iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** XI **Month of publication:** November 2026

DOI: https://doi.org/10.22214/ijraset.2026.85436

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

Exploration Research on Intelligent Generation Application of VR Virtual Scenes Driven by Artificial Intelligence and Digitalization

Xin YuYang[1], Yuan YuanLiu[2]
Guangdong Atv College Of Performing Arts, Zhaoqing Guangdong, China.
Guangdong Atv College Of Performing Arts, Zhaoqing Guangdong, China.

Abstract: ***The large-scale and efficient production of virtual reality (VR) content remains a critical bottleneck constraining industry development. Traditional scene construction methods relying on manual modeling and scripted programming exhibit prominent issues such as prolonged cycles, high costs, and limited customization. Recent breakthroughs in artificial intelligence technologies—including generative AI (AIGC), large language models, and neural rendering—have opened new pathways for intelligent VR scene generation. This study systematically examines the theoretical foundations and technical architectures of VR scene generation driven by AI technologies, delves into key technical approaches including geometric appearance generation, physical logic mapping, and ambient environment intelligence, conducts empirical analyses through domestic and international case studies, and explores cutting-edge directions such as edge intelligence, embodied intelligence, and world models. Research demonstrates that AI-driven technologies can enhance VR scene construction efficiency by over 80% while reducing costs by approximately 70%. Building on these findings, the paper proposes actionable recommendations across three dimensions: technological innovation, industrial applications, and ethical standards, aiming to provide theoretical references and practical guidance for developing next-generation VR content production paradigms characterized by high efficiency, fidelity, and strong interactivity.***

Keywords: ***Artificial Digital Intelligence, Virtual Reality, Scene Intelligent Generation, Neural Radiance Field, Multimodal Large Models, Aigc.***

I. INTRODUCTION

A. Research Background and Significance

The virtual reality industry has undergone a paradigm shift from "manual modeling" to "intelligent generation." After decades of development, VR technology has demonstrated significant application value in fields such as education and training, industrial simulation, and cultural tourism experiences. However, high-quality VR scene construction has long relied on traditional computer graphics pipelines. Artists manually perform modeling, UV mapping, texture rendering, and material customization using tools like Maya, 3ds Max, and C4D, while programmers handle scene assembly, lighting baking, and interaction logic programming. This workflow presents three major limitations: First, low production efficiency—creating a high-precision urban scene often requires months of work by dozens of team members. Second, content output capacity fails to keep pace with hardware adoption rates, hindering VR ecosystem development. Third, there is a scarcity of personalized content, making it difficult to meet users' demand for immersive experiences tailored to individual preferences.

With the rapid advancement of artificial intelligence technology, particularly breakthroughs in generative AI across text, images, and audiovisual domains, VR content production is undergoing a fundamental paradigm shift from "manual modeling" to "intelligent generation." At the core of this transformation lies the integration of human creativity with machine-generated capabilities, enabling rapid construction and dynamic evolution of 3D environments through multimodal inputs such as natural language, sketches, and voice commands.

©IJRASET: All Rights are Reserved

The breakthrough value of artificial intelligence-driven digital intelligence (AIGC) + big data + intelligent decision-making in VR content production lies in the proposed "artificial digital intelligence" concept, which represents an organic integration of AI and digital intelligence encompassing three core dimensions: First, generative AIGC endows machines with content creation capabilities from scratch. Second, big data provides massive training samples and semantic knowledge bases. Third, intelligent decision-making ensures generated content maintains logical coherence and interactive adaptability. The synergistic integration of these three elements demonstrates transformative potential for VR content production bottlenecks by achieving efficiency leaps—shortening traditional scenario construction cycles from weeks/months to minutes/hours, enabling scalable content supply while driving cost restructuring. This reduces reliance on professional modeling talent, empowers non-professional users to participate in creation, fosters prosperity in user-generated content ecosystems, achieves personalized adaptation through real-time scenario adjustments based on user preferences and behavioral data, and delivers dynamically adaptive VR experiences. Finally, it deepens virtual-real integration by leveraging neural rendering technology to bridge the digital divide between physical and virtual worlds, enabling hyper-realistic scene replication and digital twin generation.

Review of Current Research Status at Home and Abroad

B. Limitations of Traditional VR Scene Construction Technology Based on Geometry and Images

Traditional VR scene construction technologies primarily fall into two categories: geometric modeling methods and image rendering approaches. Geometry-based modeling methods rely on manual modeling processes that depend on artists' technical expertise. While offering strong controllability and high precision, these methods have notable limitations: prolonged modeling cycles, difficulty in accurately representing geometric details of complex environments, material and lighting physics accuracy dependent on rendering experience, and near-impossible implementation for large-scale scenarios. Image-based rendering methods, represented by panoramic imaging and light field technology, generate VR scenes through capturing real-world image data. Although capable of delivering rapid high-fidelity visual effects, this approach suffers from limited interactivity—users can only view scenes from fixed viewpoints without free movement or interaction with objects. Additionally, image acquisition equipment remains expensive, and post-processing requires complex workflows.

Advances in Generative AI within the 3D domain have seen explosive growth in research, particularly in text/image-to-3D model generation. Technologies like Hunyuan 3D and Stable 3D have achieved end-to-end generation from textual descriptions to 3D point clouds or mesh models. These models typically employ architectures combining diffusion models with latent space encoding, enabling rapid creation of usable 3D assets within minutes. In neural rendering and scene reconstruction, NeRF neural radiance fields and their improved variants (e.g., Mip-NeRF and Instant-NGP) have achieved high-fidelity continuous 3D scene reconstruction using sparse images. The 3D Gaussian Splatting technique further breaks through real-time rendering bottlenecks, allowing neural rendering to support interactive applications in VR headsets. With the widespread adoption of large language models like GPT-4 and LLaMA, multimodal-driven scene generation has become a key research focus. Current studies explore deep integration of language understanding capabilities with 3D generation, such as leveraging large language models to interpret user intent and intelligently orchestrate generation modules to complete creative tasks including scene layout, object placement, and material configuration.

C. Research Content, Methods, and Innovations

This study focuses on the core theme of "Intelligent Generation of VR Virtual Scenes Driven by Artificial Intelligence and Digitalization," establishing a comprehensive research framework spanning theory to practice and technology to application. At the theoretical level, it systematically analyzes the core technological support system and proposes a three-tier architecture for intelligent generation systems comprising data layers, model layers, and interaction layers. Technologically, it delves into three key technical pathways: geometric appearance generation, physical logic mapping, and ambient environment intelligent generation. Application-wise, it conducts empirical analyses through exemplary cases from domestic and international contexts to validate

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

technical feasibility. The study also explores cutting-edge directions including edge intelligence, embodied intelligence, world models, and brain-computer interfaces, while proposing implementation recommendations across three dimensions: technological development, industrial application, and ethical standards.

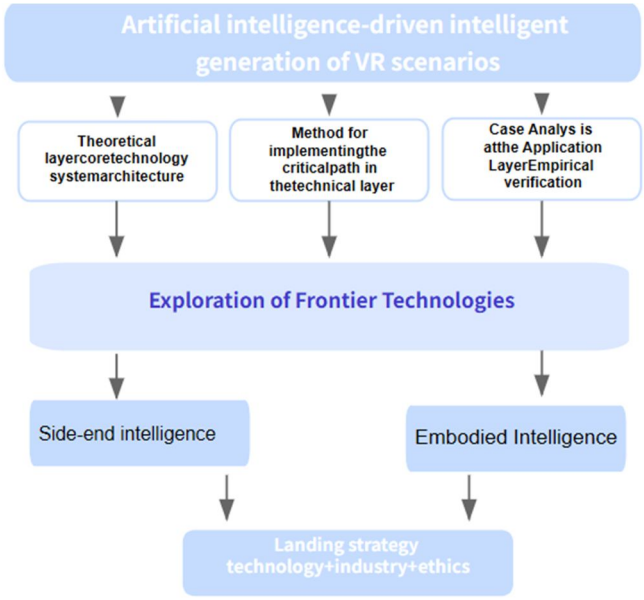


Fig. 1 Research Framework Diagram of This Study

The key technological innovations are primarily manifested in the systematic construction of an end-to-end generation paradigm, which breaks through traditional "stepwise" modeling processes. This framework proposes an intelligent end-to-end generation system that transitions from user intent input to interactive VR scene output, achieving integrated generation of geometric structures, visual aesthetics, physical properties, logical logic, and environmental atmospheres. Leveraging multimodal fusion technology, it enables efficient semantic understanding and scene mapping by unifying multi-modal inputs including text, images, audio, and gestures. Through large-scale multimodal models, the system accomplishes high-level semantic comprehension and precisely maps to critical 3D scene parameters such as layout design, material textures, lighting effects, and interactive interactions. Additionally, a closed-loop feedback mechanism integrating generation and interaction is established. By incorporating real-time user feedback and behavioral data, the system forms a "generation-experience-feedback-optimization" closed-loop system, enabling dynamic adaptive evolution of VR scenes tailored to user preferences.

II. THEORETICAL FOUNDATIONS AND TECHNICAL ARCHITECTURE OF VIRTUAL SCENE GENERATION DRIVEN BY ARTIFICIAL INTELLIGENCE AND DIGITALIZATION

A. Core Technology Support System

1) Application of Diffusion Models and Transformers in 3D Generation

Generative AI serves as the core driving force behind artificial intelligence systems. In 3D content generation, diffusion models and Transformer architectures demonstrate exceptional generative capabilities. Diffusion models achieve data generation by incrementally adding noise to input data and learning inverse denoising processes, enabling the transition from random noise to target datasets. These models are applied across multiple dimensions in 3D generation: First, for point cloud generation, Point-E utilizes text encoders combined with point cloud diffusion models to produce coarse point clouds, followed by detail enhancement through upsample networks. Second, for texture and material generation, 2D diffusion models employ multi-view projection techniques and consistency constraints to automatically map high-quality textures onto generated mesh models. Transformer architectures play a

pivotal role in 3D generation through their robust sequence modeling capabilities. By treating 3D data (e.g., point cloud sequences and mesh vertex sequences) as sequential inputs, they employ self-attention mechanisms to capture long-range dependencies, ultimately producing coherent 3D structures.

2) Principles of Neural Radiation Field and 3D Gaussian Spraying in Scene Reconstruction

Neural radiance fields implicitly represent 3D scenes through multi-layer perceptrons (MLPs), processing spatial coordinates and viewing angles to output color and density values. Their core advantage lies in reconstructing continuous, high-fidelity 3D scenes from sparse multi-view images, demonstrating exceptional capability in capturing complex lighting effects and semi-transparent materials. However, their limited real-time rendering performance makes direct application to interactive VR scene browsing challenging. 3D Gaussian splatter serves as an efficient alternative to neural radiance fields, representing scenes as collections of 3D Gaussian ellipsoids and employing tile-based rasterization rendering to achieve optimal balance between real-time rendering performance and high-fidelity quality.

3) Multimodal Large Models LMMs

Multimodal large models (LMMs)can simultaneously process diverse modalities including text, images, and speech, enabling cross-modal semantic alignment and reasoning. In VR scene generation, their core functionality manifests across multiple dimensions: By understanding user intent through textual descriptions, they accurately extract key semantic elements such as scene type, style, atmosphere, and critical objects. Leveraging cross-modal retrieval capabilities, these models match corresponding models and textures from large-scale 3D asset libraries based on semantic descriptions. Integrating multidimensional inputs like image sketches, voice commands, and gesture actions further enhances scene generation precision. Additionally, they can make interactive decisions in dynamic scenarios according to user behavior and environmental states, generating real-time dialogue content and behavioral responses for NPCs.

B. Intelligent Generation System Architecture Design

1) Preprocessing of Multi-source Heterogeneous Data

The data layer serves as the foundational support for VR virtual scene intelligent generation systems, undertaking core functions such as multi-source heterogeneous data collection, preprocessing, standardized annotation, structured storage, and efficient scheduling. It provides high-quality data assurance for upper-layer model training, scene generation, and interactive decision-making. Its data sources primarily encompass three categories: The first category consists of real-scene scanning data collected through devices like LiDAR, depth cameras, drone aerial surveys, and multi-angle photography, generating 3D point clouds, RGB images, and depth information. After processing through point cloud registration, pose optimization, and texture fusion, these datasets reconstruct high-fidelity 3D base models that serve as physical spatial references for digital twins and real-scene replication. The second category includes open-source 3D datasets such as Objaverse-XL, ShapeNet, and 3D-FUTURE, containing massive collections of pre-classified, annotated, and normalized 3D models, material textures, and scene components. These datasets enable direct use in model pre-training, fine-tuning, and cross-modal retrieval, significantly enhancing model generalization capabilities and generation efficiency. The third category comprises semantic knowledge bases integrating prior knowledge such as scene layout specifications, object spatial relationships, physical attribute constraints, and lighting-material matching logic. Stored in structured knowledge graph formats, these resources constrain the generation process to ensure spatial coherence, object relationships, and logical consistency in VR scenes.

2) Collaborative Model Evolving from "Unimodal Generation" to "Scenario-Element-Comprehensive Generation"

The model layer serves as the core computational unit of the VR virtual scene intelligent generation system, composed of multiple complementary and collaboratively functioning generation models that collectively enable intelligent construction from individual

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

assets to complete interactive environments. The 3D asset generation model primarily focuses on creating standalone 3D objects. It automatically generates grid models with standardized topological structures, high-definition texture maps, and corresponding physical material parameters based on user input text descriptions or reference images. Technically implemented using diffusion models or Generative Adversarial Networks (GAN) architectures, it balances generation quality and diversity. The scene layout generation model emphasizes holistic spatial structure planning. Leveraging Transformer or graph neural networks to analyze spatial distribution patterns and semantic constraints from massive scene datasets, it autonomously designs logical scene layouts—including furniture placement positions, scale proportions, and compositional relationships in indoor spaces—ensuring alignment with everyday cognition and spatial logic. The material and lighting generation model adopts scene semantic ambiance as its guiding principle. For stylistic scenarios like "cozy living rooms" or "mysterious corridors," it intelligently matches and generates appropriate material textures, reflection/refraction properties, and global illumination schemes to achieve seamless visual ambiance-semantics integration. The interaction logic generation model utilizes large language models' comprehension and generation capabilities to convert scene requirements into executable interaction rules—including trigger conditions, animation state machines, NPC behavior trees, and interaction logic code—endowing static 3D scenes with responsive and interactive dynamic capabilities.

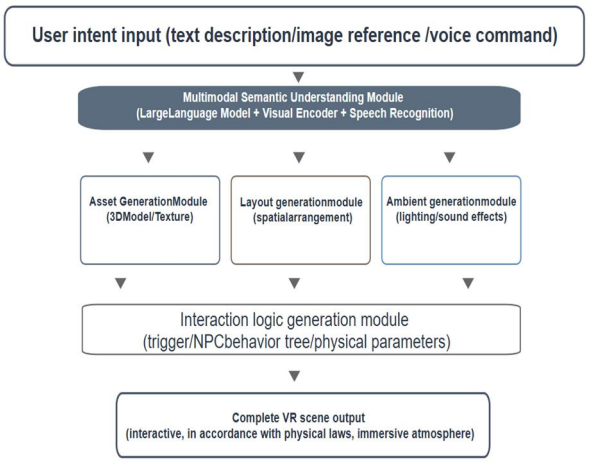


Fig. 2 Collaborative Model Architecture for Comprehensive Scene Element Generation

3) Real-time Rendering and Closed-loop User Intent Feedback

The interaction layer serves as the critical bridge between VR virtual scene generation systems and end-users, acting as the key component for achieving the closed-loop process of "generation-experience-feedback-optimization." Its core function lies in bridging the gap between system-generated capabilities and user needs, ensuring immersive and seamless VR experiences while enabling dynamic control of scene generation through user intent. It primarily handles two core functionalities: real-time rendering and intent feedback loops. The real-time rendering mechanism converts 3D scene data generated by the model layer—including mesh models, material parameters, lighting information, and interaction logic codes—into immersive visuals compatible with VR headsets through specialized rendering engines, delivering an authentic virtual reality experience. To accommodate the dynamic generation characteristics of VR environments, rendering engines must maintain tight integration with various generation models in the model layer. This integration supports two key aspects: dynamic loading of scene data (allowing immediate rendering after partial asset generation to enhance efficiency) and incremental updates (enabling real-time rendering only for localized modifications like object adjustments, material changes, or lighting optimizations based on user feedback, thus preventing lag caused by full-scene re-rendering and maintaining interactive fluidity).

©IJRASET: All Rights are Reserved

During rendering, it is essential to address core VR experience requirements such as controlling frame rates, optimizing perspective adaptation, reducing motion sickness symptoms, and ensuring image clarity, color fidelity, and spatial depth. These aspects must align with VR headsets 'display characteristics and users' visual perception habits. The intent feedback loop mechanism establishes a two-way interaction channel between users and generative systems, enabling dynamic scenario adjustments based on user preferences. Specifically, the interaction layer collects real-time multi-modal data through VR device sensors—including eye-tracking modules, controller inputs, voice commands, posture data, and gesture movements—such as eye movement patterns, joystick operations, and voice commands. This multi-modal data is then processed by large multimodal models (LMMs)to accurately extract user intent. Whether modifying existing scenes, optimizing details, or generating new content, the system can precisely identify user requirements and translate them into executable commands. Finally, these instructions feed back to the model layer, triggering the corresponding generative model to perform real-time scene modifications and optimizations. For instance, it adjusts detail precision in key areas based on user gaze duration, moves object positions according to controller operations, and modifies materials or lighting through voice commands. This creates a complete closed-loop process of "user interaction → data collection → intent parsing → scene optimization → re-rendering," enabling dynamic adaptive scene adjustments. The resulting VR environments become more aligned with users' actual needs and usage preferences.



Fig. 3Closed-loop process for user intent feedback

III. KEY TECHNOLOGIES AND IMPLEMENTATION PATHWAYS FOR INTELLIGENT GENERATION OF VR VIRTUAL SCENES

A. Automated Generation of Geometry and Appearance

1) Text and Sketch-driven 3D Asset Generation

Automated generation of 3D assets serves as the cornerstone of VR scene construction. As one of the system's core capabilities, this technology adopts a multidimensional approach that integrates various technical architectures to achieve efficient and high-quality asset creation. Text-driven 3D asset generation utilizes pre-trained text-3D mapping models to accurately interpret user inputs like "retro wooden-textured desk" or "futuristic sci-fi console," automatically generating mesh models compliant with topological requirements while applying material parameters such as roughness and reflectivity to ensure both functionality and visual appeal. Image-driven generation reconstructs real-world object forms through feature extraction from reference images, producing 3D models that maintain original style and proportions—ideal for scenarios requiring precise object replication. For batch processing needs, the system employs model batch processing capabilities to generate multiple 3D assets of diverse types and styles simultaneously, significantly enhancing scene construction efficiency. It also supports batch optimization and adjustments of generated assets to meet varied scenario requirements.

2) Programmatic and AI Hybrid Generation for Large-Scale Urban Scenarios

The creation of city-scale VR environments demands exceptional efficiency and realism. The intelligent generation process typically employs a hybrid strategy combining "programmatic generation + AI enhancement." First, based on urban layout principles, fractal algorithms and other technologies are utilized to construct a foundational urban framework encompassing road networks, building distributions, and topographical features, ensuring structural rationality and controllability. Subsequently, AI technology refines this framework through detailed optimization, employing image restoration models to generate authentic building facade textures. Generative AI then fills in street infrastructure elements like streetlights, greenery, and signage, while neural rendering techniques deliver hyper-realistic details for landmark buildings. This integrated approach ultimately achieves efficient and high-quality urban VR environment construction.

B. Intelligent Mapping of Physical Rules and Dynamic Logic

1) AI Parameter Auto-Optimization Based on Physical Engine

The core of immersion in VR environments lies in ensuring that physical behaviors such as object movement and collisions adhere to real-world physical laws. Therefore, AI-driven parameter optimization based on physics engines serves as a critical component for maintaining scene authenticity. The fundamental approach employs AI algorithms like reinforcement learning and Bayesian optimization, guided by semantic requirements and physical realism constraints. This system automatically searches and optimizes core physics engine parameters—including object mass, friction coefficients, collision thresholds, and gravitational acceleration—eliminating the need for manual parameter adjustments. Through continuous algorithmic optimization iterations combined with real-time validation from physics simulators, the system ensures that collision feedback, motion trajectories, and force effects in VR environments strictly conform to physical principles. This methodology prevents issues like object penetration or motion anomalies that defy physical logic, enabling every virtual object to exhibit realistic physical behaviors. Such precision significantly enhances the immersive experience and realism of VR environments.

2) Automated Configuration of Dynamic Interaction Logic

Automated configuration of dynamic interaction logic serves as the cornerstone for achieving interactive VR environments. This process primarily leverages multimodal input parsing and large language model (LLM) generation capabilities to automate the construction of triggers, NPC behavioral logic, and dialogue systems. Specifically, LLMs analyze user interaction requirements to automatically generate scene triggers—such as object-touch-based scene transitions and animation playback—while creating NPC behavior trees and finite state machines based on natural language descriptions. These mechanisms establish clear behavioral rules and response logic for NPCs, enabling dynamic and intelligent scene interactions. This approach transforms static visual representations into real-time interactive environments, elevating VR experiences from passive viewing to interactive dynamic scenarios.

C. Ambient Intelligent Generation of Immersive Lighting and Sound Effects

Intelligent lighting generation technology serves as a critical foundation for enhancing the realism and immersion of VR environments, encompassing three core components: Semantic-driven light source placement involves analyzing scene semantics such as "cozy living room" or "mysterious corridor" to automatically match and determine light source types (primary/secondary/ambient lighting), positioning, color parameters, and illumination intensity, ensuring optimal alignment with scene aesthetics and user needs. Global illumination parameter op4timization utilizes neural rendering technology to dynamically adjust lighting parameters based on material properties and spatial structures, balancing contrast ratios and light-shadow transitions while preventing uneven illumination and shadow distortion. Dynamic light adaptation continuously captures user perspective

changes and scene dynamics, dynamically adjusting lighting angles, intensity, and shadow effects to synchronize visual presentation with user interactions and environmental shifts, thereby significantly enhancing VR scene realism and interactive immersion.

AI-driven spatial audio generation technology serves as the cornerstone for enhancing VR immersion, encompassing three core components: Audio asset matching focuses on scene semantics and object attribute analysis, automatically retrieving and matching suitable sound effects from audio libraries to ensure perfect alignment with environmental ambiance and object characteristics. For instance, low-frequency textures are applied to simulate wooden furniture collisions, while crisp sound effects are generated for metal object interactions. 3D audio localization calculates real-time binaural rendering parameters based on sound source positions and user head orientation, accurately reproducing spatial directionality and distance perception. This enables users to pinpoint sound sources through auditory cues, significantly improving spatial depth perception. Ambient audio generation utilizes generative AI to automatically create scene-specific sound effects matching descriptive contexts like "quiet rainforest" or "noisy streets," achieving seamless integration between audio and visual environments. This technology further enhances VR immersion through realistic spatial experiences and enhanced realism.

IV. ANALYSIS OF OUTSTANDING CASES AND EMPIRICAL RESEARCH

A. Industrial Application Cases: Digital Twin and Smart City

Case 1: The "Immersive Map" project jointly developed by Google and IMMERSE represents a flagship industrial application for intelligent VR scene generation in digital twin and smart city domains. Leveraging Google's extensive geospatial data repository and integrating two core technologies—neural radiance field and 3D Gaussian splatter—it establishes a fully automated system capable of transforming 2D data into high-fidelity VR street scenes. This breakthrough effectively addresses traditional urban scene modeling challenges including low efficiency, insufficient fidelity, and high costs, enabling efficient digital twin scenario construction and immersive experience implementation for smart cities.

In terms of technical implementation, the project leverages Google's data source advantages by integrating three core datasets: Street View imagery, satellite imagery, and drone aerial photography. This integration has created a massive global real-world data repository covering multiple cities, providing a robust foundation for high-fidelity VR scene reconstruction. Street View imagery delivers close-range details including building facades, street infrastructure, and vegetation patterns. Satellite imagery ensures macro-level accuracy in urban planning, road networks, and topography, while drone aerial photography supplements high-altitude perspectives such as building rooftops and large public squares. The synergy of these three data sources delivers comprehensive, multi-angle real-world support for scene reconstruction.

In the scene generation workflow, the project adopts a collaborative approach combining neural radiance field technology with 3D Gaussian splashing techniques. First, multi-view 2D video/image data is processed using neural radiance field technology. A multi-layer perception mechanism constructs an implicit neural radiance field to accurately capture geometric structures and surface textures, enabling implicit reconstruction from 2D data to continuous 3D scenes while ensuring realism and detail fidelity. Subsequently, 3D Gaussian splashing technology optimizes the reconstructed scene using differential Gaussian elements as explicit representation units, achieving hundreds-of-times rendering acceleration. This effectively addresses the pain points of traditional neural radiance field rendering—slow processing speeds and inadequate real-time interaction capabilities for VR. Meanwhile, dynamic adaptive mechanisms optimize spatial distribution of Gaussian elements to maintain high fidelity while reducing model storage requirements, aligning with VR device performance characteristics. Additionally, the project integrates Gemini's multimodal large model capabilities to intelligently refine scene details, precisely reconstructing architectural contours, vegetation distribution, traffic signals, and other elements to further enhance realism and practicality.

From a user experience perspective, VR headsets enable seamless access to "immersive maps," allowing free exploration of virtual cities. The system meticulously replicates real-world details including architectural structures, street layouts, and natural environments.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Integrated with intelligent lighting and spatial audio generation technologies, the virtual scenes dynamically adapt lighting effects based on user perspectives. Ambient sounds like street ambiance and vehicle traffic synchronize precisely with spatial coordinates, creating a dual immersion experience that combines visual and auditory elements. Users can intuitively perceive neighborhood landscapes and architectural details as if physically present in real cities, even previewing actual conditions of unfamiliar routes and complex intersections in advance—effectively eliminating travel uncertainties and confusion.

B. Case Study of Entertainment and Social Applications: Metaverse Content Creation

Skybox AI, developed by Blockade Labs, stands as a benchmark solution in VR environment background generation. Focusing on rapid automated creation of VR panoramic skyboxes, it delivers core advantages of "low entry barriers, high efficiency, and strong adaptability," significantly reducing costs and technical barriers for VR scene construction. Widely adopted in VR social networking, game development, virtual production, and other fields, it has become a cornerstone tool for both individual creators and enterprise applications building VR environments. Its core value lies in achieving a seamless workflow: "text/sketch input → instant generation → direct deployment," fundamentally transforming the traditional manual skybox creation process characterized by low efficiency and high complexity.

In terms of technical implementation, Skybox AI integrates generative AI with 3D rendering technology through its core Latent Diffusion Model for 3D Diffusion, jointly developed by Blockade Labs and Intel Labs. This model generates depth maps via diffusion processes to create high-fidelity, seamless 360-degree panoramic images. Supporting both text-driven and sketch-driven generation modes, it balances flexibility with precision. Users can operate seamlessly across VR devices like Quest or desktop platforms without specialized modeling skills—simply input text prompts to specify scene styles, and the AI instantly interprets semantic requirements to produce high-resolution, seamless environmental textures, effectively avoiding traditional manual production issues like mosaic gaps and texture blurring. Additionally, Skybox AI features Sketch-a-Skybox functionality, allowing users to sketch basic designs on auxiliary grids (e.g., spheres/cubes) while refining scene layouts with text prompts to generate customized skybox content. Exported textures and depth maps in formats like JPEG, PNG, EXR, and HDR facilitate post-production optimization in 3D software. Notably, the tool is developing NeRF-compatible features to automatically convert generated 2D panoramas into navigable 3D environments, expanding its application scope.

In terms of functional features and user experience, Skybox AI centers on 'efficiency, convenience, and high adaptability,' with its performance and user data fully demonstrating industrial-grade application value.

V. CUTTING-EDGE TECHNOLOGY EXPLORATION AND DEVELOPMENT TRENDS

A. Real-time Rendering Technology for Mobile VR/AR Headsets

With the widespread adoption of VR/AR technologies, mainstream mobile VR/AR headsets are accelerating their evolution toward lightweight, standalone, and portable designs, breaking free from dependence on PCs or cloud platforms to become the core medium for immersive user experiences. However, the massive computational demands of AI-driven VR/AR scene generation, coupled with limited hardware resources in mobile devices and latency issues from cloud-based data transmission, have created critical technical bottlenecks that severely limit real-time scene rendering capabilities and large-scale applications. Under traditional cloud-based generation models, scene data transmission requires a multi-step process involving "cloud computing – data transfer – terminal rendering," often resulting in screen lag, excessive latency, and disconnections that fail to meet users' core needs for real-time interaction. Direct deployment of full-scale generation models on mobile devices, however, faces challenges such as low rendering efficiency, excessive power consumption, and even operational failures due to insufficient device processing capabilities.

To address this critical bottleneck, edge-side intelligent technology has emerged as the core solution for real-time scene generation in mobile VR/AR headsets. The fundamental approach involves migrating AI generation capabilities from cloud or PC environments to local mobile VR/AR headsets. By integrating dedicated neural network processing units within headsets and running

lightweight-optimized AI models, this technology enables localized real-time scene generation, updates, and rendering—completely eliminating cloud dependency while balancing generation efficiency, scene quality, and device power consumption. Implementing edge-side intelligence requires multi-technology collaborative optimization rather than simple model migration. On one hand, mainstream 3D generation models undergo lightweight transformation through model compression techniques, employing core technologies like quantization, network pruning, and knowledge distillation to significantly reduce parameter sizes and computational demands without compromising quality, ensuring compatibility with mobile headsets 'NPU architectures. On the other hand, optimized scheduling coordination between headsets' NPU, generation models, and rendering engines adopts an "on-demand generation, incremental updates" strategy. This approach generates and renders only scene content within user viewpoints in real-time, while pre-generating or compressing background content outside view ranges. This approach reduces computational overhead, minimizes generation latency, and optimizes power allocation to ensure long-term stable operation of mobile headsets.

End-device intelligence has now entered the initial implementation phase in real-time content generation for mobile VR/AR headsets. Tech giants including Google, Meta, and Liangliang Vision are accelerating strategic deployments through dual approaches: enhancing NPU hardware performance to boost local computing capabilities, while developing lightweight generation models and edge deployment technologies to optimize core innovations like model compression and collaborative scheduling. With continuous advancements in NPU computing power, breakthroughs in lightweight modeling techniques, and deep integration of multimodal interaction technologies, end-device intelligence will fundamentally overcome computational and latency bottlenecks in mobile VR/AR real-time rendering. This transformation will propel VR/AR technology from "fixed-scenario experiences" to mobile-enabled, real-time, and intelligent applications, enabling widespread adoption across smart cities, cultural tourism experiences, industrial inspections, and educational training sectors. Such progress will expand the technological boundaries of VR/AR and drive comprehensive upgrades in its content ecosystem.

VI. STRATEGIES FOR PROMOTING THE IMPLEMENTATION OF INTELLIGENT GENERATED VR APPLICATIONS

A. Recommendations at the Technical R&D Level

The VR intelligent generation field currently faces severe "data silo" issues, which have become one of the core technical bottlenecks hindering the deep integration of World Model with VR technology. Different R&D teams employ inconsistent 3D data formats, while toolchains used in model training, content generation, and rendering processes remain isolated. The lack of unified data exchange standards impedes efficient data circulation, sharing, and reuse. Virtual scene assets generated by World Model often require manual format conversion to adapt to various VR rendering engines, resulting in substantial labor costs, data precision loss, and physical attribute deviations. These challenges severely impact development efficiency and experience consistency, while exacerbating resource waste and collaboration difficulties.

To address this challenge, we propose establishing a unified data format for VR intelligent content generation and creating standardized data channels from generative models to rendering engines. The open-source high-performance 3D scene description technology does not merely replace existing formats, but provides a unified data representation framework for 3D concepts including meshes, materials, and physical properties. This solution supports non-destructive workflows and hierarchical collaboration, perfectly meeting the end-to-end requirements of world model integration with VR. By enhancing compatibility with physical simulation and real-time rendering capabilities tailored for VR intelligent generation, it effectively breaks down toolchain barriers and lays the foundation for large-scale world model applications in VR environments.

Currently, insufficient physical consistency remains a critical technical challenge in the integration of world models with VR technology, directly impacting the realism and immersion of VR experiences. Although world models can learn fundamental physical principles through video data and some models like Sora can simulate basic physical effects, their generated content often exhibits low coupling with professional physics engines. This results in issues such as logical inconsistencies in physical behavior, distorted

motion trajectories, and non-compliant physical principles—examples include virtual objects falling at speeds inconsistent with gravity laws, collisions lacking reasonable rebound angles, and flexible objects deforming in ways that violate mechanical principles. These problems create noticeable dissonance during VR interactions, preventing the realization of the "what you see is what you feel" immersive experience.

The fundamental issue stems from existing generative models that primarily rely on statistical learning to derive physical laws, lacking deep integration with physics engines. This results in their inability to accurately replicate complex physical interactions in the real world, coupled with the absence of explicit physical constraints during generation. Consequently, these models struggle to balance visual appeal with physical realism. VR environments demand far greater physical consistency than conventional video generation, particularly in specialized applications like industrial simulations and sports training. The accuracy of physical effects directly determines the practical value of such scenarios.

To address this challenge, we propose prioritizing the integration of generative models with physics engines. During model training, physical constraints should be proactively incorporated to achieve a harmonious balance between physical realism and visual fidelity. The core approach leverages a differentiable physics engine to embed fundamental physical principles—including collision detection, rigid-body dynamics, fluid mechanics, and flexible deformation—into the loss function of generative models. This enables the models to not only learn visual features but also autonomously adhere to physical laws throughout the training process.

B. Recommendations at the Regulatory Level

With the deep integration of world models and VR technology, AIGC (Generative AI) has become the core methodology for VR content creation, significantly enhancing production efficiency. However, the resulting ambiguity in copyright delineation has emerged as a critical constraint hindering industry development. Currently, the VR sector lacks clear industry standards and legal frameworks governing AIGC content copyright. The ownership of AI-generated VR content remains unclear—whether it belongs to model developers, content prompt providers, or the AI itself. Additionally, the copyright boundaries of training datasets are ambiguous, as world models often utilize vast public VR assets and 3D models during training, frequently leading to infringement disputes. Furthermore, copyright determination for derivative works proves challenging. VR content refined through manual editing from AI-generated drafts struggles with clear ownership and rights allocation, resulting in inadequate creator protection and dampening corporate enthusiasm for investing in AIGC+VR applications.

To address this issue, it is recommended that the industry proactively establish AIGC content copyright definition standards, clarify the legal status of generative AI in VR content creation, and build a comprehensive copyright protection system.

VII. SUMMARY AND OUTLOOK

A. Research Summary

This study systematically explores the theme of "Intelligent Generation of VR Virtual Scenes Driven by Artificial Intelligence and Digitalization," encompassing theoretical frameworks, technical analysis, case studies, and trend projections. The research yields the following key conclusions:

First, the integration of artificial intelligence and digital technologies is fundamentally transforming the core logic of VR content production, driving the industry's transition from "manual modeling" to "intelligent generation." Core technologies such as generative AI, neural rendering, and multimodal large models provide robust support for this shift. Research indicates that AI-assisted VR scene construction can enhance efficiency by over 80% while reducing costs by approximately 70%.

Secondly, the technical approach for intelligent VR scene generation encompasses multiple dimensions including geometric appearance generation, physical logic mapping, and ambient environment generation. These technologies must work in synergy to establish a comprehensive architecture spanning the data layer, model layer, and interaction layer.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Third, outstanding domestic and international case studies have validated the application value of intelligent generation technology in fields such as digital twins, cultural tourism experiences, and social entertainment, demonstrating the evolutionary path from 'single-asset generation' to 'full-scenario generation'.

Fourth, cutting-edge technologies such as edge-side intelligence, embodied intelligence, and world models will further expand the boundaries of VR scene generation, driving VR to evolve from 'scene generation' to 'world generation'.

B. Limitations of the Study and Future Challenges

While intelligent generation technology demonstrates significant potential in VR scene creation, it still faces several limitations and challenges. The conflict between computational costs and real-time rendering demands remains unresolved, as high-quality scene generation requires substantial computational resources that clash with the real-time rendering requirements of VR headsets. Currently, generating high-quality 3D assets still takes several minutes to tens of minutes, falling short of real-time generation capabilities.

The bottleneck in automated generation of complex interactive logic remains unresolved. Current AI-generated content primarily focuses on static scenarios and basic interactions. For VR applications involving intricate narratives, multi-NPC collaboration, and open-world exploration, automated interaction logic generation still struggles to match the quality of human-designed systems. Additionally, unclear copyright definitions for AI-generated content make it difficult to trace security liabilities, posing legal risks for industry-wide adoption.

C. Future Outlook: From "Generating Scenes" to "Creating Worlds"

Looking ahead, VR content production powered by artificial intelligence and digital intelligence will undergo three evolutionary stages:

The first phase is "scene generation," focusing on the rapid construction of static 3D scenes to achieve automated conversion from text/images to interactive VR environments. Current research has made significant progress in this stage.

The second phase is "world simulation," where the breakthrough lies in the integration of a world model and physics engine, enabling generated scenes to possess dynamic evolution capabilities, respond to external inputs, and maintain logical consistency. At this stage, the VR world is no longer a "static frame" but a "living ecosystem."

The third phase is "co-created experiences," achieving deep collaboration between humans and AI. AI not only understands user intent but also predicts user needs and stimulates creativity, jointly creating unique immersive experiences. The maturation of brain-computer interfaces and intention perception technologies will facilitate the realization of this phase.

Ultimately, artificial intelligence and digital intelligence will endow each user with the ability to "create worlds," transforming VR from a platform for "consuming content" into a stage for "world creation." This represents not only a technological revolution but also a profound liberation of human imagination and creativity.

REFERENCES

[1] Qin Zhibao. Research on 3D Scene Reconstruction and Rendering of Neural Radiation Fields for Virtual Laparoscopic Surgery [D]. Yunnan Normal University, 2024. DOI: 10.27459/d.cnki.gynfc. 2024.000002.

[2] Li Hua. Research on Digital Cultural Heritage Scene Communication in the Intelligent Era [D]. Jinan University, 2022. DOI: 10.27167/d.cnki.gjinu. 2022.002779.

[3] Chen Shuo. Research on Sparse Sampling 3D Light Field Content Generation Based on Deep Learning [D]. Beijing University of Posts and Telecommunications, 2024. DOI: 10.26969/d.cnki.gbydu. 2024.000162.

[4] Vehicle Dynamics. Research on Key Technologies for Virtual Scene Generation with Visual Perception Optimization [D]. Northwestern Polytechnical University, 2019. DOI: 10.27406/d.cnki.gxbgu. 2019.000058.

[5] Chen Dawei. Research on Realism Issues in Virtual Environment Generation Based on Scene Graph Systems [D]. Huazhong University of Science and Technology, 2006.

[6] Liu Mengya, Yan Dapeng. Research on Intelligent Optimization Methods for Virtual Film and Television Scenes [J]. Modern Film Technology, 2025, (10):14-20.

[7] Zhang Shunan, Lü Desheng. Programmatic generation method for virtual scenes of large-scale ancient architecture[J]. Computer Engineering and Applications, 2025,61(13):349-357.

[8] Chen Junyue. AI Empowering the Design and Construction of Metaverse Virtual Spaces [J]. Cultural Industry, 2023, (27):43-45.

[9] Yu Benzhi, Han Jianning, Zhao Xinsha, et al. Design and Development of a Virtual Experimental System for Liquid-like Photoacoustic Detection Based on the Unity3D Engine [J]. Journal of North University of China (Natural Sciences Edition), 2022,43(02):186-192.

[10] Zhu Mei, Chen Jianwei. Research on Virtual Scene Generation Algorithms Based on VR [J]. Software Guide, 2011,10(06):45-47.

[11] Li Yihua. Virtual scene generation technology based on OpenGL [J]. Ship Electronic Engineering, 2010,30(04):136-139.

[12] Zhang Tianxiang, Ji Dengpei, Yin Jinliang, et al. Precision target selection combining gesture and voice commands in virtual reality environments [J/OL]. Journal of Tianjin University of Technology, 1-8 [2026-03-30]. https://link.cnki.netUrlid/12.1374.N.20260321.1228.002.

[13] Liu Zhenlei, Dong Chuanxue, Li Yueyue, et al. Application of artificial intelligence technology-enabled virtual reality in experimental courses at higher education institutions [J/OL]. Laboratory Research and Exploration, 1-9 [2026-03-30]. https://link.cnki.netUrlid/31.1707.T.20260306.1250.040.

[14] Guo Taitai. Virtual Testing Technology Based on VR and Its Application Fundamentals Research [D]. Zhejiang University, 2005.

[15] Yang Q. Evaluation of the Quality of Virtual Reality (VR) Technology Development in China, the United States, Japan, South Korea, and Germany [D]. Huazhong University of Science and Technology, 2023. DOI: 10.27157/d.cnki.ghzku. 2023.004931.

[16] Shi Jing, Zhang Jianbo, Liang Zhiyong. Analysis of VR Security Inspection System Based on Big Data and Artificial Intelligence [J]. Smart China, 2024, (11):24-25.

[17] Jiang Zhengfeng, Chen Yan. Innovative Applications and Development of Virtual Reality Technology in the Media Field: A Case Study of the VR Stereoscopic Image Documentary "Traces of Time" [J]. China Media Technology, 2024, (11):137-140+144. DOI: 10.19483/j.cnki.11-4653/n.2024.11.030.

[18] Zhang Y. Analysis of the Application of Virtual Reality (VR) Technology in Museum Cultural Relics Exhibitions [J]. Collection, 2024, (10):132-134.

[19] Lu Jiawen. Research on Interactive VR Image Experience Based on Virtual Reality Technology [J]. Science and Technology Information, 2024,22(19):43-45. DOI: 10.16661/j.cnki.1672-3791.2405-5042-2258.

[20] Wang Jiajia, Cheng Wanqi, Han Ling. Application of digital twin VR technology in emergency drills for gas fields [J]. Information System Engineering, 2024, (08):44-47.

[21] Wu J ,Luan S ,Zhou D , Wu J ,Luan S ,Zhou D , et al.User perception based label layout for efficient target localization in virtual environment[J].International Journal of Human - Computer Studies,2026,211103801-103801.DOI:10.1016/J.IJHCS.2026.103801.

[22] Sun Yang, Zhang He. Virtual generation algorithm for complex landscape architectural scenes based on VR technology [J]. Computer Simulation, 2023,40(08):458-461+504.

[23] Zhu Mei, Chen Jianwei. Research on Virtual Scene Generation Algorithms Based on VR [J]. Software Guide, 2011,10(06):45-47.

[24] Chen Jie, Chen Chaoyang, Zhu Wenqi. VR Virtual Technology in Ready Player One [J]. Film Literature, 2019, (9):141-143.

[25] Jie Ji. Research on Interaction Design Applications Based on VR Virtual Reality Technology[J]. Modern Electronic Technology, 2019,42(15):86-90. DOI: 10.16652/j.issn.1004.373x.2019.15.021.

[26] Huang Qing. Research on the Application of Virtual Reality Technology in Cultural Communication Activities — Review of "Application of VR Virtual Technology in Cultural and Art Teaching Activities" [J]. China Science and Technology Papers, 2020,15(10): I0008-I0008.

[27] Li Zhenyan, Ma Xu, Zhang Shuyi, et al. Challenges in VR virtual scene development based on Unity3D and Zbrush [J]. Smart Cities, 2024,10(3):70-73. DOI: 10.19301/j.cnki.zncs. 2024.03.022.

[28] Xia Mingjin. The impact of VR virtual reality technology on traditional video advertising production [J]. Information and Computer, 2020,32(19):139-141.

[29] Yang, Zhengyu.Design Of A Visual Communication System For Animated Characters In Virtual Reality Using Sobel Edge Detection And Motion Capture[J].JOURNAL OF APPLIED SCIENCE AND ENGINEERING,2026,29(1):235-243.DOI:10.6180/jase.202601_29(1).0023.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[30] Xing, Huiya,Li, Xiangyi,Liu, Min, Xing, Huiya,Li, Xiangyi,Liu, Min, et al.Implementation of personalized customization and enhanced experiences for cultural tourism resources using genetic algorithm-based virtual reality technology[J].Systems and Soft Computing,2025,7.DOI:10.1016/j.sasc.2025.200360.

[31] Stabryla, Rafal,Grudzinska, Magdalena.Smartphone-Based Virtual Reality in Residential Architecture: Enhancing Spatial Understanding Through Immersive BIM plus VR Visualization[J].SUSTAINABILITY,2025,17(22).DOI:10.3390/su17229959.

[32] Wang, Yidan,Wang, Yan,Li, Xiang, Wang, Yidan,Wang, Yan,Li, Xiang, et al.Pedestrian Emotion Perception in Urban Built Environments Based on Virtual Reality Technology: A Comparative Review of Chinese- and English-Language Literature [J]. BUILDINGS, 2025,15(20):3713. DOI:10.3390/buildings15203713.

[33] Liu Datong, Guo Kai, Wang Benkuan, et al. Review and Prospects of Digital Twin Technology [J]. Journal of Instrumentation and Control, 2018,39(11):1-10. DOI: 10.19650/J.CNKI.CJSI. J1804099.

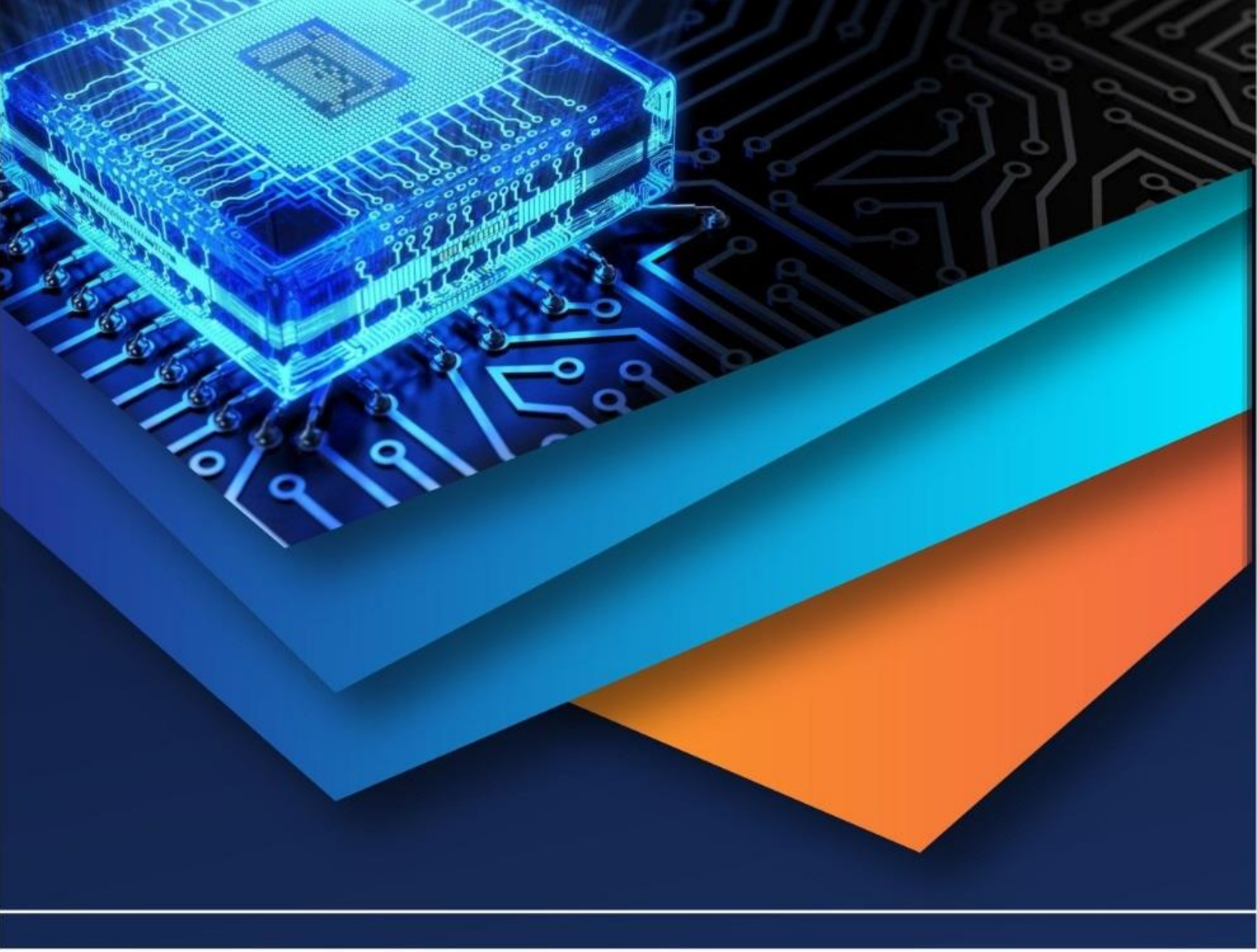
[34] Jiang Xuezhi, Li Zhonghua. Research Status of Virtual Reality Technology at Home and Abroad [J]. Journal of Liaoning Technical University (Natural Science Edition), 2004,23(2):238-240.

[35] Zou Xiangjun, Sun Jian, He Hanwu, et al. Evolution and Prospects of Virtual Reality Technology [J]. Journal of System Simulation, 2004,16(9):1905-1909.

[36] Liu Geping, Wang Xing, Gao Nan, et al. From Virtual Reality to Metaverse: A New Direction for Online Education [J]. Modern Distance Education Research, 2021,33(6):12-22. DOI: 10.3969/j.issn.1009-5195.2021.06.002.

[37] Zheng Taixiong, Huang Shuai, Li Yongfu, et al. Review on Key Technologies for Visual-Based 3D Reconstruction [J]. Journal of Automation, 2020,46(4):631-652. DOI: 10.16383/j.aas.2017.c170502.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)