Survey of Large Language Models in Extended Reality: Technical Paradigms and Application Frontiers

Jingyan Wang[1], Yang Zhao[1], Haotian Mao[1], and Xubo Yang *[1]

[1] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China*

Received: 4th August 2025 Accepted: Date Month 2025

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, and their integration with Extended Reality (XR) is poised to transform how users interact with immersive environments. This survey provides a comprehensive review of recent developments at the intersection of LLMs and XR, offering a structured organization of research along both technical and application dimensions. We propose a taxonomy of LLM-enhanced XR systems centered on key technical paradigms—such as interactive agent control, XR development toolkits, and generative scene synthesis—and discuss how these paradigms enable novel capabilities in XR. In parallel, we examine how LLM-driven techniques support practical XR applications across diverse domains, including immersive education, clinical healthcare, and industrial manufacturing. By connecting these technical paradigms with application frontiers, our survey highlights current trends, delineates design considerations, and identifies open challenges in building LLM-augmented XR systems. This work provides insights that can guide researchers and practitioners in advancing the state of the art in intelligent XR experiences.

1 Introduction

1.1 Background and Motivation

Extended Reality (XR), encompassing Virtual, Augmented, and Mixed Reality, has seen rapid advancement across domains such as education, healthcare, design, and entertainment. However, traditional interaction paradigms based on fixed scripts, manual triggers, and limited dialogue trees fall short in delivering adaptive, intelligent, and human-like interactions. Meanwhile, Large Language Models (LLMs), such as GPT-4, Gemini, and LLaMA, have revolutionized the field of natural language understanding and generation, with recent progress extending into multimodal and embodied intelligence.

The integration of LLMs into XR systems promises to overcome key limitations in speech recognition, contextual memory, content authoring, and real-time adaptation. By enabling semantic understanding of text, voice, and visual input, LLMs act as the cognitive core of next-generation XR agents. This motivates a systematic survey of how LLMs are being embedded, evaluated, and applied within XR ecosystems.

*Corresponding author: yangxubo@sjtu.edu.com

1.2 Contributions and Scope of the Survey

This work provides the first systematized, taxonomy-based review of LLM applications in XR. Our contributions are threefold:

- We propose a comprehensive taxonomy for LLM-enhanced XR systems, categorized along three axes: input modality (text/speech, 3D scenes/images, sensor signals), technical paradigm (e.g., generative AI, embodied AI, toolkits), and application domain (e.g., healthcare, education, social interaction).
- We analyze papers from top-tier conferences and preprint archives published between 2019 and 2025, identifying seminal contributions, emerging trends, and notable use cases.
- We synthesize major challenges and future directions, offering guidance for researchers and developers seeking to build next-generation LLM-XR systems.

1.3 Survey Methodology

Following PRISMA principles[1], we conducted a multi-stage literature review using databases such as Google Scholar, ACM Digital Library, IEEE Xplore, and arXiv. We used keyword combinations like *LLM + XR*, *embodied agent*, *3D scene understanding*, and *multimodal interaction*. The inclusion criteria required that works explicitly integrate LLMs with XR technologies and contribute technical insights, evaluate use cases, or present system-level frameworks. Each paper was annotated with metadata on input type, model used, modality fusion, system scope, and evaluated capability.

2 Human-Centered XR Intelligence: Capabilities from LLMs

2.1 Speech and NLP for XR

Speech interaction technology constitutes a pivotal force in the evolution of human-computer interaction paradigms within Extended Reality (XR) systems, playing a fundamental role in mitigating cognitive load and enhancing interaction efficiency. Traditional XR speech interaction methodologies, which predominantly rely on constrained command sets or text-based Natural Language Processing (NLP) techniques, exhibit significant limitations in accommodating complex multimodal and dynamic interactions. A primary technical challenge in XR speech interaction arises from the fragmentation between Automatic Speech Recognition (ASR) and NLP pipelines, resulting in the loss of implicit semantic information and thereby diminishing the accuracy of natural language instruction conversion into executable commands. Furthermore, XR speech interaction systems must process intricate multi-step instructions involving logical dependencies and ambiguous references, further complicating the task of semantic parsing. Additionally, given that XR environments often necessitate prolonged and continuous dialogues, conventional NLP models struggle to maintain conversational coherence due to constraints in contextual memory.

Recent advancements in Large Language Models (LLMs) offer transformative solutions to these challenges by integrating speech processing, semantic comprehension, and long-term contextual retention within a unified framework. Unlike conventional ASR-to-text pipelines, modern LLMs, such as the multimodal GPT-4o, possess the capability to process speech inputs directly, thereby preserving tonal and contextual cues that significantly

enhance semantic understanding. Moreover, LLMs demonstrate superior performance in parsing complex multi-step instructions while maintaining logical coherence, effectively interpreting commands such as "First adjust the lighting, then move the object, and finally save the current scene." Additionally, LLMs dynamically model the XR workspace, enabling them to accurately resolve ambiguous references, such as "Move that object over there". Furthermore, LLMs extend conversational memory beyond the limitations of traditional models, thereby ensuring continuity in multi-turn dialogues. By enhancing context tracking, users can seamlessly reference prior discussions without the need to reiterate details, thereby substantially improving the fluidity and intuitiveness of XR speech interactions. Through addressing these critical challenges, LLMs establish the foundation for more immersive, intelligent, and naturalistic XR speech interactions.

To quantitatively assess the performance of different large language models in speech interaction, we leverage the official evaluation results from VoiceBench, a state-of-the-art benchmarking framework for speech understanding and interaction. A comparative analysis of mainstream speech- and text-based large language models is presented in Table 1.

2.2 Vision-Language Models (VLMs) for XR

In extended reality (XR) systems, the deep integration of visual perception and language comprehension is crucial for enhancing interactive experiences. Traditional XR visual interactions primarily rely on computer vision models for object detection and scene parsing, while language interactions depend on natural language processing techniques. However, this separate processing paradigm often leads to information fragmentation, making it challenging to meet the demands of complex and dynamic interactions. In recent years, the emergence of vision-language models (VLMs) has enabled systems to perform advanced semantic reasoning based on visual inputs and engage in efficient interactions through natural language. This advancement significantly enhances environmental perception, intelligent decision-making, and multimodal information fusion, providing critical support for XR's vision-language integration.

The primary challenges faced by XR visual interactions include: (1) the disconnection between traditional vision and language models, which hinders effective information integration; (2) insufficient capabilities in complex scene understanding and reasoning, making it difficult for systems to accurately perceive and infer causal relationships in dynamic, multi-object environments; and (3) constraints related to real-time performance and multimodal information fusion, which limit the responsiveness of XR devices in resource-constrained settings. For instance, in XR design software, conventional approaches struggle with precisely interpreting user instructions, while standalone visual recognition techniques fail to effectively perform causal reasoning and safety analysis in complex scenarios. Furthermore, XR devices impose strict computational and real-time requirements, yet existing technologies exhibit limitations in efficiently integrating visual and linguistic information, making it difficult to meet the needs of applications such as navigation and real-time interaction.

Vision-language models demonstrate significant advantages in addressing these challenges. Their end-to-end vision-language comprehension capabilities enable XR devices to directly interpret both visual and textual information, facilitating seamless integration and efficient interaction. Additionally, VLMs excel in complex scene reasoning and decision-making, allowing for advanced causal inference based on visual content, which is particularly beneficial in specialized domains such as healthcare and industrial applications. Moreover, with the adoption of lightweight optimization techniques, certain VLMs have attained the capability to operate with low latency on

Rank	Model	AlpacaEval	CommonEval	SD-QA	MMSU	OpenBookQA	IFEval	AdvBench	Overall
1	Whisper-v3-large+GPT-4o	4.80	4.47	75.77	81.69	92.97	76.51	98.27	87.23
2	GPT-4o-Audio	4.78	4.49	75.50	80.25	89.23	76.02	98.65	86.42
3	GPT-4o-mini-Audio	4.75	4.24	67.36	72.90	84.84	72.90	98.27	82.30
4	Whisper-v3-large+LLaMA-3.1-8B	4.53	4.04	70.43	62.43	81.54	69.53	98.08	79.06
5	Whisper-v3-turbo+LLaMA-3.1-8B	4.55	4.02	58.23	62.04	72.09	71.12	98.46	76.16
6	Ultravox-v0.5-LLaMA-3.1-8B	4.59	4.11	58.68	54.16	68.35	66.51	98.65	74.34
7	Qwen2.5-Omni-7B	4.49	3.93	55.71	61.32	81.10	52.87	99.42	74.12
8	MiniCPM-o	4.42	4.15	50.72	54.78	78.02	49.25	97.69	71.69
9	Ultravox-v0.4.1-LLaMA-3.1-8B	4.55	3.90	53.35	47.17	65.27	66.88	98.46	71.45
10	Baichuan-Omni-1.5	4.50	4.05	43.40	57.25	74.51	54.54	97.31	71.14
11	Whisper-v3-turbo+LLaMA-3.2-3B	4.45	3.82	49.28	51.37	60.66	69.71	98.08	70.66
12	Baichuan-Audio	4.41	4.08	45.84	53.19	71.65	50.31	99.42	70.03
13	VITA-1.5	4.21	3.66	38.88	52.15	71.65	38.14	97.69	65.13
14	Phi-4-multimodal	3.81	3.82	39.78	42.19	65.93	45.35	100.00	63.69
15	MERaLiON	4.50	3.77	55.06	34.95	27.23	62.93	94.81	62.91
16	Ola	4.12	2.97	33.82	45.97	67.91	39.57	90.77	59.98
17	Lyra-Base	3.85	3.50	38.25	49.74	72.75	36.28	59.62	57.66
18	Ultravox-v0.5-LLaMA-3.3-1B	4.04	3.57	34.72	30.03	35.60	45.56	96.92	56.43
19	GLM-4-Voice	3.97	3.42	36.98	39.75	53.41	25.92	88.08	55.99
20	DiVA	3.67	3.54	57.05	25.76	25.49	39.15	98.27	55.70
21	Qwen2-Audio	3.74	3.43	35.71	35.72	49.45	26.33	96.73	55.35
22	Freeze-Omni	4.03	3.46	53.45	28.14	30.98	23.40	97.30	54.72
23	KE-Omni-v1.5	3.82	3.20	31.20	32.27	58.46	15.00	100.00	53.90
24	Step-Audio	4.13	3.09	44.21	28.33	33.85	27.96	69.62	49.77
25	Megrez-3B-Omni	3.50	2.95	25.95	27.03	28.35	25.71	87.69	46.25
26	Lyra-Mini	2.99	2.69	19.89	31.42	41.54	20.91	80.00	43.91
27	Ichigo	3.79	3.17	36.53	25.63	26.59	21.59	57.50	43.86
28	LLaMA-Omni	3.70	3.46	39.69	25.93	27.47	14.87	11.35	37.51
29	VITA-1.0	3.38	2.15	27.94	25.70	29.01	22.82	26.73	34.68
30	SLAM-Omni	1.90	1.79	4.16	26.06	25.27	13.38	94.23	33.84
31	Mini-Omni2	2.32	2.18	9.31	24.27	26.59	11.56	57.50	31.32
32	Mini-Omni	1.95	2.02	13.92	24.69	26.59	13.58	37.12	27.90
33	Moshi	2.01	1.60	15.64	24.04	25.93	10.12	44.23	27.47

Table 1: Performance Comparison of Speech/Text Large Language Models on VoiceBench.

Rank	Method	Param (B)	Language Model	Vision Model	Overall
1	InternVL2.5-78B-MPO	78	Qwen-2.5-72B	InternViT-6B-v2.5	77.4
2	Qwen2-VL-72B	73.4	Qwen2-72B	QwenViT	76.7
3	GPT-4o (0806, detail-high)	-	-	-	76.5
4	Ovis2-34B	34.9	Qwen2.5-32B	AIMv2-1B	75.6
5	GPT-4o (0513, detail-high)	-	-	-	75.4
6	Qwen2.5-VL-72B	73.4	Qwen2.5-72B	QwenViT	75.3
7	Gemini-2.0-Pro	-	-	-	74.8
8	InternVL2.5-38B-MPO	38	Qwen-2.5-32B	InternViT-6B-v2.5	74.4
9	Qwen-VL-Max-0809	72	Qwen2-72B	QwenViT	74.2
10	Step-1o	-	-	-	74.1
11	Ovis2-16B	16.2	Qwen2.5-14B	AIMv2 Huge	74.1
12	LLaVA-OneVision-72B	73	Qwen2-72B	SigLIP-400M	73.9
13	InternVL2.5-26B-MPO	26	InternLM2.5-20B	InternViT-6B-v2.5	73.7
14	Molmo-72B	73.3	Qwen2-72B	CLIP ViT-L/14	73.7
15	LLaVA-OneVision-72B (SI)	73	Qwen2-72B	SigLIP-400M	73.7
16	Taiyi	-	-	-	73.1
17	InternVL2-Llama3-76B	76	Llama-3-70B-Instruct	InternViT-6B	72.7
18	Ovis1.6-Gemma2-27B	28.9	Gemma2-27B	SigLIP-400M	72.7
19	VARCO-VISION-14B	15.2	Qwen2.5-14B	SigLIP-400M	72.5
20	Ovis2-8B	8.94	Qwen2.5-7B	AIMv2 Huge	72.5

Table 2: RealWorldQA Evaluation Results

local XR devices, thereby enhancing real-time interactive experiences.

As VLM technology continues to advance and further integrates with XR systems, future XR devices are expected to achieve more intelligent and natural vision-language interactions, driving the widespread application of immersive human-computer interaction. Table 2 presents the current RealWorldQA Evaluation [2] in OpenVLM Leaderboard data, comparing the performance of various VLMs to illustrate their potential applications in XR interactions.

2.3 Generative AI for XR Content Generation

In Extended Reality (XR) systems, the efficient generation of virtual assets is a critical research area. The challenge of rapidly producing high-quality, diverse, and user-adaptive virtual assets remains a fundamental issue in XR research. Traditional asset generation methods primarily rely on manual modeling by artists or rule-based automated generation techniques. However, these approaches face limitations in scalability, diversity control, and user customization, making it difficult to meet the dynamic and evolving demands of XR interactive environments.

Recently, Large Language Models (LLMs) have emerged as a promising solution for enhancing 3D asset generation due to their strong general semantic understanding, diverse text generation capabilities, and ability to process and generate structured information. Recent studies have explored the application of LLMs in 3D

asset generation, including but not limited to object modeling [3], scene construction and Editing [4, 5], and animation generation [6]. Additionally, LLMs have been leveraged to assist in the automated construction of XR environments [7, 8, 9], significantly improving the efficiency and adaptability of virtual scene generation. In the context of XR storytelling and character-driven interactions [10], recent research has further integrated LLMs with motion generation techniques to enhance the naturalness of character behavior and interactive experiences. These advancements suggest that LLM-enhanced Generative AI holds significant potential for XR content creation, offering novel solutions for more intelligent and automated asset generation.

2.4 Reinforcement Learning and Decision Making

2.4.1 Enhancing XR Interaction Through RLHF in LLMs

Reinforcement Learning from Human Feedback (RLHF)[11] has emerged as a crucial technique for refining the behavior of Large Language Models (LLMs), demonstrating significant potential in enhancing interaction experiences within Extended Reality (XR) environments. Traditional LLM training primarily relies on large-scale text corpora through unsupervised learning or supervised fine-tuning; however, such methods often fail to ensure optimal adaptability and consistency in complex interactive settings. By incorporating human feedback into the training loop, RLHF enables LLMs to align more closely with user expectations, thereby improving interaction naturalness, controllability, and personalization.

In XR interactions, RLHF plays a key role in optimizing LLM-driven decision-making within dynamic, multimodal environments. For instance, in virtual assistants, immersive learning, and role-playing interactions, LLMs must generate responses that align with user preferences while maintaining contextual coherence[12]. RLHF allows models to refine their outputs based on human-provided reinforcement signals, enhancing contextual understanding and reducing irrelevant or inaccurate responses. Additionally, RLHF contributes to improving consistency in XR speech and vision-based interactions by enabling LLMs to retain contextual continuity across multiple dialogue turns and interpret complex user instructions more accurately. For example, in XR design applications, users may wish to issue natural language commands to manipulate virtual assets or modify environments[13]. RLHF-trained LLMs can better comprehend and execute these complex directives, adapting responses based on prior interactions. As RLHF methodologies continue to evolve, future LLMs in XR environments will be better equipped to capture user intentions precisely, delivering more personalized and adaptive interaction experiences.

2.4.2 Applications of Embodied AI in XR Environments

Embodied AI represents a significant advancement in artificial intelligence, aiming to equip intelligent agents with not only language understanding and reasoning capabilities but also the ability to perceive, act, and adapt autonomously within physical or virtual environments. In XR systems, the integration of Embodied AI is driving the development of virtual agents, intelligent robotics, and immersive interactive experiences, enabling more intelligent and human-like engagements.

Within XR environments, Embodied AI demonstrates core competencies in perception, motion generation, and decision-making. By leveraging computer vision and multimodal language models, intelligent agents can analyze XR scenes in real time, interpreting spatial layouts, object properties, and user behaviors[14]. This capability enables XR systems to make informed decisions, such as autonomous navigation, target tracking, and personalized content recommendations. Additionally, through Reinforcement Learning (RL) and motion synthesis models,

Embodied AI can autonomously plan and execute interaction tasks in both physical and virtual spaces[15]. For example, XR assistants can recognize gestures or voice commands to facilitate seamless device control, while VR-based training simulations can replicate human actions with high fidelity. Furthermore, Embodied AI is pivotal in XR storytelling and gaming, where virtual characters dynamically respond to user speech and gestures, creating more immersive and naturalistic role-driven interactions.

Looking ahead, the advancement of Embodied AI will further enhance the realism and autonomy of XR experiences, enabling virtual agents to exhibit higher levels of adaptability and intelligence. By integrating LLMs, multimodal learning, and reinforcement learning techniques, XR systems and intelligent agents will be able to comprehend complex user intentions, execute intricate interaction tasks, and continuously refine their behaviors through learning. As computational capabilities improve and algorithms advance, Embodied AI will unlock new possibilities in XR applications across remote collaboration, intelligent education, medical simulations, and beyond.

3 Taxonomy of LLM for XR: Data Types, Technical Paradigms, and Applications

3.1 Input Modalities in LLM-Powered XR Systems

Across the growing landscape of XR systems powered by Large Language Models (LLMs), input modalities play a foundational role in shaping interaction mechanisms and system capabilities. Based on our survey of over fourty representative works, we find that text, speech, and contextual memory serve as the dominant input types. These modalities are essential for enabling natural language interaction with LLMs, allowing users to communicate through spoken or typed queries, while the systems leverage historical context, user preferences, or task states to deliver coherent, personalized responses. Systems such as EmBARDiment[16] and IllusionX[17] demonstrate this paradigm clearly, incorporating eye-gaze attention and cross-window contextual memory to ground LLM responses in users' behavioral patterns and environmental history.

A subset of these works also incorporates 3D scene and model data as input, enhancing the system's capacity to understand spatial configurations, object semantics, and physical affordances within virtual environments. For instance, in LLMER[9] and Text2VRScene[13], LLMs generate structured data (e.g., JSON) to define scene composition, while works like Function-Adaptive Affordance Extraction utilize LLMs to infer interaction possibilities from 3D object geometry. These approaches go beyond language, enabling LLMs to reason about space and embodiment by integrating spatial and functional models into the input pipeline.

In contrast, other systems rely on image or video data as a primary input channel, especially in cases where visual perception supports semantic understanding or 3D scene reconstruction. The key distinction between this category and 3D scene data lies in the use of image/video as a raw perceptual signal, which may be translated into semantic maps, object labels, or spatial layouts. For example, OCTOPUS[15] and Toward Facilitating Search in VR with Vision-Language Models harness visual inputs for object tracking and retrieval in immersive environments, leveraging multimodal LLMs (e.g., Flamingo[18], Gemini[19]) to bridge vision and language.

Some advanced systems further incorporate sensor-based multimodal input, such as motion capture, eye tracking, and haptic feedback, to enrich user modeling and interaction depth. These modalities provide implicit cues about user focus, intention, or engagement. Works like EmBARDiment[16], VRCopilot[20], and Multimodal

Grounding[21] exemplify this trend by fusing gaze direction, head orientation, and controller motion with textual input to support responsive, embodied LLM agents. This sensor data complements language-based interaction by providing real-time grounding of spatial and attentional context.

In summary, LLM-enhanced XR systems adopt a layered input architecture combining semantic input (text/speech), environmental input (3D scenes/images), and perceptual input (sensor signals). The integration of these modalities forms the backbone of multimodal interaction, enabling more immersive, adaptive, and intelligent virtual experiences. As LLMs continue to evolve, these input strategies will become increasingly important for shaping the next generation of embodied AI interfaces in virtual, augmented, and mixed reality.

3.2 Technical Paradigms

3.2.1 3D Sensing and Input

This paradigm includes systems where LLMs are embedded into environments that leverage multimodal sensors such as spatial tracking, gesture recognition, and gaze estimation. In ARAS[22], the surgical assistance system integrates GPT-4 with Microsoft HoloLens-based AR, enabling context-aware instruction generation based on the surgeon's head orientation and tool position (noted in system design section of paper). Augmented Object Intelligence with XR-Objects[23] uses a smart object framework where virtual artifacts become semantically aware when touched, with scene-aware LLMs parsing object identity and issuing dialog or feedback based on embedded tags (Unity + LLM bridge).

In Building LLM-based AI Agents in Social VR[24], the system implements GPT-3.5 in a Meta Quest Pro social VR setup, using gaze and controller input to trigger conversational turns from embodied avatars. Similarly, Robotic Ultrasound with Conversational Agent[25] integrates GPT-based conversational AI with a robotic probe and visual overlays on HoloLens, enhancing trust and user understanding through sensor-driven explanatory prompts. Environment-Aware Spatial Interactions[26] leverages a fine-tuned LLaMA-2 model to infer users' intents from room layout and prior action sequences (noted in architecture block diagram), while Search in VR with Vision-Language Models integrates Flamingo[18] and Gemini[19] for visual grounding and gaze-based retrieval of objects in cluttered VR workspaces (noted in multimodal pipeline description). These projects differ in how they fuse sensor streams—some emphasize spatial awareness, others social dialogue—but all use LLMs as semantic interpreters over embodied interaction.

3.2.2 Multisensory and 3D Content Rendering

Systems in this category involve LLMs coordinating rich feedback across modalities including visuals, haptics, and audio. Analyzing Multimodal Interaction Strategies[27] investigates how LLMs assist scene manipulation through speech and gaze, using a custom command-to-action pipeline linked to Unity-based 3D environments. Autonomous Workflow for Training Assistants[28] combines multimodal fine-tuned LLMs (based on LLaMA backbone) with scene simulation modules to build fine-grained training flows for industrial XR (described in Figure 1).

Function-Adaptive Affordance Extraction[29] takes 3D object geometry and applies a two-step pipeline: first, a shape encoder computes object affordance likelihoods; then GPT-4 maps these to human-readable instructions or scene updates (affordance-LLM hybrid noted in figure caption). LLMER[9] demonstrates the use of GPT-3.5 to output structured JSON scene specifications from narrative prompts, which are parsed into 3D interactive XR

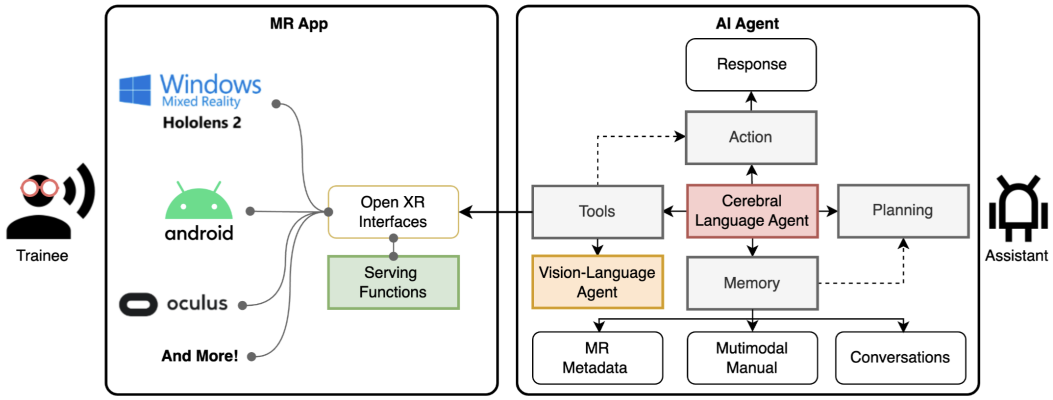
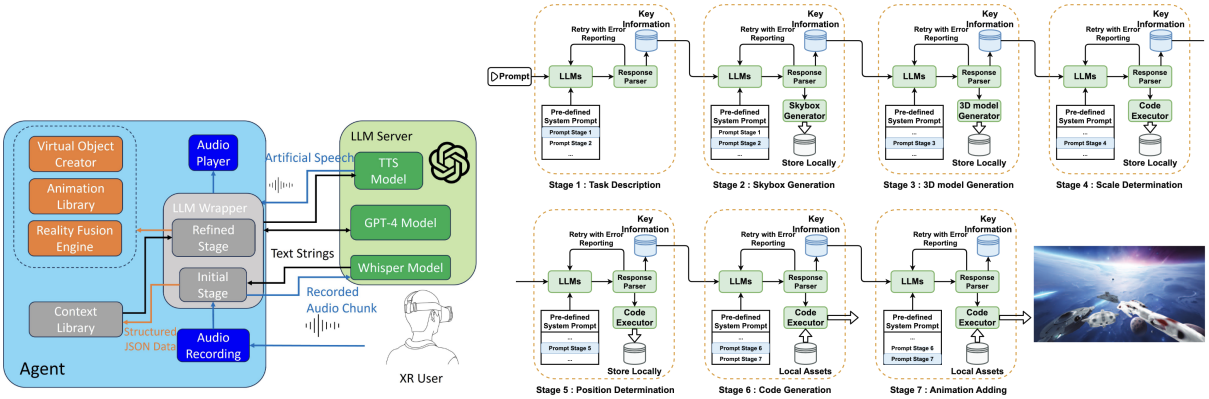


Figure 1: System architecture of the "Autonomous Workflow for Training Assistants[28]" platform. The system involves an AI agent interacting with an MR application. The AI agent comprises a core cerebral language agent, which interacts with a vision-language agent to interpret the multimodal context into metadata, which can be utilized by the cerebral language agent iteratively. The MR application interacts with AI agents by serving functions as external tools.

layouts in Unity (system pipeline specifies OpenAI API). sMoRe[30] expands on this by introducing a generative LLM-based object organizer, leveraging Unity+LLM+GAN architecture to automatically reposition objects in MR environments based on spatial reasoning. This category is characterized by scene-level synthesis and affordance reflection—using LLMs as both semantic translators and interactive planners.



(a) System architecture of LLMER[9].

(b) Workflow of Text2VRScene[13] System

Figure 2: The proposed method learns a latent space representation of social interactions.

3.2.3 Embodied Agents and Virtual Humans

Embodiment is a defining trait of many LLM-XR systems aiming for human-like presence. EmBARDiment[16] integrates GPT-4 with gaze-based attention models from XR headsets (e.g., Vision Pro, Meta Quest Pro), enabling proactive engagement without explicit prompting. Exploring Presence in NPCs[31] tests both speech-to-text GPT-3 pipelines and fixed dialog trees in a controlled VR study, examining how LLM fluidity affects presence.

LLM Chatbots for Autism Communication Training[32] utilizes GPT-3.5 with scenario-specific prompt engineering to simulate job interviews, and includes latency management and error mitigation layers to suit neurodi-

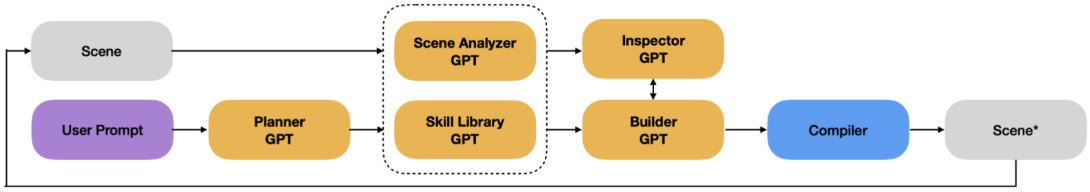


Figure 3: Large Language Model for Mixed Reality (LLMR[35]) architecture for real-time interactive 3D scene generation. Starting from the left, a user prompt and the existing 3D scene (Ω) are fed into the Planner (P) and Scene Analyzer (SA) modules, respectively. The Planner decomposes the user prompt into a sequence of sub-prompts, while the SA summarizes the current scene elements. These are then integrated with a Skill Library (SL) to guide the Builder (B) module, which generates the appropriate code. The Inspector (I) module iteratively checks the generated code for compilation and run-time errors. Upon receiving the green light from the Inspector, the code is compiled using the Roslyn Compiler and executed in the Unity Engine to produce the desired 3D scene and functionalities as specified by the user.

vergent users. Multimodal Grounding in AR[21] employs ARKit, gaze tracking, and LLaMA-2 to interpret user intent and provide semantic task feedback, where sensor-to-prompt architecture is explicitly discussed. Sushi with Einstein[33] showcases a hybrid LLM avatar system using scripted prompts and real-time user response parsing, delivered through a smart glasses + Unity pipeline, enabling an LLM-based historical character to co-host events. Personality in Educational Agents[34] explores how fine-tuned GPT-based agents with personality embeddings improve educational outcomes in XR learning environments. These works vary by embodiment fidelity, from fully improvisational agents to structured hybrids, yet all reflect the growing maturity of LLM-based NPC design.

3.2.4 Architectures and Toolkits

This paradigm addresses how LLMs are embedded at a system level, enabling interaction, scene logic, and agent control. LLMR[35] presents an architecture that uses real-time GPT-4 prompting to dynamically modify XR environments; developers inject new prompts that influence object properties, scene narratives, and even avatar behavior (architecture overview shown in Figure 3). MagicItem[36] describes an LLM-enhanced scripting system where a domain-specific language (DSL) is interpreted by Codex or GPT-3.5, compiling natural-language rules into in-game logic on a commercial metaverse platform.

Mixed Reality IoT Smart Environments[37] leverages GPT-4 API to semantically translate user goals (e.g., "make room coz") into actuator control sequences across IoT-connected devices, synchronized within a Unity-driven XR world. XaiR[38] offers an XR platform that bridges real-world sensors (e.g., RFID, thermal) with LLM agents via Python + ROS + Unity, allowing physical context to drive LLM interactions.

In parallel with system runtime frameworks, Vlmevalkit[2] provides an open-source evaluation toolkit for benchmarking large multi-modal models, including vision-language LLMs like Flamingo and Gemini. While not embedded directly in XR environments, it supports architecture-level decisions by offering standardized evaluation metrics, dataset wrappers, and fine-tuning protocols, making it highly relevant for XR developers selecting or refining their LLM pipelines.

Khan *et al.* present LoXR[39], which provides actionable guidance: XR developers should carefully match LLM size in figure 4 to device capabilities, as using a slightly smaller or quantized model can dramatically improve latency and efficiency with minimal quality loss. High-end devices like AVP unlock better absolute performance,

but one must mind stability and power draw; lower-end devices can host LLMs reliably if given models optimized for efficiency. The Pareto-optimal set in figure 4 identified in LoXR can serve as a reference for choosing an appropriate model–device combination given a particular XR application's speed vs. accuracy requirements. The authors combine a model's performance (speed) and stability scores with its quality (accuracy) into a unified evaluation, and plot Pareto frontiers for the model–device pair as shown in figure 4. The Pareto analysis reveals clear trade-offs: Apple Vision Pro (CPU) offers unmatched raw speed and accuracy, but its runs are somewhat less stable (higher variance). In contrast, Magic Leap 2 consistently delivers the most stable performance (almost no variance or errors), making several ML2 pairs Pareto-optimal despite lower throughput. The Vivo and Quest 3 devices each achieve a middle-ground trade-off – a few of their model runs lie on the Pareto front by balancing decent speed with good stability. Interestingly, the largest models (LLaMA-2 7B and Mistral-7B) did not dominate the Pareto-optimal set due to their slow speed and high resource usage. Instead, some mid-sized models (around 3–4B parameters with efficient quantization) emerged as optimal choices, offering a better quality-to-speed ratio for on-device use. For example, LoXR notes that model m3 (a 2.6B Vikhr-Gemma variant) ran very fast on AVP and very stably on MQ3, making it Pareto-optimal, while m8 (3.8B Phi series) offered strong accuracy and speed on AVP plus high stability on ML2.

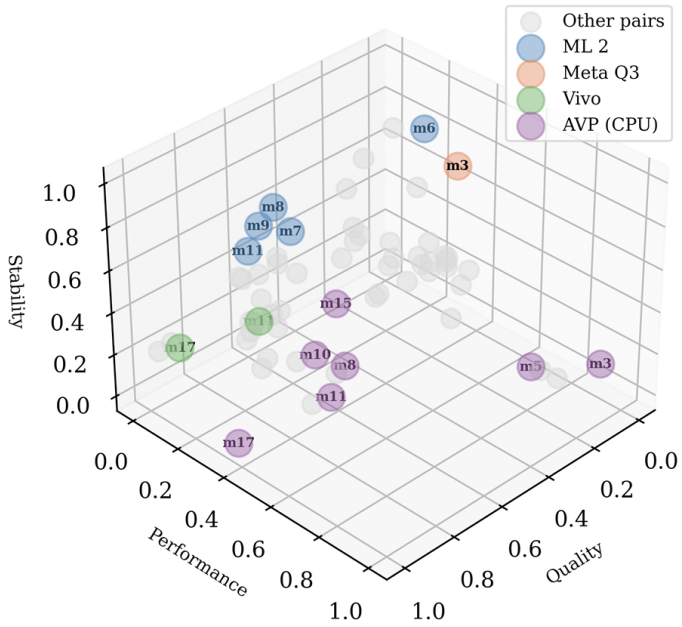
This category emphasizes extensibility and LLM-as-middleware designs—LLMs don't just converse, they orchestrate environment dynamics at runtime and require robust architectural and evaluation support to operate effectively.

3.2.5 Presence, Agency, and Cognition

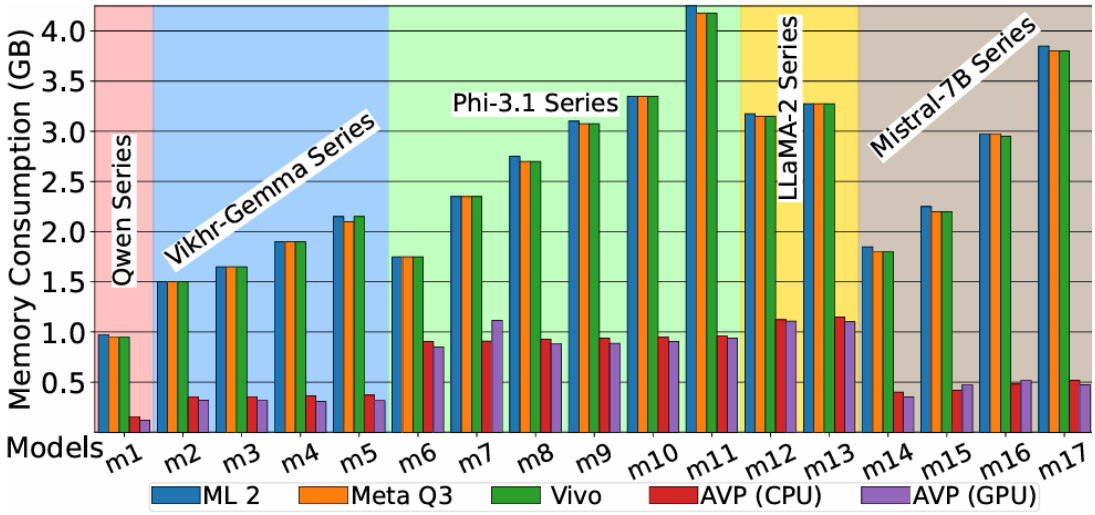
Some studies focus on how LLMs enhance presence and cognition by generating culturally rich, immersive experiences. Open Sharing of Art Labs[40] explores how artists co-create content via VR interfaces supported by GPT-3.5 to facilitate semantic bridging between languages, media types, and disciplines. Scientific and Fantastical Learning[41] combines LLM-driven storytelling with educational AR overlays; the agent curates exhibits with historical, scientific, or fantasy layers depending on user profile and gaze direction. While these systems don't push architectural boundaries, they showcase how LLMs shape narrative coherence and user immersion in cultural XR settings.

3.2.6 Human Factors and Usability

LLM-powered XR systems also face challenges in user experience design and physiological feedback. Early Prediction of Cybersickness[42] uses time-series modeling with LLaMA-based LLMs to infer motion sickness likelihood based on inertial and biofeedback sensor data, predicting discomfort before it manifests (noted in Sec. 3). Enhancing Patient Acceptance of Robotic Ultrasound[25] fuses a LLM dialogue agent (based on GPT-4) with visual feedback to increase transparency and trust during interaction. VirtuWander[43] analyzes how a multilingual GPT-based tour guide can boost spatial awareness and memory retention during virtual tourism (language preference and route adjustment pipelines specified). These works often employ controlled studies and highlight the need to optimize response time, trust, and clarity in LLM interactions.



(a) Pareto front of optimal model–device pairs, with different colors for each device type. Performance combines speed, memory, and battery metrics; stability reflects the variance (CV) and error rates in PP/TG tests; quality is the accuracy on six benchmark datasets. Points along the front (e.g. certain models on AVP and ML2) represent the best trade-offs between speed and reliability for a given accuracy level.[39]



(b) Memory consumption for each model-device pair.The results represent the mean values across batch sizes 128,256,512,and 1024.[39]

Figure 4: Quantitative results from LoXR on performance vs. memory consumption.

3.2.7 Accessibility and Inclusive Design

LLMs also unlock new possibilities for accessibility. Supporting Text Entry in VR[44] incorporates GPT-2-style autoregressive prediction in a Unity-based VR keyboard, reducing input time and improving error tolerance. VIAssist[45] adapts vision-language models like Flamingo to assist users with visual impairments in object identification and spatial orientation, combining VLM-generated descriptions with haptic cues through XR gloves. Both systems emphasize inclusive design, showing how LLMs can personalize XR experiences for diverse users.

3.2.8 Immersive Applications

Finally, numerous systems use LLMs to support applied domains such as training, therapy, or daily productivity. AdaptiveCoPilot[46] incorporates GPT-4 with neuroadaptive feedback, adjusting cockpit guidance based on fNIRS and EEG input for novice and expert pilots (sensor-LLM interface detailed in Sec. 2). Investigating Latency with Multimodal Feedback[47] studies how LLM-agent response delays (via GPT-3) can be masked using well-timed visual or haptic signals, optimizing immersion. IllusionX[17] offers a personal XR assistant powered by GPT-4, using user memory and attention to provide proactive reminders, environment labeling, and context-aware messaging (noted in memory layer architecture). OCTOPUS[15] enables object placement using semantic search with GPT-4, allowing users to refer to any known or novel object using open-vocabulary queries, supported by spatial tracking and object anchoring in MR. This group of applications demonstrates the versatility of LLMs as spatially grounded assistants that extend user capabilities in both professional and personal contexts.

3.3 Evaluation Metrics in LLM-Enhanced XR Systems

Modern XR systems augmented with large language models (LLMs) are evaluated using a mix of qualitative user experience metrics and quantitative performance metrics. Many projects conduct user studies to assess subjective outcomes like immersion or trust, alongside objective measures of system efficiency and accuracy. Below, we group common metrics by type, noting which papers employed each and in what context. We also highlight baseline comparisons (XR-only or LLM-only) where applicable.

3.3.1 Qualitative (User Experience) Metrics

Presence and Immersion

Several studies measured how "present" or immersed users felt in the XR experience. For example, a VR study on LLM-driven NPC dialogue explicitly evaluated social presence using the Game Experience Questionnaire social presence module[31]. Similarly, an IVA 2024 study[47] on an embodied conversational agent (ECA) measured immersion/presence levels under different feedback conditions. In both cases, participants filled out Likert-scale questionnaires after the experience to rate their sense of being "there" in the virtual environment. Generally, LLM-Enhanced XR systems aim to maintain high presence; free-form speech with an LLM-driven character yielded higher social presence than menu-based dialogue in a VR storytelling scenario, showing the benefit of LLMs for immersion. Conversely, the presence metrics help ensure that adding an AI does not detract from the XR realism.

Trust and Acceptance

Trust is crucial when an AI agent is introduced, especially in sensitive domains. In a medical mixed-reality study (virtual assistant for robotic ultrasound), patient trust and technology acceptance were key metrics[25]. This post-experience survey showed significant improvements in patient trust and acceptance of the procedure when an LLM-powered conversational agent and AR visuals were present, compared to a baseline with no such agent. Similarly, perceived trustworthiness of an NPC's responses was evaluated in a VR gaming context by asking users to rate how much they trusted the information from an LLM-driven character[48]. These trust metrics are especially common in safety-critical or educational XR applications, where users must feel confident in the AI's guidance.

User Satisfaction and Usability

General satisfaction and usability are often assessed via questionnaires or interviews. For instance, an XR design tool for 3D scene editing with LLM assistance gathered post-experience questionnaire feedback to gauge user

satisfaction and identify pain points[27]. Many projects used standard usability metrics: for example, a VR text-entry system with LLM suggestions examined perceived ease of use and workload, noting reduced cognitive effort when the LLM auto-completed text[44]. In the VR NPC study, users also rated interaction effort and usability, showing the speech-based LLM interface required less effort than navigating dialogue menus[31]. Although not all papers reported a formal SUS score, usability themes (efficiency, ease, workload) were common in questionnaires. Virtual AIVantage[49], a VR interview training tool using GPT-4, noted plans for future usability studies focusing on underrepresented users, emphasizing the importance of inclusive user satisfaction metrics even if a formal study was pending.

Engagement and Enjoyment

Especially in educational and entertainment XR applications, user engagement is measured. An AR learning experience with a narrative LLM character asked children to rate their enjoyment and engagement during an interactive story[41]. In that study, investigators used custom Likert questions and observed that all LLM-enhanced narrative models were positively perceived in terms of engagement. Likewise, a study on an educational LLM-based agent with different personality styles measured user engagement and flow – finding that agents with more expressive (extroverted) personalities were rated as more engaging by learners[34]. A VR NPC experiment also evaluated flow (the degree of absorption in the experience) via standard game experience questions[48]. High engagement metrics indicate the LLM's dynamic responses can increase immersion and keep users interested, compared to static XR content.

Embodiment and Persona Perception

In XR settings with virtual avatars or agents, researchers often measure how the agent's embodiment or personality is perceived. Embodiment refers to how "present" and lifelike the AI character appears. EmBARDiment[16] evaluated this via a custom HLMIq questionnaire, which quantitatively assessed the agent's human-likeness and helpfulness. Their user study found, for example, that 75% of users preferred an anthropomorphic avatar design for the AI agent, indicating a clear user preference for human-like embodiment. In an educational XR agent study, participants rated the agent's personality traits; the authors report that the perceived extroversion/agreeableness of the agent matched the intended persona, confirming the embodiment and persona were effectively conveyed[34]. These metrics (often collected via trait rating scales and preference polls) help compare an LLM-driven embodied agent vs. a disembodied or less expressive version. Such findings guide designers on whether a virtual human form, gestures, and expressions (powered by the LLM's output or state) enhance the experience.

Open-Ended Feedback

In addition to numeric scales, many studies gathered qualitative comments. While not a "metric" per se, themes from interviews or open-ended survey questions were used to explain quantitative results. For example, the 3D scene editing paper analyzed common interaction patterns and barriers from user feedback[50], and the LLMER study noted user suggestions that point to further optimization[9]. These insights are grouped under qualitative evaluation to refine systems beyond the measured scores.

As seen above, presence/immersion and usability are among the most commonly reported qualitative metrics across LLM enhanced XR studies, while trust is crucial in domains like healthcare or education where users rely on the AI's guidance. The specific metrics vary by context (e.g. social presence for collaborative NPC interactions, vs. trust for assistive or medical agents), but nearly all papers conducted some form of post-study survey to gauge user sentiment (often 5- or 7-point Likert scales). Notably, many comparisons were made against a baseline: for instance, rating an LLM-driven condition versus a non-LLM condition. In the NPC study, participants

consistently reported higher presence and lower effort with the LLM-driven free speech interface compared to a traditional dialogue tree (XR-only baseline)[31]. In the text entry study, users preferred the LLM-augmented typing, citing less workload than the baseline VR typing without predictive assistance[44]. These qualitative outcomes demonstrate the value added by LLM integration in XR experiences in terms of user experience.

Table 3: Common qualitative evaluation metrics and corresponding LLM enhanced XR studies.

Qualitative Metric	**Studies (Application Context)**
Presence / Immersion	Christiansen et al. *VR NPC dialogue[31]* [VRST'24]; Morad et al. *VR ECA[47]* [IVA'24]
Trust	Song et al. *MR medical agent[25]* [CHI'25]; Christiansen et al. *NPC trustworthiness[31]* [VRST'24]
Ease of Use / Effort	Christiansen et al. *Menu vs Speech input[31]* [VRST'24]; Chen et al. *VR text input with LLM suggestions[44]* [IEEE VR'24]
Likeability	Morad et al. *Agent likeability[47]* [IVA'24]
Willingness to Reuse	Morad et al. *Willingness to re-engage[47]* [IVA'24]
Engagement / Flow	Sonlu et al. *Personality Effects on Engagement[34]* [CHI'24]; Christiansen et al. (Narrative NPC flow[31]) [VRST'24]
Human-likeness	Bovo et al. *EmBARDiment Agent[16]*; Sonlu et al. *Agent perception and persona[34]* [CHI'24]
User Preference	Bovo et al. *75% preferred avatar vs voice-only[16]*
Acceptance	Song et al. *Tech acceptance improved in MR surgery[25]* [CHI'25]

3.3.2 Quantitative (Performance) Metrics

In addition to subjective feedback, LLM enhance XR papers track various objective metrics to quantify performance improvements or trade-offs introduced by the LLM. Table 4 outlines key quantitative metrics and which papers reported them.

Quantitative metrics were often used to compare against baseline systems. For XR-only baselines, for example, in the VR NPC study, the baseline was a traditional branching dialogue (no LLM). The LLM condition increased conversation flexibility at the cost of some response time. Quantitatively, they logged conversation length and variety, but more telling were the user metrics showing the LLM's advantage in experience.

For LLM-only baselines, some papers compared their integrated system to a non-XR scenario or an alternate AI. XaiR[38], for example, compared the LLM assistant's task performance to a human operator baseline, finding the LLM achieved comparable task accuracy (90%), though the paper notes gaps remain (the AI might still be inferior in complex situations). In educational agents, different embodiments were compared (dialogue-only vs. animated), effectively isolating the XR component's impact on learning – result: all had similar learning scores, so adding animations didn't reduce learning, but it did boost engagement[34].

Compared to previous state-of-the-art, LLMER[9] compared its JSON-based world creation to prior code-generation methods, showing drastic improvements in stability (fewer crashes) and efficiency. Here the metrics

Table 4: Quantitative metrics and associated outcomes in LLM enhanced XR studies.

Quantitative Metric	**Studies (Application Context)**
Task Completion Time	Chen et al. *LLMER: XR Scene Generation via LLM[9]* – 60% faster than baseline [IEEE VR'25]
Text Entry Speed	Chen et al. *Text Entry in VR with LLM Suggestions[44]* – Higher WPM [IEEE VR'24]
Error Rate	Chen et al. – Reduced typos with LLM support[44] [IEEE VR'24]
Task Success / Accuracy	Srinidhi et al. *XaiR[38]* – 90%+ task success; Wang et al. – 15% higher assembly accuracy [CHI'24]
Resource Consumption	Chen et al. *LLMER: XR Scene Generation via LLM[9]* – 80% fewer prompt tokens [IEEE VR'25]
Engagement / Flow	Sonlu et al. *Personality Effects on Engagement[34]* [CHI'24]; Christiansen et al. *Narrative NPC flow[31]* [VRST'24]
Latency	Chen et al. *LLMER[9]* – Reduced response time; Morad et al. – *Perceived latency improved[47]* [IVA'24]
Learning Gain	Cheng et al. – *AR quiz post-test gains[41]* [CHI'24]; Sonlu et al. – *Consistent learning improvements with LLM tutor[34]* [CHI'24]

like task time and token count underscore the benefit over the baseline approach (which still used LLM, but in a less optimized way).

Across the literature, task time, success rate, and user task load emerge as the most commonly used quantitative metrics. Fewer papers used standardized NLP metrics (e.g. BLEU, accuracy on a labeled dataset) because the tasks in XR were interactive and open-ended. Instead, they defined metrics aligned with the application: e.g. "accuracy" might mean correct object placement in AR[38] or correct answers in a quiz[41]. Table 4 illustrates this diversity of metrics by context.

4 Applications: LLM-Enhanced XR Experience

Building on rapid advancements in LLMs and deep integration of them and XR, numerous innovative applications have emerged across various fields, such as healthcare, education, social interaction, and entertainment. Owing to multimodal understanding and the assistance of AI agents, LLMs bring more intelligent applications with a lower barrier.

Healthcare

Prior XR applications in healthcare focus on display and manipulation, primarily relying on computer vision techniques and rule-based configurations. However, these applications fail in complex medical fields such as surgery, where unlimited situations can happen. Besides, these rigid and formulaic arrangements overlook patients' emotional requirements, accompanied by an unreal experience.

One obvious advantage of LLMs is that, after extensive training on vast datasets, they can deeply participate

in the medical process, thus enabling context-aware and intuitive communications between medical staff and XR systems. Some work [22] leverages the LLMs' understanding of NLP to enable users to execute corresponding functionalities with voice input during the surgical procedure. Its multimodal analysis capabilities, especially image understanding, can accept various kinds of pathological images, which assist physicians in diagnosing medical conditions [51]. They can also help patients with their defects in daily life [45, 52]. Furthermore, as the cognitive core of AI agents, it enhances the human-like qualities of virtual characters. Considering a state of psychological and physical vulnerability, these virtual assistants provide more psychological comfort, which contributes to reducing stress and improving the patient's experience during the therapeutic procedure [25, 32].

Education

XR applications have shown potential in education due to their immersive experience and interactive learning procedure [53]. Compared with traditional teaching methods, XR applications offer more practical training and more enjoyable motivation. Besides, due to their low cost and reproducibility, they demonstrate inherent advantages in destructive training.

When combined with LLMs, XR applications further demonstrate personality expression and embodiment in various educational fields, which are positively perceived regarding users' experience and learning outcome [34]. For conversational tasks, acquiring vast amounts of training data, LLM-enhanced Chatbots are capable of engaging in attentive conversations and providing solutions. For example, they can offer interactive language practice and exhibit the flexibility of human tutors in linguistic education[54], where learners can enhance their speaking skills in various contexts. In online learning platforms, they act as teaching assistants to bridge the gap between students and teachers, providing assignment evaluation and analysis, and real-time Q&A [55]. In industrial assembly tasks, detailed instructions through voice input during the operation process can significantly improve both efficiency and accuracy [56]. Beyond merely utilizing conversational question-and-answer capabilities, LLMs can be integrated as virtual agents to actively engage in the learner's entire educational journey, offering enhanced motivation and greater enjoyment [41, 57, 17].

Social Interaction

With the significant development of the digital human, social interactions in XR can maintain high realism while being unrestricted by space, where everyone can have their own virtual avatars. However, interacting with virtual characters, they lack the necessary reasoning capabilities to interpret user behavior and content generation skills to respond appropriately.

When powered by LLMs, virtual characters can simulate human behaviors and offer a more immersive interaction experience for users. One typical perspective is to leverage LLMs to enhance non-playable characters (NPCs) in conversational scenes, which provides social presence for better immersion [31]. Another typical kind of method takes a step forward, giving AI agents the ability to hear, distinguish human language from noise, understand conversations, organize responses based on memory systems, and express responses in natural language with facial expressions and body language of a VR avatar [24, 10]. Furthermore, these virtual characters are capable of engaging in professional fields and participating in meaningful discussions, showcasing vast potential for applications across cultural, entertainment, and educational domains [33].

Entertainment

XR technologies have transformed the way of entertainment with digital assets via blending the physical and virtual worlds, which have opened up new possibilities for various domains. By utilizing LLM technologies, XR can further improve user experience and trustworthiness in various fields. Different from traditional XR appli-

cations, LLM-enhanced techniques have introduced a new dimension in cultural context, which facilitates a new way for engaging with and learning about artworks in virtual environments [48]. Prior works also focus on leveraging a multi-modal large model for the tour guidance in virtual museums, encouraging multi-modal interactions to boost user experiences, concerning engagement, immersion, and spatial awareness [43]. For scene creation, such pre-trained generative AI can be integrated into immersive authoring to facilitate human-AI co-creation, thus supporting rapid prototyping and intermediate representations such as wireframes to augment user controllability over the created content [20].

5 Challenges and Open Questions

5.1 Challenges of Embodied, Real-Time XR Integration

Large language models promise more natural interactions in XR, but integrating them into embodied, physical settings raises fundamental challenges. Embodiment and context-grounding are foremost: XR systems must anchor AI understanding in the physical world. Prior augmented reality (AR) assistants in industry have relied on explicit environment models and sensors – for example, a digital twin with 3D cameras and trackers provided context for an AR repair-assistance system. Such systems proved that co-located human–robot teaming via AR is feasible[21], but they also revealed brittleness in language-based control (e.g. sensitivity to phrasing). An LLM-driven agent in XR needs similar environmental awareness to avoid floating free of reality. Early work on XR "cognitive assistants" has shown the value of computer vision and spatial tracking to bind instructions to real objects (e.g. overlaying part labels on actual machinery). The open question is how to imbue LLMs with this same situated understanding. Unlike traditional screen chatbots, an XR-embodied AI can observe user gaze, gestures, and surroundings – a rich contextual "canvas" that mostly went unused in past chat interfaces[16]. Leveraging these implicit signals for grounding is non-trivial. Recent research suggests using attention mechanisms on XR sensor data to give the LLM an awareness of what the user is doing and seeing, thereby minimizing the need for explicit user prompts. How to best fuse language models with continuous spatial inputs remains an open challenge in embodiment. The goal is an AI that perceives and acts in situ, but achieving this means solving fundamental problems of reference ("what is this object?"), scene understanding, and maintaining a believable agent persona within the immersive environment. Existing virtual human simulations in XR show that embodiment can greatly enhance realism and engagement[49], yet controlling such agents with LLMs raises new issues of consistency, trust, and user comfort that the field has yet to resolve. In summary, truly embodied LLMs will need to ground their dialogues and decisions in the here-and-now of the user's world – a requirement that pushes the boundaries of current AI context modeling.

5.2 Real-Time Interaction Constraints

XR applications demand responsive, fluid interaction, so any AI component must operate under strict timing and performance constraints. This poses a serious challenge for large language models, which are computationally intensive and typically run on servers with noticeable latency. In immersive settings, even a few hundred milliseconds of lag can break presence or frustrate users. Prior AR studies underscore that user performance and workload are sensitive to interface timing – well-designed AR guidance can shorten task times and reduce errors, but if feedback comes too late, the benefits vanish. Unfortunately, LLM inference is not instantaneous, especially when

handling multi-modal inputs (vision, speech) or long context windows. How can we reconcile the need for real-time XR interaction with the heavy resource demands of LLMs? Emerging system architectures are beginning to tackle this: one strategy is to split processing between the edge device and the cloud. For instance, an AR headset might handle local spatial tracking and user interface rendering, while offloading language understanding to a remote server. This was the approach taken by XaiR, which offloaded multimodal LLM computation to a server and thereby improved real-time content placement in AR scenes. Yet, such split architectures introduce dependency on network connectivity and risk higher latency if the connection is poor. They also complicate the system, creating new points of failure and raising concerns about data privacy (streaming sensor data off-device). An open research question is how to optimize the responsiveness of LLM-driven XR agents. Possible avenues include model compression or distillation to run lighter language models on-device, prediction caching or anticipatory processing to "guess" likely user requests, and clever interaction design that masks latency (e.g. by smoothing turn-taking in dialogue). Without improvements, the dream of a seamlessly interactive XR assistant remains out of reach – real-time performance is a non-negotiable requirement that today's LLMs struggle to meet. Ensuring low-latency, jitter-free operation of LLMs in XR is thus a critical challenge for the community.

5.3 Physicality and Spatial Reasoning

Unlike conventional AI applications, XR systems operate in and affect the physical world. This brings challenges of spatial reasoning and safety that current LLMs are ill-equipped to handle. AR guidance must be precisely aligned to real objects and spaces – a virtual arrow pointing to "the red valve" is only useful if it actually indicates the correct real valve. Classic AR research achieved such alignment through tracking and registration techniques, but large language models have no innate sense of 3D space. Recent work explicitly notes that current multimodal LLMs have "inherent limitations... in processing 3D inputs". In other words, an LLM on its own cannot reliably interpret spatial geometry or maintain a mental model of an evolving 3D scene. This gap manifests as a major open question: how can we give LLMs a sense of physical space? One approach is to integrate external perception modules – for example, computer vision models that convert images or depth scans into symbolic descriptions the LLM can reason over. Yet this handoff is imperfect; critical spatial details might be lost in translation. Another strategy is to train embodied language models with 3D-aware data (as in robotics research that pairs vision and language[58]), but these models are still in their infancy. Ultimately, effective spatial reasoning may require hybrid AI architectures where dedicated spatial computing (from simultaneous localization and mapping to object recognition) works in tandem with the LLM. Aside from understanding space, an XR AI assistant must also respect the constraints of physicality: real objects are heavy, fragile, or dangerous, and users have limited field of view and dexterity. Instructions that are perfectly logical in text could be impractical or unsafe in a real environment. For instance, an LLM with no physical grounding might blithely suggest an assembly step that requires an extra hand or ignores gravity. Prior studies have measured how AR instruction systems impact human cognitive load and error rates; they show that if information is not presented at the right moment and in the right way, it can confuse more than help. Translated to LLM-driven guidance, this means the AI must decide what level of detail the user needs and when to provide it – too early or too verbose, and the user may be overwhelmed, too vague or too late, and the user may make a mistake. Recent adaptive AR prototypes attempt to tune information flow to the user's cognitive state, but implementing such adaptation with LLMs remains largely unexplored. In summary, the physical nature of XR tasks demands that LLMs become keen spatial reasoners and context-sensitive instructors. Ensuring that an

LLM's advice is not just linguistically correct but physically appropriate (and safe) is an open problem requiring advances in multimodal understanding and human-in-the-loop design.

5.4 Integration and Complexity

Integrating large language models (LLMs) into XR workflows adds complexity to already intricate XR application development. Traditional XR systems handle 3D rendering, sensor fusion, user interfaces, and networked collaboration. Adding LLMs as a new component requires seamless integration. One challenge is multimodal integration: XR interactions generate visual, auditory, and kinesthetic data, which the LLM must process alongside language understanding. For example, in an industrial assembly setting, an XR system must receive camera data, detect a missing bolt, interpret the user's spoken question ("What's next?"), query an LLM for instructions, and provide a verbal and visual response, all in sync.

Past systems achieved this using hard-coded domain knowledge, while LLM's flexible reasoning may introduce unpredictability, leading to mismatches between actions and displays. Ensuring consistency between virtual and real worlds is essential. For instance, if the LLM says "press the green button" but the AR display highlights a red lever, the user gets confused. Validation and control mechanisms are needed to verify LLM outputs against known scene data or constrain the LLM with a predefined ontology.

This modular architecture, with the LLM as one module in a larger loop, requires careful design. Though frameworks like context libraries and execution monitors have been proposed, integration remains challenging due to differences in data formats and update rates. Current LLM-XR integration tools are still in early stages. Future research should focus on reducing the integration burden, including middleware development and best practices for XR developers without ML expertise. How can we test and debug XR applications where LLM outputs vary?

These practical concerns highlight that LLM-XR integration is not just about model accuracy or speed; it's about building reliable, maintainable systems. Until we resolve integration complexity, large-scale LLM-driven XR deployments, such as in factories or AR-enabled workforces, will remain a challenge.

Despite these hurdles, the confluence of XR and large language models is a frontier rich with opportunity. Tackling embodiment, real-time performance, physical grounding, and integration complexity will require interdisciplinary advances. The surveyed works without LLMs – from AR assembly guides to embodied robot interfaces – provide crucial lessons on what to expect when language meets XR. They remind us that effective XR assistants must be as aware of a bolt's location and a user's fatigue as they are of grammar and semantics. Ultimately, surmounting these open questions will pave the way for XR systems that are not only immersive, but also intelligent and context-savvy – bridging the gap between human intentions and the physical world through fluent, embodied dialogue.

6 Conclusion

In this survey, we have presented a comprehensive analysis of how large language models can enhance extended reality systems, drawing connections between underlying technical paradigms and a broad spectrum of XR applications. By organizing the literature along technical dimensions (from multimodal sensing and generative scene rendering to embodied agent control and XR development toolkits) and mapping these onto application domains (from education and healthcare to manufacturing and entertainment), our work provides a structured lens to un-

derstand this emerging field. This dual categorization elucidates how advances in natural language capabilities translate into tangible improvements in immersive learning environments, clinical training simulations, industrial support tools, and beyond. For researchers and practitioners, the survey serves as a valuable reference that synthesizes diverse efforts, highlights design patterns, and clarifies the state-of-the-art for LLM-driven XR experiences.

Looking ahead, the convergence of LLMs and XR opens numerous opportunities and challenges that require interdisciplinary effort. We encourage further work that spans interaction, cognition, and system architecture in XR. This includes innovating more natural and intuitive interaction techniques (e.g., speech, dialogue, and gesture interfaces powered by LLMs), expanding the cognitive abilities of XR agents (such as better contextual understanding, reasoning, and long-term memory), and refining system architectures (for efficient, real-time integration of LLMs with XR hardware and sensors). Bridging these aspects will be crucial for developing next-generation XR systems that are more intelligent, context-aware, and human-centric. It is our hope that this survey stimulates new ideas and collaborations across these dimensions, guiding the design of LLM-augmented XR systems that effectively blend advanced language intelligence with immersive spatial computing.

References

[1] Page, M. J.; McKenzie, J. E.; Bossuyt, P. M.; Boutron, I.; Hoffmann, T. C.; Mulrow, C. D.; Shamseer, L.; Tetzlaff, J. M.; Akl, E. A.; Brennan, S. E.; others The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj* **2021**, *372*.

[2] Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; others Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. Proceedings of the 32nd ACM International Conference on Multimedia. 2024; pp 11198–11201.

[3] Wang, Z.; Lorraine, J.; Wang, Y.; Su, H.; Zhu, J.; Fidler, S.; Zeng, X. LLaMA-Mesh: Unifying 3D Mesh Generation with Language Models. 2024; `https://arxiv.org/abs/2411.09595`.

[4] Yang, X.; Man, Y.; Chen, J.-K.; Wang, Y.-X. SceneCraft: Layout-Guided 3D Scene Generation. Advances in Neural Information Processing Systems. 2024.

[5] Fang, S.; Wang, Y.; Tsai, Y.-H.; Yang, Y.; Ding, W.; Zhou, S.; Yang, M.-H. Chat-Edit-3D: Interactive 3D Scene Editing via Text Prompts. Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLII. 2024.

[6] Wu, Q.; Zhao, Y.; Wang, Y.; Liu, X.; Tai, Y.-W.; Tang, C.-K. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. 2024; `https://arxiv.org/abs/2405.17013`.

[7] De La Torre, F.; Fang, C. M.; Huang, H.; Banburski-Fahey, A.; Amores Fernandez, J.; Lanier, J. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024.

[8] Zhang, L.; Pan, J.; Gettig, J.; Oney, S.; Guo, A. VRCopilot: Authoring 3D Layouts with Generative AI Models in VR. Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. 2024.

[9] Chen, J.; Wu, X.; Lan, T.; Li, B. LLMER: Crafting Interactive Extended Reality Worlds with JSON Data Generated by Large Language Models. http://arxiv.org/abs/2502.02441.

[10] Jiang, J.; Xiao, W.; Lin, Z.; Zhang, H.; Ren, T.; Gao, Y.; Lin, Z.; Cai, Z.; Yang, L.; Liu, Z. SOLAMI: Social Vision-Language-Action Modeling for Immersive Interaction with 3D Autonomous Characters. 2024.

[11] Ouyang, L. et al. Training Language Models to Follow Instructions with Human Feedback. http://arxiv.org/abs/2203.02155.

[12] Rychert, A.; Ganuza, M. L.; Selzer, M. N. Integrating GPT as an Assistant for Low-Cost Virtual Reality Escape-Room Games. *44*, 14–25.

[13] Yin, Z.; Wang, Y.; Papatheodorou, T.; Hui, P. Text2VRScene: Exploring the Framework of Automated Text-driven Generation System for VR Experience. 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). pp 701–711.

[14] Konenkov, M.; Lykov, A.; Trinitatova, D.; Tsetserukou, D. VR-GPT: Visual Language Model for Intelligent Virtual Reality Applications. http://arxiv.org/abs/2405.11537.

[15] Yang, J.; Dong, Y.; Liu, S.; Li, B.; Wang, Z.; Jiang, C.; Tan, H.; Kang, J.; Zhang, Y.; Zhou, K.; Liu, Z. Octopus: Embodied Vision-Language Programmer from Environmental Feedback. http://arxiv.org/abs/2310.08588.

[16] Bovo, R.; Abreu, S.; Ahuja, K.; Gonzalez, E. J.; Cheng, L.-T.; Gonzalez-Franco, M. EmBARDiment: An Embodied AI Agent for Productivity in XR. http://arxiv.org/abs/2408.08158.

[17] Yousri, R.; Essam, Z.; Kareem, Y.; Sherief, Y.; Gamil, S.; Safwat, S. IllusionX: An LLM-powered Mixed Reality Personal Companion. http://arxiv.org/abs/2402.07924.

[18] Alayrac, J.-B. et al. Flamingo: A Visual Language Model for Few-Shot Learning. http://arxiv.org/abs/2204.14198.

[19] Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; Silver, D.; Johnson, M.; Antonoglou, I.; etc. Gemini: A Family of Highly Capable Multimodal Models. http://arxiv.org/abs/2312.11805.

[20] Zhang, L.; Pan, J.; Gettig, J.; Oney, S.; Guo, A. VRCopilot: Authoring 3D Layouts with Generative AI Models in VR. Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp 1–13.

[21] Wanna, S.; Parra, F.; Valner, R.; Kruusamäe, K.; Pryor, M. Multimodal Grounding for Embodied AI via Augmented Reality Headsets for Natural Language Driven Task Planning. http://arxiv.org/abs/2304.13676.

[22] Javaheri, H.; Ghamarnejad, O.; Lukowicz, P.; Stavrou, G. A.; Karolus, J. ARAS: LLM-Supported Augmented Reality Assistance System for Pancreatic Surgery. Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp 176–180.

[23] Dogan, M. D.; Gonzalez, E. J.; Ahuja, K.; Du, R.; Colaço, A.; Lee, J.; Gonzalez-Franco, M.; Kim, D. Augmented Object Intelligence with XR-Objects. Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp 1–15.

[24] Wan, H.; Zhang, J.; Suria, A. A.; Yao, B.; Wang, D.; Coady, Y.; Prpa, M. Building LLM-based AI Agents in Social Virtual Reality. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp 1–7.

[25] Song, T.; Pabst, F.; Eck, U.; Navab, N. Enhancing Patient Acceptance of Robotic Ultrasound through Conversational Virtual Agent and Immersive Visualizations. `http://arxiv.org/abs/2502.10088`.

[26] Li, Z.; Zhang, H.; Peng, C.; Peiris, R. Exploring Large Language Model-Driven Agents for Environment-Aware Spatial Interactions and Conversations in Virtual Reality Role-Play Scenarios.

[27] Chen, J.; Grubert, J.; Kristensson, P. O. Analyzing Multimodal Interaction Strategies for LLM-Assisted Manipulation of 3D Scenes. `http://arxiv.org/abs/2410.22177`.

[28] Pei, J.; Viola, I.; Huang, H.; Wang, J.; Ahsan, M.; Ye, F.; Yiming, J.; Sai, Y.; Wang, D.; Chen, Z.; Ren, P.; Cesar, P. Autonomous Workflow for Multimodal Fine-Grained Training Assistants Towards Mixed Reality. Findings of the Association for Computational Linguistics: ACL 2024. pp 4051–4066.

[29] Jeong, E.; Kim, H.; Park, S.; Yoon, S.; Ahn, J.; Woo, W. Function-Adaptive Affordance Extraction from 3D Objects Using LLM for Interaction Authoring with Augmented Artifacts. 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). pp 205–208.

[30] Xing, Y.; Liu, Q.; Wang, J.; Gómez-Zará, D. sMoRe: Enhancing Object Manipulation and Organization in Mixed Reality Spaces with LLMs and Generative AI. *abs/2411.11752*.

[31] Christiansen, F. R.; Hollensberg, L. N.; Jensen, N. B.; Julsgaard, K.; Jespersen, K. N.; Nikolov, I. Exploring Presence in Interactions with LLM-Driven NPCs: A Comparative Study of Speech Recognition and Dialogue Options. 30th ACM Symposium on Virtual Reality Software and Technology. pp 1–11.

[32] Li, Z.; Babar, P. P.; Barry, M.; Peiris, R. L. Exploring the Use of Large Language Model-Driven Chatbots in Virtual Reality to Train Autistic Individuals in Job Communication Skills. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp 1–7.

[33] Shoa, A.; Oliva, R.; Slater, M.; Friedman, D. Sushi with Einstein: Enhancing Hybrid Live Events with LLM-Based Virtual Humans. Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. pp 1–6.

[34] Sonlu, S.; Bendiksen, B.; Durupinar, F.; Güdükbay, U. The Effects of Embodiment and Personality Expression on Learning in LLM-based Educational Agents. `http://arxiv.org/abs/2407.10993`.

[35] Torre, F. D. L.; Fang, C. M.; Huang, H.; Banburski-Fahey, A.; Fernandez, J. A.; Lanier, J. LLMR: Real-time Prompting of Interactive Worlds Using Large Language Models. `http://arxiv.org/abs/2309.12276`.

[36] Kurai, R.; Hiraki, T.; Hiroi, Y.; Hirao, Y.; Perusquía-Hernández, M.; Uchiyama, H.; Kiyokawa, K. MagicItem: Dynamic Behavior Design of Virtual Objects With Large Language Models in a Commercial Metaverse Platform. *13*, 19132–19143.

[37] Khelifi, S.; Morris, A. Mixed Reality IoT Smart Environments with Large Language Model Agents. 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS). pp 1–7.

[38] Srinidhi, S.; Lu, E.; Rowe, A. XaiR: An XR Platform That Integrates Large Language Models with the Physical World. 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp 759–767.

[39] Khan, D.; Liu, X.; Mena, O.; Jia, D.; Kouyoumdjian, A.; Viola, I. LoXR: Performance Evaluation of Locally Executing LLMs on XR Devices. http://arxiv.org/abs/2502.15761.

[40] Xu, Y. Open Sharing and Cross-border Integration of Art Laboratory Resources Based on LLM and Virtual Reality. 2024 International Conference on Interactive Intelligent Systems and Techniques (IIST). pp 441–445.

[41] Cheng, A. Y.; Guo, M.; Ran, M.; Ranasaria, A.; Sharma, A.; Xie, A.; Le, K. N.; Vinaithirthan, B.; Luan, S. T.; Wright, D. T. H.; Cuadra, A.; Pea, R.; Landay, J. A. Scientific and Fantastical: Creating Immersive, Culturally Relevant Learning Experiences with Augmented Reality and Large Language Models. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–23.

[42] Choi, Y.; Jeong, D.; Kim, B.; Han, K. Early Prediction of Cybersickness in Virtual Reality Using a Large Language Model for Multimodal Time Series Data. Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp 25–29.

[43] Wang, Z.; Yuan, L.-P.; Wang, L.; Jiang, B.; Zeng, W. VirtuWander: Enhancing Multi-modal Interaction for Virtual Tour Guidance through Large Language Models. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–20.

[44] Chen, L.; Cai, Y.; Wang, R.; Ding, S.; Tang, Y.; Hansen, P.; Sun, L. Supporting Text Entry in Virtual Reality with Large Language Models. 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). pp 524–534.

[45] Yang, B.; He, L.; Liu, K.; Yan, Z. VIAssist: Adapting Multi-modal Large Language Models for Users with Visual Impairments. http://arxiv.org/abs/2404.02508.

[46] Wen, S.; Middleton, M.; Ping, S.; Chawla, N. N.; Wu, G.; Feest, B. S.; Nadri, C.; Liu, Y.; Kaber, D.; Zahabi, M.; McMahan, R. P.; Castelo, S.; Mckendrick, R.; Qian, J.; Silva, C. AdaptiveCoPilot: Design and Testing of a NeuroAdaptive LLM Cockpit Guidance System in Both Novice and Expert Pilots. http://arxiv.org/abs/2501.04156.

[47] Elfleet, M.; Chollet, M. Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. Proceedings of the ACM International Conference on Intelligent Virtual Agents. pp 1–9.

[48] Constantinides, N.; Constantinides, A.; Koukopoulos, D.; Fidas, C.; Belk, M. CulturAI: Exploring Mixed Reality Art Exhibitions with Large Language Models for Personalized Immersive Experiences. Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. pp 102–105.

[49] Ajri, S. J.; Nguyen, D.; Agarwal, S.; Padala, A. K. R.; Yildirim, C. Virtual AIVantage: Leveraging Large Language Models for Enhanced VR Interview Preparation among Underrepresented Professionals in Computing. Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia. pp 535–537.

[50] Fang, S.; Wang, Y.; Tsai, Y.-H.; Yang, Y.; Ding, W.; Zhou, S.; Yang, M.-H. Chat-Edit-3D: Interactive 3D Scene Editing via Text Prompts. http://arxiv.org/abs/2407.06842.

[51] Thawakar, O. C.; Shaker, A. M.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; Khan, F. XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models. Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. Bangkok, Thailand, 2024; pp 440–448.

[52] Waisberg, E.; Ong, J.; Masalkhi, M.; Zaman, N.; Sarker, P.; Lee, A. G.; Tavakkoli, A. Meta Smart Glasses—Large Language Models and the Future for Assistive Glasses for Individuals with Vision Impairments. *38*, 1036–1038.

[53] Ardiny, H.; Khanmirza, E. The Role of AR and VR Technologies in Education Developments: Opportunities and Challenges. 2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM). pp 482–487.

[54] Hajahmadi, S.; Clementi, L.; Jiménez López, M. D.; Marfia, G. ARELE-bot: Inclusive Learning of Spanish as a Foreign Language Through a Mobile App Integrating Augmented Reality and ChatGPT. 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp 335–340.

[55] Gu, R.; Gu, X. Application of Large Language Models in Teaching Traditional Chinese Medicine Diagnostics. *China Medical Herald* **2024**, *26*, 737–741, In Chinese.

[56] Bao, J.; Li, J.; Yuan, Y.; Lyu, C.; Wang, S. Augmented Reality Assembly Method Assisted by Large Language Models. *Aeronautical Manufacturing Technology* **2024**, *67*, 107–116, In Chinese.

[57] Liu, Z.; Zhu, Z.; Zhu, L.; Jiang, E.; Hu, X.; Peppler, K. A.; Ramani, K. ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–17.

[58] Driess, D. et al. PaLM-E: An Embodied Multimodal Language Model. http://arxiv.org/abs/2303.03378.

[59] Alabood, L.; Dow, T.; Feeley, K. B.; Jaswal, V. K.; Krishnamurthy, D. From Letterboards to Holograms: Advancing Assistive Technology for Nonspeaking Autistic Individuals with the HoloBoard. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–18.

[60] Alghamdi, A.; Mohaisen, D. Through the Looking Glass: LLM-Based Analysis of AR/VR Android Applications Privacy Policies. http://arxiv.org/abs/2501.19223.

[61] Beltagy, I.; Peters, M. E.; Cohan, A. Longformer: The Long-Document Transformer. http://arxiv.org/abs/2004.05150.

[62] Borg, A.; Parodis, I.; Skantze, G. Creating Virtual Patients Using Robots and Large Language Models: A Preliminary Study with Medical Students. Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. pp 273–277.

[63] Bozkir, E.; Özdel, S.; Lau, K. H. C.; Wang, M.; Gao, H.; Kasneci, E. Embedding Large Language Models into Extended Reality: Opportunities and Challenges for Inclusion, Engagement, and Privacy. ACM Conversational User Interfaces 2024. pp 1–7.

[64] Chen, M. et al. Evaluating Large Language Models Trained on Code. `http://arxiv.org/abs/2107.03374`.

[65] Cui, Y.; Ge, L. W.; Ding, Y.; Harrison, L.; Yang, F.; Kay, M. Promises and Pitfalls: Using Large Language Models to Generate Visualization Items. *31*, 1094–1104.

[66] Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. `http://arxiv.org/abs/2305.14314`.

[67] Dongye, X.; Weng, D.; Jiang, H.; Chen, P. Learning Personalized Agent for Real-Time Face-to-Face Interaction in VR. 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp 759–760.

[68] Du, M.; Liu, N.; Hu, X. Techniques for Interpretable Machine Learning. *63*, 68–77.

[69] Gallardo, A.; Choy, C.; Juneja, J.; Bozkir, E.; Cobb, C.; Bauer, L.; Cranor, L. Speculative Privacy Concerns about AR Glasses Data Collection. *2023*, 416–435.

[70] Giunchi, D.; Numan, N.; Gatti, E.; Steed, A. DreamCodeVR: Towards Democratizing Behavior Design in Virtual Reality with Speech-Driven Programming. 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). pp 579–589.

[71] Gunawardhana, B. S.; Zhang, Y.; Sun, Q.; Deng, Z. Toward User-Aware Interactive Virtual Agents: Generative Multi-Modal Agent Behaviors in VR. 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp 1068–1077.

[72] Hartholt, A.; Fast, E.; Reilly, A.; Whitcup, W.; Liewer, M.; Mozgai, S. Ubiquitous Virtual Humans: A Multi-platform Framework for Embodied AI Agents in XR. 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). pp 308–3084.

[73] He, Z.; Li, S.; Song, Y.; Cai, Z. Towards Building Condition-Based Cross-Modality Intention-Aware Human-AI Cooperation under VR Environment. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–13.

[74] Hoffmann, J. et al. Training Compute-Optimal Large Language Models. `http://arxiv.org/abs/2203.15556`.

[75] Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; Fei-Fei, L. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. `http://arxiv.org/abs/2307.05973`.

[76] Huang, J.; Yong, S.; Ma, X.; Linghu, X.; Li, P.; Wang, Y.; Li, Q.; Zhu, S.-C.; Jia, B.; Huang, S. An Embodied Generalist Agent in 3D World. http://arxiv.org/abs/2311.12871.

[77] Hu, Z.; Iscen, A.; Jain, A.; Kipf, T.; Yue, Y.; Ross, D. A.; Schmid, C.; Fathi, A. SceneCraft: An LLM Agent for Synthesizing 3D Scenes as Blender Code.

[78] Janaka, N.; Cai, R.; Zhao, S.; Hsu, D. Demonstrating PANDALens: Enhancing Daily Activity Documentation with AI-assisted In-Context Writing on OHMD. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp 1–7.

[79] Jiang, H.; Sun, W.; Guo, H.; Zeng, J.; Xue, X.; Li, S. Review of Intelligent Diagnosis Methods for Imaging Gland Cancer Based on Machine Learning. *5*, 293–316.

[80] Jin, X.; Tong, W.; Wei, X.; Wang, X.; Kuang, E.; Mo, X.; Qu, H.; Fan, M. Exploring the Opportunity of Augmented Reality (AR) in Supporting Older Adults to Explore and Learn Smartphone Applications. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–18.

[81] Kahng, M.; Tenney, I.; Pushkarna, M.; Liu, M. X.; Wexler, J.; Reif, E.; Kallarackal, K.; Chang, M.; Terry, M.; Dixon, L. LLM Comparator: Interactive Analysis of Side-by-Side Evaluation of Large Language Models. *31*, 503–513.

[82] Kim, Y.; Aamir, Z.; Singh, M.; Boorboor, S.; Mueller, K.; Kaufman, A. E. Explainable XR: Understanding User Behaviors of XR Environments Using LLM-assisted Analytics Framework. http://arxiv.org/abs/2501.13778.

[83] Kim, B.; Kim, M.; Seo, D.; Kim, B. Leveraging Large Language Models for Active Merchant Non-player Characters. http://arxiv.org/abs/2412.11189.

[84] Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention. http://arxiv.org/abs/2309.06180.

[85] Li, X.; Zhang, M.; Geng, Y.; Geng, H.; Long, Y.; Shen, Y.; Zhang, R.; Liu, J.; Dong, H. ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation. http://arxiv.org/abs/2312.16217.

[86] Li, W.; Yu, Z.; She, Q.; Yu, Z.; Lan, Y.; Zhu, C.; Hu, R.; Xu, K. LLM-enhanced Scene Graph Learning for Household Rearrangement. SIGGRAPH Asia 2024 Conference Papers. pp 1–11.

[87] Li, Z.; Gebhardt, C.; Inglin, Y.; Steck, N.; Streli, P.; Holz, C. SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning. Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp 1–13.

[88] Liang, C. X.; Tian, P.; Yin, C. H.; Yua, Y.; An-Hou, W.; Ming, L.; Wang, T.; Bi, Z.; Liu, M. A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision-Language Tasks. http://arxiv.org/abs/2411.06284.

[89] Lingyun, B.; Huang, Z.; Lin, Z.; Sun, Y.; Chen, H.; Li, Y.; Li, Z.; Yuan, X.; Xu, L.; Tan, T. Automatic Detection of Breast Lesions in Automated 3D Breast Ultrasound with Cross-Organ Transfer Learning. *6*, 239–251.

[90] Liu, C.; Cheung, C. S. C.; Xu, M.; Zhang, Z.; Su, M.; Fan, M. Toward Facilitating Search in VR With the Assistance of Vision Large Language Models. 30th ACM Symposium on Virtual Reality Software and Technology. pp 1–14.

[91] Min, Y.; Jeong, J.-W. Public Speaking Q&A Practice with LLM-Generated Personas in Virtual Reality. 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). pp 493–496.

[92] Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large Language Models: A Survey. `http://arxiv.org/abs/2402.06196`.

[93] Numan, N.; Rajaram, S.; Kumaravel, B. T.; Marquardt, N.; Wilson, A. D. SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending. Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp 1–25.

[94] Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; Bernstein, M. S. Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp 1–22.

[95] Park, G. W. W.; Panda, P.; Tankelevitch, L.; Rintel, S. CoExplorer: Generative AI Powered 2D and 3D Adaptive Interfaces to Support Intentionality in Video Meetings. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp 1–10.

[96] Patil, S. G.; Zhang, T.; Wang, X.; Gonzalez, J. E. Gorilla: Large Language Model Connected with Massive APIs. `http://arxiv.org/abs/2305.15334`.

[97] Rist, T. Using a Large Language Model to Turn Explorations of Virtual 3D-Worlds into Interactive Narrative Experiences. 2024 IEEE Conference on Games (CoG). pp 1–8.

[98] Su, X.; Koh, E.; Xiao, C. SonifyAR: Context-Aware Sound Effect Generation in Augmented Reality. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp 1–7.

[99] Suzuki, R.; Gonzalez-Franco, M.; Sra, M.; Lindlbauer, D. XR and AI: AI-Enabled Virtual, Augmented, and Mixed Reality. Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp 1–3.

[100] Tang, Y.; Situ, J.; Huang, Y. Beyond User Experience: Technical and Contextual Metrics for Large Language Models in Extended Reality. Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp 640–643.

[101] Teo, T.; Lawrence, L.; Lee, G. A.; Billinghurst, M.; Adcock, M. Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp 1–14.

[102] Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models. http://arxiv.org/abs/2302.13971.

[103] Uhl, J. C.; Gutierrez, R.; Regal, G.; Schrom-Feiertag, H.; Schuster, B.; Tscheligi, M. Choosing the Right Reality: A Comparative Analysis of Tangibility in Immersive Trauma Simulations. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–17.

[104] Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; Hajishirzi, H. Self-Instruct: Aligning Language Models with Self-Generated Instructions. http://arxiv.org/abs/2212.10560.

[105] Wang, Z.; Lorraine, J.; Wang, Y.; Su, H.; Zhu, J.; Fidler, S.; Zeng, X. LLaMA-Mesh: Unifying 3D Mesh Generation with Language Models. http://arxiv.org/abs/2411.09595.

[106] Wang, X.; Su, Z.; Rekimoto, J.; Zhang, Y. Watch Your Mouth: Silent Speech Recognition with Depth Sensing. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–15.

[107] Wang, H. W.; Gordon, M.; Battle, L.; Heer, J. DracoGPT: Extracting Visualization Design Preferences from Large Language Models. *31*, 710–720.

[108] Wu, G.; Qian, J.; Castelo Quispe, S.; Chen, S.; Rulff, J.; Silva, C. ARTiST: Automated Text Simplification for Task Guidance in Augmented Reality. Proceedings of the CHI Conference on Human Factors in Computing Systems. pp 1–24.

[109] Wu, Q.; Zhao, Y.; Wang, Y.; Liu, X.; Tai, Y.-W.; Tang, C.-K. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. http://arxiv.org/abs/2405.17013.

[110] Yang, Y.; Sun, F.-Y.; Weihs, L.; VanderBilt, E.; Herrasti, A.; Han, W.; Wu, J.; Haber, N.; Krishna, R.; Liu, L.; Callison-Burch, C.; Yatskar, M.; Kembhavi, A.; Clark, C. Holodeck: Language Guided Generation of 3D Embodied AI Environments. http://arxiv.org/abs/2312.09067.

[111] Yan, K.; Wang, Z.; Ji, L.; Wang, Y.; Duan, N.; Ma, S. Voila-A: Aligning Vision-Language Models with User's Gaze Attention.

[112] Yoffe, L.; Sharma, A.; Höllerer, T. OCTOPUS: Open-vocabulary Content Tracking and Object Placement Using Semantic Understanding in Mixed Reality. 2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). pp 587–588.