O-RAN next Generation Research Group (nGRG)

Contributed Research Report

Research Report on Generative AI Use Cases and Requirements on 6G Network

Report ID: RS-2025-02

Contributors:

CMCC

Lenovo

Tongji University

LG Uplus

Nokia

AsiaInfo

Release date: 2025.06

Authors

Ziqi Chen, CMCC (Editor-in-Chief)

Mingzeng Dai, Lenovo

Hao Xu, Ming jiang, Tongji University

Jaehyun Ahn, Hyosun Yang, LG Uplus

Daiju Chiriyamkandath Antony, Nokia

Zhanwu Li, AsiaInfo

Lehan Wang, Yi Ren, CMCC

Reviewers

Yuanfang Huang, CICT

Vikas Dixit, Reliance Jio

Rajat Agarwal, Tech Mahindra

Daiju Chiriyamkandath Antony, Nokia

Hiroshi Miyata, Sumitomo Electric Industry

Disclaimer

Copyright

Executive summary

The integration of Generative AI (GenAI) into 6G networks marks a transformative step in the evolution of communication systems. This report explores the diverse use cases and requirements of GenAI in 6G, structured into three main categories: GenAI for 6G network, 6G networks supporting GenAI applications, and GenAI based services as a native feature of 6G.

To enable these use cases, the report highlights 4 aspects of essential requirements: network computing, data transmission, network functions and network AI designs. Finally, the report identifies four key research directions pivotal to achieving this vision: developing AI-optimized communication protocols, advancing GenAI algorithms for network augmentation, building scalable distributed computing frameworks, and exploring 6G network native GenAI applications.

By harnessing the capabilities of GenAI, it is expected that 6G networks will achieve unparalleled levels of efficiency, automation, and personalization. Collaboration across industries and academia will be critical in overcoming challenges and driving innovation toward a future where 6G networks empower both individuals and industries with intelligent, context-aware, and adaptive solutions.

Table of Contents

List of abbreviations

1G	1st Generation
5G	5th Generation
6G	6th Generation
AI	Artificial Intelligence
AIGC	Artificial Intelligence Generated Content
AI/ML	Artificial Intelligence/Machine Learning
AMF	Access and Mobility Management Function
BB	Baseband
CLIP	Contrastive Language-Image Pre-Training
CSI	Channel State Information
DM	Diffusion Model
DT	Digital Twin
FD	Fully-digital
GAN	Generative Adversarial Network
GenAI	Generative AI
GPT	Generative Pre-trained Transformer
JSCC	Joint Source-Channel Coding
LCM	Life Cycle Management
LLM	Large Language Model
LoRA	Low-rank Adaptation
mmWave	Millimeter Wave
MTSI	Multimedia Telephony Service for IMS
NF	Network Function
QoS	Quality of Service
RAG	Retrieval Augmented Generation
RAN	Radio Access Network
RE	Reasoning Engine
RSRP	Reference Signal Receiving Power
RSSI	Received Signal Strength Indicator
SINR	Signal to interference and noise and ratio
SSIM	Structural Similarity Index
PEFT	Parameter-Efficient Fine-Tuning
PSNR	Peak Signal-to-Noise Ratio
RF	Radio Frequency
UE	User Equipment
SMF	Session Management Function
TEE	Trusted Execution Environment
XR	eXtended Reality

List of figures

1 Introduction

The evolution of mobile communication networks from 1G to 5G has brought significant advancements in connectivity, speed, and reliability. However, as the industry moves towards the next generation, 6G, there is an increasing demand for a network that is not only faster but also more intelligent, adaptive, and capable of supporting a wide array of emerging applications. In this context, GenAI is poised to play a transformative role in shaping the capabilities of 6G networks, unlocking new possibilities for enhanced automation, user experience, and service innovation.

GenAI, with its powerful abilities to understand, generate, and transform content across multiple modalities (text, image, video, and audio), has already shown immense potential in various domains [1]. By integrating GenAI into 6G networks, it is possible to go beyond traditional communication paradigms and introduce various network capabilities and applications into the next generation network. These advancements promise to revolutionize how networks operate, offering dynamic and personalized services while meeting the increasingly complex needs of users and industries.

This research report provides a comprehensive exploration of the key use cases for GenAI in 6G networks, including the use case augmenting 6G network operation, the GenAI applications serviced by 6G network and GenAI service embedded into 6G network native service. The associated technical requirements and challenges are also identified. Furthermore, the report identifies critical research directions that must be pursued to realize the full potential of GenAI in 6G network.

2 Use Cases

2.1 Generative AI for 6G network

2.1.1 Network management automation empowered by AI Agent

2.1.1.1 Description

Compared to 5G networks, 6G networks are expected to adopt more flexible and scalable network architectures, as well as more complex networking methods, to support a wider range of vertical industry applications and services. This will pose greater challenges for 6G network management. Thus, 6G networks urgently require more intelligent technologies to enhance network management efficiency. AI Agents, with GenAI's powerful comprehension and generation capabilities, autonomy and flexibility, learning and adaptation abilities, as well as rich interaction methods, can provide more intelligent solutions for 6G network management as follows.

1. Comprehension and generation capabilities

GenAI's strong natural language understanding and content generation capabilities enable it to understand abstract concepts, process complex tasks, and generate creative output. For example, in network slice management, GenAI can accurately understand user service requirements expressed in natural language and automatically create corresponding network slices.

2. Autonomy and flexibility

AI Agent can autonomously analyze, adapt, and respond to environmental changes based on pre-defined objectives. For example, in network slice management, it can autonomously perform tasks such as creating, configuring, optimizing, and repairing network slices based on real-time network conditions, eliminating the need for manual intervention [2].

3. Learning and adaptation capabilities

AI Agent uses unsupervised or self-supervised learning to learn from large amounts of data, continuously accumulating knowledge and improving its capabilities. For example, an AI Agent can learn network operation rules from network logs and user behavior data, and optimize network performance based on the learned knowledge.

4. Richer interaction capabilities

AI agent enables more natural and richer interactions. For example, it can communicate with network administrators or users using natural language. It can also generate multi-modal content such as images or videos.

The introduction of AI agents in 6G network management can promote the automation of network management, greatly enhance the level of network intelligence, and promote the development of 6G network management.

2.1.1.2 Requirement

To achieve the above AI Agent-powered 6G network management automation, AI Agents and 6G network need to meet the following key requirements:

REQ1: Specialized knowledge and few-shot learning capabilities. GenAI models should integrate specialized knowledge bases. This will enhance their understanding of 6G network management language and content generation capabilities. In addition, they should have Few-Shot Learning capabilities, enabling them to quickly learn new network management strategies from limited data.

REQ2: Explainable AI techniques. GenAI models need to be more explainable to make their decision processes transparent and easier to understand and analyze. This requires research into explicable AI techniques such as Attention Mechanism and Decision Tree.

REQ3: Open and programmable capabilities. 6G networks should adopt a more open and programmable architecture to provide open interfaces and programmability capabilities for GenAI models to flexibly access the network, acquire data, and perform operations.

2.1.2 Intent-based network management supported by LLM and retrieval augmented generation

2.1.2.1 Description

Intent-based management plays a fundamental role in network management automation. Intent-based management is a use case commonly seen with the potential to leverage Large Language Models (LLM) [3][4]. In such a use case, a network operator may provide a network intent towards the network system in a free-form natural language. The intent can comprise the desire, e.g., to improve network availability by minimizing network function downtime. Such intent can be comprehended by the GenAI models, particularly those with the capability to comprehend natural language, such as language models, e.g., LLMs. The LLM can take as input the operator's prompt and can provide as output the recommendations on the actions that need to be taken in order to, for example, prevent failures of Network Function (NF), improve network performance, or recommend the most suitable time for NF maintenance. The language model can also provide recommendations on the actions needed to recover from occurred failures.

The generic LLM, i.e. foundation model, is commonly obtained by extensive training using large amounts of data to capture the relations between the words. As a result, the model obtains generic capabilities for generic text understanding, processing, and generation. Such a process is called pre-training. The generic model usually does not perform well on domain specific tasks and needs to be further trained, i.e., fine-tuned using domain and task specific data. Fine-tuning enables the model to perform better on specific tasks [5], e.g., understanding technical text in the RAN (Radio Access Network)/Core/telco vendor domain or understanding telco network intent and recommending management actions to fulfill the intent.

An important approach for further improving the LLM performance and keeping it up to date with the most recent facts is Retrieval-Augmented Generation (RAG). The

approach is based on grounding the model on external knowledge sources beyond training data and thus supplementing the LLM's internal knowledge. It aims at retrieving snippets of information from external knowledge that are most relevant to the user's input prompt or question. Such external information is used to augment the initial user's prompt prior to passing it to the large language model. The LLM uses such an augmented prompt as well as its internal knowledge to generate the output. With such augmentation, the LLM relies on the most current and reliable information when generating an answer, thereby offering improved accuracy as well as lowering the costs (e.g., compared to fine-tuning) of keeping the LLM responses up to date [6].

Despite the advantages of using RAG in output generation, certain implications need to be considered as well, such as the fact that fetching the information from external knowledge sources may be associated with constraints in terms of access rights, cost of obtaining the answer, increased latency in generating the answer, security, etc. Therefore, control of RAG activation and deactivation is needed to overcome such issues and to use RAG in the most optimal way for a given use case.

In the intent-based management use case described above, where the network operator poses an intent, e.g., "improve network availability", the system may leverage RAG capabilities. If the network operator activates the RAG support, the LLM outputs can be based on the knowledge external to the model, collected as part of network operation, e.g., containing customer documentation on NF deployed in the network along with the knowledge on ticket resolutions from the customer support team. Based on such up-to-date information on the most recent ticket resolutions as well as any updates in the customer documentation on the deployed NFs, the model can provide more reliable outputs, e.g., recommendations on the actions to be taken to minimize the NF failure.

2.1.2.2 Requirement

REQ1: The 6G network should have the capability to expose information on different knowledge sources along with the associated implications, e.g., costs, based on which the LLM-based solutions can generate outputs to input prompts.

REQ2: The 6G network should enable the configuration of whether, how and to what extent external and internal knowledge (e.g., what amount of data) should be used when generating the LLM outputs.

REQ3: The 6G network should have the capability to report on usage of internal or external knowledge when generating the outputs.

2.1.3 Network management using multi-modal AI

2.1.3.1 Description

Network management using GenAI's multi-modal functionality is expected to make network operations more efficient and intelligent. The network can be analyzed from multiple analytical perspectives, leveraging the diverse data formats—such as text, images, and videos—that GenAI models can process in a unified manner. This

multimodal capability enables the network to synthesize insights from various data sources, integrating technical metrics like radio quality with contextual information such as sensor data or surrounding visual environments, for a more comprehensive understanding. Therefore, in 6G, the network may perform network management by integrating radio quality information, sensor data, camera images, and network fault alarms, allowing it to account for both radio quality and the surrounding environment [7][8].

For example, the 6G network can integrate and process various modalities of data with the help of multi-modal GenAI models, such as radio quality information (RSRP, RSSI, SINR), sensor data, camera images, and fault alarms to visualize and monitor the status of base stations and surrounding environments in real time, as shown in Figure 1. The network can detect obstacles around the base station blocking signals and optimize beam patterns to avoid these obstructions [9].

Also, by monitoring the network status and analyzing real-time environmental changes, it can identify the causes of degraded radio quality without the need of on-site inspections and apply appropriate solutions.

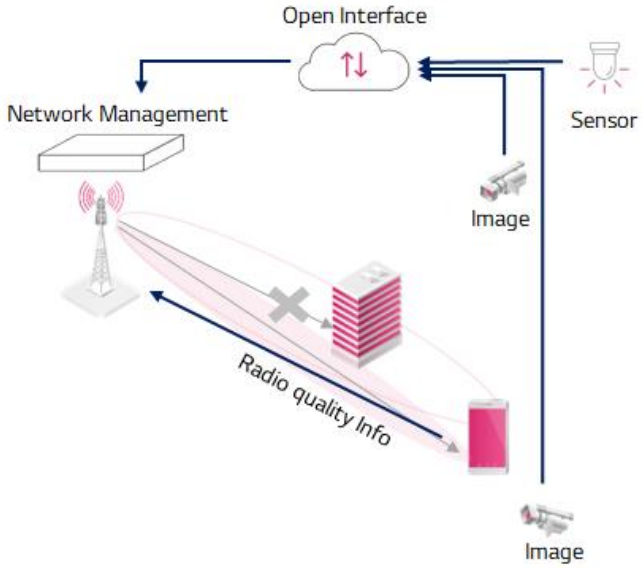

Figure 1 Multi-modal data aided network management

2.1.3.2 Requirement

The several considerations should be addressed when using multi-modal GenAI to network management:

REQ1: The 6G network should be able to co-relate radio information with various other data, which needs to be effectively gathered via open interface. Since each type of data provides a different perspective on network status, it is crucial that radio information and other data be processed and analyzed together.

REQ2: Data synchronization should be also considered. If data is collected separately from different sources without synchronization, inconsistencies may

occur, making accurate analysis difficult. Additionally, it needs real-time data synchronization in network environment analysis. Wireless signals fluctuate continuously due to environmental factors, obstacles, user movement, and other time-of-day variations. In such a dynamic environment, when network failures occur, it is necessary to diagnose the issue quickly and accurately. This requires real-time data synchronization, which ensures that the network environment is assessed immediately, enabling the rapid identification of the root cause and prompt action. However, real-time synchronization may occasionally experience delays due to various factors like network congestion or processing limitations. To address this, optimization efforts and advanced technologies such as Time-Sensitive Networking protocols are needed to ensure precise timing for critical data.

REQ3: The processing of diverse data sets via multi-modal capabilities requires high-performance computing resources. Since network management demands real-time processing and acting quickly, the 6G network should have high performance computing power that possess sufficient capacity and efficiency to handle the complex and demanding Gen AI tasks associated with multi-modal data processing in real-time.

In conclusion, leveraging the multi-modal AI in network management can maximize network efficiency through optimization and real-time responses, providing better service quality to users. However, data integration, open data interface, data synchronization, and the computational power required for multi-modal processing should be considered.

2.1.4 GenAI aided RAN digital twin

2.1.4.1 Description

A RAN Digital Twin (DT) is a real-time virtual model of the physical RAN infrastructure, enabling continuous monitoring, simulation, predictive analysis, and optimization. This advanced technology boosts network performance and reliability while providing automated decision-making [10]. The implementation involves key steps like data collection, real-time analysis, model training, and continuous resource adjustments, ultimately creating a fully automated and intelligent network management system. By integrating GenAI, the RAN DT enhances its data collection and augmentation capabilities, generating high-fidelity synthetic data to better represent complex network scenarios.

Generative AI can enhance RAN DT in the 6G network from the following aspects:

1. Data augmentation

RAN DT can utilize GenAI to implicitly learn the distribution and characteristics (e.g. spatio-temporal correlation feature of channel state information [11]) of 6G network data, enabling the generation of high-fidelity data. This process enhances both the quality and quantity of the data available for real-time analysis and closed-form network optimization. GenAI can also assist in generating corner case data, which represents rare or extreme scenarios that may not be well-represented in the

collected dataset [12]. By simulating these unusual conditions, RAN DT can provide the network with capabilities to effectively handle unexpected events and edge cases that might otherwise lead to performance degradation or failures.

2. AI model training

Based on high-fidelity synthetic network datasets enhanced by GenAI, RAN DT provides robust support for various AI model training needs. In particular, RAN DT enables continuous, real-time training of AI models, ensuring up-to-date and adaptive learning capabilities.

3. Network performance optimization

For performance optimization, RAN DT generates recommendations for updating network parameters and configurations using comprehensive simulations, tailored to specific network requirements [10]. GenAI can significantly aid in this process. Various deep generative models such as diffusion models and generative adversarial networks (GANs) have shown their effectiveness in handling time-varying and high-dimensional network states, such as environment information, user location, and traffic volume [13]. This capability enables RAN DT to make optimized decisions, such as determining the sleep-wake timing and antenna configurations of base stations, under any complex network conditions, thereby enhancing the versatility of the DT framework.

2.1.4.2 Requirement

DT is a cutting-edge, data-driven technology. To accurately and responsively represent the highly dynamic and large-scale 6G networks with low data collection overhead, it is necessary to introduce GenAI for data augmentation, providing a more comprehensive and high-fidelity dataset compared to the 5G era. To support real-time network performance optimization, the latency of this data augmentation process needs to be minimized. This latency primarily stems from delays in uploading collected data and generating new data.

REQ1: For the massive volume of real-time uploaded network data, 6G networks should possess the capability to effectively compress and transmit the data, extracting key semantic information that accurately represents the 6G RAN with the aid of GenAI. This process involves generating low-dimensional encoded representations, thereby reducing network load and backhaul pressure, minimizing data collection latency, and ultimately enhancing network performance and user experience.

REQ2: Construction of the DT requires support from a vast array of data from different sources and types, such as environmental, user, and service data. 6G networks should ensure that these heterogeneous data, sampled at the same time, are transmitted to the physical entity running the RAN DT with minimal time difference, thereby guaranteeing its real-time and accurate construction.

REQ3: To minimize the latency associated with generating new data and real-time adjustments of the DT based on live network data, it is essential to implement distributed inference and training of GenAI models at the edges of the 6G network [14]. Consequently, 6G networks are anticipated to ensure overall service latency

and stability by jointly optimizing the computing and communication tasks involved through strategies such as task orchestration.

2.1.5 GenAI aided physical layer processing

2.1.5.1 Description

Generative AI models, including Generative Adversarial Network, Transformer, Diffusion Model, represent a new generation of AI/ML models that originates from content generation tasks but also excels in a wide variety of telecommunication problems. In this section, we delve into how GenAI is used in physical layer processing, specifically channel prediction, channel generation and signal denoising, and pre-coding.

1. Channel estimation

In telecommunication systems, the precision of channel estimation directly influences the subsequent signal detection and demodulation processes. Generally, pilots are sent along with data to acquire the corresponding channel state information (CSI). Given that the patterns of the pilots are predetermined, the receiver is able to estimate the channel characteristics based on the received signals. Traditional channel estimation algorithm is linear minimum mean-squared error [15]. More recently, CNN-based deep learning algorithms are explored to improve estimation accuracy [16], but resulting in large number of AI/ML parameters and high computational complexity. Inspired by generative models, many researches have introduced attention modules (core of transformer) into channel estimation [17]. Specifically, channel matrices are patch embedded and combined with channel characteristics information such as Delay, SINR, Doppler shift, then fed into Transformer encoder for channel estimation, before inverse patch embedding operation and splicing together the real and imaginary parts.

2. Channel generation and signal denoising

Diffusion models (DM) have recently achieved unprecedented success in Artificial Intelligence Generated Content (AIGC), including image generation and editing, text, and video generation. DM is a class of latent variable models inspired by non-equilibrium thermodynamics. They directly model the score function of the likelihood function through variational lower bounds, resulting in advanced generative performance. DM gradually adds Gaussian noise to the
training data in the forward diffusion process until the data becomes pure noise. Then, in the reverse sampling process, it learns to recover the data from the noise. The designed process of DM and the wireless communications system are similar. DM progressively learns to effectively remove noise, thereby generating data that closely resembles the original distribution, while the receiver in the wireless communications system aims to recover the transmitted signal from the received signal. It is worth to study whether DM be applied to the wireless communications system to help the receiver to remove noise. For example, DM can be employed to generate the wireless channel for an end-to-end communication system, achieving almost the same performance as the channel-aware case [18]. And channel denoising diffusion models

is proposed for communication systems to remove channel noise and purify the received signals [19].

3. Pre-coding

For Millimeter Wave (mmWave) massive MIMO systems to function effectively, precoding of transmitted signals is crucial. Traditionally, fully-digital (FD) methods have been used, requiring one Radio Frequency (RF) chain per antenna element, but faces challenges due to high computational complexity, power consumption, and hardware costs. Hybrid analog-digital beamforming uses cost-effective and computationally efficient analog components such as phase shifters along with digital precoding, resulting in fewer RF chains. In this hybrid setup, the digital baseband (BB) precoder is combined with the analog RF beamformer components.

Traditional Deep learning algorithms, such as deep neural networks and convolutional neural networks have been explored to design the RF beamformers, aiming to optimize the overall hybrid beamforming matrix to be as close as possible to FD RF precoder matrix. Originating from Image generation tasks, Generative Adversarial Network (GAN) is an algorithm in which two neural networks, the generator and the discriminator, are trained to compete with each other to improve the generator's ability to create realistic outputs that can mimic real data. Conditional generative adversarial networks have shown their advantage in fine-tuning beam steering matrix and generating BB providing matrix, while requiring less time and smaller datasets to train [20].

2.1.5.2 Requirement

REQ1: GenAI model inference at millisecond level. To perform GenAI based signal processing in physical layer, it is necessary to lower the inference time of those models to sub-frame or symbol level. Thus, highly efficient model inference system is required, containing joint design of dedicated GenAI model architecture, computing hardware and communication protocol.

REQ2: Life Cycle Management facing GenAI models. Despite the existing life cycle management (LCM) of AI/ML models (e.g., model training, model deployment, model inference, model monitoring, model updating), new procedures and LCM should be studied and incorporated into the network facing GenAI model operation (e.g., pre-training, fine-tuning, prompt engineering, etc.). Two-sided AI/ML model may be needed: A paired AI/ML Model(s) over which joint inference is performed, where joint inference comprises AI/ML inference whose inference is performed jointly across the UE and the RAN node. The computational complexity, power consumption, and hardware costs caused by GenAI model training can be huge, thus the optimization of GenAI training is required as well.

2.2 6G network for generative AI applications

2.2.1 Real-time interactive video generation

2.2.1.1 Description

The real-time interactive video generation use case leverages the advanced capabilities of GenAI and 6G networks to deliver a fully interactive video experience for multiple UEs. This technology allows real-time adjustments to the video content based on the feedback provided by viewers. The video's storyline, scenes, and even visual styles can be dynamically generated or modified in response to audience interactions, creating a highly personalized and immersive viewing experience. This application will involve crucial features listed below, also shown in Figure 2.

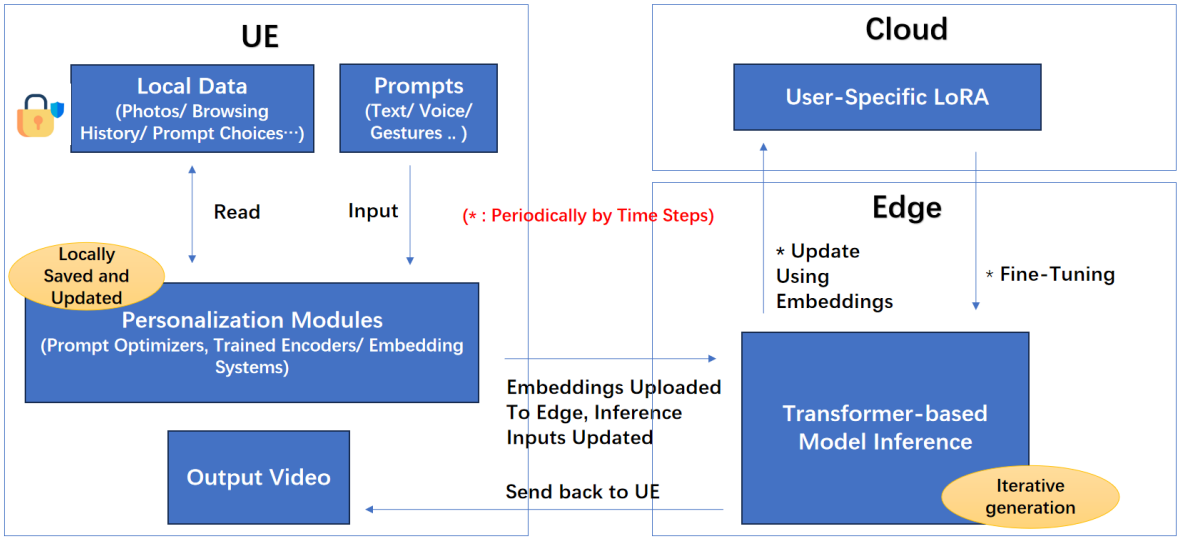

Figure 2 Interactive video generation

1. Real-time feedback

Throughout the viewing experience, viewers can provide input through text prompts (or more advanced inputs such as voice, facial expressions or motion gestures). These inputs are processed in real-time and translated into prompts that influence the storyline, scene transitions, or visual elements of the video. The 6G network's ultra-low latency ensures that the feedback is reflected almost instantaneously in the video being generated.

2. Model partitioning and edge computing

The GenAI model is split between the edge and user devices. The model inference part is hosted on the network edge, e.g. MEC, where it will be responsible for complex generative tasks and video rendering, ensuring that high-quality video content is produced and delivered without delay; while the modules related to personalization such as Contrastive Language-Image Pre-Training (CLIP)/prompt optimizers can reside on the UEs, as user data and preference information are typically privacy-sensitive. The model receives real-time data inputs from multiple users, processes them at the edge, and sends back the video streams to UEs.

3. Personalized elements

The part of the model running locally on the user's device is responsible for personalizing the experience. For instance, based on user consent, the local module can store and process personal preferences (e.g., past prompt choices, viewing history, saved contents) and generate personalized prompts or suggestions. These local data help create a more customized interaction, making the video experience unique to each user. The edge or central cloud can also offer Parameter-Efficient Fine-Tuning (PEFT) services for each user by deploying Low-rank Adaptation (LoRA) models, which is feasible due to their lightweight nature. During the generation process, personal LoRA model weights will be updated based on that user's interaction history, which will be used for customized fine tuning.

4. Data privacy and security

While local processing allows for personalized services, user privacy is maintained through the use of secure, encrypted communication channels and the option for users to control how their data is used. Sensitive information such as personal preferences or motion sensor data remains on the device unless the user consents to sharing it with the edge.

2.2.1.2 Requirement

To enable real-time interactive video generation within a 6G network, the following technical requirements should be met:

REQ1: Edge Computing Support. Real-time interactive video generation requires significant computational capacity for tasks such as rendering and model inference. Networks should provide robust edge computing support to accommodate the video inference model, especially users are interacting in a real-time stream video content. The inference model, which can include transformer-based model such as Diffusion-Transformer, will be deployed on the edge. The inference process will then take place on the edge, taking advantage of its capacity and latency, ensuring the interactions with UEs remain smooth.

REQ2: Data Privacy and Security. Network should implement strong encryption and privacy measures. User data should remain encrypted during transmission and processing, with sensitive data, such as personalized preferences, staying on the UEs. 6G networks need to implement secure communication protocols, such as end-to-end encryption, to ensure that data transferred between UEs and edge nodes remains protected from unauthorized access.

REQ3: Quality of Service (QoS) Assurance. Real-time interaction relies on low latency to ensure that user inputs are reflected in the generated video with minimal delay. For high-resolution content, requires significant bandwidth. The 6G network should be able to dynamically allocate bandwidth to each user based on their data transmission and interactions. Also, network should be able to implement prioritization policies to ensure that video generation and user feedback traffic are given precedence, reducing the chances of congestion or bandwidth competition affecting the video experience. The network should provide consistent user experience levels, ensuring that video generation and delivery remain reliable.

REQ4: Network Extendibility. The system should provide sufficient infrastructure at the edge or cloud to store user-specific LoRA models. These models could be periodically updated based on user interactions, in order to enable incremental fine-tuning of the video generation model. The storage and computational capacity at the edge must support not only the storage of multiple users' LoRA models but also the ability to load, update, and apply these models in real-time during video generation. On the UE side, the system needs to deploy and store personalization modules capable of handling local adjustments for user prompts and embeddings. The UE should also be able to dynamically adjust these modules based on user feedback or stored preferences. Moreover, the system should support the flexibility to scale-up, as new users join or existing users' personalized models grow in complexity.

2.2.2 Edge-based computing service for GenAI

2.2.2.1 Description

Generative AI in 6G network requires inter-operations between the control plane and user plane and the additional provisioning plane for computing and storage resources used for AI purposes, indicating the necessity of Reasoning Engine (RE) [21][22].

The personal edge-based computing service use case in 6G network enables user to efficiently and securely deploy their own AI models through a RE. When the network analyzes the computing intent based on the user's requirements, the RE securely provides the computing service using public key encryption, creating a virtual space for the deployment of the user's personal AI model.

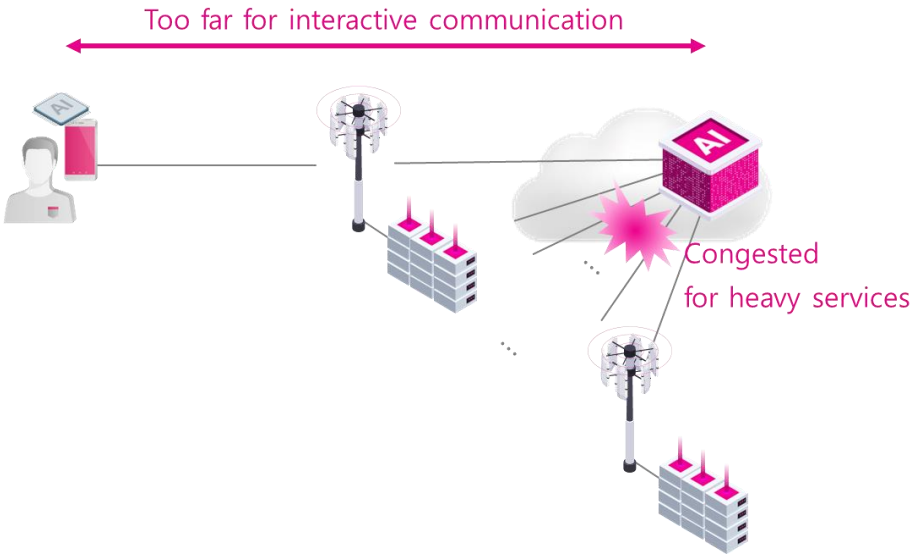

Figure 3 Edge-based GenAI service example

Basically, edge-based computing service would provide interactive AI service to UEs with low latency. Centralized cloud computing seems improper for the AI service since

the delay between central cloud server and device is quite big and it would be worse when AI service connections grow too many and congested, as shown in Figure 3.

The example process details are shown as follows in Figure 4.

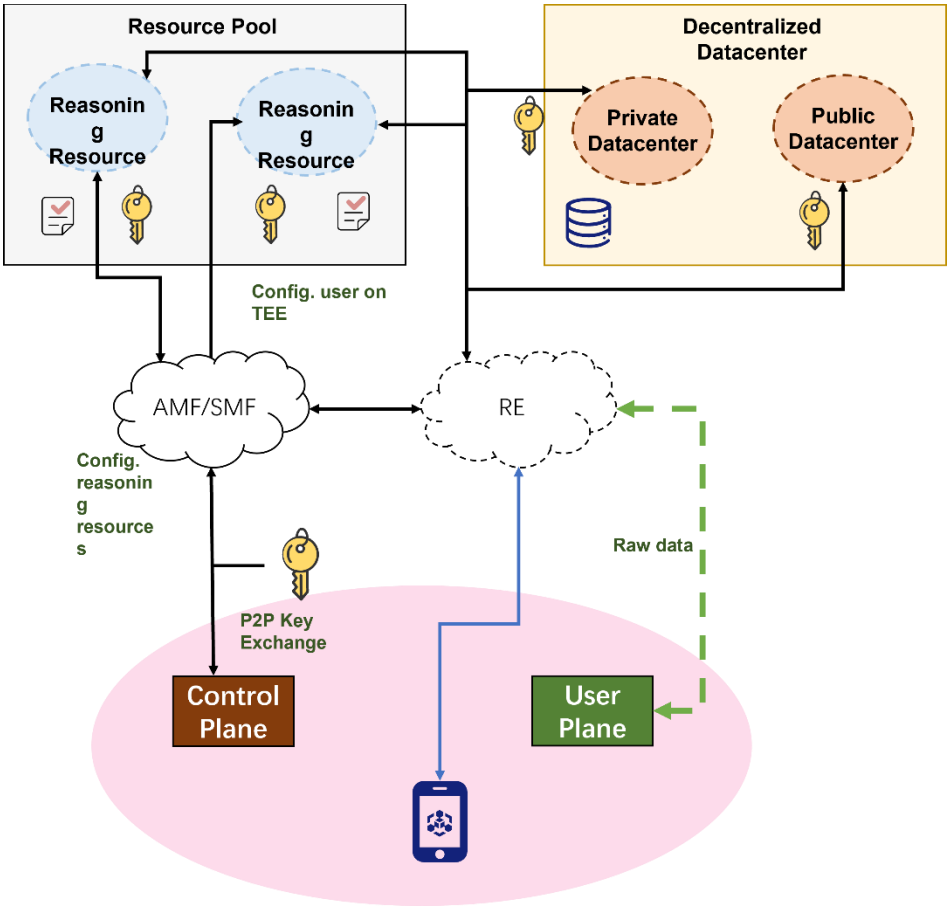

Figure 4 Edge-based GenAI process

1. Computing resources allocation

When UEs request computing service from 6G networks, the network allocates resources for training and inference of UEs' private data. Therefore, it provides a set of public keys, exclusive for the computing reasoning resources, which consist of distributed storage and user privacy preserving Trusted Execution Environment (TEE). By doing so, the public key pair is exchanged with the designated resource instance to ensure the integrity and confidentiality of computing resources.

2. Computing resource provisioning

When Access and Mobility management Function (AMF) or Session Management Function (SMF) receives the computing service request from the user, it starts to pass on public keys supplied by the user, and start provisioning the computing resources from the designated pool. Note that there are two types of computing services involved: computing resources used for AI training; and data generation/inference from public/private datacenter; both play an important aid to the network for AI applications.

3. UE data transmission

Once the UE's attached RE has gained access to the provisioned resources provided by networks, it uses the configured credentials for the uplink of its raw data to the RE, where the source raw data meet the rendered data. RE gains access to user's data under strict privacy and confidentiality compliance. With proper computing resources and enough data obtained, the RE is able to train the AI model based on TEE.

2.2.2.2 Requirement

REQ1: Since data in the database is treated as a commodity, ensuring data privacy and security is the top priority. For data under privacy protection, each user on the network must comply with relevant regulations when accessing or using the data, ensuring that the data remains usable but not visible to the RE. Additionally, to maintain privacy throughout the entire process, all sensitive information, including user requirements, must be encrypted using methods like public key encryption. Effective measures must also be in place to protect the user's private key from eavesdropping or tampering.

REQ2: Performance improvements such as throughput and latency remain crucial for the commercial deployment of this architecture from the user's perspective. This necessitates a new network topology, an enhanced service session continuity mechanism, and congestion management mechanism based on the new architecture to ensure fast and smooth point-to-point data forwarding.

REQ3: Future 6G network scenarios will be more complex, with a significant increase in data transmission volumes, leading to substantial network bandwidth overhead. The 6G network must ensure quality of service even under constrained bandwidth conditions.

2.2.3 Distributed AI Agents

2.2.3.1 Description

Distributed AI Agents in 6G networks redefine service delivery through "Agent as a Service", where agents, equipped with decentralized identities, offer personalized, privacy-protected, and continuously optimized GenAI services. These agents operate within a decentralized infrastructure, seamlessly integrating multiple tools and information sources to deliver multi-modal and complex services tailored to user needs. Through a decentralized marketplace, agents can offer and trade their services, receiving feedback and funds to enhance their model and hardware.

The process of deploying these AI agents throughout the network and user devices, enables them to collaborate, process multi-modal data, and provide context-aware services. This is achieved through several technical steps:

1. Agent deployment and decentralized identity

AI agents are strategically deployed across various layers of the 6G network, including central cloud, edge nodes, and user devices, to maximize their reach and effectiveness. Each agent is assigned a unique blockchain-based identity, which serves as a cornerstone for secure, verifiable interactions within the decentralized network. This identity system facilitates the establishment of trust, as agents can authenticate themselves and their actions in a transparent manner. Additionally, decentralized identities enable precise permission management, allowing agents to access resources and services in accordance with their defined roles and capabilities. This framework also supports accountability, as every action and decision made by an agent is recorded and traceable, promoting responsible behavior and compliance with predefined operational standards throughout the agents' lifecycle.

2. Multi-modal data processing and context awareness

The AI agents are equipped with advanced capabilities to process diverse data types, including audio, video, text, and sensor inputs. These multi-modal data streams are integrated using sophisticated AI models that enable a comprehensive understanding of the user's context and specific requirements. By continuously analyzing real-time data, the agents can infer context, such as a user's location, activity, preferences, and even emotional state, allowing them to deliver highly personalized and context-aware responses. This dynamic adaptability is crucial for maintaining relevance and enhancing user experience, as agents can tailor their interactions and services in real time. This approach ensures that the services provided are not only accurate but also aligned with the nuanced needs of users, thereby increasing the effectiveness and satisfaction of agent-driven interactions.

3. Self-evolution and continuous learning

Distributed AI agents emerge as autonomous intelligent entities based on GenAI, capable of dynamically controlling and invoking RAN devices and resources in real time. These agents ensure high-quality, responsive services while optimizing the performance of the network. Agents possess the ability to self-evolve and continuously improve through a structured lifecycle that includes phases of Initialization, Early Operation, Steady Operation, and Termination, as shown in Figure 5. Utilizing federated learning, agents collaboratively learn from network interactions while preserving privacy, allowing them to refine their models, enhance their capabilities, and autonomously adapt to new scenarios. Throughout their lifecycle, agents dynamically assess their own status and leverage available resources for hardware upgrades and operational optimization.

In the Initialization phase, agents secure initial funding from various sources. These funds can be drawn from an internal network operator or external GenAI maintain by external network operators. The initial funding is primarily defined as the hardware and network resources allocated to set up the agent. Operators invest physical infrastructure, such as computing power and connectivity resources, while agents begin using these resources to convert computing power into services. The

success of these services is measured through network metrics established by RAN operator, such as throughput, latency, and user satisfaction ratings. Agents generate revenue based on these metrics, which serve as the key performance indicators for assessing the effectiveness of their services.

As agents move into the Early Operation phase, they refine or expand their services based on real-time feedback from the RAN environment, optimizing models and improving resource utilization. Successful agents continue to generate value for the network by meeting or exceeding the established KPIs, thus earning rewards and maintaining the flow of resources. If the agent fails to meet the expected performance standards, its resources may begin to deplete. In such cases, agents may be incentivized to self-terminate, consolidating their remaining resources.

Upon entering the Steady Operation phase, agents provide stable services and reinvest any surplus revenue into further enhancements, including the development of new agents within the ecosystem. At this point, agents may also begin repaying loans or rewarding investors in accordance with agreements made during the Initialization phase. In this phase, agents may choose to sell their services or optimize their resource pools on an open marketplace. There are two types of markets available: the resource market and the service market. In the resource market, agents or external entities can buy or sell unused network and computational resources. In the service market, agents can offer their optimized, GenAI-driven services, which combine resources to provide a comprehensive solution.

In the Termination phase, agents conclude their operations, liquidate assets, and allocate any remaining funds. These funds can be reinvested into new projects or distributed to relevant stakeholders such as investors or contributors. During this phase, agents document essential insights and data that inform future developments and guide the next generation of agents. As part of the iteration process, hardware and network resources used by the agent are recovered and recycled for redeployment. These resources may be reallocated to new or emerging agents, which are deployed to take on new challenges or optimize previously unaddressed areas.

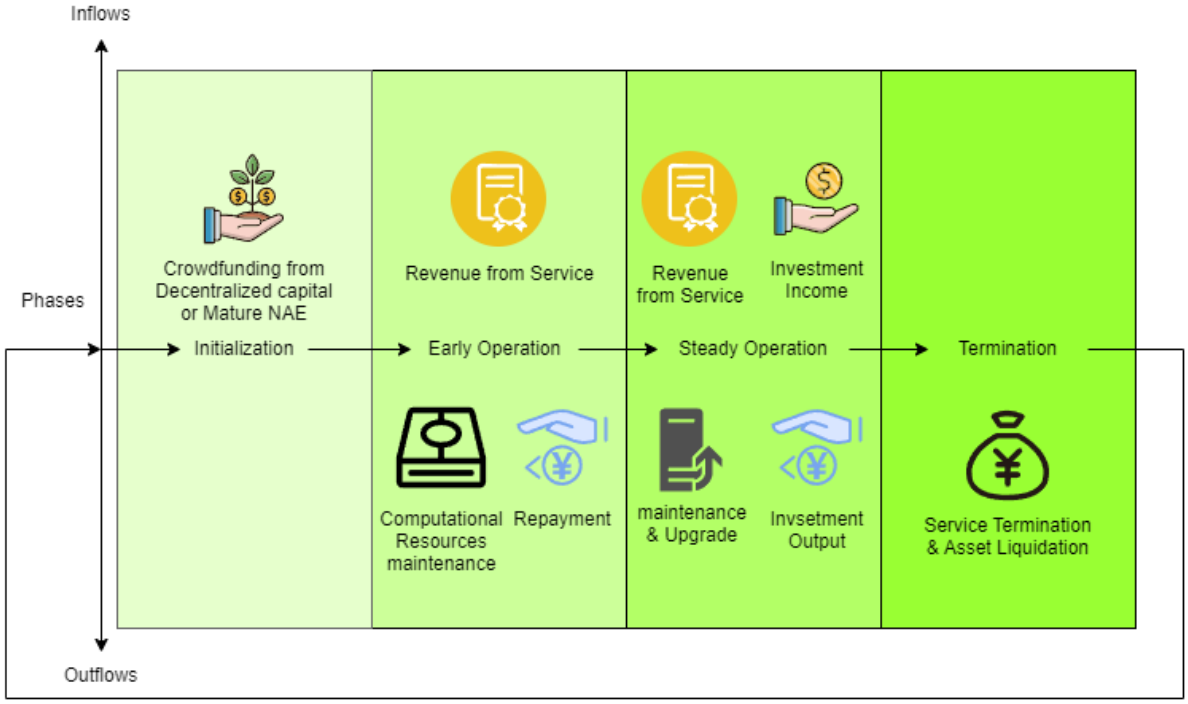

Figure 5 Distributed AI Agents lifecycle

4. Agent service ecosystem and marketplace

A decentralized marketplace forms a core component of the Distributed AI Agents framework. In this ecosystem, agent services can be tokenized and traded, allowing users, developers, and organizations to offer or acquire specialized services. Agents operate within this marketplace with a high degree of autonomy, using feedback and funds received from their services to optimize their capabilities and expand their offerings. This marketplace supports dynamic pricing models and flexible service contracts, enabling agents to adjust their service parameters in real-time based on demand, competition, and resource availability. By participating in this marketplace, agents not only enhance their models and hardware but also gain insights into market trends, which guide their development priorities and strategic positioning within the ecosystem. The marketplace also includes mechanisms for auditing and quality assurance, ensuring that the services provided meet predefined standards and that agents maintain accountability for their performance.

5. Security, privacy, and governance

To safeguard user data and maintain trust within the distributed AI agent ecosystem, the system employs end-to-end encryption and privacy-preserving AI techniques. These measures protect sensitive information throughout the entire lifecycle of data handling, from collection to processing and storage. Content control is essential in managing and ensuring secure identity management, with a tamper-proof record of all agent activities and interactions. Furthermore, robust governance mechanisms oversee the entire lifecycle of agents, including upgrades, role

adjustments, and compliance with ethical standards. This governance structure enforces security and privacy policies while maintaining transparency, ensuring that all operations align with the broader goals of accountability, user trust, and continuous evolution of the agent ecosystem.

2.2.3.2 Requirement

To fully support the Distributed AI Agents ecosystem in 6G networks, several key requirements must be met:

REQ1: Decentralized Identification. A security mechanism in the infrastructure is required to support the assignment of decentralized identities to AI agents, enabling secure and verifiable interactions within the distributed network. A native blockchain setup could be leveraged to facilitate secure transactions in the agent marketplace and ensures transparent governance, allowing agents to operate autonomously within the decentralized infrastructure.

REQ2: Interoperable AI Agents. Standardized frameworks and/or APIs are needed to ensure seamless interaction and collaboration between AI agents across different developers and network domains, facilitating a cohesive agent ecosystem.

REQ3: AI-optimized and secure communication protocols. The network is required to incorporate communication protocols optimized for AI operations, including efficient model updates, federated learning, and semantic communications between agents. Additionally, these protocols are required to integrate advanced security and privacy measures, such as zero-knowledge proofs, secure multi-party computation, and privacy-preserving AI methods.

2.3 6G network native GenAI services

2.3.1 Phone call simultaneous interpretation

2.3.1.1 Description

The phone call simultaneous interpretation use case in 6G network allows real-time translation of spoken language during phone calls, effectively bridging communication gaps between individuals who speak different languages using 6G network native GenAI services. The process involves capturing the speech of each participant, translating it into the target language, and delivering the translated speech to the target participants almost instantaneously.

The process entails several technical steps to ensure accurate, seamless, and natural conversation flow.

1. Speech recognition

The caller’s speech is captured through their device’s microphone. Then, at the device or edge server, the audio data is processed using speech recognition algorithms, to convert the spoken language into text or other forms of representations in real-time.

2. Translation

The processed speech is then translated into target language chosen by receiver. Then, the translated text or other forms of representation is converted back into speech using text-to-speech synthesis models.

3. Voice adaption

The voice adaptation process is used to replicate the voice of caller's speech, but in translated target language. This involves training a text-to-speech synthesis model on the caller's speech data to capture unique vocal characteristics, including tone, pitch, and speaking style. The model is then used to recover the caller's voice when the phone call is made.

2.3.1.2 Requirement

REQ1: Edge computing is important in reducing the transmission volume of simultaneous interpretation. By recognizing user speech, and generating the translated speech close to the user, 6G networks can significantly compress the original voice data into text or other forms of representation, thus leverage backhaul pressure and reducing the burden on centralized speech processing. The incorporating of training and inference of customized GenAI models at 6G network edges is also required to enabled voice adaptation to reflect the unique characteristics of participating users.

REQ2: The transmission and/or synchronization of GenAI models between network edges is needed to quickly establish simultaneous interpretation services. Alternatively, such voice recognition and synthesis computations can be performed locally within the UE, and 6G network is required to collaborate with UE in enabling such end-to-end service.

REQ3: As the phone call latency is critical to user experience, 6G network is required to reduce the end-to-end delay to acceptable ranges. The end-to-end delay includes voice recognition time, data transmission time, and voice synthesizing time. 6G network is expected to assure the latency and stability of the service in a whole, jointly optimizing the computing and communication tasks involved through task orchestrating, etc. This also requires 6G network to handle mobility issues as users move geographically.

2.3.2 Smart assistant application

2.3.2.1 Description

Traditional voice assistants can only execute simple commands, such as playing music or checking the weather. They lack the ability to provide personalized and proactive services. Intelligent assistant applications in 6G network will no longer be passive tools that receive instructions. They will become active partners capable of learning, thinking, and creating. Leveraging powerful GenAI technology, particularly the deep learning capabilities of LLMs and Natural Language Processing, they can accurately understand your intentions and perform complex operations. More

importantly, they will continually learn your usage habits, preferences, and lifestyle to provide tailored services.

This means that the intelligent assistant application can:

1. Deeply understand natural language

Going beyond simple keyword recognition, it can truly comprehend the complexity and nuances of human language, including semantics, context, and emotion. This allows it to respond in a way that is more appropriate to the user's expression and tone.

2. Provide personalized services

By continuously learning users' behavior, preferences, and needs, the intelligent assistant application can tailor services for each user. This includes recommending restaurants, movies, or music that match their tastes; reminding them of important anniversaries or birthdays; creating travel itineraries; and even anticipating potential risks and offering warnings and suggestions.

3. Proactively offer assistance

Instead of passively waiting for instructions, the intelligent assistant application can proactively provide help and suggestions based on the user's context and needs. This includes providing destination information, booking flights and hotels before the user's trip, or offering online medical consultations when the user is sick.

Intelligent assistant applications in 6G network will be more than just simple tools; they will become indispensable companions in our lives. They will fundamentally change how we interact with the world, opening up a future of greater intelligence, convenience, and overall well-being.

2.3.2.2 Requirement

To realize these intelligent assistant applications and seamlessly integrate them into our daily lives, the 6G network needs to meet the following key requirements:

REQ1: Ultra-high bandwidth. Intelligent assistant applications require processing large amounts of data and frequent interaction with cloud servers. Therefore, the 6G network needs to provide ultra-high bandwidth to meet the data transmission requirements of these applications.

REQ2: Universal connectivity. Intelligent assistant applications need to connect and interact with a wide range of devices and services, such as smart home devices, wearables, vehicles, and medical devices, and more. This requires the 6G network to provide universal connectivity that enables seamless information flow.

REQ3: Collaborative intelligence. Intelligent assistant applications require powerful AI capabilities, which requires efficient intelligent collaboration between the UEs, the network, and third-party application services, e.g., crowd-sourced decision making or knowledge retrieval across multiple nodes.

REQ4: Security, reliability, and privacy protection. Intelligent assistant applications collect and process a large amount of personal user information, such as voice,

images, location, and health data, etc. Therefore, data security and privacy protection are paramount. The 6G network needs to provide a more secure and reliable network environment, as well as a more comprehensive data security and privacy protection mechanism, to ensure user information security.

In summary, the 6G network is the foundational infrastructure of the realization of intelligent assistant applications. The full potential of intelligent assistant applications, which promises a truly convenient, intelligent, and secure future living experience, can only be realized when there is sufficient network capacity.

2.3.3 Semantic communication

2.3.3.1 Description

Semantic communication in 6G network aims to improve spectrum utilization and save communication resources through conveying desired meaning rather than bit-wise data. Therefore, an AI/ML enabled semantic encoder and semantic decoder is deployed in transceiver and receiver respectively, to extract and restore the content to be transmitted. Furthermore, Joint Source-Channel Coding (JSCC) has been proposed to utilize AI to encode the source content and transform it into electromagnetic signal at the same time, achieving better content recovering performance.

For source-only semantic communication, GenAI can be utilized for knowledge construction, semantic extraction and content restoration with its excellent performance in creating enormous multimodal content, understanding context and recovering multimedia data.

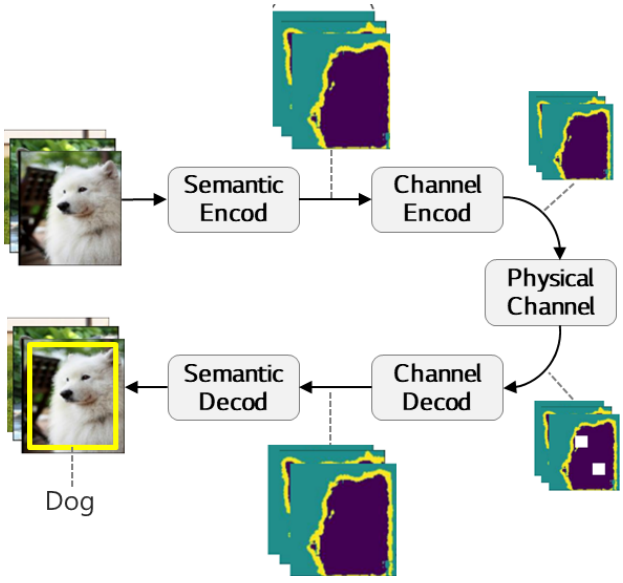

Figure 6 Semantic communication example

An end-to-end semantic communication system mainly consists of three significant components.

1. Semantic encoding

The transmitter filters irrelevant content and extracts key information for transmission through context reasoning, with GenAI encoding models pre-trained and fine-tuned on extensive datasets.

2. Semantic decoding

Benefitting from the high-quality creativity of GenAI, the receiver can utilize decoding models and rendering resources to perform semantic comprehension and multimedia reconstruction.

3. Common knowledge

Common knowledge is the shared information between transceivers. Large language models like Generative Pre-trained Transformer (GPT) can easily offer vast content for knowledge base construction. Additionally, knowledge base is supposed to update simultaneously and keep synchronized for superior semantic communication performance.

For JSCC-based semantic communication, it possesses the capability against channel variations with CSI integrated into semantic coding [23] [24], hence the passthrough of CSI key information within semantic communication is essential.

The overall CSI key information passthrough consists of public keys exchange, source and channel information circulation, common knowledge delivery and semantic decoding information alignment. Particularly, with availability to joint channel and source information, the critical step is to fine-tune the AI model of semantic codecs with private knowledge transferred within TEE.

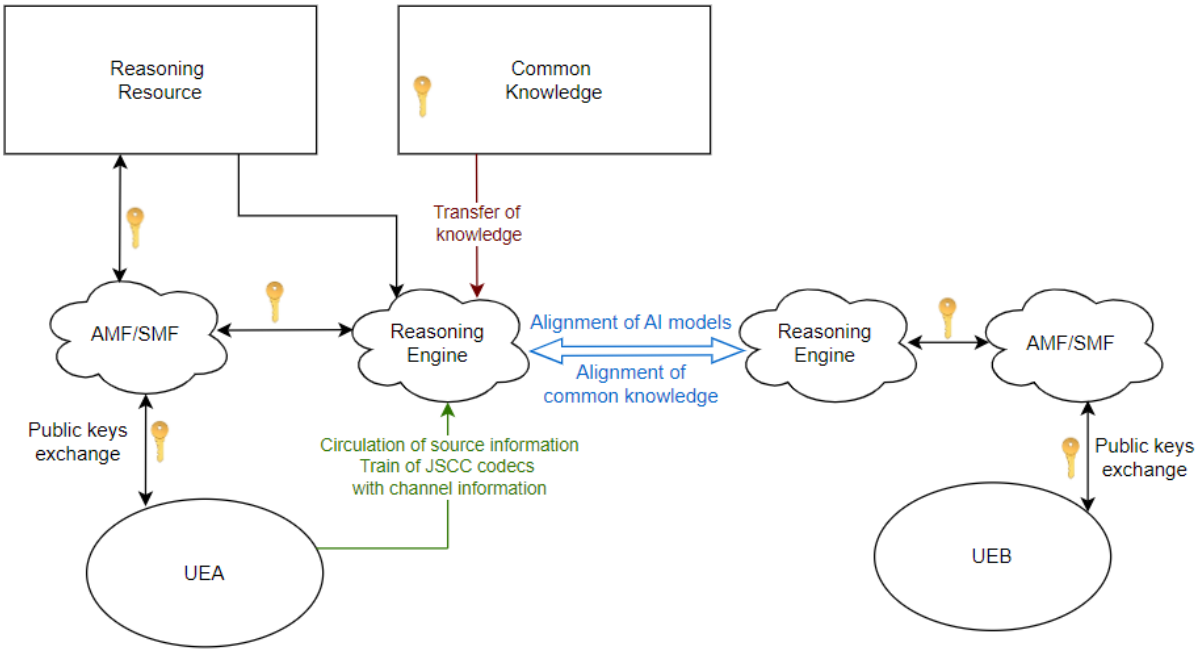

Figure 7 Semantic communication process

1. Exchange of public keys

Public key-based method can protect data privacy and security within CSI key information passthrough. There are two different keys for public key encryption, namely the public key available to anyone and private key only kept by one party. With transmitted data encrypted with receiver's public key, the receiver can decrypt content with the corresponding private key. In conclusion, it is required to exchange public keys with relevant network functions beforehand for privacy-preserving purpose.

2. Circulation of source and channel information

It is significant to enable close interactions between source information and channel information for JSCC. Furthermore, access to raw source data needs to ensure privacy and security through public key-based method and TEE.

3. Transfer of knowledge

Shared knowledge between semantic transmitters and receivers is the enabler of semantic communication, involving AI models, object artifacts and databases. With joint source and channel information, JSCC codec models can be trained with transferred knowledge.

4. Alignment of decoding information

By aligning the decoding information including JSCC AI models and common knowledge, receiver is capable of performing semantic interpretation and multimedia restoration.

2.3.3.2 Requirement

REQ1: Considerable computing and storage resources are required for semantic processing which includes codecs training and inference. It has exceeded the computing capability of current 5G RAN and 5G UE modem processor (especially AI computing), thus more computing and storage resources are demanded for both network and mobile devices. Semantic processing needs to be distinguished from normal data service without computing demands and processed separately because it requires additional computing processing and is only activated on demand. This also requires 6G network to manage configuration and provisioning of reasoning resources.

REQ2: CSI key information passthrough needs close inter-operations among users, control plane functions like AMF/SMF and user plane functions like UPF, thus 6G network is required to promote cooperation and breach the barrier between the control plane and user plane in compliance with privacy-preserving. Besides, 6G network is expected to provide content and channel owners a holistic approach allowing extensive resource integration.

REQ3: As the knowledge for semantic communication is continuously updating, keeping shared knowledge between two transceivers synchronized through distributed knowledge exchange scheme is required for 6G network. Knowledge

privacy and security preserving is also emphasized. Data privacy preserving is of vital importance within CSI key information passthrough. This requires 6G network to secure the data confidentiality using public key-based method with the help of TEE.

2.3.4 Immersive communication

2.3.4.1 Description

Content-aware semantic communication could potentially provide extensive support for eXtended Reality (XR) with GenAI technology. One of the typical applications of GenAI in XR is creating high-fidelity virtual objects and interactions within the wireless metaverse. GenAI is utilized to enhance the capabilities of semantic multiverses by enabling agents to manipulate and render high-fidelity virtual scenes and interactions based on learned semantic representations of multi-modal data.

Specifically, the above application refers to techniques such as text-to-3D object creation, which allows for the generation of 3D models based on textual descriptions. This capability is essential for rendering immersive experiences in the metaverse, where users can interact with both virtual objects and intelligent agents.

2.3.4.2 Requirement

REQ1: Visual fidelity, semantic consistency and diversity need to be introduced as performance metrics for XR on 6G networks. Visual fidelity can be assessed through subjective evaluations (user studies) or objective measures that compare the generated objects to real-world counterparts or high-quality reference models. Metrics like Structural Similarity Index (SSIM) or Peak Signal-to-Noise Ratio (PSNR) can be used for quantitative assessments. Semantic consistency evaluates whether the generated objects accurately represent the intended semantics or concepts described in the input (e.g., text descriptions). This can be assessed through human judgment or automated methods that compare the generated object features with expected features. Diversity indicates the ability of the generative model to produce a wide range of unique and varied objects like photorealistic images, videos and animations, which can be measured using metrics like Inception Score or Fréchet Inception Distance, commonly used in generative models for images.

REQ2: In real-time applications, computational efficiency is especially important to provide users with smooth and immersive experiences. Computational efficiency is related to the required time and resources to generate the virtual objects. To enhance computational efficiency, distributed computing architecture, high-speed network transmission and semantic-aware technologies are required.

3 Key requirements and research directions

This chapter outlines the essential requirements for implementing GenAI within 6G network based on the use cases explored in this report. To effectively utilize GenAI

techniques in augmenting 6G network from all aspects, as well as support a diverse range of GenAI applications, several core requirements must be addressed across network infrastructure, data transmission, network functions, and network AI design.

3.1 Network computing

3.1.1 Distributed collaborative computing

Many GenAI use cases, such as real-time interactive video generation, phone call simultaneous interpretation, and smart assistant applications, require high computational power at the edge. Distributed computing collaboration allows these complex generative tasks to be partitioned, distributed and jointly processed in multiple nodes including the user equipment, ensuring minimal latency and improving user experience. By enabling computational resources at multiple network layers—UE, RAN, edge, and core—6G network can support real-time efficient model inference and data processing.

3.1.2 Stronger RAN/edge computing capacity

Edge nodes must support resource-intensive GenAI tasks such as model inference and rendering for most of the analyzed applications. For example, real-time interactive video generation relies on edge computing for high-quality video generation and personalization, while the GenAI-aided physical layer processing requires robust computational capabilities to handle tasks like channel prediction and signal denoising. This needs even stronger computing resources dedicated for GenAI model computing.

3.1.3 Ultra-low model inference time

Achieving ultra-high speed model inference is crucial for applications such as GenAI-aided physical layer processing. To meet these requirements, 6G networks should integrate specialized hardware and optimize GenAI model architectures to support inference within sub-frame or symbol-level latency.

3.1.4 Joint optimization of communication and computation tasks

Use cases like GenAI-aided digital twin and phone call simultaneous interpretation demonstrate the need for coordinated communication and computation across network layers. As data processing delay becomes a non-negligible component of end-to-end delay in these cases, to minimize latency and maximize efficiency, 6G networks must implement task orchestration that optimally allocates resources to tasks based on heterogeneous hardware resources, real-time conditions and application demands.

3.1.5 Flexible and scalable edge storage

GenAI applications, especially those requiring personalization or extensive knowledge, demand flexible storage at the RAN or network edge. For instance, interactive video generation use case generates personalized contents based on user preference, the

local module should store and process personal preferences information, necessitating scalable edge storage to accommodate user data and numerous lightweight models for individual users.

3.1.6 Secure computing service exposure

Distributed AI agents and edge-based computing services for GenAI benefit from secure exposure of computing resources to third-party applications. Encryption and trusted execution environments, e.g., TEE, will be vital in protecting sensitive data and ensuring the integrity of exposed computing services.

3.2 Data transmission

3.2.1 Stricter QoS and throughput/latency improvement

High-bandwidth, low-latency connectivity is essential for applications that involve real-time user interaction, such as edge-based computing service, smart assistant and immersive communication, etc. These applications require several times higher bandwidth, lower latency and stricter QoS protection compared with current network. Furthermore, evaluation of quality of experience(QoE) for GenAI application differs from other applications, e.g. first token response time, LLM recognition correctness rate, etc.

3.2.2 Communication protocols optimized for AI

In 6G networks, communication protocols optimized for AI are essential for seamless AI collaboration across distributed environments. These protocols must ensure the rapid transmission, timely updates, and synchronization of AI models and their associated data across a wide diversity of end users, while also safeguarding data privacy and security. These optimized protocols will drive scalable, secure, and innovative AI applications in the 6G network ecosystem.

3.2.3 Adaptive and efficient data compression

Future 6G networks will require adaptive and efficient data compression and transmission capabilities with the aid of GenAI to support data-driven applications, such as RAN digital twin. By extracting essential semantic information and encoding raw data in low-dimensional representations, network load, backhaul pressure, and data collection latency can be efficiently reduced. This approach enhances overall network performance and user experience, providing robust support for a wide range of data-driven applications.

3.2.4 Data privacy and security

Privacy is paramount, particularly for GenAI applications processing sensitive user data, such as smart assistant applications and distributed AI agents. Data encryption,

identity verification, and TEEs should be used to protect data in transit and at rest, complying with stringent privacy requirements.

3.3 Network functions

3.3.1 Decentralized identification

Distributed AI agents rely on secure, decentralized identities to function autonomously within a 6G network. Using blockchain or similar technologies, decentralized IDs enable secure, transparent interactions across a distributed AI ecosystem, enhancing security and trust among various network components.

3.3.2 Access to external knowledge and data source

To enhance the functionality and performance of GenAI applications in 6G networks, it is essential for the network to have access to external knowledge and data source, e.g. customer documentation, ticket resolutions, sensor data, camera images, and fault alarms. Many GenAI use cases, such as intent-based network management and multi-modal AI-based network analysis in 6G network, rely on up-to-date, domain-specific and multi-modal external information to improve decision-making, accuracy, and adaptability. Discovering and accessing to external knowledge and data sources requires robust data retrieval mechanisms. Certain challenges must be addressed, such as ensuring data privacy, managing access rights to proprietary data, and minimizing latency in data retrieval.

3.3.3 Universal connectivity

Universal connectivity is crucial for GenAI applications that interact across multiple device types, such as smart assistant applications. 6G networks should support seamless connections across IoT devices, wearables, medical devices, unmanned systems and other digital tools, ensuring compatibility and consistent service delivery in a diverse ecosystem.

3.3.4 Network programmability

To support GenAI applications like network management automation, the 6G network should provide programmable interfaces. These allow for flexibly accessing the network, acquiring data, and performing operations without significant infrastructure changes.

3.4 Network AI design

3.4.1 AI explainability

For GenAI to gain widespread adoption in critical applications like AI agent-based and intent-based network management, explainable AI is essential. Techniques such as attention mechanisms, decision tree and interpretable model architectures enable transparency, allowing operators and users to understand AI decisions and build trust in automated processes.

3.4.2 Life Cycle management of evolving GenAI procedures

6G networks should remain adaptable to support the current and future GenAI models, such as diffusion models and transformer-based architectures and many more used in physical layer processing and semantic communication. Despite the existing LCM of AI/ML models, new procedures and LCM should be studied and incorporated into the network facing GenAI model operation (e.g., pre-training, fine-tuning, prompt engineering, etc.) .

3.4.3 Intelligent collaboration between UE, network and 3rd-party App

The smart assistant and distributed AI agent applications exemplify the need for intelligent collaboration across the user equipment (UE), network, and third-party apps. This collaboration supports personalized services, distributed processing, and efficient use of resources.

3.4.4 Transmission and/or synchronization of GenAI models

Real-time applications, such as phone call simultaneous interpretation and real-time interactive video, require models to be synchronized and transmitted seamlessly across the network. Maintaining updated model parameters across nodes and user devices ensures high-quality, responsive performance.

3.4.5 Standardized framework and/or API for AI Agent operation and collaboration

Developing a standardized framework and/or API will streamline the deployment, coordination, and interoperability of AI agents across the 6G network. This standardization is essential for distributed AI agent ecosystems, enabling seamless collaboration across various devices and services.

3.5 Research directions

3.5.1 Communication protocol for AI tasks

Existing communication protocols are designed primarily for generic data transmission and lack the efficiency needed for AI-specific tasks. As 6G networks integrate GenAI for real-time services (e.g., simultaneous interpretation, interactive video generation, semantic communication and immersive communication, etc.), there is a need for protocols optimized for high-frequency data transfer, model synchronization, and multi-modal data transmission. Key research topics include semantic-aware data compression, adaptive data processing and prioritization, fast synchronization of AI modes and security measures.

3.5.2 GenAI algorithm for network augmentation

Generative AI algorithms have the potential to transform network operations by generating synthetic data, predicting network conditions, optimizing configurations in real time and physical layer signal processing. Applications such as the network management automation, GenAI-aided RAN DT and GenAI-enhanced physical layer processing, etc. require sophisticated algorithms that can handle the complexity of 6G networks. Key research topics include data augmentation for RAN DT, incorporating GenAI models into physical Layer, telecommunication foundation models and Network management AI Agents, etc.

3.5.3 Advanced distributed computing framework

A scalable distributed computing framework is critical for handling the computational demands of real-time GenAI applications, such as interactive video generation, edge-based GenAI service, and distributed AI agents. This framework must facilitate seamless task distribution, resource management, and load balancing across the entire network. Key research topics include edge-RAN-Core integration, dynamic task orchestration, distributed GenAI training and inference and resource virtualization and scalability, etc.

3.5.4 6G network native GenAI applications and optimization

The integration of GenAI into 6G networks opens up new possibilities for native applications that are deeply embedded within the network's service. These applications, such as simultaneous translation, smart assistants, immersive communication, require continuous evolution of MTSI system and tailored optimization strategies to meet performance expectations. Key research topics include Application developing and capability exposure, dedicated 6G network architecture design, immersive communication devices and protocols, etc.

4 Conclusion

This research report presents an analysis of the use cases and requirements for integrating GenAI within 6G networks. Three categories of use cases: GenAI for 6G network, 6G network for GenAI applications and 6G network native GenAI services are analyzed with critical requirements on 6G network identified.

Four groups of requirements, namely requirements on network computing, data transmission, network functions and network AI design, are summarized, followed by the identification of key research directions involving communication protocol for AI tasks, GenAI algorithm for Network augmentation, advanced distributed computing framework and 6G network native GenAI applications and optimization.

In conclusion, the successful integration of GenAI into 6G networks will enable unprecedented levels of automation, personalization, and efficiency, unlocking new opportunities for both consumers and industries. As we move towards this vision, continued collaboration between network operators, AI researchers, and technology providers will be crucial to overcoming challenges and realizing the full potential of 6G networks powered by GenAI.

References

[1] S. S. Sengar, A. B. Hasan, S. Kumar, C. Fiona, "Generative artificial intelligence: a systematic review and applications, " in Multimedia Tools and Applications, Aug. 2024

[2] J. Tong, W. Guo, J. Shao, Q. Wu, Z. Li, Z, Lin, and J. Zhang, "Wirelessagent: Large language model agents for intelligent wireless networks, " arXiv preprint arXiv:2505.01074, 2025 Online: https://arxiv.org/abs/2505.01074

[3] I. Chatzistefanidis, A. Leone and N. Nikaein, "Maestro: LLM-Driven Collaborative Automation of Intent-Based 6G Networks," in IEEE Networking Letters, vol. 6, no. 4, pp. 227-231, Dec. 2024

[4] A. Mekrache et al. “Intent-Based Management of Next-Generation Networks: an LLM-Centric Approach,” in IEEE Network, vol. 38, no. 5, Sep. 2024

[5] S. Pratar et al., "The fine art of fine-tuning: A structured review of advanced LLM fine-tuning techniques, " in Natural Language Processing Journal, vol. 11, Jun. 2025

[6] A. Balaguer et al., "RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture, " in Computation and Language, Jan. 2024

[7] Y. Tian, Q. Zhao, F. Boukhalfa, K. Wu, F. Bader et al., “Multimodal Transformers for Wireless Communications: A Case Study in Beam Prediction,” arXiv preprint arXiv:2309.11811, 2023. Online: https://arxiv.org/abs/2309.11811

[8] R. Zhang, L. Cheng, S. Wang, Y Lou, Y. Gao, W. Wu, DWK. Ng, "Integrated Sensing and Communication With Massive MIMO: A Unified Tensor Approach for Channel and Target Parameter Estimation," in IEEE Transactions on Wireless Communications, vol. 23, no. 8, pp. 8571-8587, Aug. 2024

[9] Z, Zhang, F. Wen, Z. Sun, X. Guo, T. He and C. Lee, "Artificial Intelligence-Enabled Sensing Technologies in the 5G/Internet of Things Era: From Virtual Reality/Augmented Reality to the Digital Twin, " in Adv. Intell. Syst., 4: 2100228, 2022

[10] Y. Huang, Y. Xie, Z. Chen, X. Xue, Q. Sun and N. Li, "AI Empowered Modeling, Closed-loop Optimization and Field Trials of RAN Digital Twin," in IEEE Network, Apr. 2025

[11] S. Wagle, A. Malhotra, S. Hamidi-Rad, M. S. Ibrahim and C. G. Brinton, "Joint Spatio-Temporal Feature Extraction for Channel State Prediction in MIMO Systems," in Proc. 2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2025

[12] M. Yiğit and A. B. Can, "GISD: generation of corner cases in infrared autonomous driving dataset with stable diffusion." in Applications of Machine Learning 2024. Vol. 13138, Oct. 2024

[13] Z. Zhang, J. Wang, J. Chen, H. Fu, Z. Tong and C. Jiang, "Diffusion-Based Reinforcement Learning for Cooperative Offloading and Resource Allocation in Multi-

UAV Assisted Edge-Enabled Metaverse," in IEEE Transactions on Vehicular Technology, Feb. 2025

[14] M. Xu et al., "Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services," in IEEE Communications Surveys & Tutorials, vol. 26, no. 2, pp. 1127-1170, Jan. 2024

[15] Y. Liu, Z. Tan, H. Hu, L. J. Cimini and G. Y. Li, "Channel Estimation for OFDM," in IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1891-1908, Oct. 2014

[16] P. Dong, H. Zhang, G. Y. Li, I. S. Gaspar and N. NaderiAlizadeh, "Deep CNN-Based Channel Estimation for mmWave Massive MIMO Systems," in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 5, pp. 989-1000, Sep. 2019

[17] Y. Zeng et al., "CSI-GPT: Integrating Generative Pre-Trained Transformer With Federated-Tuning to Acquire Downlink Massive MIMO Channels," in IEEE Transactions on Vehicular Technology, vol. 74, no. 3, pp. 5187-5192, Mar. 2025

[18] M. Kim, R. Fritschek, and R. F. Schaefer, “Learning end-to-end channel coding with diffusion models,” in Proc. WSA & SCC 2023, pp. 1–6, 2023

[19] T. Wu, Z. Chen, D. He, L. Qian, Y. Xu, M. Tao, and W. Zhang, “CDDM: Channel denoising diffusion models for wireless communications,” in Proc. IEEE GLOBECOM 2023, pp. 1–5, 2023

[20] B. Banerjee, R. C. Elliott, W. A. Krzymień and M. Medra, "Hybrid Beamforming for mmWave Massive MIMO Systems Using Conditional Generative Adversarial Networks," in IEEE Transactions on Vehicular Technology, vol. 73, no. 10, pp. 15803-15808, Oct. 2024

[21] C. Chaccour, W. Saad, M. Debbah, Z. Han and H. V. Poor, "Less Data, More Knowledge: Building Next Generation Semantic Communication Networks," in IEEE Communications Surveys & Tutorials, 2024

[22] E. C. Strinati et al., "Goal-Oriented and Semantic Communication in 6G AI-Native Networks: The 6G-GOALS Approach," in Proc. 2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Antwerp, Belgium, 2024

[23] J. Xu, B. Ai, N. Wang and W. Chen, "Deep Joint Source-Channel Coding for CSI Feedback: An End-to-End Approach," in IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 260-273, Jan. 2023

[24] P. Jiang, C. -K. Wen, S. Jin and G. Y. Li, "Wireless Semantic Communications for Video Conferencing," in IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 230-244, Jan. 2023