International Journal of Technology

http://ijtech.eng.ui.ac.id

Research Article

Performance Evaluation of an XR-based Immersive Room for Climate Change Storytelling Using Multisensory Interaction

Hestiasari Rante[1*], M. Agus Zainuddin[1], Sonki Prasetya[2], Haolia Rahman[2], Cahya Miranto[3], Ardiman Firmanda[3], Sritrusta Sukaridhoto[1], Norhaida Mohd Suaib[4], Dwi Mulyo[1]

[1]Department of Informatics and Computer Engineering, Politeknik Elektronika Surabaya, Jl. Raya ITS, Keputih, Sukolilo, Surabaya, East Java 60111, Indonesia

[2]Department of Mechanical Engineering, Politeknik Negeri Jakarta, Jl. Prof. DR. G.A. Siwabessy, Kukusan, Kecamatan Beji, Kota Depok, Jawa Barat 16425, Indonesia

[3]Department of Informatics Engineering, Game Technology, Politeknik Negeri Batam, Jl. Ahmad Yani, Tlk. Tering, Kec. Batam Kota, Kota Batam, Kepulauan Riau 29461, Indonesia

[4]Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, Jalan Iman, 81310 Skudai, Johor Darul Ta'zim, Malaysia

*Corresponding author: hestiasari@pens.ac.id; Tel.: +62 811-331-015

Abstract: Climate change communication increasingly requires innovative approaches to enhance public awareness. However, existing XR-based climate communication systems rarely provide rigorous technical performance evaluation of integrated multimodal environments. This study presents the design and technical performance evaluation of an XR-based immersive system integrating three-wall projection mapping, gesture and voice interaction, and multisensory physical feedback. The system is implemented using a unified pipeline combining Unity, MediaPipe, TouchDesigner, Resolume, and Arduino to enable real-time multimodal interaction. Experimental evaluations were conducted to assess system responsiveness and interaction reliability, including latency analysis using a 240 FPS high-speed camera, gesture recognition accuracy under varying lighting and distance conditions, and voice recognition accuracy across multiple noise levels. Results show a consistent user-perceived latency of 16.68 ms, gesture accuracy of 90–100% in bright conditions and 90% at 160 cm distance in low-light conditions, and voice recognition accuracy ranging from 100% (30 dB) to 30% (80 dB). These findings demonstrate that the system achieves stable and responsive multimodal interaction in an XR immersive environment. The study highlights key design considerations and operational constraints, providing a technical foundation for future development of XR-based interactive systems. This study is positioned as an initial technical performance evaluation of a multimodal XR system.

Keywords: Climate change storytelling; Extended reality; Gesture recognition; Multisensory interaction; Voice interaction

1. Introduction

Climate change represents one of the most severe and pressing challenges confronting the global community, with profound implications for both human societies and natural ecosystems. The scientific consensus, articulated by the Intergovernmental Panel on Climate Change (IPCC), confirms that human activities are the primary driver of global warming. The average global surface temperature between 2011 and 2020 rose by 1.1°C above pre-industrial levels (1850–1900), a trend that continues to accelerate (Lee and Romero, 2023). This warming is fuelled by a relentless increase in greenhouse gas (GHG) emissions, with atmospheric carbon dioxide (CO_2) concentrations now exceeding 420 parts per million (ppm)-a 150% increase from

https://doi.org/10.14716/ijtech.v17i3.8400
Received January 2026; Revised March 2026; Accepted April 2026

the pre-industrial era (Climate Division of the Meteorological and Geophysical Agency (BMKG), 2024). Indonesia, an archipelagic nation of immense ecological and social diversity, is particularly vulnerable to the impacts of climate change. The country experienced its hottest year on record in 2024, with a national average temperature anomaly of +0.8°C relative to the 1991-2020 baseline (Coaction Indonesia, 2024). As the world's sixth-largest GHG emitter, Indonesia contributes approximately 2% of the global total, yet it also ranks among the top 15 countries most affected by climate-related risks, according to the Global Climate Risk Index 2021 (Baroroh and Agarwal, 2022; Ecstein et al., 2021).

Despite the urgency of this threat, public awareness and engagement with climate change issues in Indonesia remain alarmingly low. A recent analysis of social media conversations revealed that only 1.3% of popular digital discourse in the country pertains to climate change, highlighting a critical communication gap that hinders collective action (Slater and Sanchez-Vives, 2016).

Innovative and engaging educational strategies are required to address this communication gap. Conventional science communication approaches, such as reports and documentaries, often face limitations in conveying the scale, urgency, and experiential consequences of climate change in ways that resonate emotionally and motivate behavioral change. In this context, immersive technologies such as Extended Reality (XR), have emerged as promising tools for delivering impactful learning experiences. XR, which encompasses Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), enables the simulation of complex phenomena within vivid, interactive, and multisensory environments (Zhang et al., 2025). Recent studies have demonstrated that XR can enhance user engagement and improve information retention by enabling interactive and immersive experiences by allowing users to directly experience climate-related scenarios, such as rising sea levels, deforestation, and extreme weather events (Bustos-Lopez et al., 2024; Sherman and Craig, 2018). Embodied cognition theories strongly support these immersive experiences, which emphasize that learning is deepened when abstract concepts are coupled with physical and sensory interaction (Dede, 2009; Ioannou et al., 2021; Jerald, 2015; Kang et al., 2021; Radianti et al., 2020). Moreover, the increasing adoption of immersive technologies in Indonesia, particularly in education, tourism, and public engagement, indicates a strong potential for XR-based climate storytelling initiatives' broader application (Bailenson, 2018). Immersive simulations have also been shown to promote pro-environmental attitudes and behavioral intention when users first experience environmental degradation (Fernández Galeote et al., 2023; Kizhevska et al., 2023). XR-based learning is strongly supported by embodied cognition theories, which emphasize that physical and sensory interactions enhance conceptual understanding (Dede, 2009; Ioannou et al., 2021; Kang et al., 2021).

Despite these advances, the application of XR for large-scale public climate change communication remains limited, particularly in developing countries, such as Indonesia. Most existing XR-based climate education studies emphasize narrative immersion or user experience and typically focus on qualitative assessment or conceptual design. As a result, rigorous and quantitative technical evaluations, such as interaction latency, gesture recognition robustness, and voice command reliability under real-world conditions, are rarely reported.

Therefore, this study targets the Indonesian public as the primary audience, aiming to deliver climate change content through an interactive and multimodal XR system with a focus on technical performance and system robustness. Immersive environments enable users to interact with dynamic content through multiple sensory channels, thereby improving the clarity and interactivity of complex information (Ignacio et al., 2015; Spence and Gallace, 2011; Treal et al., 2021). XR also offers opportunities for spatial narratives that cannot be replicated through traditional screens, especially when addressing abstract phenomena such as climate systems, sea-level rise, and extreme weather patterns (Makransky and Petersen, 2021).

In response to this need, this study develops an XR-based immersive room for climate change awareness. The system integrates dynamic visual elements, user interaction through gestures and voice, and physical environmental effects that respond directly to the storyline to

create an interactive and multisensory XR experience (Kopcha et al., 2021; Lee-Cultura and Giannakos, 2020; Viswakumar et al., 2022). The proposed approach combines digital content with real physical space to deliver climate change content through a technically integrated XR system with multimodal interaction and multisensory feedback.

This work also represents the continuation of a previous project (Phase 1) conducted by the TRM PENS consortium in 2021, which focused on developing a virtual reality (VR) platform for climate change awareness. The earlier system was implemented using the Vircadia open-source metaverse platform, combined with Blender-based 3D assets and Ready Player Me avatars, and integrated with IoT sensors powered by solar panels for real-time water quality monitoring (Rante et al., 2024). While Phase 1 demonstrated the feasibility of VR-based climate communication, it was limited to headset-based interaction and virtual-only environments.

In this second phase, the approach is extended beyond headset-based VR toward an XR immersive room that incorporates spatial, physical, and multisensory interaction. This includes gesture recognition, voice commands, and physical environmental effects such as lighting, airflow, and smoke that dynamically respond to the narrative flow (Jeong et al., 2014; Mufarroha and Utaminingrum, 2017; Pham et al., 2020). By expanding into a projection-based XR environment, this study aims to support interactive exploration of climate change scenarios through a more compelling, embodied, and technically robust storytelling experience.

2. Related Works

Several studies have explored the role of XR technologies in education and environmental storytelling, and others have addressed technical challenges related to 3D asset optimization. Research that combines immersive system engineering (interaction, multisensory feedback, projection mapping) with high-performance asset pipelines for climate change-focused storytelling is less common.

In the field of education, (Huang and Tseng, 2025) found that of the 56 empirical studies between 2010 and 2024, about 53% used VR, 42% used AR, and a small fraction (5%) used MR, often combining systems. The results indicate the effectiveness of XR in improving motivation, learning outcomes, and engagement, although many studies have reported short-term deployment and challenges in hardware accessibility and asset performance.

Several recent studies have highlighted asset optimization in immersive applications. The paper "Performance analysis of 3D assets in virtual reality simulations for climate change" (Miranto et al., 2025) provides empirical evidence showing how different asset properties directly affect real-time performance on standalone VR headsets. The study reports that transparent tree foliage, despite having relatively low triangle counts, caused early and significant FPS degradation due to overdraw and poor batching efficiency, dropping from 120 FPS at 20 trees to 78 FPS at 40 trees, and reaching 39 FPS at 100 trees. In contrast, high-polygon animated characters (>16,000 triangles each) maintained smoother performance because they used shared opaque materials that enabled efficient batching, sustaining 120 FPS up to 40 characters and remaining above the VR comfort threshold ($\geq$72 FPS) until around 70 characters. These findings demonstrate that material complexity, especially transparency, imposes a greater performance penalty than geometric complexity, emphasizing the importance of optimizing materials, batching, and LOD strategies in the development of immersive applications.

Similarly, (Jamil et al., 2024) in "Optimizing 3D Assets and Character Modeling of the Mixed Reality Simulator in a Disaster Mitigation Learning Using Vertex Decimation and Depth-of-Field Algorithm" proposed constraints such as maximum triangle or vertex count for levels of detail (LoD) to balance performance versus visual realism. The rendering times for the assets ranged from 0.8 and 7.7 ms using these optimization strategies.

In projects specifically focused on climate change storytelling, "Immersive Climate Stories" by Indiana University's Environmental Resilience Institute merges real-world data, case studies, and XR video elements to evoke emotional responses and promote public engagement with climate impacts such as heatwaves and flooding.

Meanwhile, "Visualizing the Causes and Impacts of Climate Change with XR Technologies" by Scott Birch (Birch et al., 2023) demonstrates the effective use of XR to show both cause and effect in immersive visualizations, enabling users to explore climate scenarios interactively.

The gap in these studies remains clear: there are individual studies on interactive immersive storytelling, studies on 3D asset optimization, and some climate-centered XR projects, but very few that (a) integrate gesture/voice/multisensory feedback in a physical immersive room; (b) enforce strict 3D asset optimization to ensure smooth performance in such multi-component setups; and (c) evaluate both technical performance and user experience in one combined system for climate change awareness.

Therefore, this study is the first to integrate gesture-based control, voice interaction, multisensory physical actuators, real-time projection mapping (Liu et al., 2022; Mikawa et al., 2018; Wakunami et al., 2016), and optimized 3D assets into a unified XR room specifically designed for climate change storytelling.

However, most existing studies focus on user experience or conceptual design, with limited emphasis on rigorous technical performance evaluation of integrated XR systems.

3. Methods

This section provides a detailed description of the methodology used to design, develop, and evaluate the XR-based immersive room for climate change storytelling. Our approach is centered on creating a robust and integrated system that combines multimodal interaction, multisensory feedback, and a high-performance rendering pipeline. We provide sufficient detail to enable other researchers to replicate our study, citing established methods where appropriate and focusing on this work's specific adaptations and innovations.

3.1 System Overview

The cornerstone of our immersive room is a tightly integrated system architecture that synchronizes multiple software and hardware components to deliver a cohesive and responsive user experience. The system was designed as a 3x3-meter booth, equipped with a three-wall projection, an array of physical effect devices, and a sophisticated interactive system. Figure 1 shows the overall architecture, which integrates Unity as the central rendering engine, TouchDesigner for real-time generative visuals, Resolume for projection mapping, Arduino for physical effect control, and Google's MediaPipe for AI-powered gesture recognition. Unity integrates voice command recognition directly to enable a seamless multimodal interaction framework (Hu et al., 2018; Kim and Shin, 2021; Nossier et al., 2021). The system follows an input-process-output workflow, as shown in Figure 1.

3.1.1 Input Layer

The system captures user interactions through two primary modalities: voice and gesture. Voice commands are captured by a microphone and processed by Unity's integrated speech recognition engine, allowing participants to trigger narrative events or make choices within the story. Simultaneously, a camera feeds video data to the MediaPipe framework, which uses a sophisticated artificial intelligence (AI) model to detect and classify hand and body gestures in real-time. These natural and intuitive inputs are converted into digital signals, which are then passed to the processing layer.

3.1.2 Processing Layer

Unity serves as the processing layer's central hub, orchestrating the entire interactive experience. It manages 3D assets, including environmental models, animated characters, and visual effects, and executes narrative logic based on user inputs. Unity communicates with TouchDesigner via the Spout protocol to create a dynamic and visually rich environment, allowing for the

seamless integration of real-time generative visuals such as flowing water, atmospheric haze, and particle effects. For the projection mapping, Unity sends the final rendered frames to Resolume, which warps and blends the images to fit perfectly across the immersive booth's three walls. In parallel, Unity sends Open Sound Control (OSC) messages to an Arduino microcontroller, which triggers a series of relays to activate the physical actuators in the environment.

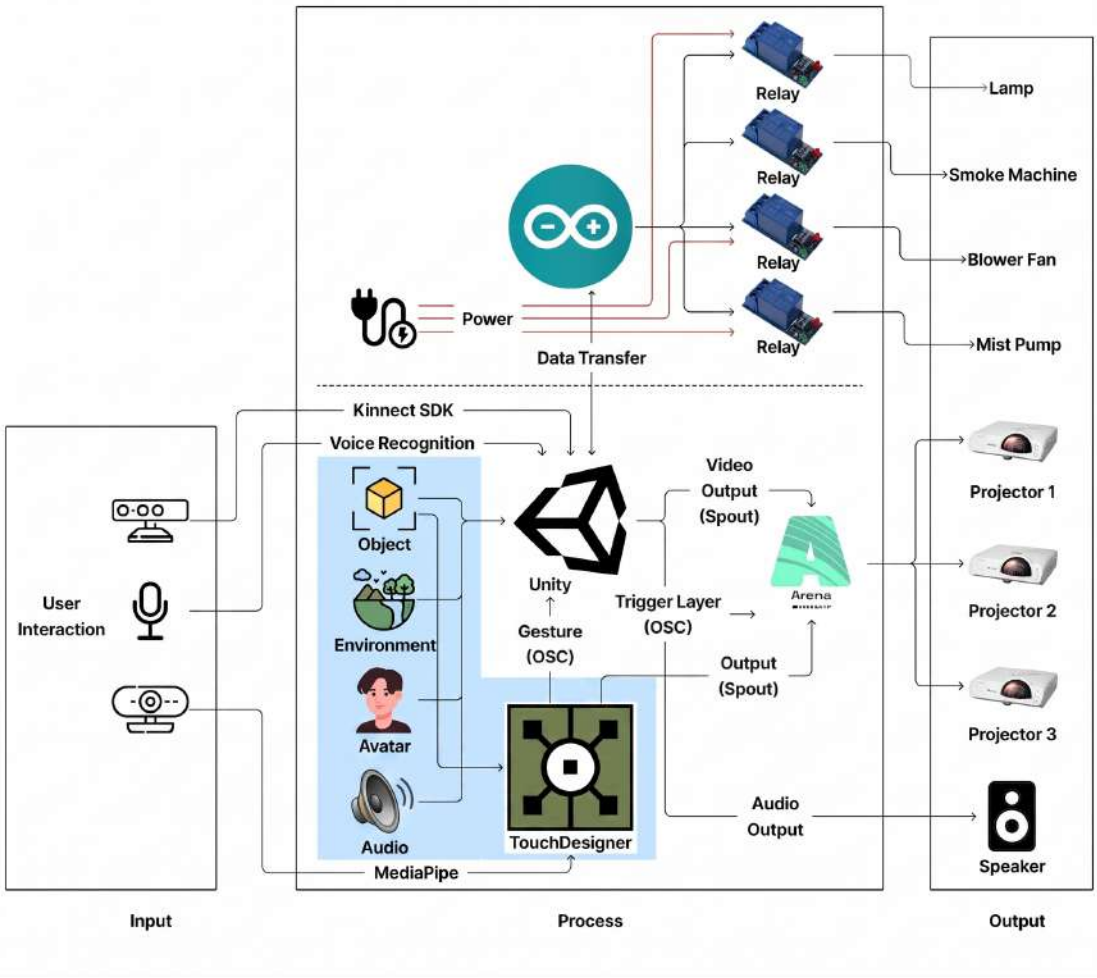

Figure 1 System architecture of the XR immersive room

3.1.3 Output Layer

The output layer delivers a multisensory experience that combines visual, auditory, and physical feedback. The three-wall projection is the primary visual output, which creates a seamless and immersive panoramic view of the virtual environment. This is complemented by a surround sound system that provides directional audio cues and an immersive soundscape. The output layer's most innovative aspect is the array of physical actuators, which are synchronized with the on-screen narrative. These include blower fans to simulate wind, a water spray system to create the sensation of rain, a smoke machine to generate haze and fog, and a dynamic lighting system to mimic changes in the time of day or the intensity of a storm. The system aims to create a powerful sense of presence and deepen the user's emotional connection to the climate change narrative by engaging multiple senses.

3.2 Hardware and physical setup

The immersive room was physically implemented as a 3×3-meter booth, designed to simulate an enclosed interactive space with multisensory feedback. The booth consists of three

projection walls, audio systems, and environmental effect devices strategically positioned to create an engaging storytelling environment.

The visual system uses three high-lumen projectors mounted on the ceiling frame, each directed toward a wall to form a panoramic projection. Surround audio is delivered through four speakers mounted in the upper corners of the booth, ensuring realistic spatial soundscapes such as rainfall, storm winds, or forest ambience. A motion capture camera and sensors are also included to detect gesture-based interactions. As shown in Figures 2 and 3, the projectors and speakers are symmetrically arranged within the booth to maximize coverage and minimize visual distortion.

Environmental actuators are integrated to provide multisensory immersion, including a smoke machine, blower fan, mist pump, and controllable lamps. These devices are connected to an Arduino-based controller that synchronizes their activation with Unity events. For example, during a storm, the mist pump sprays fine water droplets while blower fans simulate wind, and lamps flicker to represent lightning.

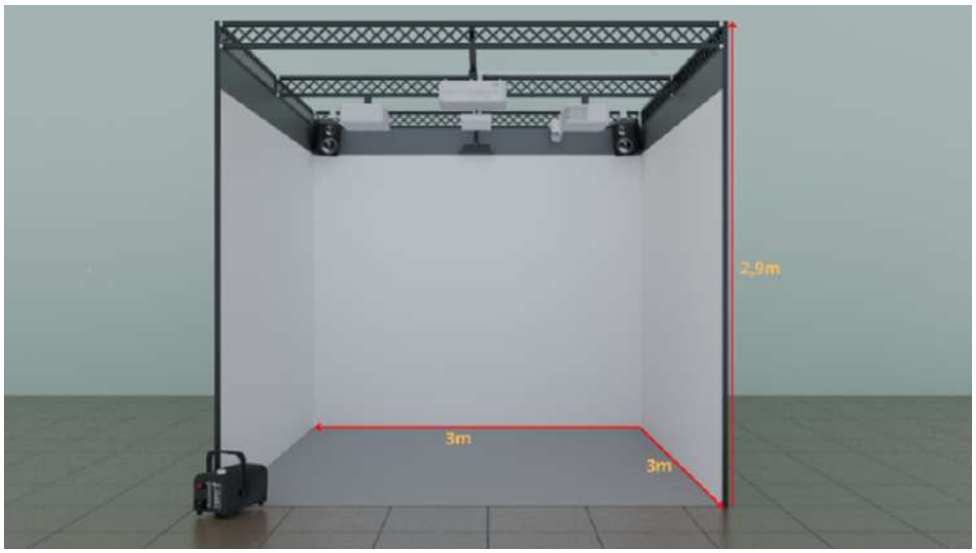

Figure 2 Front view of the 3×3-meter immersive booth design, illustrating projection walls, actuator placement, and user position

Figure 3 Top-corner perspective of the immersive booth setup, showing the projectors, speakers, and environmental device arrangement

Table 1 summarizes the technical specifications of the hardware components in addition to

the spatial layout, which provides details on the devices used for projection, audio, control, and physical effect simulation.

Table 1 Hardware specifications of XR immersive room

Component	Model	Key Specification	Function
Projector	Epson EB-L210SW	WXGA (1280×800), 4000 lumens, Laser light source, 3LCD technology, 20,000 hours lifespan	Displays visuals on walls/screens with high brightness and clarity
Speaker	Edifier MR4 Powered Studio Monitor Speakers 4”	Frequency response: 60 Hz–20 kHz, 21 W RMS each, 4” woofer + 1” silk dome tweeter, Balanced TRS/XLR + RCA + AUX input	Produces clear audio output for immersive experience
Arduino Board	Arduino Uno R3	Microcontroller: ATmega328P, 5 V operating voltage, 14 digital & 6 analog I/O pins, 16 MHz clock speed, USB interface	Controls physical devices (triggering environmental effects)
Relay	Relay Module 4 Channel	5V trigger voltage, optocoupler isolation, 10A 250VAC / 10A 30VDC load	Switches high-power devices (fans, floodlights, pumps) from Arduino
Web Camera	Logitech C270	720p HD video, 30 FPS, fixed focus, built-in noise-reducing mic, USB 2.0	Captures user presence for photobooth or interaction input
Depth Camera	Xbox Kinect V2	Depth camera: 512×424 @ 30 FPS, RGB camera: 1080p, Field of View: 70° × 60°, Tracks up to 6 people	Captures gesture and body movement for interactivity
USB Extender	Websong	Extends USB signal up to 10 m (active repeater cable)	Extends USB device connection distance
HDMI Cable	HDMI Active Optical Fiber 4K UHD	4K @ 60 Hz, 18 Gbps bandwidth, length up to 100 m, HDR compatible, HDCP 2.2	Provides stable long-distance video transmission without signal loss
Audio Cable	Vention AUX Splitter 3.5mm Male to Dual 6.5mm Male	Gold-plated connectors, stereo 3.5mm to dual 6.35mm TS (L/R)	Connects audio from PC to powered speakers (left/right channel)
Fog Machine	900 W	Power 900 W	Generates fog/mist effect for immersive ambiance
Mist Pump	DC Slinelader 80 Psi, 0.1 mm nozzle	High-pressure fine mist spray, 12 V DC 80 Psi	Produces light mist or rain-like environmental effect
Blower Fan	8” Inline Fan	8-inch diameter, adjustable air output	Creates wind or cooling air effects
Lamps	Led Flood Light 12V 50W	12 V DC, 50 W, LED type	Provides spotlight or ambient/flash lighting effects

3.3 Software integration workflow

The immersive room system required a robust software integration workflow to ensure synchronization between visual content, user interaction, and physical feedback (Figure 4). Unity served as the primary development platform, responsible for real-time rendering, scene management, and input and output orchestration. Unity received gesture data from MediaPipe and voice commands via its speech recognition API, which were then processed to trigger scene changes or interactive responses. TouchDesigner was used to generate procedural visual effects, such as smoke, flowing water, and abstract transitions, which were transmitted to Unity using the Spout protocol. These outputs were further routed through Resolume, which handled projection mapping to precisely align visual content with the three-wall display in the booth. Physical actuators, including fans, lamps, and a smoke machine, were managed via Arduino controllers, which received Unity Open Sound Control (OSC) messages. This modular workflow allowed each software component to operate independently while maintaining seamless communication across the system.

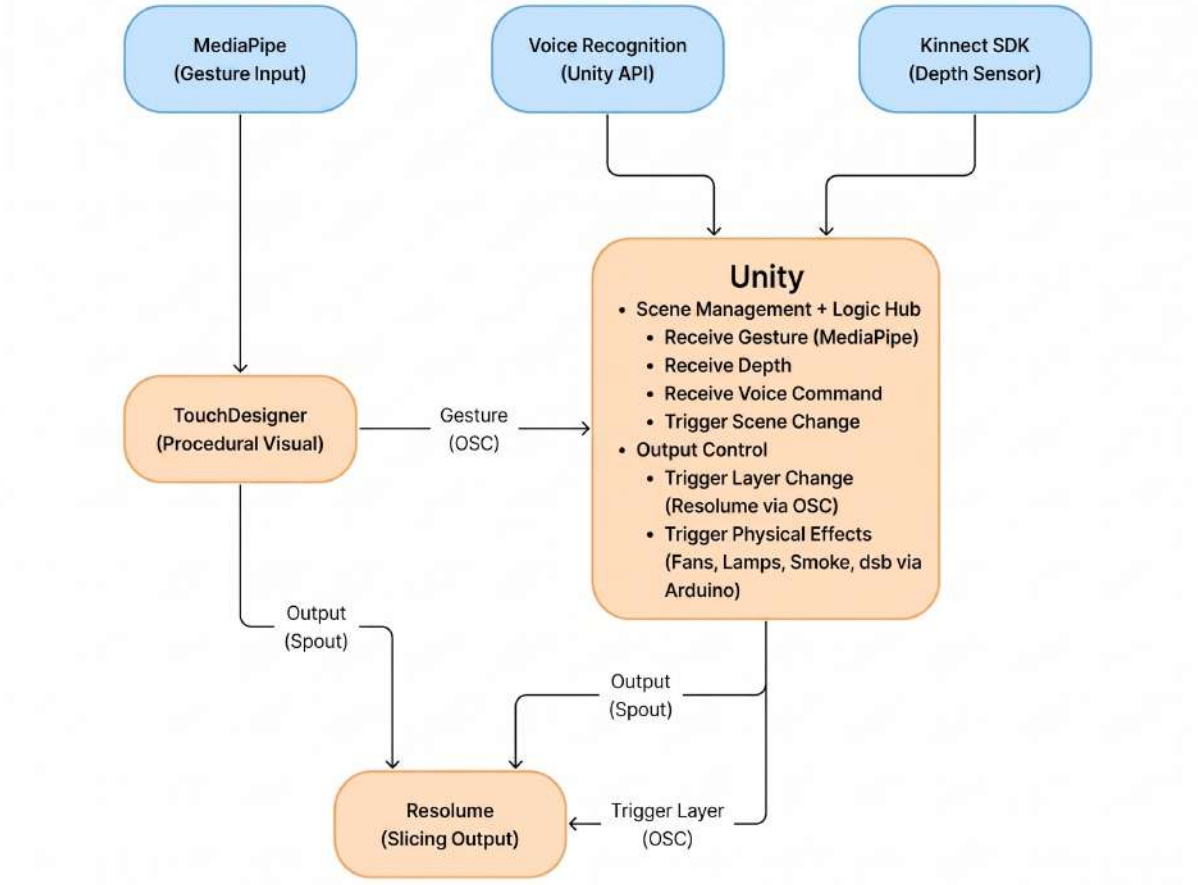

Figure 4 Software integration workflow showing communication between Unity, MediaPipe, TouchDesigner, and Resolume

3.4 3D Asset Creation and Optimization

The creation of high-quality, performant 3D assets is a critical component of any immersive XR experience. We developed a meticulous asset creation and optimization pipeline to ensure a smooth and stable frame rate, as illustrated in Figure 5. This workflow balances visual fidelity with the demanding real-time rendering requirements of a multisensory, multi-projection

environment.

The process began with the creation of 3D models in Blender, a powerful open-source modeling and animation tool. Our artists created a range of assets, including environmental elements (trees, rocks, and buildings), animated characters, and props relevant to the narrative of climate change. Following the initial modeling phase, the assets were textured in Adobe Substance Painter, allowing creating realistic materials with detailed surface properties. We employed several techniques to optimize performance, including the use of the Decimate Modifier in Blender to reduce polygon counts without sacrificing significant visual detail. We also pre-baked lighting and shadow information into the textures, a process known as "texture baking," which significantly reduces the Unity rendering engine's computational load. These texture maps were typically generated at a resolution of 2048x2048 pixels.

We created simple rigs for assets that required animation, such as characters or dynamic objects, to enable basic movements such as walking, pointing, or collapsing. Once the assets were fully modeled, textured, and optimized, they were exported to Unity and integrated into the game engine. Animator Controllers were then used to link the animations to the narrative timeline, ensuring that the movements of characters and objects were synchronized with the unfolding story. This comprehensive optimization strategy was crucial for maintaining a stable frame rate and a high immersion level within the XR environment.

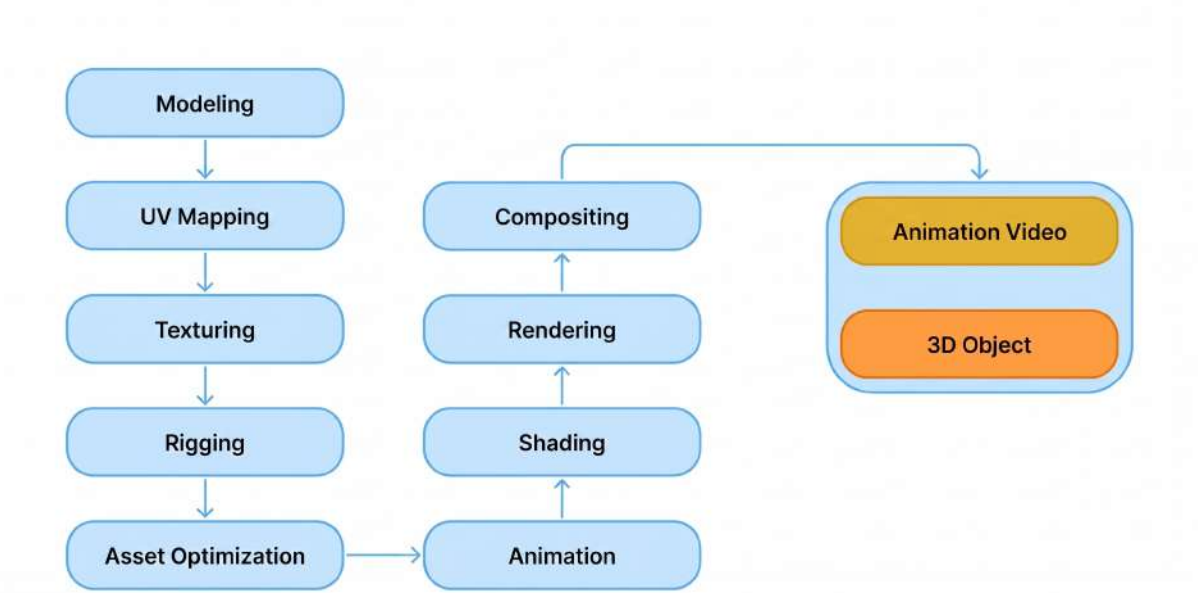

Figure 5 Workflow of 3D asset creation and optimization in Blender

3.5 Storytelling and Interaction Design

The narrative framework of our immersive room was structured around a sequence of compelling climate change scenarios, including deforestation, rising global temperatures, intense flooding, and widespread air pollution. As shown in the story flow diagram in Figure 6, each scene was designed to present information and actively engage the participant through intuitive gesture and voice-based interactions. This participatory approach transforms the user from a passive observer into an active agent within the narrative, fostering a deeper sense of connection and responsibility.

For example, in one scene, the user can raise their hands to summon a rainstorm, triggering a cascade of synchronized effects: falling rain projection, downpour sound, fine mist sprayed from the ceiling, and cool breeze from the blower fans. In another scene, the user's spoken responses to yes-or-no questions can alter the narrative's course, leading to different outcomes and reinforcing the idea that individual choices have consequences. The integration of physical effects, such as the flickering of lamps during a fire or smoke dispersal to simulate pollution,

further enhances the sense of presence and realism.

Our system creates a powerful and memorable experience by weaving together a compelling narrative with multimodal interaction and multisensory feedback. This approach is designed to improve the interactivity and clarity of information delivery, making the abstract concepts of climate change more tangible and immediate. By allowing users to actively participate in the story, we aim to inspire a sense of agency and encourage reflection on the real-world implications of climate change.

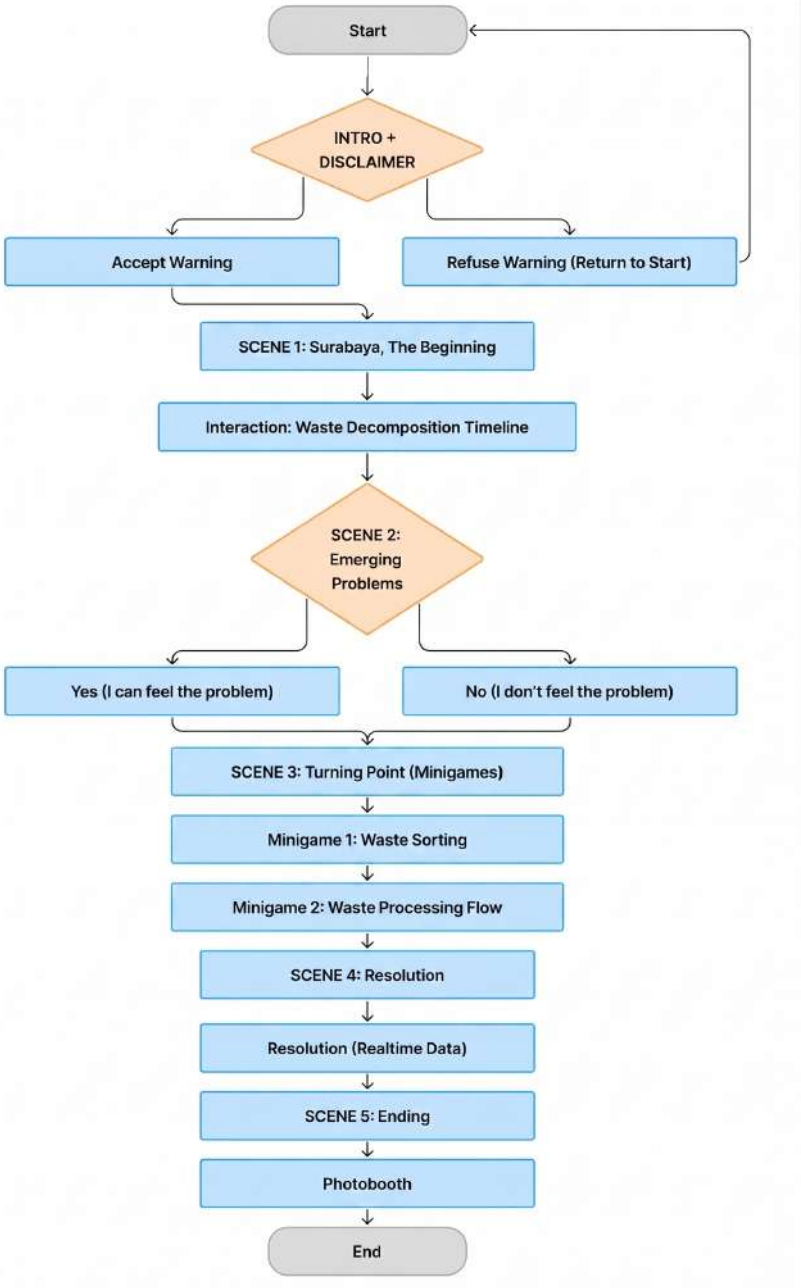

Figure 6 Story flow diagram illustrating scene progression and mapping of user interaction

4. Experiments

To validate the performance and robustness of our XR immersive room, we conducted a series of experiments to evaluate three critical aspects of the user experience: (1) latency and

responsiveness, (2) gesture recognition accuracy under different environmental conditions, and (3) voice recognition reliability in the presence of background noise. All experiments were conducted within the fully constructed 3x3-meter immersive booth, the real-world implementation of which is shown in Figures 7 and 8. This setup provided a controlled environment in which the synchronized operation of the projection mapping, gesture tracking, environmental actuators, and integrated audio-visual systems could be tested.

Figure 7 Real-world implementation of the XR immersive booth (left view)

Figure 8 Real-world implementation of the XR immersive booth (right view)

4.1 Experimental Setup

The immersive system integrates multiple hardware and software components, including Unity for scene rendering, MediaPipe and a Kinect V2 for gesture and body tracking, TouchDesigner and Resolume for real-time visual processing and projection mapping, and Arduino-based

controllers for the environmental effects (fans, lamps, fog, and mist). To rigorously evaluate the system's performance, we used the following equipment and parameters:

- A high-speed camera operating at 240 frames per second was used to capture user actions and system responses with a temporal resolution of 4.17 ms per frame.
- Gesture recognition was tested at two distinct user distances: 160 and 200 cm from the camera.
- Lighting conditions were varied to simulate different narrative scenarios, with "bright" scenes representing clear daytime conditions and "dark" scenes representing stormy or low-visibility environments.
- The accuracy of voice commands was evaluated under three different ambient noise levels: 30 dB (a quiet room), 60 dB (a normal conversation), and 80 dB (a noisy public space). The noise levels (30, 60, and 80 dB) were simulated using external audio sources and do not fully represent the combined noise generated by the system's actuators during operation.

These experiments were designed to assess the real-world robustness and reliability of the interactive XR storytelling experience under a range of challenging conditions.

Due to the time and operational constraints of the immersive system, the number of trials in this study is limited. However, the experiments were conducted in a controlled environment with consistent system behavior across trials. The results are intended to provide an initial technical evaluation rather than a generalized statistical performance analysis. Future work will expand the number of trials to improve statistical robustness.

4.2 Latency Measurement

Latency refers to the time elapsed between a user's action and the moment the system produces a visible or physical response. The high-speed camera allowed frame-accurate marking of three key timestamps:

- P1 (Input Frame): The frame in which the user's gesture is fully performed.
- P2 (Processing Start Frame): The frame in which the system updates the user interface (UI) or other internal visual elements in response to the input.
- P3 (Output Frame): The frame in which the final visible or physical response (e.g., a change in the projection or actuator activation) is observed.

The decomposition of latency into separate processing intervals follows the real-time pipeline timing model described by Gregory in Game Engine Architecture (Gregory, 2018) frame-based temporal analysis is widely applied to evaluate responsiveness in interactive graphical systems. In this model, each stage of the processing pipeline contributes a measurable delay, allowing the total perceived latency to be calculated by summing the corresponding frame differences.

Therefore, latency can be decomposed according to Equations (1) through (4):

$$T_{frame} = \frac{1000}{FPS} = \frac{1000}{240} = 4.17\ ms \tag{1}$$

$$T_{12} = (P_2 - P_1) \times 4.17\ ms \tag{2}$$

$$T_{23} = (P_3 - P_2) \times 4.17\ ms \tag{3}$$

$$T_{13} = (P_3 - P_1) \times 4.17\ ms \tag{4}$$

Table 2 presents the user-perceived latency, indicating the complete delay from user input to the final visual or physical output.

Table 2 Latency decomposition at 240 FPS

Trial	P1	P2	P3	Δ12	Δ23	Δ13	T12 (ms)	T23 (ms)	T13 (ms)
1	227	230	231	3	1	4	12.51	4.17	**16.68**
2	191	195	195	4	0	4	16.68	0.00	**16.68**
3	101	106	105	4	-1*	4	16.68	-4.17*	**16.68**
4	324	328	328	4	0	4	16.68	0.00	**16.68**
5	077	081	081	4	0	4	16.68	0.00	**16.68**
6	211	214	215	3	1	4	12.51	4.17	**16.68**

Note: Minor inconsistencies (P3 < P2 by 1 frame) are attributed to frame-level synchronization uncertainty in high-speed video capture, where frame alignment between processing and output events may result in ±1 frame discrepancy. This does not indicate actual negative latency but reflects measurement limitations.

The interpretation thresholds presented in Table 3 are adapted from psychophysical findings on multisensory temporal perception, particularly the temporal binding window described by Vroomen and Keetels, 2010 and Wallace et al., 2020:

Table 3 Latency interpretation

Visual Delay	Perceptual Effect
< 15–20 ms	Brain fuses events → *delay not perceived*
> 20–30 ms	Timing mismatch becomes noticeable
< 40–60 ms	Clearly unsynchronized

Since the XR system consistently achieves 16.68 ms, the latency falls below the perceptual threshold, ensuring a seamless interaction experience. The consistent latency values observed across trials indicate that the system operates within a frame-locked rendering pipeline, where the responses are synchronized with a fixed frame rate. Variations smaller than one frame (4.17 ms) cannot be captured due to the measurement system's temporal resolution.

4.3 Gesture recognition accuracy

Gesture recognition was tested by performing ten repeated gestures at two distances (160 and 200 cm) under bright and dark lighting scenarios. The gestures tested in this study include basic hand-raising, hand waving, and directional pointing, as well as interaction-based gestures such as reaching, grasping, and placing virtual objects within the XR environment. These gestures are mapped to the system's interaction triggers. Detection was marked as successful (v) or unsuccessful (x) depending on whether the MediaPipe correctly triggered the intended event. Gesture recognition is highly reliable in bright conditions, with nearly perfect detection across both distances (Table 4).

Table 4 Gesture accuracy in bright conditions

Lighting	Distance	Success Count	Accuracy
Bright	200 cm	10/10	100%
Bright	160 cm	9/10	90%

Accuracy decreased under dark scenes at 200 cm due to reduced hand contour contrast and depth noise (Table 5). At 160 cm, the performance remains strong.

Table 5 Gesture accuracy in dark conditions

Lighting	Distance	Success Count	Accuracy
Dark	200 cm	4/10	40%
Dark	160 cm	9/10	90%

4.4 Reliability of Voice Recognition

The voice commands were tested under quiet (30 dB), moderate (60 dB), and loud (80 dB) environments. The voice commands tested consisted of both simple predefined keywords, such as "yes," "no," and "let's start," as well as short natural phrases, such as "I want to go to Surabaya," which were used to control narrative branching and interaction flow. Ten trials were performed for each condition (Table 6).

Table 6 Voice recognition under noise

Condition	Noise Level	Success Count	Accuracy
Quiet	30 dB	10/10	100%
Noisy	60 dB	5/10	50%
Noisy	80 dB	3/10	30%

The performance is excellent in quiet conditions, partially reliable in moderate noise, and significantly impaired at 80 dB due to masking effects on speech frequency ranges.

The system supports a multimodal fallback mechanism to address interaction reliability issues under challenging conditions such as low lighting or high noise. When gesture recognition performance decreases due to insufficient lighting or increased distance, voice commands can be used as an alternative input modality. Conversely, gesture-based interaction remains available in noisy environments where voice recognition accuracy is reduced. This complementary interaction design ensures user interaction continuity across varying environmental conditions.

4.5 Overall Discussion

The experimental results confirm that the XR immersive room is a technically robust and responsive platform for multimodal interaction in an XR environment. The system's low latency, high gesture recognition accuracy under optimal conditions, and reliable voice command performance in quiet environments demonstrate its ability to provide a seamless and engaging user experience. This section discusses the implications of these findings, contextualizes them within the broader field of XR and environmental communication, and elaborates on this work's novelty and contribution.

The consistently low latency of 16.68 ms is a significant achievement, as it falls well below the human perception threshold. This ensures that the system's responses to user actions feel instantaneous and natural, which is crucial for maintaining a sense of presence and agency within the immersive environment. This level of responsiveness is on par with, or even exceeds, that of many commercial gaming engines and high-end VR systems, underscoring the integrated software and hardware pipeline's effectiveness.

The gesture recognition results highlight the importance of environmental factors in designing natural user interfaces. The high accuracy achieved in bright lighting conditions demonstrates the power of the MediaPipe framework for real-time hand and body tracking. The performance degradation observed in dark conditions and at greater distances is a known challenge in computer vision. However, our findings provide a clear operational guideline: maintaining an optimal interaction distance of approximately 160 cm can preserve high accuracy even in low-light scenes. This has important implications for the design of XR interaction systems, suggesting that gesture-based interactions should be designed in consideration of the environmental

context.

Similarly, the voice recognition results underscore the trade-off between NLI and environmental noise. The system's flawless performance in quiet settings confirms the viability of voice commands as an intuitive input modality. However, the sharp decline in accuracy at higher noise levels indicates that voice input may not be suitable for all narrative moments, particularly those involving loud sound effects such as storms or industrial noise. Therefore, a well-designed experience should leverage a multimodal approach, allowing the user to seamlessly switch between gesture and voice commands depending on the context. This adaptability is a key strength of the proposed integrated system.

Our work advances the state of the art in several key respects:

- Integrated Multimodal and Multisensory System: The technical feasibility of a complex system that synchronizes three-wall projection mapping, artificial intelligence (AI)-powered gesture recognition, voice commands, and a suite of physical actuators was demonstrated. This integration creates a level of immersion that is not possible with traditional screen-based media or VR alone.
- The physical feedback, such as the sensation of wind and rain, further enhances this embodied experience, making the abstract concepts of climate change more tangible and immediate.
- Comprehensive Performance Evaluation: We have provided a rigorous and quantitative evaluation of the performance of our system, including key metrics such as latency, gesture accuracy, and voice reliability. This data provides a valuable benchmark for future research and development in the field of XR system development and evaluation.

The discussion of our findings highlights both the technical achievements of our work and its broader implications for XR and XR system design and evaluation. This study provides a compelling model for the future development of multimodal XR systems by demonstrating the power of an integrated, multimodal, and multisensory approach.

This study focuses on the XR system's technical performance, including latency, gesture recognition, and voice interaction reliability. The evaluation of user-centered outcomes, such as engagement, learning effectiveness, or behavioral impact, is beyond the scope of this work and should be considered in future research.

5. Conclusions

This study demonstrated the successful design, development, and performance evaluation of an XR-based immersive room for climate change communication by integrating three-wall projection mapping, AI-based gesture recognition, voice interaction, and synchronized multisensory physical feedback. Experimental results confirmed that the system achieves low user-perceived latency (16.68 ms), high gesture recognition accuracy under optimal lighting and interaction distance, and reliable voice recognition performance in low-noise environments, ensuring responsive and immersive interaction. The primary contribution of this work lies in the integration of multimodal interaction and multisensory feedback within a physically grounded XR environment, supported by a rigorous quantitative performance evaluation. While performance degradation was observed under low-light and high-noise conditions, these findings provide important operational insights and design guidelines for future immersive installations. Future work will focus on improving system robustness through alternative sensing modalities, enhanced noise handling, and expanded multisensory channels. User-centered evaluation such as learning effectiveness and user engagement will be explored in future studies.

First, the experimental evaluation is based on a limited number of trials (6 for latency measurement and 10 for gesture and voice interaction tests), which may affect statistical reliability and should be interpreted as an initial technical validation rather than a statistically generalizable result. Second, the evaluation focuses only on technical performance metrics (latency, gesture accuracy, and voice recognition) and does not include user-centered assessments such as engagement or learning outcomes. Third, the experimental conditions may not fully

represent real-world operational scenarios, particularly regarding dynamic lighting and environmental noise generated by the system. These limitations will be addressed in future work through expanded experiments and user studies. Despite these limitations, the study provides a solid technical foundation for future research on multimodal XR systems.

Conflict of Interest

The authors declare no conflicts of interest.

References

Bailenson, J. (2018). *Experience on Demand: What Virtual Reality Is, How It Works, and What It Can Do*. W. W. Norton; Company.

Baroroh, D. K., & Agarwal, A. (2022). Immersive Technologies in Indonesia Faces 'New Normal' COVID-19. *International Journal of Technology*, *13*(3), 633–642. https://doi.org/10.14716/ijtech.v13i3.5220

Birch, S., Wernert, E. A., Danielson, K., Filippelli, G., & Hines, J. (2023). Visualizing the Causes and Impacts of Climate Change with XR Technologies. *PEARC 2023 - Computing for the Common Good: Practice and Experience in Advanced Research Computing*. https://doi.org/10.1145/3569951.3603630

Bustos-Lopez, G., Aguirre-Villalobos, E. R., & Meingast, K. (2024). XR for Transformable and Interactive Design. *Media and Communication*, *12*. https://doi.org/10.17645/mac.8603

Climate Division of the Meteorological and Geophysical Agency (BMKG). (2024). Indonesia's Climate and Air Quality Notes 2024.

Coaction Indonesia. (2024). Dynamics of Indonesia's Climate Agenda.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66–69. https://doi.org/10.1126/science.1167311

Ecstein, D., Kunzel, V., & Schafer, L. (2021). Global Climate Risk Index 2021.

Fernández Galeote, D., Legaki, N. Z., & Hamari, J. (2023). Climate Connected: An Immersive VR and PC Game for Climate Change Engagement. *CHI PLAY 2023 Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. https://doi.org/10.1145/3573382.3616053

Gregory, J. (2018). *Game Engine Architecture*. Taylor; Francis. http://taylorandfrancis.com

Hu, H., Tan, T., & Qian, Y. (2018). Generative adversarial networks based data augmentation for noise robust speech recognition. *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings*. https://doi.org/10.1109/ICASSP.2018.8462624

Huang, T. C., & Tseng, H. P. (2025). Extended Reality in Applied Sciences Education: A Systematic Review. *Applied Sciences*. https://doi.org/10.3390/app15074038

Ignacio, J., Dolmans, D., Scherpbier, A., Rethans, J. J., Chan, S., & Liaw, S. Y. (2015). Comparison of standardized patients with high-fidelity simulators for managing stress and improving performance in clinical deterioration: A mixed methods study. *Nurse Education Today*, *35*(12). https://doi.org/10.1016/j.nedt.2015.05.009

Ioannou, A., Bhagat, K. K., & Johnson-Glenberg, M. C. (2021). Guest Editorial: Learning Experience Design: Embodiment, Gesture, and Interactivity in XR. *Educational Technology and Society*, *24*(2).

Jamil, M., Hadiyanto, & Sanjaya, R. (2024). Optimizing 3D Assets and Character Modeling of the Mixed Reality Simulator in a Disaster Mitigation Learning Using Vertex Decimation and Depth-of-Field Algorithm. *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-981-99-6547-2_17

Jeong, J., Lee, C.-K., Hong, K., Yeom, J., & Lee, B. (2014). Projection-type dual-view three-dimensional display system based on integral imaging. *Applied Optics*, *53*(27). https://doi.org/10.1364/AO.53.000G12

Jerald, J. (2015). Human-Centered Interaction. In *The VR Book*. Morgan; Claypool Publishers. https://doi.org/10.1145/2792790.2792821

Kang, J., Diederich, M., Lindgren, R., & Junokas, M. (2021). Gesture Patterns and Learning in an Embodied XR Science Simulation. *Educational Technology and Society*, *24*(2).

Kim, H., & Shin, J. W. (2021). Target exaggeration for deep learning-based speech enhancement. *Digital Signal Processing*, *116*. https://doi.org/10.1016/j.dsp.2021.103109

Kizhevska, E., Ferreira-Brito, F., Guerreiro, T., & Luštrek, M. (2023). Using Virtual Reality to Elicit Empathy: A Narrative Review. *CEUR Workshop Proceedings*.

Kopcha, T. J., Ocak, C., & Qian, Y. (2021). Analyzing children's computational thinking through embodied interaction with technology: A multimodal perspective. *Educational Technology Research and Development*, *69*(4). https://doi.org/10.1007/s11423-020-09832-y

Lee, H. L., & Romero, J. R. (2023). IPCC, 2023: Sections. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Edited by H. Lee and J. Romero].

Lee-Cultura, S., & Giannakos, M. (2020). Embodied Interaction and Spatial Skills: A Systematic Review of Empirical Studies. *Interaction Design and Architecture(s)*. https://doi.org/10.1093/iwcomp/iwaa023

Liu, S., Xu, X., & Claypool, M. (2022). A Survey and Taxonomy of Latency Compensation Techniques for Network Computer Games. *ACM Computing Surveys*, *54*(11s). https://doi.org/10.1145/3519023

Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A Theoretical Research-Based Model of Learning in Immersive Virtual Reality. *Educational Psychology Review*. https://doi.org/10.1007/s10648-020-09586-2

Mikawa, Y., Sueishi, T., Watanabe, Y., & Ishikawa, M. (2018). VarioLight: Hybrid Dynamic Projection Mapping Using High-Speed Projector and Optical Axis Controller. *SIGGRAPH Asia 2018 Emerging Technologies*. https://doi.org/10.1145/3275476.3275481

Miranto, C., Firmanda, A., Rante, H., Sukaridhoto, S., Zainuddin, M. A., & Rahman, H. (2025). Performance analysis of 3D assets in virtual reality simulations for climate change: A case study in sustainable energy systems. *Bulletin of Electrical Engineering and Informatics*, *14*(5), 3659–3670. https://doi.org/10.11591/eei.v14i5.9532

Mufarroha, F. A., & Utaminingrum, F. (2017). Hand gesture recognition using adaptive network based fuzzy inference system and K-nearest neighbor. *International Journal of Technology*, *8*(3), 559–567. https://doi.org/10.14716/ijtech.v8i3.3146

Nossier, S. A., Wall, J., Moniri, M., Glackin, C., & Cannings, N. (2021). An experimental analysis of deep learning architectures for supervised speech enhancement. *Electronics*, *10*(1). https://doi.org/10.3390/electronics10010017

Pham, H. H., Salmane, H., Khoudour, L., Crouzil, A., Velastin, S. A., & Zegers, P. (2020). A unified deep framework for joint 3D pose estimation and action recognition from a single RGB camera. *Sensors*, *20*(7). https://doi.org/10.3390/s20071825

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers and Education*, *147*. https://doi.org/10.1016/j.compedu.2019.103778

Rante, H., Achmad, Z., Suaib, N., Prasetya, S., Avianto, T., Hermanu, A., Alfarezi, F., & Wijaya, R. (2024). Preliminary Development of Vircadia Virtual Reality Platform for Monitoring Water Quality Powered by Solar Panels. http://www.joiv.org/index.php/joiv

Sherman, W. R., & Craig, A. B. (2018). *Understanding Virtual Reality: Interface, Application, and Design* (2nd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2013-0-18583-2

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*, 74. https://doi.org/10.3389/frobt.2016.00074

Spence, C., & Gallace, A. (2011). Multisensory design: Reaching out to touch the consumer. *Psychology and Marketing*, *28*(3), 267–308. https://doi.org/10.1002/mar.20392

Treal, T., Jackson, P. L., Jeuvrey, J., Vignais, N., & Meugnot, A. (2021). Natural human postural oscillations enhance the empathic response to a facial pain expression in a virtual character. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-91710-5

Viswakumar, A., Rajagopalan, V., Ray, T., Gottipati, P., & Parimi, C. (2022). Development of a Robust, Simple, and Affordable Human Gait Analysis System Using Bottom-Up Pose Estimation With a Smartphone Camera. *Frontiers in Physiology*, *12*. https://doi.org/10.3389/fphys.2021.784865

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, and Psychophysics*, *72*(4), 871–884. https://doi.org/10.3758/APP.72.4.871

Wakunami, K., Hsieh, P.-Y., Oi, R., Senoh, T., Sasaki, H., Ichihashi, Y., Okui, M., Huang, Y.-P., & Yamamoto, K. (2016). Projection-type see-through holographic three-dimensional display. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms12954

Wallace, M. T., Woynaroski, T. G., & Stevenson, R. A. (2020). Multisensory integration as a window into orderly and disrupted cognition and communication. *Annual Review of Psychology*, *71*, 193–219. https://doi.org/10.1146/annurev-psych-010419-051112

Zhang, X., Yang, H., Liu, C., Tong, Q., Xiu, A., Kong, L., Dan, M., Gao, C., Gao, M., Che, H., Wang, X., & Wu, G. (2025). XR-based Interactive Visualization Platform for Real-time Exploring Dynamic Earth Science Data [SSRN Preprint]. https://ssrn.com/abstract=4769475