Xu, T., Gao, M.N., Liu, X.D., Ye, Q., Ma, L., & Zhu, B.Y. (2025). Intelligent travel avatar: An LLM-based tourism quadrupedal robot. In Li, E.Y. *et al.* (Eds.) *Proceedings of The International Conference on Electronic Business, Volume 25* (pp. 247-256). ICEB'25, Hanoi, Vietnam, August 21-25, 2025.

Intelligent Travel Avatar: An LLM-based Tourism Quadrupedal Robot

Tao Xu [1]
Manni Gao [2,*]
Xudong Liu [3]
Qiang Ye [4]
Lin Ma [5]
Bingyu Zhu [6]

*Corresponding author
[1] Harbin Institute of Technology, Harbin, China, hitxutao@163.com
[2] Harbin Institute of Technology, Harbin, China, mannigao@stu.hit.edu.cn
[3] Harbin Institute of Technology, Harbin, China, cameran@hit.edu.cn
[4] University of Science and Technology of China, Hefei, China, yeqiang@ustc.edu.cn
[5] Harbin Institute of Technology, Harbin, China, upallnight2019@163.com
[6] Harbin Institute of Technology, Harbin, China, 179328331@qq.com

ABSTRACT

The digital transformation of tourism, while enhancing efficiency and marketing, faces persistent challenges in delivering truly immersive and authentic remote experiences. While immersive technologies like VR mitigate temporal-spatial constraints, current implementations often lack real-time intelligence, personalized co-creation, and deep contextual understanding. This study introduces the Intelligent Travel Avatar, an embodied AI paradigm that combines an advanced LLM with a quadrupedal robot, which integrates physical environment perception, real-time data processing, and multimodal interaction to deliver context-aware, personalized guidance through natural language dialogue. Through the combination of embodied robotics and LLM-driven intelligence, this study offers both theoretical contributions and practical insights into the design of AI-powered tourism agents, while extending the application of GenAI in the tourism domain.

Keywords: Virtual tourism, travel avatar, LLM, quadrupedal robot.

INTRODUCTION

As digital technologies increasingly shape contemporary life, they have profoundly transformed the tourism industry and opened new avenues for future development (Huang *et al.*, 2016). Integrating advanced technologies into tourism significantly improves productivity, efficiency, effectiveness, and marketing capabilities. Travel avatars in virtual tourism have emerged as promising solutions to overcome temporal and spatial constraints inherent in traditional tourism, including public health crises, physical disabilities, and heritage preservation concerns (Momani *et al.*, 2022). Although existing tourism avatars support real-time remote sightseeing, they are often limited in their ability to provide diverse content, personalized guidance, and intelligent interaction. These constraints reduce user immersion and satisfaction in both virtual tourism and remote visit scenarios (Chamola *et al.*, 2024b). Recent advancements in Generative Artificial Intelligence (GenAI) offers an opportunity to address these limitations by enabling intelligent, context-aware, and user-centric travel experiences (Chamola *et al.*, 2024a).

To this end, this study proposes the Intelligent Travel Avatar, a large language model-powered tourism quadrupedal robot equipped with gimbal-mounted cameras, enabling teleoperation and telepresence via VR, PC and mobile devices. Embedded with Doubao large language model (LLM) and ResNet-50 image recognition model, the avatar achieves physical environment perception, real-time data processing, and multimodal interaction, thereby delivering personalized, interactive, and intelligent tour services. By integrating state-of-the-art technologies, such as quadrupedal robots, telerobotics, LLMs and image recognition models, the avatar not only addresses the key limitations of current virtual tourism systems but also represents a novel application of LLMs in the tourism domain. This study contributes to the growing literature on GenAI-powered tourism and offers design insights for next-generation virtual travel experiences.

LITERATURE REVIEW

Technology has long served as a transformative lever in the tourism sector (Sharma *et al.*, 2021). The digitalization of tourism, empowered mainly by information and communication technologies, has led to the emergence of various paradigms, including e-tourism, digital tourism, smart tourism, and virtual tourism. As virtual tourism gains prominence, the integration of enabling technologies such as drones, live-streaming, and virtual and augmented reality (VR/AR) has accelerated the development of travel avatars (Pestek & Sarvan, 2020). More recently, the rise of GenAI further expands the potential of travel avatars by enhancing their interactivity, contextual awareness, and ability to deliver personalized services.

Travel Avatars in Virtual Tourism

Virtual tourism has emerged as a viable alternative to physical travel, particularly in response to crises such as the COVID-19 pandemic (Pestek & Sarvan, 2020). Facilitating virtual or remote site visits typically requires an intermediary technological platform, and several travel avatars have been introduced to meet these demands. A range of innovative systems demonstrate the potential of avatars in enhancing user engagement, satisfaction, and behavioral intention (Tussyadiah *et al.*, 2018). The simplest implementations offer video streams of tourist destinations that users can view on smartphones, tablets, or desktops, providing real-time remote access to global attractions (Verma *et al.*, 2022). For example, the Avatar Tourist Visit model enables interpretive guided tours of heritage sites by integrating on-site human guides with remote virtual participants (Viñals *et al.*, 2021). Similarly, remotely operated drones (e.g., DJI Avata 2) and metaverse-based tourism platforms allow users to explore digital replicas of physical locations or interact with avatars in immersive virtual environments (Shin & Kang, 2024). Despite these advancements, existing travel avatar systems largely rely on static content and lack contextual intelligence and interactive capabilities, thereby limiting their ability to deliver personalized and immersive tourism experiences.

GenAI Applications in Tourism

Recent breakthroughs in GenAI have opened new possibilities across sectors such as education, healthcare, emotional support domains, as well as tourism (Hasan *et al.*, 2023; Wong *et al.*, 2023). Its applications in the tourism industry are able to boost customer engagement and loyalty, customize travel plans, optimize the management of tourism companies, and provide other benefits (Carvalho & Ivanov, 2023).For example, the AI-driven TravelAgent system delivers personalized, comprehensive, and rational travel planning services. Similarly, LLM-based applications such as ChatGPT have enhanced decision-making across the pre-trip, en-route, and post-trip stages, enabling real-time planning, round-the-clock assistance, natural communication, and enriched travel experiences (Wong *et al.*, 2023).

Generative AI generated content, including text, images, audio, 3D objects, and interactive narratives, plays a pivotal role in unlocking the immersive potential of the Metaverse. Text generation supports dynamic narration and environmental description, while image and video generation facilitate the creation of realistic and personalized virtual experiences (Chamola *et al.*, 2024a). For example, tourists can virtually explore historical landmarks or museums where GenAI-powered agents offer real-time commentary on cultural significance, thereby increasing engagement and memorability (Dubourg *et al.*, 2024).

Although existing tourism avatars support remote and real-time exploration, they fall short in delivering interactive, adaptive, and intelligent experiences. Most lack the ability to sense and respond to user context dynamically, and they operate without the embodied intelligence necessary to meet evolving user needs. GenAI holds promise in bridging this gap, yet its integration with embodied devices remains underdeveloped in tourism industries. To address these limitations, this study proposes the Intelligent Travel Avatar that integrates quadrupedal robots, telerobotics, and LLMs to achieve physical environment perception, real-time data processing, and multimodal interactions.

RELATED TECHNOLOGIES FOR INTELLIGENT TRAVEL AVATAR

To bridge the foresaid gaps in the tourism industry, we prepare to design an intelligent travel avatar, which need several technical supports. We briefly review the related technologies including their basic definition, advancement and limitations. Specifically, the device–cloud collaboration architecture is adopted to realize the overall design of physical environment sensing, data processing, and multimodal interaction, with the quadrupedal robot and the drone camera as the main body of the avatar, the telerobotics technology and user terminal devices to provide the function of remote manipulation of the avatar, and the image recognition model and LLM used to enable intelligent interaction.

Device-Cloud Collaboration Architecture

With the help of deep neural networks (DNNs), robots endowed with self-X (aware, optimizing, and learning) ability become smarter. However, since DNNs often requires high-performance computing resources (GPUs, CPUs and storage devices) for model training and execution on massive data, exiting robots may not fulfill this stringent requirement on computing capability. A common practice is to offload part of computation tasks from the robot to cloud by way of device–cloud collaboration (Hu *et al.*, 2023).

The device–cloud collaboration architecture represents a two-tiered distributed computing paradigm comprising cloud computing, and device layers. Cloud computing enables organizations to dynamically scale computing resources based on demand, leveraging infrastructure provided by third-party service providers. It offers several advantages, including elasticity, cost efficiency, and enhanced security, thereby reducing operational overhead and idle hardware investment.

Thus, this structure retains the centralized coordination advantages of traditional mainframe systems while distributing computation across multiple nodes, facilitating distributed data storage, localized data processing, and seamless interaction with end devices (Wang *et al.*, 2020).

Quadrupedal Robots

Quadrupedal robots are robots that mimic the locomotion of quadrupedal animals, realizing movement and attitude adjustment through four legs. Its core features include bionic structures, dynamic balancing capabilities, and strong adaptability to

complex terrains. Compared to tracked or wheeled robots, quadrupedal robots exhibit superior adaptability to varied real-world terrains such as muddy paths, slopes, and staircases. Moreover, quadrupedal robots offer enhanced stability over bipedal robots due to their additional support points, smaller form factor, longer battery life, and more mature technological implementations.

Traditional control methods for quadrupedal robots have been successfully applied across diverse operational environments (Jenelten *et al.*, 2019), while these approaches generally depend on extensive domain expertise and manual tuning. There has been significant attention on applying model-free reinforcement learning to control quadrupedal robots for agile locomotion skills in the real world (Hwangbo *et al.*, 2019). Such systems employ sim-to-real transfer techniques, enabling policies trained in simulated environments to be effectively deployed onto physical robots. However, these reinforcement learning applications predominantly focus on developing low-level locomotion capabilities, such as basic walking behaviors. Recent advancements in learning-based quadrupedal robot control systems have demonstrated considerable success across various scenarios (Hu *et al.*, 2019). For example, Tan *et al.* (2024) introduced a hierarchical reinforcement learning framework enabling agile and terrain-adaptive locomotion without pre-training, thus significantly improving the ease of system deployment and adaptability to real-world environments. Unitree Robotics' B1 and Go2 have found applications in military contexts, specifically for battlefield reconnaissance and engagement tasks.

Telerobotics

Telerobotics, also known as remote-controlled robotics, refers to the field within robotics focused on remotely operating semi-autonomous robots, typically via wireless networks such as Wi-Fi, Bluetooth, Deep Space Networks, or tethering network connections. Telerobotics integrates two primary subfields: teleoperation and telepresence (Martínez-Romero *et al.*, 2015).

Teleoperation involves master-slave systems wherein robot movements are remotely controlled in synchronization with human actions. Researchers have proposed various teleoperation approaches, including kinesthetic teaching, joystick-based control, virtual reality interfaces, mobile-device terminals, RGB camera tracking, exoskeleton systems, and motion capture technologies. Teleoperation technology enables humans to extend their interactive range with physical environments through remote robots. For instance, in telesurgery, physicians can perform precise surgical procedures by remotely controlling robotic systems (Tian *et al.*, 2020).

Telepresence refers to the immersive experience of controlling a remote robot, allowing users to perceive themselves as physically present in the robot' s environment through somatic feedback and real-time visual transmission (Kikuchi *et al.*, 2022). By delivering continuous video streams from the remote environment, telepresence creates a sense of embodiment and situational awareness. For instance, telepresence enables immersion in VR through sensory modalities such as camera-based vision and pressure-driven pseudo-haptics (Desnoyers-Stewart *et al.*, 2023).

The application scenarios of telerobotics are diverse and socially impactful. For example, individuals with physical disabilities can remotely perform work tasks through teleoperated robots (Tsuchiya & Koizumi, 2020). In disaster response, telepresence robots can be operated remotely in hazardous environments, reducing human risk and improving efficiency in rescue operations (Vaz *et al.*, 2024).

Multimodal Interaction

Multimodal interaction refers to the process of converting information across different modalities, such as text, image, and audio, thereby enhancing expressive capabilities of AI agents and interactivity between AI agents and users (Baltrušaitis *et al.*, 2018). Typical applications include speech-to-text transcription, AI image generation, and multimodal medical diagnostics. This technology has been widely applied in human-computer interaction scenarios including healthcare and autonomous driving. In this study, the proposed avatar system is designed to support interaction between text, audio, and image. To this end, we briefly introduce the three core components that enable such multimodal processing: text generation via large language models, text-to-speech synthesis, and image recognition models.

Large language model

Large Language Models (LLMs) generally refer to Transformer-based models characterized by extensive parameter, often in the hundreds of billions or beyond, and trained on massive-scale textual datasets, such as ChatGPT, PaLM, and LlaMA.

LLMs have demonstrated strong capabilities in natural language understanding and performing complex tasks through text generation. Since the great development of ChatGPT-3.5, LLMs show great influence and wide utilization in the all industries, including education, creativity and tourism (Gao *et al.*, 2024). Moreover, integrating LLMs with Text-To-Speech (TTS) technologies has substantially enhanced user interactivity. High-quality synthesized voice outputs coupled with minimal latency have created seamless conversational experiences for users interacting with LLMs. For instance, ChatGPT's voice interaction capability closely mimics natural human speech, while other open-source LLMs have similarly incorporated TTS packages to facilitate accessible interactions.

Image recognition model

Image recognition models have gained significant momentum in recent years, with different types having their own advantages for various tasks. Since Residual Neural Network (ResNet, hereafter) is made up of layers, these networks can be arbitrarily deep

for an arbitrary level of spatial representation, it has been successful in diverse tasks, including image classification, object detection, and semantic segmentation. ResNet has emerged as highly effective deep-learning models, characterized by ease of training and robustness in maintaining or even improving accuracy with increasing network depth (Zhang *et al*., 2022).

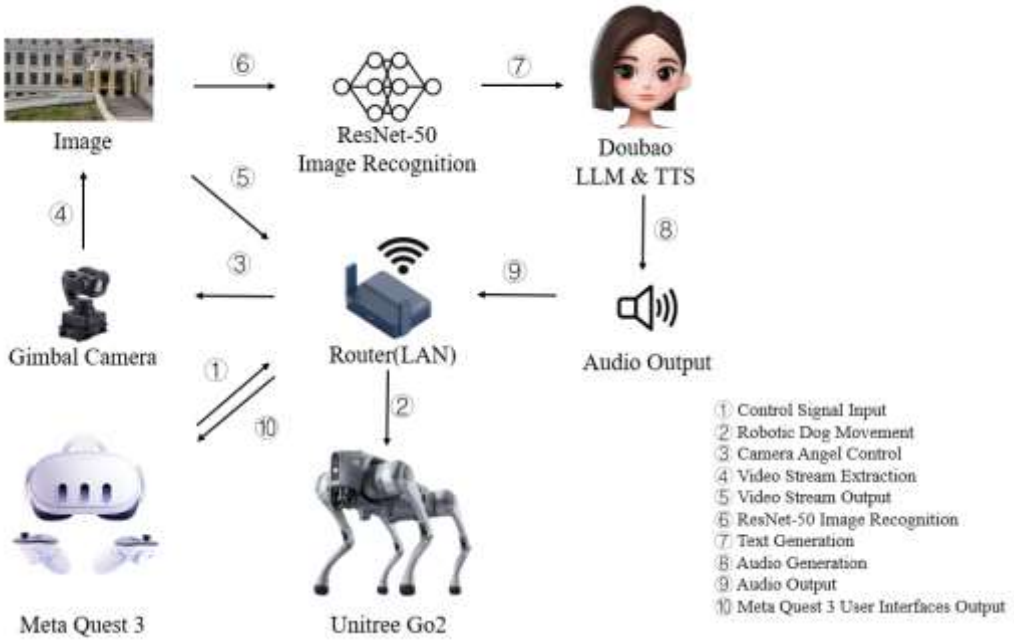
The core innovation of ResNet lies in its approach to learning residual mappings rather than direct outputs for each network layer. This architecture employs skip connections, allowing layers to bypass certain intermediate layers and directly combine their outputs. The underlying mechanism resembles that of Highway Networks, where gating mechanisms with significantly positive bias weights facilitate efficient information flow. Additionally, the skip connection structure of ResNet has also been incorporated into Transformer, which has been widely applied to text and image processing tasks (Liu *et al.*, 2021).

DESIGN PRACTICES

Design Concept

Considering the current absence of intelligent avatars in the tourism industry and the maturation of relevant technologies, we propose an Intelligent Travel Avatar, an LLM-based quadrupedal robots with multimodal information processing and interaction. As illustrated in Figure 1, our system leverages a Unitree quadrupedal robot as the main body of the avatar, equipped with an external gimbal-mounted camera that streams live video to remote users. As illustrated in Figure 2, The avatar is designed to operate within selected attractions or urban environments, such as famous university campuses, while enabling remote users to freely navigate the robot through VR headsets or web-based interfaces. This avatar precisely identifies attractions using a ResNet-50 image classification model, then employs the Doubao LLM to generate real-time, context-aware narrations and interactions about the detected attractions. The multimodal interaction loop seamlessly integrates image or voice inputs converted into text, and subsequently transformed into natural speech outputs. More importantly, integrating the LLM facilitates dynamic interactions, including engaging dialogues, contextual explanations, and itinerary planning, thereby empowering a highly personalized and interactive remote visit or virtual tourism. To realize this design, we define the system's core functional requirements, including telerobotics and intelligent interaction.

Source: This study.
Figure 1: Exterior Rendering of Intelligent Travel Avatar



Source: This study.
Figure 2: Workflow of the Avatar' s Intelligent Tourism Service

Function Requirements

Telerobotics

To realize personalized remote travel, the design of the avatar should facilitate the teleoperation and telepresence, including the teleoperation of the camera's viewpoint and the robot's movement, as well as telepresence methods. Specifically, the avatar supports two distinct teleoperation methods: interaction via VR, PC and mobile devices. Additionally, telepresence is realized through the screen of the VR headsets and the web interface, specifically through the use of the VLC player plugin, which is a free and open source software.

We develop a Unity3D-based client APP and deploy it on users' devices to collect, compute and transmit users' actions and instructions. More specifically, users view real-time video streams from the gimbal-mounted camera through VR headsets, and when they rotate their heads to adjust viewing angles, motion changes are computed by the APP within the VR headset and transmitted to the camera, triggering corresponding angle adjustments. Additionally, the selected camera should be optimized for stabilization, wide-angle coverage, high-definition imaging, low latency, and stable signal transmission over considerable distances. For mobility control, users manipulate VR joysticks to direct the quadrupedal robot's movements. Movement data from the joysticks are processed by the APP and subsequently transmitted to the robot' s server. The server translates this high-level motion data into low-level motion commands, utilizing a motion-control model trained through reinforcement learning by Unitree Robotics. This enables robust locomotion capabilities across complex terrains, including stair climbing and obstacle avoidance.

Considering hardware constraints associated with VR devices, two alternative teleoperation modes via the website and the mobile APP is provided. Users can access real-time video streams directly through the web interface or the APP player, using mouse to adjust camera angles and keyboard to remotely navigate the quadrupedal robot, ensuring broader accessibility and convenience for diverse user environments.

Intelligent interaction

To facilitate intelligent interactions and personalized services, our proposed travel avatar integrates multiple deep learning models for data processing and content generation. First, to meet the demands of high-performance computation and low-latency service delivery, we adopt a device-cloud collaboration architecture. Second, the avatar is designed to recognize attractions and offer contextualized guidance and narrations such as attraction highlights and historical backgrounds. Lastly, it provides additional intelligent interaction services like daily chat and itinerary planning.

The device-cloud collaboration architecture employed by the avatar effectively optimizes operational costs based on service scale, supporting distributed deployment. The end device is responsible for collecting image data and control signals. The server primarily handles lightweight data processing, performing tasks like image recognition to reduce data volume and minimize transmission latency. The cloud undertakes complex computations, enabling advanced functionalities such as speech recognition, text generation, and speech synthesis to deliver intelligent services. This architecture satisfies both the computational requirements of large-parameter models and user' expectations for minimal response delays.

The attraction explanation service first captures real-time video streams through a gimbal-mounted camera, and then the trained image recognition model processes the video frames to extract visual features and recognize attractions. Once text labels from image recognition models are obtained, these labels are passed to the Doubao LLM, generating contextually appropriate explanatory text. Subsequently, the Doubao TTS module converts this textual content into natural and fluent multilingual narrations. Then, users can hear the voice narration through the VR headset or computer audio. This integrated workflow represents a multimodal information interaction pipeline succinctly summarized as "visual perception–semantic understanding–speech generation." Additionally, leveraging the general knowledge, advanced semantic comprehension, and conversational capabilities inherent to the LLM, the avatar further provides personalized services, including daily conversational interactions, retrieval of local information, and customized itinerary recommendations, thereby substantially enriching user experience and enhancing interactive engagement.

Design Implementation Process

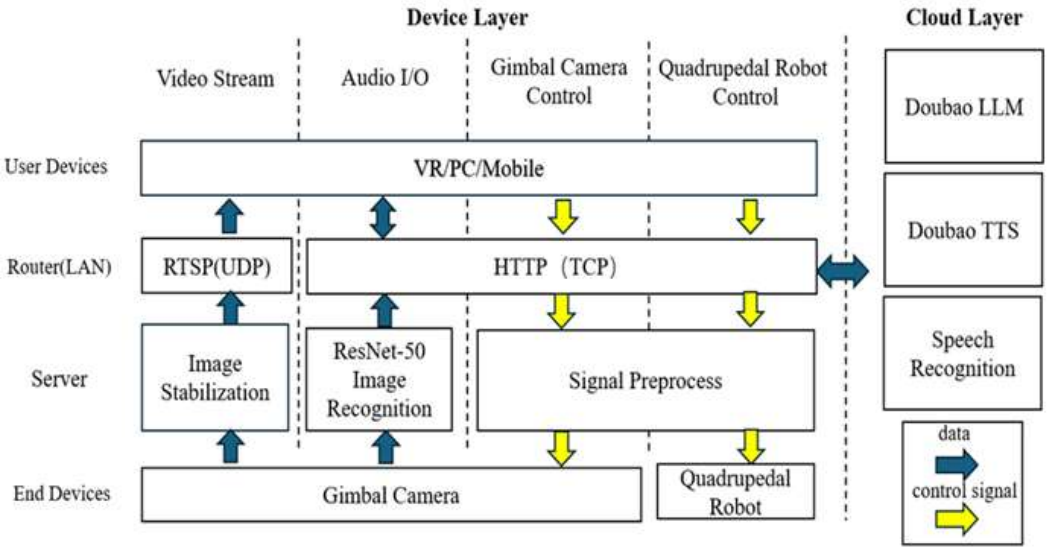
To realize the two main functions, our implementation focuses on three key areas: overall system design, teleroboics, the LLM-based multimodal interaction.

Device-cloud collaboration architecture

The avatar system adopts a device-cloud collaboration architecture to support physical environment perception, real-time data processing and multimodal interaction. As shown in Figure 3, the overall architecture consists of a device layer and a cloud layer, with system functions distributed between the two layers.

The device layer serves as the critical interface where the avatar perceives the physical environment and enables user's teleoperation and telepresence, including both the end devices and user devices. On the end devices, the quadrupedal robot operates as the mobile platform, equipped with a gimbal-mounted camera that captures real-time video streams while responding to movement and camera adjustment commands. This visual data forms the foundation for environmental perception. The user

devices include VR, PC, and mobile devices, through which users can access the telepresence interface via the Unity3D application or website. For the camera control, VR headsets, PC mouses or Mobile screen touch are employed, while VR joysticks or PC keyboards serve to convert users' movement intentions into operational commands. In all three types of user devices, audio input and output are handled by headsets in these three kind of devices, thus completing the immersive experience loop. The server in the quadrupedal robot is responsible for data processing to reduce the computational load and latency associated with transmitting large volumes of data to the cloud. Centered on the quadrupedal robot' s integrated development board, the server performs video stabilization on the gimbal-mounted camera, filters control signals, and runs the ResNet-based image recognition model to identify attractions and generate corresponding text labels, which are then uploaded to the cloud for further processing and computing.



Source: This study.

Figure 3: Device-Cloud Collaboration Architecture of the Intelligent Travel Avatar

The cloud layer handles large-scale data computation and enables intelligent interaction between the digital human and users. We depoly the Doubao large language model along with speech recognition and speech synthesis models on the Volcengine cloud platform. When a user requests attraction interpretation, the cloud-based LLM generates explanatory text based on image recognition labels, which is then synthesized into natural speech and delivered back to the user. For voice interactions, the user's audio input is transmitted to the cloud, where the LLM processes it to generate a textual response. This response is then converted into speech and returned to the user.

Throughout this architecture, data transmission between the cloud and device layers relies on a network communication layer. This layer includes Routers and local area networks (LAN, hereafter), where Routers relay traffic from LANs. Users and end devices communicate via LAN, while remote users access the LAN using NAT traversal (e.g., port forwarding) over the internet. In terms of communication protocols, control signals between devices and the server are transmitted using HTTP over TCP. Video streams, encoded with H.264, are transmitted via the RTSP protocol over UDP. The VR devices, implemented using Unity3D, handles the bidirectional transmission of control commands, audio, and video data.

Telerobotics

Based on the device–cloud architecture, the user side of the device layer controls the end devices, thereby realizing telerobotics. We designed the server using Python, primarily implementing the functions of five modules: the network communication module is responsible for transmitting data to end devices and cloud servers; the instruction processing module parses instructions received from the client; the motion control module drives the quadrupedal robot based on instructions; the state management module maintains the quadrupedal robot's posture and status; and the safety protection module monitors the system's operational status in real time.

We developed a Unity3D APP to realize users' teleoperation and telepresence of the avatar. The APP mainly implements the functions of three major modules: network communication, instruction processing, and state management.

First, create a VR project and user object in Unity3D, set the server address and port, enabling the user side to receive and display the video streams sent by the quadrupedal robot and the gimbal-mounted camera. The gimbal-mounted camera we use is enhanced with mechanical structure stability and equipped with a shock-absorbing ball to reduce shaking caused by mechanical looseness, ensuring the camera remains stable even under significant vibration. Additionally, the rotation of the gimbal-mounted camera is compensated based on the quadrupedal robot's motion state. When the quadrupedal robot moves, the gimbal-mounted

camera's roll angle is automatically adjusted to counteract the impact of the quadrupedal robot's shaking, ensuring a horizontal viewing angle and stable, clear images.

Second, relying on Unity3D's input system, the APP real-time collects motion data from VR headsets and joysticks, converts the data into JSON-formatted instructions recognizable by the quadrupedal robot side, and transmits them stably via the TCP protocol. The quadrupedal robot and gimbal-mounted camera then perform corresponding actions and adjustments. To enable the gimbal-mounted camera to track the user's viewing angle, we also designed camera control functions in the APP. The rotation actions of the headset are bound using the Input Action Property component in Unity3D. When the user turns their head, the system automatically executes data acquisition operations in the Update method of each frame to real-time obtain the quaternion rotation data of the VR headset. Due to the limited degrees of freedom of the gimbal-mounted camera, the angle data are processed with the following restrictions:
1. Map the angle values to the interval [-180°, 180°] to eliminate ambiguities caused by angle periodicity.
2. Limit the pitch angle (x-axis) to [-45°, 135°].
3. Limit the yaw angle (y-axis) to [-160°, 160°], covering the user's normal horizontal viewing angle rotation range.
4. Fix the roll angle (z-axis) at 0
The processed angle data are packaged and transmitted, and the developed camera control interface is called to adjust the gimbal-mounted camera angle.

Following the same development process as VR, we have also developed a PC website and mobile APP to enable users without VR devices to access the avatar. The user usage logs recorded by the APP assist us in monitoring and managing the system's operational status, ensuring the APP's response efficiency in high-concurrency input scenarios while reducing the coupling between functional modules to facilitate future functional expansion and maintenance.

In order to control and enhance the quadrupedal robot's motion stability, we designed a weighted filtering function and sensitivity curve. A weighted sliding average filtering algorithm based on dead-zone filtering is used to process motion instructions. This algorithm effectively filters instruction noise with low computational complexity by reasonably setting the dead-zone range and constructing a weighted sliding window, significantly improving the quadrupedal robot's motion control stability and positioning accuracy. Specifically, by zeroing smaller data to form an operational dead zone, the system does not respond to operations within the dead zone. The weighted sliding window maintains a data sequence: as new data are added, the sequence window slides, discarding old data, and finally averaging the numbers in the sequence as the control signal to buffer and enhance system stability.

In terms of interactive sensitivity adjustment, a nonlinear curve as shown in Equition (1) is selected as the mapping function.

$$y = sign(x)|x|^{\frac{3}{2}} \tag{1}$$

Where x represents the input control signal and y denotes the output signal, this nonlinear curve demonstrates excellent mapping characteristics across varying input intervals. The value range of the control signal x is (0, 1), deviating more from the linear signal when the control signal is weak and approaching the linear signal when the control signal is strong, to enhance the quadrupedal robot's stability when transitioning from rest to motion and solve the problem of abrupt responses when the quadrupedal robot moves from rest to motion.

Additionally, a state lock is added to the state management module, requiring users to confirm twice during the quadrupedal robot's state transition. For example, after the quadrupedal robot switches from a lying posture to a standing mode, it cannot move and needs to transition from standing to balanced standing before starting to move. Similarly, a moving quadrupedal robot cannot lie down directly but needs to switch to normal standing mode first and then to the lying posture.

Intelligent interaction

Intelligent interaction requires the avatar to perceive the physical environment, provide real-time contextual interpretation, and support multimodal interaction. We trained an image recognition model specific to attractions. To accurately identify landmark buildings from complex perspectives, this study fine-tunes an attraction-specific image recognition model based on the ResNet-50 pre-trained model.

We constructed an attraction dataset for a certain scenic area through aerial photography using a DJI NEO drone to fine-tune the model. The dataset contains 8,500 images, capturing aerial and horizontal views of attractions under different weather and lighting conditions. The dataset adopts a hierarchical classification system: attraction names serve as first-level labels, and ordered numbers for different sides of attractions act as second-level labels. The dataset includes 6 first-level attraction labels, 1 first-level negative sampling label (i.e., street view images without attractions), and 17 second-level labels. The model training process is as follows: first, adjust the ResNet-50 model structure by modifying the fully connected layer to reduce the number of output categories from 1,000 to 17; then, divide 70% of the dataset as the training set and the remaining 30% as the test set; finally, train the model and save models at different training stages.

The statistics in Table 1 illuminate that the pre-trained model without fine-tuning has almost no ability to recognize attractions. As the number of training epochs increases, the accuracy of the trained model in recognizing attractions significantly improves, with the accuracy of both first-level and second-level labels exceeding 99%, indicating that the model can accurately recognize attractions in most cases.

Table 1: ResNet-50 Accuracy across Epochs of Training

Model	First-Level Label Accuracy	Second-Level Label Accuracy
Pre-trained	3.99%	2.99%
1 epoch	76.04%	70.11%
3 epoch	97.30%	94.25%
20 epoch	99.82%	99.59%

Source: This study.

Deploying an LLM on the avatar enables intelligent interactions with users, such as attraction knowledge retrieval, interactive chatting, and personalized itinerary planning. We selected the Doubao LLM, due to ByteDance's Doubao intelligent agent featuring comprehensive tools, advanced technology, rapid product updates, and a large user base.

At first, we deployed Doubao LLM services, speech synthesis model services, and speech recognition model services on the Volcengine cloud. Then, we packaged the functional interfaces provided by Volcengine and integrated them into the edge layer server. Interface packaging requires user permission authentication and input data alignment, including image format interaction and commentary text cleaning. Specifically, the BGR channel order of images captured by the gimbal-mounted camera is rearranged to the RGB order required by the image recognition model; additionally, the output results of the large model are cleaned to avoid meaningless speech output caused by Markdown content.

After completing the training and deployment of the image recognition model and the deployment of the LLM and related functions, the intelligent guidance and interaction functions are mainly managed by the server deployed on the quadrupedal robot's development board. The server program implements a real-time request monitoring mechanism. When the client sends a service request, the development board server immediately triggers the response process: first, call the image recognition model to identify the attraction, generate attraction introduction text through the Doubao large model based on the model output results, and then convert it into an audio file. The development board transmits the audio file back to the client terminal via network communication protocols, and the Unity3D client APP or website performs speech decoding and audio playback operations to achieve real-time voice commentary on attractions. For dialogue functions, after users input voice in the Unity3D client APP or website, the voice file is transmitted to the edge side, which calls the packaged interface to upload it to the cloud-based speech recognition model, converts the voice into text, inputs it into the Doubao LLM to generate corresponding response text, and finally synthesizes the speech.

The Significance of The Avatar

GenAI has demonstrated transformative potential in the tourism industry due to its capabilities in contextual understanding, content generation, and predictive modeling. However, current applications of GenAI in tourism remain limited. Most implementations focus on the pre-trip phase, such as hotel bookings, itinerary planning, and customer service while applications in the en-route and post-trip stages remain underexplored. This study addresses this gap by introducing a LLM-based telerobotic avatar that enhances the en-route phase of tourism. By deploying the avatar at attractions, users can experience personalized and interactive virtual tours and remote visits.

Integrating LLMs with telerobotic systems addresses longstanding challenges in virtual tourism, particularly the generation of authentic, diverse, and context-aware content (Chamola *et al.*, 2024a). This design enhances accessibility for individuals unable to travel due to physical disabilities, health concerns, natural disasters, or geopolitical instability. Through VR devices, desktop platforms, or mobile applications, users can remotely control the avatar in real time. More importantly, they can interact with the LLM to receive personalized guidance, engage in natural language conversations, and dynamically plan itineraries, effectively overcoming the limitations of static content and rigid interaction formats that characterize most existing virtual tourism systems.

SUMMARY

Considering the emerging AI technologies such as GenAI, AgenticAI, and other cutting-edge systems driving the future of intelligence and automation, the tourism industry is undergoing a paradigm shift, including delivering highly personalized experiences, enhancing decision-making, and interpreting vast amounts of traveller feedback, which makes tourism more adaptive, engaging, and insight-driven (Wong *et al.*, 2023). Virtual tourism and remote visits have particularly benefited from the development of GenAI and other technologies. While prior tourism avatars allow for basic virtual tours, their inability to process real-time data or generate dynamic content significantly limits user engagement. In response, we design the Intelligent Travel Avatar, a telerobotic system enhanced with LLMs.

The avatar integrates physical environment perception, data processing, and multimodal interaction within a device-cloud collaboration architecture. Built on a quadrupedal robotic with a gimbal-mounted camera, the avatar supports both VR, PC and mobile teleoperation and telepresence. Leveraging the capability of LLMs, such as real-time content generation, speech

recognition, and speech synthesis, the system offers users personalized, interactive, and intelligent services. To the best of our knowledge, this is the first avatar system that integrates LLMs, robotics, and telerobotics for use in virtual tourism and remote visits. While we present several use-case scenarios, the potential for long-term application and continuous evolution remains significant. Theoretically, this work contributes to literature on technological innovations in virtual tourism and GenAI-enabled tourism systems. Practically, it offers a good example and guidance for the design of AI-powered tourism agents.

REFERENCES

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443.

Carvalho, I., & Ivanov, S. (2023). ChatGPT for tourism: Applications, benefits and risks. *Tourism Review*. https://doi.org/10.1108/tr-02-2023-0088

Chamola, V., Bansal, G., Das, T. K., Hassija, V., Sai, S., Wang, J., ... & Niyato, D. (2024). Beyond reality: The pivotal role of generative ai in the metaverse. *IEEE Internet of Things Magazine*, *7*(4), 126-135.

Chamola, V., Sai, S., Bhargava, A., Sahu, A., Jiang, W., Xiong, Z., ... & Hussain, A. (2024). A comprehensive survey on generative AI for metaverse: Enabling immersive experience. *Cognitive Computation*, *16*(6), 3286-3315.

Desnoyers-Stewart, J., Stepanova, E. R., Liu, P., Kitson, A., Pennefather, P. P., Ryzhov, V., & Riecke, B. E. (2023, April). Embodied telepresent connection (ETC): Exploring virtual social touch through pseudohaptics. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-7).

Dubourg E, Thouzeau V, Baumard N. A step-by-step method for cultural annotation by LLMS. *Frontiers in Artificial Intelligence*. 2024;7:1365508.

Gao, M., Liu, J., & Ye, Q. (2024). Empowering Innovation: An Empirical Study on the Impact of Generative AI on Online Freelancers' Performance. *ICIS 2024 Proceedings.* 41.

Hasan, M., Ozel, C., Potter, S., & Hoque, E. (2023). SAPIEN: Affective virtual agents powered by large language models. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*(pp. 1-3). ACIIW'11, Cambridge, MA, USA, September 10-13.

Hu, B., Shao, S., Cao, Z., Xiao, Q., Li, Q., & Ma, C. (2019). Learning a faster locomotion gait for a quadruped robot with model-free deep reinforcement learning. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (pp. 1097-1102). IEEE.

Hu, Y., Li, Z., Chen, Y., Cheng, Y., Cao, Z., & Liu, J. (2023). Content-aware adaptive device–cloud collaborative inference for object detection. *IEEE Internet of Things Journal*, *10*(21), 19087-19101.

Huang, Y. C., Backman, K. F., Backman, S. J., & Chang, L. L. (2016). Exploring the implications of virtual reality technology in tourism marketing: An integrated research framework. *International Journal of Tourism Research*, *18*(2), 116–128.

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., & Hutter, M. (2019). Learning agile and dynamic motor skills for legged robots. *Science Robotics*, *4*(26), eaau5872.

Jenelten, F., Hwangbo, J., Tresoldi, F., Bellicoso, C. D., & Hutter, M. (2019). Dynamic locomotion on slippery ground. *IEEE Robotics and Automation Letters*, *4*(4), 4170-4176.

Kikuchi, Y., Ojima, Y., Kato, R., Unno, M., Yem, V., Nagai, Y., & Ikei, Y. (2022). Dual robot avatar: Real-time multispace experience using telepresence robots and walk sensation feedback including viewpoint sharing for immersive virtual tours. In *ACM SIGGRAPH 2022 Emerging Technologies* (pp. 1-2).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer vision* (pp. 10012-10022).

Martínez-Romero, A., Quesada-Arencibia, A., Rodríguez-Rodríguez, J. C., Hernández-Sosa, J. D., García, C. R., & Moreno-Díaz, R. (2015). A robotic platform prototype for telepresence sessions. In *Computer Aided Systems Theory–EUROCAST 2015: 15th International Conference, Las Palmas de Gran Canaria, Spain, February 8-13, 2015*, Revised Selected Papers 15 (pp. 706-713). Springer International Publishing.

Momani, A. M., Alsakhnini, M., & Hanaysha, J. R. (2022). Emerging technologies and their impact on the future of the tourism and hospitality industry. *International Journal of Information Systems in the Service Sector (IJISSS)*, *14*(1), 1–18.

Pestek, A., & Sarvan, M. (2020). Virtual reality and modern tourism. *Journal of Tourism Futures*, *7*(2), 245–250.

Sharma, R., Kumar, A., & Chuah, C. (2021). Turning the Blackbox into a glass box: An explainable machine learning approach for understanding hospitality customer. *International Journal of Information Management Data Insights*, *1*(2), Article 100050.

Shin, H., & Kang, J. (2024). How does the metaverse travel experience influence virtual and actual travel behaviors? Focusing on the role of telepresence and avatar identification. *Journal of Hospitality and Tourism Management*, *58*, 174-183.

Tan, W., Fang, X., Zhang, W., Song, R., Chen, T., Zheng, Y., & Li, Y. (2021). A hierarchical framework for quadruped locomotion based on reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 8462-8468). IEEE.

Tian, W., Fan, M., Zeng, C., Liu, Y., He, D., & Zhang, Q. (2020). Telerobotic spinal surgery based on 5G network: The first 12 cases. *Neurospine*, *17*(1), 114.

Tsuchiya, K., & Koizumi, N. (2020, March). An optical design for avatar-user co-axial viewpoint telepresence. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 108-116). IEEE.

Tussyadiah, I. P., Wang, D., Jung, T. H., & tom Dieck, M. C. (2018). Virtual reality, presence, and attitude change: Empirical evidence from tourism. *Tourism Management*, *66*, 140–154.

Vaz, J. C., Kosanovic, N., & Oh, P. (2024). ART: Avatar Robotics Telepresence—The future of humanoid material handling loco-manipulation. *Intelligent Service Robotics*, *17*(2), 237-250.

Verma, S. (2022). Sentiment analysis of public services for smart society: Literature review and future research directions. *Government Information Quarterly*, *39*(3), 101708.

Viñals, M. J., Gilabert-Sansalvador, L., Sanasaryan, A., Teruel-Serrano, M. D., & Darés, M. (2021). Online synchronous model of interpretive sustainable guiding in heritage sites: The avatar tourist visit. *Sustainability*, *13*(13), 7179.

Wang, W., Kumar, N., Chen, J., Gong, Z., Kong, X., Wei, W., & Gao, H. (2020). Realizing the potential of the internet of things for smart tourism with 5G and AI. *IEEE Network*, *34*(6), 295-301.

Wong, I. A., Lian, Q. L., & Sun, D. (2023). Autonomous travel decision-making: An early glimpse into ChatGPT and generative AI. *Journal of Hospitality and Tourism Management*, *56*, 253-263.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., ... & Smola, A. (2022). Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2736-2746).