ITUPublications

International Telecommunication Union

Standardization Sector

ITU Focus Group Technical Specification

(12/2023)

ITU Focus Group on metaverse

Capabilities and requirements of generative artificial intelligence in metaverse applications and services

Working Group 2: Applications & Services

Technical Specification ITU FGMV-22

Capabilities and requirements of generative artificial intelligence in metaverse applications and services

Summary

As technology continues to evolve, there is an increasing demand for generative artificial intelligence (GAI) technology in the metaverse. GAI is crucial for creating immersive and interactive experiences in the metaverse. It has numerous capabilities in metaverse applications and services, from creating personalized avatars and environments to generating more immersive and personalized services. These capabilities can enrich the content of metaverse in more forms and significantly enhance the user experience within metaverse, providing a more engaging and immersive environment.

This Technical Specification describes the capabilities and requirements of GAI in metaverse applications and services. This document specifies four common capabilities of GAI in metaverse applications and services and analyses the description, assumption and service scenario. It also specifies the requirements of GAI in metaverse applications and services.

Keywords

Capability, generative artificial intelligence, metaverse, requirement.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1.0 of the ITU Technical Specification on "*Capabilities and requirements of generative artificial intelligence in metaverse applications and services*" approved at the 4th meeting of the ITU Focus Group on metaverse (ITU FG-MV), held on 4–7 December 2023 in Geneva, Switzerland.

Acknowledgements

This Technical Specification was researched and written by Qiuhong Zheng (China Telecommunications Corporation, China) and Liang Wang (ZTE Corporation, China) as a contribution to the ITU Focus Group on metaverse (FG-MV). The development of this document was coordinated by Yuntao Wang (CAICT, China) and Yuan Zhang (China Telecom), as FG-MV Working Group 2 Co-chairs, and by Qiuhong Zheng (China Telecom) as Chair of Task Group on Generative Artificial Intelligence in the metaverse.

Additional information and materials relating to this report can be found at: https://www.itu.int/go/fgmv. If you would like to provide any additional information, please contact Cristina Bueti at tsbfgmv@itu.int.

Editor & Task Group Chair:	Qiuhong Zheng China Telecommunications Corporation China	Tel: +86 17300137101 E-mail**:** zhengqh@chinatelecom.cn
Editor:	Liang Wang ZTE Corporation China	Tel: +86 25 88014641 E-mail: wang.liang12@zte.com.cn
WG2 Co-chair	Yuntao Wang CAICT China	E-mail: wangyuntao@caict.ac.cn
WG2 Co-chair	Yuan Zhang China Telecom China	E-mail: zhangy666@chinatelecom.cn

Table of contents

Technical Specification ITU FGMV-22

Capabilities and requirements of generative artificial intelligence in metaverse applications and services

1 Scope

This Technical Specification specifies the capabilities and requirements of Generative Artificial Intelligence (GAI) in metaverse applications and services.

The scope of this Technical Report includes:

– Overview of GAI

– Capabilities of GAI in metaverse applications and services

– Requirements of GAI in metaverse applications and services

2 References

None.

3 Definitions

3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

3.1.1 application [b-ITU-T Q.1741.7]: An application is a service enabler deployed by service providers, manufacturers or users. Individual applications will often be enablers for a wide range of services.

3.1.2 avatar [b-ISO/IEC 23005-4]: Entity that can be used as a (visual) representation of the user inside the virtual environments.

3.1.3 digital human [b-ITU-T F.748.15]: A computer application that integrates the technologies of computer graphics, computer vision, intelligent speech and natural language processing. It can be used for digital content generation and human-computer interaction to help improve content production efficiency and user experience.

3.1.4 service [b-ITU-T Y.2091]: A set of functions and facilities offered to a user by a provider.

3.1.5 text-to-speech synthesis (TTS) [b-ITU-T P.10]: A TTS process generates a speech signal from text codes. It is usually composed of two parts:

– A language-dependent text processing part (the high-level processing part), which generates from the character string (by reading rules, vocabulary and semantic analysis) a set of phonetic, prosodic, etc., parameters which are used by

  – an acoustical signal generating part, the synthesizer itself, which generates the audible speech.

3.2 Terms defined in this Technical Specification

This Technical Specification defines the following terms:

None.

4 Abbreviations and acronyms

This Technical Specification uses the following abbreviations and acronyms:

3D	Three-Dimensional
GAI	Generative Artificial Intelligence
NeRF	Neural Radiance Fields
NPC	Non-playable characters

5 Conventions

In this Technical Report:

– The keywords "**is required**" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this this Technical Report is to be claimed.

– The keywords "**is recommended**" indicate a requirement which is recommended but which is not absolutely required. Thus, this requirement needs not be present to claim conformance.

6 Overview of GAI

GAI refers to a broad field of research and development that focuses on creating intelligent systems that can generate new, original content, such as images, videos, music, text and even entire conversations. These systems use machine learning algorithms to learn patterns and structures within the data they are trained on, and then use this knowledge to generate new content that resembles the original data but is not necessarily identical to it.

The classification of GAI technology includes many aspects. In terms of the modality and domain of content generated by GAI technology, Table 1 lists the classification, subclass and use cases of GAI technologies.

Table 1 – Classification and use cases of GAI

GAI technical classification	Subclass	Use cases
Text generation	Non-interactive text generation	Structured writing such as news content generation Unstructured writing such as marketing text generation
	Interactive text generation	Generation of interactive content such as dialogues
Speech generation	Text-to-speech synthesis	Digital human speech synthesis
	Voice conversion	Voice cloning
	Music synthesis	Music composition, arrangement, performance and recording
Image generation	Image editing tools	Watermark removal Resolution improvement
	Image creating	Creative image generation such as painting Functional image generation such as posters and logos
Video generation	Video editing	Specific subject deletion Special effects generation
	Video clips	Detection and synthesis of specific fragments

Table 1 – Classification and use cases of GAI

GAI technical classification	Subclass	Use cases
	Video section editing	AI face swapping
3D model generation	3D digital human modelling	Digital human construction
	3D object modelling	Digital asset generation
	3D space modelling	Digital environment generation
Cross modal generation	Text generated image	Generate creative images based on text prompts
	Image generated video	Splicing images to generate videos
	Text generated video	Script based video synthesis
	Image/video generated text	Automatic subtitle generation
Strategy generation	Rule-based strategy generation	Strategy generation for simple issue
	Deep reinforcement learning-based strategy generation	Strategy generation for complex issue

7 Capabilities of GAI in metaverse applications and services

7.1 Personalized avatar creation

GAI has the capability of creating avatars that are unique and personalized based on a variety of factors such as user preferences, facial features and body type. This can help users feel more connected to their avatar and enhance their overall engagement with the metaverse.

7.1.1 Description

For the capability of personalized avatar creation, users can obtain their own exclusive avatars by providing audio and video data that can describe personal appearance characteristics and personal preference information to the terminal. Personalized avatar creation intelligently generates decisions by analysing the personalized information obtained from the user's provided audio and video data, combined with the user's personal preference information. It calls the personalized avatar creation model to perform high-fidelity image modelling based on the user's personalized information using a portrait generation model. It also utilizes a speech generation model to learn the specific avatar's speech characteristics based on the input voice features. The animation generation engine learns the potential mapping relationship between the user's voice, lip movements, facial expressions and body posture parameters, forming their respective driving models and driving methods. With sufficient driving key points and high-precision driving models, it can accurately restore subtle changes in the user's skeleton and muscles, obtaining realistic human body driving model parameters. Finally, through real-time rendering with the rendering engine, it is presented to the user via the terminal. Figure 1 shows the concept of the personalized avatar creation capability.

Personalized avatar creation can help users generate avatars that match their own appearance characteristics and personal preferences, allowing users to have a closer connection with their avatars in the metaverse. This enhances the overall engagement of users in the metaverse.

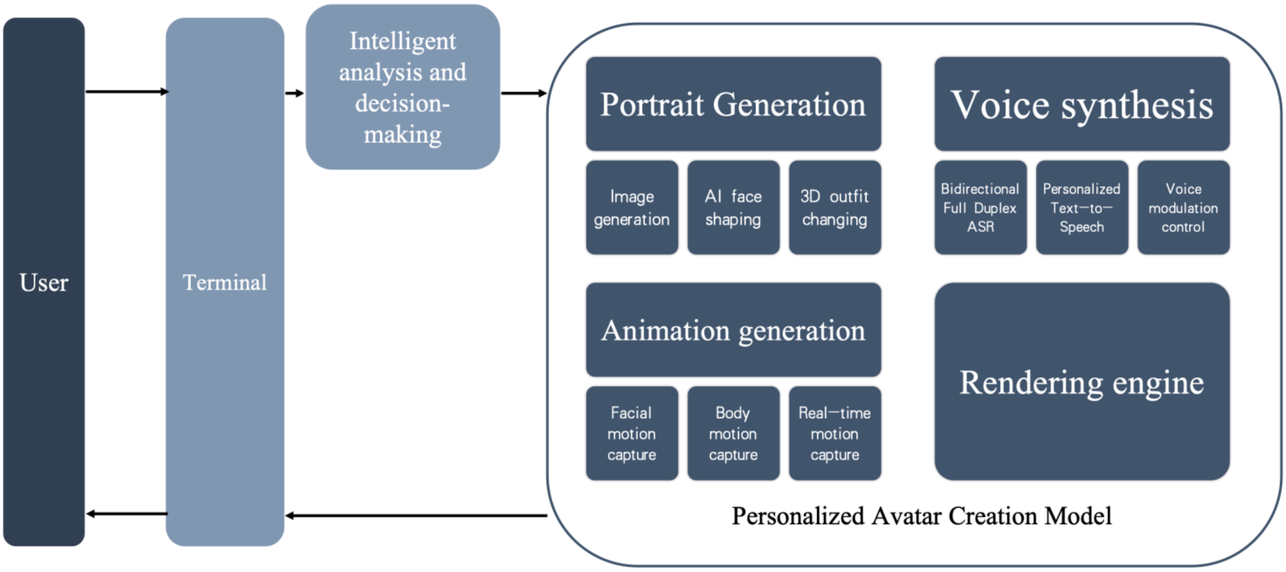

Figure 1 – Concept of the personalized avatar creation capability

7.1.2 Assumptions

The assumptions related to this capability include the following:

- That users can provide accurate audio and video data describing their personal appearance characteristics.
- That users can selectively provide preference information or specific suggestions for personalized avatar.
- That users can have smooth and effective interaction with the terminal.
- That the terminal has the capability to display avatars and present the personalized avatar to the user.

7.1.3 Service scenario

Figure 2 describes a sample service flow for personalized avatar creation.

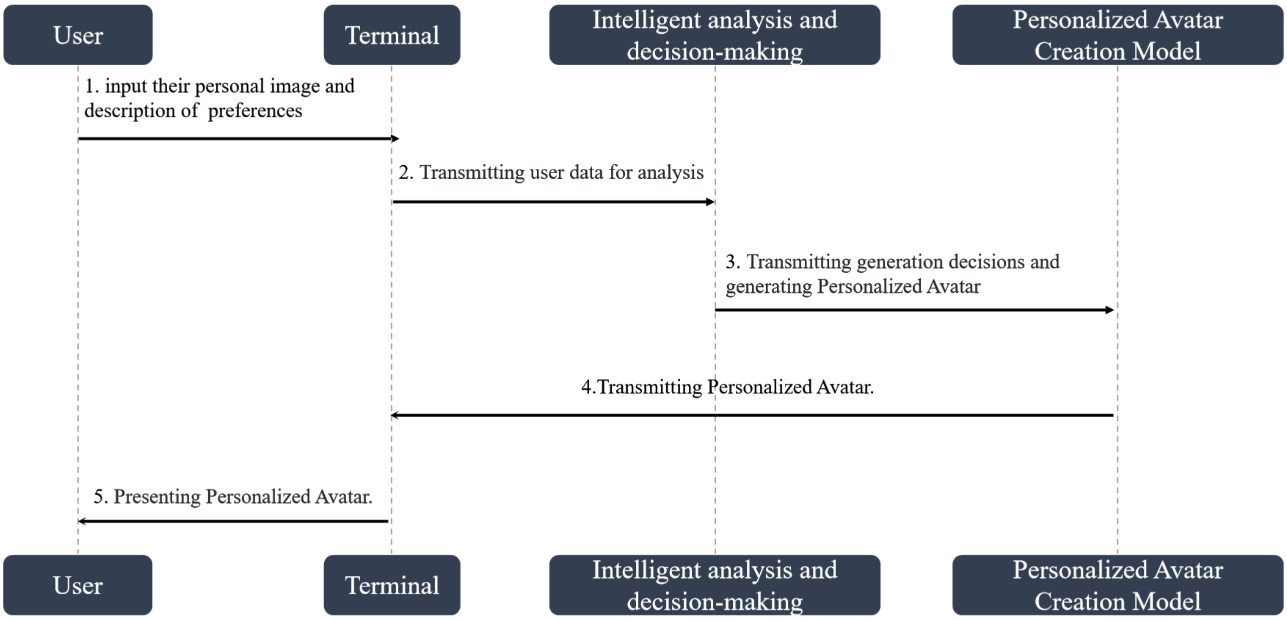
User
Terminal
Intelligent analysis and decision-making
Personalized Avatar Creation Model
1. input their personal image and description of preferences
2. Transmitting user data for analysis
3. Transmitting generation decisions and generating Personalized Avatar
4.Transmitting Personalized Avatar.
5. Presenting Personalized Avatar.
User
Terminal
Intelligent analysis and decision-making
Personalized Avatar Creation Model

Figure 2 – Service flows for the personalized avatar creation capability

1. The user interacts through the terminal and inputs audiovisual data that can describe their personal image features as well as personal preferences.
2. The terminal sends the input to the intelligent analysis and decision-making module, which analyses the user's description and generates decisions for personalized avatar creation.
3. The decisions are sent to the personalized avatar creation model for generating a personalized avatar.
4. The personalized avatar creation model utilizes four engines: the portrait generation engine, voice generation engine, animation generation engine and rendering engine, to generate the personalized avatar. The generated avatar is then transmitted to the terminal.
5. The personalized avatar created by the personalized avatar creation model is presented to the user on the terminal.

7.2 Dynamic environment generation

GAI has the capability of generating realistic and diverse environments within the metaverse. For example, it can create different types of terrain, weather patterns and lighting conditions to simulate real-world environments. This can help create a more immersive experience for users.

7.2.1 Description

Dynamic environment generation utilizes technologies such as three-dimensional (3D) model generation to create diverse dynamic environments for users, providing them with immersive experiences that include realistic weather changes and diverse terrain variations. Figure 3 shows the concept of the dynamic environment generation capability.

Dynamic environment consists of four main aspects:

1. Layout generation: Using 3D model generation technology, a diffusion model is trained to learn real-world outdoor scene layouts, such as terrains and road networks. The GAI model can quickly generate realistic and diverse outdoor scene layouts based on user suggestions. It also allows real-time modifications for generating various rich variations. Users can fine-tune the generated results to achieve scene layouts that better meet their requirements.
2. Architectural generation: Building upon the generated scene layout, this module generates diverse and realistic architectures by leveraging 3D object modelling technology to learn patterns from a large datasets of real-world building data. It applies GAI techniques to give unique appearances to architectural designs and adds details such as windows and balconies to the buildings.
3. Indoor mapping generation: This module utilizes photos of real-world houses as training data to train a neural radiance fields (NeRF) model. Based on the colour and depth maps generated by NeRF from given viewpoints, it creates pseudo-3D mapping materials to achieve a sense of depth and reconstruct indoor scenes. Finally, with procedural UV calculation, these pseudo-3D rooms are filled into the previously generated building exteriors.
4. Dynamic element: The core elements required for realistic virtual scenes, such as diverse layouts and various architectural styles, are already in place. In this module, dynamic elements are added using procedural generation. Standard elements such as roads and vegetation are generated, and then all elements are transferred to the engine to introduce dynamic changes such as weather, traffic flow and pedestrians.

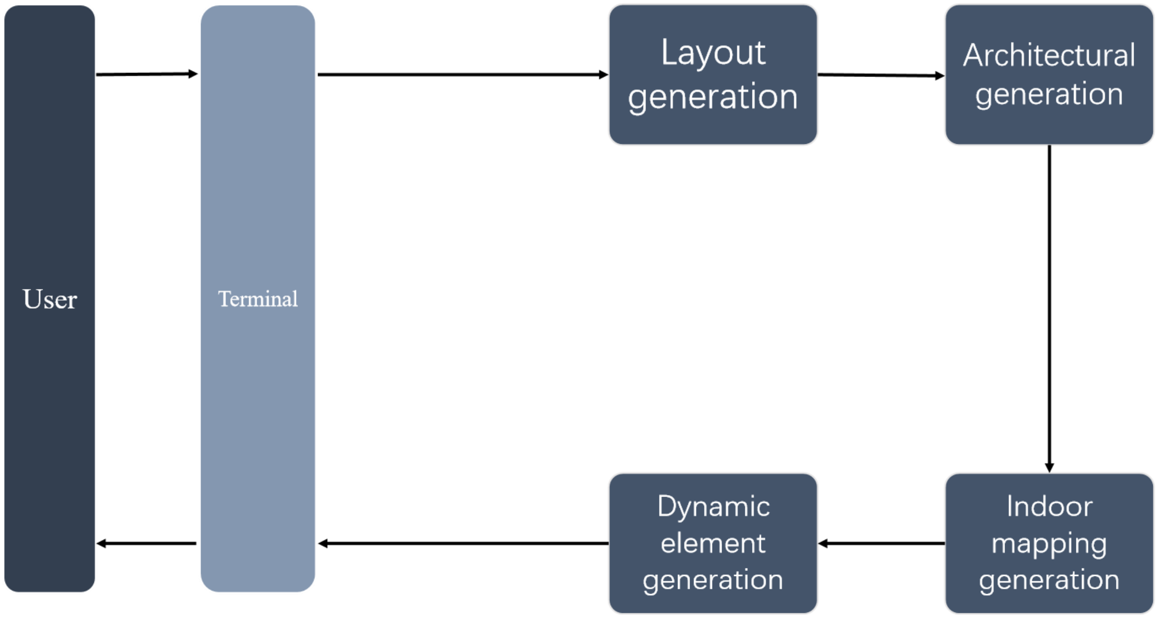

Figure 3 – Concept of the dynamic environment generation capability

7.2.2 Assumptions

The assumptions related to this capability include the following:

- That the system has sufficient computing resources with high-performance hardware and ample memory to handle real-time complex computations and simulations.
- That there is a comprehensive dataset containing real-world terrain, road networks, building layouts and indoor mappings available for effective model training.
- That users are able to provide suggestions regarding desired scene layouts, architectural styles and environmental elements. Users should be able to effectively communicate their preferences to generate customized and satisfactory dynamic environments.
- That the dynamic environment generation system is compatible with the target platform or application. It should seamlessly integrate with existing software tools, game engines or virtual reality platforms to easily incorporate the generated environments into the desired context.
- That real-time rendering techniques are available to display the dynamic environment with smooth animations, realistic lighting effects and high visual fidelity.

7.2.3 Service scenario

Figure 4 shows a sample service flow for dynamic environment generation.

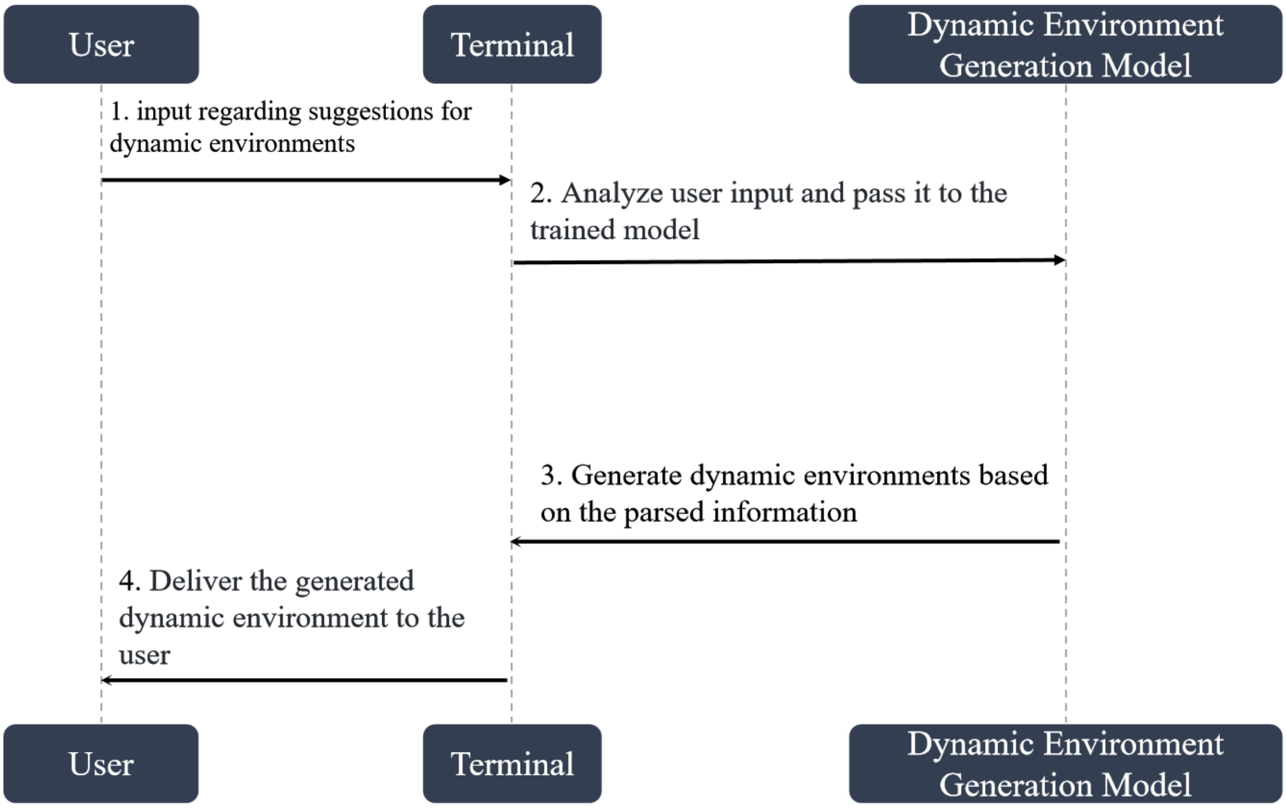
User
Terminal
Dynamic Environment Generation Model
1. input regarding suggestions for dynamic environments
2. Analyze user input and pass it to the trained model
3. Generate dynamic environments based on the parsed information
4. Deliver the generated dynamic environment to the user
User
Terminal
Dynamic Environment Generation Model

Figure 4 – Service flows for the dynamic environment generation capability

1. User suggestions for dynamic environment, such as scene layout, architectural style and environmental elements, and even specific requirements or limitations for the dynamic environment are made.
2. The system parses and analyses the user input, including scene layout, architectural style and mapping of environmental elements, extracting key information and transforming it into a format understandable by underlying machine learning models.
3. The models generate the dynamic environment based on the parsed information. 3D model generation technology is utilized to predict and generate the desired environment based on the given input. These models consist of components for terrain generation, road network layout, building placement, interior design and other environmental factors.
4. The generated dynamic environment is delivered to the user, who can explore and interact with the generated dynamic environment.

7.3 Immersive interaction

GAI can generate non-playable characters (NPC) within the metaverse that can interact with users based on predefined rules or AI algorithms. These NPC can provide personalized experiences based on user behaviour and preferences. Moreover, text generation technology can enhance natural language processing capabilities within the metaverse. This can enable users to communicate more effectively with NPC and other users, and allow for more complex interactions within the virtual world.

7.3.1 Description

Immersive interaction allows the generation of NPC with human-like interactive capabilities using GAI technology. NPC generated by immersive interaction can engage in natural, smooth and interesting voice, text and gesture interactions with users, thereby enhancing the user's immersion and engagement in the virtual world. Immersive interaction not only provides NPC with rich knowledge and expertise but also enables NPC to respond to user actions through devices such as a mouse or controller or by inputs of voice and text, based on their own motivations.

Figure 5 shows the concept of the immersive interaction capability. A brief summary to the interaction mechanism of immersive interaction would be as follows: Firstly, immersive interaction generates NPCs based on the NPC knowledge base using the NPC generation model, where users can supplement their own envisioned NPC settings to create more personalized NPCs that align with their preferences. After generating the NPC, immersive interaction utilizes the NPC interaction model to generate appropriate interactive behaviours in response to user inputs of text, voice and human-computer interaction signals provided through devices such as a mouse or controller. The NPC outputs textual, vocal and expressive actions to complete the immersive interaction between the user and the NPC. This ensures that the interaction output from the NPC is not rigid but offers users a personalized experience tailored to their preferences.

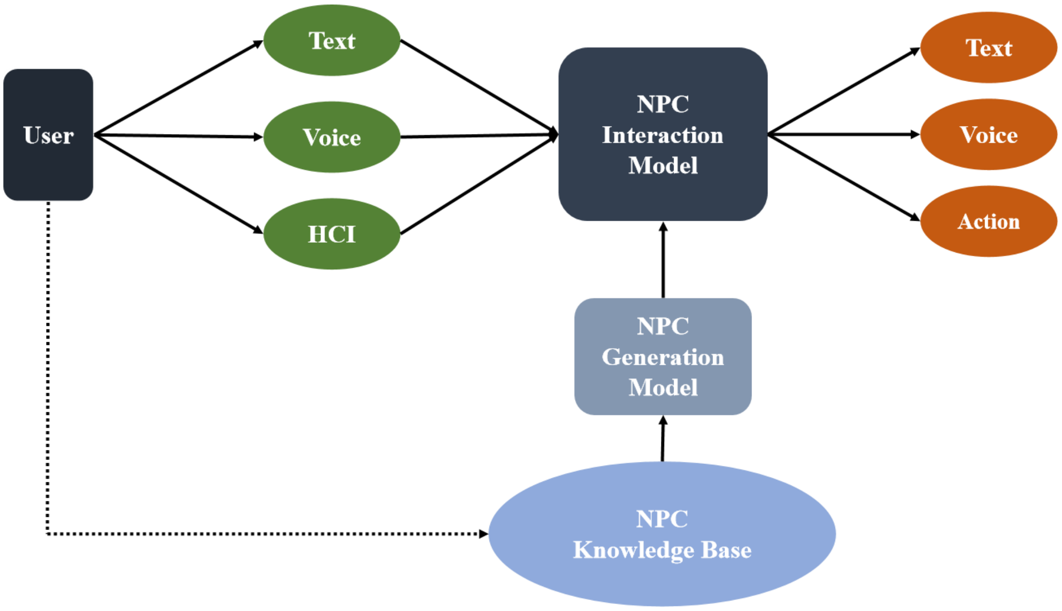

Figure 5 – Concept of the immersive interaction capability

7.3.2 Assumptions

The assumptions related to this capability include the following:

– The existence of a comprehensive NPC knowledge base containing NPC attributes, background knowledge and behavioural information, serving as the foundation for generating NPC.

– That users can specify the desired characteristics and attributes of the NPCs. These settings can include personality traits, character knowledge and background to shape the generated NPC.

– That users can provide inputs for human–computer interaction through text, voice or the use of devices such as a mouse or controller.

– That the immersive interaction system can generate NPC responses in real-time or near real-time to ensure prompt responses to user inputs and provide a seamless interactive experience.

7.3.3 Service scenario

Figure 6 shows a sample service flow for immersive interaction.

1. The user inputs NPC settings to generate an NPC. The user supplements their envisioned NPC settings to the NPC character repository, and the NPC generation model analyses the knowledge in the NPC character repository to generate the NPC.

2. The NPC generation model outputs NPC data, and after generating the NPC, the NPC generation model transfers the data to the NPC interaction model for loading.
3. The user provides interaction information, such as text, voice or human–computer interaction to the NPC interaction model.
4. The NPC interaction model outputs interaction information. It comprehends the user's text, voice, human–computer interaction and other interaction information. Based on the loaded NPC data, it intelligently generates appropriate responses and behaviours for the NPC, ensuring a dynamic and engaging interactive experience.

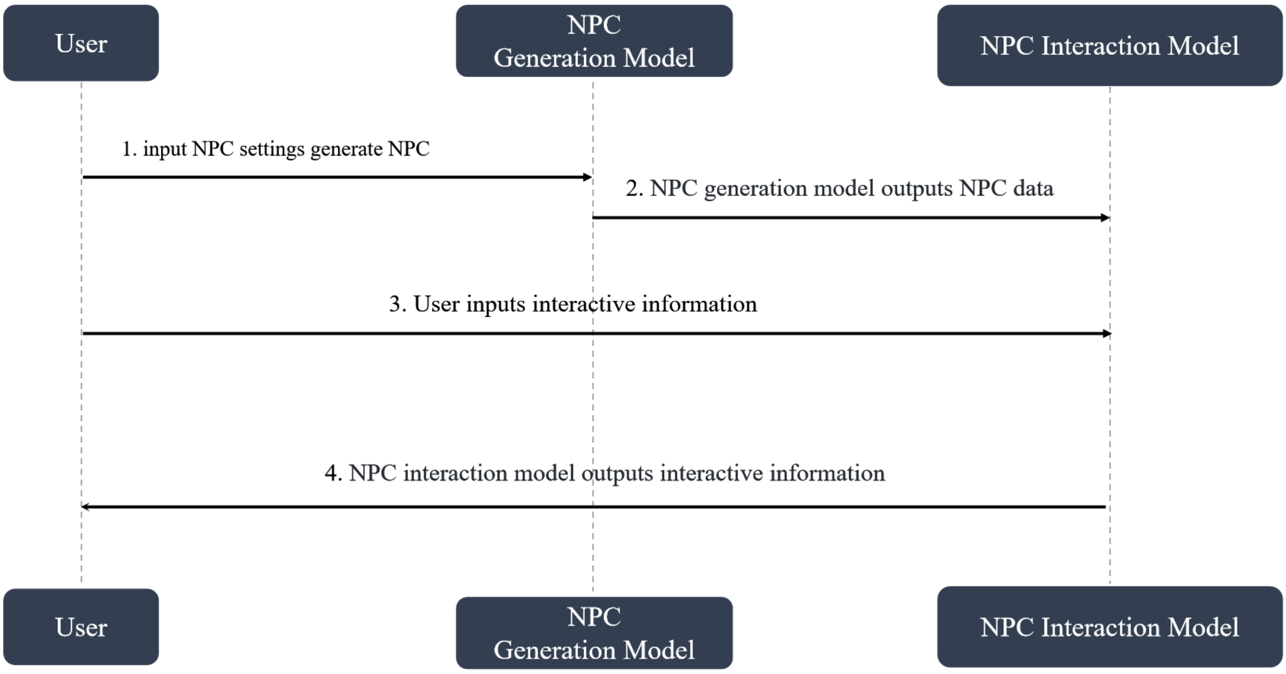
User
NPC Generation Model
NPC Interaction Model
1. input NPC settings generate NPC
2. NPC generation model outputs NPC data
3. User inputs interactive information
4. NPC interaction model outputs interactive information
User
NPC Generation Model
NPC Interaction Model

Figure 6 – Service flows for the immersive interaction

7.4 Personalization

GAI can personalize the user experience by generating content that is tailored to each user's preferences. For example, it can suggest specific quests or missions based on past activity or generate personalized rewards for completing certain tasks.

7.4.1 Description

Personalization can generate personalized tasks in real-time based on user preferences, either by suggesting existing tasks or creating new personalized tasks. Personalization leverages GAI technology to analyse users' multimodal inputs such as natural language, audio and video to understand their personalized needs before recommending existing tasks or generating new personalized tasks. Figure 7 shows the concept of the personalization capability. The entire personalization system consists of components such as the user interaction database, user historical interaction data, preference analysis, personalization and task repository. The user interaction database integrates real-time user inputs, as well as historical multimodal interaction data such as audio, images and video, which are then fed into preference analysis for user preference analysis to guide the personalization process of task generation. Personalization recommends specific tasks or generates personalized tasks based on user preferences and the existing task repository, and stores the generated personalized tasks in the task repository. The preference analysis module utilizes GAI technology to customize exclusive image descriptions for users, such as "loving adventure" or "pursuing excitement", so that the system's understanding of users is not the same for everyone. The personalization module utilizes rule-based strategy generation technology to generate tasks exclusive to each user, enhancing their immersive experience.

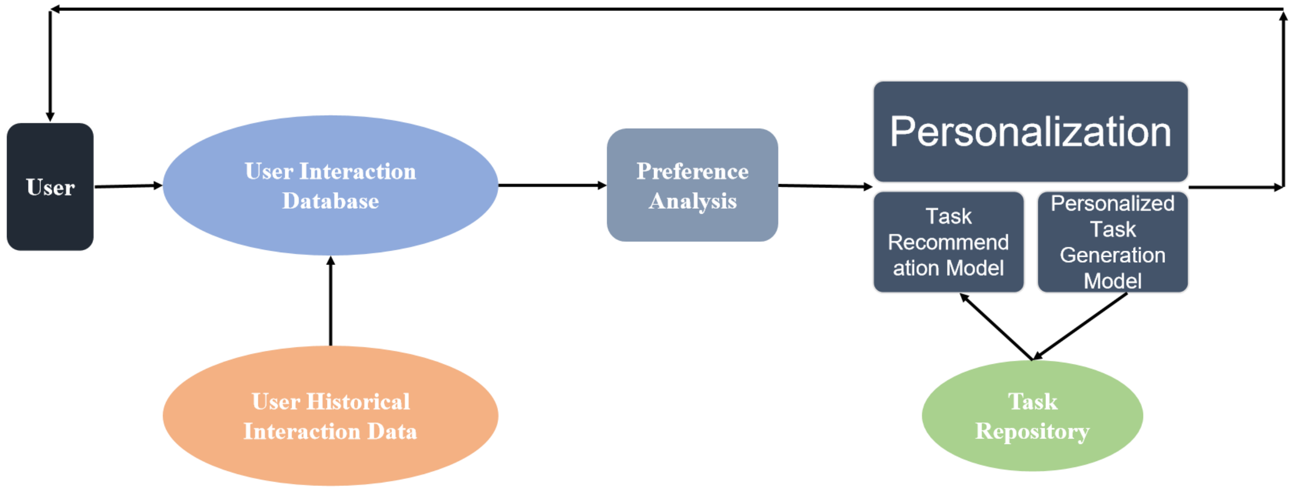
User
User Interaction Database
Preference Analysis
Personalization
Task Recommend ation Model
Personalized Task Generation Model
User Historical Interaction Data
Task Repository

Figure 7 – Concept of the personalization capability

7.4.2 Assumptions

The assumptions related to this capability include the following:

- The existence of a reliable source or database that can collect and store user interaction data in various formats, including natural language, audio, images and videos.
- The existance of a significant amount of historical user interaction data available for training the preference analysis component.
- The presence of a comprehensive task repository that can recommend existing tasks or serve as a foundation for generating personalized tasks. This task repository should encompass various types of tasks and have relevant task information available for recommendation.
- That the task repository has a reliable storage system for storing and retrieving personalized tasks generated by the personalization component.

7.4.3 Service scenario

Figure 8 describes a sample service flow for personalization.

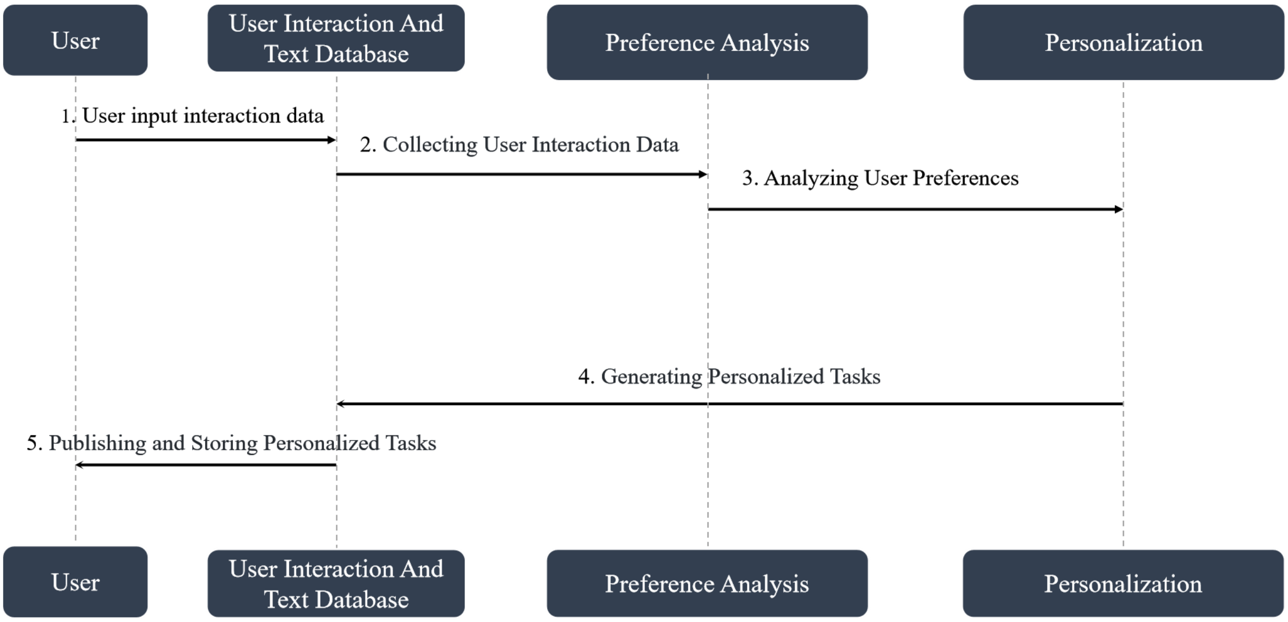
User
User Interaction And Text Database
Preference Analysis
Personalization
1. User input interaction data
2. Collecting User Interaction Data
3. Analyzing User Preferences
4. Generating Personalized Tasks
5. Publishing and Storing Personalized Tasks
User
User Interaction And Text Database
Preference Analysis
Personalization

Figure 8 – Service flows for personalization

1. User input interaction data: Users input interaction data such as audio, images and natural language into the user interaction database.
2. Data collection: The system collects and stores the user's interaction data in the user interaction database. This includes storing the user's current inputs as well as historical interaction data.
3. Analysing user preferences: The preference analysis component extracts key information from the user's interaction data to identify their interests, needs and preferences. It analyses the user's historical interaction data to understand their preferences.
4. Generating personalized tasks: Based on the results of preference analysis and the available task repository, the personalization component recommends existing tasks or generates personalized tasks. It considers user preferences, task relevance and other factors to customize tasks for specific user needs.
5. Publishing and storing personalized tasks: The generated personalized tasks are published to the user and stored in the task repository for future use. Ensuring the ease of retrieval, they can be recommended to the user in subsequent interactions.

8 Requirements of GAI in metaverse applications and services

8.1 Requirements of personalized avatar creation

a) Terminal:

- It is required to support the capture of user audio and video data for supporting personalized avatar creation.
- It is recommended for the terminal to have an intuitive interface for data collection.
- It can optionally provide real-time feedback during the avatar creation process.

b) Intelligent analysis and decision-making:

- It is required to be able to analyse and interpret user input data in order to extract information and respond accordingly.
- It is recommended to use advanced deep reinforcement learning algorithm for accurate decision-making to improve the quality of decisions.

- It can optionally utilize machine learning techniques to enhance the quality of decisions.

c) Personalized avatar creation model:

- It is required to integrate avatar image generation, speech generation, animation generation and rendering engines.
- It is recommended for the generated avatars to have a high degree of match with the user's input data.
- It can optionally provide customization options to adjust the features of appearance and sound of the avatar.

8.2 Requirements of dynamic environment generation

a) Layout generation:

- It is required to train a GAI diffusion model using real-world data to learn outdoor scene layouts such as terrain and road networks.
- It is required to generate diverse and realistic outdoor scene layouts that match user input and suggestions.
- It is recommended to support real-time modification of the layouts, version management and backup.
- It is recommended to allow users to adjust the generated layouts to better match their specific requirements.

b) Building generation:

- It is required to learn patterns from a large datasets of real-world building data using 3D object modelling technology.
- It is required to generate diverse and realistic building layouts based on the generated scene layout.
- It is recommended to provide unique appearances to the buildings using GAI technology and to add details such as windows and balconies.

c) Indoor mapping generation:

- It is required to train an AI model (such as NeRF) using real-world house photos to enable mapping and reconstruction of indoor scenes.
- It is recommended to create pseudo-3D effects by generating colour and depth maps based on a given viewpoint.
- It is recommended to fill the previously generated building walls with these pseudo-3D effects.

d) Dynamic element:

- It is required to generate standard elements such as roads and vegetation using procedural generation techniques.
- It is recommended to integrate all the generated elements into an engine and include dynamic variations like weather, traffic and pedestrians.

8.3 Requirements of immersive interaction

a) NPC knowledge base service:

- It is required to provide an NPC priori knowledge base, including knowledge in various relevant professional domains and general knowledge.

- It is recommended that the NPC knowledge base allows users to expand and customize it according to their personalized needs to generate more personalized NPC that better meet the user's expectations.

b) NPC generation model service:

- It is required to generate NPCs that have natural and realistic appearance and behavioural characteristics.
- It is recommended to support users in enhancing the personalized capability of the NPC generation model by supplementing their own predefined NPC settings.

c) NPC interaction model service:

- It is required to be able to generate interactive behaviours based on user inputs such as text and voice information, as well as human–computer interaction signals provided through devices such as a mouse or controller.
- It is recommended to be able to understand context and intent, and generate accurate and interesting outputs in terms of text, speech and facial expressions.

d) Immersive interaction service:

- It is required to provide natural, smooth and engaging voice, text and gesture interactions that allow users to have an immersive experience interacting with NPCs.
- It is recommended to enhance the user's engagement and immersion in the virtual world, enabling them to feel the pleasure and enjoyment of genuine interaction with NPCs.

8.4 Requirements of personalization

a) User interaction:

- It is required to support users to interact with the system through natural language, audio, images or videos.
- It is required to be able to collect and store user interaction data, including the user's current input and historical interaction data.

b) Task repository:

- It is required to be able to store task descriptions, tags and relevance ratings as metadata.
- It is required to support task retrieval and recommendation based on user preferences and task relevance.

c) Preference analysis component:

- It is required to be able to analyse the user's historical interaction data and extract information such as their interests, needs and preferences.

d) Personalization component:

- It is required to be able to generate personalized tasks based on user preferences and tasks in the task repository.
- It is required to be stored in the task repository for future use.

e) Task recommendation:

- It is required to be able to recommend personalized tasks suitable for the user based on their current input and past interactions, and to make adjustments to the tasks according to changes in user input information.
- It is required to be based on user preferences and task relevance.

f) User privacy protection:

- It is required to protect user’s personal data, in compliance with applicable privacy policies and regulations.

g) Scalability:

- It is required to be able to support multiple users and multiple tasks, and effectively handle high concurrency situations.

Bibliography

[b-ITU-T F.748.15]	Recommendation ITU-T F.748.15 (2022), *Framework and metrics for digital human application systems.*
[b-ITU-T P.10]	Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience.*
[b-ITU-T Q.1741.7]	Recommendation ITU-T Q.1741.7 (2011), *IMT-2000 references to Release 9 of GSM-evolved UMTS core network*.
[b-ITU-T Y.2091]	Recommendation ITU-T Y.2091 (2011), *Terms and definitions for next generation networks*.
[b-ISO/IEC 23005-4]	ISO/IEC 23005-4:2018, *Information technology – Media context and control – Part 4: Virtual world object characteristics*.