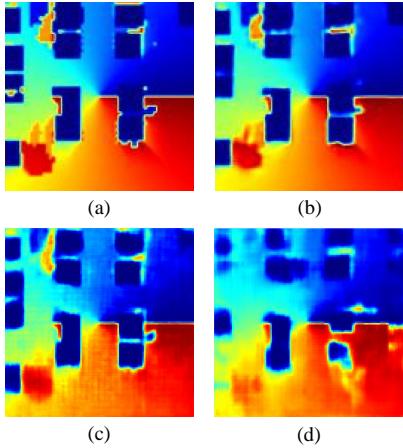


Fig. 8. Comparisons of multipath parameter results. (a) Ray-tracing-based result. (b) The WiCo-MG scheme. (c) The LLM-based scheme. (d) The ResNet-based scheme.

TABLE V
THE NUMBER OF NETWORK PARAMETERS (TRAINING
PARAMETERS/TOTAL PARAMETERS) AND THE INTERFERENCE TIME PER
BATCH (BATCH SIZE IS SET TO 8) OF CASE STUDY 2 FOR ROADMAP 2

	Parameters (M)	Inference time (ms)
WiFo (zero-shot)	0/21.60	8.74
WiFo (CSI)	2.16/21.60	8.74
WiFo (CSI+RGB)	2.68/33.71	10.90
LLM4WM (CSI)	2.20/84.10	8.67
LLM4WM (CSI+RGB)	4.50/98.10	9.08
Task-specific (CSI)	14.84/21.60	8.74
Task-specific (CSI+RGB)	15.36/33.71	10.90

scenario is utilized. For the network architecture determination, we utilize VQGAN and Transformer attributed to its visual discretized representation and autoregressive generation capability. For the pre-training strategy selection, we also exploit the self-supervised learning tailored for the high-fidelity and efficient multipath parameter generation. For the model adaption, the pre-trained network is fine-tuned on different scenarios and frequency bands by leveraging the pre-trained knowledge efficiently. As shown in Table IV, the WiCo-MG scheme and the ResNet-based scheme are also with the same order of magnitude regarding storage and computational overhead.

The WiCo-MG scheme is trained on the SynthSoM dataset [8] with multipath parameters, i.e., departure of angle (DoA), under 28 GHz frequency band and UAV images. Figs. 8(a) and (b) show a close consistency between the ray-tracing-based DoA and the generated DoA via the WiCo-MG scheme through the efficient cross-modal generation ability of FMs. Compared with the LLM-based scheme based on GPT-2 in Fig. 8(c), the WiCo-MG scheme in Fig. 8(b) achieves over 5% higher accuracy in multipath parameter generation. In addition, in Figs. 8(a) and (d), the ray-tracing-based DoA and the generated DoA via the task-specific model, i.e., ResNet, are exceedingly different, where the building edges are blurred and the DoA in the occluded regions behind buildings is notably inaccurate.

C. Case Study 2: Wireless Foundation Model Empowered SoM-enhanced Transceiver Design

In this case study, we consider vision-aided frequency-domain channel prediction powered by wireless foundation models. Due to the limited scale of existing multi-modal sensing and communication datasets, it is challenging to pre-train a multi-modal wireless foundation model from scratch. Therefore, we first build a CSI-oriented wireless foundation model, termed WiFo [104], and then fine-tune it using vision data for frequency-domain channel prediction in new scenarios, as shown in Fig. 6.

- *Step 1: Pre-training Dataset Construction.* CSI datasets can be obtained through real-world measurements, ray-tracing simulations, and statistical channel modeling. Existing measurement datasets [148, 149] for channel prediction are limited in scale and diversity, constraining the performance of pre-trained models. While ray-tracing simulations offer flexibility, they come with high computational costs. Therefore, we leverage the QuaDRiGa channel generator to generate a large-scale 3D CSI dataset compliant with 3GPP standards, containing over 160k samples. The dataset covers 16 heterogeneous scenarios and system configurations, with more details provided in [104].
- *Step 2: Network Architecture Determination.* It is worth noting that CSI data and video are quite similar in type, both being structured and continuous 3D data. Inspired by the success of masked autoencoders (MAE) in image and video pre-training, we propose an MAE-based network for CSI reconstruction. As shown in Fig. 6, diverse CSI data is first transformed into varying token numbers via 3D patching and embedding, facilitating processing by transformer blocks. For both the encoder and decoder, we introduce a novel positional encoding structure (STF-PE) to capture the 3D positional information.
- *Step 3: Pre-training Strategy Selection.* Noticing that the CSI prediction and the masking reconstruction pre-training task are similar, we adopt a masking-based self-supervised pre-training approach to enable WiFo with general reconstruction capabilities. Specifically, in addition to random masked reconstruction, we also design time and frequency domain masked reconstruction pre-training tasks to enhance the model's ability for both the time and frequency domain channel prediction.
- *Step 4: Adaptation.* The pre-trained WiFo can be directly used or fine-tuned for frequency-domain channel prediction on specific scenarios. We consider a fine-tuning approach with visual information enhancement, utilizing aligned RGB and CSI sample pairs. Since WiFo is designed to handle 3D CSI, we first concatenate the 2D CSI along the time dimension to transform it into a three-dimensional format. Specifically, the visual information, processed by the pre-trained ResNet-18 [125], is mapped to a token through a fully connected layer and concatenated to the input tokens of the WiFo decoder. During fine-tuning, only the additional fully connected layer of ResNet-18, the first layer of the WiFo decoder,

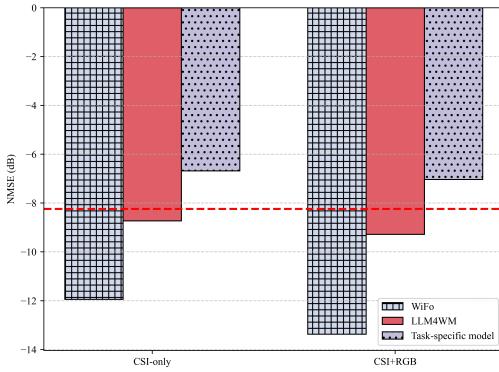


Fig. 9. NMSE performance comparison of WiFo-based, LLM-based, and task-specific model-based schemes.

and the final output layer are trainable, while the rest of the network is frozen to retain general knowledge.

We utilize the SynthSoM dataset [8] for fine-tuning, which includes 1,500 aligned CSI and RGB sample pairs with 16 antennas and 64 subcarriers. We aim to predict the CSI of the successive 32 subcarriers based on the first 32 subcarriers, using NMSE as the metric for prediction accuracy. For the task-specific models used as baselines, the WiFo encoder is randomly initialized and participates in training, while the other components are the same as those in the WiFo-based scheme. Simulation results of WiFo-based, LLM-based, and task-specific model-based schemes are illustrated in Fig. 9. It can be observed that whether using only CSI or combining it with RGB, the WiFo-based approach significantly outperforms the task-specific approach and the LLM-based scheme. This indicates that the pre-trained wireless foundation model has stronger few-shot learning capabilities than task-specific models and can quickly adapt to specific scenarios and demonstrates its superior performance compared to the LLM-based scheme. Furthermore, the zero-shot performance of WiFo achieves better results than the task-specific models and performance comparable to LLM4WM, suggesting that even without fine-tuning, WiFo can still deliver acceptable performance in new scenarios. The parameter size and inference time of the above schemes are shown in Table V. It can be observed that the increase in model parameters and inference time brought by the introduction of visual information is relatively limited, demonstrating the feasibility of applying multi-modal sensing information to practical transceiver design.

D. Case Study 3: Wireless Foundation Model Empowered SoM-enhanced Cooperative Perception

Given the massive and diverse data transmission demands in multi-agent communication networks, it is essential to develop FMs capable of supporting multi-modal data transmission for cooperative perception. Compared to training separate models for each modality, FMs offer greater generality and scalability, substantially reducing the cost and complexity of deployment. To this end, we propose a wireless cooperative perception foundation model, named WiPo, which supports modality-agnostic feature transmission, as illustrated in Fig. 6.

- *Step 1: Pre-training Dataset Construction.* Modality-agnostic feature transmission aims to learn shared encod-

ing rules across heterogeneous data. To achieve this, data diversity is essential during pre-training. By leveraging existing open-source datasets from various modalities, we construct a heterogeneous multi-modal dataset. In each pre-training iteration, a sample is randomly selected from this dataset to promote generalization across modalities.

- *Step 2: Network Architecture Determination.* For modality-agnostic feature transmission, network architecture selection must consider both universality and performance. Vanilla transformers, such as ViT, offer strong universality due to their flexibility with respect to input token lengths and dimensionalities. We further adopt a window-based attention mechanism to capture detailed features, as used in the Swin Transformer.
- *Step 3: Pre-training Strategy Selection.* The foundation model consists of three components: lightweight tokenizers, a unified backbone, and specific task heads. The modality-specific tokenizers and task heads introduce inductive biases tailored to each modality and generate tokens compatible with the shared backbone. During pre-training, different tokenizers and task heads are employed for each modality, while the backbone remains shared across all modalities. The primary objective of pre-training is to obtain a unified backbone capable of reconstructing heterogeneous multi-modal data under the influence of wireless channel distortions. A similar heterogeneous pre-training strategy has been applied in the field of robotic manipulation [134].
- *Step 4: Adaptation.* Once a unified backbone is pre-trained, it can be directly applied to new datasets and even unseen modalities for feature transmission. Adaptation requires training only a small number of parameters in modality-specific tokenizers and task heads. The backbone can either be frozen or fine-tuned using PEFT techniques. Specifically, we insert adapter modules [150] into the transformer layers to leverage the pre-trained knowledge effectively, as shown in Fig. 6.

We pre-train the model on large-scale heterogeneous multi-modal datasets, including ImageNet, CsiNet-Outdoor and ShapeNet, and then fine-tune it on the SynthSoM dataset [8]. For comparison, we adopt the same network architecture for baseline models but train them individually for the respective modality. Simulation results for CSI feedback and image transmission tasks are presented in Fig. 10. Across all SNR levels, WiPo consistently outperforms the task-specific models, demonstrating its superior one-for-all and generalization capabilities. For inference efficiency, WiPo does not incur additional inference costs, as it shares the same network architecture as task-specific models. As shown in Table VI, the number of stored parameters in WiPo is only 13.65M across three tasks, which is 59.1% less than that of task-specific models. Moreover, adaptation for new datasets only needs to train a small number of parameters and freeze the pre-trained backbone, demonstrating the flexibility and versatility of WiPo. WiPo achieves competitive performance with fewer parameters by sharing backbone weights, demonstrating its effectiveness in multi-modal data transmission scenarios.

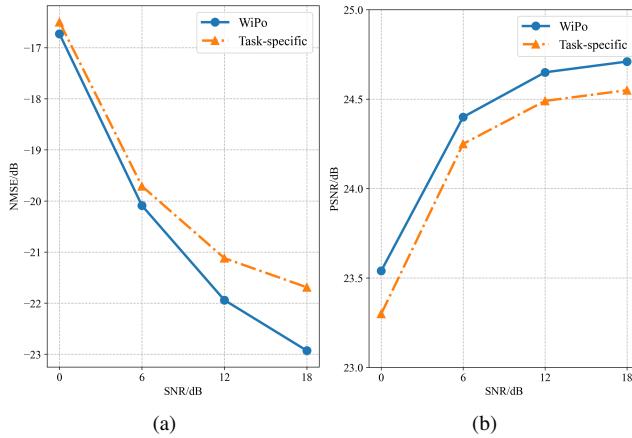


Fig. 10. Reconstruction performance comparison of WiPo and task-specific models. (a) NMSE performance of CSI feedback. (b) Peak signal-to-noise ratio (PSNR) performance of image transmission.

TABLE VI
THE NUMBER OF NETWORK PARAMETERS (TRAINING PARAMETERS/TOTAL PARAMETERS) OF CASE STUDY 3 FOR ROADMAP 2

Parameters (M)	WiPo	Task-specific
Image Transmission	1.03/11.69	11.18/11.18
CSI Feedback	0.92/11.58	10.98/10.98
Point Cloud Transmission	1.04/11.70	11.19/11.19
Stored Parameters	13.65	33.35

V. DISCUSSIONS

In this section, we first summarize and compare the three existing AI-empowered SoM system design paradigms, including task-specific AI models and the two proposed FM-empowered schemes. In addition, several open issues and potential directions for future research on FM-empowered SoM system design are discussed.

A. Paradigm Comparison of AI-empowered SoM System Design

As shown in Table VII, we comprehensively compare the three AI-empowered SoM system design paradigms from several perspectives. It can be observed that the proposed two foundation model-based schemes have significant advantages in terms of modeling capability, generalization, universality, and generative capability. Moreover, wireless foundation models not only outperform LLMs in the above aspects but also have fewer parameters, which helps reduce storage overhead. Although wireless foundation models require an additional training process, the training is offline and does not incur extra overhead during the actual deployment process. In addition, existing simulations [66, 104] show that compared to task-specific models, LLMs-based and wireless foundation models-based schemes do not significantly increase inference latency, making them promising for application in real-time systems.

In practical SoM systems, the choice of scheme should align with system requirements and hardware capabilities. Task-specific models are ideal for SoM problems with lower task difficulty, such as multi-modal sensing-aided beam prediction [10] in low-speed scenarios, and are well-suited for deployment on user-side devices with limited computational power. Pre-trained LLM-based schemes are better suited for

TABLE VII
COMPARISON OF THREE AI-EMPOWERED SOM SYSTEM DESIGN PARADIGMS, I.E., TASK-SPECIFIC AI MODELS, LLMs, AND WIRELESS FOUNDATION MODELS

	Task-specific AI models	LLMs	Wireless foundation model
Parameters	Small	Large	Medium
Inference time	Low	Medium or low	Medium or low
Pre-training requirement	No	No	Yes
Modeling capability	Weak	Medium	Strong
Generalization	Weak	Medium	Strong
Universality	Weak	Medium	Strong
Generative capability	Weak	Medium	Strong

SoM problems of moderate complexity, like frequency-domain channel prediction [66] in high-speed scenarios, and are more suitable for deployment on base stations with ample storage and computing resources. Wireless foundation model-based schemes excel in addressing high-difficulty SoM challenges, such as zero-shot channel prediction [104], while requiring fewer computational resources than LLM-based models.

B. Future work

In this part, future work is categorized by research direction, highlighting key considerations for SoM system design empowered by foundation models.

1) *Multi-modal dataset construction for SoM*: Since the dataset scale and quality determine the ultimate performance limit of AI-native systems, it is necessary to construct a massive and high-quality multi-modal sensing-communication dataset for SoM research. Towards this objective, based on the generation ability of foundation models, massive multi-modal sensing-communication data can be generated efficiently. To further ensure the quality of generated multi-modal sensing-communication data, it is essential to conduct real-world data injection via digital twin to efficiently guide the process of data generation. As a consequence, models trained on the generated data can be directly deployed in the real world, thus achieving zero-shot generalization.

2) *SoM mechanism exploration*: For the LLM-empowered SoM mechanism exploration, although fine-tuning is an intuitive approach for SoM mechanism exploration, prompt engineering techniques are yet to be utilized to explore SoM mechanisms. Leveraging such methods, based on the LLM, the SoM mechanism exploration can be explored more accurately. For the WiCo empowered scheme, by collecting real-world data, SoM mechanisms between communications and multi-modal sensing can be accurately explored through real-world injection in diverse scenarios, various frequency bands, and different conditions.

3) *SoM-enhanced transceiver design*: On one hand, for LLM-empowered SoM transceiver design, a key direction is developing transceivers that retain LLM language capabilities through prompt engineering or other innovative methods. This approach significantly enhances the utilization of LLMs, enabling them to function both as AI service providers and

as tools for optimizing transceiver design and improving communication quality, thereby significantly increasing the feasibility of practical deployment. On the other hand, for the wireless foundation model empowered scheme, one promising approach is to integrate multi-modal sensing information, as demonstrated in case study 1 of Section IV-C. However, since multi-modal sensing information is not considered during the pre-training phase, the wireless foundation models pre-trained solely on CSI data lack native capabilities for aligning and jointly processing multi-modal information, restricting their performance on SoM tasks. In future research, it is necessary to consider constructing multi-modal native wireless foundation models for SoM transceiver design, which explores the general mechanisms of multi-modal sensing-assisted transceiver design and demonstrates great generalization and universality.

4) *SoM-enhanced cooperative perception*: On one hand, for LLM-empowered SoM-enhanced cooperative perception, although many pretrained LLMs exist for perception, they largely ignore communication constraints. As the number of collaborating agents increases, limited bandwidth becomes the bottleneck for perception performance. Therefore, it is significant to design communication-efficient LLMs for cooperative perception, which can achieve both communication bandwidth efficiency and reliable perception capability. On the other hand, for wireless foundation models supporting data transmission in collaborative perception, existing approaches primarily focus on coordinating heterogeneous tasks, typically under the assumption of single-input single-output (SISO) communication over AWGN or Rayleigh channels. In reality, communication systems feature diverse physical-layer configurations, such as varying numbers of subcarriers, antennas, and users. Therefore, it is imperative to develop novel, physical-layer-aware pretraining strategies that can adapt to these heterogeneous configurations.

5) *Network system support for SoM*: LLMs can serve as intelligent orchestrators that interpret dynamic network states and provide data-driven, real-time decisions for complex resource allocation tasks. Beyond language models, specialized foundation models could learn from large-scale network data to intelligently schedule tasks and dynamically reconfigure network parameters across diverse sensing-computing-communication workloads. For example, graph-oriented or physics-informed generative models (e.g., diffusion networks) could be developed as dedicated base models that capture network topologies and physical constraints, enabling near-optimal control strategies through sampling from learned solution distributions. Integrating such generative models into decision pipelines opens new possibilities for automated service placement, adaptive topology management, and cross-layer optimization, ultimately yielding a self-optimizing, AI-driven network system that fully realizes the SoM paradigm.

VI. CONCLUSIONS

In this paper, we conducted a comprehensive study of FM-empowered SoM system design and established a complete theoretical framework. In light of existing FM-empowered

SoM-related studies, we categorize FMs for SoM system design into general-purpose FMs, i.e., LLMs, and domain-specific FMs, i.e., wireless foundation models. In light of this, we identified the key motivations for leveraging FMs to address the existing challenges in SoM systems and for the first time proposed corresponding two research roadmaps, i.e., LLMs-based and wireless foundation model-based SoM system design. For the first roadmap, we introduced the proposed design framework empowered by LLMs, including LLM selection and adaptation technology determination, and then presented two case studies. Specifically, we proposed LLM4PG and LLM4SG for SoM mechanism exploration and LLM4WM for SoM-enhanced transceiver design. Similarly, for the second roadmap, we gave a framework to illustrate the specific steps involved in building a wireless foundation model from scratch, including pre-training dataset construction, network architecture determination, pre-training strategy selection, and adaptation. Furthermore, we presented WiCo, WiFo, and WiPo, for SoM mechanism exploration, SoM-enhanced transceiver design, and SoM-enhanced cooperative perception, respectively. For each case study, preliminary simulation results were given to demonstrate the superiority of FMs over task-specific models in SoM systems. Finally, we compared the existing paradigms for AI-enabled SoM system design and outlined potential future research directions.

ACKNOWLEDGMENT

The authors would like to thank Mingran Sun and Zengrui Han for their assistance in the simulation experiments of case study 1 for roadmap 1 and roadmap 2.

REFERENCES

- [1] X. You *et al.*, “Towards 6G Wireless Communication Networks: Vision, Enabling Technologies, and New Paradigm Shifts,” *Sci. China Inf. Sci.*, vol. 64, pp. 1–74, Nov. 2021.
- [2] IMT-2030 (6G) Promotion Group, “White paper on 6G vision and candidate technologies,” *China, CAICT*, 2021.
- [3] F. Liu *et al.*, “Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond,” *IEEE J. Select. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.
- [4] X. Cheng *et al.*, “Intelligent Multi-Modal Sensing-Communication Integration: Synesthesia of Machines,” *IEEE Commun. Surv. Tutorials*, vol. 26, pp. 258–301, Firstquarter 2024.
- [5] F. Liu *et al.*, “Seventy Years of Radar and Communications: The road from separation to integration,” *IEEE Signal Process Mag.*, vol. 40, no. 5, pp. 106–121, Jul. 2023.
- [6] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, “A Survey on Vision-Language-Action Models for Embodied AI,” *arXiv preprint arXiv:2405.14093*, 2024.
- [7] L. Bai, Z. Huang, M. Sun, X. Cheng, and L. Cui, “Multi-Modal Intelligent Channel Modeling: A New Modeling Paradigm via Synesthesia of Machines,” *IEEE Commun. Surv. Tutorials*, 2025.
- [8] X. Cheng *et al.*, “SynthSoM: A Synthetic Intelligent Multi-Modal Sensing-Communication Dataset for Synesthesia of Machines (SoM),” *Sci. Data*, vol. 12, no. 819, May 2025.
- [9] M. Sun, L. Bai, Z. Huang, and X. Cheng, “Multi-Modal Sensing Data Based Real-Time Path Loss Prediction for 6G UAV-to-Ground Communications,” *IEEE Wireless Commun.*, vol. 13, no. 9, pp. 2462–2466, Sept. 2024.

- [10] H. Zhang, S. Gao, X. Cheng, and L. Yang, "Integrated Sensing and Communications Towards Proactive Beamforming in mmWave V2I via Multi-Modal Feature Fusion (MMFF)," *IEEE Trans. Wireless Commun.*, vol. 23, pp. 15 721–15 735, Nov. 2024.
- [11] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," *arXiv preprint arXiv:2108.07258*, 2021.
- [12] W. X. Zhao *et al.*, "A Survey of Large Language Models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023.
- [13] A. Liu *et al.*, "Deepseek-v3 Technical Report," *arXiv preprint arXiv:2412.19437*, 2024.
- [14] Y. Liang *et al.*, "Foundation Models for Time Series Analysis: A Tutorial and Survey," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, Barcelona, Spain, Aug. 2024, pp. 6555–6565.
- [15] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3D neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, Jul. 2023.
- [16] X. Sun *et al.*, "RingMo: A Remote Sensing Foundation Model With Masked Image Modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–22, Jul. 2022.
- [17] M. Xu *et al.*, "When large language model agents meet 6G networks: Perception, grounding, and alignment," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 63–71, Dec. 2024.
- [18] X. Qin *et al.*, "Generative AI Meets Wireless Networking: An Interactive Paradigm for Intent-Driven Communications," *IEEE Trans. Cognit. Commun. Networking*, early access 2025.
- [19] F. Jiang *et al.*, "Large Language Model Enhanced Multi-Agent Systems for 6G Communications," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 48–55, Dec. 2024.
- [20] H. Zhou *et al.*, "Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities," *IEEE Commun. Surv. Tutorials*, early access 2024.
- [21] L. Long *et al.*, "On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey," *arXiv preprint arXiv:2406.15126*, 2024.
- [22] Z. Liang *et al.*, "A Survey of Multimodel Large Language Models," in *Proc. Int. Conf. Comput. Artif. Intell. Control Eng. (CAICE)*, Xi'an, China, Jan. 2024, pp. 405–409.
- [23] A. Vaswani *et al.*, "Attention Is All You Need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Long Beach, CA, USA, 2017, pp. 6000–6010.
- [24] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying Large Language Models and Knowledge Graphs: A Roadmap," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3580–3599, Jul. 2024.
- [25] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
- [26] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin, "Large Language Models for Mathematical Reasoning: Progresses and Challenges," *arXiv preprint arXiv:2402.00157*, 2024.
- [27] T. Guo *et al.*, "What Indeed Can GPT Models Do In Chemistry? A Comprehensive Benchmark On Eight Tasks," *arXiv preprint arXiv:2305.18365*, 2023.
- [28] Q. Zhang *et al.*, "Scientific Large Language Models: A Survey on Biological & Chemical Domains," *ACM Comput. Surv.*, vol. 57, no. 6, pp. 1–38, Feb. 2025.
- [29] X. Hou *et al.*, "Large Language Models for Software Engineering: A Systematic Literature Review," *ACM Trans. Software Eng. Methodol.*, vol. 33, no. 8, pp. 1–79, Dec. 2024.
- [30] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey," *arXiv preprint arXiv:2403.14608*, 2024.
- [31] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," *arXiv preprint arXiv:2402.07927*, 2024.
- [32] S. Zhang *et al.*, "Instruction Tuning for Large Language Models: A Survey," *arXiv preprint arXiv:2308.10792*, 2023.
- [33] Z. Shen, "LLM With Tools: A Survey," *arXiv preprint arXiv:2409.18807*, 2024.
- [34] M. Goyal and Q. H. Mahmoud, "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI," *Electronics*, vol. 13, no. 17, p. 3509, Sep. 2024.
- [35] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations," *arXiv preprint arXiv:2310.07849*, 2023.
- [36] Y. Abdullin, D. Molla-Aliod, B. Ofoghi, J. Yearwood, and Q. Li, "Synthetic Dialogue Dataset Generation using LLM Agents," *arXiv preprint arXiv:2401.17461*, 2024.
- [37] C. Whitehouse, M. Choudhury, and A. F. Aji, "LLM-powered Data Augmentation for Enhanced Cross-lingual Performance," *arXiv preprint arXiv:2305.14288*, 2023.
- [38] A. Kang, J. Y. Chen, Z. Lee-Youngzie, and S. Fu, "Synthetic Data Generation with LLM for Improved Depression Prediction," *arXiv preprint arXiv:2411.17672*, 2024.
- [39] R. Tang, X. Han, X. Jiang, and X. Hu, "Does Synthetic Data Generation of LLMs Help Clinical Text Mining?" *arXiv preprint arXiv:2303.04360*, 2023.
- [40] L. He, H. Zhang, J. Liu, K. Sun, and Q. Zhang, "Zero-Shot Relation Triplet Extraction via Knowledge-Driven LLM Synthetic Data Generation," in *Int. Conf. Intell. Comput. (ICIC)*. Springer, 2024, pp. 329–340.
- [41] Z. Zhou *et al.*, "LawGPT: Knowledge-Guided Data Generation and Its Application to Legal LLM," *arXiv preprint arXiv:2502.06572*, 2025.
- [42] G. Kumichev *et al.*, "MedSyn: LLM-based Synthetic Medical Text Generation Framework," in *Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*. Vilnius,Lithuania: Springer, Sep. 2024, pp. 215–230.
- [43] T. Fedoseev, D. I. Dimitrov, T. Gehr, and M. Vechev, "Constraint-Based Synthetical Data Generation for LLM Mathematical Reasoning," in *Workshop on Math. Reasoning AI at NeurIPS'24*, Vancouver, Canada, Dec. 2024.
- [44] D. Yang, N. Monaikul, A. Ding, B. Tan, K. Mosaliganti, and G. Iyengar, "Enhancing Table Representations with LLM-powered Synthetic Data Generation," *arXiv preprint arXiv:2411.03356*, 2024.
- [45] J. Qin *et al.*, "DiffusionGPT: LLM-Driven Text-to-Image Generation System," *arXiv preprint arXiv:2401.10061*, 2024.
- [46] X. Cao *et al.*, "Medical Video Generation for Disease Progression Simulation," *arXiv preprint arXiv:2411.11943*, 2024.
- [47] X. Zhou, Q. Jia, Y. Hu, R. Xie, T. Huang, and F. R. Yu, "GenG: An LLM-Based Generic Time Series Data Generation Approach for Edge Intelligence via Cross-Domain Collaboration," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*. Vancouver, Canada: IEEE, May 2024, pp. 1–6.
- [48] Y. Wang *et al.*, "HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection," *arXiv preprint arXiv:2408.02927*, 2024.
- [49] H. Gani, S. F. Bhat, M. Naseer, S. Khan, and P. Wonka, "LLM Blueprint: Enabling Text-to-Image Generation with Complex and Detailed Prompts," *arXiv preprint arXiv:2310.10640*, 2023.
- [50] R. Zhang *et al.*, "Generative AI-Enabled Vehicular Networks: Fundamentals, Framework, and Case Study," *IEEE Network*, vol. 38, no. 4, pp. 259–267, Jul. 2024.
- [51] R. Zhang *et al.*, "Generative AI agents with large language model for satellite networks via a mixture of experts transmission," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3581–3596, Dec. 2024.
- [52] J. Shao *et al.*, "WirelessLLM: Empowering Large Language Models Towards Wireless Intelligence," *arXiv preprint arXiv:2405.17053*, 2024.

- [53] M. Kotaru, "Adapting Foundation Models for Information Synthesis of Wireless Communication Specifications," *arXiv preprint arXiv:2308.04033*, 2023.
- [54] Z. He *et al.*, "Designing Network Algorithms via Large Language Models," in *Proc. ACM Workshop Hot Topics Networks (HotNets)*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 205–212.
- [55] P. Medaranga, D. Shah, S. V. Kandala, and A. Varshney, "POSTER: Simplifying the Networking of Wireless Embedded Systems using a Large Language Model," in *Proc. ACM SIGCOMM Posters Demos*, Sydney, NSW, Australia, Aug. 2024, pp. 78–80.
- [56] S. Taheri, A. Ihalage, P. Mishra, S. Coaker, F. Muhammad, and H. Al-Raweshidy, "Domain Tailored Large Language Models for Log Mask Prediction in Cellular Network Diagnostics," *IEEE Trans. Netw. Serv. Manage.*, early access 2025.
- [57] H. Zou *et al.*, "TelecomGPT: A Framework to Build Telecom-Specific Large Language Models," *arXiv preprint arXiv:2407.09424*, 2024.
- [58] Y. Lin *et al.*, "Empowering Large Language Models in Wireless Communication: A Novel Dataset and Fine-Tuning Framework," *arXiv preprint arXiv:2501.09631*, 2025.
- [59] K. B. Kan, H. Mun, G. Cao, and Y. Lee, "Mobile-LLaMA: Instruction Fine-Tuning Open-Source LLM for Network Analysis in 5G Networks," *IEEE Network*, vol. 38, pp. 76–83, Sep. 2024.
- [60] P. Gajjar and V. K. Shah, "ORANSight-2.0: Foundational LLMs for O-RAN," *arXiv preprint arXiv:2503.05200*, 2025.
- [61] Z. Xiao *et al.*, "LLM Agents as 6G Orchestrator: A Paradigm for Task-Oriented Physical-Layer Automation," *arXiv preprint arXiv:2410.03688*, 2024.
- [62] H. Zhou *et al.*, "Large Language Model (LLM)-Enabled In-Context Learning for Wireless Network Optimization: A Case Study of Power Control," *arXiv preprint arXiv:2408.00214*, 2024.
- [63] M. Abbas, K. Kar, and T. Chen, "Leveraging Large Language Models for Wireless Symbol Detection via In-Context Learning," *arXiv preprint arXiv:2409.00124*, 2024.
- [64] C. Hu, H. Zhou, D. Wu, X. Chen, J. Yan, and X. Liu, "Self-Refined Generative Foundation Models for Wireless Traffic Prediction," *arXiv preprint arXiv:2408.10390*, 2024.
- [65] H. Noh, B. Shim, and H. J. Yang, "Adaptive Resource Allocation Optimization Using Large Language Models in Dynamic Wireless Environments," *arXiv preprint arXiv:2502.02287*, 2025.
- [66] B. Liu, X. Liu, S. Gao, X. Cheng, and L. Yang, "LLM4CP: Adapting Large Language Models for Channel Prediction," *J. Commun. Inf. Networks*, vol. 9, no. 2, pp. 113–125, Jun. 2024.
- [67] S. Fan, Z. Liu, X. Gu, and H. Li, "CSI-LLM: A Novel Downlink Channel Prediction Method Aligned with LLM Pre-Training," *arXiv preprint arXiv:2409.00005*, 2024.
- [68] Y. Sheng, K. Huang, L. Liang, P. Liu, S. Jin, and G. Y. Li, "Beam Prediction Based on Large Language models," *IEEE Wireless Commun. Lett.*, early access 2025.
- [69] Y. Cui, J. Guo, C.-K. Wen, S. Jin, and E. Tong, "Exploring the Potential of Large Language Models for Massive MIMO CSI Feedback," *arXiv preprint arXiv:2501.10630*, 2025.
- [70] J. Xue *et al.*, "Large AI Model for Delay-Doppler Domain Channel Prediction in 6G OTFS-Based Vehicular Networks," *arXiv preprint arXiv:2503.01116*, 2025.
- [71] D. Wu *et al.*, "Netllm: Adapting Large Language Models for Networking," in *Proc. ACM SIGCOMM Conf. (SIGCOMM)*, Sydney, NSW, Australia, Aug. 2024, pp. 661–678.
- [72] X. Liu, S. Gao, B. Liu, X. Cheng, and L. Yang, "LLM4WM: Adapting LLM for Wireless Multi-Tasking," *arXiv preprint arXiv:2501.12983*, 2025.
- [73] T. Zheng and L. Dai, "Large Language Model Enabled Multi-Task Physical Layer Network," *arXiv preprint arXiv:2412.20772*, 2024.
- [74] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic Communications: Principles and Challenges," *arXiv preprint arXiv:2201.01389*, 2021.
- [75] H. Nam, J. Park, J. Choi, M. Bennis, and S.-L. Kim, "Language-Oriented Communication with Semantic Coding and Knowledge Distillation for Text-to-Image Generation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Seoul, Korea: IEEE, Apr. 2024, pp. 13 506–13 510.
- [76] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic Importance-Aware Communications Using Pre-Trained Language Models," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2328–2332, Sep. 2023.
- [77] Z. Wang *et al.*, "Large Language Model Enabled Semantic Communication Systems," *arXiv preprint arXiv:2407.14112*, 2024.
- [78] S. R. Pokhrel and A. Walid, "On Large Language Model Based Joint Source Channel Coding for Semantic Communication," in *Int. Conf. Found. Large Lang. Models (FLLM)*. Dubai, United Arab Emirates: IEEE, 2024, pp. 322–329.
- [79] P. Jiang, C.-K. Wen, X. Yi, X. Li, S. Jin, and J. Zhang, "Semantic communications using foundation models: Design approaches and open issues," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 76–84, Jun. 2024.
- [80] F. Jiang *et al.*, "Large AI Model Empowered Multimodal Semantic Communications," *IEEE Commun. Mag.*, Jan. 2025.
- [81] H. Du *et al.*, "Generative AI-aided Joint Training-free Secure Semantic Communications via Multi-modal Prompts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Seoul, Korea: IEEE, Apr. 2024, pp. 12 896–12 900.
- [82] W. Chen *et al.*, "Semantic Communication Based on Large Language Model for Underwater Image Transmission," *arXiv preprint arXiv:2408.12616*, 2024.
- [83] L. Qiao, M. B. Mashhadni, Z. Gao, C. H. Foh, P. Xiao, and M. Bennis, "Latency-Aware Generative Semantic Communications With Pre-Trained Diffusion Models," *IEEE Wireless Commun. Lett.*, vol. 13, no. 10, pp. 2652–2656, Oct. 2024.
- [84] Y. Zhao, Y. Yue, S. Hou, B. Cheng, and Y. Huang, "LaMoSC: Large Language Model-Driven Semantic Communication System for Visual Transmission," *IEEE Trans. Cognit. Commun. Networking*, vol. 10, no. 6, pp. 2005–2018, Dec. 2024.
- [85] Z. Chen, H. H. Yang, K. F. E. Chong, and T. Q. Quek, "Personalizing Semantic Communication: A Foundation Model Approach," in *IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. Lucca, Italy: IEEE, Sep. 2024, pp. 846–850.
- [86] X. Cheng *et al.*, "M³SC: A Generic Dataset for Mixed Multi-Modal (MMM) Sensing and Communication Integration," *China Commun.*, vol. 20, no. 11, pp. 13–29, Nov. 2023.
- [87] D. Yang *et al.*, "UniAudio: An Audio Foundation Model Toward Universal Audio Generation," *arXiv preprint arXiv:2310.00704*, 2023.
- [88] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "MaskGIT: Masked Generative Image Transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11 315–11 325.
- [89] C. Bluethgen *et al.*, "A Vision-Language Foundation Model for The Generation of Realistic Chest X-ray Images," *Nat. Biomed. Eng.*, vol. 9, pp. 494–506, Aug. 2024.
- [90] W. Wang, K. Wu, Y. B. Li, D. Wang, X. Zhang, and J. Liu, "Mitigating Data Scarcity in Time Series Analysis: A Foundation Model with Series-Symbol Data Generation," *arXiv preprint arXiv:2502.15466*, 2025.
- [91] Y. Wu *et al.*, "VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation," *arXiv preprint arXiv:2409.04429*, 2024.
- [92] H. Wu, Z. Zhao, Y. Zhang, W. Xie, and Y. Wang, "MR-Gen: Diffusion-based Controllable Data Engine for MRI Segmentation towards Unannotated Modalities," *arXiv preprint arXiv:2412.04106*, 2024.
- [93] Z. Yang *et al.*, "Revolutionizing wireless networks with self-

- supervised learning: A pathway to intelligent communications,” *IEEE Wireless Commun.*, early access 2025.
- [94] A. Salihu, S. Schwarz, A. Pikrakis, and M. Rupp, “Low-dimensional Representation Learning for Wireless CSI-based Localisation,” in *Int. Conf. Wireless Mobile Comput. Netw. Commun. (WiMob)*. IEEE, Oct. 2020, pp. 1–6.
- [95] P. Ferrand, A. Decurninge, L. G. Ordóñez, and M. Guillaud, “Triplet-Based Wireless Channel Charting: Architecture and Experiments,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2361–2373, Aug. 2021.
- [96] N. Naderializadeh, “Contrastive Self-Supervised Learning for Wireless Power Control,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. IEEE, 2021, pp. 4965–4969.
- [97] I. Chafaa, R. Negrel, E. V. Belmega, and M. Debbah, “Self-Supervised Deep Learning for mmWave Beam Steering Exploiting Sub-6 GHz Channels,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8803–8816, Oct. 2022.
- [98] K. Davaslioglu, S. Boztas, M. C. Ertem, Y. E. Sagduyu, and E. Ayanoglu, “Self-Supervised RF Signal Representation Learning for NextG Signal Classification With Deep Learning,” *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 65–69, Jan. 2022.
- [99] R. Zhao, Y. Ruan, Y. Li, T. Li, R. Zhang, and P. Xiao, “A Transformer based Self-supervised Learning Framework for Robust Time-frequency Localization in Concurrent Cognitive Scenario,” *IEEE Trans. Wireless Commun.*, 2025.
- [100] Z. Zhang, T. Ji, H. Shi, C. Li, Y. Huang, and L. Yang, “A Self-Supervised Learning-Based Channel Estimation for IRS-Aided Communication Without Ground Truth,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5446–5460, Aug. 2023.
- [101] Z. Liu *et al.*, “Leveraging Self-Supervised Learning for MIMO-OFDM Channel Representation and Generation,” *arXiv preprint arXiv:2407.07702*, 2024.
- [102] Y. Huangfu *et al.*, “Realistic Channel Models Pre-training,” in *IEEE Globecom Workshops (GC Wkshps)*. Hawaii, USA: IEEE, Dec. 2019, pp. 1–6.
- [103] S. Alikhani, G. Charan, and A. Alkhateeb, “Large Wireless Model (LWM): A Foundation Model for Wireless Channels,” *arXiv preprint arXiv:2411.08872*, 2024.
- [104] B. Liu, S. Gao, X. Liu, X. Cheng, and L. Yang, “WiFo: Wireless Foundation Model for Channel Prediction,” *Sci. China Inf. Sci.*, early access 2025.
- [105] F. O. Catak, M. Kuzlu, and U. Cali, “BERT4MIMO: A Foundation Model using BERT Architecture for Massive MIMO Channel State Information Prediction,” *arXiv preprint arXiv:2501.01802*, 2025.
- [106] J. Guo, Y. Cui, C.-K. Wen, and S. Jin, “Prompt-Enabled Large AI Models for CSI Feedback,” *arXiv preprint arXiv:2501.10629*, 2025.
- [107] A. Salihu, M. Rupp, and S. Schwarz, “Self-Supervised and Invariant Representations for Wireless Localization,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8281–8296, Aug. 2024.
- [108] A. Aboulfotouh, A. Eshaghbeigi, and H. Abou-Zeid, “Building 6G Radio Foundation Models with Transformer Architectures,” *arXiv preprint arXiv:2411.09996*, 2024.
- [109] Z. Zhao, T. Chen, F. Meng, H. Li, X. Li, and G. Zhu, “Finding the missing data: A bert-inspired approach against package loss in wireless sensing,” in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*. IEEE, 2024, pp. 1–6.
- [110] Z. Zhao, F. Meng, H. Li, X. Li, and G. Zhu, “Mining Limited Data Sufficiently: A BERT-inspired Approach for CSI Time Series Application in Wireless Communication and Sensing,” *arXiv preprint arXiv:2412.06861*, 2024.
- [111] J. Jiang, W. Yu, Y. Li, Y. Gao, and S. Xu, “A MIMO Wireless Channel Foundation Model via CIR-CSI Consistency,” *arXiv preprint arXiv:2502.11965*, 2025.
- [112] T. Jiao *et al.*, “6G-Oriented CSI-Based Multi-Modal Pre-training and Downstream Task Adaptation Paradigm,” in *Int. Conf. Commun. Workshops (ICC Workshops)*. Denver, CO, USA: IEEE, Jun. 2024, pp. 1389–1394.
- [113] Y. Liu, H. Zhang, C. Li, X. Huang, J. Wang, and M. Long, “Timer: Generative Pre-trained Transformers Are Large Time Series Models,” *arXiv preprint arXiv:2402.02368*, 2024.
- [114] X. Shi *et al.*, “Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts,” *arXiv preprint arXiv:2409.16040*, 2024.
- [115] B. Zhang, Z. Liu, C. Cherry, and O. Firat, “When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method,” *arXiv preprint arXiv:2402.17193*, 2024.
- [116] X. Xu *et al.*, “A Survey on Knowledge Distillation of Large Language Models,” *arXiv preprint arXiv:2402.13116*, 2024.
- [117] H. Yin, H. Wang, Y. Liu, and D. Gesbert, “Addressing the Curse of Mobility in Massive MIMO With Prony-Based Angular-Delay Domain Channel Predictions,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2903–2917, Dec. 2020.
- [118] Z. Huang, L. Bai, M. Sun, and X. Cheng, “A LiDAR-aided channel model for vehicular intelligent sensing-communication integration,” *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 20105–20119, Dec. 2024.
- [119] Z. Huang, L. Bai, Z. Han, and X. Cheng, “Scatterer recognition for multi-modal intelligent vehicular channel modeling via Synesthesia of Machines,” *IEEE Wireless Commun.*, early access 2025.
- [120] Z. Han, L. Bai, Z. Huang, and X. Cheng, “Llm4sg: Large language models for scatterer generation via synesthesia of machines,” *arXiv preprint arXiv:2505.17879*, 2025.
- [121] A. Radford *et al.*, “Language Models are Unsupervised Multitask Learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, Feb. 2019.
- [122] W. Qi, Y.-P. Ruan, Y. Zuo, and T. Li, “Parameter-Efficient Tuning on Layer Normalization for Pre-trained Language Models,” *arXiv preprint arXiv:2211.08682*, 2022.
- [123] M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh, “Deep Learning-Based Channel Estimation,” *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 652–655, Apr. 2019.
- [124] A. Salihu, S. Schwarz, and M. Rupp, “Attention Aided CSI Wireless Localization,” in *IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. Oulu, Finland: IEEE, Jul. 2022, pp. 1–5.
- [125] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
- [126] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, “Cross-Stitch Networks for Multi-task Learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3994–4003.
- [127] Y. Liu, J. Cao, C. Liu, K. Ding, and L. Jin, “Datasets for Large Language Models: A Comprehensive Survey,” *arXiv preprint arXiv:2402.18041*, 2024.
- [128] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision Meets Robotics: The Kitti Dataset,” *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, Nov. 2013.
- [129] A. Alkhateeb *et al.*, “DeepSense 6G: A Large-Scale Real-World Multi-Modal Sensing and Communication Dataset,” *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, Sep. 2023.
- [130] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, “ViWi: A Deep Learning Dataset Framework for Vision-Aided Wireless Communications,” in *Proc. IEEE Veh. Technol. Conf. (VTC2020-Spring)*. Antwerp, Belgium: IEEE, May 2020, pp. 1–5.
- [131] E. Aiello, D. Valsesia, and E. Magli, “Cross-modal learning for image-guided point cloud shape completion,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 37349–37362, 2022.
- [132] T. Zheng, A. Li, Z. Chen, H. Wang, and J. Luo, “Autofed: Heterogeneity-aware federated multimodal learning for robust

- autonomous driving,” in *Proceedings of the 29th annual international conference on mobile computing and networking*, 2023, pp. 1–15.
- [133] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *arXiv preprint arXiv:2010.11929*, 2020.
 - [134] L. Wang, X. Chen, J. Zhao, and K. He, “Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, Vancouver, Canada, Dec. 2024, pp. 124 420–124 450.
 - [135] W. Fedus, B. Zoph, and N. Shazeer, “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity,” *J. Mach. Learn. Res.*, vol. 23, no. 120, pp. 1–39, Jan. 2022.
 - [136] A. Q. Jiang *et al.*, “Mixtral of Experts,” *arXiv preprint arXiv:2401.04088*, 2024.
 - [137] D. Dai *et al.*, “DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models,” *arXiv preprint arXiv:2401.06066*, 2024.
 - [138] N. Shazeer, “Glu Variants Improve Transformer,” *arXiv preprint arXiv:2002.05202*, 2020.
 - [139] H. Touvron *et al.*, “LLaMA: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, 2023.
 - [140] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, “FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, Vancouver, Canada, Jun. 2022, pp. 16 344–16 359.
 - [141] J. Ainslie, J. Lee-Thorp, M. De Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghi, “GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints,” *arXiv preprint arXiv:2305.13245*, 2023.
 - [142] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 10 012–10 022.
 - [143] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A Large-Scale Hierarchical Image Database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Miami, FL, USA: IEEE, Jun. 2009, pp. 248–255.
 - [144] J. Guo, T. Chen, S. Jin, G. Y. Li, X. Wang, and X. Hou, “Deep Learning for Joint Channel Estimation and Feedback in Massive MIMO Systems,” *Digital Commun. Networks*, vol. 10, no. 1, pp. 83–93, Feb. 2024.
 - [145] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 16 000–16 009.
 - [146] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in *Int. Conf. Mach. Learn. (ICML)*, Jul. 2020, pp. 1597–1607.
 - [147] J. Zhang, Z. Wei, B. Liu, X. Wang, Y. Yu, and R. Zhang, “Cloud-Edge-Terminal Collaborative AIGC for Autonomous Driving,” *IEEE Wireless Commun.*, vol. 31, no. 4, pp. 40–47, 2024.
 - [148] I. Yaman *et al.*, “The LuViRA Dataset: Synchronized Vision, Radio, and Audio Sensors for Indoor Localization,” in *Proc. Int. Conf. Robot. Automat. (ICRA)*. Yokohama, Japan: IEEE, May 2024, pp. 11 920–11 926.
 - [149] F. Euchner, M. Gauger, S. Dörner, and S. ten Brink, “A Distributed Massive MIMO Channel Sounder for “Big CSI Data”-driven Machine Learning,” in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, Eurecom, France, Nov. 2021, pp. 1–6.
 - [150] N. Houlsby *et al.*, “Parameter-Efficient Transfer Learning for NLP,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Los Angeles, CA, USA, Jul. 2019, pp. 2790–2799.