GTI

5G-A IGNITES

THE THREE-TYPES OF

NEW INTELLIGENT SERVICES

5G-A Ignites the Three-Types of New Intelligent Services

Version	V1.0.0
Deliverable Type	□ **Procedural Document** ☑ **Working Document**
Confidential Level	□ **Open to GTI Operator Members** □ **Open to GTI Partners** ☑ **Open to Public**
Working Group	**5G Technology and Product Program**
Source members	China Mobile, Huawei, Leju Robotics
Support members	ZTE, Nokia, CICT Mobile, DEEPRobotics, HONOR, Qualcomm, MTK, Fibocom, TD Tech, ASR, Quectel, UNISOC
Last Edit Date	14-06-2025
Approval Date	

Document History

Date	Meeting #	Version #	Revision Contents
06-14-2025		V1.0.0	

Content

Introduction

At present, the world is accelerating towards a new era of digitization and intelligence in which everything is connected, and the pattern of information services ushering in structural changes driven by 5G-Advanced (5G-A) technologies. As a key stage in the evolution from 5G to 6G, 5G-A is becoming the core engine for information services through a tenfold upgrade of technical capabilities and multi-dimensional integration and innovation. This white paper focuses on the three -types of new intelligent services—intelligent robots, intelligent devices, and intelligent connected vehicles—and systematically compiles their typical scenario requirements, with the aim of providing forward-looking guidance for the intelligent upgrading of the industry. In addition, this white paper also focuses on the scale development of the three-types of new intelligent services enabled by 5G-A. Through standardization, cross-area collaboration, and ecological co-construction, 5G-A will accelerate the transition from capability enhancement to value creation, injecting value into the digital economy. In the future, the in-depth 5G-A×AI convergence will further unleash the potential of “Human-Vehicles -Device” interconnection and open a new era of intelligent society.

Chapter 1 Synergistic Progression of Technology Revolution & Mobile Communication

1.1. Evolutionary Journey of Mobile Communication Services

Communication is an old but modern word. Before the 19th century, people communicated mainly by letter. In the 19th century, with the enlightenment and rapid development of communication theories and basic sciences, communication technology witnessed great progress entered the era of mobile communication.

In the 1980s, the First generation analog mobile communication system (1G) based on the concept of “cellular” achieved large-scale commercialization, using

Frequency Division Multiple Access (FDMA) technology to realize analog modulation of voice signals. The Second Generation digital mobile communication system (2G) is based on Time Division Multiple Access (TDMA) technology, to transmit voice and low-speed data services, and realize global roaming. With the further development of data and multimedia communications, the Third Generation mobile communication system (3G) came into being, using Code Division Multiple Access (CMDA) technology to enhance the security of data communications, not only to provide high-quality voice services, simultaneous transmission of voice and data information, but also to support multimedia services and the access to the mobile Internet.

3G has realized the goal of mobile broadband multimedia communication, but it has not stopped people from further research on communication technology. The 4th Generation (4G) mobile communication technology is represented by 3GPP LTE/LTE-A system, which introduces Orthogonal Frequency Division Multiplexing), MIMO (Multi-Input & Multi-Output (OFDM), and other key technologies. In 4G, benefiting from the improvement in network performance (such as rate and delay), the IP-based multimedia communication services and the integration of mobile communication and WLAN services are realized and bring new experiences for users.

On this basis, along with the vigorous development and innovation of upper-layer applications (including the needs of individual consumers and industry customers), the mobile communication network needs to provide diversified, flexible, and customizable services, and the rate is no longer the only pursuit. The 5th Generation (5G) mobile communication technology contains three typical application scenarios, enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communications (uRLLC), and massive Machine Type of Communication (mMTC), which put forward brand new requirements on rate, delay, reliability, connection scale and other capabilities.

Table 1 Summary of the evolution of mobile communication network and services

Mobile communication technology	Network capabilities	Services and applications
1G	/	Analog voice services
2G	Rate: 10~×100kbps	Digital voice, text
3G	Rate：(DL) 3.6 Mbps (UL) 382kbps	High-quality voice services, multimedia services, mobile Internet services
4G	Rate: 10Mbps~1Gbps Radio delay: ≧10ms Traffic density: 0.1~1Mbps/m^2 Connection density: 10^5/km^2 Mobility: 350km/h	IP-based multimedia communication services, VoLTE, IoT
5G	Rate: 100Mbps~10Gbps Radio delay: ms level Traffic density: 10Mbps/m^2 Connection density: 10^6/km^2 Mobility: 500km/h	eMBB：HD videos、VR/AR uRLLC：verticals (V2X, smart factory, etc.) mMTC：AIoT (smart logistics, smart cities, etc.)

Throughout the development of mobile communications, network and business, terminal interdependence between the three, promote each other, the evolution of the network to promote the rapid development of terminals and services, incubating more new needs of users, and the growing demand for a better life for the user has injected the driving force for the continuous development of mobile communications technology.

1.2. The Technological Revolution and the Synergistic Evolution of Mobile Communications

The technological revolution and mobile communication technology are reshaping the global industrial landscape with synergistic evolution spirally. Scientific and technological changes centered on artificial intelligence (AI), the Internet of Things (IoT), and cloud computing are driving service forms to deep intelligence. AI

models break through the boundaries of cognition and decision-making, equipping machines with human-like capabilities, and mobile communication technology (5G/5G-A) is becoming the "neural network" of intelligent transformation through the characteristics of ultra-low latency, ultra-large bandwidth, and ubiquitous connectivity. In the future, service development will be characterized by intelligence-infused applications, 3D-enabled content, and cloud-migrated services.

1. Intelligence-infused applications

Breakthroughs in AI technology and the AI-communication network convergence give rise to "AI +" application scenarios. AI applications with powerful data processing capabilities and deep learning technology can more accurately understand user needs and behavior, so as to provide more personalized, intelligent services. The integration of technologies in multiple fields has promoted the development of mobile AI services, and incubated three new types of information consumption services such as embodied intelligent robots, intelligent devices, and intelligent Internet-connected vehicles, etc. With the wide application of emerging AI services, the demand for connectivity and computing power has also shown significant growth.

2. 3D-enabled content

With the continuous maturity of 3D display technology and content generation technology, service content will gradually shift from 2D to 3D. 3D applications offer a richer experience for users, thanks to their unique immersive, high-quality display, and interactive capabilities. At present, the main forms are XR, glass-free 3D, etc. This shift from 2D to 3D not only helps to improve the efficiency of users' work and life, but also signals that future entertainment applications will pay more attention to user experience and interactivity, and promote the innovative development of the whole industry.

3. Cloud-migrated services

With the continuous maturation and popularization of cloud computing

technology, cloud-based services are gradually developing with the widespread deployment of 5G networks. The bandwidth and latency guarantees provide strong support for cloud-based services, enabling users to access cloud-based services anytime and anywhere, realizing a truly mobile work and life. In addition, the development of cloud-based services not only reduces operation and maintenance costs and improves resource utilization efficiency, but also enables services to respond to market changes more flexibly. In the future, cloud-based services will be applied in more fields, bringing users more convenient, efficient, and rich experiences.

Chapter 2 Typical Scenarios and Requirements

2.1. Overview of New Information Consumption Services

2.1.1. The Three-Types of New Intelligent Services Incubated by Mobile Communications

1. Embodied intelligent robots: breakthroughs in embodied intelligence

Embodied AI realizes the qualitative change of the closed loop of "Perception-Decision-Motion" by empowering robots with the ability to interact with real-time environments, promoting the transformation of robots from a single tool to an autonomous collaborative partner with an "embodied brain", and realizing autonomous learning and complex environment interaction. The technological breakthroughs of embodied intelligent robots focus on three major directions. The first is the integration of multi-modal perception, with the help of LiDAR, visual sensors, and tactile feedback to build human-like senses, for example, Boston Dynamics robots dynamically adapt to complex terrain, and China Mobile's "Body Brain" for robot training based on the Jiutian Big Model can integrate multi-source data to improve scene adaptability. The second is autonomous learning and generalization ability, through deep reinforcement learning and simulation training, robots can quickly migrate skills to real tasks, such as Tesla Optimus learning housework operations. The third is the upgrading of human-machine emotional

interaction, breaking through the traditional command mode, for instance, the pensioner robot can feel the user's emotions and provide active health support.

The embodied intelligent robot architecture includes the robot “brain” for planning and decision, the robot “cerebellum” for motion control, and the robot “body”. There are two types of architecture models, the hierarchical decision-making model and the end-to-end model. The hierarchical decision-making model breaks down the task into different layers, trains with multiple neural networks respectively, and then combines them in the way of a pipeline. There are data transfer, interaction, and coordination needs between different layers in the hierarchical decision-making model, with relatively less need for training data and a stronger scene generalization ability. Take Figure 1 robot as an example. The robot brain accesses AI multi-modal big model and provides visual inference and language understanding, the robot cerebellum acts as the cerebellum for motion control and generates torso sensing and execution, and the robot body accepts the motion commands from the neural network strategy for control execution. The end-to-end model combines the robot brain and the robot cerebellum into one, and directly converts the task objectives into control signals through a single neural network training, realizing a seamless connection from input to output, which has the advantage of data and computing power, and can support the complex training needs of the end-to-end model; however, the amount of data required is estimated to be at the level of hundreds of billions, and the speed of reasoning and response is slow. At present, most robotics companies employ the hierarchical decision-making model, such as Figure, AgiBot, Leju, etc., since the multi-layer industrial collaboration reduces product development and application cycles. Tesla and Google employ the end-to-end model, which has lower development complexity and can save the transmission and coordination costs brought about by multi-layer collaboration.

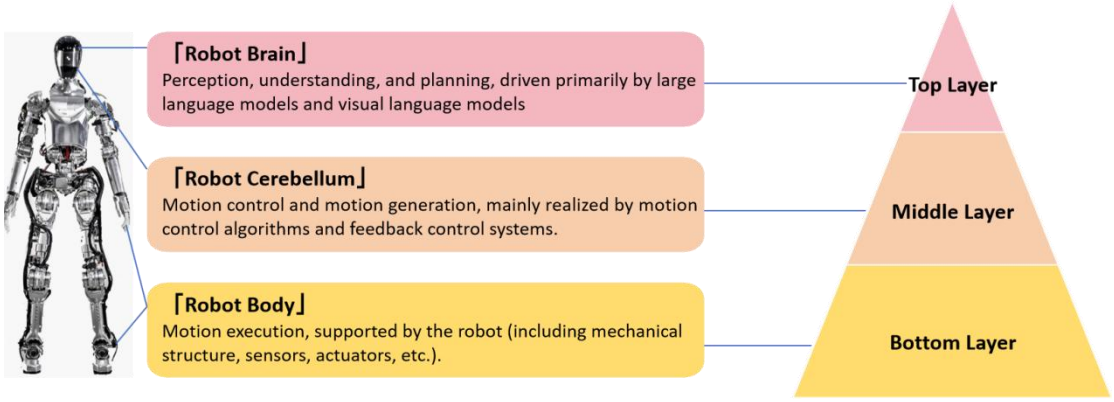

Figure 1 The embodied intelligent robot architecture

For the applications, embodied intelligent robots are extending from basic services to high-value scenarios. In family scenarios, AI dietitian robots achieve personalized health management. In the industry, multi-robot collaborative assembly improves automobile manufacturing efficiency. In dangerous environments, robots replace humans to carry out high-risk tasks, such as overhauling nuclear power plants. In the future, with the 5G-A/6G network optimizing delay and quantum computing to accelerate decision-making, 2030 is expected to enter the era of ubiquitous intelligent agents, in which robots will become the core carrier of the convergence of the digital economy and the entity industry.

2. Intelligent devices: accelerating ubiquitous intelligence

As the core entrance of information consumption, intelligent devices have evolved from communication tools to AI-driven multi-modal interaction platforms. Ubiquitous intelligence aims at the seamless convergence and active service, relying on the technological integration and ecological synergy of intelligent devices, to promote intelligent services to be seamlessly integrated into human life and realize the deep interconnection and active response of Human-Machines-Device-Environments. As a carrier, intelligent devices are accelerating this process through technological innovation and scene expansion. At the technological level, the end-side computing leap (AI cell phone NPU computing power up to 100TOPS) and multi-modal interaction upgrade (AR navigation, intent understanding) support tens

of billions of parameters of large models for localized operation. The end-cloud synergistic architecture (Hongmeng OS cross-equipment latency less than 0.5 seconds) and AI native operating system reconstruct the logic of the service, realizing the transformation from a “human-controlled device” to “equipment predicting demand”.

For the applications, in the ToC market, intelligent devices have entered our daily life. For example, whole-house smart devices are actively linked to handling security and energy consumption, and the AI cell phone has become an intelligent personal assistant. In the ToB scenarios, industry AI devices optimize the production line through digital twins, and the smart city promotes vehicle-road-cloud synergy to reduce congestion. Moreover, the emerging applications, such as the low-altitude UAV cluster scheduling and meta-universe XR, are building integrated virtual-reality experience and continuing to expand the boundaries.

The accelerated realization of ubiquitous intelligence is essentially the triple resonance of technology, scene, and ecology. From the breakthrough of computing power on the end side to the synergy of the whole scene, the intelligent device is evolving from an isolated device to a ubiquitous intelligent agent, reconfiguring the interaction mode and productivity form of human society. In the future, with the integration of 6G communication, quantum computing, and other technologies, intelligent devices will be more deeply integrated into the physical world, reshape the social operation paradigm, and become the core hub for the convergence of the digital economy and the entity industry.

3. Intelligent connected vehicles: reconstructing the traffic ecology

Intelligent connected vehicles take the deep integration of vehicle, road, cloud, network, and map as the core, and through technological breakthroughs and ecological restructuring, upgrade the traditional travel mode to an efficient, safe, and sustainable intelligent transportation system. At the technological level, multi-modal sensing and decision-making (LIDAR + high-precision maps to realize centimeter-level

environment recognition) and vehicle-road-cloud synergy (5G-A/6G network to support millisecond-level interactions) break the data silos, and the automotive software system (Tesla FSD, Hongmeng OS cross-end synergy) promotes function iteration and ecological openness. For the applications, Robotaxi's single-kilometer cost is reduced to RMB 1.8, and the intelligent cockpit reshapes the driving experience through AR-HUD and emotional interaction.

Intelligent connected vehicles are not only the product of technology integration but also the engine of travel ecological reconstruction. Through vehicle-road-cloud collaboration, data drive, and ecological openness, it is promoting the transformation of the transportation system from “people adapting to vehicles” to “vehicles serving people”. In the future, with the integration of 6G communication, quantum computing, and other technologies, intelligent connected vehicles will be deeply embedded in the smart city, and become the core node of the synergistic development of the digital economy and the entity industry.

2.1.2. Development Drivers and Policies of New Information Consumption Services

As a new type of growth point in the era of digital economy, the development of emerging businesses represented by the “New Three Kinds” of information consumption has been strongly supported by the Chinese government and local policies, and accelerated by multiple driving forces, such as technological breakthroughs, market demand and industry chain synergy.

1. National policy support

China's national policies have provided all-round support for the “three new types” of information consumption through a combination of top-level design, financial incentives, scenario opening and commercial support. Intelligent terminals, robots and intelligent networked cars are accelerating towards scale and industrialization under the dual drive of technological breakthroughs and market demand. With continued policy support and steady growth in market size, China's

intelligent emerging business is about to enter a period of rapid development.

(1) Top-level design and strategic planning

- **National strategic positioning**: The Chinese government's work report in 2025 explicitly lists smart grid-connected new energy vehicles, artificially intelligent cell phones and computers, and intelligent robots as new-generation intelligent terminals, and positions them as the "three new things" of information consumption, aiming at releasing the potential of consumption and fostering new quality of productivity.

- **Chinese Local policy support**: Local governments of Shenzhen, Chongqing and other places have issued special documents to support the research and development of embodied intelligent robots, for example, the "Shenzhen Action Plan for Technological Innovation and Industry Development of Embodied Intelligent Robotics (2025-2027)", issued in 2025, mentioned that the focus on supporting embodied intelligent robotic core components, AI chips, bionic manipulator and other key core technology research and development.

(2) Application scenarios and commercial support

- **Scenario-based trialing and commercialization support**: Many Chinese local governments promote the trialing and demonstration of the application of the three-types of new intelligent services in logistics, industry, transportation, and other scenarios.

- **Government procurement and benchmarking demonstration**: Many Chinese local governments have included intelligent robots and intelligent terminals in their government procurement catalogs, for example, Hefei plans to promote intelligent robotic solutions in education, healthcare, and other fields.

2. Industry development drivers

(1) Technology integration and innovation breakthroughs: The deep

integration of 5G networks, AI big models, cloud computing, and other multi-disciplinary technologies promotes the upgrading of intelligent devices in the direction of low-power consumption and high computing power, such as end-side big models significantly improving the natural interaction capability of cell phones and wearable devices. Intelligent connected vehicles rely on the development of vehicle-grade chips and the integration of vehicle-road-cloud architecture to realize the leap from single-vehicle intelligence to full-domain collaboration. Intelligent robots benefit from the technology iteration of the "Perception-Decision-Control" chain, combined with multi-modal AI breakthroughs in the ability to adapt to complex scenes. Moreover, the collaborative innovation of the upstream and downstream of the industry chain accelerates the commercialization of technology landing, while the cloud collaborative architecture can further reduce the computing threshold, giving rise to the scale application of lightweight devices and high-precision services, forming a technology-driven multiplier effect.

(2) Market demand and consumption upgrading: The first is aging and intelligent demand. Under the aging trend, the demand for home robots and intelligent health devices is surging. In addition, the market expansion of intelligent connected vehicles is accelerated due to the increase in new energy penetration rate. Secondly, the upgrading of consumer electronics is promoting the device market iteration. Cell phones are transforming into AI phones, and wearables are extending to more scenarios such as health monitoring.

(3) Industry chain synergy and ecological construction: The first is the transformation of the role of operators. From "pipeline provider" to "ecological builder", operators are gradually building the industrial ecosystem through the pan-alliance integration of terminal vendors and chip companies. The second is the effect of the regional industry clusters. Local governments are actively laying out industrial cluster demonstration areas, such as Shenzhen to create a "4+3" intelligent robotics industry layout, Hangzhou Qiantang District and Ningbo Qianwan New District form the Hangzhou Bay Intelligent Connected Vehicle Cluster, reducing the

cost of enterprise collaboration.

2.2. Typical Scenarios and Requirements of Intelligent Robots

2.2.1. Typical Scenarios

Through the in-depth integration of the physical body and intelligent decision-making, embodied intelligent robots are accelerating their penetration into multiple scenarios of consumers and industry scenarios, and reconfiguring the way of human life and production.

1. Scenarios for consumers

(1) Life housekeeper: Humanoid robots with dexterous hands and upright walking ability can complete clothing folding, cleaning, cooking, carrying and delivering household chores, etc. In addition, bionic interactive robots can provide emotional support through facial expressions and voice interaction, which can be used for health management and emotional comfort in retirement communities.

(2) Intelligent toys: The embodied intelligent robot is deeply integrated into the field of intelligent toys, reconfiguring children's growth experience through programming education, emotional accompaniment, and robot performance and entertainment, and accelerating the integration of the IP economy (such as the materialization of the secondary characters) and the upgrading of the emotional computing, to promote toys from “short-term entertainment” to “long-term companionship”, becoming a new carrier of family life and education.

2. Scenarios for industries

(1) Industry manufacturing: In automotive production, electronics manufacturing, and other industry fields, embodied intelligent robots can replace manual labor to carry out sorting and handling, safety inspection, precision operation, quality control, etc., in the production line, and are suitable for complex, dangerous industry environment to carry heavy objects, manufacturing, and other tasks, to enhance production efficiency and safety.

(2) Transmission and distribution for logistics: Through the integration of LiDAR

and deep vision, the embodied intelligent robot can autonomously plan obstacle avoidance paths, realize intelligent sorting, handling, and transportation of goods, and improve the efficiency of warehousing.

(3) Park inspection: The embodied intelligent robot can realize all-weather autonomous inspection through multi-modal perception (LiDAR, infrared thermal imaging, etc.), edge computing, etc., to reduce artificial risks. In addition, it can be combined with self-driving vehicles to expand the robot's scope of action, with dynamic obstacle avoidance and real-time monitoring capabilities, to become the core node of the intelligent park operation and maintenance.

(4) Medical surgery: Surgical robots integrate medical models and high-precision robotic arms, intelligently generate surgical navigation recommendations by analyzing surgical impact data to promote minimally invasive and precise operation innovation, and in the future will be combined with AI to realize multi-disciplinary collaborative decision-making and accelerate the construction of intelligent medical ecosystem.

(5) Emergency rescue: The embodied intelligent robots can break through the physical limitations, including temperature difference, poisonous gas, and other extreme environmental limitations, and replace the manual operation in community patrol, detonation, reconnaissance, data collection, etc., in dangerous scenarios such as firefighting, nuclear power plant inspection, chemical combustion detection, etc. Meanwhile, it can be applied to earthquake rubble search and rescue, fire rescue, and other critical scenes, shorten the golden rescue time, and enhance the safety of high-risk tasks.

(6) Welcome guide: The robot optimizes the service experience through multi-modal interaction (voice recognition, emotional recognition), providing users with problem consultation, route guidance, merchandising, and other services, reshaping the efficiency of public services and interaction temperature.

2.2.2. The Workflow

The workflow of embodied intelligence can be divided into four parts, **environment perception**, **decision analysis**, **motion generation**, and **motion execution**. The workflow achieves learning and adjustment optimization through interaction with the environment to complete task planning and command execution.

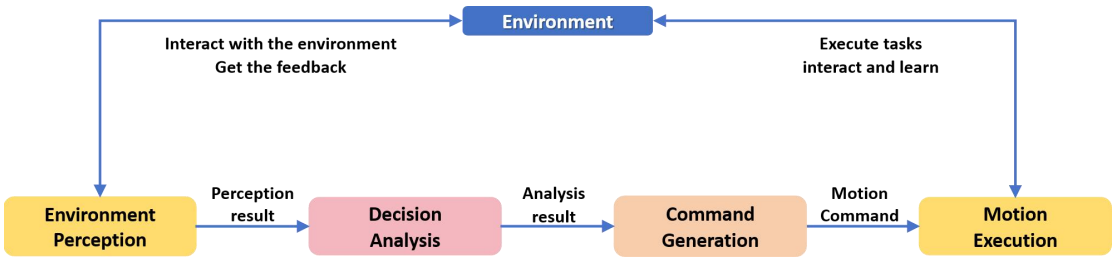

Figure 2 The workflow of embodied intelligent robots

Among them:

1. **Environment perception**: mainly performed by the robot body (such as limbs, hands, feet, skin, etc.), including object and scene perception in the external environment, as well as behavioral and expression perception of the user. The perception results are used for data acquisition and inputted for subsequent decision analysis.

2. **Decision analysis**: mainly performed by the robot brain, analyzing data and decision-making results for three categories, task planning (combining atomic skills to generate motion commands), environment understanding and analysis (analyzing the real environment information in route planning/navigation scenarios), and intelligent Q&A (human-robot Q&A interaction).

3. **Motion generation**: mainly executed by the robot cerebellum, generating motion commands based on the decision-making results, including commands based on fixed rules (dancing choreography), commands based on interactive learning (adjusting hand posture to complete the item grasping), and commands based on the

big model (question answering).

4. **Motion execution**: mainly completed by the robot body, including the motion or Q&A according to the commands, and interactions with the environment and the user for feedback before the next round of process.

As the robot brain and the robot cerebellum of the embodied intelligent robot need high computing power support in decision analysis and command generation, to minimize the complexity and cost of robot hardware deployment and enhance the flexibility of the robot itself, the capability improvement of the robot brain and the robot cerebellum will be the future development trend when they are detached from the robot body and deployed remotely. Based on the hierarchical decision-making model, the current deployment method offers two options: end-cloud collaboration and cloud-edge-end collaboration. The end-cloud collaboration is the current mainstream deployment option. In the future, for multi-robot collaboration scenarios, the cloud-edge-end collaboration deployment option will become one of the possible evolution directions. By deploying the motion model of the robot cerebellum in the edge nodes (e.g., base stations, edge servers, etc.) and making full use of their computing power resources, the communication and work efficiency of embodied intelligent robots can be effectively improved.

Table 2 The comparison between end-cloud collaboration and cloud-edge-end collaboration

	End-cloud collaboration	**Cloud-edge-end collaboration**
Cloud	Robot Brain: human-robot interaction, intent understanding, task planning, etc.	Robot Brain: human-robot interaction, intent understanding, task planning, etc.
Edge	/	Robot Cerebellum: generation of motion commands and trajectory
End	● Robot Cerebellum: generation of motion commands and trajectory ● Robot Body: motion execution	Robot Body: motion execution
Pros	● Suitable for single-robot scenarios	● Suitable for multi-robot networking

	● Fast deployment and scenario generalization	scenarios ● The motion model inference of multiple robots can use edge computing resources at the same time, effectively reducing the power consumption of the robots themselves
Cons	Demanding High Manufacturing Precision and Stringent Cost Requirements for Robot Bodies	Multi-robot tasks place high demands on network connectivity and computing power
Industry supporting	Main option for the industry currently	/

2.2.3. Service Requirements

Typical services of embodied intelligence include real-time interaction, designed motion, and remote operation, in which the robot obtains recognizable multi-modal commands to complete the execution of motions and interactions. Designed motion does not require network transmission of data, whereas real-time interaction and remote operation require the cloud server to parse the commands, and therefore require mobile communication networks to transmit the commands and data.

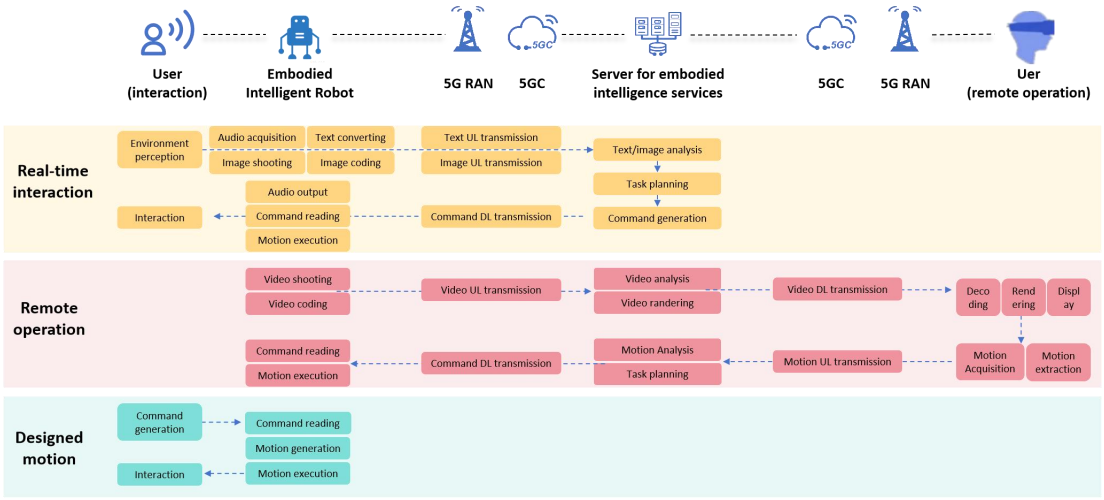

Figure 3 The service flow of embodied intelligent robots

1. Requirements for network connection

Embodied intelligent robots can be applied to different scenarios of ToC and ToB, from the daily life of the public to industry production. According to the analysis of service flow (shown in Figure 3), the downlink transmission data of embodied intelligent robots are mostly motion commands, while the uplink transmission data are slightly different according to different types of demands. Real-time interaction, as a rigid demand of embodied intelligent robot services and requesting high network performance (rate, delay, etc.), includes voice Q&A, environment analysis, etc. Voice Q&A involves audio acquisition, text conversion, text transmission, audio parsing, etc. Environment analysis refers to human-robot interaction Q&A after recognizing the physical world, objects, etc. In addition to voice Q&A-related tasks, it involves the tasks of capturing, encoding, transmitting, and parsing images or videos. The remote operation requires the transmission, decoding, and rendering of images captured by remote headsets or cameras, and therefore requires mobility and reliability in addition to rate and latency.

This white paper focuses on analyzing the latency and rate requirements of the network for embodied intelligence robots. Embodied intelligence end-to-end delay consists of the segmented delay of each part. Currently, the response delay of a single task is 1~5 seconds, of which the large model processing delay accounts for more than 50%.

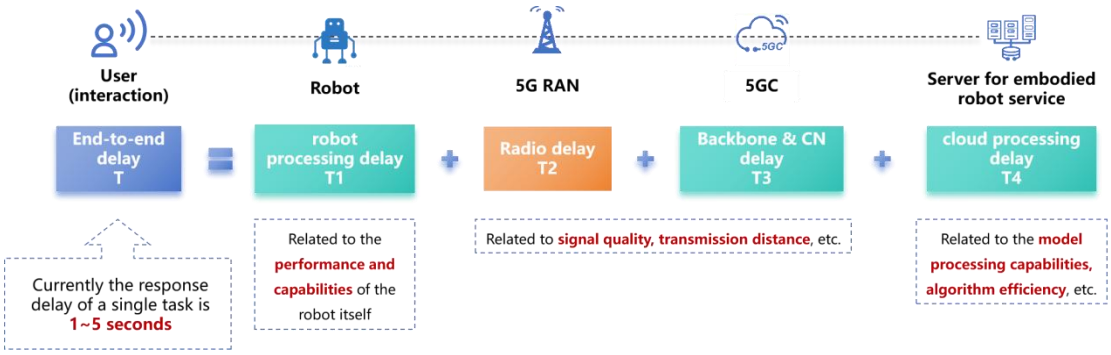

Figure 4 The end-to-end delay analysis of the embodied intelligent robots

In the laboratory test environment, for the multi-modal Q&A, the test results

are as follows, which are basically in line with the theoretical analysis.

(1) Rate test

- Service model: transferring 3 images (720P, 400KB) per second
- Test results: as Table 3

Table 3 The network test result of the embodied intelligent robot application

Round-trip delay	Estimation of radio delay	UL frame-level rate
2s (feeling good)	UL：1s DL：180ms	3.25Mbps
1.2s (felling normal)	UL：350ms DL：30ms	9.4Mbps

Note: In the real test, it is proved that The delay of a single picture analysis of the large model is about 800ms, and Transmission delay of backbone network in laboratory environment is about 20ms.

(2) Delay test

- Service model: capture and store human & robot audio, analyze the intervals between two audio segments and derive the round-trip response delay
- Test result: the Q&A delay of the voice interaction prototype is 2.15 seconds (acceptable good response time).

2. Requirements for computing power

(1) Home service robots: The home environment has random obstacles (such as pets, toys), light changes, etc., and the robots need to adapt to the flexible but have low real-time requirements (hundred milliseconds).

(2) Industrial robots: Industrial scenarios (e.g. automotive welding, precise assembly) require closed-loop control of robot motions and sensor feedback to be

completed within microseconds (μs), with stringent real-time requirements.

The autonomous action and cooperative operation of intelligent robots in complex manufacturing scenarios rely on high-intensity AI computation, including environment perception (e.g., target recognition, obstacle detection, three-dimensional reconstruction), dynamic planning (e.g., path planning, task assignment), and intelligent decision-making (e.g., real-time strategy adjustment, motion prediction). Due to the power consumption, computing power, and weight constraints of the robot body, it is difficult for the robot itself to independently undertake high-intensity computational tasks. Therefore, the embodied intelligent robots put forward a clear demand for offloading computing and AI processing to the network side:

(1) Edge-side AI arithmetic requirements: Offloading the robot brain functions to the edge side requires that the edge nodes have strong AI computing capability for real-time processing of multi-channel high-resolution video streams, LiDAR point cloud data, and completing the complex AI model inference tasks.

(2) AI fast inference: The edge needs to deploy a high-performance AI inference engine to support large-scale AI model real-time inference. The delay needs to be controlled at the millisecond level to meet the needs of the robot's real-time decision-making.

(3) Dynamic update and training of AI models: The changes in the industrial production environment and the diversity of tasks require that the AI model can be updated quickly and iteratively according to the field data, and the edge side needs to have online learning and model updating capabilities to quickly respond to the environmental changes and task requirements.

The network can play the advantage of being close to users and perceiving users, and make full use of the computing advantage of network edge nodes through the computing power sharing of the cloud-edge-end collaboration to provide low-latency and high-reliability computing services.

3. Requirements for data

The training data collection and storage of embodied intelligence large models is difficult, for example, the multi-modal dataset of Google's intelligent robots (RT1, RT2) needs to collect millions of data from 160,000 tasks in 22 robots, so there is a need for data collection, storage and sharing for the network.

(1) Data collection and aggregation: The edge network needs to receive sensor data uploaded by multiple robots in real time and quickly perform preprocessing and integration.

(2) Data preprocessing and compression: Edge nodes need to support data preprocessing, feature extraction, data compression, etc., to reduce the burden of transmission and storage, and improve the efficiency of subsequent AI calculations.

(3) Data storage and synchronization: The edge side needs to have certain data storage and synchronization mechanisms to ensure the consistency and synchronization of the data when multiple robots collaborate, and to avoid collaborative errors caused by data delay or desynchronization.

The network can provide data collection, storage, and sharing, and build intra-network data management to support data sharing among multiple embodied intelligent robots, realizing real-time analysis and processing of data in the vicinity, and reducing data transmission and processing overhead.

2.3. Typical Scenarios and Requirements of AI Agent

2.3.1. Introduction of AI Agent

An AI agent is a software entity or system that is capable of autonomously perceiving, making decisions, and executing actions in order to achieve specific goals. It has the capability to gradually accomplish given goals through independent thinking and tool invocation.

The collaboration between humans and AI can be divided into three modes. AI

embedding mode, AI copilot mode, and AI Agent mode. Compared to the first two modes, the AI Agent mode is more efficient and will be the main mode of collaboration between humans and AI in the future.

1. **AI Embedding Mode**: Users communicate with AI through language, use prompts to set goals, and then AI assists users in achieving these goals. The role of AI is a tool for executing commands, while humans play the roles of decision makers and commanders.

2. **AI Copilot mode**: Humans and AI participate in the workflow together, each playing their own roles and complementing each other's abilities. AI intervenes in the workflow, from providing suggestions to assisting. The role of AI is a knowledgeable partner.

3. **AI Agent Mode**: Humans set goals and provide necessary resources (such as computing power), then AI independently undertakes most of the work, and finally humans supervise the process and evaluate the final results. In this mode, AI fully embodies the interactive, autonomous, and adaptive characteristics of intelligent agents like independent actors, while humans play more roles as supervisors and evaluators.

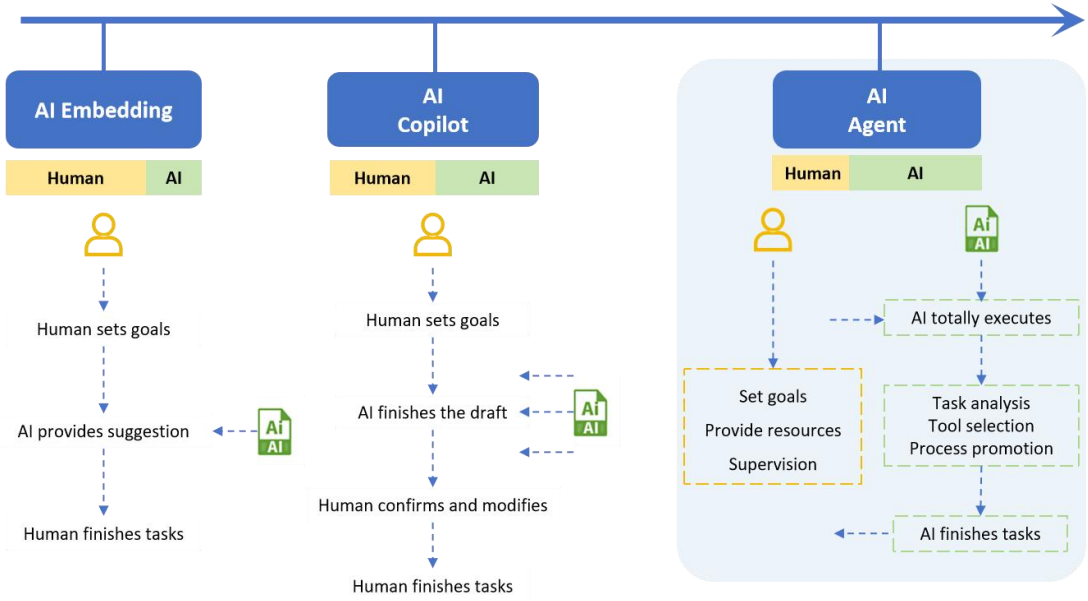

Figure 5 The development of AI Agent

Currently, multi-modal dialogue applications are a companion feature of AI native applications launched by OTT. Mainstream AI native applications already support multi-modal dialogue, such as Alibaba's Tongyi, Zhipu AI's ChatGLM, Doubao, etc., which have all launched vision-based multi-modal dialogue. Users can capture images on their mobile phone cameras and have real-time conversations with AI. Nevertheless, currently, the resolution of applications is still relatively low, basically maintained at 720P level.

Table 4 AI native applications supporting multi-modal real-time dialogue

	Doubao	**Zhipu AI**	**iFLYTEK SPARK**	**Alibaba Tongyi**
Users	118 Million	8 Million	8 Million	5 Million
Resolution	720P	720P	720P	720P

2.3.2. Typical Scenarios

The combination of AI Agent with different device forms can be widely applied in personal consumption and industry production, improving user experience and production efficiency. The typical application scenarios of AI agents include:

1. **Mobile AI Assistant**: AI mobile phones have reached the stage of AI primitive biology, with personal AI assistants as the core. This is the first typical application of AI Agent landing. After the user issues a voice command, the mobile AI assistant can complete tasks such as e-commerce shopping, ordering takeout, booking train tickets, and interacting with friends on social media.

2. **Embodied Intelligence**: AI Agent can interact with the environment through physical entities to perceive the environment, recognize information, make autonomous decisions, and take action. It is applied in scenarios such as life

management, industrial manufacturing, park inspections, etc.

3. **Intelligent customer service**: AI Agent technology can automatically handle high-frequency scenarios such as bank wealth management consultation and e-commerce after-sales disputes based on multi-round dialogue management and intention recognition, improving service efficiency and enhancing user satisfaction.

4. **Driving assistant**: AI Agent can combine real-time road condition analysis and driving behavior monitoring to achieve functions such as dangerous lane change warnings and dynamic optimization of freight routes, reducing the incidence of traffic accidents.

5. **Financial assistant**: AI Agent can use big data correlation analysis and dynamic risk modeling to complete high-value decisions such as intelligent stock portfolio adjustment and credit anti-fraud detection, optimizing asset allocation returns.

6. **Marketing Assistant**: AI Agent can rely on user profile mining and generative AI to automatically generate social media hot copy and predict advertising effectiveness, improving the return on investment of marketing.

7. **Medical Assistant**: AI Agent can integrate cross-modal medical data and diagnosis and treatment knowledge graphs, assist in CT imaging lesion localization, and generate personalized medication plans, to improve clinical diagnosis efficiency and accuracy.

2.3.3. The Workflow

The workflow of AI Agent can be divided into three steps: perception and input, inference and decision-making, and execution and feedback. Through the use of large models and memory data for decision inference, execution plans are generated and stored for the next decision execution.

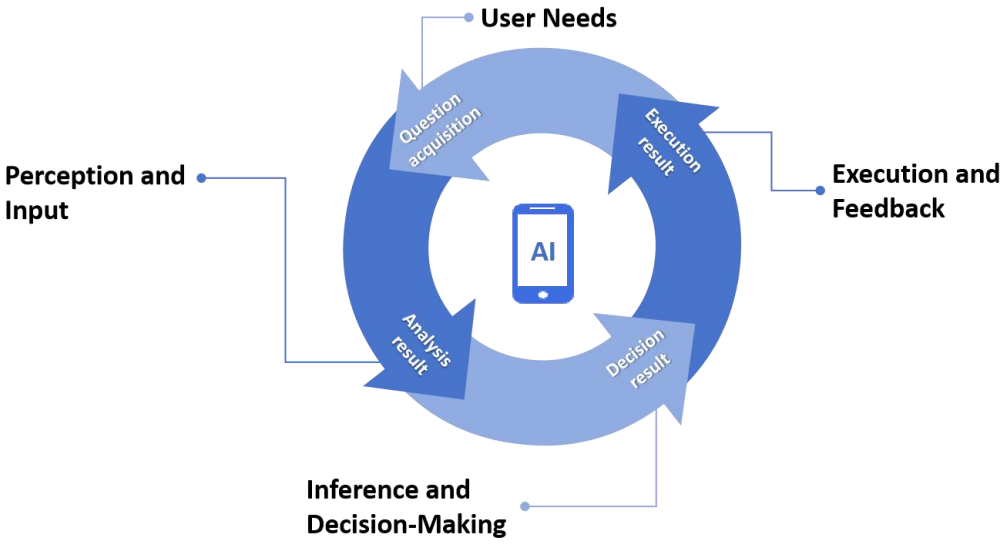

Figure 6 The workflow of AI Agent

1. Perception and Input

(1) Intention understanding: With the help of large model, accurately understand user's questions and intentions.

(2) Problem analysis: Task decomposition, breaking down the question into several sub questions.

2. Inference and decision-making

(1) Parameter matching: Utilizing cloud and local memory to learn and summarize problem-solving experience before.

(2) Decision planning: Determining the solutions and steps and generating the order of execution tool calling according to the solutions.

3. Execution and Feedback

(1) Execution call: Calling various tool components (such as third-party apps, native functions on mobile phones, etc.) to execute.

(2) Memory storage: Observing user's feedback, adjusting execution in the next interaction, and storing the execution results for subsequent reasoning and

decision-making.

The implementation of AI Agent mainly includes three solutions: device AI, cloud AI, and hybrid AI.

1. **Device AI**: Both data processing and model inference are completed at the device, with fast response speed, high privacy, low network requirements, but weak capability to handle complex tasks and low level of intelligence.

(1) Advantages:

- **Low latency response**: No need to upload data to the cloud and wait for processing results, providing instant feedback.
- **High privacy**: All data processing is carried out locally.
- **Offline availability**: Even if the phone is in a network-free state, the AI assistant can still work normally.

(2) Disadvantages:

- **Limited computing resources:** The hardware performance of smartphones is relatively limited, making it difficult to run complex and large-scale AI models.
- **Difficulty in updating models**: Due to limitations in mobile phone storage space and processing power, updating local models requires a significant amount of time and traffic.

2. **Cloud AI**: All computing and data storage are placed in the cloud, utilizing the powerful computing resources of the cloud to handle complex tasks with high network requirements and privacy and security risks.

(1) Advantages:

- **Powerful computing power**: Capable of running complex large-scale AI models and handling various complex tasks.

- **Convenient model updates**: The cloud can update and optimize AI models at any time.

- **Rich knowledge resources**: The cloud can integrate a large amount of knowledge and data, providing AI Agent with more comprehensive and accurate information.

(2) Disadvantages:

- **High latency**: Data transmission requires a certain amount of time, especially in unstable networks or weak signals, where latency becomes longer.

- **Privacy risk**: Users' voice and interaction data need to be uploaded to the cloud, which poses a certain risk of privacy leakage;

- **High cost**: Long term use of cloud services is expensive.

3. Hybrid AI solution (Device AI + Cloud AI): Simple tasks are executed on the device side, while complex tasks are uploaded to the edge/cloud, balancing performance, privacy, and network dependencies.

(1) Advantages:

- **Balance performance and privacy**: Utilize the low latency and privacy protection advantages of Device AI to handle local tasks, and leverage the computing power of Cloud AI to handle complex tasks.

- **Improve resource utilization**: Flexibly allocate tasks based on the characteristics of tasks and the resource status of devices.

- **Enhanced reliability**: In the event of network instability or interruption, Device AI can continue to complete some basic tasks.

(2) Disadvantages:

- **High system complexity**: Requiring coordination of task allocation and data

exchange between the device side and the cloud side, resulting in high complexity.

- **High development difficulty**: Development requires consideration of both the device side and the cloud side technologies and environments, resulting in relatively high development cycles and costs.

Hybrid AI solution combines the advantages of both Device AI and Cloud AI, which not only protects privacy and has low latency, but also has powerful computing power to support complex tasks, making it the mainstream solution for personal AI assistants today.

Table 5 The comparison of AI Agent solutions

Solution	Pros	Cons	Application Scenarios	Computing Requirements
Device AI	Low latency and high privacy	Limited computing capability	Face recognition, voice assistant	Device: 1~10 billion model parameters
Cloud AI	Strong computing capability and rich data resources	Relying on network, existing latency problem	Search engine	Server: >100 billion model parameters
Hybird AI	Low latency and high computing capabilities	Complex task allocation	Smart voice assistant	Device (1~10 billion model parameters) +Server (>100 billion model parameters)

2.3.4. Service Requirements

1. The smartphone assistant service based on the device AI solution

The smartphone assistant is based on the device AI solution that obtains voice commands through the phone microphone. The smartphone then performs text conversion, intent understanding, planning and decision-making, and generates commands that are subsequently executed.

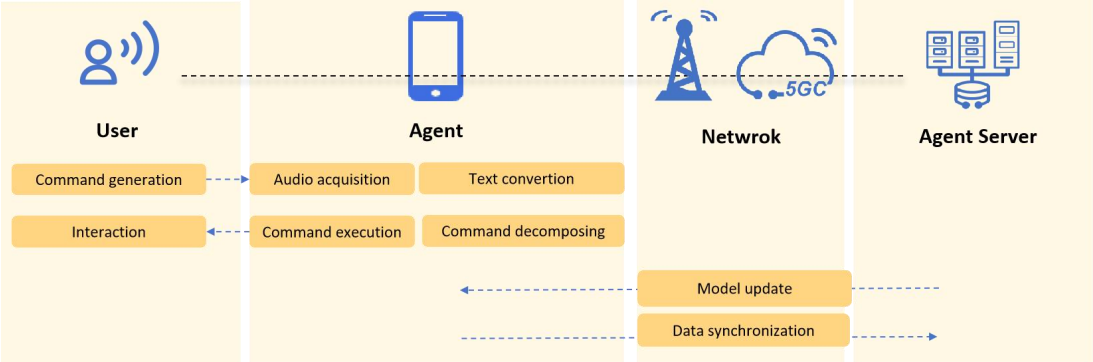

Figure 7 The service flow of smartphone assistant based on the device AI

The service has low network requirements; mainly, the timed model updates and data synchronization will request the network. The downlink rate for model update is approximately 1Gbps and the uplink rate for data synchronization is about 100Mbps.

2. The smartphone assistant service based on the cloud AI solution

The smartphone assistant service based on the cloud AI solution involves obtaining voice commands through the cell phone's microphone, decomposing the commands at the device side, and uploading them to the cloud server via the mobile communication network for data and command parsing, task planning and execution. The server then generates downlink response information and sends it back to the smartphone to display the results for the user to supervise.

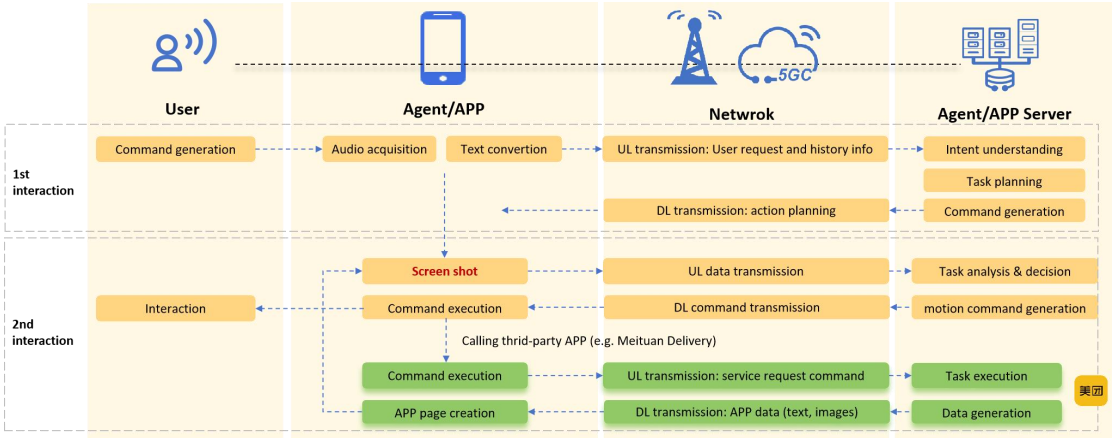

Figure 8 The service flow of smartphone assistant based on the cloud AI (Take

take-out delivery ordering as an example)

This service places high demands on the network. Its main service data is uplink-transferred, including user requests, user profiles, screenshots, service request commands, text frames, etc.

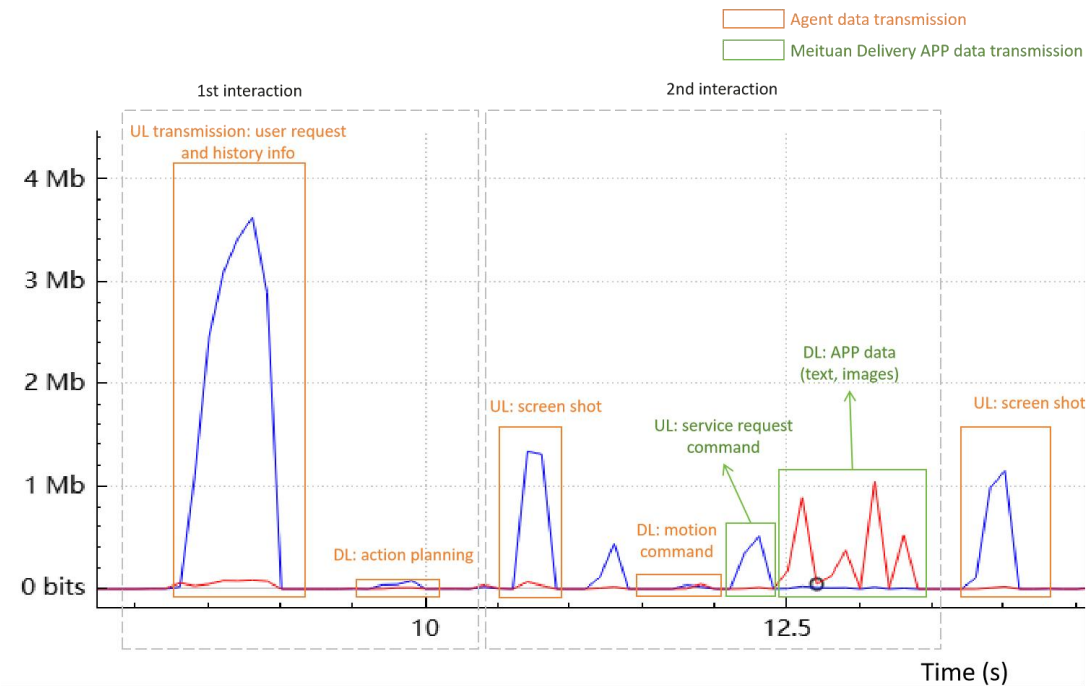

Figure 9 The test results of smartphone assistant based on the cloud AI

The test results of the AI assistant service based on the cloud AI solution shows that for the takeout ordering service, the uplink rate is up to 10.9Mbps (one interaction) and 4.4Mbps (consecutive interactions), and the downlink rate is up to 3.1Mbps. During the test, to finish one takeout ordering task, the system needs 15 interactions for average and the total time for one task is 60s for average, hence the latency for single interaction is about 4s. In the future, with the improvement of cloud processing capacity and the reduction of processing delay, the single interaction delay will drop to 2~3 seconds.

Table 6 Test results of the smartphone assistant service based on the cloud AI solution

Interaction type	Data type	UL Rate	DL Rate	Data Volume
Agent Data Interaction	User requests and user history information	10.9Mbps	/	3.5Mb
	Motion plans/command	/	0.2Mbps	70kb
	Screenshots	4.4Mbps	/	1.4Mb
Service App Data Interaction	Commands for service requests	1.6Mbps	/	500kb
	Text frames + images	/	1.6-3.1Mbps	500kb-1Mb

Note: Radio delay is calculated at 320ms (for the smooth interaction)

3. The smartphone assistant service based on the hybrid AI solution

The smartphone assistant based on the hybrid AI solution obtains the user's voice commands through the cell phone microphone. The device agent first pre-processes the commands and converts them to text. The text is then uploaded to the agent server via the mobile communication network for intent understanding, task planning, and command generation, and commands are sent back in the downlink transmission. The device agent then calls a third-party APP (e.g. Meituan Delivery) to execute the commands. The pages and data of the third-party APP are generated according to commands, and then the device agent processes the screenshot of the APP page and uploads the processed data to the agent server via the mobile communication network for task analysis and decision making. The server then generates commands to be sent back to the device-side agent in the downlink transmission.

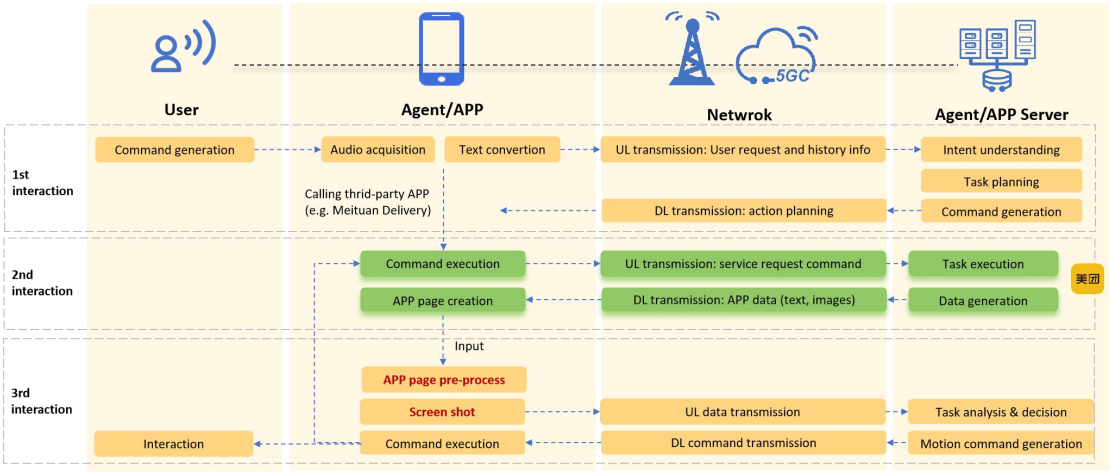

Figure 10 The service flow of smartphone assistant based on the hybrid AI (Take take-out delivery ordering as an example)

This service places high demands on the network. Its main service data is uplink-transferred, including screenshots, service request commands, text frames, etc.

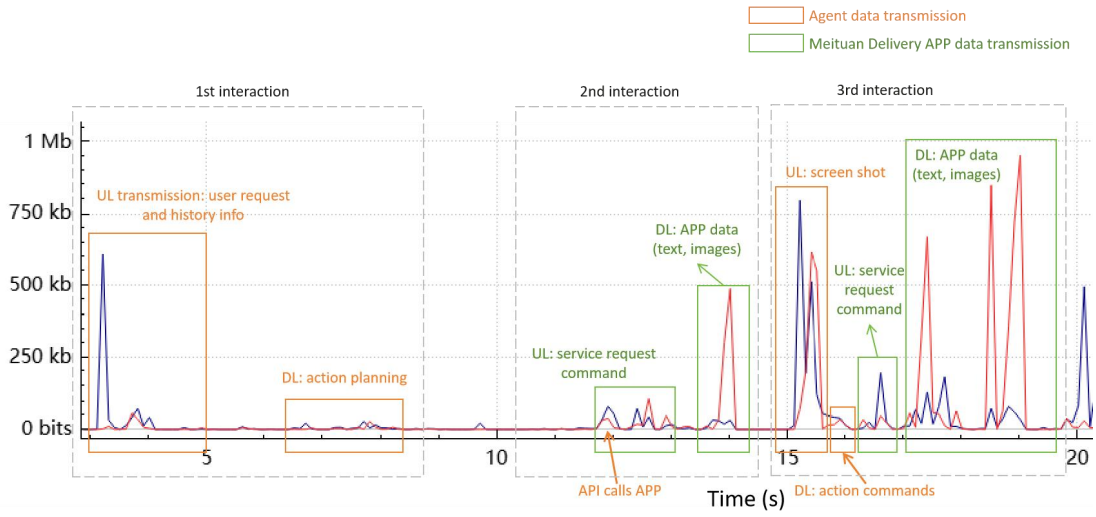

Figure 11 The test results of smartphone assistant based on the hybrid AI

The test results of the AI assistant service based on the cloud AI solution shows the for the takeout ordering service, the uplink rate is up to 2.5Mbps and the downlink rate is up to 3.1Mbps. During the test, to finish one takeout ordering task,

the system needs 10 interactions for average and the total time for one task is 22s for average, hence the latency for single interaction is about 2.2s. In the future, with the improvement of the processing capability of the cloud and the reduction of the processing latency, the single interaction latency will be reduced to 1~1.5 seconds.

Table 7 Test results of the smartphone assistant service based on the hybrid AI solution

Interaction type	Data type	UL Rate	DL Rate	Data volume
Agent Data Interaction	User requests and user history information	10.9Mbps	/	3.5Mb
	Motion plans/command	/	0.2Mbps	70kb
	Screenshots	4.4Mbps	/	1.4Mb
Service App Data Interaction	Commands for service requests	1.6Mbps	/	500kb
	Text frames + images	/	1.6-3.1Mbps	500kb-1Mb

Note: Radio delay is calculated at 320ms (for the smooth interaction)

2.4. Typical Scenarios and Requirements of AI Glasses

2.4.1. Typical Scenarios

AI glasses are the product form of traditional glasses iterated to AR/MR glasses, and they are lightweight wearable devices integrating AI and intelligent hardware. The core function is to combine AI technology with glasses form, and through the core modules such as cameras, sensors, and LLM, AI glasses can realize the functions of environment perception, intention analysis, AI and large language models, and superposition of virtual and real information, and provide users with voice assistants, real-time translations, navigation reminders, etc., becoming an interactive portal for users to connect to the digital world.

The application scenarios of AI glasses are expanding to penetrate into many specialized fields such as medical treatment, industry, education, etc. The application scenarios of AI glasses include:

1. **Consumer market**: AI glasses can be used for environment and picture

recognition, and the real-time translation of personal life and entertainment activities to enhance user experience.

2. **Industry applications**: AI glasses can provide real-time information and guidance in equipment maintenance, production management, and quality inspection.

3. **Education and training**: AI glasses can be used for online education and skills training to provide an immersive learning experience.

4. **Medical field**: In telemedicine and surgical assistance, AI glasses can help doctors access real-time data and guidance.

5. **Augmented reality meeting**: Through AI and AR technology, users can interact in a virtual environment to enhance the sense of participation in the meeting.

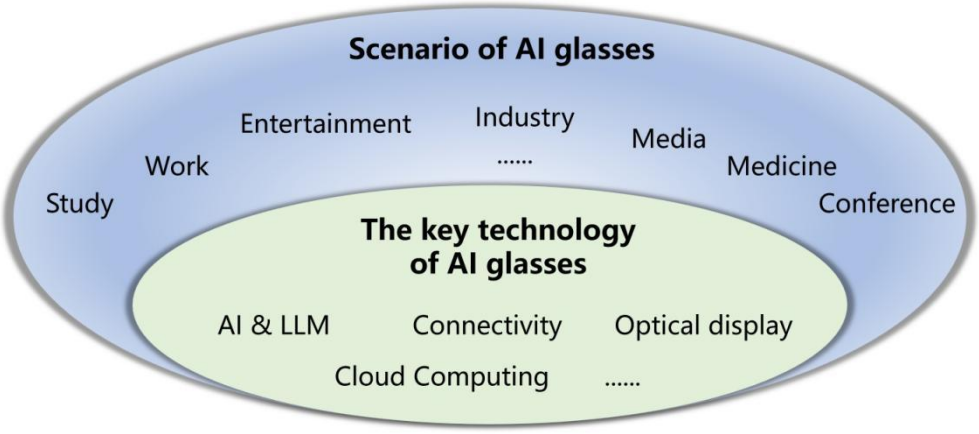

Figure 12 The panoramic view of AI glasses applications

These application scenarios demonstrate the potential and value of AR glasses in multiple fields, and the market growth of AI glasses provides strong support.

2.4.2. The Workflow

The workflow of AI glasses can be divided into four steps, information acquisition, information recognition and command inference, decision-making and

execution, and feedback and presentation. Through the LLM and information-based data decision inference, AI glasses generate response information and present feedback to the user.

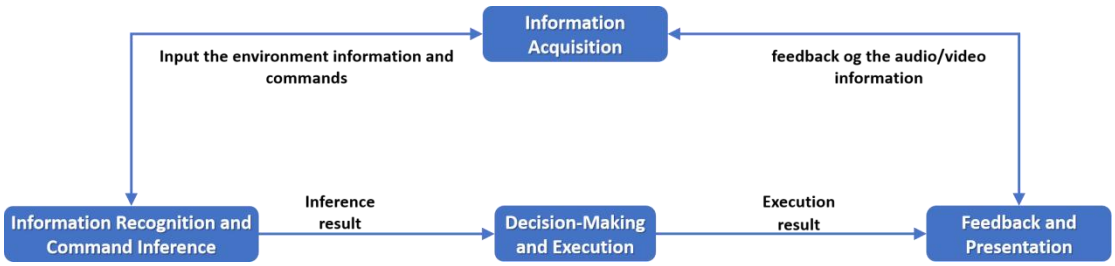

Figure 13 The workflow of AI glasses

1. Information Recognition and Command Inference

(1) Information recognition: Utilizing AI capabilities to parse the environment and commands to accurately extract key command information.

(2) Command inference: Utilizing AI big model and knowledge to understand the user's intent.

2. Decision-making and execution

(1) Decision-making: Making decisions based on the intent, planning the task flow, and decomposing the required steps.

(2) Task execution: Executing task steps, calling various functional components, generating task results, and making decision adjustments based on the results.

3. Feedback and presentation: Judging the results and generating the corresponding audio and video information, and presenting the information and results on the user's device.

2.4.3. Service Requirements

1. AI glasses service flow

The camera and microphone of glasses body are activated to obtain image and

voice command by voice commands, and the data information is uploaded to the cloud server through mobile networks for content parsing, intent understanding, and planning and decision-making. The downlink response information is generated to be sent back to the side of glasses for voice playback.

From the point of view of the data content of interaction between AI glasses and the server, the uplink data mainly is voice commands and image snapshot (for photo Q&A)/video (for video Q&A), while the downlink data mainly is the voice response data. In the future, AI glasses in the downlink direction may also include other data information such as AR-augmented image.

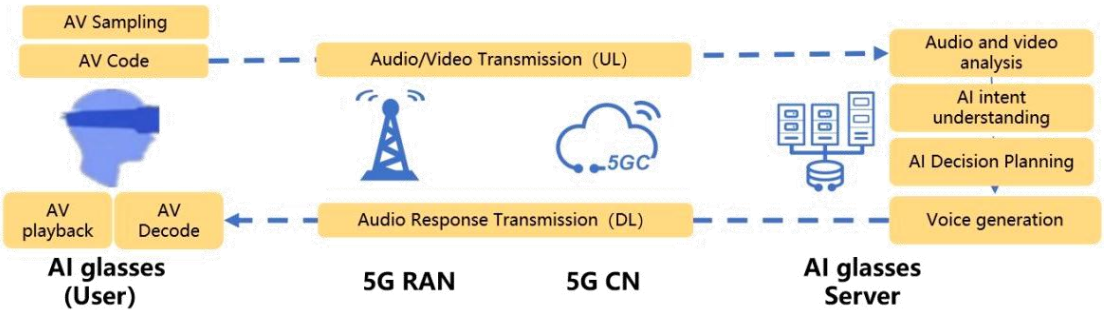

Figure 14 The service flow of AI glasses

2. Network requirements

Table 8 Estimated statistics of data transmission in different service of AI glasses

Service	Data Content	UL Traffic Volume	Rate	Protocols
Image Q&A	Voice command + Image (1 snapshot)	~100KB (The duration of a single voice question: ~1.5s)	545Kbps (ave) 840Kbps (max)	TCP
Video Q&A	Voice command + Video (2fps)	~8.3MB (Continuous shooting duration:~140s)	475Kbps (ave) 1.6Mbps (max)	UDP

Note: The data is sourced from laboratory validation

(1) AI glasses image Q&A

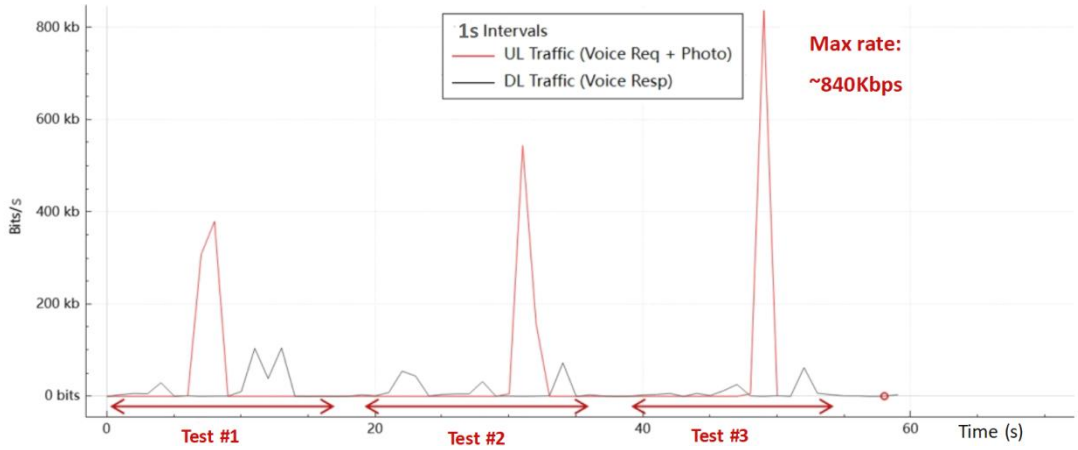

Figure 15 The test result of AI glasses image Q&A

(2) AI glasses video Q&A

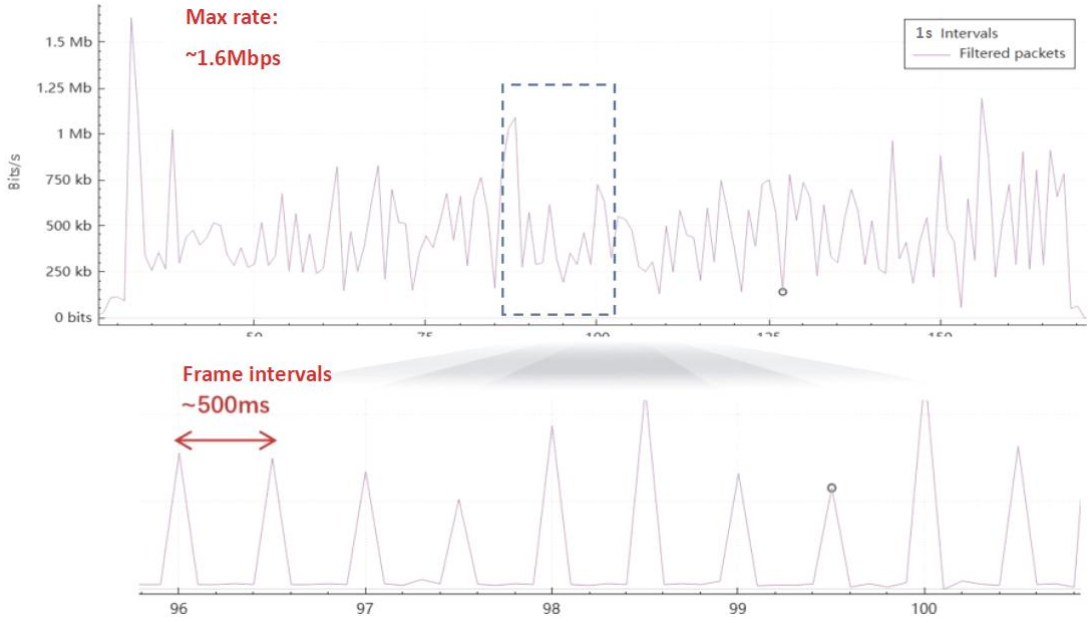

Figure 16 The test result of AI glasses video Q&A

The above data are the test results of AI glasses service in laboratory. In the future, AI smart glasses will have a richer variety of service types and a wider range of application scenarios. In order to meet the requirement of the real-time human-machine dialogue experience, AI glasses and server platforms will have more frequent interactions and higher-quality content transmission, which will also put

forward higher network bandwidth requirements for future networks.

2.5. Typical Scenarios and Requirements of Intelligent Connected Vehicles

2.5.1. Typical Scenarios

Intelligent connected vehicles are intelligent transportation tools that integrate automotive engineering, smart technologies, and connectivity capabilities. They are built upon vehicles equipped with perception, computing, and execution capabilities, centered around intelligent systems capable of understanding, analyzing, and learning, and rely on interconnected technologies enabling communication, interaction, and collaborative connectivity as their core foundation.

These vehicles find applications in toC scenarios such as intelligent cockpits, AR navigation, and assisted driving, as well as toB scenarios like autonomous taxis and unmanned logistics delivery, including:

1. **Intelligent Cockpit:** Leveraging AI-driven voice interaction and facial recognition technology to deliver personalized services and seamless multi-modal human-vehicle collaboration, dynamically optimizing the in-vehicle environment and entertainment systems.

2. **AR Navigation**: Utilizing AI-based real-time environmental perception and dynamic path planning to overlay virtual navigation information onto real-world traffic imagery, enhancing driving decision accuracy in complex road conditions.

3. **Assisted Driving**: By integrating multi-sensor data through AI, functionalities such as lane-keeping, adaptive cruise control, and emergency obstacle avoidance are achieved, with continuous optimization of driving security and scenario adaptability via deep learning.

4. **Autonomous Taxi Services**: Relying on AI algorithms for full-scenario perception and multi-objective decision-making, enabling vehicles to autonomously accept orders, avoid obstacles, and optimize routes on urban roads, providing 24/7

efficient mobility services.

5. **Unmanned Logistics Delivery**: AI-powered high-precision positioning and dynamic obstacle detection ensure delivery vehicles can navigate in complex urban environments autonomously, perform real-time obstacle avoidance, and complete the "last-mile" delivery.

6. **Campus Unmanned Transportation**: AI-coordinated scheduling systems integrate fleet control and route planning to achieve all-weather automated material transportation within industrial parks, significantly reducing labor and energy costs.

7. **Autonomous Sanitation Vehicles**: Employing AI-based visual recognition to identify garbage distribution and road boundaries. These vehicles autonomously generate efficient cleaning routes and dynamically adjust operational modes in response to weather changes, enhancing the intelligence level of urban operations.

2.5.2. The Workflow

1. The intelligent cockpit with the co-pilot AI assistant

The intelligent cockpit refers to a digital in-vehicle space centered around the in-vehicle infotainment system, integrating technologies such as LCD instrument clusters, heads-up displays, voice- and gesture-based multi-modal interaction, and network connectivity features. It delivers personalized services through multi-modal interaction (combining voice, touch, gestures, etc.). The AI co-pilot assistant, powered by AI Agent technology, enhances the intelligent cockpit experience by enabling multi-modal interaction, proactive engagement, and personalized development.

The workflow of the intelligent cockpit with the co-pilot AI assistant can be divided into three steps: perception & analysis, inference & decision-making, and execution & feedback. This closed-loop system ensures continuous adaptation to user behavior, environmental changes, and evolving service demands, driving the intelligent cockpit toward seamless, intuitive, and user-centric experiences.

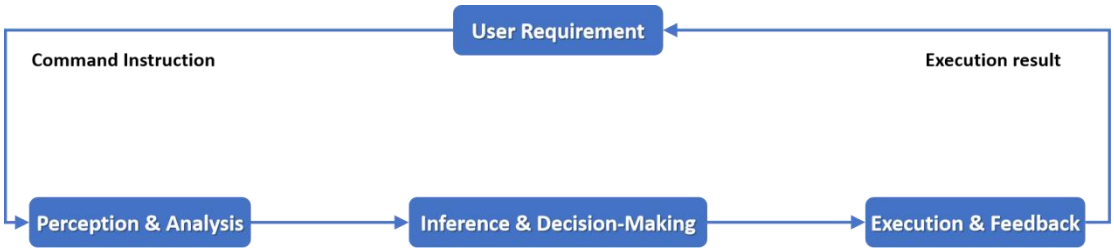

Figure 17 The workflow of the intelligent cockpit with the co-pilot AI assistant

(1) Perception and Input

- **Intention understanding**: With the help of large model, accurately understand user's questions and intentions.

- **Problem analysis:** Task decomposition, breaking down the question into several sub questions.

(2) Inference and decision-making

- **Parameter matching:** Utilizing cloud and local memory to learn and summarize problem-solving experience before.

- **Decision planning**: Determining the solutions and steps and generating the order of execution tool calling according to the solutions.

(3) Execution and Feedback

- **Execution call**: Calling various tool components (such as third-party apps, native functions on mobile phones, etc.) to execute.

- **Memory storage**: Observing user's feedback, adjusting execution in the next interaction, and storing the execution results for subsequent reasoning and decision-making.

2. Cloud-assisted intelligent driving

Currently, autonomous driving is evolving from "single-vehicle intelligence" to

"vehicle-road-cloud collaborative driving". **Single-vehicle intelligence** relies on onboard sensors to collect external environmental data and uses algorithms for autonomous vehicle control. This approach demands high-performance onboard sensors and computational platforms. **Vehicle-road-cloud collaboration**, by contrast, gathers external information from multiple endpoints (vehicle-side, road-side, and cloud-side) and leverages synergistic algorithmic processing between onboard and cloud-based systems to achieve high-level autonomous driving.

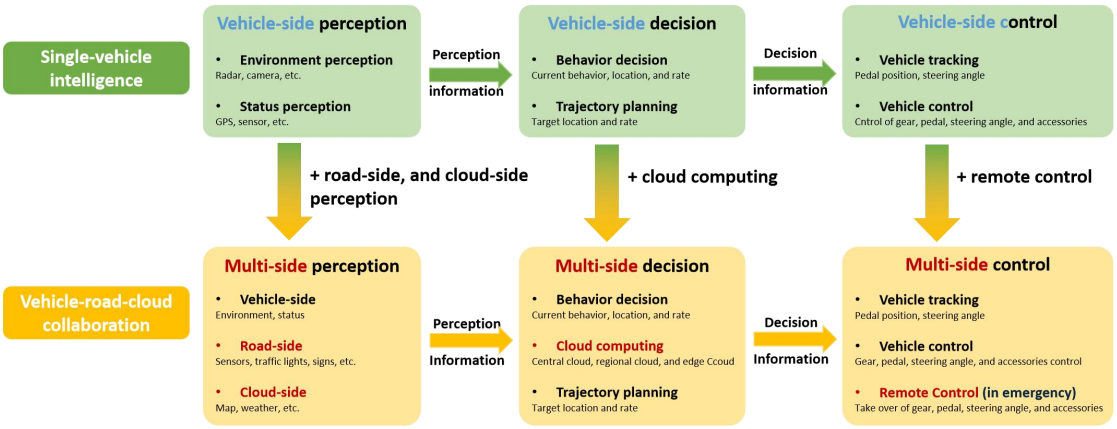

Figure 18 The development of intelligence driving

Cloud-assisted intelligent driving services refer to the real-time uploading of complex semantic traffic information (e.g., irregular traffic signs, tide lanes, or ambiguous road markings) that vehicles cannot locally interpret. A cloud-based large AI model processes and identifies this data in real-time, then instantly transmits the results back to the vehicle. This reduces traffic accidents, enables real-time optimal route selection, and enhances autonomous driving efficiency.

The workflow can be divided into four stages: environment perception, Decision Analysis, command Generation, and Command Execution. The vehicle executes commands locally (e.g., rerouting, braking, or steering adjustments) and continuously learns and optimizes through interaction with the environment (e.g., adapting to new traffic patterns). This integrated framework combines local execution with cloud intelligence, creating a robust solution for next-generation

autonomous driving.

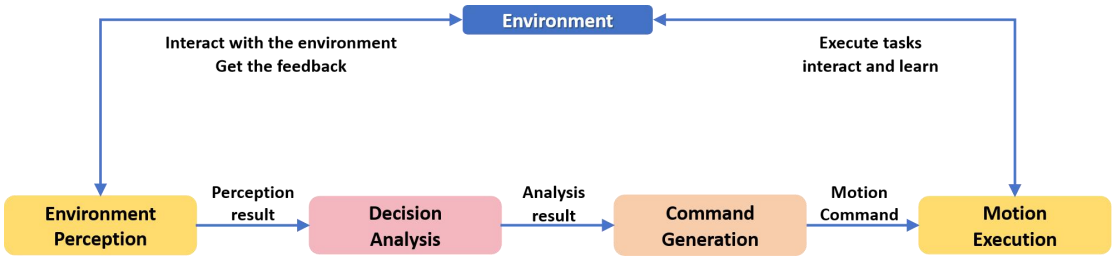

Figure 19 The workflow of the cloud-assisted intelligent driving

(1) Environment perception

- **Road**: Perception of scenario and physical world.
- **Thing**: Perception of things around vehicles and on the road.
- **Human:** Perception of human voice and facial expression.

(2) Decision analysis

- **Task Planning**: Combining capabilities to generate commands.
- **Environment analysis**: Analyzing the real environment information, such as the road, trajectory, navigation.

(3) Command generation

- Commands based on rules.
- Commands based in learning.
- Commands based on large models.

2.5.3. Service Requirement

1. The intelligent cockpit with the co-pilot AI assistant

The intelligent cockpit service can utilize any AI Agent solution. Take the hybrid AI as an example. The AI Agent assistant on the vehicle obtains the user’s voice commands through the cell phone microphone, pre-processes the commands, and converts them to text. The text is then uploaded to the agent server via the mobile communication network for intent understanding, task planning, and command generation, and commands are sent back in the downlink transmission. The vehicle agent then calls a third-party APP to execute the commands. The screenshots of the third-party APP are processed and uploaded to the agent server via the mobile communication network for task analysis and decision making. The server then generates commands to be sent back to the vehicle-side agent in the downlink transmission, and the vehicle agent calls the local system to execute the commands.

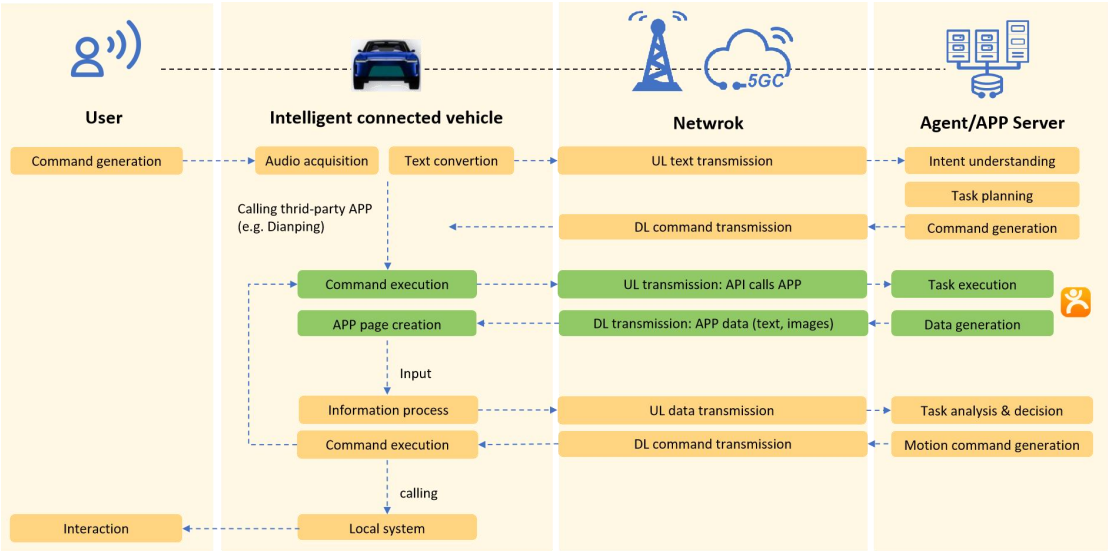

Figure 20 The service flow of the intelligent cockpit with the co-pilot AI assistant

This service places high demands on the network. Its main service data is uplink-transferred, including screenshots, service request commands, text frames, etc.

2. Cloud-assisted intelligent driving

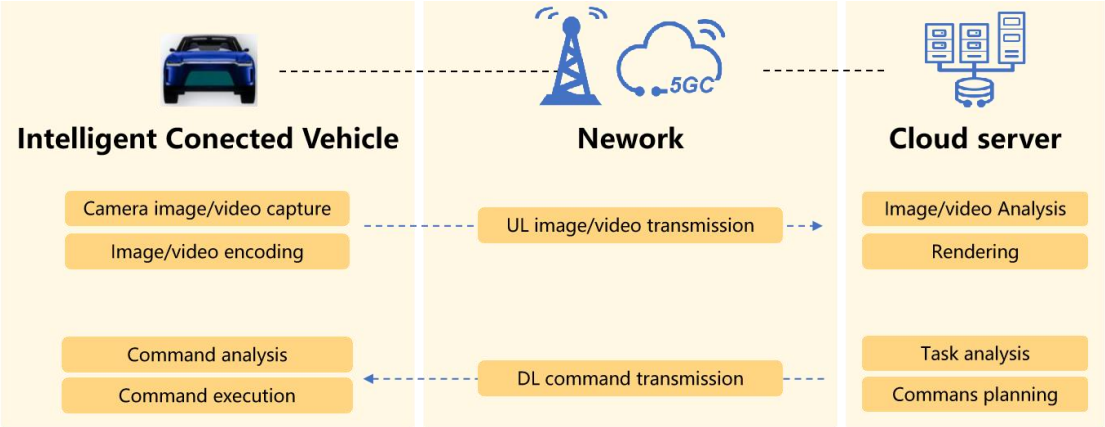

Figure 21 The service flow of the cloud-assisted intelligent driving

The cloud-assisted intelligent driving service collects and encodes images and videos from car cameras, which are then uploaded to cloud servers via mobile communication networks for image and video analysis and task analysis. The service then generates commands to be transmitted back to the vehicle for execution.

This service places high demands on the network, with uplink image transmission being the main type of data transferred.

Chapter 3 5G-A Empowers New Information Consumption Services

3.1. 5G-A Technology Scope

5G-Advanced (5G-A) narrowly refers to the new technologies defined in the 3GPP R18 specification. Broadly, it has become a general term for a new development phase. All new features defined in the 3GPP R18 standard or innovative technical solutions based on product implementation are collectively termed as 5G-A.

5G-A acts as a bridge between 5G deployment and next-generation 6G communication, driving innovation in industries such as autonomous driving,

Ambient IoT, and immersive digital experiences.

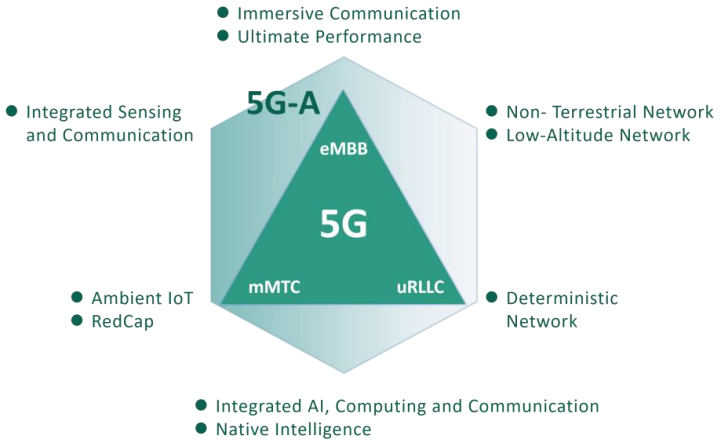

Figure 22 5G-A Top 10 Innovations

3.2. Enhanced Connectivity Guarantee

New information consumption services such as embodied intelligence, AI glasses, AI Agent, and intelligent connected vehicles are characterized by diverse service models, crossing indoor and outdoor scenarios, and stringent network performance requirements (e.g., bandwidth, delay, reliability). 5G-A, with its significantly enhanced connectivity capabilities, provides robust support for the development of these new services through five core capabilities. **Service perception** identifies high-priority transmission demands and dynamically allocates resources to ensure critical data flows receive priority. **Rate improvements** guarantee seamless delivery of ultra-HD video streams and critical data, ensuring timely and smooth transmission. **Latency assurance** enables instantaneous response to commands and real-time interaction, critical for applications like remote control and immersive experiences. **Mobility enhancement** ensures seamless handover in high-speed scenarios, maintaining low latency and stability during transitions between networks or environments. **Cluster communication** realizes that efficient device-to-device coordination supports high-density device interactions, enabling synchronized operations in scenarios like smart cities or industry automation. The synergistic effect of these five core capabilities collectively constructs a highly reliable,

high-performance connectivity foundation, which empowers the innovation and application of diverse AI-driven services across broader domains, driving advancements in areas such as autonomous systems, immersive media, and intelligent infrastructure.

1. Service perception

(1) 5G slicing: Leveraging the Traffic Descriptor in 5G network slicing to identify and distinguish service traffic enables the network to perceive the service type and characteristics, associate with the corresponding slice identifier, establish an appropriate slice session connection, and configure end-to-end connectivity including core network, transport network, and radio access network components. This mechanism instructs each network element node to provide service assurance. The 5G slicing architecture offers precise service identification with diverse and flexible granularity levels. It features an immediate effect capability where service initiation automatically triggers transmission service assurance. Service assurance is simultaneously implemented across radio access, transport, and core network domains, achieving comprehensive end-to-end network-service collaboration assurance throughout the entire service lifecycle.

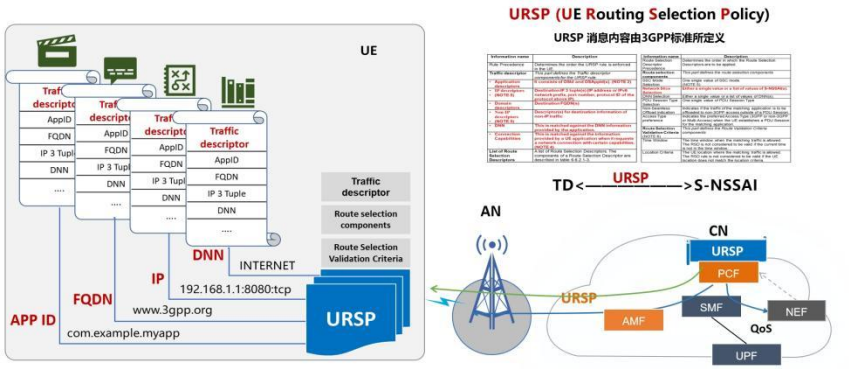

Figure 23 The technique principle of 5G slicing

(2) QoS: The QoS Flow represents the smallest differentiation granularity of QoS (Quality of Service), where different QoS Flows provide varying levels of service

assurance capabilities. AI-driven intelligent services typically involve multiple correlated data streams, each with distinct QoS requirements. The network can identify service types by analyzing the 5QI (5-Quantum Identifier) carried in the QoS Flow and prioritize traffic based on importance. By assigning different QoS requirements to data streams according to their criticality, the network allocates appropriate transmission resources. For instance, basic-layer data is configured with high-reliability QoS to ensure priority transmission of critical information, while non-critical data employs standard QoS. Under the premise of maintaining user experience, the air interface may tolerate certain transmission errors for non-priority data, thereby enhancing transmission efficiency and improving both service quality and the effectiveness of wireless communication.

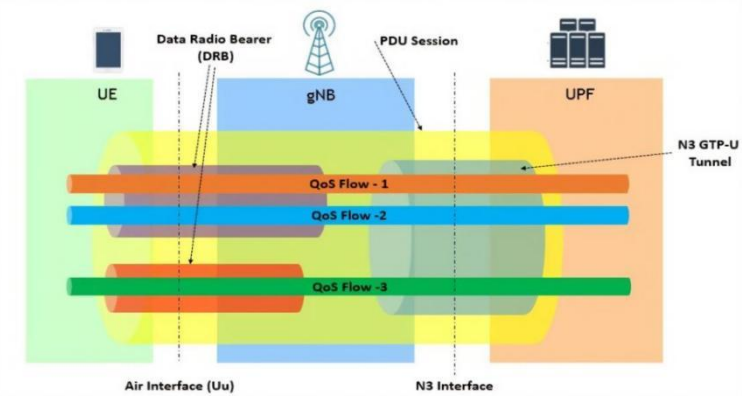

Figure 24 The technique principle of QoS

2. Rate improvements

(1) 3CC: 3CC enhances network performance by bundling three distinct frequency band carriers into a unified channel. It combines joint scheduling algorithms to dynamically allocate data streams, precisely synchronize inter-carrier timing, and leverages terminal-side multi-carrier processing capabilities. This technology significantly boosts peak network throughput, reduces latency, and enhances capacity. This approach provides critical support for emerging applications like embodied intelligence: its ultra-wide bandwidth and low-latency characteristics enable real-time transmission of multi-channel high-definition environmental

perception videos, massive sensor data, and precise control commands from intelligent devices such as robots. It facilitates seamless autonomous navigation, human-robot collaboration, and immersive remote operation, delivering essential connectivity assurance for next-generation services.

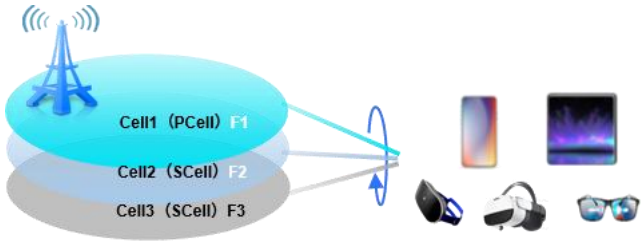

Figure 25 The technique principle of 3CC

(2) SUL (Supplementary Uplink): SUL is to deploy an independent uplink carrier on a low-frequency band while sharing the same downlink carrier with mid/high-frequency primary carriers. By leveraging the strong coverage and penetration capabilities of low-frequency signals, the network dynamically schedules terminals to switch to the SUL carrier for data transmission in weak signal scenarios (e.g., when downlink RSRP falls below a threshold). This addresses the challenges of limited 5G high-frequency uplink coverage and terminal transmit power constraints. Through high-low frequency coordination and time-frequency domain resource aggregation, SUL significantly improves uplink edge rates and reduces delay. It enables emerging applications such as embodied intelligence robots (real-time synchronization of multi-sensor data streams for autonomous navigation and human-robot collaboration), intelligent connected vehicles (massive traffic video backhaul for real-time road condition monitoring and AI-based decision-making), industry AI quality inspection, and XR (Extended Reality) immersive interactions. This technology provides critical uplink connectivity assurance for AI-driven human-machine collaboration scenarios, ensuring reliable and efficient data transmission in resource-constrained environments.

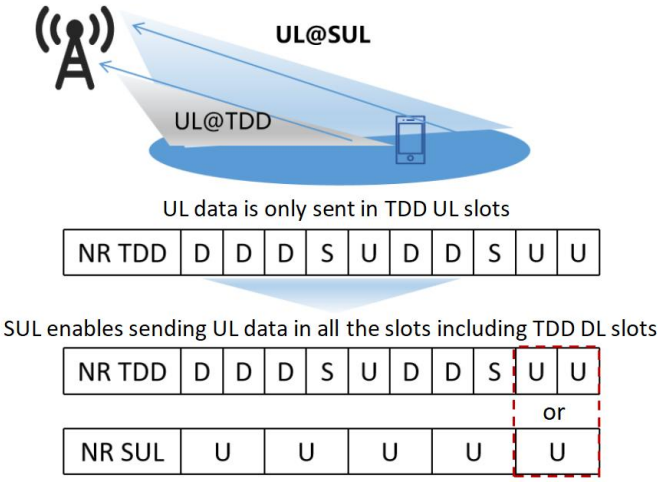

Figure 26 The technique principle of SUL

3. Latency assurance

(1) L4S (Low Latency, Low Loss, Scalable throughput):

For AI-driven intelligent applications, interactive response latency is a critical parameter to ensure smooth user experience during cloud-AI interaction. If the response delay cannot be maintained at a low level, users will find AI-based services intolerable. Therefore, reliable and deterministic latency performance must be guaranteed for such delay-sensitive applications.

L4S (Low Latency, Low Loss, Scalable Throughput) congestion control is specifically designed for delay-sensitive services. It alleviates congestion by performing real-time congestion detection and proactive congestion control across end-to-end (E2E) transmission links, ensuring a seamless user experience.

The L4S mechanism operates as follows: network devices evaluate congestion status by analyzing the waiting time of data packets in the transmission queue. This congestion information is then transmitted to the application layer, enabling it to adjust service transmission rates dynamically based on network conditions.

L4S represents an end-to-end perception-feedback-response framework under network-service-terminal collaboration. By dynamically and adaptively matching transmission rates with the perceived congestion state of the link in real time, it

reduces the waiting duration of L4S packets in the buffer queue, ensuring interactive latency and preventing severe user experience degradation caused by full congestion.

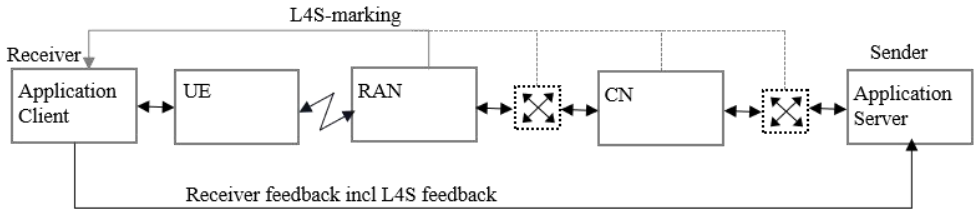

Figure 27 The technique principle of L4S

(2) PSDB: In practical network transmission, AI-based services involve diverse data formats. Application-layer data is fragmented into multiple packets, which collectively form a data packet set. These packets within the set exhibit interdependencies, meaning the loss or delay of any single packet can compromise the accurate decoding of the entire application-layer data. Additionally, retransmission or delayed reception of packets introduces prolonged waiting latency. Therefore, scheduling data transmission at the granularity of the packet set as a fundamental unit can significantly enhance latency assurance performance, thereby improving the overall user experience and increasing user satisfaction with service quality.

4. Mobility enhancement

(1) Frame-level mobility enhancement: In AI-driven applications such as intelligent connected vehicles and outdoor smart robotic dog inspections, mobility scenarios are prevalent. However, cell handover during movement significantly impacts service throughput and user experience. Frame-level mobility enhancement technology primarily mitigates these effects by controlling handover timing to minimize disruptions to data transmission. When the currently transmitted data stream exhibits frame-level characteristics, the technology identifies features of frame header and trailer packets. It initiates the handover process during the interval

between two frames, thereby reducing the impact of handover interruption delay on the complete transmission of a single frame.

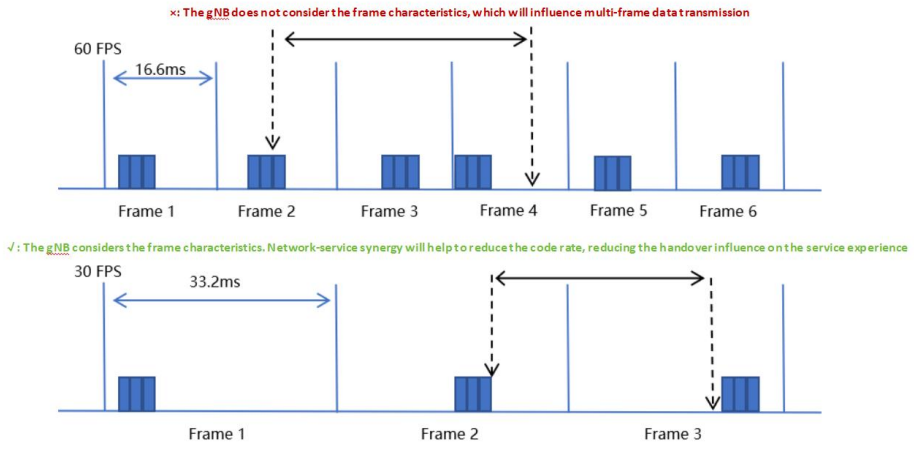

Figure 28 The technique principle of frame-level mobility enhancement

(2) Link Adaptation Optimization in Mobility Scenarios: The MCS (Modulation and Coding Scheme) optimization selection strategy for mobility scenarios addresses the issue of overly conservative initial MCS selection after handover. Two approaches are employed: First, at the target base station, historical MCS data collected before and after previous handovers is utilized to set the initial MCS after the handover. Second, the device reports measurement reports or CQI (Channel Quality Indicator) information from the source cell to the target cell during the reconfiguration completion command (e.g., Physical Channel Reconfiguration Completion). The target cell estimates SINR (Signal-to-Interference-plus-Noise Ratio) based on the measurement report or references the CQI measurements from the source cell prior to handover to determine the post-handover initial MCS. By implementing these methods, the initial MCS selection level can be elevated, effectively mitigating the abrupt rate drop and excessive service latency during handover processes.

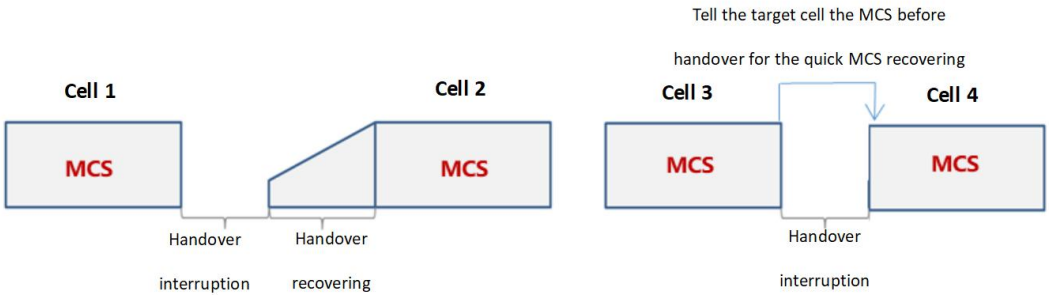

Figure 29 The technique principle of link adaptation optimization

5. Cluster communication

Currently, intelligent connected vehicles basically support the update of in-vehicle systems and software based on the over-the-air (OTA) download technology, which can improve the response speed of the in-vehicle system, enhance the battery's endurance, and enrich the functions of the entertainment system through OTA upgrades. The problem is that when automobile manufacturers carry out OTA updates for intelligent connected vehicles, they generally choose specific times for large-scale upgrades, and the upgrade packets are large, resulting in high-flow downlink transmission at the same time and taking up a large portion of the network bandwidth, leading to the problem of the network congestion. For a single vehicle, the limited resources allocated by the network result in a long upgrade time for a single vehicle.

5G MBS technology adopts broadcast or multicast to realize point-to-multi-point distribution and can perform OTA upgrades for a large number of intelligent connected vehicles at the same time. It not only effectively saves network bandwidth, but also significantly shortens the upgrade time of the single vehicle. 5G MBS technology can also be used for OTA upgrades for a large number of intelligent robots at the same time.

3.3. Computing Power Resources Opening and Sharing

In the 5G-A era, base station computing power is undergoing a profound transformation from dedicated physical hardware to the integration of general-purpose processors and dedicated physical hardware, supporting increasingly diversified service requirements. Traditionally, base stations mainly rely on dedicated chips such as ASICs, DSPs, and FPGAs to handle communication tasks. With the emergence of new applications such as AI, high-definition video, intelligent connected vehicles, and embodied intelligence, the service demand for computing power has shown exponential growth. For this reason, 5G-A base station computing power begins to strengthen general-purpose processors such as CPUs, GPUs, and NPUs. With the advantages of wide distribution and proximity to users, 5G-A base stations have become the core carrier of wireless connection services and edge computing power.

As a core driving technology, the sharing of wireless computing power resources plays a key enabling role in empowering service development. Currently, 5G-A wireless computing power sharing faces two major challenges. The first challenge is how to uniformly manage and orchestrate wireless computing power with a large number of wireless computing power nodes, heterogeneous forms, and distributed deployments. The second challenge is how to provide wireless computing power through the wireless access network.

In terms of unified management and orchestration of wireless distributed heterogeneous computing power, the wireless access network management layer needs to enhance computing power orchestration capabilities. After the distributed wireless computing power is connected to the network, it actively registers with the management and orchestration layer to achieve automatic discovery and global unified management and orchestration of wireless computing power. Considering that wireless computing power has obvious characteristics of distribution and dynamic fluctuation, the management and orchestration layer can adopt resource pooling technology to combine discrete wireless computing power nodes into a logically unified deep edge computing power pool, effectively improving the

utilization rate and flexibility of wireless computing power resources.

In terms of wireless computing power sharing, wireless computing power nodes can provide wireless computing power services and wireless communication-computing integrated services. For the wireless computing power service, the management and orchestration layer realizes the logical isolation and on-demand allocation of wireless computing power through virtualization technology and establishes a full-lifecycle management mechanism including resource registration, discovery, reservation, elastic expansion, and release. OTT service providers and vertical industry users obtain available computing power resources through service discovery, make optimal choices based on factors such as geographical location and computing capabilities, and realize the application deployment in wireless computing power.

The wireless communication-computing integrated service realizes the deep hosting and intelligent maintenance of applications on the wireless access network side by constructing a new type of network service with deep integration of communication and computing. The wireless management and orchestration layer can obtain multi-dimensional requirements such as service experience requirements through standardized interfaces, such as key SLA indicators such as end-to-end delay, throughput, jitter, and packet loss rate. The wireless management and orchestration layer first translates the application SLA requirements into requirements for two types of resources: wireless communication and computing, deploys applications on matching wireless computing power, and generates service quality guarantee strategies. The wireless management and orchestration layer monitors the status indicators of applications, the wireless network environment, and the status information of wireless computing power. When it predicts or finds that the application experience may deteriorate, it can achieve multi-dimensional experience guarantee through the joint adjustment of wireless communication resources and wireless computing power.

The sharing of wireless computing power is the key to supporting the

development of 5G-A services. Through the unified management and orchestration of wireless computing power resources, wireless computing power services, and wireless communication-computing integrated services, it can support the flexible deployment of services and the joint optimization of communication and computing, and effectively guarantee the quality of business experience.

3.4. Data Acquisition and Storage

Facing the data requirements of the differentiated communication-computing integrated application scenarios, the traditional centralized data resource supply mode represented by “data centers + data pipelines” can no longer meet the requirements in terms of real-time performance, synchronism, usability, security, and privacy. Therefore, the wireless communication-computing integrated 5G-A network needs to achieve real-time data collection, routing and forwarding, and processing and management functions in dynamic spatial, temporal, and system environments. It is advisable to introduce the following data-related functions.

1. **Data collection**: Through customized data collection capabilities, the wireless communication-computing integrated 5G-A network can collect cross-protocol layer (such as PHY, MAC, PDCP) and cross-system domain data from various devices and wireless networks according to native AI task requirements and unified rules. Meanwhile, it can provide data distribution services based on multiple collection modes to support distributed AI computing tasks.

2. **Data preprocessing**: This function preprocesses the collected data, such as data cleaning, denoising, and feature extraction, to enable subsequent computing tasks to be executed more efficiently. It can also dynamically deploy data preprocessing-related functions according to the customized dataset requests from computing tasks, and perform integration processing on data from different sources. In addition, data privacy protection processing is an important function in the preprocessing process. Data services must comply with corresponding privacy

protection regulations, protecting user privacy through de-identification, anonymization, and other capabilities.

3. **Data routing and forwarding**: Considering the diversity of data sources, the complexity of transmission links, and the flexibility of transmission modes, this function ensures the timeliness, reliability, and integrity of collected data during transmission and forwarding in all scenarios (such as mobility scenarios). It can intelligently select the most suitable path and transmission scheme to transfer data from data sources to computing nodes among clouds, edges, and devices.

4. **Data storage**: In the wireless communication-computing integrated 5G-A network environment, heterogeneous computing nodes have significant performance differences, and some intelligent devices have weak data caching capabilities, making it difficult to synchronize multi-source received data, while data synchronization is crucial for maintaining data consistency. The data caching and synchronization can solve the data synchronization problem from different times and spaces through a data caching mechanism, achieving synchronization of communication-computing data collected from clouds, edges, and devices, and ensuring that computing nodes can obtain multi-source synchronized data at low cost, as well as ensure the reliability and availability of stored data.

Chapter 4 Summary and Outlook

The new generation of information consumption services is profoundly reshaping personal life, industrial ecosystems, and social operation. For individuals, AI assistants enable efficient and natural interactive services, AI glasses expand immersive virtual-real integration experiences, and embodied intelligence promotes personalized companionship and collaboration of robots. For industries, embodied intelligent robots become the hub of mobile productivity and empower remote control and decision-making in scenarios such as industry campus and medical healthcare. For society, the new intelligent services, such as intelligent connected

vehicles, will jointly promote the universal accessibility of barrier-free services, the refinement of urban governance, and low-carbon development. The capabilities of new services, such as multi-modal perception, real-time decision-making, and cross-domain collaboration, rely on the network foundation of 5G-A with ultra-low latency, ultra-large bandwidth, high reliability, and mobility guarantee. 5G-A provides core support for intelligent services to evolve from single-point innovation to universal penetration by constructing an integrated foundation of "connection + computing power + data". Facing the requirements of mobile communication networks for services, this white paper clarifies the processes and network metrics of different intelligent services, providing technical references for the development of intelligent emerging services in the current network.

Looking to the future, with the development of complex service scenarios such as embodied intelligence and networked autonomous driving, multi-technology integration and cross-industry collaboration will become the core driving forces. First, we need to continuously deepen 5G-A capabilities, strengthen the foundation of 5G-A networks through technologies such as communication-sensing integration, network intelligence, and immersive communication, build a collaborative ecosystem, work with industry partners from multiple fields including communications, automotive, robotics, and cloud computing to deeply explore network capabilities and service requirements, and create benchmark applications. Second, for the 6G era, technological layout needs to further leap towards multi-agent collaboration. Through technological research in fields such as intelligent agent collaboration networks, breakthroughs in digital twin interaction, and exploration of communication-sensing-intelligence-computing integrated architectures, we need to promote the development and implementation of closed loops for universal intelligent agent cognition, decision-making, and action. Only by advancing both the continuous deepening of current network capabilities and the forward-looking layout of next-generation technologies can we achieve leapfrog development from "single-agent intelligence" to "collective intelligence" and from "scenario

empowerment" to "social reshaping".

Appendix: The Theoretical Analysis of Embodied Intelligence Service Requirement

For the theoretical analysis of requirements for emerging services such as embodied intelligence, the radio delay is first estimated through segmented delay, and then the radio rate is calculated based on the radio delay. The derivation formula for the radio delay is

$$\text{Radio Delay } \boldsymbol{T2} = \text{E2E Delay } \boldsymbol{T} - \text{Robot Processing Delay } \boldsymbol{T1} - \text{Backbone \& CN Delay } \boldsymbol{T3} - \text{Cloud Processing Delay } \boldsymbol{T4} \quad (1)$$

The radio rate can be estimated based on the radio delay. For data types such as voice, text, commands, and images, based on Formula (2), the derivation formula for the air interface rate is:

$$\text{radio rate} \geq \frac{\text{transmission data} + \text{communication payload}}{\text{radio delay}} \quad (2)$$

wherein, communication overhead≈ transmission data×0.2.

When the robot's uplink transmission service is a video-based service, since video is transmitted in a frame-level continuous manner, and different resolutions and frame rates correspond to different data volumes per frame and different radio delays, the derivation formula for the radio rate of video-based services is:

$$\text{radio rate} \geq \frac{\text{data per frame} + \text{communication payload}}{\text{radio delay}} \quad (3)$$

Wherein,

communication overhead≈data per frame×0.2,

data per frame = bit rate / frame rate, and

radio delay = 1 / frame rate.

Based on the above formulas, Table 9 provides the theoretical estimated values of air interface transmission delay and air interface rate by taking voice transmission, 1080p images, and 1080p/30fps video as examples. The service scenario for video transmission is set as remote control of robots via VR headsets:

Table 9 The theoretical analysis of embodied intelligence service requirement

Transmission Data		**End-to-end Delay**	**Service Data**	**UL Radio Delay**	**UL Radio Rate**
Voice/Text/Command		1s	~100KB	~400ms[1]	~2Mbps
Image (1080P)		2s	~400KB	~900ms[2]	~4.5Mbps
Transmission Data		**Bit Rate**	**Frame Rate**	**Radio Delay**	**Dario Rate**
Video (1080P, 30fps)	UL Video Transmission	~8Mbps	30fps	33.33ms	~10Mbps
	DL VR Presentation	~12Mbps[3]		~30ms[4]	~29Mbps[5]

Note:

1. Assume the delay of each part: terminal 50ms, backbone network 50ms, large model 500ms.

2. Assume the delay of each part: terminal 50ms, backbone network 50ms, large model 1000ms.

3. After 3D rendering and encoding, it is approximately 1.5 to 2 times the original video bit rate.

4. Taking XR as an example, the single-loopback delay within 100ms can basically meet user experience requirements; assume the delay of each part: terminal 10ms, backbone network 30ms, cloud rendering 30ms.

5. Video frame-level services exhibit burst characteristics, with the instantaneous rate approximately 2 to 3 times the average rate.