# Realizing Seamless Connection Between Real and Virtual Spaces via Adaptive Virtual Reality: Objective Description of Cognitive Discrepancies Between These Spaces

Katsuko T. Nakahira
*Nagaoka University of Technology*
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

Taichi Nakagawa
*Nagaoka University of Technology*
Nagaoka, Niigata, Japan
Email: s223308@stn.nagaokaut.ac.jp

Thalpe Liyanage Amila Nirosh Chandrasiri
*Nagaoka University of Technology*
Nagaoka, Niigata, Japan
Email: s245065@stn.nagaokaut.ac.jp

Nobuyuki Ogawa
*National Institute of Technology(KOSEN), Gifu College*
Motosu, Gifu, Japan
Email: ogawa@gifu-nct.ac.jp

Kuniaki Yajima
*National Institute of Technology(KOSEN), Sendai College*
Sendai, Miyagi, Japan
Email: yajima@sendai-nct.ac.jp

Muneo Kitajima
*Nagaoka University of Technology*
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

***Abstract*— Although the implementation of "Adaptive Virtual Reality" is becoming feasible, understanding the main effects of its realization on users based on cognitive models is essential. Here, we first described a model of the flow of information obtained by actual human perception through avatars in virtual reality (VR) and the resulting human reactions, and confirm the validity of the user models proposed thus far. We also considered the degree of immersion predicted due to the integration of multimodal information. The cognitive processes of VR experiences are largely categorized into "perception and recognition of information (attention, memory, and decision making)" and "perception-based physical actions and interactions with VR objects". Based from this, we describe a cognitive model of VR experiences. In addition, as examples of the discrepancies in sensory perception experienced in real/VR spaces, we briefly describe the phenomena that occur in communication. We describe the cognitive models for these phenomena and qualitatively consider the degree to which sensory information obtained from the real/VR space affects the degree of chunks activation. The intensity of human sense is expressed as a logarithm according to Weber-Fechner's Law, suggesting that human senses can distinguish differences even with weak sensory information. We argue that the "slightly different from the real world" sense felt in VR content is caused by such slight differences in sensory information. Overall, we advance the cognitive understanding of the immersive experience particularly in the VR space, and qualitatively describe the possibility of designing highly immersive VR content which are adapted to each individual.**

***Keywords*— sensory perception; cognitive model; virtual reality; experience.**

## I. Introduction

This paper is based on the previous work originally presented in AIVR2024 [1]. The following changes were made: (1) a restructuring of the model diagram, (2) the addition of preliminary experimental results, and (3) the addition of corresponding discussions.

A concept called "Adaptive Virtual Reality (Adaptive VR)" has been discussed in recent years. Baker and Fairclough [2] described it as follows: Adaptive VR monitors human behavior, psychophysiology, and neurophysiology to create a real-time model of the user. This quantification is used to infer the emotional state of individual users and induce adaptive changes within the virtual environment during runtime. Therefore, the authors argued that the efficacy of the emotional experience can be increased by modeling individual differences in the way users interact within a particular virtual environment as a system.

Several methods exist for inferring emotional states. The most classical approach, following Russell's circular model [3][4], involves measuring a person's valence and arousal states to predict their emotions. These are often predicted through measurements such as pupil response or heart rate. Specifically, regarding visual behavior, Bao et al. [5] proposed a method to recognize learners' emotional states during distance learning, suggesting emotion recognition techniques for relatively simple emotions such as interest, happiness, confusion, and boredom. Furthermore, Sun et al. [6] suggested that arousal related to cognitive effort interacted informationally with luminance, and that the strongest pupil response due to arousal occurred at luminances below 37 cd/m$^2$.

Given this background, the implementation of Adaptive VR is becoming feasible. However, the main effects of its implementation on users should be understood based on a cognitive model. Studies have mainly focused on bottom-up content design with an awareness of Adaptive VR. However, it is difficult for empirical developments to provide effects that create a new phenomena. Hence, not only a bottom-up but also a top-down approach is necessary.

On the other hand, with the growth in Virtual Reality (VR)

goggles and the low cost of equipment for shooting omnidirectional video, VR content has attracted substantial attention. In addition to games, a wide range of VR contents have been developed, including omnidirectional video playback, education, sightseeing, property previews, and shopping. VR systems that enable these contents to be viewed are also growing rapidly. For example, the following innovations have emerged in content design. VR systems using Head-Mounted Displays (HMDs) sold to general consumers cover the user's field of vision; thus, the user cannot see their own body. Therefore, VR systems using HMDs typically display a virtual body drawn from the user's first-person perspective. A mechanism for realizing the user's first-person perspective is the implementation of avatars. The effects of avatars have been described by researchers. Steed et al. [7] suggested that the use of avatars that follow the user's movements can reduce the cognitive load of certain tasks in the VR space. People around the world have been using VR social networking services, such as VRChat, where users enjoy interacting with other users using avatars that they have selected and edited to their liking. This shows that avatars are a means of self-expression in VR communication.

There are many research approaches to VR contents and systems, including research from the perspective of Human Computer Interaction (HCI), research on the relation between VR and working memory (WM), research on the differences in sensory perception between the real world and VR, and research on Adaptive VR that incorporates individual adaptability into VR contents.

Among the studies from the perspective of HCI, Mousavi et al. [8] integrate Emotion Recognition (ER) and VR to provide an immersive and flexible environment in VR. This integration can advance HCI by allowing the Virtual Environment (VE) to adapt to the user's emotional state.

According to Batra et al. [9], the following requirements for VR are listed: First, the primary component called "visualization" enables human-machine interaction to approximate real life; Additionally, VR requires removing the barrier between the real world and the virtual world. Through these means, a series of simulation technologies must generate artificial tactile, auditory, olfactory, and sensory experiences grounded in reality. For this simulation, it is crucial to capture human cognitive characteristics multi-modally.

As a stepping stone to this goal, we do the following in this study. We describe a model of the flow of information obtained by actual human perception through avatars in VR and the resulting human reactions, and confirm the validity of the user models proposed so far. The degree of immersion predicted because of the integration of multi-sensory information is also discussed. Understanding the role of multi-sensory information can enable us to design VR contents for individual users and how we can control sensory perception.

The remainder of this article is organized as follows. We describe the sense-perception cognitive model on VR in Section II. Section III describes experiments conducted to investigate the behavioral characteristics of participants' information acquisition and attention direction within VR spaces, as well as the relationship between the degree of recall and the variety of perceptual information quantity and quality within the VR space. Section IV argues the relationship between the variety of perceptual information quality combinations and cognitive load.

## II. Description of the Cognitive Model for Sensory in VR

In general, physical information in the VR space is represented as follows. Objects in the VR space (VR objects) are represented by computer graphics, and their behavior is based on a program previously written to interact with the environment and other objects. The sound in the VR space is provided by artificially preparing audio data that is predicted in advance to be uttered in the space, and is played continuously in a background music-like manner, or by using a sound engine controlled by the user. Specifically, in the latter case, it can be attached to a VR object and played when certain conditions are met. Comprised of these elements, all human activities and virtual experiences in the VR space are performed by using the avatar as one's own body. The avatar's movement is performed by tracking the user's real-world body movements. Tracking methods include three-point tracking, which consists of an HMD and two hand controllers, and full-body tracking, which uses motion capture and a tracking suit.

Consequently, the human experience in the VR space differs slightly from perception and cognition in the real world, and can be said to be the result of the interaction between avatar and VR objects, as well as the perception of the accompanying environment such as sound linked to these objects. Considering this, the model of human perception, cognition, and behavior in the VR space should be described with an awareness of the various interactions in the VR space with those in the real world.

Based on the above, the integration process of information perceived in both the real world and VR is shown in Figure 1. The concept is as follows.

### *A. Transformation of Perceptual Information Provided by Objects*

The perceptual information of an object existing in real space is expressed through the five senses—visual, auditory, touch (somatosensory), smell, and taste—in a form adapted to our sensory organs. Perceptual information directly received by humans from the real world is received without attenuation beyond the capabilities of the individual's sensory organs. However, in VR, analog-to-digital conversion is applied to the perceptual information possessed by real-world objects. Consequently, this information exists within the VR space in a form where some information is missing. This means that in Figure 1, the chunk $c_j$(transferred) provided by the object transferred to the virtual world—resulting from the analog/digital conversion of the analog perceptual information in chunk $c_j$(real) provided by the real-world object—exists in a form where information is missing.

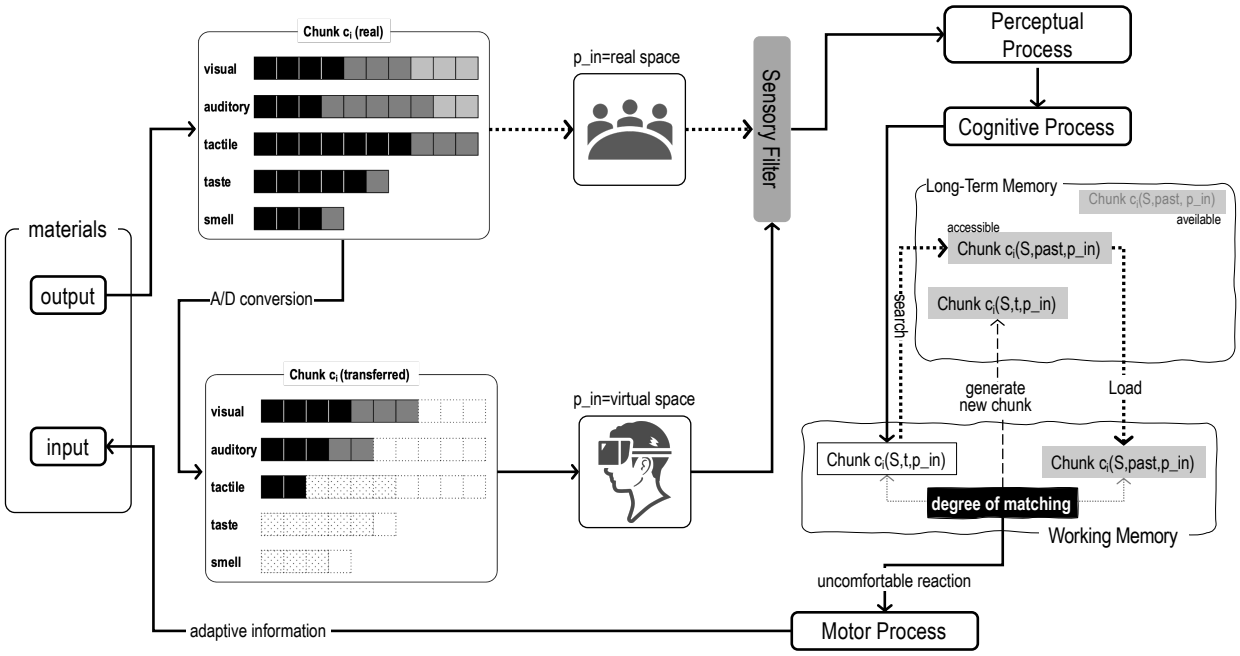



Figure 1. The information flow received by the senses in different environments.

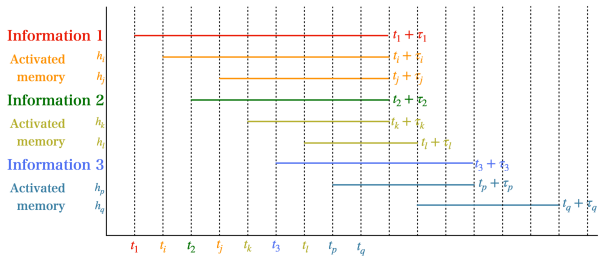



Figure 2. Staying timeline of sensory information stored in working memory and information invoked from long-term memory.

### B. Perception of Information Provided by Objects in Virtual World

In general VR experiences using current HMDs, visual, auditory, and somatosensory information are used as perceptual information. The VR experience begins when the user puts on the HMD and views the images displayed on the lenses; by moving their head while wearing the HMD, the user can perceive the virtual space in the same way as they perceive the real world. Auditory information is output from the HMD's built-in or external speakers, and audio is played in response to the behavior of VR objects. The somatosensory information is used to make operations in the VR space clearer by vibrating the controllers in both hands to generate tactile feedback when operating the User Interface (UI) in the VR space or selecting VR objects.

### C. Cognition of Information in Virtual World

The perception of information in virtual world is fundamentally no different from that in the real world. However, even when perceiving the same object, the provided information from the object is incomplete compared to that in the real world. The process differs in that it involves cross-referencing with similar memories. Thus, we describe the sequence of events as below.

*1) Attention:* Perceptual information moves to the sensory register, and then only the information to which the user's attention is directed passes through the selective filter and into the WM. Here, each sensory information does not completely enter the WM at the same time, but one piece of information passes through per processing.

*2) Memory:* If the sensory information obtained in the VR space is similar to that obtained in the real world, the user perceives the VR space as if it were a real space. In addition, based on the information in the Long-Term Memory (LTM), the user anticipates and expects the response of objects in the VR space to his or her actions, and engagement is generated.

*3) Decision:* Based on the perceptual information, the next action is determined. Here, when actions on a VR object are performed via a controller, the actions in the real world are converted into the corresponding controller operations.

### D. Body Movement Based on Perception

The operator (actual body) moves, and the avatar in the VR space moves in response to the movement. There are two methods for incorporating human motion into VR:

- Image sensing by the camera attached to the HMD: Basic UI operations (clicking and screen scrolling) and grasping VR objects (realized by holding something with a hand gesture) are possible. The high degree of synchronization between the actual hand and the avatar's hand motion is an advantage of this method. Conversely,

TABLE I. The difference between past or current input information and situations, and the *degree of matching*

| difference of situation | difference of input information almost never | little | greatly |
|---|---|---|---|
| almost never | − | + | ++ |
| little | + | ++ | ++ |
| greatly | ++ | ++ | +++ |

precise manipulation, movements large enough to cause both hands to move out of the camera's field of view, and very fast hand movements are weaknesses.

- Yaw, pitch, roll + relative position by controller:
  The accurate tracking of position, posture, and motion information by sensors is possible, and the sense of actual body motion is directly reflected during the operation, resulting in a high sense of immersion. However, if the reflection of body motion by the HMD is not synchronized with the actual body motion, it may cause a sense of discomfort and reduce the immersiveness of the VR experience.

*1) Interaction with VR Objects:* VR objects not only appear to be three-dimensional, but can also be actually manipulated. Examples include playing a musical instrument or a push-button switch. Here, the immersiveness of the VR experience can be enhanced by providing not only a visual 3D effect, but also contextual information that one's actions affect the VR object.

### *E. Integration of Information Obtained in the Virtual World and Past Experiences*

Based on Figures 1 and Table I, we consider the perception of a phenomenon in the real $(R)$ or virtual $(V)$ space as follows. The chunk $C_j$ stored in the LTM is constructed from the information group $I_i^{env}(t)$ obtained from sensory organ $i$ $(1 \leq i \leq 5)$ in the past. Here, $i$ refers to the five sensory organs possessed by a person. Each $I_i^{env}(t)$ passes through the attention filter $F_i^{env}(t)$ via the sensory register. And at time $t$, only the information obtained from a specific sensory organ passes through. $C_j$ contains the information obtained from each sensory organ as a set $I(t)$ and is denoted as $C_j(I(t))$. Here, $I(t)$ is represented as follows:

$$I(t) = \{\boldsymbol{I} \mid F_i^{env}(t) I_i^{env}(t),\ 1 \leq i \leq 5\}.$$

The information that has passed through the attention filter is stored in the WM for a specific time $\tau_k$, and a set of information $I(t)$ is sent to the LTM at the same time or with a time lag. In the LTM, $C_j(I(t))$ is matched with $C_j(I(t))$ based on the information in $I(t)$, and the closest or matching $C_j(I(t))$ is used as knowledge. The used knowledge is overwritten in the LTM through the WM in the form that the information in $I(t)$ is enhanced. Here, we target two sensory organs – visual and auditory. We consider how the information flows through these three types of sensory organs in turn.

Suppose that at a certain time, a specific amount of information $I_i^{env}(t)$ $(1 \leq i \leq 5)$ is received from the external environment. $I_i^{env}(t)$ correspond to Information $N$ in Figure 2. Information $N$ simultaneously activates several chunks. Although the degree of chunk activation varies, $F_i^{env}(t) I_i^{env}(t)$ is integrated into a single piece of information and sent to the LTM. This difference, the integrated information $I^{syn}(t)$, can be expressed using the integration operator $G$ as follows. However, since $G$ depends on the individual, it does not take a unique form.

$$I^{syn}(t) = G(i, j, F_i^{env}(t) I_i^{env}(t), C_j(I(t)))$$

For the sake of simplicity, we simply add the amount of information and the degree of chunk activation as follows.

$$G^{env}(t) = \sum_{j}^{m} \sum_{i}^{n} F_i^{env}(t) I_i^{env}(t) C_j(I(t)) \qquad (1)$$

## III. Information Acquisition and Attention Direction in Metaverse Space

If the cognitive framework described in Section II is correct, differences in sensory perception should be observed between the real and virtual worlds. The following are examples of what these differences in sensory perception might cause:

- Differences in memory quantity/quality:
  Information easily memorized in the real world may be difficult to contextualize in the virtual world, or vice versa.
- Differences in reaction:
  In the real world, even minor changes can trigger significant reactions. Conversely, in the virtual world, reactions may be difficult to elicit without substantial changes. Or the opposite may occur.

We consider the differences in sensory perception is caused by depending on the *degree of matching* within working memory, described in Figure 1, namely the value of $I^{syn}(t)$. We also consider $I^{syn}(t)$ as the cognitive load incurred during the integration of perceived similar information of past and current, the greater the divergence between the two pieces of information, the higher the load required to generate $I^{syn}(t)$. To realize this divergence, this research sets the number of objects in the virtual world as the excess or deficiency of integration targets for visual information, and the audio quality of explanatory narration for specific objects as the difficulty level of integration targets for auditory information. We confirm the possibility of measuring the load on information integration through the combination of these two factors. Based on the above, we propose the following hypotheses regarding the quality of visual and auditory information:

- Hypothesis 1: Different combinations of quality result in different cognitive load for information integration, and an optimal combination exists that provides the least load.
- Hypothesis 2: Consequently, differences are observed in the memory of information perceived during activities in a virtual world.
- Hypothesis 3: When the combination provides optimal load, the perceived cognitive load is closer to that experienced in the real world, leading to a sense of immersion.

In this study, we design experiments to gain insights into Hypotheses 1 and 2 and verify their validity.

To verify this, one method involves recreating real-world objects within a virtual world with appropriate explanations, then conducting visual behavior analysis and memory depth analysis using variables such as fixation time on the object and depth of memory for the explanation. The experimental method for this is described below.

### A. The Configuration of the Target Virtual World and Experimental Conditions

To conduct experiments testing hypotheses, it is necessary to construct a virtual world and set the quality and quantity of objects. In this study, to perform trend analysis for the hypothesis, we designed the space as follows as a preliminary experiment.

The virtual world is structured with sightseeing in mind. Consequently, activities within the virtual world are as follows:

- Free exploration (primarily focused on acquiring visual information)
- Discovery of distinctive objects (discovery through visual information). We focuse on architectural structures.
- Receiving supplementary knowledge through explanatory narration on architectural styles and structures (learning involving auditory information).

The primary object is content featuring the construction of shrines—people often seen but rarely understood in detail in real world. A screenshot of the VR space used in the experiment and the intensity condition waveform of the sound source are shown in Figure 3.

The virtual world space was constructed using Unity. Within the VR space, a shrine model and an explanatory audio track about the shrine's construction were placed. For the shrine's 3D model, a commercially available standard architectural style was used. However, since the focus was specifically on learning architectural styles, decorative items that should be placed inside were excluded. The commentary audio is designed to play when the participant pushes the speaker icon. The content consists of standard explanatory text combined and read aloud by a automated voice, with only amplitude adjustments made. However, for the lower quality setting, a lowpass filter is applied to achieve telephone-like audio quality. Three explanatory audio clips were prepared for each shrine, with their auditory quality set to three varieties. For the building exteriors, objects related to shrines—such as trees, *torii* gates, and *chozuya* purification fountains—were placed to enhance visual and contextual information, making it closer to the real thing.

In this virtual world, each participant freely moves through the space, exploring both inside and outside the shrine. They memorize the space sometimes using only visual information, sometimes only auditory information, and sometimes by integrating both types of information. Therefore, a recall test for the memorized content can serve as an indicator of how information acquired in various ways is expressed.

For visual and auditory information, their quality was set according to the conditions shown in Table II. For visual information variations, three patterns were prepared:

- A simple plane with only a shrine (condition $L$),
- A simple plane with a shrine and related objects (trees, *torii* gate, *chozuya*) which represents condition $N$, and
- A shrine located within a forest, accompanied by a *torii* gate and *chozuya* which represents condition $R$.

For auditory information variations, we prepared three patterns:

- poor audio quality for the explanatory narration (with lowpass filter for normal sound) which represents condition $L$,
- standard commentary audio (default automated voice and no customize) which represents condition $N$, and
- consistently high volume (amplifying power) which represents condition $R$.

### B. Experimental Procedure

The experiment was conducted as follows. The overall flow is shown in Figure 4. The HMD used for the experiment was the Meta Quest 2, and the Tobii Pro Glasses 3 were used for eye tracking. Data acquired with the Tobii Pro Glasses 3 was processed in Tobii Pro Labo to identify saccades, fixations, and obtain gaze point coordinates. Furthermore, to mitigate VR sickness, teleportation was adopted as the method of movement within the VR space. Before the experiment began, subjects were asked to answer questions regarding:

- Previous VR experience,
- Prior knowledge of architecture, and
- Prior knowledge of shrine construction.

Additionally, subjects were asked to answer several questions before the experiment began, including their knowledge of shrines, learning experiences, and whether they had recently visited a shrine. Sessions were conducted for each visual condition, and at the end of viewing each session,

- Did you feel as if you were actually present there?
- Did you feel the VR world was so realistic that you forgot the outside world?
- Did you feel like you were watching a video, or did you feel like you were actually in the space?

They were asked to rate their responses on an eight-point scale.

After the questionnaire, participants were given a tutorial and then experienced VR content with sensory information appropriately altered. To measure participants' gaze information, they wore an HMD over an eye tracker while experiencing VR content. Participants experienced one session consisting of an audio playback task with three different auditory conditions under the same visual condition, completing a total of three sessions for different visual conditions. At the end of each session, they answered questions about the content. After completing the three sessions, a recall test was conducted.

In the recall test, a free-response section was included where subjects were asked to write about what they remembered

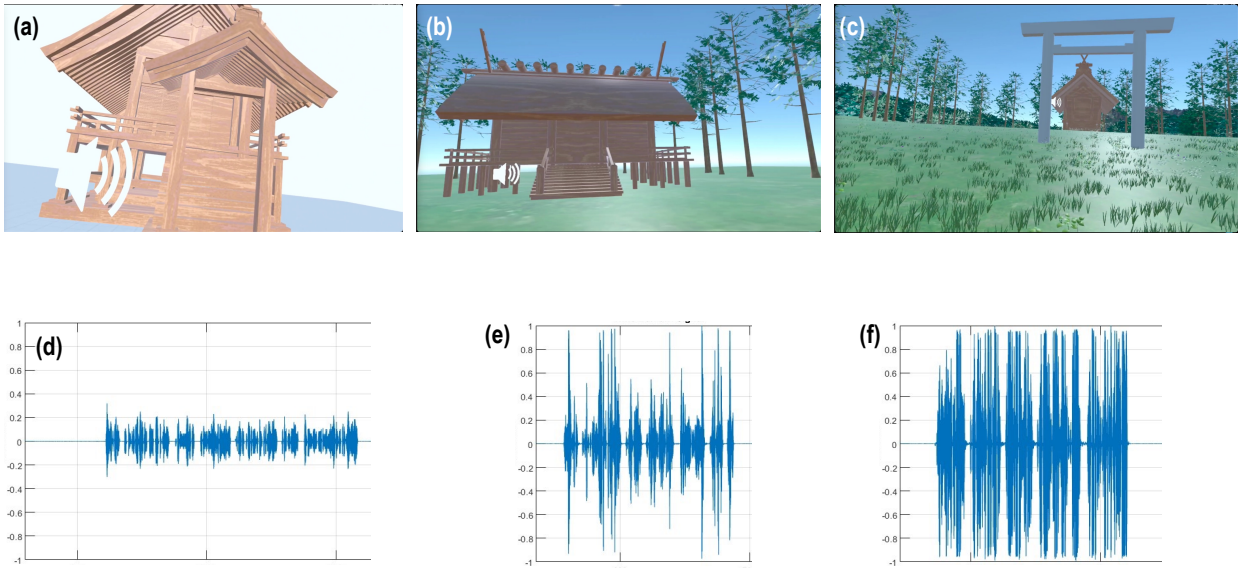



Figure 3. VR content presented to participants. Visual conditions $(V)$ are: (a) less (weak) stimulus $(L)$, (b) normal stimulus $(N)$, (c) rich (strong) stimulus $(R)$. Audio conditions $(A)$ are: (d) low quality $(L)$, (e) normal quality (baseline), $(N)$), and (f) rich quality $(R)$.

TABLE II. VR content presentation stimulus change patterns.

| | condition1 : weak stimuli $L$ | condition2 : normal stimuli $N$ | condition3 : strong stimuli $R$ |
|---|---|---|---|
| visual $V$ | Only the shrine is placed on the flat surface. | Arrange a shrine, 47 trees, a *torii* gate, and ground texture on a flat surface. | Arrange a shrine, 75 trees, three-dimensional terrain, grass, *torii* gates, a *chozuya*, and background on a flat surface. |
| auditory $A$ | A muffled sound quality like over the phone, which apply a low-pass filter to achive the situation. | Unadjusted audio. | The volume is excessively loud, which change amplitude to achive the situation. |

about the content and what they felt during the experience. The questions were:

- Please freely write down what you remember about the content.
- Please freely describe what you felt while experiencing the VR content.

The free-response session had no time limit, and subjects were permitted to write down as much information as they could recall. A 3-minute break followed the free-response session, during which subjects spent time with the HMD removed.

### C. Experimental Conditions

The experimental conditions for visual and auditory information are as shown in Table II, with three variations prepared for each. For variations in visual information, three patterns were prepared: a simple plane with only a shrine, a simple plane with a shrine and shrine-related objects (trees, *torii* gate, *chozuya*), and a shrine located within a forest, accompanied by a *torii* gate and *chozuya*. For variations in auditory information, three patterns were prepared: a case with poor audio quality for the explanatory narration, a case with normal audio quality for the explanatory narration, and a case with high volume for the explanatory narration. Examples of scenes presenting each condition are shown in Figure 3.

TABLE III. Average fixation time for each subject across combinations of visual variety and auditory variety.

| | | | visual $(V)$ | | |
|---|---|---|---|---|---|
| | | | $R$ | $N$ | $L$ |
| auditory$(A)$ | $R$ | $S_a$ | 213.5 | 183.1 | 225.5 |
| | | $S_b$ | 159.1 | 124.7 | 172.3 |
| | | $S_c$ | 144.6 | 154.8 | |
| | $N$ | $S_a$ | **210.0** | *198.5* | *190.2* |
| | | $S_b$ | **198.5** | | *164.6* |
| | | $S_c$ | **190.2** | *138.1* | *101.1* |
| | $L$ | $S_a$ | **190.2** | *208.8* | **193.5** |
| | | $S_b$ | **208.8** | *173.0* | **177.6** |
| | | $S_c$ | **193.5** | *115.6* | **144.9** |

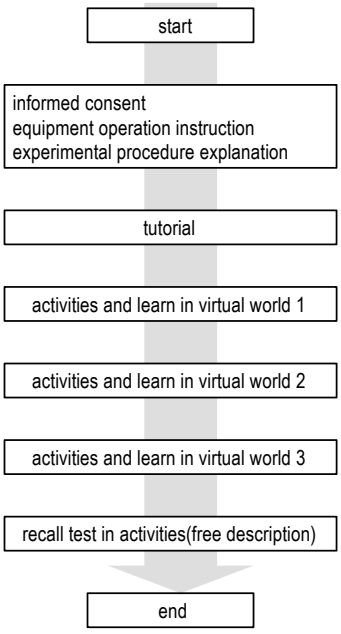



Figure 4. Experimental design.

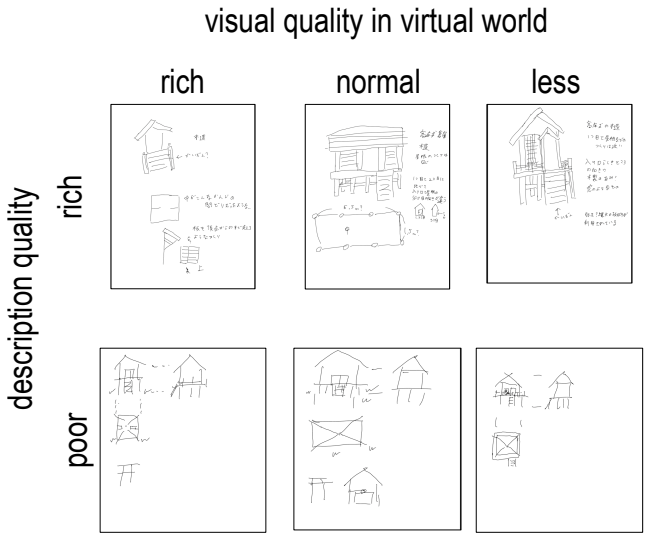



Figure 5. The result of free description in recall test.

### *D. Results (1) : Effect of Fixation Time and Visual/Auditory Quality*

We focused on three participants $S_a$, $S_b$, $S_c$ whose recall test results were particularly distinctive as a case study to confirm the validity of this experiment. Among them, $S_a$ and $S_b$ both had almost no interest in architecture, while $S_c$ was a graduate of an architecture department. The characteristics of their respective descriptions were as follows.

- $S_a$ described information obtained within the virtual world simply but faithfully, as seen in the rich class of Figure 5,
- $S_b$ described information obtained within the virtual world solely through deformed drawings (poor class of Figure 5), and
- $S_c$ described information obtained within the virtual world using both simple text and detailed drawings.

Table III shows the average fixation time for each participant across visual and auditory condition combinations. The two blank entries indicate missing data due to malfunction of the auditory information presentation program. Although the values varied considerably across subjects, several trends were observed. When examining the (visual, auditory) conditions, fixation times were generally longer for $(R,\ N)$, $(R,\ L)$, and $(L,\ L)$. Furthermore, for $(N,\ R)$, fixation times tended to be shorter overall.

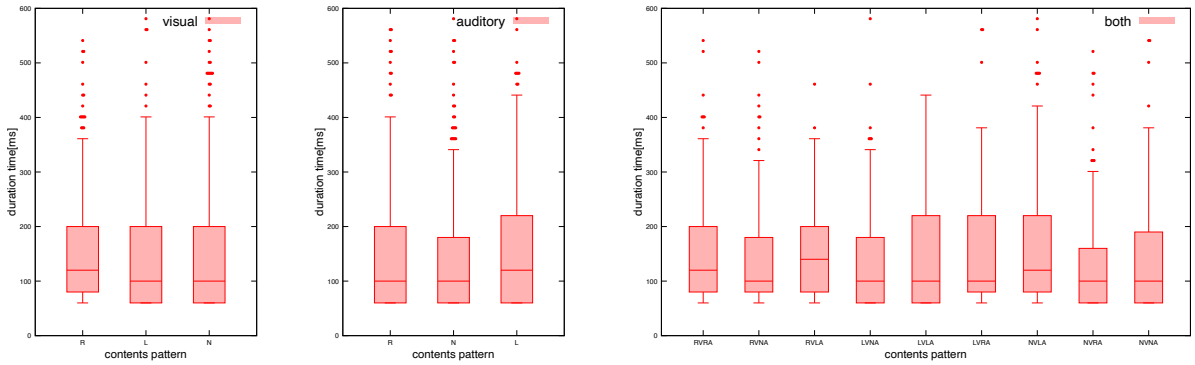
Figure 6 shows a boxplot of fixation time for combinations of manipulated perceptual information. In the figure, $V$ denotes visual, $A$ denotes auditory, and $L,\ N,\ R$ denote the quality of each perceptual information as low/normal/high. Figure 6(a) is a boxplot of fixation time for shots grouped by visual quality on the left and auditory quality on the right. Figure 6(b) is a boxplot of fixation time for shots segmented by visual quality-auditory quality combinations.

Figure 6 shows as follows: First, the distribution of fixation times for visual $V$ was slightly longer for the $R$ condition, with a median of approximately 120 ms; The distribution of fixation times for auditory $A$ was slightly longer for the $L$ condition, with a median of approximately 120 ms. These results suggest that fixation times are longer when the perceptual information condition is $RVLA$. Indeed, examining Figure 6(b) and confirming the median for $RVLA$ in Figure 6(a), it was approximately 150 ms, a value clearly larger than the medians for the $V$ or $A$ groups alone. Furthermore, both $RVRA$ and $NVLA$ were around 120 ms, larger than the median for the $V$ or $A$ single-stimulus groups. Considering the above, it is reasonable to conclude that for the perceptual conditions $RVLA$, $RVRA$, and $NVLA$ in the shot, information integration takes longer compared to a single perceptual condition.

To verify this, we conducted an analysis of variance (ANOVA) on the distribution of fixation times for the independent variables visual variety and auditory variety. First, a two-way ANOVA on the fixation times for three participants revealed a weak tendency toward an interaction effect $(F(2, 1201) = 2.258,\ p = 0.06)$. A weak tendency was also observed for the main effect of auditory variety $(F(2, 1201) = 2.414,\ p = 0.09)$. Next, focusing specifically on participant $S_c$ who drew with particular precision in the recall test, a two-way ANOVA was performed using only $S_c$'s data. No interaction was confirmed $(F(2, 402) = 1.65,\ p = 0.177)$, a significant main effect was observed for auditory variety $(F(2, 402) = 4.081,\ p = 0.018)$. Furthermore, multiple comparisons revealed a significant tendency for (visual variety, auditory variety) $= (N, R)$ and $(L, N)$ $(p = 0.087)$.

### *E. Result (3): Distribution of Shot Duration for Visual/Auditory Variety*

Next, we discuss the pupil diameter changes during auditory information listening for each of $S_a$, $S_b$, and $S_c$. Table IV




(a) For visual and auditory stimuli respectively, the left side shows visual stimuli grouped together, and the right side shows auditory stimuli grouped together.

(b) Combination of visual/auditory.

Figure 6. Boxplot of fixation time for combinations of manipulated perceptual information. In the figure, $V$ denotes visual, $A$ denotes auditory, and $L, N, R$ denote the quality of each perceptual information as low/normal/high.

TABLE IV. Auditory Experimental Conditions Results and Measurements in Two-Second Time Period. The under number of each visual condition represents lightness for each condition which calculated by Matlab.

| | | Visual Condition | | |
|---|---|---|---|---|
| | | $L(166.98)$ | $N(129.14)$ | $R(122.99)$ |
| Auditory Condition | $L$ | 0.196 | 0.341 | 0.172 |
| | $N$ | 0.163 | 0.261 | 0.125 |
| | $R$ | 0.172 | 0.291 | 0.148 |

shows the APD during auditory information listening for each visual variety.

The pupil diameter change was calculated as follows. First, it is necessary to determine the baseline $r_b$ for the pupil diameter acquired simultaneously during gaze measurement, which is obtained for each gaze ($\sim$20 [ms]). $r_b$ was set as the average pupil diameter from 500 [ms] before the time $t_{sa}$ of entering the actual space after the tutorial in the experiment until $t_{sa}$. Using the pupil diameter $r_p(t)$ measured at time $t$, $\bar{r_p}$ is calculated as follows.

$$\bar{r_p} = \frac{1}{n_{r_p}} \sum_{i}^{n_{r_p}} (r_p(t_i) - r_b),$$

where $n_{r_p}$ represents the number of data observed $r(t)$ from $t_{sa}$ to $t_{sa} + \Delta t$. We set $\Delta t$ to 2000 [ms], which is considered sufficient for the response to the presented stimuli to settle.

In Table IV, when visual condition is $N$, $\bar{r_p}$ shows a larger value compared to the others. Particularly in $(N,\ L)$, $\bar{r_p}$ shows a large value. For $(N,\ L)$, the value is nearly twice of that of the other $\bar{r_p}$'s. Furthermore, comparing $\bar{r_p}$ in auditory condition, all variety of visual condition have large $\bar{r_p}$ when auditory condition $L$.

### F. Result (2): Distribution of Shot Duration for Visual/Auditory Variety

Next, we perform a preliminary analysis of fixation behavior among participants. For participants' fixation behavior, we classified visual actions according to the conceptual diagram shown in Figure 7. Participants' eye movement behavior is broadly categorized into saccades and fixations. Regarding fixations, participants repeatedly make very short fixations to acquire information from the target object. In this process, the distribution of fixations that can occur can be categorized into the following three types:

- Remaining stationary on a specific point of a specific object for an extended period ((a) in Figure 7)
- Remaining stationary on the same object while shifting gaze to several parts of it ((b) in Figure 7)
- Repeating very short stationary periods and saccades, each time stationary on a different object ((c) in Figure 7)

We defined the three visual action classifications shown in Figure 7 as "1 shot", and examined the distribution of the time $t_{shot}$ required for each shot to reveal the distribution of visual actions among participants. Table V shows the statistics for this. A clear difference in trend is that the visual behavior of $S_c$ differs significantly from that of $S_a$ and $S_b$. For $S_c$, the median $t_{shot}$ was $(R,\ N,\ L) = (581,\ 621,\ 641)$ [ms], approximately half the time compared to $S_a$ and $S_b$. Additionally, $S_c$ exhibited very small values for other metrics such as $Q_1, Q_2,\ Q_3$, and $IQR$ compared to the other groups.

## IV. The Relationship Between the Variety of Perceptual Information Quality Combinations and Cognitive Load

Based on the above results, we consider the relationship between the combination of variety of perceptual information quality and the cognitive load experienced by participants.

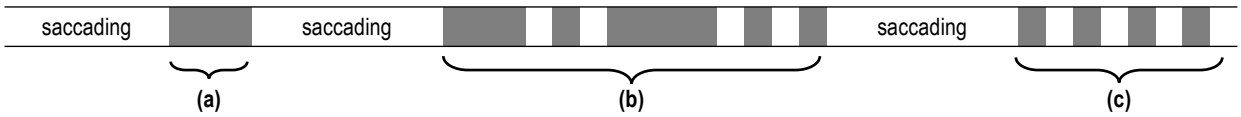



Figure 7. Setting fixation variety. (a) represents a state of continuously fixating on the same object, (b) represents a state of continuously fixating on different locations within the same object with saccades in between, and (c) represents a state of rapidly shifting gaze(short fixation) between different objects with saccades in between.

TABLE V. Visual variety-individual shot distribution statistics for each participant.

| | $S_a$ | | | $S_b$ | | | $S_c$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $R$ | $N$ | $L$ | $R$ | $N$ | $L$ | $R$ | $N$ | $L$ |
| $Q_1$ | 426.0 | 481.0 | 721.0 | 530.8 | 761.0 | 641.5 | 290.5 | 281.0 | 270.5 |
| $Q_2$ | 891.5 | 761.0 | 1323.0 | 1282.5 | 1683.0 | 1302.0 | **581.0** | **621.0** | **641.0** |
| $Q_3$ | 2083.3 | 1884.0 | 2124.0 | 2519.0 | 2424.0 | 4077.5 | 1202.0 | 1503.0 | 1542.5 |
| average | 1505.3 | 1443.4 | 1810.7 | 2335.6 | 2215.8 | 2336.2 | 1133.6 | 1087.9 | 1264.0 |
| stdev | 1578.0 | 1698.9 | 1631.6 | 2929.2 | 2208.3 | 2246.6 | 1863.3 | 1132.4 | 1472.5 |
| min | 60.0 | 161.0 | 40.0 | 100.0 | 201.0 | 160.0 | 80.0 | 80.0 | 20.0 |
| max | 8556.0 | 9056.0 | 8055.0 | 13425.0 | 10800.0 | 9598.0 | 15729.0 | 5631.0 | 7434.0 |

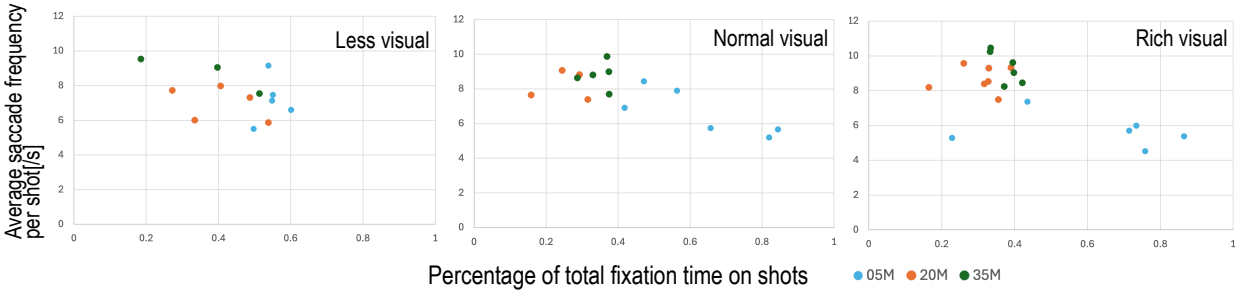



Figure 8. Relationship between the ratio of total fixation-saccadic time per memorized shot and the number of saccades per unit time.

TABLE VI. Summary of experimental results.

| | | | visual condition | | |
|---|---|---|---|---|---|
| | | | $L(++)$ | $N(-)$ | $R(-)$ |
| Auditory Condition | $L$ | Fixation time | N/A | N/A | N/A |
| | | ANOVA | N/A | N/A | N/A |
| | | $\bar{r_p}$ | $+$ | $++$ | $+$ |
| | $N$ | Fixation time | $++$ | N/A | $--$ |
| | | ANOVA | $*$ | N/A | N/A |
| | | $\bar{r_p}$ | $+$ | $+$ | $--$ |
| | $R$ | Fixation time | $++$ | $--$ | $++$ |
| | | ANOVA | N/A | $*$ | N/A |
| | | $\bar{r_p}$ | $+$ | $+$ | $-$ |

### A. Relationship between Average Fixation Time, Information Integration, and Cognitive Load

Characteristics of fixation time distributions across different perceptual information qualities revealed a fundamental interaction effect dominated by auditory information. Furthermore, fixation time statistics showed that combining visual and auditory qualities tend to result in longer fixation durations compared to single information quality varieties, relative to the average fixation time for either visual or auditory quality alone. This phenomenon is explained based on Figure 1 as follows. In multimodal information processing, as depicted in Figure 1, integrating $I^{sym}(t)$ at the *degree of matching* requires searching for objects within long-term memory that contain multiple perceptual information types. In the example from Section III, focusing on the quality of the input information (visual and auditory), Table VI shows various results for the combination of variety.

The main effect was observed for auditory information, so we will examine each auditory information category.

First, for the auditory information condition $L$, there was no significant trend in fixation time, while $\bar{r_p}$ showed a tendency toward dilation. This suggests complex information processing is occurring due to the auditory information. Specifically, $(N,\ L)$ exhibited significant pupil dilation. If this auditory condition is appropriate for the listener, it indicates cognitive load arising from language processing.

Next, for auditory information with condition $N$, the tendency differs depending on the visual condition. For $(L,\ N)$, pupil dilation occurred, while for $(R,\ N)$, pupil constriction occurred. Additionally, for $(L,\ N)$, fixation time was longer, while for $(R,\ N)$, fixation time was shorter. From them, it suggests that pupil response and fixation are linked. Therefore, depending on the quality of visual information, the following behavioral differences are expected:

- When quality is low, auditory information is obtained in a state where nothing else is visible. To confirm which part the explanation refers to, longer fixation times occur.

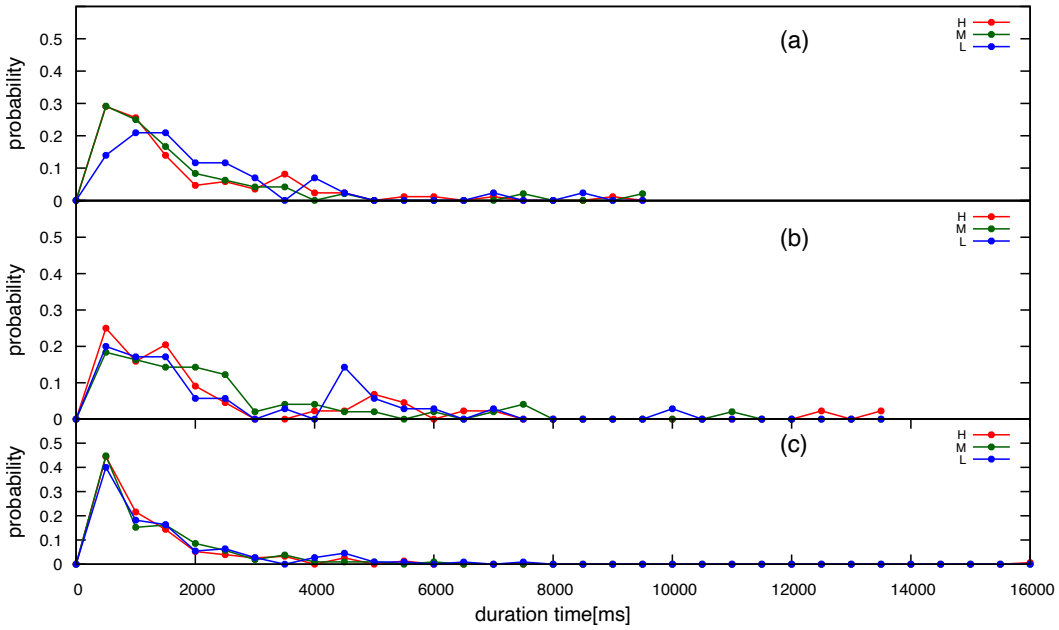



Figure 9. Distribution of fixation activity time per subject. (a) and (b) represent subjects with no interest in architecture ($S_a$, $S_b$), (c) represents subjects with interest in architecture ($S_c$).

- Under the conditions of this study, excessively high quality is approximately equivalent to having too much visual information (objects). That is, obtaining auditory information amidst numerous visual targets leads to unstable visual point, resulting in shorter fixation times. This cognitive overload may explain the smaller $\bar{r_p}$ values.

Therefore, under the $(N,\ N)$ condition, where a moderate cognitive load is applied, the results suggest that the value of $\bar{r_p}$ may show a slight upward trend.

Next, the condition where auditory information was $R$ also had a very large effect on fixation time. Condition pair $(N,\ R)$ had very short fixation times, while the others had long ones. Furthermore, the pupil response in condition pairs $(L,\ R)$ and $(N,\ R)$ was somewhat large. This suggests that when visual information is excessively scarce or abundant, longer fixations may be maintained to integrate it with auditory information. In our case, the origin of complexity is considered to stem from the integration of auditory and visual information. On the other hand, for condition pair $(R,\ R)$, where there is much to integrate, it is suggested that subjects may abandon memory of what they saw and heard in the VR space, primarily as a result of receiving excessive stimuli from visual/auditory information.

### *B. Individual Differences in Eye-Movement Behavior*

Due to individual differences in human behavior, we focus on eye-movement behavior within the virtual world to perform individual-level analysis. Figure 8 shows the total fixation time - total saccadic time ratio and the average saccade frequency per second for shots recalled in the recall test. Although individual differences exist, a general trend shows a slight increase in saccade frequency as visual information moves from left to right. From the figure, the fixation-to-saccade ratio per shot was generally around 0.4, and the saccade frequency averaged approximately 8 to 10 saccades per second. We consider this trend to show no significant variation.

Figure 9 shows the shot time distribution for each participant. The number of shots for $S_a$ was $(L,\ N,\ R) = (87, 50, 44)$, for $S_b$ it was $(45,\ 50,\ 36)$, and for $S_c$ it was $(153,\ 106,\ 111)$. Looking at the distributions for $S_a$ and $S_b$, while there are differences in peak locations and visual conditions, they show distinct distributions for each condition. Generally, the peak is around 500ms, but the shot distribution extends relatively far up to about 1500ms. Additionally, there is a second peak around 4000ms. This trend is particularly pronounced when the visual condition is $L$.

In contrast, $S_c$ consistently changed shot scenes at similar time intervals regardless of visual condition, showing no variation based on visual condition. On the other hand, during the recall test, $S_c$ provided quite detailed descriptions regarding drawing but used few verbal expressions.

This suggests that the way information is acquired in the virtual world is significantly influenced by the participant's timing for selecting specific scenes—that is, the shots. Participants like $S_c$, who possess an interest in architecture and are skilled at information acquisition, extract shots at consistent intervals regardless of visual appearance. They acquire a large amount of information in relatively short, fragmented intervals, enabling the extraction of detailed features. In contrast, participants like $S_a$ and $S_b$, who lack interest in architecture and are beginners in information acquisition, likely attempt to gather as much information as possible in a single shot,

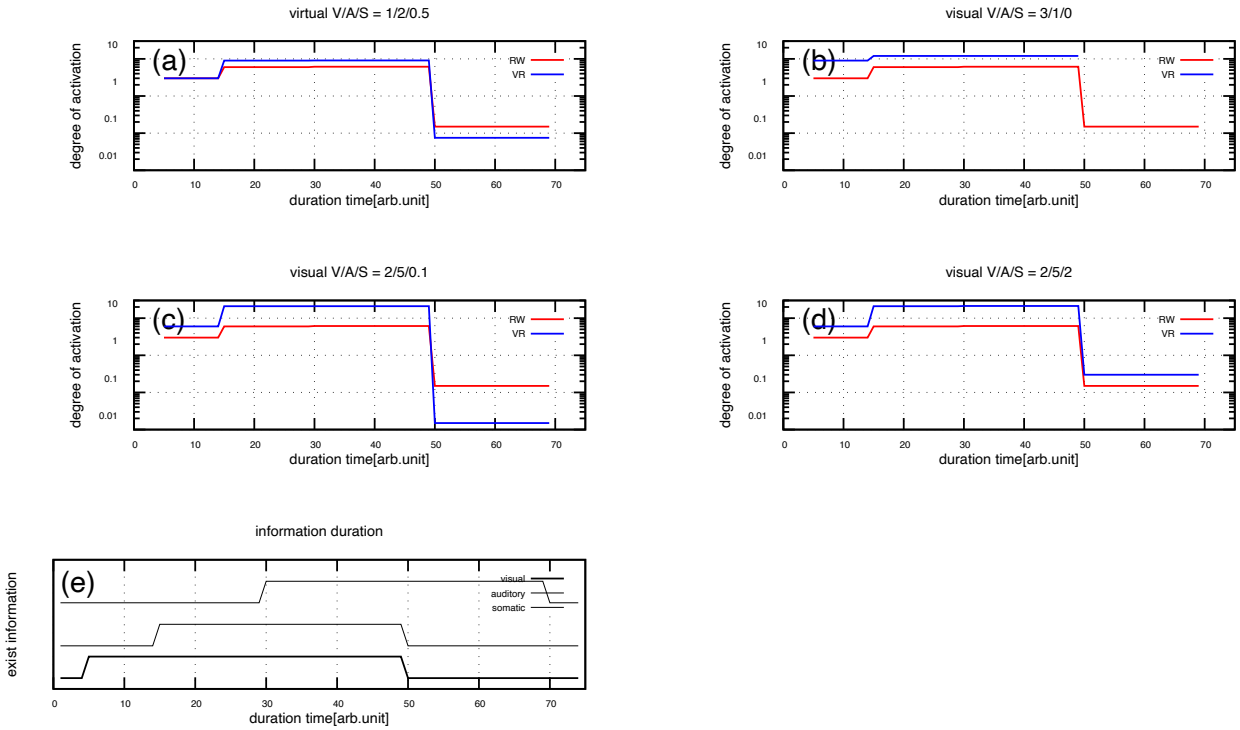



Figure 10. The trends of estimated $I^{syn}(t)$ which are changed three perceptual information(visual, auditory, somatic) amplified in Virtual Reality space.

resulting in a more variable shot time distribution.

Specifically, focusing on $S_c$ and analyzing the details, when the objects of interest were $A, B, C$, a tendency was frequently observed where the fixation point would shift from the object of interest to another object—such as $A - B$, $A - C$, or $B - C$—before returning to the original location. This suggests a connection to the findings of Kurihara et al. [10], who demonstrated that temporarily shifting the fixation point away and then returning it to the same location enhances memory consolidation.

### C. Perception in the Real/Virtual World

So far, we discussed that the impact of combining multimodal perceptual information in virtual spaces on cognitive load. Finally, we will mention what can be expected regarding the relationship between perception and cognition in real and virtual spaces. Figure 10 shows the trend of $I^{syn}(t)$ when the degrees to which visual, auditory, and somatic information are emphasized in VR are varied. The solid red line in the figure shows $I^{syn}(t)$ when visual, auditory, and somatic information are received in the real world. Here, we set $j = 1, 2$. Both visual and auditory information equal 1 for one, and 2 for the other. The somatic information is set to 0.5 on one side and 0.3 on the other. The solid blue lines indicate the degree to which the same information is distorted in VR.

Figure 10 (e) shows the duration of information obtained from each sense. In contrast, Figures 10 (a)∼(d) show the degree of integrated information activation calculated by Equation (1). Figure 10 (a) shows the case where auditory is multiplied by a factor of 2 and somatic by a factor of 0.5. For $t < 50$, the VR space is slightly more chunk activated, but the characteristics are almost same. However, at $t \geq 50$, when only somatic information is perceived, the chunk activation in the VR space is lower. In Figure 10 (b), the visual information is markedly increased, while the somatic information is not reproduced in the VR space. For $t < 50$, the activation of chunk in the VR space is markedly increased, but at $t \geq 50$, the somatic information is lost; Hence, there is no chunk activation in the VR space. In Figure 10 (c), the somatic information is lowered to 0.1 and the information is emphasized in the form of visual<auditory. In particular, at $t \geq 50$, the somatic information is still present, but its effect is much smaller. Figure 10 (d) is the case where the somatic information is also doubled. Compared with Figures 10 (b) and (c), chunk activation remains high at $t \geq 50$.

The intensity of human sensation is expressed as a logarithm according to Weber-Fechner's Law. Therefore, as shown in Figure 10, even if the difference in sensory information is very slight, it suggests that the human senses can distinguish this difference. The sense of "slightly different from the real world" felt in VR content is thought to be caused by such slight differences in sensory information. The sensory information obtained in real space is not necessarily large, as shown in the example in Section III. However, it is easy to understand that these small differences lead to a sense of discomfort, which in turn indicates a decrease in immersive perception.

In the present case, we only dealt with a very simple integration of information. To advance our understanding of human sensory perception and use knowledge in VR spaces,

scholars should develop a new approach that uses operators in Equation (1), such as Adaptive Control of Thought—Rational (ACT-R) [11] and Model Human Processor with Realtime Constraints (MHP/RT) [11] which incorporate Two Minds, to integrate information in a cognitive architecture [12][13][14].

## V. Conclusion and Future Work

To realize adaptive VR, we need to design deeper immersion resulting from human interaction with real/VR spaces. As a first step, this study describes a sensory-cognitive model for VR spaces. Based on the described model, we analyzed how information is acquired in a virtual world, focusing on visual and auditory information, and how behavior changes when conditions are altered, using results of recall test. The results suggested that some malfunction occurs under conditions other than $(N,\ N)$. Furthermore, it suggested that the degree of proficiency in acquiring information from space may influence eye-movement behavior and, consequently, the state of memory. Connecting the two issues, multimodal information and chunk activatin, we undertake the research qualitatively and explain the phenomenon that can occur when one or more types of information (visual, auditory, or somatic) is overemphasized or surpressed in a VR space. Expressing human sensory intensity as a logarithm according to Weber-Fechner's Law, we suggest that human senses can distinguish differences in sensory information, even if the differences are very slight. Considering these points, we are able to deepen our understanding of how the VR space realizes the immersive effect with impressive each other. Moreover, we are able to design "adaptive" immersive contents. In the future, it is necessary to investigate in experiments whether the degree of immersion felt by users changes when they experience VR content by changing the degree of emphasis of each sensory information. The metrics used to judge the degree of similarity between the real and virtual worlds can be defined as the overlap between the information held in the WM and the information in the LTM that has been activated up to that point in time. As the activation of information in the LTM is considered to be reflected in biological information, future experiments could be conducted using eye gaze and skin resistance measurements and subjective evaluation by means of questionnaires. Hysteresis can be considered based on the impact of inputs from the environment on the memory of the time series.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 20H04290, 22K12284, 23K11334, and National University Management Reform Promotion Project. The authors would like to thank Editage (www.editage.com) for the English language editing.

## References

[1] T. Nakagawa, M. Kitajima, and K. T. Nakahira, "Model-Based Analysis of the Differences in Sensory Perception between Real and Virtual Space : Toward "Adaptive Virtual Reality"," in AIVR 2024 : The First International Conference on Artificial Intelligence and Immersive Virtual Reality, 2024, pp. 39–44.

[2] C. Baker and S. H. Fairclough, "Chapter 9 - adaptive virtual reality," in Current Research in Neuroadaptive Technology, S. H. Fairclough and T. O. Zander, Eds. Academic Press, 2022, pp. 159–176. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128214138000142

[3] J. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, 12 1980, pp. 1161–1178.

[4] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." Journal of personality and social psychology, vol. 76 5, 1999, pp. 805–19. [Online]. Available: https://api.semanticscholar.org/CorpusID:14362153

[5] J. Bao, X. Tao, and Y. Zhou, "An emotion recognition method based on eye movement and audiovisual features in mooc learning environment," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, 2024, pp. 171–183.

[6] N. Sun and Y. Jiang, "Eye movements and user emotional experience: a study in interface design," Frontiers in Psychology, vol. Volume 16 - 2025, 2025. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1455177

[7] A. Steed, Y. Pan, F. Zisch, and W. Steptoe, "The impact of a self-avatar on cognitive load in immersive virtual reality," in 2016 IEEE Virtual Reality (VR), 2016, pp. 67–76.

[8] S. M. H. Mousavi et al., "Emotion recognition in adaptive virtual reality settings: Challenges and opportunities," CEUR Workshop Proceedings, vol. 3517, jan 2023, pp. 1–20. [Online]. Available: https://sites.google.com/view/wamwb/

[9] T. Batra and P. Chunarkar-Patil, "Virtual reality in bioinformatics," Open Access Journal of Science, vol. 3, no. 2, 2019, pp. 63–70.

[10] Y. Kurihara, M. Shino, K. Nakahira, and M. Kitajima, "Visual behavior based on information foraging theory toward designing of auditory information," in Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 1: HUCAPP, INSTICC. SciTePress, 2024, pp. 530–537.

[11] F. E. Ritter, F. Tehranchi, and J. D. Oury, "ACT-R: A cognitive architecture for modeling cognition," WIREs Cognitive Science, vol. 10, no. 3, 2019, p. e1488. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1488

[12] M. Kitajima, Memory and Action Selection in Human-Machine Interaction. Wiley-ISTE, 2016.

[13] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)," Behaviour & Information Technology, vol. 31, no. 1, 2012, pp. 41–58.

[14] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)," Biologically Inspired Cognitive Architectures, vol. 5, 2013, pp. 82–93.