GPT-VR Nexus: ChatGPT-Powered Immersive Virtual Reality Experience

Jiangong Chen† Tian Lan‡ Bin Li†
†Department of Electrical Engineering, The Pennsylvania State University, Pennsylvania, USA
‡Department of Electrical and Computer Engineering, George Washington University, Washington, USA
{jiangong, binli}@psu.edu, tlan@gwu.edu

***Abstract*—The fusion of generative Artificial Intelligence (AI) like ChatGPT and Virtual Reality (VR) can unlock new capabilities to interact with VR environments through natural language, e.g., automatically generating and animating 3D scenes using only audio input. However, significant gaps exist in supporting this vision: 1) limited AI data processing ability for accurate VR context comprehension; 2) AI "hallucinations" that leads to misaligned responses in VR; and 3) the absence of tools for directly translating AI's responses into VR scene creation and animated interactions. To address these challenges, we introduce GPT-VR Nexus, a novel framework creating a truly immersive VR experience driven by an underlying generative AI engine. It employs a two-step prompt strategy and robust post-processing procedures, without the need of fine-tuning the complex AI model. Our experimental results show quick responses of the VR environment to a diverse range of user audio requests/inputs in merely a few seconds.**

***Index Terms*—Virtual Reality, Generative AI, ChatGPT**

I. Introduction

The advent of generative Artificial Intelligence (AI) such as ChatGPT has ignited a wave of exploration into its diverse application scenarios, such as content creation and software development. The fusion of Virtual Reality (VR) with generative AI has the potential to extend user interaction beyond the conventional realms of text, vision, and voice, thus creating a truly immersive experience. However, enabling this vision of the GPT-VR nexus must address several key challenges: 1) the need to improve generative AI's contextual understanding of user requests/inputs in VR settings, by effectively utilizing the vast amounts of VR data to generate relevant and accurate responses; 2) the hallucination problem (see [1]) due to possibly inapplicable responses from generative AI, violating physical constraints and significantly degrading user experience; and 3) the lack of tools mapping generative AI's responses directly to drive VR scene creation and animated interactions.

In this demo, we present a novel GPT-VR nexus to bridge the gap between generative AI and the VR environment. It enables a ChatGPT-powered VR experience – e.g., automated generation of 3D scenes and interaction with VR objects – from VR user audio inputs/commands. In particular, we propose a two-step strategy to process VR contextual data. It first categorizes user requests/inputs and then queries relevant data for precise prompts. To achieve the most relevant response, we develop an additional processing layer for response validation and adjustment. It creates VR environment/scene and animated interaction directly from ChatGPT responses. By showcasing the novel capabilities by integrating advanced generative AIs like ChatGPT into VR, this demo underscores the immense potential of building a GPT-VR nexus in creating novel interactive experiences, as well as significantly reducing content/scene development costs.

II. System Architecture

Fig. 1 depicts the system architecture of GPT-VR Nexus. The user-initiated interaction begins with an audio recording, activated via the VR controllers. This audio input is first transcribed into text using OpenAI's Whisper model, which is then combined with a custom-designed prompt and relayed to the GPT-4 Turbo model. Upon receiving the processed response from the server, the Nexus undertakes different strategies based on the categorization of the response. For responses that are in plain text, the Nexus employs OpenAI's text-to-speech (TTS) model to convert the text responses into voice outputs, which are then emitted from the virtual avatar's audio source within the VR environment, thereby simulating a natural human conversation. For responses that involve more complex commands, the Nexus parses these commands and distributes them to various specialized modules within the system. Note that generating an entire response at once can be time-consuming; hence, we opt for delivering the response in smaller, manageable chunks. These chunks are subsequently merged into coherent sentences for further processing. In the next few sections, we will introduce the core design of GPT-VR Nexus to deal with the unique challenges.

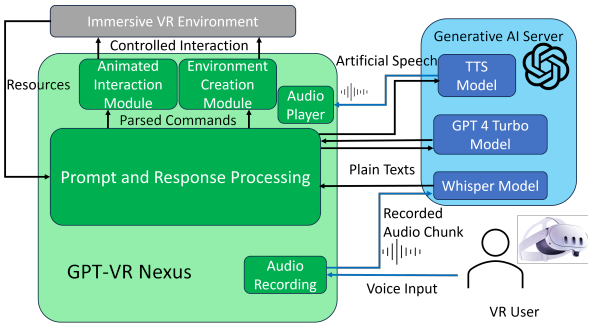

Fig. 1: System architecture.

Prompt and Response Processing. To effectively utilize VR contextual data for user requests, GPT-VR Nexus employs