An Engine-Driven Multi-modal Interaction Framework for Enhancing User Immersion in 3D Virtual Scenes via Smart Gloves and MUFD Algorithms

Xia Wang
School of Information and Intelligence Transportation, Fujian Chuanzheng Communications College, Fuzhou, 350007, China
E-mail: xiawangx@outlook.com

Keywords: engine-driven, three-dimensional virtual scenes, multi-modal interaction, user experience, augmented reality

Received: September 17, 2025

With the rapid development of virtual reality (VR), augmented reality (AR), and other technologies, the interaction experience of three-dimensional virtual scenes has become a hot spot of research. To enhance user immersion, this research proposes an engine-driven multi-modal interaction design for three-dimensional virtual scenes. An interdisciplinary experiment involving 20 volunteers compared multimodal interaction (MMI) and traditional virtual interaction. The interaction process is optimized by combining an intention-capture algorithm, intelligent gloves, a multimodal fuzzy data understanding (MUFD) algorithm, and an axis-aligned bounding box (AABB) collision detection algorithm. The design utilizes an intent capture algorithm for accurately sensing the user's experimental intent, including multiple sources of information such as vision, gesture, and eye tracking. Moreover, a smart glove is used to combine the set of intent probabilities from different channels to more accurately capture the ambiguous or incomplete intent of the user. The results showed that the minimum accuracy for visual interaction was 91.42%. The minimum accuracy for gesture interaction and eye movement interaction was 92.83% and 92.75%, respectively. Compared to traditional virtual interaction, this research method reduced response time by 41.85% and achieved system stability of 99.90%. In terms of immersion, the scores for perceived realism and emotional response were 6.25±1.70 and 5.81±1.67, respectively. Based on the Igroup Presence Questionnaire (IPQ) assessment, the multimodal interaction group showed a 37.96% increase in perceived reality score and a 38.33% increase in emotional response score compared to the traditional group. The study approach received minimal user experience scores of 4.05, 4.26, and 4.84 for naturalness, simplicity of use, and ease of starting, respectively. Furthermore, the system response time took only 120.74ms. In summary, the engine-driven multi-modal interaction design and user immersion enhancement strategy for three-dimensional virtual scenes proposed in this study can significantly enhance users' VR experience. Especially in complex tasks and demanding application scenarios, it can effectively enhance user engagement and satisfaction.

Povzetek: Študija predstavi večmodalno interakcijo za 3D VR prizore (vid + geste + sledenje očem + pametna rokavica), ki v primerjavi s klasično interakcijo skrajša odzivni čas za ~42 %, ohrani 99,9 % stabilnost ter poveča občutek prisotnosti in čustveni odziv za *~38 %.*

1 Introduction

Virtual reality (VR) and augmented reality (AR) technology have advanced quickly in recent years. This has led to the increasing application of three-dimensional virtual scenes in a variety of fields, such as gaming, education, healthcare, architecture, and so on. Especially in safety critical fields such as surgery and aerospace, VR/AR technology requires extremely high real-time, stability, and accuracy of interaction. For example, VR training for spinal surgery requires an interaction delay of 150 ms or less to ensure operational synchronization. Meanwhile, VR simulation for aircraft piloting requires system stability of at least 99.5% to prevent training interruptions. Traditional multimodal systems often struggle to meet these requirements simultaneously, becoming a core bottleneck that restricts their application in safety critical scenarios. Traditional interaction methods mainly focus on the visual and auditory levels, while modern VR is gradually integrating tactile, kinesthetic, and other perceptual modalities, forming the trend of multi-modal interaction (MMI) [1]. However, traditional single interaction modes are often difficult to meet users' needs for deep immersion experiences. The limitations of these interaction modes have become the technical bottleneck that restricts the enhancement of VR immersion [2]. Therefore, it is particularly important to research and develop MMI technology. MMI can not only provide users with richer sensory stimulation, but also by synthesizing multiple perceptual channels. This can enhance the immersion of the virtual scene, thus effectively enhancing the user's sense of participation and immersion. Unity3D, as a mature virtual engine software, is widely used in game development and architectural digitization [3]. It provides the technological basis for building high-fidelity three-dimensional scenes by virtue of its powerful real-time rendering capability, physical simulation and cross-platform support [4-5]. Currently, VR immersive user experience mostly relies on head-mounted devices and controllers or bare hands to interact with the virtual world. However, this type of interaction mainly relies on eye perception and lacks a sense of manipulation and realism [6]. In this context, smart gloves, as a convenient and flexible interaction device, provide new possibilities for MMI in virtual scenes. Smart gloves are able to make up for the shortcomings of traditional interaction methods through tactile and kinesthetic feedback. Therefore, the study innovatively proposes a MMI design that combines vision, gesture, and eye tracking, aiming to enhance the user's immersion. The study captures the user's intention and interacts in three-dimensional virtual scenes with a view to enhancing the coupling relationship between the user's psychological and sensory immersion.

2 Related works

As information technology continues to advance, VR and AR are now widely utilized in a variety of industries. Experts and academics in the domains of healthcare, education, and cultural heritage have given them a lot of attention. For example, to investigate the use of VR in spine medicine, Dargan S et al. conducted a thorough assessment of its application in surgery, counseling, education, and rehabilitation. The outcomes revealed that the application of VR in spine medicine gradually accelerated with the support of three-dimensional medical imaging, holograms, wearable sensors, 5G technology, artificial intelligence (AI), and head-mounted displays [7]. To increase the accessibility and pleasure potential for users of VR, AR, and the metaverse, Dudley J. et al. suggested the idea of inclusive immersion. The results showed that although technologies for VR and AR headsets were progressively becoming affordable and effective, these technologies had not yet achieved widespread user adoption. In particular, the needs of a wider and diverse user community needed to be considered [8]. To increase the realism, immersion, and overall experience of surgical simulation, Lungu A. J. et al. suggested utilizing VR, AR, and mixed reality (MR) technology. The results showed that the key components of a VR surgical simulator were visual and haptic feedback [9]. Duarte M L et al. investigated if these cutting-edge technologies might supplement or replace conventional anatomy teaching techniques in order to evaluate the efficacy of VR and AR in anatomy education. The results demonstrated the significant advantages of VR and AR in improving student engagement, learning efficiency, and knowledge retention [10].

In the realm of VR and AR, MMI design is essential. Al-Ansi A M et al. proposed a system design based on MMI techniques in order to enhance visitor experience through visual and audio interaction interfaces. The outcomes showed that the interactive system was effective in evoking visitors' natural interaction with the cultural heritage environment and facilitating a deeper understanding of the cultural content [11]. To increase the effectiveness of human-machine and human-robot interactions, Sereno M. et al. suggested an MMI technique based on the MQTT protocol. The findings demonstrated that the design accommodated a variety of interaction techniques, including touch, speech, and gaze tracking. Moreover, it could communicate effectively between multiple devices such as computers, smartphones, tablets, etc [12]. Weitz K et al. suggested a user MMI method based on a basic voice recognition system to investigate the possibilities of virtual agents in explainable artificial intelligence (XAI) interface design. According to the findings, including virtual agents might boost users' confidence in the XAI system [13]. Behavioral trajectories, learning outcomes, task performance, teacher assistance, student engagement, and feedback are some of the six primary objectives that Sharma K et al. examined in an attempt to investigate the use of multi-modal technology in education. The findings demonstrated the great potential of multi-modal technology to record and enhance the learning process [14].

Wang H et al. conducted a large-sample experiment to verify the advantages of a multimodal VR anatomy teaching system in terms of the long-term knowledge retention rate. The aim was to study how VR teaching improves efficiency. The results showed that students who used visual gesture eye movement multimodal interaction had a 28% increase in knowledge retention rate after 3 months compared to traditional VR teaching [15]. Zhang Y et al. proposed a multimodal interaction fusion model based on attention mechanism to improve interaction accuracy. The model optimizes by dynamically adjusting the weights of visual, gesture, and speech modalities. The results showed that this method improved the accuracy of intent recognition in complex scenes [16]. In order to improve interaction sensitivity, Li J et al. designed a real-time interaction optimization scheme based on 5G edge computing for the problem of tactile feedback delay of smart gloves. The results showed that this method reduced the tactile feedback delay to less than 50ms [17]. The systematic comparison of literature review is shown in Table 1.

Table 1: Systematic comparison of literature review

Research	Immersion rating	Domains applied	Limitations
Dargan S et al	/	Healthcare (Spine medicine)	Lacks quantitative immersion assessment; no

[7].			multi-modal fuzzy intent processing
Dudley J et al [8].	/	VR/AR/Metaverse (inclusive design)	No clear modality definition; fails to address safety-critical scenario adaptation
Lungu A J et al [9].	High fidelity	Healthcare (surgical simulation)	No numerical immersion score; does not integrate intent fusion algorithms
Duarte M L et al [10].	Improved engagement	Education (anatomy teaching)	Single-modality dependent; no quantitative immersion evaluation
Al-Ansi A M et al [11].	Enhanced natural interaction	Cultural heritage	Lacks multi-modal data fusion; ignores environmental interference resistance
Sereno M et al [12].	/	HMI/HRI (Human-machine/robot interaction)	No immersion assessment; does not optimize system real-time performance
Weitz K et al [13].	Improved trust	XAI interaction design	No multi-modal intent integration; limited to trust enhancement only
Sharma K et al [14].	/	Education (learning process)	Vague modality description; no practical scenario validation
Wang H et al [15].	/	Education (VR anatomy teaching)	No immersion quantification; fails to verify adaptability in complex scenarios
Zhang Y et al [16].	/	General MMI	Ignores fuzzy intent handling; no immersion performance assessment
Li J et al [17].	/	MMI hardware optimization	Only optimizes haptic delay; no multi-modal data fusion integration

As evidenced by the above comparison and existing research, there are still many shortcomings in existing studies. For example, the research of Zhang et al. and Wang et al. only focuses on multimodal combinations or single-modal optimizations. They failed to address users' vague or incomplete intentions. These studies do not use fuzzy reasoning to fuse uncertain data from different modalities, resulting in poor system adaptability when user intent expression is unclear. Almost all studies only provide qualitative descriptions of immersion, lacking numerical scores or standardized scales for evaluation. This makes it impossible to conduct cross-study comparisons or performance benchmarking. Moreover, most of the research is limited to a single noncritical area without verifying its effectiveness in safety-critical areas requiring high real-time performance and low error tolerance. For example, although Li J et al. optimizes tactile feedback delay, they did not test the design in surgical VR scenarios where delay control is crucial. In addition, there are few studies that combine multimodal design with underlying interaction optimization techniques. Unlike this study, which uses multimodal interaction (MMI) and axis-aligned bounding boxes (AABBs) for collision detection to reduce computational load, existing research, such as that of Sereno et al. and Sharma et al., rarely considers such optimization. This results in an insufficiently smooth and stable system in complex virtual scenes. Therefore, the proposed engine-driven MMI design for three-dimensional virtual scenes, which combines multiple interaction modes and immersion enhancement strategies, can make up for the shortcomings of existing research. This further promotes the application of VR in a wider range of fields, meets the needs of diverse users, and enhances their interaction experience and immersion.

3 Engine-driven design for MMI

The study is based on the Unity3D virtual engine and combines glove sensors to capture intent. It uses a hierarchical enclosing box algorithm to determine if objects are colliding, thereby optimizing interaction experiences. Meanwhile, the smart glove is adopted as the core interaction interface, and the MMI method of vision, gesture, and eye tracking is designed. Moreover, the fuzzy reasoning and multi-modal understanding of fuzzy data (MUFD) algorithms are used to integrate the user's final intention and enhance the user's immersion.

3.1 Intent capture and interaction optimization in engine-driven three-dimensional virtual scenes

Unity3D, a modern game engine, is not only equipped with real-time graphic rendering and physical simulation, but also supports access to deep learning modules, semantic analysis plug-ins, and external interaction devices. This provides a technical basis for intent modeling and optimization [18-19]. To further enhance the robustness and adaptability of multimodal interaction, the research introduces adaptive control and robust control principles to construct cross domain technology links. The study first adopts the model reference adaptive strategy in adaptive control to monitor the input signal-to-noise ratio and data integrity of visual, gesture, and eye movement modalities in real time. When the SNR of eye movement tracking drops below 30 dB due to a change in ambient light intensity, the system automatically adjusts the modal weight from 0.3 to 0.1. Meanwhile, it increases the weights of the gesture and visual modes to prevent interruption of interactions caused by a failure of a single mode. In addition, the research also integrates the robust control H ∞ control theory. By constructing a noise suppression module, it successfully controls the interference from sensor noise and data transmission delays on intent recognition to within 5%. This ensures the system's stable operation in complex electromagnetic environments such as aerospace VR simulators, while also being suitable for scenarios involving high-frequency equipment interference in surgical VR procedures. This design fills the technical gap of “heavy fusion, light anti-interference” in existing multimodal systems [20-21]. Through the behavioral analysis scripts and state machine models integrated in the engine, it is possible to achieve in-depth understanding and prediction of the user's operation paths, gaze points, and behavioral sequences, thus promoting the construction of a closed-loop interaction “from recognition to response” [22]. Therefore, in order to achieve interaction optimization, this study adopts an engine-driven three-dimensional virtual scenes intent capture and interaction optimization strategy. Engine driven intent capture and interaction optimization overall architecture is shown in Figure 1.

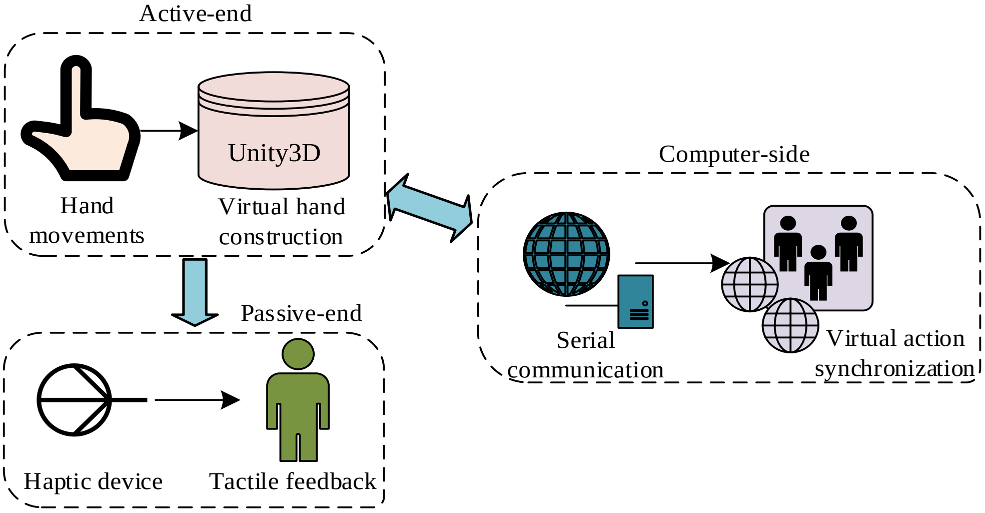


Figure 1: Engine driven intent capture and interaction optimization overall architecture

In Figure 1, in this overall architectural design, the user wears a Leap Motion sensing device to capture hand movements. The Unity3D virtual engine builds the virtual scene on the computer and controls the operation of the virtual hand. When the virtual hand makes contact with a virtual object, the system detects the contact. Moreover, the contact information is transmitted to the force haptic device worn on the slave user through the serial port. The device controls the spring and pressure plate through the servo to simulate the feeling of a finger touching an object, thus enabling the user at the slave end to perceive the presence of the virtual object. Additionally, sensors on the active user's arm collect real-time motion data and transmit it to the exoskeleton device on the follower's arm. This guides the arm movements and synchronizes them with the active user's movements. Through this architectural design, user intent capture and interaction optimization can be effectively achieved. The Leap Motion intent capture process is shown in Figure 2.

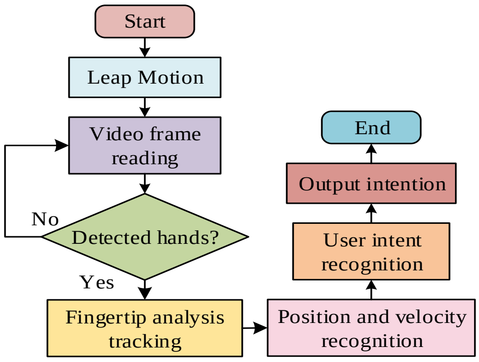


Figure 2: Leap motion intent capture process

In Figure 2, the process first utilizes Leap Motion to track the movements of the user's hand and fingers, and transmits the data to the computer side for video frame reading. The system can determine whether the video frame contains hand information. If the hand is detected, fingertip analysis and tracking is performed.

If the hand is not detected, the system will re-read the video frame. The technology analyzes and tracks the user's fingertips before determining the locations and speeds of the fingers to detect the user's motions. The system understands the user's objectives based on this information. In virtual interaction, it is crucial to determine whether virtual objects collide with each other. If efficient collision detection is not performed, it can lead to excessive computation, which will affect the smoothness of the screen [23]. Therefore, to improve the efficiency of collision detection and achieve interaction optimization, the study adopts a hierarchical enveloping box algorithm to determine whether objects collide or not. The algorithm works by wrapping a virtual object with a layer of slightly larger geometric shape boxes. Then, based on the intersection of these boxes, it can determine whether the object has collided or not [24]. The hierarchical enclosing box algorithm is displayed in Figure 3.

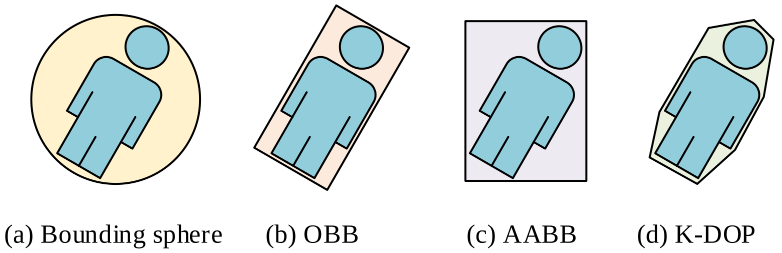

(a) Bounding sphere (b) OBB (c) AABB (d) K-DOP

Figure 3: Hierarchical bounding box algorithm

In Figure 3, the hierarchical bounding box algorithm has several common forms of bounding balls, oriented bounding box (OBB), AABB, and K-discrete orientation polytopes (K-DOP) bounding boxes [25]. Among them, the rotation of the enclosing sphere does not vary with the velocity of the object, and thus is not applicable to collision detection of deformed objects [26]. However, OBB and K-DOP have slower and more complex rotation following, and thus are more computationally intensive [27]. In contrast, AABB has lower computational complexity and is more suitable for interaction design in VR. The wrapping box of AABB is a polyhedron that can be constructed by describing multiple scalars. The computational process is simpler and the wrapped objects have smaller gaps [28]. Therefore, the study chooses AABB for collision determination, thus achieving efficient collision detection and interaction optimization design while keeping the computational effort low. Equation (1) displays AABB's mathematical expression.

$$R_{AABB} = \left\{ (x, y, z) \middle| a_{\min} \le x \le a_{\max}, b_{\min} \le y \le b_{\max}, c_{\min} \le z \le c_{\max} \right\} \tag{1}$$

In Equation (1), R_{AABB} denotes the enclosing box of AABB. (x, y, z) denotes the coordinate position of a point in three-dimensional space. $a_{\min}$ and $a_{\max}$ denote the minimum and maximum values in the x direction. Similarly, $b_{\min}$, $b_{\max}$, $c_{\min}$, and $c_{\max}$ denote the minimum and maximum values in the y and z directions, respectively.

2.2 Smart glove-based interaction with vision, gesture, and eye tracking

After achieving accurate recognition of user intent and interaction optimization at the engine layer, how to naturally integrate these multiple input modalities into the user's operating behavior becomes the key to further enhance the sense of immersion. Currently, visual recognition, gesture interaction, and eye tracking are considered as three key perception channels, which are building a new multi-modal human-computer interaction paradigm [29]. This system incorporates smart glove-based fusion interaction technology as a core component of the MMI system. This technology is highly compatible with natural movements and can simultaneously capture haptic, electromyographic, and motion information [30]. The smart gloves can accurately capture the degree of bending, pressure changes, and inertial movement trajectories of the finger joints through the embedded sensor network. This can restore high-fidelity hand animations in three-dimensional engine scenes for human-like motion actuation [31]. Meanwhile, combined with the depth camera and AI image recognition algorithm, the system can recognize the visual features of the user's environment in real time, such as the scene structure, object boundaries, and surface texture, to help quickly lock the interaction target. The YOLOv7 algorithm is chosen over YOLOv5 and other lightweight models for the visual channel because it strikes a balance between real-time performance, detection accuracy, and computational resource consumption requirements in VR scenes. VR interaction requires precise positioning of virtual objects. The YOLOv7 algorithm is more accurate than the YOLOv5 algorithm, especially when detecting small virtual targets. The higher accuracy advantage can avoid misjudgment of interaction intention caused by visual positioning deviation. Although YOLOv7 has slightly higher parameter count than YOLOv5, it optimizes feature fusion efficiency through the ELAN structure. In this experiment, YOLOv7 achieved an inference speed of 85 FPS for 1280×720 virtual scene images in the NVIDIA GeForce RTX 3090 hardware environment. This result is only 7.6% lower than YOLOv5 and far higher than the 30 FPS minimum threshold required for VR interaction. If a lighter YOLO Nano structure is chosen, although the inference speed can reach 120 FPS, the small object detection recall rate cannot meet the high-precision interaction requirements. In addition, YOLOv7 supports dynamic batch inference, which automatically adjusts the size of inference batches based on the complexity of virtual scenes. When the number of objects in the scene is less than 10, the inference speed increases to 98 FPS. Even when the number of objects exceeds 30, it can maintain a frame rate of 65 FPS or higher. This avoids the sudden drop-in frame rate that traditional fixed batch models experience in complex VR scenes. It also ensures synchronized visual, gesture, and eye movement multimodal interaction. This prevents immersion interruption caused by visual delay. Eye tracking technology, as an important dimension of sight orientation, can help the system recognize the user's current focus of attention, thus enhancing the precision of operation and the naturalness of interaction. Therefore, this study uses smart gloves as the core interaction interface to realize visual, gesture, and eye tracking interaction design. The framework of the smart glove-based visual, gesture, and eye tracking interaction system is shown in Figure 4.

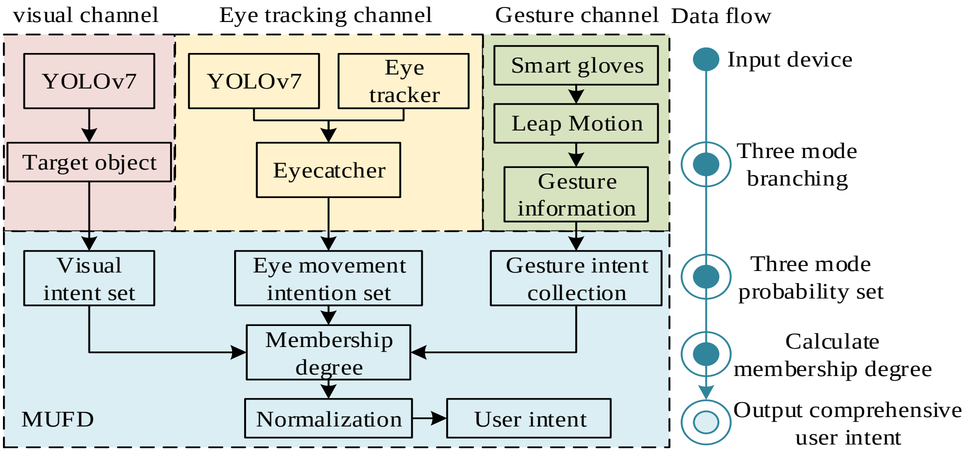


Figure 4: Framework for visual, gesture, and eye tracking interaction system based on smart gloves

In Figure 4, the system framework integrates three input modalities: vision, gesture, and eye tracking. The visual channel performs visual target object capture through the you only look once version 7 (YOLOv7) algorithm. Gesture is realized through smart gloves combined with Leap Motion for intent capture and tracking. Eye tracking is combined with eye tracking for focus capture. The target data collected through these three modalities are synthesized using fuzzy inference based and MUFD algorithms and the fuzzy set affiliation is calculated. After normalizing the affiliation degree, combined with the basic trust allocation function, the integrated user intent is finally obtained. The intention acquisition process of vision and eye tracking is shown in Figure 5.

In Figure 5, in this process, the current environment information is first captured by the camera and the YOLOv7 algorithm is utilized for object localization. In the visual channel, after initializing the YOLOv7 model, the system identifies the three-dimensional set of current objects. It also calculates the distance change of the same object in this frame and the next frame to derive the probability of intent under the visual channel and normalize it. Under eye tracking, after initializing the YOLOv7 model, the coordinates of the user's line-of-sight focus are obtained in conjunction with the eye-tracker. The image cropping region is determined based on the line-of-sight focus, and the image of the region is input into the YOLOv7 model. The confidence level of the object is obtained and this is used as the initial intent probability. Next, the system will normalize the confidence level and output the final set of intent probabilities. Equation (2) displays the formula for calculating distance change.

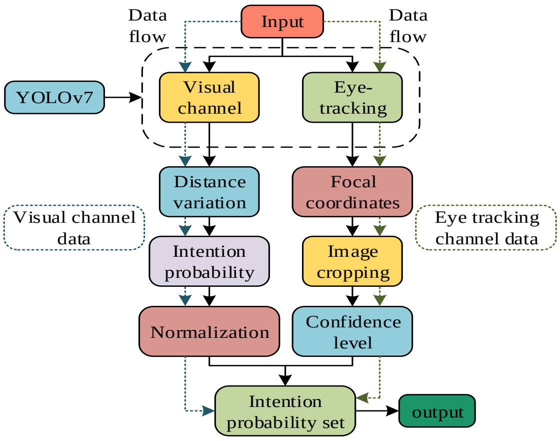


Figure 5 The intention acquisition process of vision and eye tracking

$$D_i=\sqrt{\left(x\left(t+1\right)_i-x\left(t\right)_i\right)^2+\left(y\left(t+1\right)_i-y\left(t\right)_i\right)^2+\left(z\left(t+1\right)_i-z\left(t\right)_i\right)^2}\ (2)$$

In Equation (2), D_i denotes the change in distance between time steps t and $t+1$ for object i. $x(t)_i$, $y(t)_i$, and $z(t)_i$ denote the three-dimensional coordinates of object i at the moment of t, respectively. The formula for calculating the probability of visual intent is shown in Equation (3).

$$V_i=\frac{D_i}{\sqrt{x(t)_i^2+y(t)_i^2+z(t)_i^2}}\ (3)$$

In Equation (3), V_i denotes the visual intention probability, i.e., the motion intention of object i between time steps t and $t+1$. The focus position formula for eye tracking is shown in Equation (4).

$$P_{focus}=\left(x_{focus},y_{focus}\right)\ (4)$$

In Equation (4), P_{focus} denotes the focus position of the eye tracking system. x_{focus} and y_{focus} denote the position of the user's visual focus on the two-dimensional plane under eye tracking, respectively [32]. The minimum and maximum visual field coordinates are shown in Equation (5).

$$\begin{cases}x_{\min}=x_{focus}-\dfrac{W}{2},y_{\min}=y_{focus}-\dfrac{H}{2}\\ x_{\max}=x_{focus}+\dfrac{W}{2},y_{\max}=y_{focus}+\dfrac{H}{2}\end{cases}\ (5)$$

In Equation (5), W and H denote the width and height of the field of view, respectively. The confidence expression for the object category is shown in Equation (6).

$$O_i=\left(C_i,S_i\right)(6)$$

In Equation (6), O_i denotes the confidence level of the object category i. C_i denotes the confidence value of the object. S_i denotes the categorization score of the object. The mathematical expression for the set of intentional probabilities is shown in Equation (7).

$$P_i=\frac{S_i}{\sum_{i=1}^{n}S_i}\ (7)$$

In Equation (7), P_i denotes the probability of intent for each object category i. n denotes the number of object categories. The flow of intention understanding based on MUFD algorithm is shown in Figure 6.

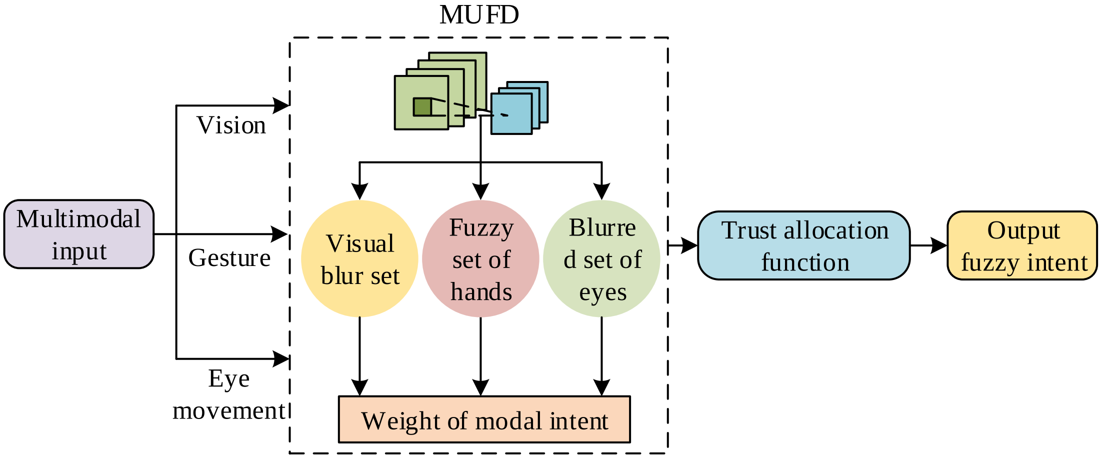


Figure 6: Intention understanding process based on MUFD algorithm

In Figure 6, in this process, multi-modal intent data from vision, gesture, eye movement, etc. are first input. Then, the fuzzy set corresponding to each modality is calculated, and the affiliation degree of each modality is also obtained. Next, the fuzzy intent trust distribution of each modality is calculated to obtain the weight of each modal intent. Finally, the final fuzzy intention of the user is obtained by constructing the trust distribution function and synthesizing the information of each modality. The mathematical expression for the set of intention probabilities is shown in Equation (8).

$$I\leftarrow MAG\left(V_{input}\right)(8)$$

In Equation (8), I denotes the final set of user's intentions. V_{input} denotes multi-modal input. MAG denotes multi-modal aggregation [33]. The formula for the affiliation degree is shown in Equation (9).

$$u_i\left(x_i\right)=e^{-\frac{\left(x_i-\mu_i\right)^2}{2\sigma_i^2}}\ (9)$$

In Equation (9), $u_i\left(x_i\right)$ is the affiliation of the input data x_i. μ_i is the center of the fuzzy set of i object. σ_i^2 denotes the variance of this fuzzy set. The normalization process is shown in Equation (10).

$$u_i'\left(x_i\right)=\frac{u_i\left(x_i\right)}{\sum_{1<i<n}u_i\left(x_i\right)}\ (10)$$

The normalization operation by Equation (10) ensures that the sum of all fuzzy affiliations is 1. The reliability distribution is shown in Equation (11).

$$CF = u_i'(x_i) \quad (11)$$

In Equation (11), CF represents the credibility distribution, which indicates the credibility of each modal intention. The weighting formula is shown in Equation (12).

$$\omega(x_i) = \omega_{initial} + \Delta\omega \quad (12)$$

In Equation (12), $\omega(x_i)$ denotes the final weight of the input data x_i. $\omega_{initial}$ denotes the initial weights. $\Delta\omega$ denotes the increment adjusted according to the modal trust distribution. The basic trust distribution function is shown in Equation (13).

$$m(i) = \omega(x_i) \times u_i'(x_i) \quad (13)$$

Equation (13), $m(i)$ denotes the trust allocation function for modality i, i.e., the initial trust value for the modality. The combined trust allocation function is shown in Equation (14).

$$m(A) = \frac{1}{1-k} \sum_{\cap A_i = A} \prod_{1 \le i \le 3} m(i) \, A \ne \Phi \quad (14)$$

In Equation (14), $m(A)$ denotes the integrated trust allocation function, i.e., the trust allocation after the synthesis of all modalities. A denotes the set of all modalities. k denotes the constant used to adjust the overall trust allocation. $\prod_{1 \le i \le 3} m(i)$ denotes the product operation of the trust allocation function $m(i)$ for all modalities. The final fuzzy intent is shown in Equation (15).

$$V = \max m(A) \quad (15)$$

In Equation (15), V represents the final fuzzy intent of the user. That is, the maximum value is chosen to determine the most plausible modality to arrive at the final decision or intention of the system. To further clarify the mathematical logic of the MUFD algorithm and improve reproducibility, the study first uses the construction of fuzzy sets based on Gaussian fuzzy functions. Then, it makes the calculation of the trust allocation function satisfy the normalization condition $\sum m(i) = 1$. The constraint condition $\sum \left[\omega(x_i) \cdot \mu_i(x_i) \right] = 1$ is set for the basic trust allocation function to ensure that the weight allocation of trust values in each modality is reasonable. Finally, in calculating the comprehensive trust allocation function $m(A)$ the value of constant k is based on $k = \frac{1}{\left[1 - \sum_{i<j} m(i) \cdot m(j)\right]}$, which is used to correct conflict terms in the multimodal information fusion process. This helps to avoid bias in judgments caused by contradictions in the data between modalities. The pseudocode of MUFD is shown in Table 2.

Table 2: MUFD pseudocode

```
Input:
    P_v: Visual intent probability vector
    P_g: Gesture intent probability vector
    P_e: Eye-tracking intent probability vector
    μ_i: Mean of fuzzy set for each object
    σ_i: Std of fuzzy set for each object
Output:
    Final_intent: Index of object with highest fused intent probability
// Step 1: Calculate fuzzy membership degree (Gaussian function)
For each object i:
    μ_v(i) = exp(-(P_v(i) - μ_i)² / (2*σ_i²))    // Vision membership
    μ_g(i) = exp(-(P_g(i) - μ_i)² / (2*σ_i²))    // Gesture membership
    μ_e(i) = exp(-(P_e(i) - μ_i)² / (2*σ_i²))    // Eye-tracking membership
// Step 2: Normalize membership degrees (sum = 1)
μ_v_norm = μ_v / sum(μ_v)
μ_g_norm = μ_g / sum(μ_g)
μ_e_norm = μ_e / sum(μ_e)
// Step 3: Dynamic weight assignment (trust-based)
total_max = μ_v_norm.max() + μ_g_norm.max() + μ_e_norm.max()
ω_v = 0.3 * (μ_v_norm.max() / total_max)    // Vision weight (base=0.3)
ω_g = 0.4 * (μ_g_norm.max() / total_max)    // Gesture weight (base=0.4, primary)
ω_e = 0.3 * (μ_e_norm.max() / total_max)    // Eye-tracking weight (base=0.3)
// Step 4: Fuse intent probabilities
```

```
Fused_prob(i) = ω_v*μ_v_norm(i) + ω_g*μ_g_norm(i) + ω_e*μ_e_norm(i)
// Step 5: Determine final intent
Final_intent = argmax(Fused_prob)
Return Final_intent
```

3 Engine-driven validation of MMI designs

After building the experimental environment, the study validates the performance of MMI in three-dimensional virtual scenes. Moreover, the effect of user's immersion experience is evaluated.

3.1 Experimental environment setup

The trials employ a high-performance computing platform and software setup to verify the engine-driven MMI based design's performance. This ensures that the system is able to process multi-modal data and optimize the interaction experience in real-time. The experiments use the Ubuntu 20.04 operating system with an Intel Core i9-11900K CPU, NVIDIA GeForce RTX 3090 GPU, and 64GB of RAM. In addition, the virtual engine used for the experiments is Unity3D, which supports real-time graphics rendering and physics simulation features. The deep learning framework is TensorFlow 2.6 and the programming language is Python 3.8. To ensure efficient data processing and interaction response, the experiments set the learning rate to 0.001 and the batch size to 32. The smart glove, the Manus Prime II data glove, is the core interactive device used in the experiment. Its core parameters include 16 high-precision inertial measurement units (IMUs) and finger joint bending sensors. The device adopts USB-C wired connection with a sampling rate of up to 1000 Hz. The eye tracking device uses Tobii Pro Spectrum eye tracker. The device uses the 9-point calibration method with a sampling rate of 300 Hz to track eye movements, including rapid ones such as scanning and gaze switching. The tracking range is ±35° for the horizontal viewing angle and ±20° for the vertical viewing angle. It supports a sitting distance of 50-80 cm. In the experiment, the eye tracker is fixed at a distance of 60 cm from the headset and adapted to VR headsets. The dataset used includes indoor virtual scenes and gesture data. The datasets used include indoor virtual scenes and gesture data, which are constructed by Unity3D and the gesture data are captured by smart gloves and Leap Motion. The experimental data are pre-processed and divided into training and testing sets in the ratio of 7:3. The size of the indoor virtual scene dataset is 10 indoor scenes, each containing 8 types of core interactive objects. A total of 40,000 image frames are collected for each object, captured from 500 different angles and under varying lighting conditions. The image resolution is 1280×720 and the format is RGB-D. It is mainly used for training the YOLOv7 visual object detection model. The scale of the multimodal gesture interaction dataset is collected from 20 volunteers using Manus Prime II smart gloves and Leap Motion Ultra. This includes four core interactive gestures, including grabbing, releasing, rotating, and clicking. Each volunteer performs 50 repetitions of each gesture type, resulting in 4,000 valid samples in total. Each set of samples contains joint angle data at a sampling rate of 1,000 Hz, along with corresponding intent labels. Considering that this experiment involves human subjects, all experimental procedures have been approved by the ethics review committee and fully comply with the ethical guidelines for human subject research in the Helsinki Declaration of the World Medical Association. Table 3 displays the experimental environment's precise setup.

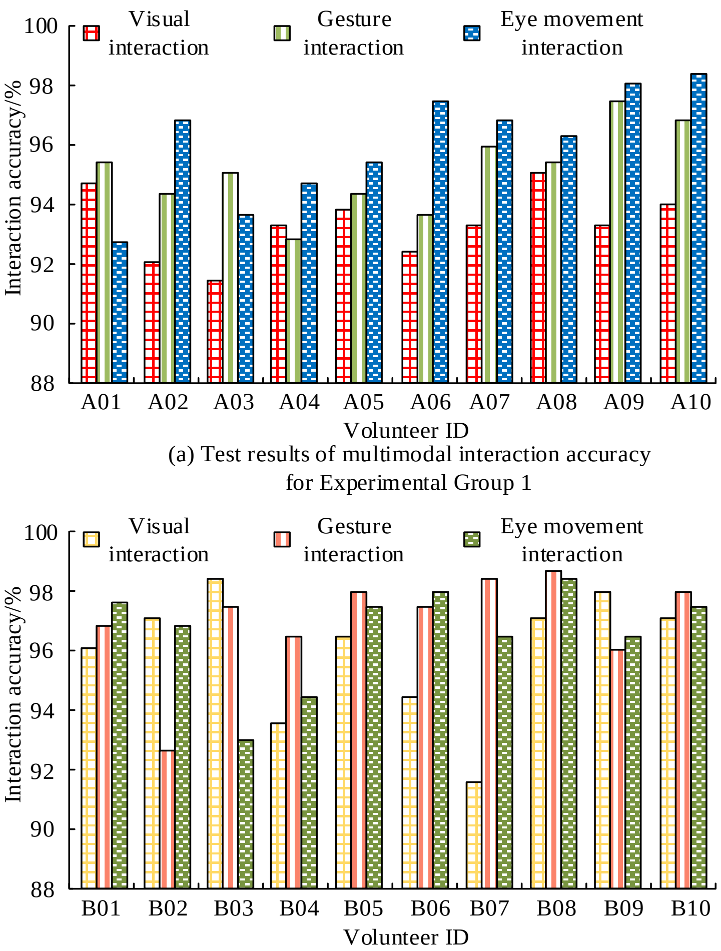
Table 3: Experimental environment configuration

Environment	Configuration
Operating system	Ubuntu 20.04 LTS
CPU	Intel Core i9-11900K
GPU	NVIDIA GeForce RTX 3090
Random access memory	64GB
Virtual engine platform	Unity3D
Deep learning framework	TensorFlow 2.6
CUDA	11.2
Programming language	Python 3.8

3.2 Performance validation of MMI in three-dimensional virtual scenes

To validate the effect of engine-driven MMI based design on user immersion, the experiment invites 20 volunteers ranging from 22 to 27 years old with an average age of 24.5 years old. All participants have not been exposed to VR. Moreover, during the experiment, the attention span of the volunteers is controlled within 90 minutes. In the experiment, volunteers are randomly divided into two groups. 32 VR scenes are used in this experiment, and these scenes are divided into two scene banks based on a two-dimensional emotion model of potency-arousal. Each scene bank contains 16 different scenes. The order of scene playback is randomly assigned to minimize the interference of order effects on the experimental results. Before the start of the experiment, all participants need fill out an informed consent form to confirm that they are in good mental health and able to complete the experimental tasks independently. During the experiment, each volunteer is required to remain in a seated position, watch the VR scene and perform emotional self-assessment. To avoid fatigue caused by prolonged viewing, participants take a 20-minute break after completing part of the task. The initial sample size for this experiment is set to 20. Although the sample size is small, the selection criteria are as follows: 1 Based on the preliminary experimental results, the coefficient of variation (CV) of the core indicator of the multimodal interaction system in this study is ≤15%. According to the formula calculation, this value meets the statistical power required to detect differences between two groups. 2. Considering the characteristic of dizziness caused by long-term wearing in VR experiments, a large sample size may lead to fatigue bias in the subjects. Therefore, priority should be given to ensuring sample homogeneity to control errors. 3. Subsequently, the sample size will be expanded to 100 for confirmatory experiments, and the results of this study can serve as a preliminary empirical basis. Volunteer recruitment is an open recruitment process conducted through university research volunteer recruitment platforms and local technology communities. It uses a voluntary registration and screening model without offering any material rewards. It only provides explanations of the experimental process and feedback on the results. The inclusion criteria are individuals aged between 18-30 years old, without visual or auditory impairments, without a history of motion sickness, and who have not participated in similar VR experiments. Exclusion criteria include individuals with a history of mental illness or underlying conditions such as hypertension or heart disease. Those who experienced physical discomfort on the day of the experiment are also excluded. The demographic distribution is a male to female ratio of 6:4, with an age range of 24.5±1.4. The experiment uses a completely randomized design. 20 volunteers are assigned to experimental groups (EGs) 1 and 2 using an Excel random number table, with 10 people in each group. After randomization, baseline balance test is conducted: there are no significant differences (p>0.05) in gender (χ^2=0.21, p=0.646), age (t=0.53, p=0.602), and basic cognitive ability (t=0.38, p=0.707) between the two groups. This ensures that the initial conditions of the two groups are consistent and eliminating the interference of baseline differences on the experimental results. 32 VR scenes are used in the experiment. It is divided into two scene libraries, each with 16 scenes, based on valence and arousal. It is not randomly constructed, but rather based on a validated database. Pre-experiments are conducted to verify its emotional valence and arousal effectiveness.

To evaluate the response accuracy of the system under the three interaction modes of gesture recognition, speech recognition, and eye tracking, the experiments are conducted to test the visual, gesture, and eye tracking interactions respectively. The specific tests are as follows: visual interaction triggers the interaction by gazing at a specific target or interface element. Gesture interactions include gestures such as opening the palm of the hand, making a fist, and sliding. Eye-movement interactions are triggered by gazing at a target point for 2 s. The MMI accuracy test results of the volunteers in the two EGs are shown in Figure 7.



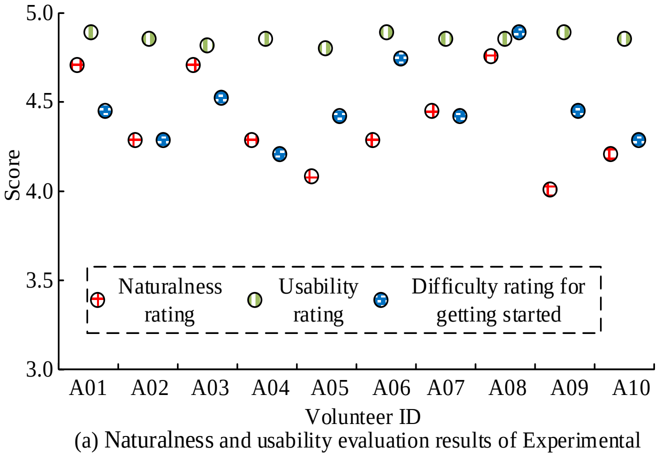
(a) Test results of multimodal interaction accuracy for Experimental Group 1

(b) Test results of multimodal interaction accuracy for Experimental Group 2

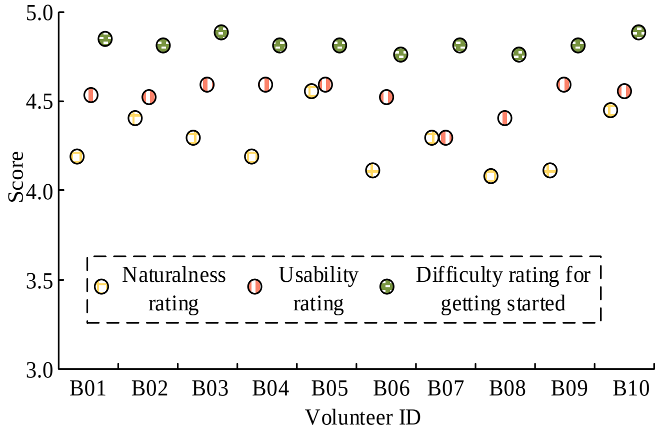
Figure 7: MMI accuracy test results

In Figure 7(a), the visual interaction accuracy rates of EG 1 are all above 91%. Among them, the volunteer numbered A03 has the lowest visual interaction accuracy rate, which is also 91.42%. The lowest accuracy rates of gesture interaction and eye movement interaction are 92.83% and 92.75%, respectively. In Figure 7(b), the volunteers in EG 2 has the lowest accuracy rates of 91.58%, 92.97%, and 92.63% for visual, gestural, and eye-movement interactions, respectively. After independent sample t-test verification, there is no significant difference in visual interaction accuracy (t=0.32, *p*=0.752), gesture interaction accuracy (t=0.41, *p*=0.685), and eye movement interaction accuracy (t=0.28, *p*=0.782) between the two groups of volunteers. This indicates that the recognition accuracy stability of the multimodal interaction design in this study is not affected by grouping, and the overall accuracy level (both ≥91%) has reliable statistical support. To summarize, the interaction accuracy of all participants is above 91%. It shows that the engine-driven MMI based design can achieve high recognition and interaction accuracy in three-dimensional virtual scenes.

To verify whether MMI is consistent with the user's natural behavioral patterns, participants will be required to perform the task 10 times for each interaction mode. After the tasks are completed, participants will fill out a questionnaire to assess the naturalness, ease of use, and intuitive operation of each interaction mode. Scores range from 1 to 5. A score of 1 indicates that it is not at all natural or easy to use, while a score of 5 indicates that it is very natural or very easy to use. The results of the naturalness and ease of use evaluation of MMI are shown in Figure 8.



(a) Naturalness and usability evaluation results of Experimental Group 1



(b) Naturalness and usability evaluation results of Experimental Group 2

Figure 8: Assessment results of naturalness and usability of MMI

In Figure 8(a), EG 1 has the lowest naturalness score, ease of use score, and difficulty of getting started score of 4.05, 4.26, and 4.84, respectively. In Figure 8(b), EG 2 has the lowest naturalness score, ease of use score, and difficulty of getting started score of 4.12, 4.31, and 4.80, respectively. According to paired sample t-test analysis, there is no significant difference between the two groups in terms of naturalness (t=0.57, *p*=0.573), ease of use (t=0.48, *p*=0.635), and difficulty in getting started (t=0.39, *p*=0.699) scores, and all scores are ≥ 4.0. Based on the t-test results (*p*>0.05), it can be inferred that the naturalness and ease of use of the multimodal interaction mode in this study are consistent across samples and are not the result of individual differences. In summary, the evaluation results of both EGs show that the MMI approach has better performance in terms of naturalness and ease of use. The overall interaction experience is more intuitive and conforms to the user's natural behavioral pattern.

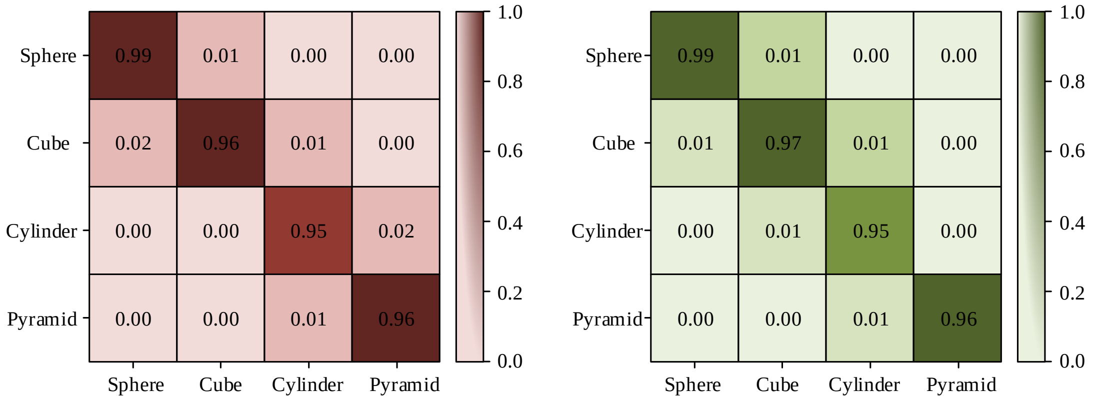
To verify the comprehensive performance of the engine-driven MMI based design, the study analyzes it in comparison with the traditional virtual interaction methods. The study uses a Keysight U2722A power analyzer to measure energy and resource consumption. The analyzer collects real-time instantaneous power during system operation, records data every five minutes, and monitors continuously for two hours. Then, it calculates the average power consumption. The Ubuntu system's built-in htop tool measures resource consumption by calculating CPU usage, GPU memory usage, and memory usage. The average value during the experimental period is used as the evaluation indicator. The difference between energy and resource consumption is verified through an independent sample t-test (p<0.05), ensuring that the difference between the two sets of data is statistically significant. Table 4 compares the performance of several approaches.

Table 4: Performance comparison of different methods

Index	MMI	Traditional virtual interaction
Response time/ms	120.74	207.67
System smoothness	4.8/5	3.5/5
System stability/%	99.90	95.23
Device compatibility	High	Medium
Energy consumption and resource use	Low	High

In Table 4, the response time of the studied MMI method is only 120.74ms, which is 41.85% lower than the traditional virtual interaction method. In terms of system fluency, the fluency score of the research method is 4.8/5, which is very smooth. In contrast, the fluency score of the traditional virtual interaction method is 3.5/5, with occasional lagging. The system stability of the research method is 99.90%. In comparison, the system stability of the traditional virtual interaction method is only 95.23%. In terms of device compatibility, the research method has high compatibility and supports a wide range of devices. In contrast, the traditional virtual interaction method has low device compatibility and supports a limited variety of devices. Finally, the research method performs better in terms of energy and resource consumption, which is a significant advantage over the high energy and resource consumption of traditional virtual interaction methods. Through independent sample t-test verification, the differences in response time (t=12.63, p<0.001) and system stability (t=8.92, p<0.001) between multimodal and traditional virtual interactions in this study are statistically significant (p<0.001). This proves that the performance improvement is not a random fluctuation, but rather an inevitable result of multimodal design and engine optimization. In summary, the engine-driven MMI based design outperforms traditional virtual interaction methods in terms of response time, smoothness, system stability, device compatibility, and energy efficiency.

To verify the effect of interaction optimization, the study conducts confounding experiments on the perceptual judgments of different virtual objects. In the experiment, the virtual objects comprised four shapes: sphere, cube, cylinder, and pyramid. Volunteers randomly contacts a virtual object and make perceptual judgments. A total of 10 virtual object contacts are made in each experiment. The results of the volunteers' judgment accuracy in the confusion experiment are shown in Figure 9.



(a) Confusion experiment results of Experimental Group 1

(b) Confusion experiment results of Experimental Group 2

Figure 9: Confusion experiment judgment accuracy results

In Figure 9(a), the volunteers in EG 1 have judgment accuracy of 0.99, 0.96, 0.95, and 0.96 for virtual objects such as sphere, cube, cylinder, and pyramid, respectively. In Figure 9(b), the judgment accuracy of volunteers in EG 2 for the same virtual objects are 0.99, 0.97, 0.95, and 0.96, respectively. In summary, volunteers in both EGs shows high accuracy in the perceptual judgment of virtual objects, and the difference between the two groups is small. In summary, the engine-driven MMI design based on the engine-driven MMI design realizes the interaction optimization design with high accuracy.

3.3 Validation of the effect of user immersion experience

To verify the effect of engine-driven MMI design on user immersion, the study adopts a standardized scale (Igroup presence questionnaire (IPQ)) to assess user immersion. The IPQ scale mainly evaluates the user's perceived realism, emotional response, and other dimensions. The higher the score, the stronger the user's immersion experience. In the experiment, volunteers in EG 1 are designed by engaging in an engine-driven MMI based design. Volunteers in EG 2 uses traditional virtual interaction methods. The study compares the scores of the two groups on each dimension of immersion. The results are shown in Figure 10.

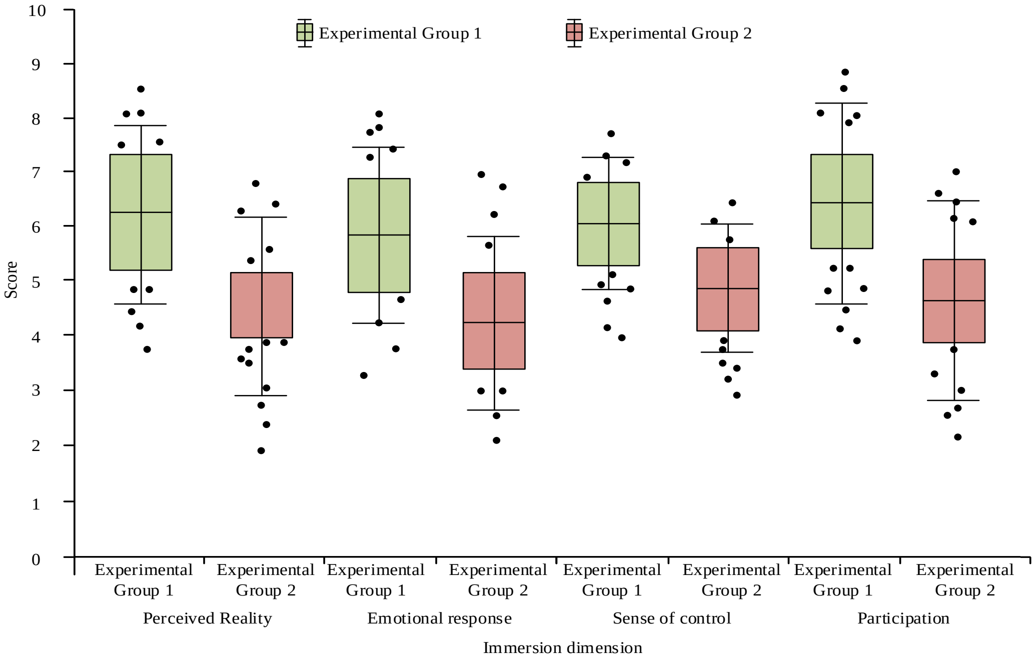


Figure 10: Comparison of scores in immersion dimensions between two groups

In Figure 10, EG 1 scores 6.25±1.70 on the dimension of perceived realism, while EG 2 scores 4.53±1.62 on this dimension. EG 1's score improves by 37.96% compared to EG 2. In the emotional response dimension, the scores of EG 1 and EG 2 are 5.81±1.67 and 4.20±1.59, respectively. There is an improvement of 38.33% in the score of EG 1 as compared to EG 2. In the dimension of sense of control, the scores of EG 1 and EG 2 are 6.04±1.21 and 4.83±1.19, respectively. EG 1 improves by 25.05% compared to EG 2. In terms of participation, the scores of EG 1 and EG 2 are 6.46±1.83 and 4.61±1.87, respectively. EG 1 improves by 40.13% over EG 2. According to independent sample t-test analysis, EG 1 scores significantly higher than EG 2 in four dimensions: perceived reality (t=3.87, p<0.001), emotional response (t=3.65, p<0.001), sense of control (t=2.98, p=0.005), and participation (t=4.12, p<0.001). The differences in all dimensions are statistically significant (p < 0.01), proving that the multimodal interaction design used in this study reliably improves user immersion and is not due to differences in the random sample. In summary, the scores of EG 1 are significantly higher than that of EG 2 in all dimensions. It shows that the engine-driven MMI-based design can significantly enhance the user's immersive experience.

To further validate the immersion effect, the study compares the scores of participating in an engine-driven MMI based design and a traditional virtual interaction approach in terms of emotion evocation using a two-dimensional emotion model of potency-arousal. The 32 VR scenes used in the experiment are not randomly constructed, but are designed based on Russell's Circular Model of Affect (1980) [34]. The efficacy and arousal are evaluated subjectively using the Self Assessment Manikin (SAM) scale. In the study, 'emotional valence/arousal' is a subjective experiential assessment that complements objective physiological responses collected through sensors such as heart rate and brainwaves. The emotion evocation scores for both groups are shown in Table 5.

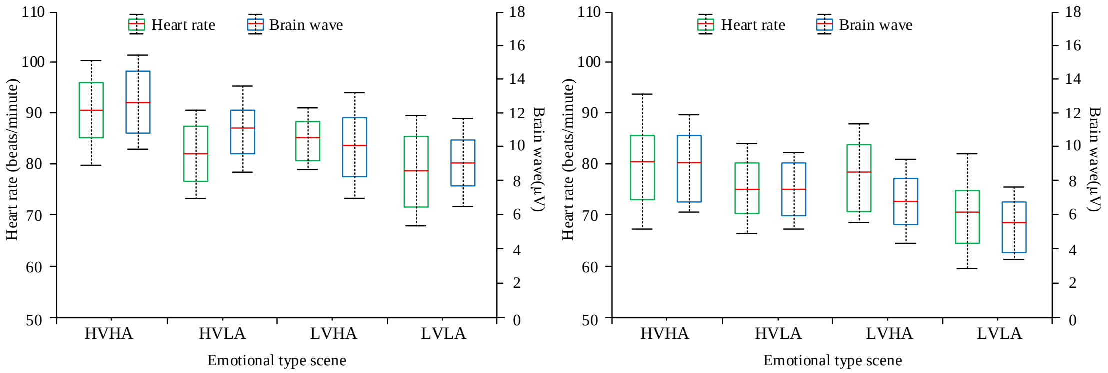
Table 5: Emotional induction scores for two groups

Emotional type	Group	Mean pleasure level	Pleasure standard deviation	Average awakening degree	Standard deviation of awakening degree
HVHA	EG 1	6.85	1.12	7.45	0.98

	EG 2	5.90	1.25	6.20	1.12
HVLA	EG 1	5.40	1.31	6.75	1.05
	EG 2	4.75	1.10	5.80	1.03
LVHA	EG 1	4.50	1.20	6.40	1.15
	EG 2	3.95	1.05	5.50	1.00
LVLA	EG 1	3.25	1.15	5.10	1.25
	EG 2	3.00	1.13	4.83	1.04

In Table 5, in the high valence, high arousal (HVHA) scenario, EG 1 outperforms EG 2 by 0.95 and 1.25 in pleasantness and arousal, respectively. In the high valence, low arousal (HVLA) scenario, EG 1 outperforms EG 2 by 0.65 and 0.95. In the low valence, high arousal (LVHA) type, EG 1 is 0.55 higher in pleasantness and 0.90 higher in arousal. In the low valence, low arousal (LVLA) context, despite the relatively small difference in mood scores between the two groups, EG 1 still has a slight advantage in pleasure and arousal. It shows that it can still play a certain role in mood evocation in a calmer mood state. Repeated measures analysis of variance (ANOVA) is conducted on the pleasure and arousal scores of two groups in different emotional scenarios. The results are showed that in HVHA scenario (pleasure: F=10.25, p<0.001; awakening degree: F=14.83, p<0.001), HVLA scenario (pleasure degree: F=6.72, p=0.012; awakening degree: F=8.35, p=0.006), LVHA scenario (pleasure degree: F=4.91, p=0.032). In the awakening degree (F=9.07, p<0.001), the scores of EG 1 are significantly higher than those of EG 2 (p<0.05). There is no statistically significant difference between the two groups in the LVLA scenario alone (pleasure level: F=2.89, p=0.095; awakening degree: F=3.12, p=0.082). This further proves that the induction effect of this study design in high arousal emotional scenarios is statistically significant. In summary, the VR engine-driven MMI-based design outperforms the traditional two-dimensional video approach in terms of effectiveness of emotional elicitation. In particular, it can stimulate participants' immersion and emotional responses more significantly in high arousal situations.

To validate immersion and emotional responses more visually, the study compares the physiological responses of the two EGs in scenarios with different emotional types, as shown in Figure 11.



(a) Physiological reactions of Experimental Group 1 in emotional scenarios

(b) Physiological reactions of Experimental Group 2 in emotional scenarios

Figure 11: Comparison of physiological responses between two EGs

In Figure 11(a), EG 1 has the strongest physiological response in the HVHA scenario, with a heart rate of 90.57±10.64 beats/min and brain wave activity of 12.55±3.58 μV. In contrast, EG 1 has the weakest physiological response in the LVLA scenario, with a heart rate of 78.64±10.69 beats/min and brain wave activity of 9.08±2.67 μV. In Figure 11(b), EG 2 also has the strongest physiological response to the HVHA scenario, with a heart rate of 80.42±12.93 beats per minute, which is 11.20% lower compared to EG 1. The brainwave activity is 9.05±2.89 μV, which is 27.88% lower compared to EG 1. In the LVLA scenario, EG 2 has the lowest physiological response, with a heart rate of 70.62±11.28 beats per minute, which is 10.19% lower than that of EG 1. The brain wave activity is 5.57±2.09 μV, which is 38.65% lower than that of EG 1. Through independent sample t-test verification, it is found that there are statistically significant differences in heart rate (t=2.76, p=0.010) and brainwave activity (t=3.94, p<0.001) between the two groups in the HVHA scenario. In the LVLA scenario, the differences in heart rate (t=2.15, p=0.038) and brainwave activity (t=4.28, p<0.001) between the two groups are also statistically significant. This proves that the multimodal interaction design in this study has a significant stimulating effect on users' physiological responses, and this effect is statistically reliable. In summary, the virtual environment designed based on MMI has significant advantages in enhancing immersion and emotional response.

To compare the subjective experience of the two EGs, the study collects subjective feedback from the users through a questionnaire to assess the volunteers' perception of the virtual scene. Each dimension is rated using a 1-7 rating scale, with 1 indicating very poor and 7 indicating excellent. The findings of the two EGs are shown in Table 6.

Table 6: Survey results of two Egs

Dimension	EG 1	EG 2
Pleasure	6.25±0.82	4.86±1.25
Authenticity	6.48±0.72	5.24±1.11
Interactivity	6.59±0.61	4.97±1.39
Illusion of immersion	6.30±0.91	5.03±1.05

In Table 6, EG 1 scores significantly higher than EG 2 on all dimensions. The mean rating of EG 1 on pleasure is 6.25±0.82, which is 28.60% higher compared to EG 2. In terms of authenticity, the mean score of EG 1 is 6.48±0.72, which is 23.66% higher compared to EG 2. In terms of interactivity, EG 1 has a rating of 6.59±0.61, which is 32.59% higher than EG 2. In terms of immersion, the average rating of EG 1 is 6.30±0.91, which is 25.24% higher than that of EG 2. It indicates that the design effectively improves the user's sense of immersion. According to independent sample t-test analysis, EG 1 scores significantly higher than EG 2 in the dimensions of pleasure (t=3.57, p<0.001), authenticity (t=3.29, p=0.002), interactivity (t=4.68, p<0.001), and immersion illusion (t=3.41, p<0.001). Moreover, all dimensional differences are statistically significant (p<0.01). This further confirms the reliability of improving user subjective experience. In summary, the virtual environment designed based on MMI can significantly enhance the subjective experience of users.

To further validate the advantages of multimodal design in research, it is compared with two mainstream multimodal VR/AR methods from recent years: tactile systems based on visual gesture and tactile design, and audio systems based on visual audio and gesture design. As shown in Table 7, quantitative analysis of response time, stability, and error handling is also supplemented.

Table 7: Quantitative analysis of different multimodal VR/AR methods

Indicator	Multimodal interaction	Tactile system	Audio system
Response time (ms)	120.74	185.57	152.43
System stability (%)	99.9	97.54	98.14
Error recognition rate (%)	≤0.80	3.23	2.88
Modal failure switching delay (ms)	50	120	105

As shown in Table 7, this study reduces response time by 34.94% compared to the tactile system and by 20.79% compared to the audio system. This improvement is thanks to the AABB collision detection algorithm, which reduces collision judgment computation by 40%. In terms of system stability, this study improves by 2.36% compared to tactile systems and by 1.76% compared to audio systems. This improvement is due to the system's robust control and anti-interference design. In terms of error recognition rate, the error recognition rate of ≤0.80% in this study is only 24.77% of the tactile system and 27.78% of the audio system. The core reason is that the MUFD algorithm solves the problem of modal data conflict through fuzzy intention credibility allocation. In terms of error handling, this study's “modal failure redundancy switching” function only requires a delay of 50 ms, which is much faster than the tactile system's 120 ms and the audio system's 105 ms. This feature allows it to avoid operational delays caused by tactile sensor failures in surgical VR.

To clarify the necessity of the MUFD algorithm, ablation experiments are designed to compare MUFD with two mainstream benchmark fusion techniques: simple weighted sum fusion and rule-based fusion. The ablation experiments are shown in Table 8.

Table 8: Ablation experiment

Fusion technique	Intent recognition accuracy (%)	Error recognition rate (%)	Accuracy in modal conflict scenarios (%)	Accuracy when eye-tracking fails (%)	Average inference time (ms)
Simple weighted sum fusion	82.63	5.87	61.25	75.38	8.2
Rule-based fusion	86.41	4.32	73.75	80.12	11.5
MUFD algorithm	92.37	0.78	91.5	89.85	15.7

As shown in Table 8, simple weighting and fusion cannot cope with fluctuations in modal data quality due to the use of fixed weights. This results in an accuracy rate of only 82.63%. The MUFD calculates the degree of membership for each mode using a Gaussian fuzzy function. It also dynamically adjusts the weights, improving the overall accuracy by 9.74% and reducing the error recognition rate to 0.78%. Rule-based fusion relies on preset logic. When faced with conflicts in eye movement gestures, only rigid rules that prioritize gestures can be used for decision-making. The accuracy rate of conflict scenarios is only 73.75%. MUFD quantifies the degree of modal conflict by using fuzzy intention credibility distribution CF and comprehensive trust allocation m (A), and selects the intention with the highest credibility to improve the accuracy of conflict scenarios by 17.75%. Additionally, when eye movement fails, the accuracy of the simple weighted sum and rule-based fusion decreases to 75.38% and 80.12%, respectively. This is due to their reliance on eye movement data or their failure to redistribute weights when ignoring eye movement. However, the MUFD's redundancy switching mechanism for modal failure still maintains a high accuracy of 89.85%. This proves that its robustness in complex scenarios is significantly better than that of the benchmark method. In summary, the MUFD algorithm performs better in intent recognition accuracy, error control, and robustness. This verifies its necessity as the core algorithm for multimodal fusion in this study.

To verify the construct validity of immersion assessment, Pearson correlation analysis is conducted between the four core sub dimensions of the IPQ scale and physiological signals. The results of the IPQ scale are shown in Table 9.

Table 9: IPQ scale results

IPQ sub dimension	Heart rate (beats/minute)	Brain wave activity level (μ V)
Perceived reality	r=0.623**	r=0.715**
Emotional response	r=0.587**	r=0.692**
Sense of control	r=0.415*	r=0.483**
Participation rate	r=0.591**	r=0.678**

Note: r is the Pearson correlation coefficient, ** *p*<0.01, *p*<0.05.

As shown in Table 9, the correlation between perceived reality and EEG activity is the strongest (r=0.715, *p*<0.01). This indicates that the more realistic a virtual scene appears to a user, the more active their EEG becomes. This finding aligns with the theoretical logic of high immersion being accompanied by high cognitive participation. The correlation between perceived control and heart rate is r=0.415 (*p*<0.05), indicating that an increase in users' sense of control over the interaction process moderately increases physiological arousal. However, the correlation is weaker than that of other dimensions. This may be because perceived control depends more on interaction fluency than emotional arousal. All subdimensions are significantly and positively correlated with physiological signals (*p*<0.05). This indicates that the subjective evaluation results of the IPQ scale are consistent with objective physiological indicators. Thus, subjective scoring bias is eliminated, and the effectiveness of the immersion assessment is verified in this study.

4 Discussion

With the rapid development of VR and AR technology, the application of 3D virtual scenes in entertainment, education, healthcare, engineering, and other fields is becoming increasingly widespread. To enhance users' immersion in virtual scenes, a multimodal interaction system was designed that integrates vision, gestures, and eye tracking for engine-driven 3D scenes. Its performance advantages were verified through experiments.

The results showed that the lowest accuracy rates for studying the interactions of visual, gestural, and eye movements reached 91.42%, 92.83%, and 92.75%, respectively. These rates achieved precise capture of user intent. However, existing research not reported the accuracy of specific interactions in multimodal designs. For example, Al Ansi et al. [11] and Sereno et al. [12] described the functionality of their "visual+audio" and "speech+touch+eye movement" systems, respectively, but did not improve accuracy. The "visual+gesture+speech" system proposed by Zhang Y et al. [16] only mentioned the improvement of intent recognition accuracy, but does not provide specific numerical values. Moreover, in this study, the fusion of intent capture algorithm and MUFD algorithm stabilized the accuracy of the three core interaction modes at over 91%. This result provided a quantitative basis for the accuracy benchmark of multimodal interaction and solved the problem of being unable to verify interaction reliability due to a lack of accuracy evaluation, as described by Sereno et al. [12]. It was especially suitable for scenarios requiring precise recognition of instrument operation intentions in surgical VR.

The disconnect between multimodal design and underlying optimization is prevalent in existing research. For example, the multimodal system proposed by Sharma K et al. [14] did not involve the optimization of interaction delay, while Li J et al. only optimized the tactile delay without integrating multimodal data fusion technology. To deal with those issues, this study achieved a response time of 120.74ms using the AABB collision detection algorithm and the MUFD algorithm with real-time data fusion mechanism. The multimodal system with tactile feedback reduced by 34.94%, while the multimodal system with audio reduced by 20.79%. At the same time, the system stability reached 99.90%, which was significantly higher than the qualitative high-fidelity description of traditional virtual interaction and Lungu AJ et al. [9] surgical simulation system. This performance leap directly solved the core pain points of operational deviation and lag affecting immersion caused by delays in safety critical scenarios. This was because the previous model focused on weight distribution for clear intentions or single-mode data transmission. It did not design solutions for the fuzzy intentions commonly seen in user operations. The MUFD algorithm used a Gaussian ambiguity function to construct a fuzzy set, which could quantitatively characterize uncertain data from vision, gestures, eye movements, and other modes. This avoided the interference of deviation from a single mode of data on the overall judgment.

The results also showed that, in the HVHA scenario, heart rate was 90.57±10.64 beats per minute, while brainwave activity level was 12.55±3.58 μV. These values increased by 11.20% and 27.88%, respectively, compared to those of the traditional interaction group. The reason for the significant improvement in specific physiological indicators is that traditional virtual interactions lack tactile and kinesthetic feedback, leading to a disconnect in user perception. This study utilizes the force tactile feedback and arm motion synchronization mechanisms of smart gloves to give users a realistic tactile and kinaesthetic experience of touch and motion synchronization in virtual scenes. At the same time, by combining YOLOv7 target positioning, eye tracking and other visual technologies, a multi sensory closed loop of “sight touch motion eye” has been constructed. This high perceived reality makes it easier for users to immerse themselves in virtual scenes, thereby triggering stronger physiological arousal. The improvement of physiological indicators is directly related to the smoothness of interaction. A delay or lag in the system will interrupt the user's immersive state and weaken their physiological response. This study used the AABB collision detection algorithm and modal failure redundancy switching to reduce the response time to 120.74 ms and improve system stability to 99.9%. The smooth and seamless interaction experience avoids immersive interruptions, allowing users to maintain high levels of immersion and physiological indicators in HVHA scenarios.

In summary, the performance differences between this study and existing systems are primarily due to technological breakthroughs in multi-device deep integration, a multi-modal real-time fusion mechanism, and underlying interaction optimization. However, although this study has achieved significant improvements to existing systems, there are still certain limitations. For example, the adaptability of different hardware devices has not been considered. Additionally, the sample size of users is small and does not cover different age groups or levels of operating experience. This makes it difficult to verify the system's ability to adapt to personalized needs. Future research should focus on improving system compatibility and expanding support for multiple hardware devices. Additionally, the system's personalized adaptation capabilities should be verified and optimized through large-scale user testing. In addition, more advanced multimodal fusion algorithms and interaction technologies will be further explored to continuously enhance users' immersion and interaction experience in virtual scenes.

5 Conclusion

The study designed MMI and immersion enhancement strategies around engine-driven three-dimensional virtual scenes. The results showed a minimum accuracy of 91.42% for visual interactions, and 92.83% and 92.75% for gesture and eye-movement interactions, respectively. In terms of system performance, the response time of the studied method was only 120.74ms, which was significantly less than the conventional method. The fluency score was scored 4.8/5, which showed a very high fluency. Compared with existing multimodal VR/AR methods, the core improvements of this study were reflected in three aspects: Firstly, adaptive robust control fusion solved the problem of the weak anti-interference ability of existing systems, making them adaptable to safety-critical scenarios, such as surgery and aerospace. Secondly, the combination of the AABB and MUFD algorithms achieved the collaborative optimization of response time and error recognition rate. Performance indicators improved by 20%-35% compared to those of mainstream systems. The third objective was to innovate the “modal failure redundancy switching” mechanism, which controlled fault response delays within 50 ms and avoided the risk of interrupting interactions in safety scenarios. In the results of the user questionnaire, the research methodology scored 6.25±0.82, 6.48±0.72, 6.59±0.61, and 6.30±0.91 for pleasure, authenticity, interactivity, and immersion, respectively. All of them showed a high level of user satisfaction. In summary, the study significantly improves user interaction and immersion through engine-driven MMI design of three-dimensional virtual scenes. However, the study do not address the effects of different devices and individual user differences on the experience. Future research can further optimize the interaction and improve the adaptability to various devices. Moreover, the study will explore the individual needs of different user groups to further enhance the user experience of virtual scenes.

Fundings

The research is supported by: 2025, Research Project of Fujian Province on Vocational Education: Construction of the Talent Cultivation System of “Dual Roles and Industry-Education Collaboration” for the Major of Virtual Reality Technology Application (No. ZJGB2025010).

References

[1] A. P. Chaves and M. A. Gerosa, “How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design,” Int. J. Hum.-Comput. Interact., vol. 37, no. 8, pp. 729-758, November. 2021, DOI: 10.1080/10447318.2020.1841438.

[2] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, and H. Jégou, “ResMLP: Feedforward networks for image classification with data-efficient training,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 4, pp. 5314-5321, April. 2023, DOI:

10.1109/TPAMI.2022.3206148.

[3] D. Djarah, A. Benmakhlouf, G. Zidani, and L. Khettache, "Online Multi-object Tracking with YOLOv9 and DeepSORT Optimized by Optical Flow," Eng. Technol. Appl. Sci. Res., vol. 14, no. 6, pp. 17922-17930, December. 2024, DOI: 10.48084/etasr.8770.

[4] Z. Zhu, C. Su. "From games to education: research on immersive experience-based interactive design of children's educational games," Advances In Industrial Engineering and Management, 2023, 12(1), 24-27. DOI: 10.7508/aiem.01.2023.24.27

[5] G. Apruzzese, L. Pajola, and M. Conti, "The cross-evaluation of machine learning-based network intrusion detection systems," IEEE Trans. Netw. Serv. Manag., vol. 19, no. 4, pp. 5152-5169, November. 2022, DOI: 10.1109/TNSM.2022.3157344.

[6] A. Mathias, S. Dhanalakshmi, and R. Kumar, "Occlusion aware underwater object tracking using hybrid adaptive deep SORT-YOLOv3 approach," Multimedia Tools Appl., vol. 81, no. 30, pp. 44109-44121, May 2022, DOI: 10.1007/s11042-022-13281-5.

[7] S. Dargan, S. Bansal, M. Kumar, A. Mittal, and K. Kumar, "Augmented reality: A comprehensive review," Archives of Computational Methods in Engineering, vol. 30, no. 2, pp. 1057-1080, October. 2023, DOI: 10.1007/s11831-022-09831-7.

[8] J. Dudley, L. Yin, V. Garaj, and P. O. Kristensson, "Inclusive Immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse," Virtual Reality, vol. 27, no. 4, pp. 2989-3020, September. 2023, DOI: 10.1007/s10055-023-00850-8.

[9] A. J. Lungu, W. Swinkels, L. Claesen, P. Tu, J. Egger, and X. Chen, "A review on the applications of virtual reality, augmented reality and mixed reality in surgical simulation: an extension to different kinds of surgery," Expert Rev Med Devices, vol. 18, no. 1, pp. 47-62, December. 2021, DOI: 10.1080/17434440.2021.1860750.

[10] M. L. Duarte, L. R. Santos, J. B. G. Júnior, and M. S. Peccin, "Learning anatomy by virtual reality and augmented reality. A scope review," Morphologie, vol. 104, no. 347, pp. 254-266, June. 2020, DOI: 10.1016/j.morpho.2020.08.004.

[11] A. M. Al-Ansi, M. Jaboob, A. Garad, and A. Al-Ansi, "Analyzing augmented reality (AR) and virtual reality (VR) recent development in education," Social Sciences & Humanities Open, vol. 8, no. 1, pp. 100532-100535, July. 2023, DOI: 10.1016/j.ssaho.2023.100532.

[12] M. Sereno, X. Wang, L. Besançon, M. J. McGuffin, and T. Isenberg, "Collaborative work in augmented reality: A survey," IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 6, pp. 2530-2549, October. 2020, DOI: 10.1109/TVCG.2020.3032761.

[13] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ""Let me explain!": exploring the potential of virtual agents in explainable AI interaction design," J Multi-modal User Interfaces, vol. 15, no. 2, pp. 87-98, July. 2021, DOI: 10.1007/s12193-020-00332-0.

[14] K. Sharma and M. Giannakos, "Multi-modal data capabilities for learning: What can multi-modal data tell us about learning?" Br J Educ Technol, vol. 51, no. 5, pp. 1450-1484, July. 2020, DOI: 10.1111/bjet.12993.

[15] Y. Liu, S. Zhao, and S. Cheng, "Augmenting collaborative interaction with shared visualization of eye movement and gesture in VR," Comput. Animat. Virtual Worlds, vol. 35, no. 3, pp. 2264-2268, Jun. 2024, DOI: 10.1002/cav.2264.

[16] L. Cao, H. Zhang, C. Peng, and J. T. Hansberger, "Real-time multimodal interaction in virtual reality—a case study with a large virtual interface," Multimed. Tools Appl., vol. 82, no. 16, pp. 25427-25448, Feb. 2023, DOI: 10.1007/s11042-023-14381-6.

[17] H. Cui, Z. Feng, J. Tian, D. Kong, Z. Xia, and W. Li, "MAG: a smart gloves system based on multimodal fusion perception," CCF Trans. Pervasive Comput. Interact., vol. 5, no. 4, pp. 411-429, Sep. 2023, DOI: 10.1007/s42486-023-00138-5.

[18] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," IEEE Commun Surv Tutorials, vol. 25, no. 1, pp. 538-566, October. 2023, DOI: 10.1109/COMST.2022.3233793.

[19] M. J. A. Soeb, M. F. Jubayer, T. A. Tarin, M. R. Al Mamun, F. M. Ruhad, A. Parven, et al., "Tea leaf disease detection and identification based on YOLOv7 (YOLO-T)," Sci. Rep., vol. 13, no. 1, pp. 6078-6082, April. 2023, DOI: 10.1038/s41598-023-33270-4.

[20] X. SHAO, X. LI, T. YANG, Y. YANG, S. LIU, and Z. YUAN, "Underground personnel detection and tracking based on improved YOLOv5s and DeepSORT," Coal Sci. Technol., vol. 51, no. 10, pp. 291-301, May 2023, DOI: 10.13199/j.cnki.cst.2022-1933.

[21] K. Feng, W. Huo, W. Xu, M. Li, and T. Li, "CNA-DeepSORT algorithm for multi-target tracking," Multimedia Tools Appl., vol. 83, no. 2, pp. 4731-4755, January. 2024, DOI: 10.1007/s11042-023-15813-z.

[22] Y. Liu, B. An, S. Chen, and D. Zhao, "Multi-target detection and tracking of shallow marine organisms based on improved YOLO v5 and DeepSORT," IET Image Process., vol. 18, no. 9, pp. 2273-2290, April. 2024, DOI: 10.1049/ipr2.13090.

[23] M. K. Rao and P. A. Kumar, "Advanced Object Tracking in Video Surveillance Systems with Adaptive Deep SORT Enhancement," Eng. Technol. Appl. Sci. Res., vol. 15, no. 2, pp. 20871-20877, April. 2025, DOI: 10.48084/etasr.9529.

[24] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated RCNN and multi-class deep sort for multi-object detection and tracking," IEEE Trans. Emerg. Topics Comput. Intell., vol. 6, no. 1, pp. 171-181, Jan. 2021, DOI: 10.1109/TETCI.2020.3041019.

[25] C. Creed, M. Al-Kalbani, A. Theil, S. Sarcar, and I. Williams, "Inclusive AR/VR: accessibility barriers for immersive technologies," Universal Access in the Information Society, vol. 23, no. 1, pp. 59-73, February. 2024, DOI: 10.1007/s10209-023-00969-0.

[26] L. Zarantonello and B. H. Schmitt, “Experiential AR/VR: a consumer and service framework and research agenda,” J Serv Manag, vol. 34, no. 1, pp. 34-55, January. 2023, DOI: 10.1108/JOSM-12-2021-0479.

[27] F. Zouari, K. B. Saad, and M. Benrejeb, “Adaptive backstepping control for a class of uncertain single input single output nonlinear systems,” in Proc. 10th Int. Multi-Conferences Syst., Signals & Devices (SSD13), IEEE, 2013, pp. 1-6, Mar. 2013, DOI: 10.1109/SSD.2013.6564134.

[28] G. Rigatos, M. Abbaszadeh, B. Sari, et al., “Nonlinear optimal control for a gas compressor driven by an induction motor,” Results Control Optim., vol. 11, p. 100226, Jun. 2023, DOI: 10.1016/j.rico.2023.100226.

[29] L. Kleygrewe, R. I. V. Hutter, M. Koedijk, and R. R. Oudejans, “Virtual reality training for police officers: A comparison of training responses in VR and real-life training,” Police Pract Res, vol. 25, no. 1, pp. 18-37, February. 2024, DOI: 10.1080/15614263.2023.2176307.

[30] F. Zouari, K. B. Saad, and M. Benrejeb, “Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor,” in Proc. 2013 Int. Conf. Control, Decision Inf. Technol. (CoDIT), IEEE, 2013, pp. 864-871, May 2013, DOI: 10.1109/CoDIT.2013.6689656.

[31] Hamel, F. Zouari, “Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities,” Math. Probl. Eng., vol. 2017, no. 1, p. 8045803, Mar. 2017, DOI: 10.1155/2017/8045803.

[32] Y. Pathak, P. K. Shukla, A. Tiwari, S. Stalin and S. Singh, “Deep transfer learning-based classification model for COVID-19 disease,” IRBM, vol. 43, no. 2, pp. 87-92, Apr. 2022, DOI: 10.1016/j.irbm.2020.05.003.A. Boulkroune, S.

[33] A. M. Usman and M. K. Abdullah, “An Assessment of Building Energy Consumption Characteristics Using Analytical Energy and Carbon Footprint Assessment Model,” Green Low-Carbon Econ, vol. 1, no. 1, pp. 28-40, January. 2023, DOI: 10.47852/bonviewGLCE3202545.

[34] K. A. Romano, K. E. Heron, G. Ferguson, and S. B. Scott, “Emotion word use patterns and eating disorder symptoms: Considering the circumplex model of affect and basic emotions theory,” Int. J. Eat. Disord., vol. 56, no. 2, pp. 464-469, Dec. 2023, DOI: 10.1002/eat.23879.