Multimodal Technologies and Interaction

MDPI

Article

NariTan: Enhancing Second Language Vocabulary Learning Through Non-Human Avatar Embodiment in Immersive Virtual Reality

Shogo Fukushima *,†, Keigo Sakamoto † and Yugo Nakamura

Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan; sakamoto.keigo@arakawa-lab.com (K.S.); y-nakamura@ait.kyushu-u.ac.jp (Y.N.)
* Correspondence: shogo@ait.kyushu-u.ac.jp; Tel.: +81-92-802-3579
† These authors contributed equally to this work.

Abstract: With the rise of head-mounted displays (HMDs), immersive virtual reality (IVR) for second-language learning is gaining attention. However, current methods fail to fully exploit IVR's potential owing to the use of abstract avatars and limited human perspectives in learning experiences. This study investigated IVR's novel potential by using non-human avatars to understand complex concepts. We developed a system for learning English vocabulary through the actions of non-human avatars, offering a unique learning perspective. This paper presents an IVR vocabulary learning environment with a dragon avatar and compares word retention rates (immediate and one-week memory tests), subjective workload, and emotional changes with traditional methods. We also examined the vocabulary ranges that are teachable using this system by varying the number of avatars. The results showed that the proposed method significantly reduced forgotten English words after one week compared to traditional methods, indicating its effectiveness in the long term.

Keywords: virtual reality; enactment effect; non-human avatar; second language vocabulary learning; bilingual dual coding theory

Citation: Fukushima, S.; Sakamoto, K.; Nakamura, Y. NariTan: Enhancing Second Language Vocabulary Learning Through Non-Human Avatar Embodiment in Immersive Virtual Reality. *Multimodal Technol. Interact.* **2024**, *8*, 93. https://doi.org/10.3390/mti8100093

Academic Editor: Julius Nganji

Received: 12 September 2024
Revised: 8 October 2024
Accepted: 10 October 2024
Published: 18 October 2024

1. Introduction

Vocabulary is fundamental to language learning, and a limited vocabulary impedes language-related activities, such as conversing and multimedia content comprehension [1,2]. Learning-by-doing [3] and the enactment effect [4], also known as the subject-performed task (SPT) effect [5], are traditional methods that have been widely utilized for learning vocabulary through physical experience. The enactment effect (or SPT effect) has been extensively studied, particularly in the fields of psychology [6] and second-language acquisition [7,8], which is believed to be effective in extending vocabulary retention by performing movements related to learning targets.

Recently, with the proliferation of head-mounted displays (HMDs), research incorporating immersive virtual reality (IVR) into language learning using these technologies has been increasing. According to The American Heritage Dictionary, the term "Virtual" is defined as "Existing or resulting in essence or effect though not in actual fact, form, or name", meaning that VR can be defined as a technology that reproduces the essence of reality. However, IVR is described as "technology that allows the user to experience physical and/or behavioral simulations" [9] and is primarily realized through devices such as HMDs or Cave Automatic Virtual Environments. In IVR, it is possible to learn vocabulary that is rarely experienced in the real world by freely changing objects and spaces. Studies that utilize IVR for vocabulary learning have presented words related to the virtual environment [10,11] and virtual objects [12,13]. Moreover, because IVR can easily create immersive environments, learning through enactment is advantageous for vocabulary retention. However, these prior studies have focused on changing objects or

environments, and there has been little examination of variables such as the first-person perspective of the learner or the avatar used. The avatars used were either human or reproduce parts, such as the hands. Therefore, the possible virtual experiences are limited to "learning experiences from a human perspective", and the vocabulary assumed in these studies is fundamentally related to humans.

We propose NariTan, which provides a "learning experience from the perspective of a non-human avatar, for memorizing English vocabulary through actions suitable for that avatar (Figure 1)". By changing the avatar, it is possible to teach English vocabulary that cannot be included in learning experiences based on human perspectives. The name combines "Nari-kiri" (becoming) and "Tan-go" (word). Non-human avatars refer to avatars that are not human, such as dogs or cats. By learning from a non-human perspective, it is possible to incorporate English words that cannot be included in a human perspective-based learning experience (see experience video in the Supplementary Materials). We hypothesized that the episodic memory of experiencing unusual body movements and perspectives that humans do not typically perform would be beneficial for the retention of English vocabulary.

Figure 1. (**Left**) Experiencing the verb "ravage" through a dragon avatar. (**Right**) Learning verbs using the whole body while becoming a dragon avatar.

Plausibility in an IVR environment is considered a key factor for maintaining presence in IVR experiences [14]. Presenting vocabulary that matches the appearance of a customized avatar and performing actions corresponding to the avatar's characteristics may contribute to the enhancement of presence in IVR. Smith et al. have reviewed and organized the knowledge regarding the formation of episodic memory in IVR environments, highlighting the importance of maintaining presence [14]. Furthermore, it is considered that embodying a non-human avatar and performing unfamiliar body movements are crucial for structuring memories and retaining them as long-term memories.

In this study, we developed the NariTan learning system and conducted experiments comparing the results of memory tests, emotional changes, and subjective load during English vocabulary learning with those of general learning methods, such as flashcards, to evaluate the effectiveness of the proposed system. Additionally, we examined how well the learning system covered the target English vocabulary. The contributions of this study are as follows:

- We proposed and implemented a system where users dress-up non-human avatars in an IVR environment to learn English vocabulary. We developed new interactions to utilize IVR spaces as learning environments and created unique vocabulary-selection methods tailored to each avatar.

- We validated the effectiveness of this body dress-up system through quantitative assessments, including vocabulary tests and episodic memory tests. Additionally, we quantified the workload and usability of this learning method.
- We also explored the range of vocabulary that can be acquired with this learning method using a large language model (LLM).

2. Related Work

Bilingual dual-coding theory [15] is a model for second-language vocabulary acquisition. This theory posits the existence of two systems in the cognitive processing of words: verbal symbolic systems and nonverbal objects and events. Linguistic information is encoded by the former system, whereas nonverbal information, such as illustrations and experiences associated with words, is encoded by the latter system. It has been suggested that images that are doubly encoded by imagery and language result in a better memory performance than words alone [16]. Several studies supporting the bilingual dual-coding theory have recognized the additive effect [17,18], which refers to the phenomenon in which learning utilizing both linguistic and imagery systems (e.g., learning vocabulary with pictures) results in approximately twice the number of word reproductions when compared with learning that uses only the linguistic system (e.g., copying second-language words).

Multimedia learning, which focuses on the connection between linguistic and imagery systems, has also been studied. Classical multimedia learning typically uses images. Studies of image-based learning have shown that adding explanations to illustrations improves comprehension and problem-solving abilities [19], whereas explanations unrelated to illustrations negatively affect learning [20]. Additionally, videos that provide more information than images are used for learning. Extensive research has been conducted on the effects of incorporating short animations in learning [21]. Moreover, recent studies on vocabulary learning systems using videos have made notable progress. Zhu et al. developed Vivo [22], a system that uses videos related to word meanings instead of paper dictionaries. Furthermore, several studies have discussed the most appropriate subtitle presentation methods within videos for viewers [23–25].

It is important to enhance the traces of episodic memory through experience to increase the amount of information in an imagery system. Episodic memory is a type of declarative memory that involves personal experiences and is characterized by information retention accompanied by various contextual information (temporal, spatial context, bodily, and psychological states) associated with the time of the event [26]. One method to enhance memory trace during an experience is learning through bodily movements, known as the enactment effect (SPT effect) [4,5]. The enactment effect refers to the phenomenon in which learning through self-generated actions (enactment) to express a concept results in better recall test performance than learning through word presentation or verbal instruction [5]. The effectiveness of using enactment in second-language learning for memory retention and understanding of learned words has been confirmed in studies using both artificial languages [27–29] and natural languages [30,31]. By applying this to second-language learning systems, Nishida et al. developed a learning-by-doing system that uses ultrasonic sensors to remember the positions of objects and people and presented educational content based on those positions [3].

IVR has also been used to enrich visual cues during the learning experience. In Ogma, Swedish words are embedded in IVR environment objects; learners then move close to the objects using an armband-style controller, whereupon the Swedish text and pronunciation are displayed for learning. An experiment with 19 participants learning 10 Swedish words showed that, while the immediate test performance was lower for Ogma, the word retention rate was higher than that of text-only learning in a one-week test [32]. Vazquez et al. developed Words In Motion, incorporating the enactment effect into IVR-based learning. In Words In Motion, learners perform appropriate actions for words in IVR space while learning them. For example, when learning “paint”, a brush object follows the controller and moving the controller along the indicated path displays

the word. Although the immediate test scores for Words In Motion were lower than those of the other two conditions, the word retention rate was higher in the one-week test [12].

Previous research has shown that actions and gestures can be cognitively distinguished in the real world [33]. Actions, defined as movements on or using objects, have a different cognitive framework and present evidence of different cognitive outcomes than those of gestures, defined as movements about objects. Additionally, actions are known to promote stronger memory than gestures [34]. IVR studies have suggested that interactions in the form of actions and gestures lead to different cognitive outcomes and that actions provide stronger memory effects. Furthermore, embodied controller actions in IVR have been suggested as being more similar to real-world actions in terms of the benefits of verb memory [13].

In this study, we implemented an IVR system for learning verbs that are difficult to learn through real-world experiences by performing various actions in the IVR space. We clarified the memory retention effects of this system. Additionally, we defined verbs specific to avatars and investigated the vocabulary coverage achieved by changing the types of avatars, which constitutes the academic contribution of this study (Figure 2). Recently, there has been a growing body of research focusing on plausibility and presence in IVR, with many studies utilizing non-human avatars such as dogs [35], cats [36], and robots [37]. While these studies do not specifically focus on vocabulary learning systems, the IVR systems they employ are similar to a medium, suggesting that this paper could offer relevant insights.

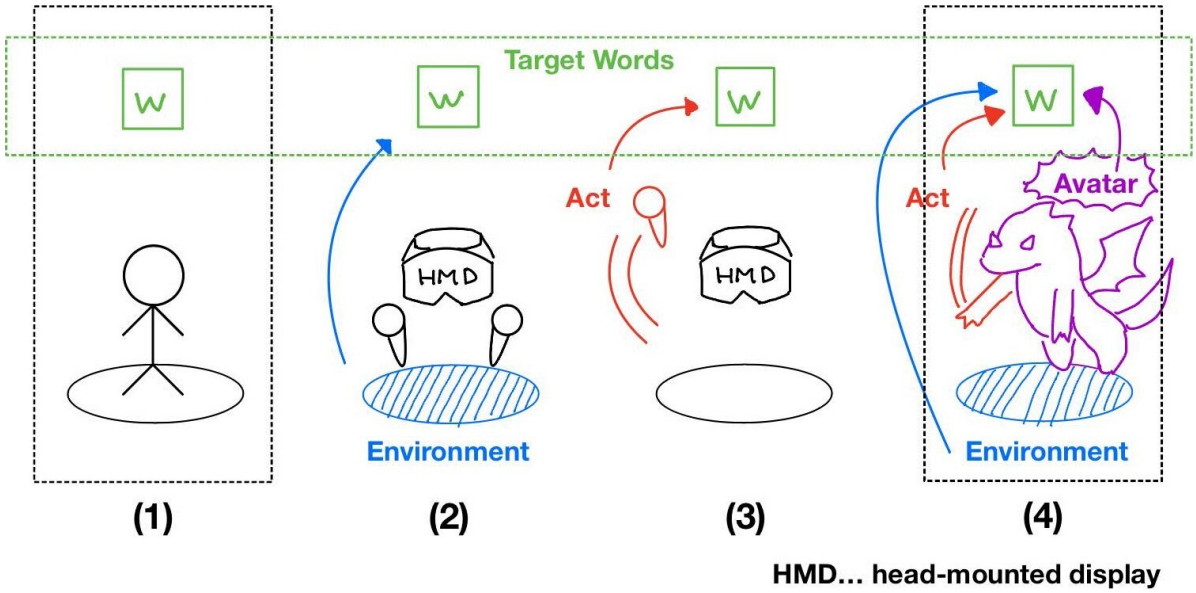


Figure 2. Academic significance of this research: (**1**) A method for memorizing words in the real world without using immersive virtual reality (IVR). (**2**) A method for selecting and learning words that match the IVR environment. (**3**) A method for learning words by enacting them by matching the movements of IVR objects. (**4**) A method for selecting and learning words typically attributed to the IVR avatars.

3. Proposed Method

NariTan is a learning system aimed at realizing an "IVR learning experience from the perspective of a non-human avatar by memorizing English words through actions suitable for that avatar". The name combines "Nari-kiri" (becoming) and "Tan-go" (word). Non-human avatars refer to avatars that are not human, such as dogs or cats. By learning from a non-human perspective, it becomes possible to incorporate English words that cannot be included in a human perspective-based learning experience. We hypothesized that experiencing actions and perspectives that humans cannot typically perform would facilitate the memorization of words.

3.1. Software for Implementation

This system was developed using Unity 2021.3.24f1 https://unity.com/ja (accessed on 11 September 2024). XR Interaction Toolkit 2.0 https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@2.5/manual/index.html (accessed on 11 September 2024) for implementing the core interaction system, Final IK 1.9 https://assetstore.unity.com/packages/tools/animation/final-ik-14290 (accessed on 11 September 2024) for implementing the avatar's IK, and Dynamic Bone 1.3.4 https://assetstore.unity.com/packages/tools/animation/dynamic-bone-16743 (accessed on 11 September 2024) for expressing bone sway were used. The HMD used was PICO4 Enterprise https://business.picoxr.com (accessed on 11 September 2024), which was connected to the PC via a wired connection. Although PICO4 can be connected wirelessly to the PC using WiFi, a wired connection was used to ensure stable image quality and reduce screen shaking.

3.2. Implementation of the Avatar

To realize "becoming a nonhuman avatar", it is necessary to have a 3D model of the non-human avatar that follows the movements of the HMD and controllers. In this study, a dragon avatar was used as an example of a non-human avatar. This avatar was chosen because its significantly taller stature than that of humans provides a different self-perspective and allows for the inclusion of many actions that are not typically experienced by humans, such as flying and breathing fire. The 3D model of the dragon is shown in Figure 3. It was a quadrupedal 3D model and difficult to implement as an avatar in its original form. Therefore, a humanoid rig was created using Blender. The edited 3D model was exported as an FBX file and imported into Unity, where it was implemented as a dragon avatar using 3-point tracking with VRIK, a feature of Final IK. The wings were designed to move in conjunction with the arms, and the tail was designed to sway with body rotation.

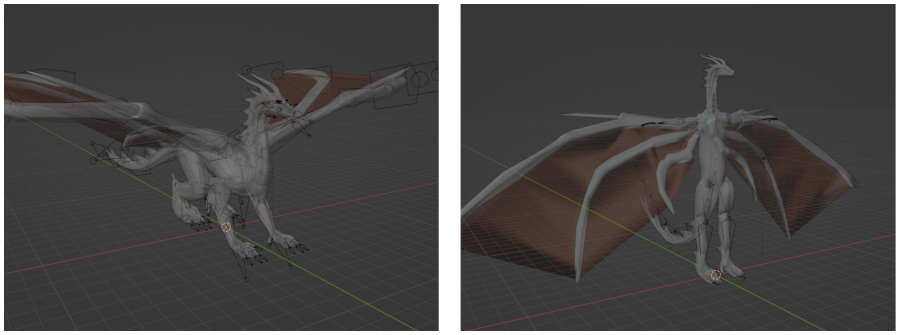


Figure 3. 3D model of the dragon used as an avatar in this study. The **left** and **right** panels show the model before and after processing in Blender and fitting into a humanoid rig, respectively.

3.3. Implementation of Interaction

To realize "memorizing English words through actions suitable for a non-human avatar", interactions that allowed the enactment of the English words were used. To associate the IVR environment experienced as an episodic memory with the related English words, an independent IVR environment was implemented for each word (Figure 4). The IVR environment for each English word included a sufficiently large area for learners to move around freely and a skybox as the background. Additionally, several objects necessary for the enactment were placed. The following operations were possible within the IVR environment:

- Movements using the left and right sticks on the controller;
- Grabbing/releasing objects by pressing/releasing the side buttons on the controller;
- Performing actions by pulling the trigger on the controller (for specific English words only).

To determine whether each English word was enacted, the following criteria were used, allowing the expected process to be executed immediately upon enactment:

- Collision detection (e.g., the dragon's hand colliding with a rock);
- Detection of entry into an area (e.g., dropping a stone into the sea);
- Detection of extent of movement (e.g., swinging the arm down significantly).

As feedback indicating that the enactment was performed, the English word and its Japanese translation were displayed in a text window for three seconds, and the pronunciation of the English word was presented. To prevent words from being displayed in every frame, the system was designed such that the words would not be displayed redundantly during the entire word-display period. Additionally, visual and auditory effects were implemented as feedback. For example, when a building was destroyed with a beam, explosion effects and sounds were incorporated. The reason for implementing such feedback was to associate it with English words as episodic memory.

Figure 4. Some examples of the IVR environments and objects (where English words and their Japanese translations are displayed).

A rectangular, semi-transparent text window showing instructions for the actions and time limit was always displayed in the learner's view. The text window followed the movement of the HMD and was always displayed in front of the learner. However, to prevent hindering the IVR experience, the user interface (UI) was designed to be semi-transparent, such that objects could still be seen through the UI. When the user performed the actual action, the English word and its Japanese translation were displayed within the text window (see Figure 5).

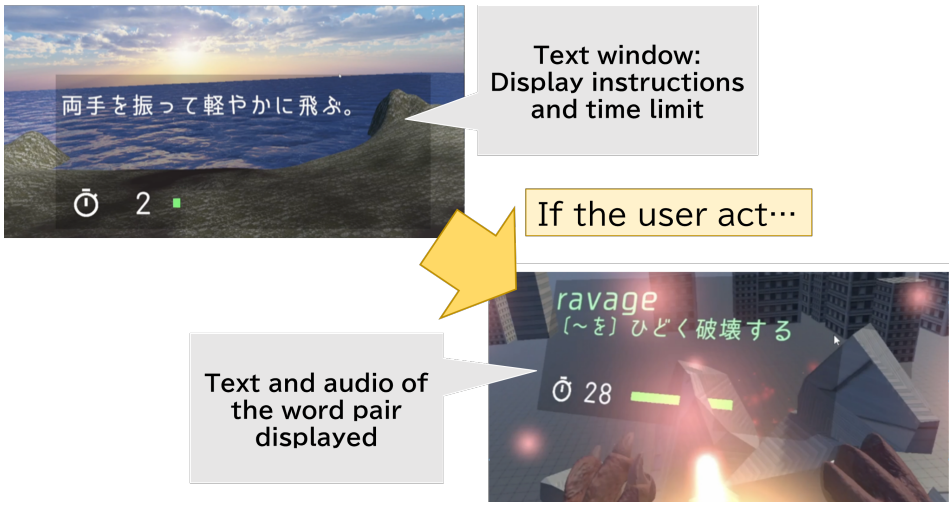


Figure 5. Text window. Instructions for actions are generally displayed in Japanese. When the user performs an action, a pair of English and Japanese text appears in the text window, and the corresponding audio is played simultaneously.

3.4. Implementation of Tutorial

A tutorial environment was implemented to help the learners become sufficiently accustomed to the avatar's body (Figure 6). The learners were first moved to this environment before entering the IVR space. The tutorial environment included cubic objects that could be grabbed and moved, and a green area indicated the place to which the objects should be moved. Additionally, a mirror was set up to allow learners to see the entire body of the dragon avatar moving in conjunction with their own movements, thus making them constantly aware that they were a dragon.

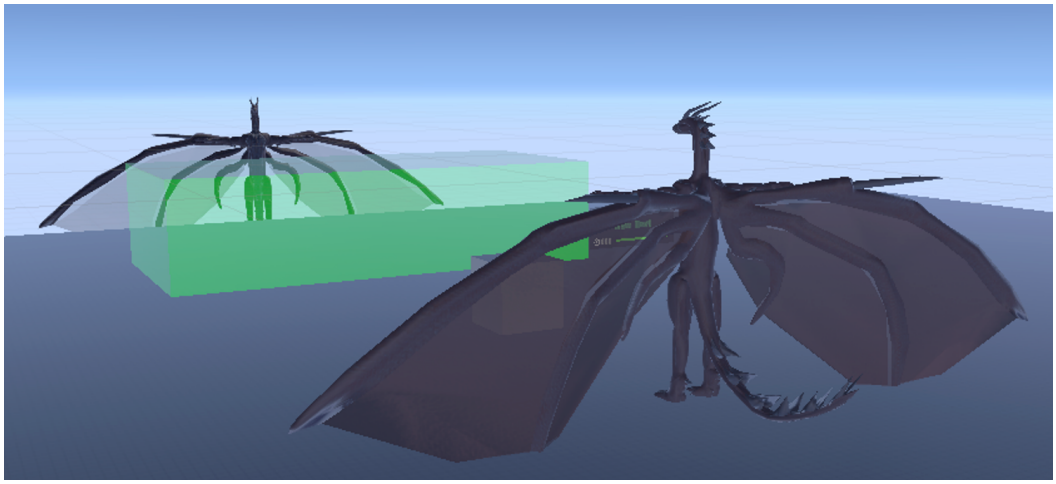


Figure 6. IVR environment for the tutorial. Users can see themselves in a mirror while practicing the task of moving a box to the green area.

3.5. Word Selection

This method involves memorizing certain transitive verbs specific to a particular non-human avatar by becoming that avatar. It is necessary to select avatar-specific English words and reduce them to words that match the proficiency levels of the participants. This procedure is illustrated in Figure 7.

We used ChatGPT [38] as the LLM and a combination of the BNC/Corpus of Contemporary American (COCA) word family lists [39] (BNC/COCA) and CEFR [40] as the word dataset. The BNC/COCA word family lists [39] are frequency-ordered English word lists created by the nation based on the British National Corpus (BNC) [41] and the COCA [42]. The lists contain 25,000 words (+derivatives/additional words) which are categorized into levels from 1 to 25 in units of 1000 words, with higher levels indicating less frequent words in the corpus. In this study, 25,000 words were used, excluding derivatives and additional words.

Although transitive verbs suitable for each avatar can be selected manually, the sheer number of words makes this impractical. Therefore, ChatGPT was used. The steps for selecting English words using ChatGPT are as follows: To select avatar-specific English words, the prompt "List 200 unique words that are verbs of movements possible for AVATAR". was input, and this task was repeated until fewer than ten new words appeared in the BNC/COCA list. The term "AVATAR" in the prompt was replaced with the name of the avatar being used. For example, "the creature dragon" was substituted for the dragon avatar used in this study.

In the experiment, it was necessary to use words that matched the participants' vocabulary levels. BNC/COCA is strictly divided by frequency and is not categorized by the language proficiency level of the user. Therefore, the Common European Framework of Reference for Languages (CEFR) [40] was introduced as a criterion for level categorization based on user proficiency.

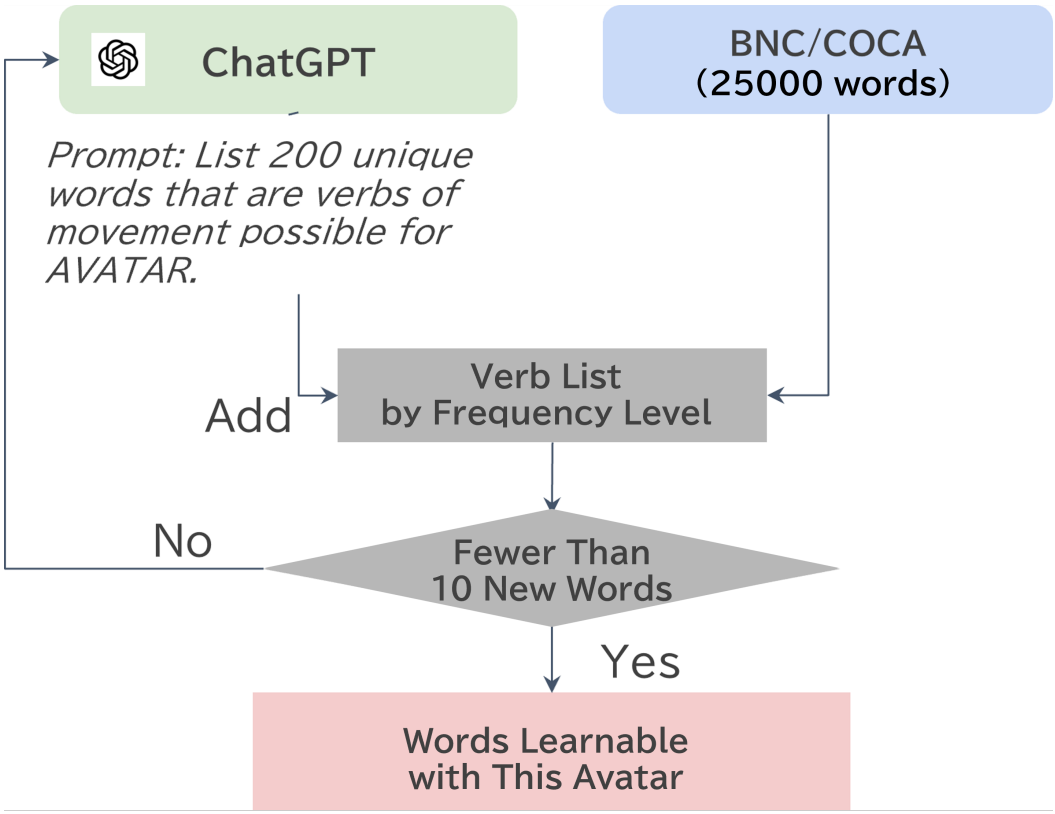


Figure 7. Process of generating avatar-specific verbs using a large language model (LLM).

4. Experiment 1: Memory Retention with Proposed Method

To clarify the effectiveness of the proposed method for vocabulary retention, the participants memorized English words, and the forgetting rate was calculated by conducting word tests immediately after learning and one week later. The results were compared with those obtained using flashcards as the baseline method (R1). The reason for using flashcards as the baseline is that they are a commonly used application for English vocabulary learning and the results would provide valuable insights for general English learners. Furthermore, flashcards have been used as a baseline for the evaluation of vocabulary learning systems that utilize IVR environments [12,32], 3D audio [43,44], and those that focus on the enactment effect [3], allowing for comparisons with these systems. The proposed method aims to enable the memorization of the meanings of English words using episodic memory as nonverbal imagery through physical actions. Therefore, the correlation between episodic memory and vocabulary learning scores was investigated (R2). Additionally, the effect of learning through enactment from the perspective of a non-human avatar on the subjective load and emotions was compared with that of the baseline method (R3).

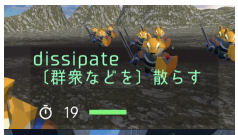
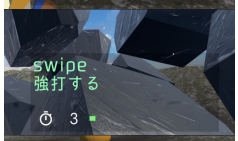
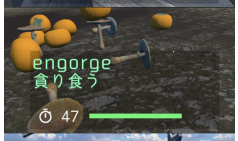
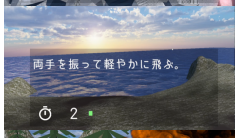
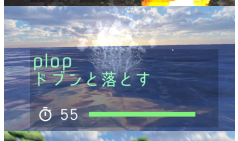
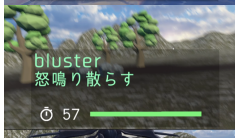
4.1. Participants

The participants were 28 university students (25 men and 3 women) with a median age of 22 years (first quartile: 20, third quartile: 23). However, the data of one participant were excluded owing to VR sickness during the experiment. Three participants did not take any English proficiency tests. The Test of English for International Communication (TOEIC) scores of the remaining participants ranged from 410 to 790 points. Ten participants had no experience of using VR HMDs, whereas 18 had previous experience of using the device. The 28 participants were divided into two groups of 14 each for learning using flashcards and NariTan.

4.2. Vocabulary and Materials

The experiment used ten words (Table 1) selected according to the word-selection method described in Section 3.5. We focused on verbs that could be learned using the dragon avatar, which were then reduced to verbs of CEFR level C2 or higher (406 words). This level of adjustment was made to minimize the influence of prior knowledge of English words, considering that the participants were university students. Subsequently, 406 words were further reduced to 10, considering the balance of interactions that could be expressed with those verbs (Table 1).

Table 1. Target English vocabulary used in the experiment, the scene images with English words and their Japanese translations, the instructional sentences for learning them, and how the user moves their body.

Scene	Target Verbs	Instruction	Interaction
	dissipate	Scattering soldiers with a hand.	Swinging arms left and right
	swipe	Striking a rock hard	Swinging a hand down
	engorge	Devouring food greedily	Grabbing food with a hand and putting it in the mouth
	ravage	Destroying buildings with a beam	Pointing the stomach at the buildings
	flit	Flying lightly by waving both hands	Vigorously swinging hands down
	incinerate	Burning a tree to ashes	Aiming flames from the mouth at the tree
	plop	Dropping something into the sea with a plop	Grabbing a stone with a hand and throwing it into the sea
	bluster	Yelling loudly	Inputting a loud voice into a microphone
	wag	Wagging the tail	Twisting the body
	elude	Skillfully dodging enemy projectiles	Moving the body left and right

4.3. Procedure

4.3.1. Day 1: Experiment on the First Day

In the pre-experiment questionnaire, the participants took a pre-test to determine whether they already knew the ten words to be learned in this experiment by answering the Japanese translations of the words (Figure 8). They also answered questions regarding their current subjective emotion. Next, the participants practiced using the learning applications. In the flashcard method, participants learned a practice set of two words. In the proposed method, the participants entered the tutorial environment, recognized themselves as the dragon avatar in the mirror, and learned the operations of the HMD and controllers.

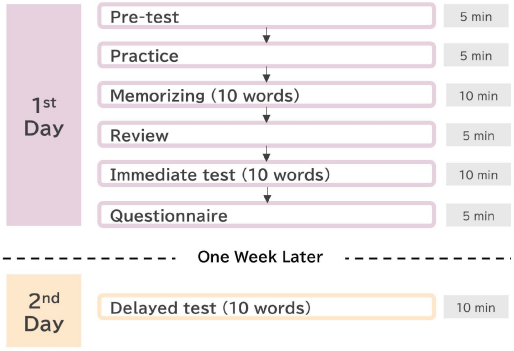


Figure 8. Experimental procedure. Experimental flows on days 1 and 2 are shown.

After the practice, the participants proceeded to the main experiment and learned ten words. In both conditions, there was no limit on the number of times the words could be presented; instead, the learning time for each word was limited to one minute, resulting in a total of 10 min of learning. As this was a between-subjects experiment, the participants learned using either the flashcard or the proposed method.

In the flashcard method, English words and their Japanese translations were presented, and the participants could check their pronunciation by pressing the playback button below. The proposed method involved learning through interactions, as explained in the previous section. The time limit was one minute per word, and the participants could always check the remaining time by means of a gauge UI that shrank over time in both methods.

After learning, the effectiveness of each test condition was reviewed using a tablet. The review was conducted by presenting multiple-choice questions, which is a feature of Flashcard Deluxe. In the questions, Japanese translations were presented with four options alongside the English word. The correct choices were highlighted in green, whereas incorrect choices were highlighted in red. Subsequently, the English word and its Japanese translation were presented again, and the participants could check their pronunciation by pressing the playback button. In the NariTan-based experiment, a screenshot of the IVR environment in which the word was learned was also presented when the English word and its Japanese translation were shown. This presentation was intended to associate the learned IVR environment with related English words as episodic memory.

A test was conducted immediately after the review. First, a mental arithmetic task of answering ten two-digit addition problems was performed, followed by answering the Japanese translations of the ten learned words. After the test, a post-experiment questionnaire was administered to assess the current subjective emotional evaluation, subjective workload, and usefulness of the system. Finally, oral questions were asked, and the participants provided feedback on the system through interviews.

4.3.2. Day 2: Experiment on the Second Day Conducted One Week Later

The second day of the experiment was conducted approximately a week later at the same location as on the first day. Among the participants, 23 took the test on the seventh day, one on the sixth day, and three on the eighth day. We intended all participants to complete the experiment at the same time exactly one week later; however, scheduling

at the same time one week later proved to be difficult. Therefore, we determined that, as long as it was the same day, in cases where it was not possible to conduct the experiment exactly seven days later, we allowed the participants to complete it the day before or the day after, and we adjusted the time of day to minimize any time differences as much as possible. Additionally, participants were instructed in the consent form not to share the details of the experiment with others

First, a one-week post-test was conducted to answer the Japanese translations of the ten words learned a week earlier. A questionnaire was then administered to check how much of the IVR experience episode in the proposed method was remembered. The participants were shown the English word alone and were asked which elements of the IVR experience they remembered. The participants answered yes/no to the following seven items:

- Background: Remembering the background.
- Objects: Remembering the objects seen/used.
- Events: Remembering what was performed/what happened.
- Body parts: Remembering the body parts used in the learning movements.
- Sound effects: Remembering the sound effects heard during the learning.
- Pronunciation: Remembering the pronunciation of the word.
- Instructions: Remembering the instructions given.

To eliminate the learning effect due to the order of word presentation, the order of words in the pre-test, learning, review, immediate test, and one-week post-test were all randomly arranged. All answers were collected using Google Forms. To avoid predicting the correct answers through predictive conversion, the keyboard's predictive conversion option was turned off and the browser's private mode was used.

4.4. Scoring of Word Tests

The participants wrote Japanese translations corresponding to English words using free text in the word tests. Therefore, it was necessary to clearly set the scoring criteria for the Japanese translations. Based on a study [45] that considered allowing differences in parts of speech as correct answers, we allowed for mistakes in parts of speech, kana/kanji mistakes, and differences in particles/auxiliary verbs. Additionally, this study considered synonyms to be correct answers. This was because the aim of the proposed method was to establish memory traces of episodic memory rather than to ensure strict semantic memory retention. Two external evaluators agreed on what could be considered synonymous.

The evaluation method for the word tests followed that adopted in previous studies [12,43], and the scores of the immediate word test and one-week post-test were used for the evaluation. Because all participants scored zero in the pre-test, the correct answer rates of the word tests conducted immediately after learning and one week after learning were used as word test scores. The forgetting count was obtained by subtracting the immediate test score from the one-week test score.

4.5. Other Evaluations

The following four items were measured:

- Subjective workload: Measured after the first day's experiment;
- System Usability: Measured after the first day's experiment;
- Emotional Changes: Measured before and after the first day's experiment.

The workload questionnaire used the NASA-TLX [46] index. The NASA-TLX consists of six scales: mental demand (MD), physical demand (PD), temporal demand (TD), performance (OP), frustration (FR), and effort (EF). The overall workload was assessed by combining these scales. This study used the adaptive weighted workload (AWWL) method to evaluate the overall workload. AWWL assigns weights of 6, 5, 4, 3, 2, and 1 in the order of the score magnitude (averaging for ties), and the comprehensive workload was evaluated by dividing the total weighted scores by 21 (sum of the coefficients).

The system evaluation was conducted using the system usability scale (SUS) [47], which is a commonly used standard measure for system evaluation. It consists of ten questions, with odd-numbered questions being positive and even-numbered questions being negative. The system usability scale (SUS) scores were calculated by subtracting 1 from the odd-numbered question scores and subtracting the even-numbered question scores from 5, summing these scores, and multiplying by 2.5 to obtain a score out of 100. A score of 70 or above indicated that the system was acceptable.

Subjective emotional evaluation was conducted using the Affect Grid [48], which evaluates the nature of emotions on a two-dimensional coordinate system, typically on a 9 × 9 grid. The vertical axis evaluates alertness (alertness-sleepiness), and the horizontal axis evaluates emotional value (pleasure-displeasure). In this experiment, the participants answered the Affect Grid before and after the experiment on the first day to evaluate the changes in subjective emotions.

4.6. Result

4.6.1. Pre-Test Score and TOEIC Score

None of the 27 participants answered any of the questions correctly during the pre-test. In this experiment, participants were randomly assigned to either the baseline or proposed method conditions. Of the 27 participants, 4 did not have English test scores. To validate data normality, we used the Shapiro–Wilk normality test, which indicated that the TOEIC scores (with TOEFL and IELTS data converted using a conversion table) for both the baseline ($p = 0.88$) and proposed method ($p = 0.18$) were normally distributed. A two-tailed *t*-test was used for statistical analyses. There was no significant difference between the participants' TOEIC scores for the proposed method ($M = 714.82$, $SD = 145.47$) and those for the baseline method ($M = 675.83$, $SD = 129.79$) (t[20] = 0.6456, ns, d = 0.28).

4.6.2. Vocabulary Test Score and Forgetting Rate

Figure 9 (left) shows that under the baseline approach, the average number of correct answers decreased from 8.15 words to 3.69 words in 1 week, that is, the average forgetting rate was 4.46 words. However, in the proposed method, the average number of correct answers decreased from 6.93 words to 5.00 words in one week, that is, the average forgetting rate was 1.93 words.

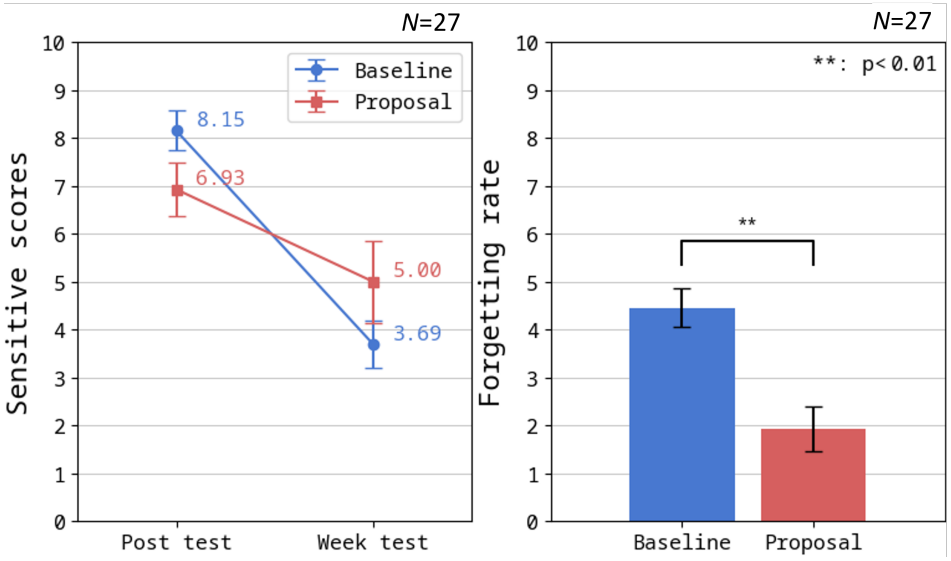


Figure 9. (**Left**) Average scores of the vocabulary test immediately and one week later. (**Right**) Comparison of the forgetting rate calculated by subtracting the immediate scores from the scores of the assessment conducted a week later. The error bar means standard error (SE).

To compare the forgetting rate between the baseline and proposed methods, their respective scores were subtracted from one another for each subject (Figure 9). To validate data normality, we used the Shapiro–Wilk normality test, which indicated that the forgetting rate data for both the baseline (p = 0.50) and the proposal (p = 0.10) were normally distributed, and we used a two-tailed t-test for the statistical analyses. The participants' forgetting rate score for the proposed method (M = 1.93, SD = 1.79) was significantly lower than that of the baseline method (M = 4.46, SD = 1.45): t[24] = 3.90, $p < 0.01$, d = 1.69). This suggests that the participants were able to memorize and retain more English language words using the proposed method than with the baseline method (RQ1).

4.6.3. Correlation Between Episodic Scores and Vocabulary Scores

After the delayed test (one week later), participants who experienced IVR were given a questionnaire to check which fragments of episodic memory they could recall. Figure 10 shows the recall rate of word episodes when only English words were presented. The episode score in Figure 10 refers to the total score based on seven categories (background, objects, events, body parts, sound effects, pronunciation, and instructions), with each category counting as one point. As each participant memorized 10 words, the maximum total episode score was 70 points.

The left side of Figure 10 is a scatter plot showing the relationship between the episode score that each participant who experienced IVR remembered and their word test score one week later. The correlation coefficient is 0.672, which is significant ($F(1,12) = 9.89$, $p < 0.01$). The right side of Figure 10 shows a scatter plot of the relationship between the episode score for each word and the correct answer rate in the word test one week later. The correlation coefficient is 0.700, which is significant ($F(1,8) = 7.69, p < 0.05$) (RQ2).

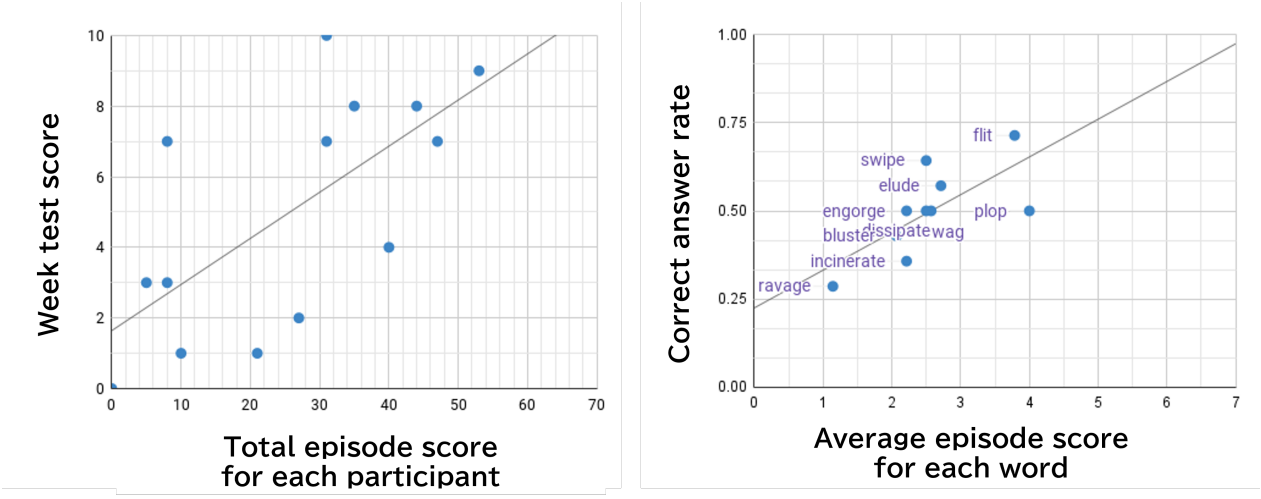


Figure 10. (**Left**) Total episode scores for ten words and test scores of the assessment conducted after a week for each participant. (**Right**) Correlation between the average episode score for each word and the accuracy rate of the test conducted after a week.

4.6.4. NASA-TLX Score and SUS Score

To validate the normality of the data, we used a Shapiro–Wilk test, which indicated that some of the baseline data were not normally distributed: MD ($p = 0.11$), PD ($p < 0.01$), TD ($p < 0.01$), performance ($p < 0.01$), stress ($p < 0.01$), effort ($p = 0.37$), and subjective workload (AWWL) ($p = 0.31$). Similarly, for the baseline data, MD ($p = 0.20$), PD ($p < 0.01$), TD ($p < 0.01$), performance ($p < 0.05$), stress ($p = 0.47$), effort ($p < 0.05$), and subjective workload (AWWL) ($p = 0.14$) were not normally distributed. Therefore, we used the non-parametric Mann–Whitney U test. No significant differences were found in any of the categories (Figure 11 left).

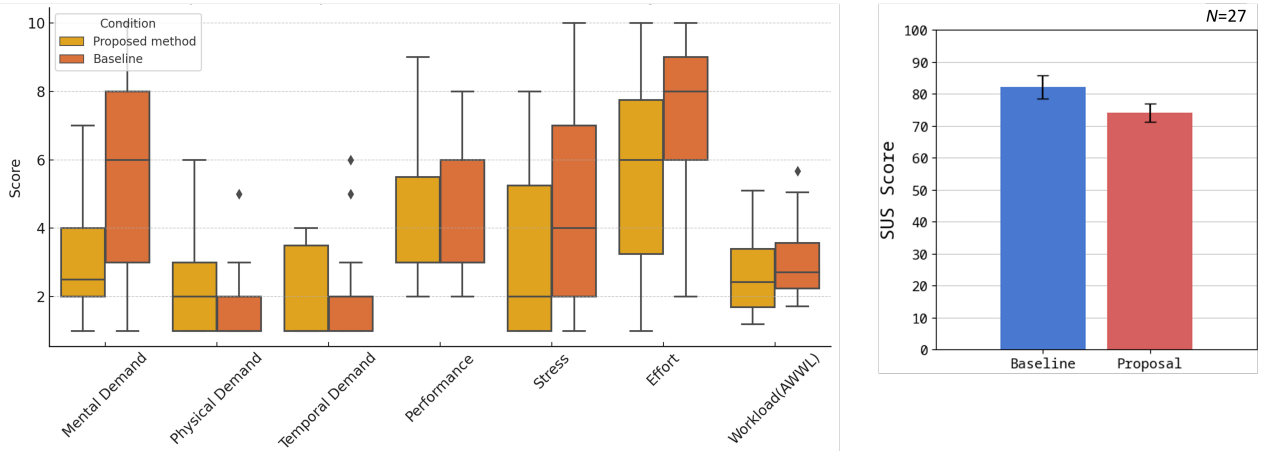


Figure 11. (**Left**) Median scores for each item of the NASA-TLX, where higher scores indicate a greater subjective workload for the user. (**Right**) Average system usability scale (SUS) scores, where a score of 68 or above generally indicates good usability.

To validate data normality, we also performed a Shapiro–Wilk normality test, which indicated that the SUS scores for both the baseline (p = 0.37) and the proposed condition (p = 0.18) were normally distributed. Therefore, a two-tailed t-test was used for statistical analysis. There was no significant difference between the SUS scores of the flashcard condition ($M = 82.12, SD = 12.97$) and NariTan condition ($M = 74.11, SD = 11.04$) ($t(23) = 1.6561, ns(0.10 < p)$). Although no significant difference was found between the conditions, both had an average score of >68 points. Because a score of 68 or above is considered an indicator of good usability, it can be said that the proposed method had good usability that was comparable to that of the commercially available Flashcard Deluxe app (Figure 11 right).

4.6.5. Emotional Changes

To test the normality of the emotional change data before and after learning, a Shapiro–Wilk test was conducted. Normality was confirmed for arousal ($p = 0.46$) and valence ($p = 0.07$) in the proposed method and for arousal ($p = 0.06$) in the baseline. However, normality was not confirmed for valence at the baseline ($p < 0.05$). Therefore, a Mann–Whitney U test was performed, and no significant differences were found between the proposed method and baseline for either arousal or valence (RQ3) (Figure 12).

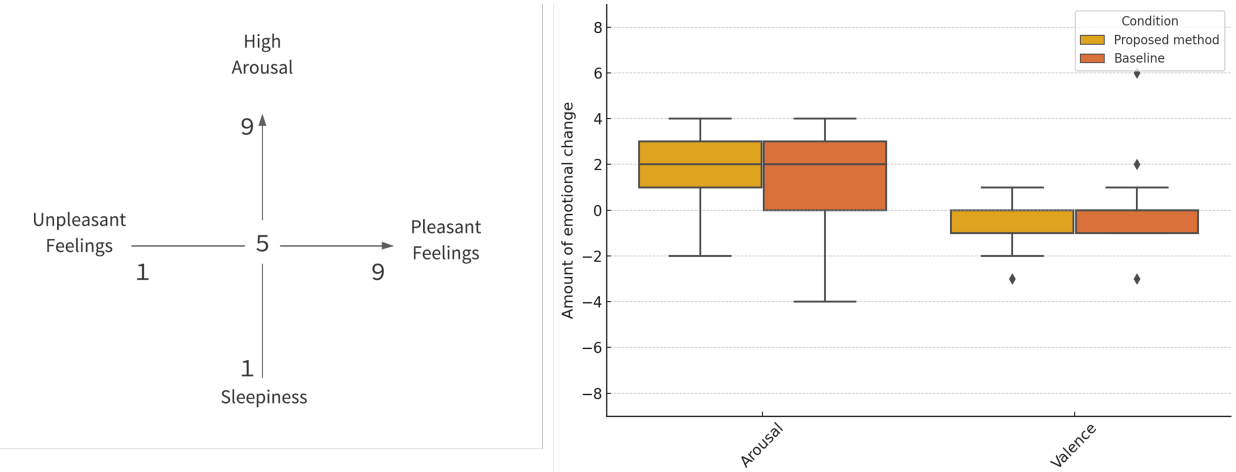


Figure 12. (**Left**) An illustration explaining the meaning of the values in the Affect Grid used to quantify emotions. (**Right**) A comparison of the median values for emotional changes on the Affect Grid before and after the experiment.

5. Experiment 2: Verbs Covered by the Proposed Method

We investigated verb retention using the proposed method. Specifically, we examined how many avatars NariTan needed to use to sufficiently learn English verbs and how many English verbs NariTan could include when considering actual use cases. In this experiment, we used 543 words extracted from the "animals" category of WordNet [49] to select non-human avatars for use in NariTan learning. For the selected English words, verbs that could be learned with the corresponding avatar were chosen by following the same procedure used for word selection in Section 3.

5.1. Procedure

Two datasets were used in this experiment. First, we verified the extent to which a given number of avatars could be used to learn English words until saturation. By using the 25,000 English words included in the BNC/COCA word family [39], we investigated the coverage rate of BNC/COCA based on the number of avatars used in NariTan, categorized by the frequency level (Figure 13).

Next, we examined how many of the 25,000 words in BNC/COCA could be covered by learning with NariTan while changing the avatars daily for one or four weeks, assuming a real-use case.

As the second dataset, we used a combination of five corpora from the New General Service List Project https://www.newgeneralservicelist.com (accessed on 11 September 2024) to conduct a similar verification as with BNC/COCA and investigated the nature of the words covered by NariTan. The five corpora included 2809 words from the New General Service List 1.2 (NGSL) https://www.newgeneralservicelist.com/new-general-service-list (accessed on 11 September 2024), 721 words from NGSL Spoken 1.2 (NGSL-S) https://www.newgeneralservicelist.com/ngsl-spoken (accessed on 11 September 2024), 957 words from the New Academic Word List 1.2 (NAWL) https://www.newgeneralservicelist.com/new-general-service-list-1 (accessed on 11 September 2024), 1250 words from the TOEIC Service List 1.2 (TSL) https://www.newgeneralservicelist.com/toeic-service-list (accessed on 11 September 2024), and 1700 words from the Business Service List 1.2 (BSL) https://www.newgeneralservicelist.com/business-service-list (accessed on 11 September 2024), totaling 7437 words. However, because BNC/COCA is the basis for word selection, words not included in the 25,000 words of BNC/COCA were excluded from the dataset.

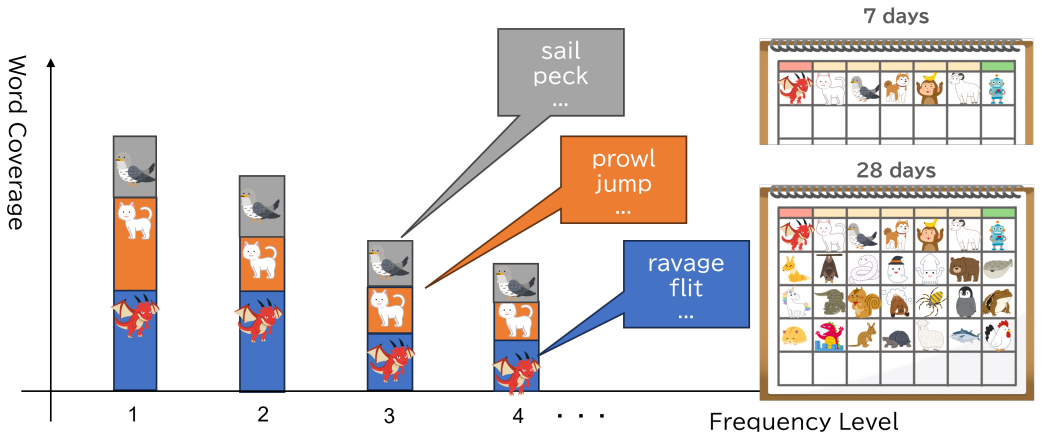


Figure 13. Illustrated steps for graphing the word coverage rate. Assuming that a different avatar is used each day, and the unique verbs of that avatar are learned, the goal is to visualize how much vocabulary can be learned in the end.

5.2. Result

Figure 14 shows the coverage rate of BNC/COCA when learning with NariTan for one or four weeks. It was shown that more than 30% of the words with frequency level 5 or

below (equivalent to CEFR C1) could be covered after one week of learning and that more than 40% could be covered after four weeks of learning.

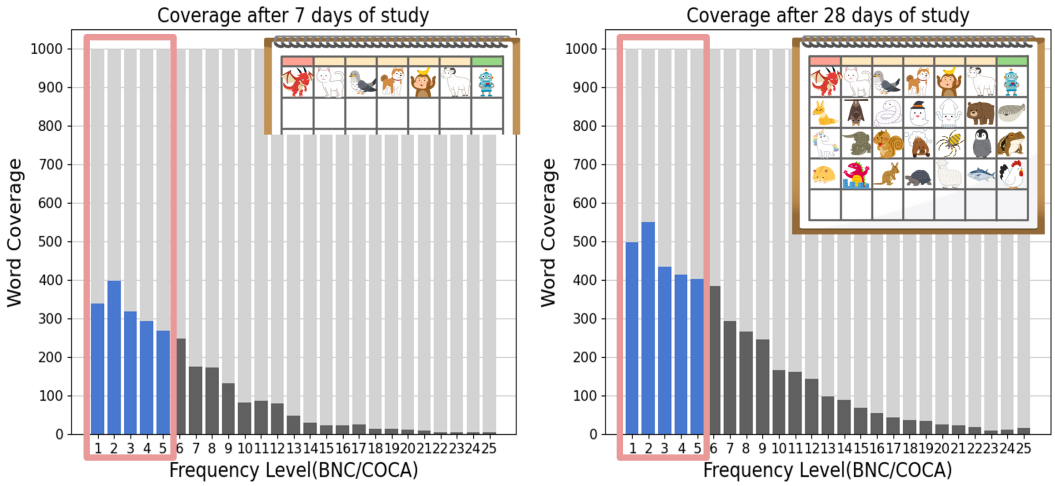


Figure 14. Coverage rate results for the British National Corpus (BNC)/Corpus of Contemporary American (COCA) datasets. The left side shows the vocabulary that can be learned when using seven types of animal avatars (assuming a different avatar each day for a week) and the right side shows what can be learned when using 28 types of animal avatars (assuming a different avatar each day for a month). Dark blue and gray represent learnable vocabularies, whereas the light gray indicates the remaining vocabulary that cannot be covered.

Figure 15 shows the coverage rate of each NGSL dataset when learning with NariTan for one week or four weeks. It was shown that approximately 40% of NGSL could be covered after one week of learning, whereas approximately 20–30% of the other datasets could be covered. After four weeks of learning, approximately 50% of NGSL and NGSL Spoken were covered, whereas approximately 30–40% of the other datasets were covered.

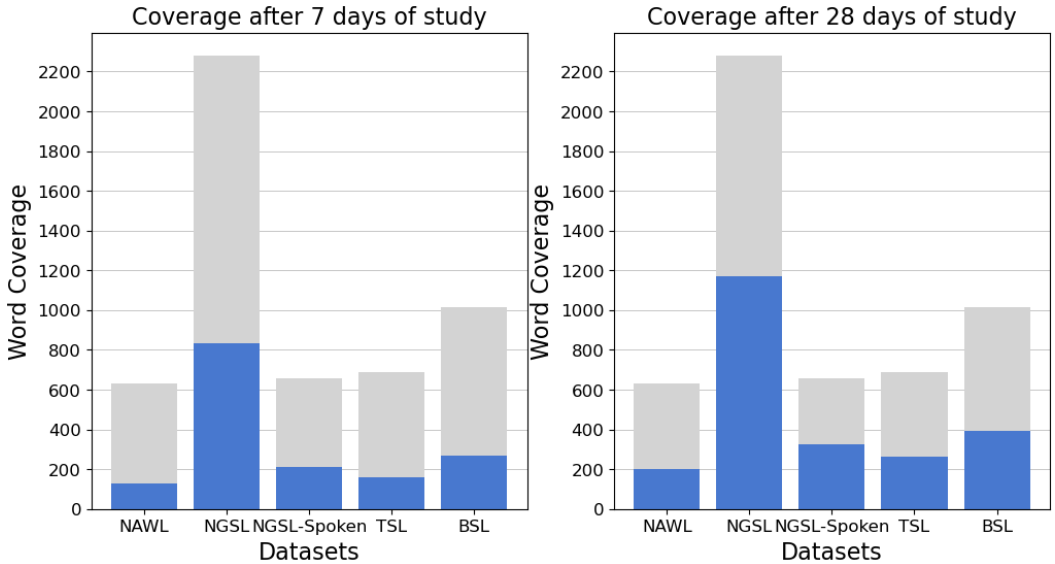


Figure 15. Coverage rate results for the New General Service List 1.2 (NGSL) dataset. Dark blue represents the learnable vocabulary, and light gray indicates the remaining vocabulary that cannot be covered.

In this experiment, we considered the upper limit to be 543 avatars, corresponding to the number of words in the animal category of WordNet. Figure 16 shows the coverage rate results when the number of avatars is increased up to this upper limit. This figure shows the coverage rate for each of the 25 levels, where each level is a group of 1000 words ordered

by frequency from the 25,000 words in the BNC/COCA dataset. It can be observed that words at higher frequency levels were mostly covered, whereas those at lower frequency levels were more difficult to cover. Therefore, this method is not suitable for covering low-frequency words, and a different learning method is required to cover these vocabulary levels. Similarly, Figure 17 shows the theoretical coverage rate of NGSL. It can be seen that approximately 80% of the vocabulary in NGSL and NGSLSpoken can be covered; however, for other datasets, the coverage rate plateaued at approximately 50%.

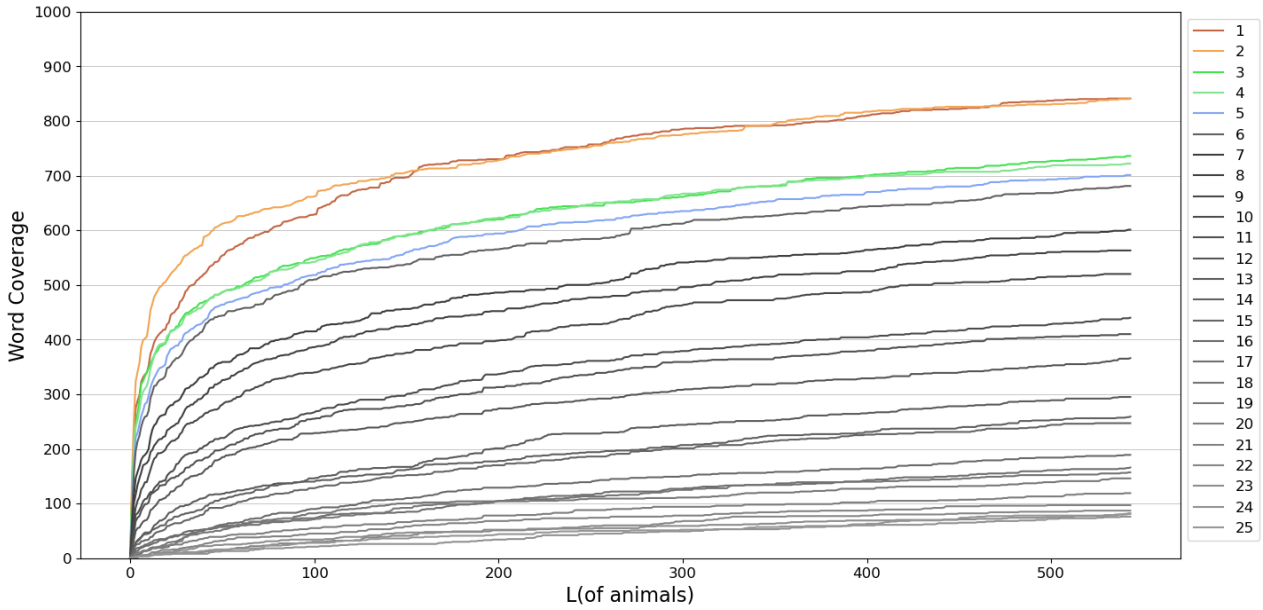


Figure 16. The results showing which level of words from the 25,000 words in the BNC/COCA dataset can theoretically be covered by this method. This indicates the coverage rate when learning words by assuming the role of all avatars in WordNet's animal category. There are 25 series, each containing 1000 words.

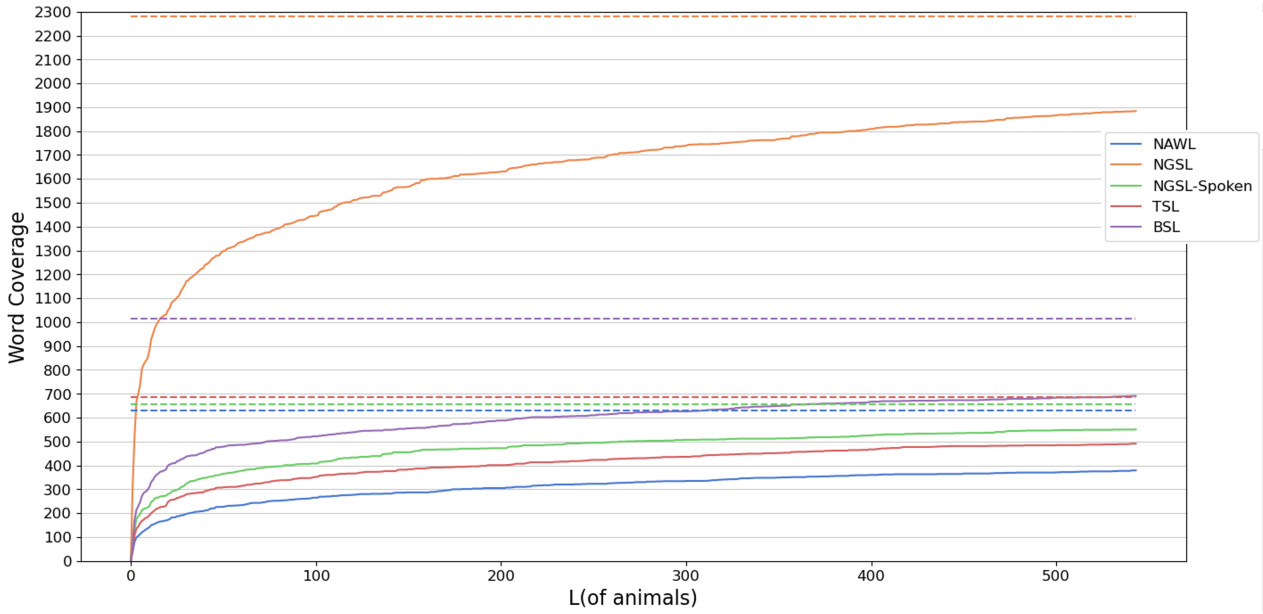


Figure 17. Results showing the genres of words from the NGSL dataset can be theoretically covered by this method. Since the number of words varies for each series, the number of words in each series is indicated with a dotted line.

6. Discussion and Limitation

6.1. Word Tests and Episodic Memory Trace

The proposed method demonstrated a significantly lower forgetting rate than the baseline. Similarly, prior studies [3,12,32,43,44,50] on experiential English vocabulary learning systems have shown a tendency toward reduced forgetting rates compared to flashcards. However, because the range of words covered by each system varies, it is necessary to combine the various methods when expanding vocabulary using experiential learning systems. In this system, instructions and word presentations were always displayed within the learner's field of view, but some participants absorbed in the IVR experience either did not read or ignored the displayed content. Focusing too much on actions or feedback effects might lead to using the IVR space solely for enjoyment without contributing to learning. Although making the experiment enjoyable is a good approach, the strong appeal of such stimuli could also become an issue.

A positive correlation was observed between the episodic memory and vocabulary test scores one week later. This suggests that the cognitive model of bilingual dual-coding theory may have been constructed as learning progressed. For words with lower episodic memory scores, multiple factors may be involved, such as fewer bodily movements and lower consistency between words and actions. Because learning in IVR involves evaluating a system with multiple sensory stimuli, identifying specific factors would require more controlled experiments in a strictly regulated environment. The experiment conducted in this study was an evaluation of the proposed system and did not compare the effectiveness of human avatars with that of non-human avatars. To verify the effectiveness of non-human avatars, it is necessary to compare the learning effects of using both human and non-human avatars in the same IVR environment.

Related studies have reported vocabulary learning and memory evaluation experiments using IVR [12,32,51]. However, these results do not always show an effect that promotes memory retention. For example, Mizuho et al. examined the context dependency on memory using IVR. However, they reported that the results sometimes contradicted those observed in the real-world context-dependency of memory. Mizuho et al. mentioned the lack of immersion as a factor, suggesting that, when sufficient content immersion is not achieved, the experience of wearing an HMD in the lab tends to be more memorable than the IVR content itself, making the presence or absence of immersion an important factor [51]. Recent studies on immersion in IVR have defined three immersion components: place, copresence, and placeability illusions [14]. In the proposed method, the place illusion is fulfilled by allowing users to freely walk around in the IVR space and the plausibility illusion is achieved through avatar-specific actions. It is possible that sufficient immersion provided by these elements contributed to memory retention.

6.2. Evaluations Beyond Word Tests

There were no significant differences in alertness-related emotional changes. After the flashcard-based experiment, the participants' feedback on emotional changes included comments such as "no emotional movement at all", "felt sleepy at first because the experiment was in the morning but felt alert after the experiment", "felt positive because I could remember the words surprisingly well", and "felt positive because I could answer the test unexpectedly well". However, in the proposed method, many participants reported "discomfort due to IVR sickness" and "fatigue from moving their body". The results of the Affect Grid might have included noise from the emotional changes due to the tests. Measuring emotions while learning remains a challenge for future research. Additionally, many participants who reported a shift to unpleasant emotions attributed this to IVR sickness. One participant stopped the experiment because of IVR sickness. Because IVR sickness can hinder learning, it is essential to address it in IVR second-language learning systems.

6.3. Coverage of English Words

In the experiment, based on actual use cases, it was shown that it is possible to teach approximately 30% of CEFR C1 level words in one week and approximately 40% in four weeks. By iterating until the vocabulary coverage rate saturates, with the number of avatars on the horizontal axis, it is suggested that more than 70% of the 25,000 words in the BNC/COCA dataset with a CEFR level of C1 or below can be covered. Furthermore, the coverage rate for each NGSL dataset did not vary significantly across the datasets. Furthermore, the coverage rate for each NGSL dataset did not significantly vary across the datasets. Therefore, NariTan could potentially achieve a certain level of coverage with any dataset. However, learning business English (BSL) with non-human avatars can result in unusual scenarios.

6.4. Use of the Avatar

We used a dragon avatar as an example of a non-human avatar. There is room for discussion regarding the learning effects of other avatars. In the post-experiment interviews, participants were asked about other avatars they would like to use. Responses included aliens, robots (which have similar structures but are completely different from humans), worms, octopuses, and birds (which have structures that are completely different from those of humans). In addition, optimal control methods for avatars in IVR must be investigated. In this study, avatar control was implemented with 3-point tracking using an HMD and two-hand controllers. If the aim is to faithfully embody a quadrupedal (i.e., moving on all fours) or bipedal creature, full-body tracking can be considered for avatar control. Recently, some studies implemented independent control of quadrupedal avatars and avatars with completely different structures from those of humans [52]. However, these implementation methods that enhance the freedom of user control beyond that possible in 3-point tracking increase the implementation cost and reduce portability. Hence, it is necessary to select the implementation method after considering actual use cases.

7. Conclusions

We propose a system in which users embody non-human avatars in IVR spaces to learn vocabulary. We developed new interactions for this learning system and created unique vocabulary-selection methods tailored to each avatar. We validated the effectiveness of this system through quantitative assessments, including English vocabulary and episodic memory tests. The results showed a significant reduction in the forgetting rate of the English vocabulary after one week. We also found a positive correlation between the amount of information retained in episodic memory and scores on the English vocabulary test one week later. Moreover, we explored the vocabulary range that can be acquired with this learning method using LLMs. We found that, for vocabulary below CEFR level C1, 30% could be covered by changing the avatar daily for seven days and 40% could be covered by changing the avatar daily for 28 d.

In Experiment 1, the evaluation was limited to long-term memory scores one week later. However, when considering actual vocabulary learning, it is also important to conduct experiments that investigate memory fragmentation after one month or longer and clarify the differences between IVR and traditional flashcard methods. Additionally, to further strengthen vocabulary retention, it is necessary to explore new learning methods and applications, including reviews. In particular, when using IVR, episodic memory can be heavily utilized, making it crucial to design applications that differ from traditional flashcard methods that focus primarily on semantic memory.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/mti8100093/s1.

Author Contributions: Conceptualization, S.F., K.S. and Y.N.; methodology, S.F., K.S. and Y.N.; software, K.S.; validation, S.F., K.S. and Y.N.; formal analysis, S.F., K.S. and Y.N.; investigation, S.F., K.S. and Y.N.; resources, S.F., K.S. and Y.N.; data curation, S.F., K.S. and Y.N.; writing—original draft

preparation,S.F., K.S. and Y.N.; writing—review and editing, S.F., K.S. and Y.N.; visualization, S.F., K.S. and Y.N.; supervision, S.F. and Y.N.; project administration, S.F., K.S. and Y.N.; funding acquisition, S.F. All authors have read and agreed to the published version of the manuscript.

Funding: Japan Society for the Promotion of Science: JSPS KAKENHI No. JP20K19936 and No. JP24834281.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Kyushu University (ISEE 2023-24, approval date: 25 January 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data collected during the study are available from the corresponding author reasonable upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nation, I.S.P. How Large a Vocabulary Is Needed For Reading and Listening? *Can. Mod. Lang. Rev.* **2006**. *500*, 59–82. [CrossRef]
2. van Zeeland, H.; Schmitt, N. Lexical Coverage in L1 and L2 Listening Comprehension: The Same or Different from Reading Comprehension? *Appl. Linguist.* **2013**, *34*, 457–479. [CrossRef]
3. Nishida, Y.; Kusunoki, F.; Hiramoto, M.; Mizoguchi, H. Learning by Doing: Space-Associate Language Learning Using a Sensorized Environment. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 3636–3641.
4. Engelkamp, J.; Krumnacker, H. Imaginale und motorische Prozesse beim Behalten verbalen Materials. *Proc. Z. Exp. Angew. Psychol.* **1980**, *27*, 511–533.
5. Cohen, R.L. On the generality of some memory laws. *Scand. J. Psychol.* **1981**, *22*, 267–281. [CrossRef]
6. Kormi-Nouri, R.; Nyberg, L.; Nilsson, L.G. The effect of retrieval enactment on recall of subject-performed tasks and verbal tasks. *Mem. Cogn.* **1994**, *22*, 723–728. [CrossRef]
7. Engelkamp, J.; Zimmer, H. Free recall and organization as a function of varying relational encoding in action memory [Electronic version]. *Psychol. Res.* **2002**, *66*, 91–98. [CrossRef]
8. Saltz, E.; Donnenwerth-Nolan, S. Does motoric imagery facilitate memory for sentences? A selective interference test. *J. Verbal Learn. Verbal Behav.* **1981**, *20*, 322–332. [CrossRef]
9. Parong, J.; Mayer, R.E. Cognitive and affective processes for learning science in immersive virtual reality. *J. Comput. Assist. Learn.* **2021**, *37*, 226–241. [CrossRef]
10. Isarida, T.; Isarida, T.K. Environmental context effects of background color in free recall. *Mem. Cogn.* **2007**, *35*, 1620–1629. [CrossRef]
11. Wälti, M.J.; Woolley, D.G.; Wenderoth, N. Reinstating verbal memories with virtual contexts: Myth or reality? *PLoS ONE* **2019**, *14*, e0214540. [CrossRef]
12. Vázquez, C.; Xia, L.; Aikawa, T.; Maes, P. Words in Motion: Kinesthetic Language Learning in Virtual Reality. In Proceedings of the 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), Mumbai, India, 9–13 July 2018; pp. 272–276. [CrossRef]
13. Ratcliffe, J.; Ballou, N.; Tokarchuk, L. Actions, not gestures: Contextualising embodied controller interactions in immersive virtual reality. In Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology, Osaka, Japan, 8–10 December 2021; pp. 1–11.
14. Brübach, L.; Westermeier, F.; Wienrich, C.; Latoschik, M.E. Breaking Plausibility Without Breaking Presence—Evidence For The Multi-Layer Nature Of Plausibility. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 2267–2276. [CrossRef] [PubMed]
15. Paivio, A.; Desrochers, A. A dual-coding approach to bilingual memory. *Can. J. Psychol. Can. Psychol.* **1980**, *34*, 388. [CrossRef]
16. Paivio, A. *Imagery and Verbal Processes*; Psychology Press: London, UK, 1971.
17. Paivio, A.; Csapo, K. Picture superiority in free recall: Imagery or dual coding? *Cogn. Psychol.* **1973**, *5*, 176–206. [CrossRef]
18. Arnedt, C.S.; Gentile, J.R. A test of dual coding theory for bilingual memory. *Can. J. Psychol. Can. Psychol.* **1986**, *40*, 290. [CrossRef]
19. Mayer, R. Systematic Thinking Fostered by Illustrations in Scientific Text. *J. Educ. Psychol.* **1989**, *81*, 240–246. [CrossRef]
20. Harp, S.; Mayer, R. How Seductive Details Do Their Damage: A Theory of Cognitive Interest in Science Learning. *J. Educ. Psychol.* **1998**, *90*, 414–434. [CrossRef]
21. Mayer, R.; Anderson, R. The Instructive Animation: Helping Students Build Connections Between Words and Pictures in Multimedia Learning. *J. Educ. Psychol.* **1992**, *84*, 444–452. [CrossRef]
22. Zhu, Y.; Wang, Y.; Yu, C.; Shi, S.; Zhang, Y.; He, S.; Zhao, P.; Ma, X.; Shi, Y. ViVo: Video-augmented dictionary for vocabulary learning. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 5568–5579.
23. Hong, R.; Wang, M.; Xu, M.; Yan, S.; Chua, T.S. Dynamic captioning: Video accessibility enhancement for hearing impairment. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 421–430.

24. Hu, Y.; Kautz, J.; Yu, Y.; Wang, W. Speaker-following video subtitles. *ACM Trans. Multimed. Comput. Commun. Appl.* **2015**, *11*, 1–17. [CrossRef]
25. Brown, A.; Jones, R.; Crabb, M.; Sandford, J.; Brooks, M.; Armstrong, M.; Jay, C. Dynamic subtitles: The user experience. In Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video, Brussels, Belgium, 3–5 June 2015; pp. 103–112.
26. Tulving, E. *Elements of Episodic Memory*; Oxford University Press: Oxford, UK, 1983.
27. Macedonia, M.; Knösche, T.R. Body in Mind: How Gestures Empower Foreign Language Learning. *Mind Brain Educ.* **2011**, *5*, 196–211. [CrossRef]
28. Macedonia, M.; Müller, K.; Friederici, A.D. The impact of iconic gestures on foreign language word learning and its neural substrate. *Hum. Brain Mapp.* **2011**, *32*, 982–998. [CrossRef]
29. Krönke, K.M.; Mueller, K.; Friederici, A.D.; Obrig, H. Learning by doing? The effect of gestures on implicit retrieval of newly acquired words. *Cortex* **2013**, *49*, 2553–2568. [CrossRef] [PubMed]
30. Tellier, M. The effect of gestures on second language memorisation by young children. *Gesture* **2008**, *8*, 219–235. [CrossRef]
31. Zhang, X.; Zuber, S. The effects of language and semantic repetition on the enactment effect of action memory. *Front. Psychol.* **2020**, *11*, 515. [CrossRef] [PubMed]
32. Ebert, D.; Gupta, S.; Makedon, F. Ogma: A virtual reality language acquisition system. In Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments, Corfu Island, Greece, 29 June–1 July 2016; pp. 1–5.
33. Wakefield, E.M.; Hall, C.; James, K.H.; Goldin-Meadow, S. Gesture for generalization: Gesture facilitates flexible learning of words for actions on objects. *Dev. Sci.* **2018**, *21*, e12656. [CrossRef] [PubMed]
34. Hostetter, A.B.; Pouw, W.; Wakefield, E.M. Learning From Gesture and Action: An Investigation of Memory for Where Objects Went and How They Got There. *Cogn. Sci.* **2020**, *44*, e12889. [CrossRef]
35. Vargas, M.F.; Fribourg, R.; Bates, E.; McDonnell, R. Now I Wanna Be a Dog: Exploring the Impact of Audio and Tactile Feedback on Animal Embodiment. In Proceedings of the 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, 16–20 October 2023; pp. 912–921. [CrossRef]
36. Li, S.; Gu, X.; Yi, K.; Yang, Y.; Wang, G.; Manocha, D. Self-Illusion: A Study on Cognition of Role-Playing in Immersive Virtual Environments. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 3035–3049. [CrossRef]
37. Ogawa, N.; Narumi, T.; Kuzuoka, H.; Hirose, M. Do You Feel Like Passing Through Walls?: Effect of Self-Avatar Appearance on Facilitating Realistic Behavior in Virtual Environments. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; CHI '20; pp. 1–14. [CrossRef]
38. OpenAI. Introducing ChatGPT. 2022. Available online: https://openai.com/blog/chatgpt (accessed on 11 September 2024).
39. Nation, P. Vocabulary Analysis Programs. 2017. Available online: https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs (accessed on 11 September 2024).
40. Persons, D.L.I. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*; Cambridge University Press: Cambridge, UK, 2001.
41. Consortium, B. The British National Corpus, XML Edition. 2007. Available online: https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554?show=full (accessed on 11 September 2024).
42. Davies, M. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Lit. Linguist. Comput.* **2010**, *25*, 447–464. [CrossRef]
43. Fukushima, S. EmoTan: Enhanced flashcards for second language vocabulary learning with emotional binaural narration. *Res. Pract. Technol. Enhanc. Learn.* **2019**, *14*, 16. [CrossRef]
44. Shimizu, K.; Fukushima, S.; Naemura, T. Effects of Binaural Audio on English Vocabulary Learning. In Proceedings of the 30th International Conference on Computers in Education Conference, Kuala Lumpur, Malaysia, 28 November–2 December 2022; Volume 2, pp. 665–667.
45. Nakata, T.; Suzuki, Y.; He, X.S. Costs and Benefits of Spacing for Second Language Vocabulary Learning: Does Relearning Override the Positive and Negative Effects of Spacing? *Lang. Learn.* **2023**, *73*, 799–834. [CrossRef]
46. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.
47. Brooke, J. Sus: A "quick and dirty'usability. *Usability Eval. Ind.* **1996**, *189*, 189–194.
48. Russell, J.A.; Weiss, A.; Mendelsohn, G.A. Affect grid: A single-item scale of pleasure and arousal. *J. Pers. Soc. Psychol.* **1989**, *57*, 493. [CrossRef]
49. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
50. Hautasaari, A.; Hamada, T.; Ishiyama, K.; Fukushima, S. VocaBura: A Method for Supporting Second Language Vocabulary Learning While Walking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *3*, 1–23. [CrossRef]

51. Mizuho, T.; Narumi, T.; Kuzuoka, H. Exploratory Study on the Reinstatement Effect Under 360-Degree Video-Based Virtual Environments. In Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology, Christchurch, New Zealand, 9–11 October 2023; Association for Computing Machinery: New York, NY, USA, 2023; VRST '23. [CrossRef]
52. Krekhov, A.; Cmentowski, S.; Krüger, J. Vr animals: Surreal body ownership in virtual reality games. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts, Melbourne, VIC, Australia, 28–31 October 2018; pp. 503–511.