applied sciences

MDPI

Article

MFIRA: Multimodal Fusion Intent Recognition Algorithm for AR Chemistry Experiments

Zishuo Xia [1,2], Zhiquan Feng [1,2,*], Xiaohui Yang [1,2], Dehui Kong [1,2] and Hong Cui [1,2]

[1] School of Information Science and Engineering, University of Jinan, Jinan 250022, China; 17860631906@163.com (Z.X.); kindhui62@gmail.com (D.K.); cuihong19980115@gmail.com (H.C.)
[2] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China
* Correspondence: ise_fengzq@ujn.edu.cn

Abstract: The current virtual system for secondary school experiments poses several issues, such as limited methods of operation for students and an inability of the system to comprehend the users' operational intentions, resulting in a greater operational burden for students and hindering the goal of the experimental practice. However, many traditional multimodal fusion algorithms rely solely on individual modalities for the analysis of users' experimental intentions, failing to fully utilize the intention information for each modality. To rectify these issues, we present a new multimodal fusion algorithm, MFIRA, which intersects and blends intention probabilities between channels by executing parallel processing of multimodal information at the intention layer. Additionally, we developed an augmented reality (AR) virtual experiment platform based on the Hololens 2, which enables students to conduct experiments using speech, gestures, and vision. Employing the MFIRA algorithm, the system captures users' experimental intent and navigates or rectifies errors to guide students through their experiments. The experimental results indicate that the MFIRA algorithm boasts a 97.3% accuracy rate in terms of interpreting users' experimental intent. Compared to existing experimental platforms, this system is considerably more interactive and immersive for students and is highly applicable in secondary school experimental chemistry classrooms.

Keywords: multimodal fusion; virtual experiment; intelligent teaching; human-computer interaction; intention understanding

Citation: Xia, Z.; Feng, Z.; Yang, X.; Kong, D.; Cui, H. MFIRA: Multimodal Fusion Intent Recognition Algorithm for AR Chemistry Experiments. *Appl. Sci.* **2023**, *13*, 8200. https://doi.org/10.3390/app13148200

Academic Editor: Dimitris Mourtzis

Received: 23 June 2023
Revised: 10 July 2023
Accepted: 11 July 2023
Published: 14 July 2023

1. Introduction

Currently, in remote areas of Northwest China, many secondary schools face challenges due to limited staff, making it difficult for teachers to teach all students simultaneously [1] Chemical experiments have the characteristics of high reagent contamination and a relatively dangerous operational process. Consequently, some students ignore the main points of the experiments, leading to dangerous and unregulated behavior. Many schools cancel students' practical operation of some chemistry experiments to avoid this danger, which in turn weakens the teaching effect and makes it more difficult for students to practice and understand the experiment's contents. Therefore, it is necessary to design an intelligent, operable, and low-danger experiment platform system for secondary school experiments in order to solve the aforementioned problems.

The commercially available teaching experiment platforms can be divided into two types. The first is a web-based virtual experimental platform [2], which can solve the problems of serious reagent contamination and high danger in real experimental teaching. However, most of the existing web-based virtual experimental platforms use mouse and keyboard input and monitor output to conduct experiments, which greatly reduces the students' sense of operation. The second is to use virtual reality technology to establish VR or AR experimental platforms in order to allow students to feel the experiment immersively [3]. However, the current VR or AR experimental platform is operated in a

single mode of interaction, or equipped with a handle and other equipment to assist in the operation, requiring students to remember a large number of operating commands. In addition, they are rarely seen to be able to actively obtain and utilize the students' experimental intentions, which leads to a high operational load for students and makes it difficult to achieve the purpose of experimental teaching.

To address the limitations of the above schemes, we chose the Microsoft Hololens 2 device as the platform of our AR chemistry experiment system, and we designed a Hololens-based AR experimental system that can fuse information from multiple modalities, and designed an MFIRA algorithm to obtain students' experimental intentions by fusing information from multiple modalities. Students wore Hololens 2 device, and the system guided them based on the experimental intentions obtained by the algorithm, assisting them in completing the experiment.

The main contributions of this paper are as follows:

- For the limitations of the traditional virtual experiment system with a single mode of interaction and a high memory load:

We designed an AR virtual experiment system based on Hololens 2. Users can complete an entire experiment naturally by speech, gestures, and vision. The system collects the information of the three modalities in real time during the experiment, senses the user's operation intention, and guides or corrects their operation; this system significantly improves the user's operation immersion and effectively reduces the user's operation load.

- To address the limitations of traditional multimodal fusion algorithms:

In the context of intelligent experiments, traditional multimodal fusion algorithms, by nature, use unimodal information to analyze the user's experimental intent serially, while failing to fuse the intent of multiple channels in parallel to infer the user's current behavior. Therefore, we innovatively propose a multimodal fusion algorithm, MFIRA, that extracts the user's experimental intent by parallel processing of multimodal information and fusion at the intent layer. This algorithm utilizes the user's experimental intent to complete AR chemistry experiments by performing cross-fusion between intent probability sequences via parallel processing of information from three channels: speech, gesture, and vision. The experimental results show that the MFIRA algorithm achieved an accuracy of 97.3% in understanding the user's experimental intention.

This article is structured as follows: Section 2 comprehensively discusses the related work, Section 3 describes the construction of the virtual experimental system, Section 4 introduces the general framework of the article and the multimodal fusion algorithm, Section 5 analyzes and discusses the experimental results, and Section 6 provides an overview of the conclusions.

2. Related Work

2.1. Virtual Experiment

Virtual experiments utilize computer and virtual reality technologies to build a digital experimental environment for experimentation and analysis, thereby simulating and substituting for real experiments. These experiments are frequently characterized by 3D graphics and interactive operations, which can simulate various aspects of real experiments and their results while performing multiple experimental operations and experimental analysis.

As computer multimedia graphic image technology continues to develop, virtual experiments are being incorporated more and more into education. Early forms of virtual experiment teaching relied on flat web technology. For example, Aljuhani et al. [4] built a virtual laboratory platform on a web platform which allowed users to conduct virtual experiments with a mouse. Morozov et al. [5] developed a 2D virtual laboratory, MarSTU, which improved students' understanding of some chemical experiments. More recently, virtual reality technologies, including VR, AR, and MR, have become the mainstream in virtual experiments. For instance, Tingfu et al. [6] proposed a 3D interactive framework

for virtual chemistry experiments based on VRML, which uses 3D modeling to display the real experimental environment and enhance students' immersion. This framework aims to address the lack of interactivity of most chemistry experiment systems using 2D technology. Other examples include Bogusevschi et al. [7], who used virtual reality technology to simulate the water cycle system in nature; Salinas et al. [8], who designed a virtual platform using Augmented Reality (AR) technology which achieved good results in spatial geometry teaching; and Rodrigues D et al. [9], who utilized artificial intelligence technology to create an educational game interface that would customize itself based on the player's profile and adjust its components in real time to improve the correct-answer rate of the players participating in the experiment. Additionally, Lenz et al. [10] developed an MR speech lab that combined a realistic classroom with a virtual environment able present the number of students and the various noises that may be generated, while the teacher monitored students' progress through MR monitors.

Currently, virtual experiments are increasingly being used in education. Wörner et al. [11] analyzed 42 different studies and found that virtual experiments not only enhance users' interest in learning and help them understand knowledge, but also reduce consumables and hazards.

2.2. A Multimodal Fusion Approach to Intent Understanding

Understanding user intent is fundamental to all human–computer interactions. However, enabling machines or systems to comprehend user intent is a challenging task in human–computer interaction research. Multimodal fusion involves fusing multiple senses and utilizing more than one input channel, i.e., gestures, speech, vision, touch, etc., in one system to interact with the machine. Chhabria et al. [12] demonstrated that the use of multimodal methods in virtual reality scenarios enhances the naturalness and efficiency of interactions compared to single-modality techniques.

Holzapfel et al. [13] proposed a fusion structure for multimodal input streams to eliminate speech information ambiguity by using gesture information. In educational games, Corradini et al. [14] utilized speech and gesture information fusion to prevent inconsistencies between modal information. Mollaret et al. [15] proposed extracting user intent based on a hidden Markov model of probabilistic discrete states which integrated multimodal information such as head posture, shoulder direction, and sound, making it easier for robots to comprehend user intent. Ge et al. [16] developed an intent-driven system that could accurately understand users' ideas through observing the actual operational process and analyzing intent expression. Experimental verification established that the intent-driven system is more effective than traditional event-driven systems. Mounir and Cheng [17] utilized complex event processing (CEP) technology and methods on multimodal system input events to decrease users' cognitive and operational burdens in virtual environments. The system generates events transformed into intent based on rules, which enhances efficiency in human–machine interactions in virtual environments.

Currently, multimodal fusion strategies (Yang, M., and Tao, J. [18]) primarily comprise feature-level fusion and decision-level fusion. Concerning feature-level fusion, Jiang et al. [19] introduced a multimodal biometric recognition method based on the Laplacian subspace which fused low-level facial and speech features. Hui and Meng. [20] fused user speech and pen input information at the feature level, enhancing system robustness. Alameda-Pineda et al. [21] utilized user head and body feature data to achieve exceptional posture estimation results. Liu et al. [22] recommended a multimodal fusion architecture based on deep learning, fusing speech, gestures, and body movement at the feature level. These experiments demonstrated the superiority of the multimodal fusion model compared to three single-modal fusion models. With respect to decision-level fusion, Vu et al. [23] used weight standards and the best probability fusion method to fuse speech and gestures at the decision level, designing a bimodal emotion recognition method. Wang et al. [24] suggested a multimodal fusion method for the spatiotemporal feature system, which elevated the accuracy of the emotion recognition system by fusing visual and

auditory data at the decision level. Zhao et al. [25] developed a human–computer interaction prototype system combining facial features, gestures, and speech, allowing the system to better comprehend user needs via decision-level fusion.

2.3. Development of Multimodal Fusion Technology in Virtual Experiments

In order to improve the effectiveness of teaching, some scholars have attempted to apply multimodal fusion technology to virtual experiment teaching. Zhangpan et al. [26] designed a MagicChemMR interaction platform based on chemistry experiments that integrated multiple sensing channels, such as vision, touch, and smell, to provide students with a more immersive experience. Wang et al. [27] proposed a smart glove-based multimodal fusion algorithm (MFA) that integrated speech, vision, and sensor data to capture the user's experimental intent and produce better teaching results. Marín D et al. [28] proposed a multimodal digital teaching method based on the VARK (visual, auditory, reading/writing, kinesthetic) model, which matches different learning modalities with different student styles. The experimental results demonstrated that this teaching method was more effective and improved students' performance.

Our study of existing virtual laboratories revealed that the current number of virtual experiment systems using multimodal fusion as the interaction method is small, and many of these systems have limited abilities to perceive and understand the user's intention, which results in the absence of systems for guiding and correcting user operations. In secondary school experimental teaching, where students are not familiar with the experimental operations, they require an extensive amount of practice to achieve the teaching objectives. Single-modal information acquisition and the provision of complex operation commands further intensify the operational load on students, thereby weakening the development of students' practical hands-on skills. This paper proposes a multimodal fusion AR chemistry experiment system based on Hololens 2, which extracts information from the user's gestures, speech, and visual modalities to deduce the experimenter's intention and guides the user to complete the experiment according to the derived experimental intention. This system has the capability to provide reminders and correct student mistakes, thus improving the efficiency of human–computer interactions and the intelligence of the experiment system.

3. Construction of an AR Chemistry Experiment System Based on Hololens 2

3.1. Hardware

We selected the Microsoft HoloLens 2 device as the platform for our AR chemistry experiment system. The HoloLens 2, released by Microsoft in 2019, offers an immersive experience for users. A physical image of the HoloLens 2 is presented in Figure 1.

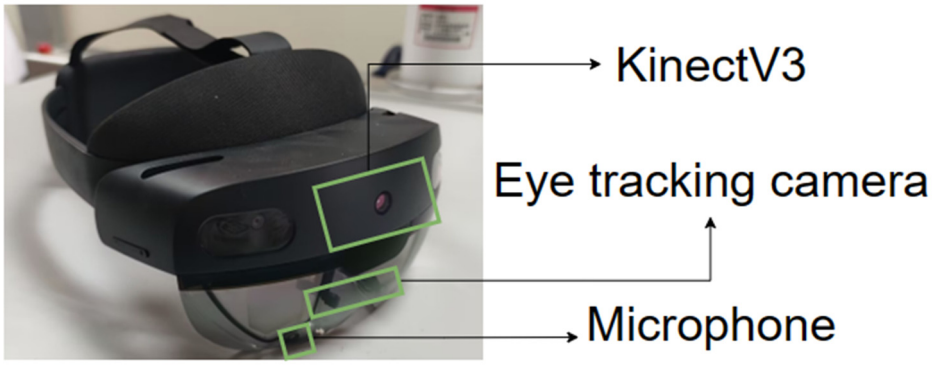


Figure 1. HoloLens 2 physical image.

The following Hololens 2 hardware components were mainly employed in our system:

- Hand tracking component: contains KinectV3 camera, located above the display of Hololens 2.

This camera provides real-time information about the user's hand movements, including the position and relative movement of 25 joints. Obtaining information about hand movements is crucial in understanding the user's experimental intentions for chemistry experiments.

- Microphone assembly: contains a 5-channel microphone array and a speaker with built-in spatial sound effects.

The microphone assembly records the user's speech input in real time and processes it. The speaker device guides the user's voice while simulating the sound of the experiment.

- Eye-tracking component: contains 2 infrared (IR) cameras located inside the Hololens 2 display.

These cameras capture the user's visual information, such as gaze point, position in the AR environment, and duration of the gaze. Visual information reflects the user's operational intent, and its integration into the reasoning of the experimental intent helps to reduce the user's operational load.

3.2. Scene Building

We utilized the 64-bit version of Unity (2019.4.38f) to create a virtual laboratory environment inclusive of a chemistry lab bench, as well as various reagents and instruments necessary for conducting the experiment. Subsequently, we transferred the virtual laboratory to Hololens 2 through a USB connection. The users donned the Hololens 2 apparatus in order to execute the experiment, as demonstrated in Figure 2 for illustration.

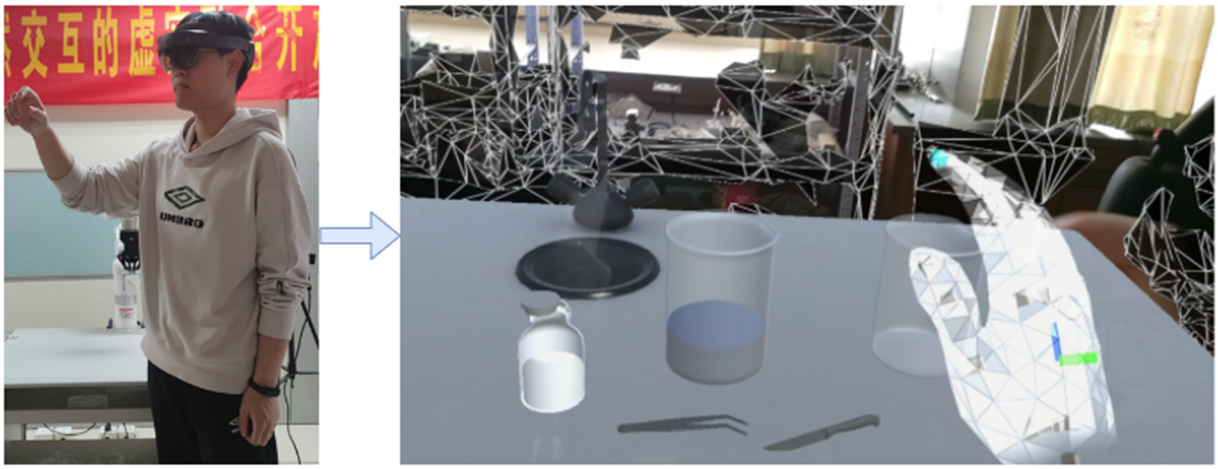


Figure 2. Operational scene.

4. Intent-Understanding Algorithm for Multimodal Fusion to Teach AR Chemistry Experiments

4.1. General Framework

In this study, a multimodal fusion intention understanding algorithm was constructed and applied to Microsoft HoloLens 2 to build an augmented-reality experimental platform. The platform system enables users to complete augmented-reality chemistry experiments through their gestures, speech, and visual gaze information. Unlike traditional augmented reality experiments, this paper selects three pieces of modal information for fusion, resulting in a higher accuracy in terms of user's operation intentions. In addition, the user does not have to stick to tedious unimodal interaction methods, which enhances the immersion and effectively reduces the user's operational load. The article can be divided into four layers: input, information processing, fusion, and application, and the overall framework is shown in Figure 3.

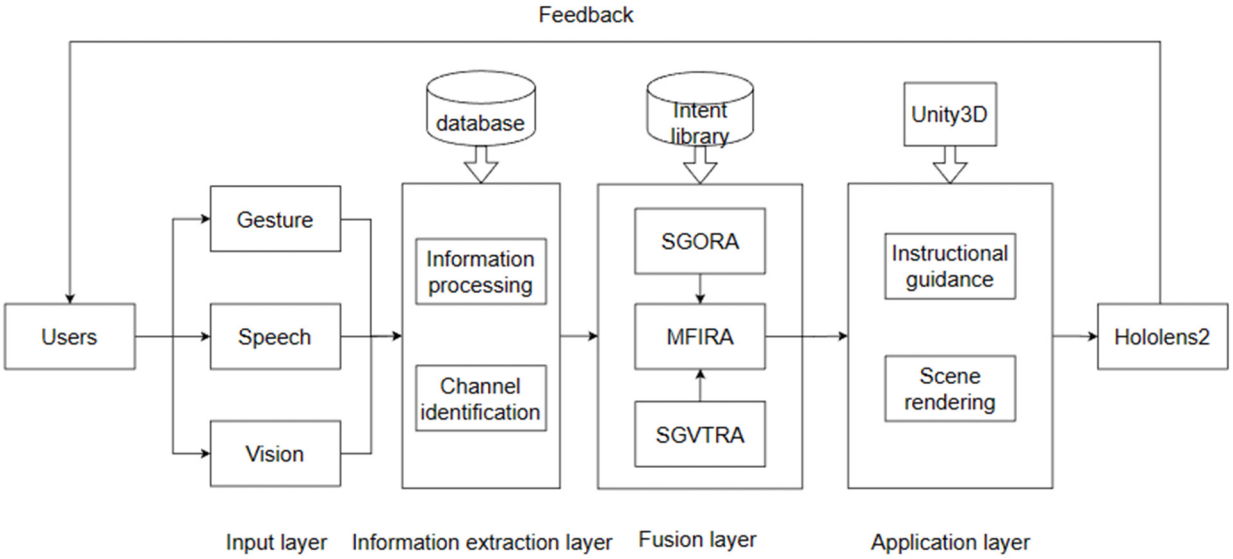


Figure 3. Overall framework of the article.

The input layer contains the user's information for three modalities: respectively, gesture information, speech information, and visual information. The gesture information is obtained through the KinectV3 camera on top of the Microsoft HoloLens 2, and includes the position information of 25 joints of the user's hands and the displacement amount. The speech information is obtained through the microphone device and includes the user's speech audio, and the visual information in the system includes the user's visual gaze point and gaze time. At the same time, the position relationship of the hand in the virtual scene and virtual objects in the unity system interface are obtained.

In the information processing layer, machine learning and mathematical modeling are used in parallel to process the information from the user's gestures, speech, and visual channels and convert it into a mathematical representation. In the fusion layer, we designed two algorithms: SGORA (Speech- and Gesture-based Operation Recognition Algorithm) and SGVTRA (Speech-, Gesture-, and Visual Attention-based Target object Recognition Algorithm). SGORA utilizes gesture and speech information to obtain the probability sequence of the user's "action", while SGVTRA integrates the user's gesture, speech, and visual information to obtain the probability sequence of the user's "target objects". Finally, based on the results of the aforementioned two algorithms, we designed the MFIRA to infer the user's experimental intent. In the application layer, based on the user's intention and the operational steps of the experiment, the system judges whether the user's intention in the fusion layer meets the experimental requirements, guides the user or corrects the wrong operation, and presents the entire augmented reality chemistry experiment scene on Microsoft HoloLens 2.

4.2. Multimodal Fusion for Intent Understanding Algorithms

The goal of this paper was to obtain the users' intentions in virtual chemistry experiments, and the key to achieve this goal was to build an algorithm which would understand multimodal fusion intentions. However, multimodal fusion intention understanding presents challenging problems, such as simple fusion mode, which is unable to reflect the correlation between each modality and makes it difficult to solve the conflict of inconsistent intention between different modalities. To this end, in this paper, we address the operational characteristics of chemical experiments; process the information of three modalities (gesture, speech, and vision) separately; build a fusion model; extract the semantic connection between each modality; and, finally, fuse this information and obtain the user's intention, as follows:

Firstly, we divided the users' experimental intentions into "action" + "target object" for the operation characteristics of chemical experiments, e.g., in the chemical experiment of "sodium-water reaction". The experimental intention of the experimenter: "拿起烧杯" (pick up the beaker) can be divided as follows: the operational action is "拿起" (pick up) and the target item is "烧杯" (beaker). Therefore, the extraction of the user's experimental intention can be divided into the acquisition of the user's action and the acquisition of the user's target item.

Thus, the intent-understanding algorithm for multimodal fusion can be divided into three parts: using the SGORA algorithm to obtain the user's action probability sequence through speech and gesture information; using the SGVTRA algorithm to obtain the user's target object probability sequence through speech, gesture, and visual attention information; and using the MFIRA algorithm to check for conflicts between modal expressions based on the inference results of the SGORA and SGVTRA algorithms and to infer the user's experimental intention.

In the process of the actual experiment, the information regarding multiple modalities of the user is not necessarily input at the same time, and the order of their input is random. After the analysis of the experimental data, when the gesture, speech, and visual information are used to express the same semantic meaning, the unimodal information is generated in the time period T. The block diagram of the multimodal fusion algorithm for intention understanding is represented in Figure 4.

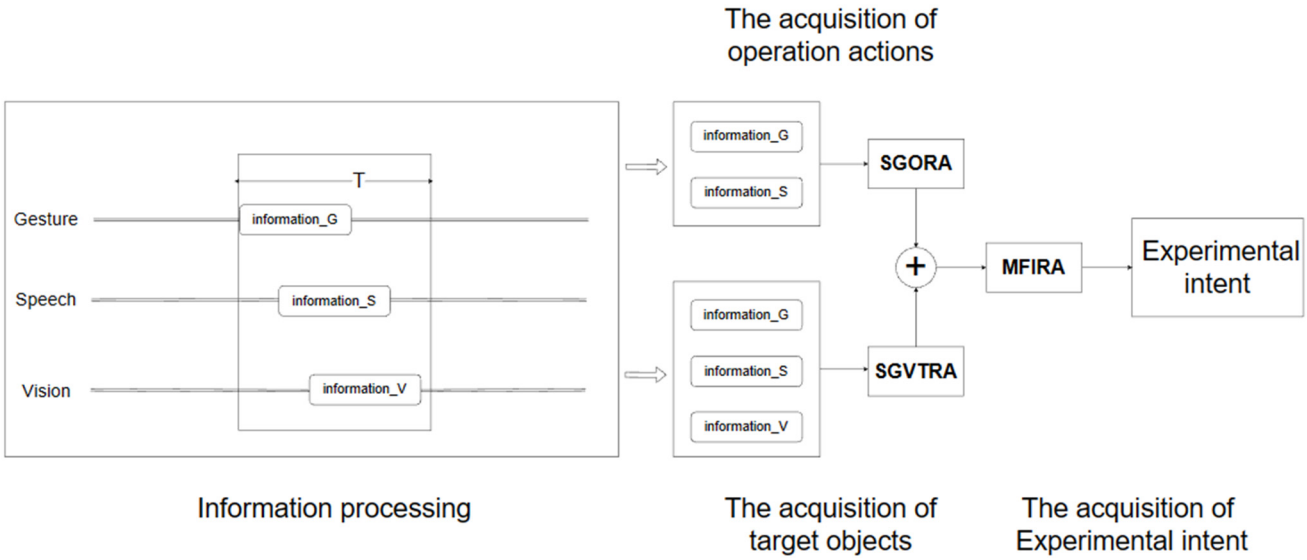


Figure 4. Algorithm diagram. (Note: T is obtained from the data measurement. In the selected chemical experiment, "sodium-water reaction", the data obtained from 20 students measured T at about 5S).

4.2.1. SGORA: Operation Action Acquisition

In chemical experiments, the user's actions are limited to basic actions such as "pick up", "put down", and "pour/rotating". Consequently, we consider the acquisition of actions as a machine learning problem. Firstly, two key challenges need to be addressed: (1) the modality inputted by the user in the actual experimental operation is not consistently unique; and (2) the user's misoperation causes the system to make inaccurate judgements.

When the user only inputs gestures, we assign gestures to specific operation actions based on chemical experiment operation habits. An LSTM network is used for determining the user's operation actions. To address the issue of user misoperation while performing operations (e.g., the user may make an unintentional gesture), a machine learning feature layer fusion method is utilized to train a model that combines speech and gestures. We used convolutional neural networks (CNN) to extract speech features, as CNN has shown

promising performance in speech recognition [29], and long short-term memory (LSTM) networks to extract gesture features and then concatenate them for modeling. Incorporating speech features can increase the precision of intention comprehension and reduce misjudgments of user gesture misoperations. The specific approach is as follows.

Unimodal (Gesture)

The dataset for gesture recognition is defined by this paper as the time-dependent relative motion sequences between finger and palm nodes, with a strong sequence correlation. Therefore, we selected the LSTM network to process these sequences. A Hololens 2-mounted KinectV3 camera was used to acquire the gesture dataset, which was later trained on LSTM for classification.

Based on the traits of the "sodium water experiment" and students' practical experience, the experiment's gestures were segregated into three different operational sequences, namely: "pick up", "put down", and "pour\rotating". The input sequence for these sequences constituted the position of finger nodes, except the wrist, relative to that of the wrist node at time "t" which was then added to the LSTM network training. A typical LSTM cell c_t. contains three gate units: input gate i_t, forgetting gate f_t, and output gate o_t, connected with recursive and feedforward links. The final state, h_t is controlled by the unit output gate o_t. By vertically layering the LSTM layers, such that the previous LSTM layer's output serves as the next layer's input, we discovered more advanced temporal features of the stacked LSTM model. This is illustrated in Figure 5, which depicts the classified and trained gesture operation processes.

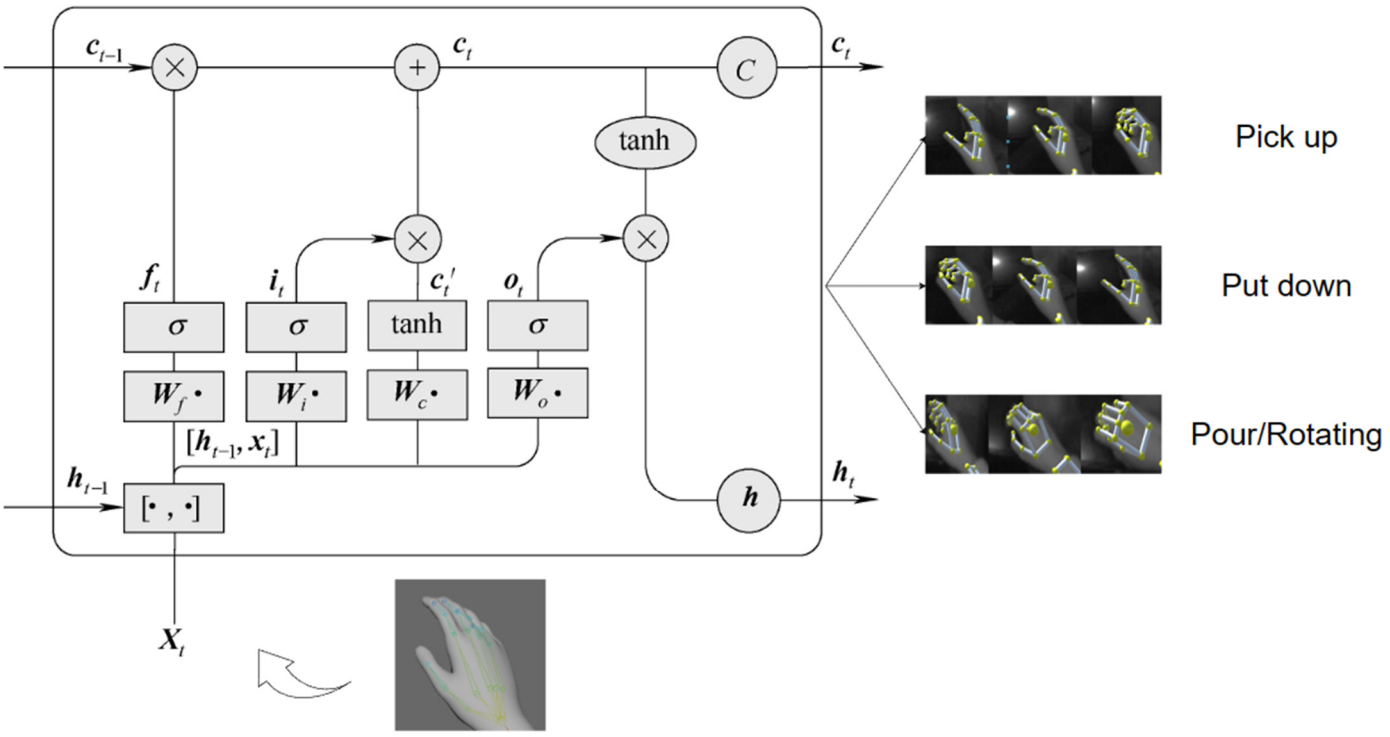


Figure 5. Gesture classification and training process.

Multimodal (Gesture + Speech)

Due to the semantic similarity between speech and gestures in AR chemistry experiments, we developed a machine learning model that amalgamates both features. To identify users' actions, such as "pick up", "put down", and "pour\rotating", we must first recognize the corresponding speech. In this paper, we utilized a CNN network to recognize and process speech information. In CNN speech command recognition, the speech command data set X^S is preprocessed by a fast Fourier transform method to generate a two-dimensional spectrogram before being fed to the CNN. Equation (1) defines the convolution operation formula:

$$y^{(j)} = ReLU\left(\sum_i a^{(ij)} \cdot x^{(i)} + b^{(j)}\right) \quad (1)$$

where $x^{(i)}$ and $y^{(i)}$ denote the ith input map and the jth feature map, respectively. $x^{(i)}$ is the local region that shares weights between each convolutional neuron $a^{(ij)}$. $a^{(ij)}$ denotes the convolutional neuron between the i_th input map and the j_th feature map. $b^{(j)}$ denotes the bias of the convolutional neuron $a^{(ij)}$. The activation function uses $ReLU(y = \max(0, x))$. The maximum pool outputs the maximum value of each local neighbor so that each feature map remains invariant to the local panning in the input map. In model training, this paper uses categorical cross-entropy as the loss function $\mathcal{J}$, defined as Equation (2):

$$\mathcal{J} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\log\left(p_{ij}\right) \quad (2)$$

where y_{ij} is the binary indicator of whether the observation X_i^S belongs to class c_i; p_{ij} is the prediction probability of whether the observation X_i^S corresponds to class c_i; n is the number of training samples; and k is the number of classification labels. The training process for speech is shown in Figure 6:

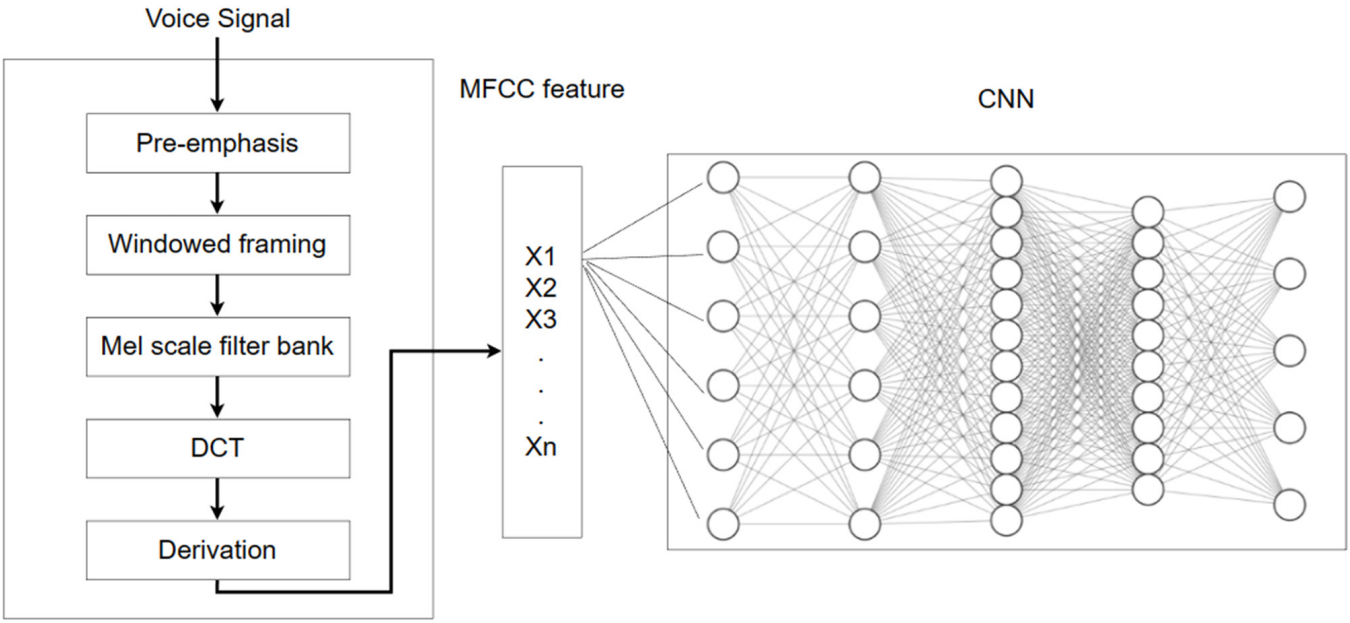


Figure 6. Mel-frequency cepstral coefficient (MFCC) feature extraction and CNN training. After processing the speech, we removed the last layer of the CNN model for speech recognition and the LSTM model for hand motion recognition, respectively, after which the models could be represented as follows: speech model: $r^S\left(X_i^S;\theta^S\right)$; hand gesture model: $r^H\left(X_i^H;\theta^H\right)$.

where X_i^S and X_i^H are training samples from the speech command recognition dataset and the hand motion dataset. θ^S and θ^H are network parameters from the speech command recognition model and the hand motion recognition model. In this paper, we define functions to fuse the features of the two modalities; the expressions were as shown in Equation (3):

$$\mathcal{G} = F\left(r^S(X_i^S;\theta^S), r^H(X_i^H;\theta^H)\right) \tag{3}$$

where $\mathcal{G}$ is the representation of the fused features, and the fusion model is further trained by minimizing the loss function (defined by the cross-entropy softmax). Finally, the optimized network was connected to the softmax function to normalize the output results. The training process of the fusion model is represented as Equation (4).

$$\min \sum_{i=1}^{n} \underset{W^F,\theta^S,\theta^H}{\mathcal{L}}(sqrtmax(W^F\mathcal{G}), c_i) \tag{4}$$

where W^F denotes the weight of the softmax layer after multimodal fusion. The training process of the model is shown in Figure 7:

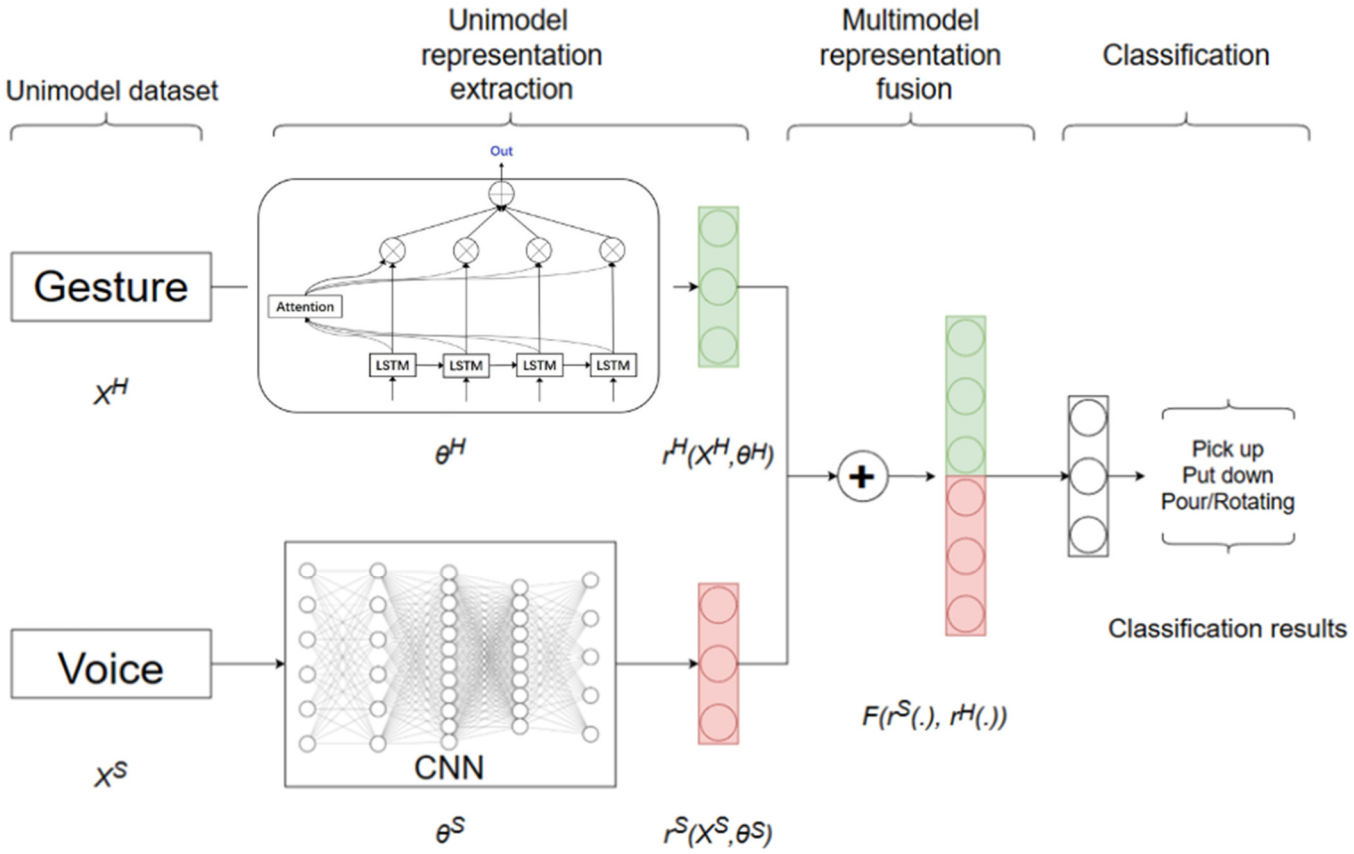


Figure 7. Training process of fusion models.

The Speech and Gesture-based Operation Recognition Algorithm (Algorithm 1) is described as follows:

Algorithm 1. SGORA: Speech and Gesture-based Operation Recognition Algorithm

Input: speech, gesture
Output: user's action probability sequence OP = [OP1,OP2, ... OPN]
1: While the system acquires information (speech or gesture) ! = NULL:
2: T0 = current system time
T1 = T0 + T
3: Within the time period: [T0, T1], the number of channels is determined for the acquired modalities:
4: IF only single modal information is available, Gesture ! = NULL:
Gesture -> LSTM -> Classification;
Obtain gesture classification similarity, compare gesture library:
OP = [OP1,OP2, ... OPN] after normalization of recognition results
5: Otherwise
(Speech ! = Null and Gesture ! = Null):
(Gesture + Speech) -> fusion model ($\mathcal{G}$) -> MLP -> Softmax -> Classification;Based on the classification result of the fusion model, the recognition result is normalized to OP = [OP1,OP2, ... OPN]Return OP
End

4.2.2. SGVTRA: Target Object Acquisition

This study aimed to extract information regarding target items by considering the characteristics of users in chemical experiments. Specifically, we focused on three modalities: gesture, speech, and visual attention.

Using a novel approach in the gesture portion, we utilized the vector angle of hand motion and the changes in proximity between the hand and the virtual object to establish a mathematical model that led to the extraction of the target item probability sequence, OBJ_A.

In the speech portion, we used text conversion and the LTP platform (Che, W. et al. [30]) to calculate cosine similarity with the pre-set item database, leading to the extraction of the target item probability sequence, OBJ_B.

Based on the user's visual fixation time on each object during operation, a mathematical model was established which led to the derivation of the probability sequence OBJ_C.

Eventually, to objectively allocate weights to each channel and obtain the weight coefficients of the three channels, we employed the coefficient of variation method. This method was further used to extract the target item probability sequence, OBJ.

Unimodal (Gesture)

Our current study has revealed a universal pattern in which users conform with the experimental specification for gestures when selecting a target object:

- The hand will be close to the target when the user selects the target object;
- The movement speed of the wrist when the user selects the target object will be less than a threshold value δ, and the whole process consists of deceleration.

Therefore, we modeled the operation process of the user by selecting the target object by gesture, and first, the mathematical expression of the restriction is

$$\begin{cases} dis(hand, j)_{T_0} - dis(hand, j)_{T_1} < 0 \\ V_{hand} < \delta \\ a_{hand} < 0 \end{cases} \quad (5)$$

where $dis(hand, j)_{T_0}$ and $dis(hand, j)_{T_1}$ denote the distance between the hand and the item j at the moments of T_0 and T_1 respectively. V_{hand} denotes the moving speed of the wrist joint point; δ is a speed threshold; and a_{hand} denotes the acceleration of the wrist joint point movement at this time. We calculated the probability of selecting each item through ges-

tures, using OBJ_A_j to represent each object. The calculation is shown in Equations (6) and (7):

$$I_j = \frac{1}{\theta_j + d_j} \tag{6}$$

$$OBJ\_A_j = \frac{I_j}{\sum_{i=1}^{m} I_i} \tag{7}$$

where θ_j is the angle between the direction of the motion of the human hand and the vector between the human hand and the virtual object j. A smaller angle indicates that the direction of motion of the human hand is more inclined to object j. d_j is the distance between the human hand and the virtual experimental object j, and is calculated as Equation (8):

$$d_j = \sqrt[2]{\left(H_x^U - j_x\right)^2 + \left(H_y^U - j_y\right)^2 + \left(H_z^U - j_z\right)^2} \tag{8}$$

Here, H^U denotes the mapped position of the human hand in the AR environment and j denotes the virtual experiment item. After performing normalization, the target item probability sequence was obtained as OBJ_A:

$$OBJ\_A = normalization[OBJ_{A1}, OBJ_{A2}, \ldots, OBJ\_A_n]$$

Note: during the gesture operation, when the system recognized the "Pick up" gesture, we used the above method to calculate the "Target Object" A. We then mapped A's coordinate to the position of the user's hand, enabling it to move with the hand. At this point, the target object remained constant. When the system recognized the "Pour/Rotating" gesture, A moved along with the hand's rotation, and we recalculated the target object using the above method. For example, if the user were to pour water from a narrow-mouthed bottle into a beaker, with the bottle being A, the "beaker" would become the new target object, and after the completion of the "Pour/Rotating" gesture, A would revert back to being the target object. When the system recognized the "Put down" gesture, we canceled the mapping between the target object A and the user's hand.

Unimodal (Speech)

We utilized the Baidu Voice API interface to transcribe the users' speech to text. Subsequently, the LTP word segmentation technology was employed to divide the user's speech into lexical representation. Then, we extracted the noun phrase after the verb as the user's intended action. For example, if the speech input was "I want to pick up the beaker", the classification would consist of "I", "want", "pick up", and "beaker", with n corresponding to the word "beaker." To calculate the cosine similarity with the "experimental items" in our database, n was normalized to obtain the probability sequence of the target object: $OBJ\_B = normalization[OBJ_{B1}, OBJ_{B2}, \ldots, OBJ\_B_n]$, as shown in Equation (9) below.

$$OBJ\_B_j = \frac{\sum_{i=1}^{n}\left(X_i \times Y_j\right)}{\sqrt{\sum_{j=1}^{n}\left(X_j\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(Y_j\right)^2}} \tag{9}$$

where $X = [X_i]$ is the extracted noun and $Y = [Y_i]$ is the phonetic form of item j, used as a comparison. The whole process is schematically illustrated in Figure 8:

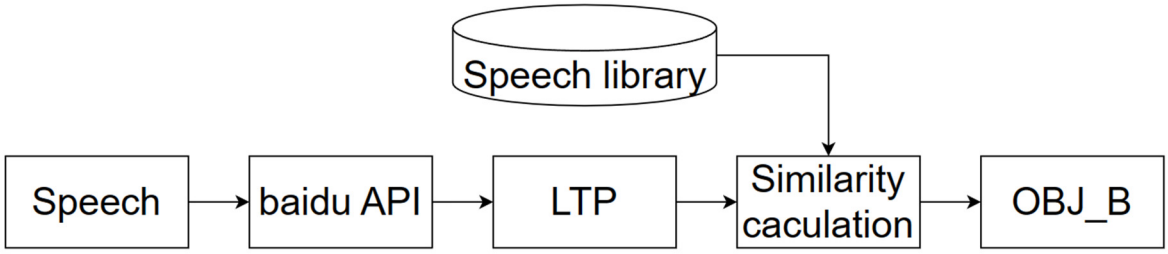


Figure 8. LTP segmentation + similarity calculation.

Unimodal (Visual Attention)

According to Ludwig and Gilchrist (Ludwig, C. J., and Gilchrist, I. D. [31]), individuals tend to perform operations within their area of focus, and visual attention can reflect operational intentions. Therefore, for our AR experiment, we chose the smallest external sphere as the attention range (range(j)) for item j. To determine the user's potential "target object of operation," we employed modeling and analysis of visual attention. Specifically, we examined how long the user's gaze point remained within the attention range (range(j)) of the item. The longer the user's gaze point stayed within the attention range, the greater the likelihood became that object j was the intended target.

In the time period $T = [T_0,\ T_1]$, the "gaze point" and "gaze time" of the user were obtained in real time by the Hololens 2 device. The specific calculation method is as follows:

$$\begin{cases} gazepoint\ is\ on\ range(j) \\ gazetime > 0.2\ s \end{cases} \tag{10}$$

$$I[j] = I[j] + 1 \tag{11}$$

$$OBJ\_C_j = \frac{I[j]}{\sum I[j]} \tag{12}$$

where $I[j]$ is the count of the gaze point within the attention range $range(j)$ of the object j, initially 0 and self-adding 1 each time the condition (10) is satisfied, and OBJ_C_j is the probability that object j is the target object.

When users scan an object, the number of fixations may increase, but the duration of each fixation is relatively short. In such cases, the probability of the object becoming the "target object" is lower. To eliminate the influence of visual scanning during the experiment, we referred to Gezeck et al.'s statistical analysis of eye fixation reaction times. According to their research findings [32], the slow regular mode of human eye fixations is approximately 200 ms. Therefore, we incorporated $gazetime > 0.2\ s$ as a limiting condition and finally obtained OBJ_C.

Multimodal (Gesture + Speech + Visual Attention)

After processing the information from the above three modes, we obtained the target item probabilities from the "speech information," "gestures," and "visual attention information." We objectively weighed the information obtained from these modes using the coefficient of variation method, normalized the weights, and fuse them to obtain a sequence of probabilities. We carried out this process as follows:

First, the three channels' information were spliced into a matrix: $OBJ = [OBJ\_A,\ OBJ\_B, OBJ\_C]$, whose dimension was 3 × n, expressed in Equation (13):

$$OBJ = \begin{pmatrix} OBJ\_A1 & \cdots & OBJ\_C1 \\ \vdots & \ddots & \vdots \\ OBJ\_An & \cdots & OBJ\_Cn \end{pmatrix} \tag{13}$$

Each column stored a single channel probability sequence of "Target object", and the three channel probability sequences were weighted to calculate the mean $\overline{x}_j$ and standard deviation S_j of each object using Equation (14):

$$\begin{cases} \overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ S_j = \sqrt{\frac{\sum_{i=1}^{n}\left(x_j - x_j\right)^2}{n-1}} \end{cases} \tag{14}$$

The coefficient of variation of the evaluation index of the jth term was obtained as

$$v_j = \frac{s_j}{\overline{x}_j}, j = 1, 2, \cdots, p \tag{15}$$

Normalizing them, the weights of the three channels were obtained as

$$w_j = \frac{v_j}{\sum_{j=1}^{3} v_j} \tag{16}$$

Finally, the probability sequence of "Target object" of the fused user was:

$$Target\ object = w_1 \times OBJ\_A + w_2 \times OBJ\_B + w_3 \times OBJ\_C \tag{17}$$

The algorithm description for the Speech, Gesture, and Visual Attention-based Target object Recognition (SGVTRA) is as follows (Algorithm 2):

Algorithm 2. SGVTRA: Speech, Gesture, and Visual Attention-based Target object Recognition Algorithm

Input: Speech, Gesture, Visual attention
Output: user's action probability sequence
OBJ = [OBJ1,OBJ2, . . . OBJN]
1: While system obtains information (Gesture, Speech or Visual) ! = NULL:
2: T0 = current system time
T1 = T0 + 5s
3: During the time period: [T0, T1],
fuse the acquired modalities with the information:
4: IF Gesture ! = Null:
Determine whether the Formula (5) is satisfied or not.
According to Formulas (6)–(8),
(θ_j, d_j) -> I_j -> OBJ_A_j
OBJ_A = normalization [OBJ_A_j]
5: IF Speech ! = Null:
According to the result of speech recognition,
Use Equation 9 to perform cosine similarity calculations for word segmentation results with the item library.
Speech -> API -> LTP -> $\cos(\theta)$ -> OBJ_B_j
OBJ_B = normalization [OBJ_B_j]
6: IF Visual attention ! = Null:
Determine whether the Formula (10) is satisfied or not.
Update OBJ_C_j according to Equations (11) and (12)
OBJ_C = [OBJ_C_j]
7: Using the coefficient of variation method:
Weight OBJ_A, OBJ_B, OBJ_C, update the weights to
w_1, w_2, w_3
OBJ = $w_1 \times$ OBJ_A + $w_2 \times$ OBJ_B + $w_3 \times$ OBJ _C
Return OBJ
End

4.2.3. MFIRA: Multimodal Fusion Intent Recognition Algorithm

The machine learning-based model fusion method (SGORA) was employed to obtain the probability sequence OP of the user's operational actions, and an algorithm (SGVTRA) was designed to obtain the probability sequence OBJ of the user's intended target object. After that, we obtained a Cartesian product of the two probability sequences to determine the probability sequence of the user's intention, which is represented as follows:

$$Intention = \text{“}Operate\text{”} \times \text{“}Target\ object\text{”}$$

$$INT = OP \times OBJ \tag{18}$$

Since the experimental subjects of this paper were many junior high school students, there were problems with misuse due to unfamiliarity with the experimental operations and conflicting expressions between modalities. There were methods implemented to remove the conflicts in the modalities, as follows:

After utilizing the Cartesian product to obtain the intention probability sequence INT, it was rearranged from largest to smallest to achieve: $INT = [int1, int2, \ldots intm]$, where $int1 > int2 > \cdots > intm$. Setting the threshold ε, if $int1 - int2 < \varepsilon$, the modal input was considered incompatible, the system cleared the current state of all items, and the user was prompted by voice to re-enter.

The MFIRA is described as follows, (Algorithm 3):

Algorithm 3. MFIRA: Multimodal Fusion Intent Recognition Algorithm

Input: SGORA, SGVTRA
Output: Intent Int
1: while SGORA (G, S) ! = NULL and SGVTRA (G,S,V) ! = NULL:
2: obtain OP, OBJ
3: INT = OP × OBJ
4: Reorder the INT sequence from largest to smallest:
INT = [int1, int2, . . . intm]
5: IF int1 − int2 < ε:
System Clear
Voice prompt: Intent conflict between modalities,
please redo the operation
Break;
6: else:
Int = MAX(INT)
Return Int
End

5. Experiment

The experimental portion contains the following aspects: 1. recognition rate of the fusion model; 2. MFIRA algorithm analysis; 3. experimental guidance method based on MFIRA algorithm; 4. comparison experiments; and 5. operational load and user satisfaction analysis.

5.1. Experimental Setup

We proposed a multimodal intent understanding algorithm for AR chemistry experiments and selected the "sodium-water reaction" experiment as the algorithm validation experiment according to the secondary school chemistry experiment syllabus. The system was built with Unity 2019.4.38f (64-bit), the scene was deployed on Hololens 2, and users wore the Hololens 2 to conduct the experiment. The entire scene of the AR chemistry experiment is depicted in Figure 9.

Figure 9. User interface for the sodium–water reaction experiment.

Following the experimental procedure of the "sodium-water reaction" experiment, we established the following database, shown in Table 1:

Table 1. Database of sodium–water reaction.

Experimental Steps	Target Object	Operation Action
1. Pick up the beaker	Beaker	Pick up
2. Put down the beaker	Beaker	Put down
3. Pick up the narrow-mouthed bottle containing water	Narrow-mouthed bottle	Pick up
4. Pour water into the beaker	Beaker	Pour\Rotating
5. Put the narrow-mouthed bottle containing water on the table	Narrow-mouthed bottle	Put down
6. Pick up the phenolphthalein reagent bottle	Phenolphthalein reagent bottle	Pick up
7. Put down the phenolphthalein reagent bottle	Phenolphthalein reagent bottle	Put down
8. Pick up the dropper with a rubber head	Rubber-headed dropper	Pick up
9. Drop phenolphthalein into the beaker	Beaker	Pour\Rotating
10. Put the dropper with rubber head back into the phenolphthalein reagent bottle	Rubber-headed dropper	Put down
11. Pick up the metallic sodium reagent bottle	Metallic sodium reagent bottle	Pick up
12. Put down the metallic sodium reagent bottle	Metallic sodium reagent bottle	Put down
13. Pick up the forceps	Forceps	Pick up
14. Take metallic sodium from the metallic sodium reagent bottle with forceps	Metallic sodium reagent bottle	Pour\Rotating
15. Put metallic sodium into the beaker	Beaker	Pour\Rotating
16. Put the forceps back on the table	Metallic sodium reagent bottle	Put down

The operation interface is shown in Figure 10.

5.2. Experimental Results

5.2.1. Recognition Rate of Fusion Model

We invited 14 secondary school students, each providing 60 sets of (gesture + speech) data, and divided the training set, validation set, and test set according to the ratio of 7:2:1. Using the method described in Section 4.2.1 an operational action recognition model (G) based on dynamic gestures was trained using an LSTM network. An operational action recognition model (G + S) combining gestures and speech was obtained using feature layer fusion. The performances of the two models on the test set are shown in Table 2:

Table 2. Performances of the two models.

	Pick Up	Put Down	Pour\Rotating
G	98.77	99.16	98.23
G + S	97.92	98.01	97.43

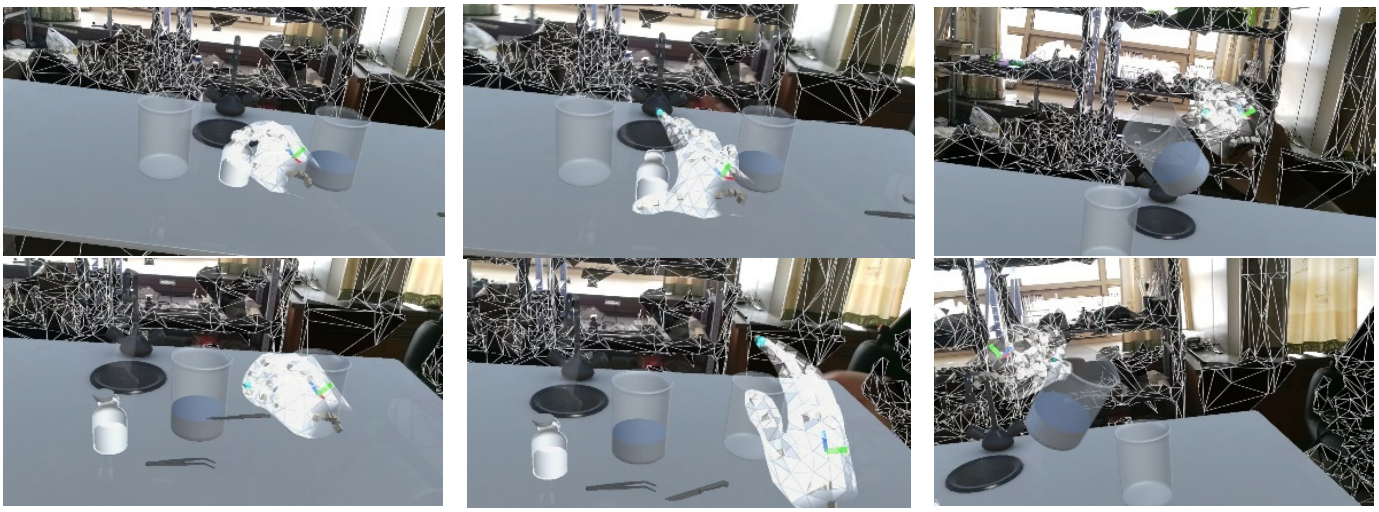


Figure 10. Operational schematic diagrams for the sodium–water reaction experiment.

The fusion model G + S requires training the features of both gesture and speech, thus necessitating a larger and more diverse dataset for training. For this experiment, we invited 14 high school students to provide gesture and speech data. Due to the diverse vocal characteristics of the students, the speech dataset samples were relatively imbalanced, which had some impact on the feature learning of the fusion model G + S. However, even with this challenge, after 400 rounds of training, the average accuracy of the fusion model reached 97.79%, meeting the basic requirements for intent inference.

Both models achieved a high level of accuracy in terms of recognizing operational actions. Figure 11 illustrates the fluctuation of accuracy and loss of the fusion model during the training process. The model's faster convergence was attributed to the pre-training weights for the gesture and speech models, achieving approximately 99.38% accuracy and 0.0109 loss in about 70 iterations.

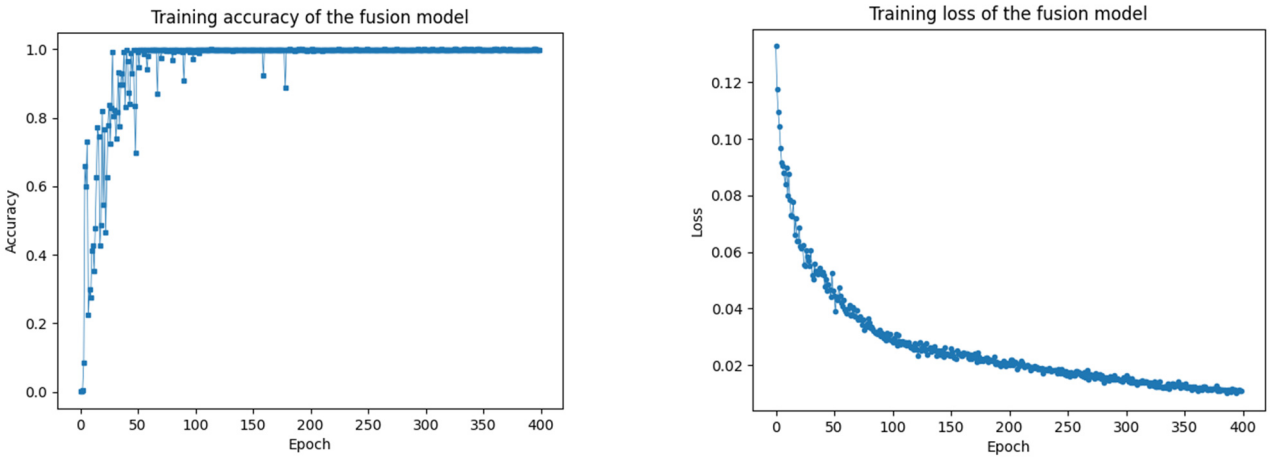


Figure 11. Training accuracy vs. training loss for multimodal fusion.

The classification accuracy of the model for each category and the classification confusion between different categories can be derived by looking at the confusion matrix of the model, as shown in Figure 12:

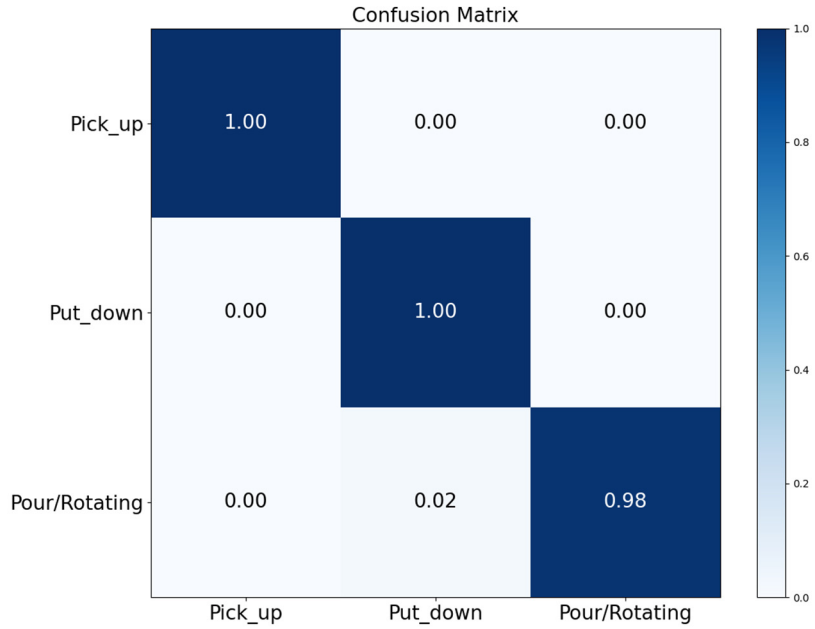


Figure 12. Confusion matrix of the fusion model.

The fused model achieved 100% accuracy for both "Pick up" and "Put down", and 98% accuracy for "Pour\Rotating". In general, the action recognition model showed a good level of accuracy, providing the basis for the intention-understanding algorithm.

5.2.2. MFIRA Algorithm Analysis

The MFIRA algorithm incorporates multiple modalities to elicit the user's experimental intent, and is evaluated in two ways: first, to validate its accuracy in terms of interpreting the intent, and second, to assess the impact of multi-channel fusion on its intent interpretation accuracy.

For the testers, eight junior high school students were invited to participate in the algorithm test. They were shown a demonstration of the "sodium-water reaction" experiment and were given an introduction to the basic operation of the experiment before performing it. The "sodium-water reaction" experiment comprised 16 steps (see Table 2), and the testers were instructed to perform 5 complete experiments, resulting in a total of 80 experimental operations per person.

The MFIRA algorithm was modified using the test software to investigate the effect of multi-channel fusion on intent interpretation accuracy by dividing it into three groups: group A acquired only the user's gesture information; group B acquired the user's gesture and speech information, and group C acquired the user's gesture, speech, and visual information. During the experimental test, the three groups of programs ran simultaneously. After the operator performed an operation, the system paused to record the number of "accurate" and "inaccurate" results for each group. We calculated the accuracy rate of each group by the formula "number of accurate results"/"number of tests results". The intent inference was considered accurate when the student's current experimental intent aligned with the intent inferred by the MFIRA algorithm. Therefore, even if the user were to perform an incorrect operation that deviated from the standard procedure, as long as the MFIRA algorithm correctly inferred this error, it would be considered a correct inference. In summary, the user's own actions do not affect the accuracy of the algorithm; only the results inferred by the algorithm impact the accuracy.

The results of the test are presented in Figure 13.

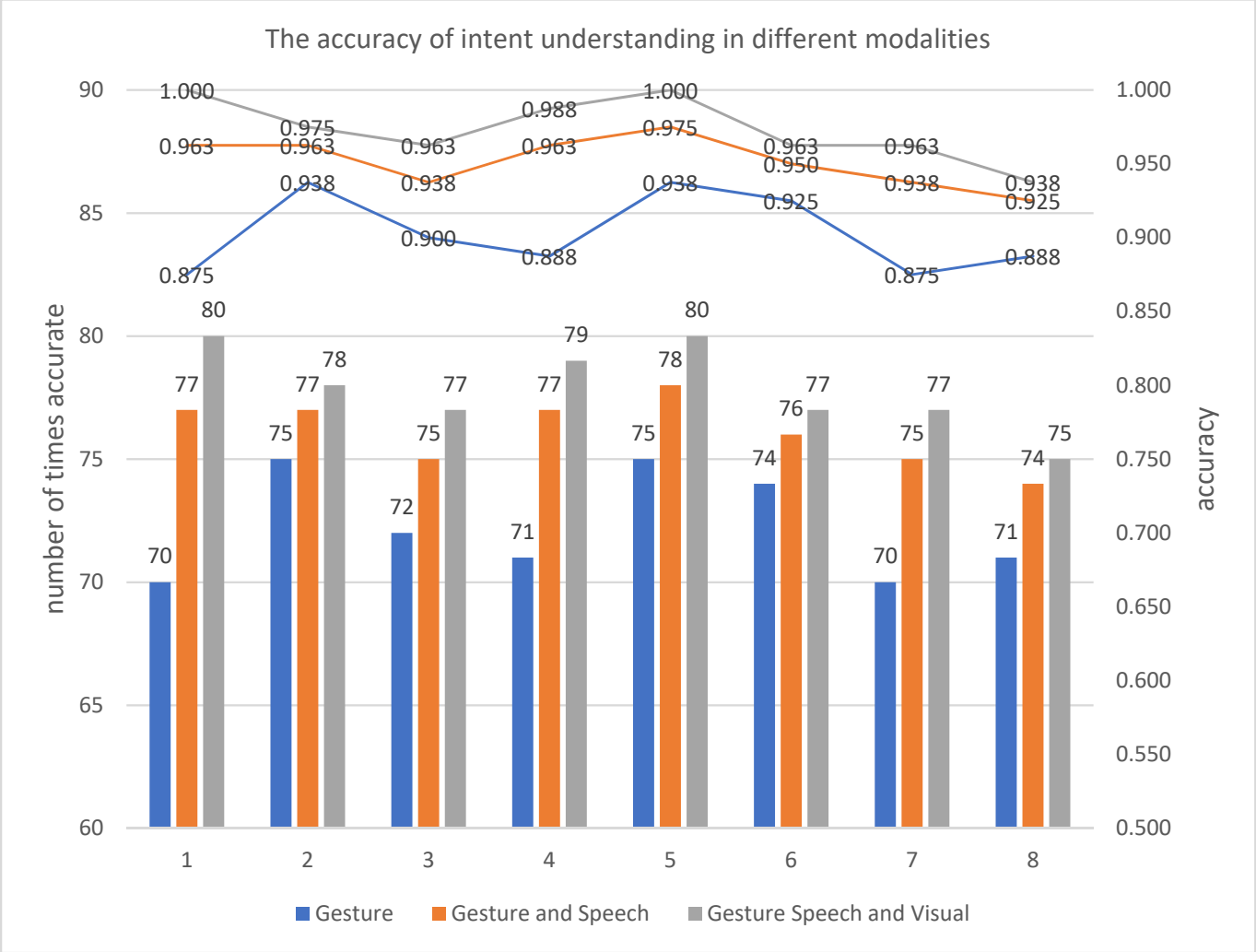
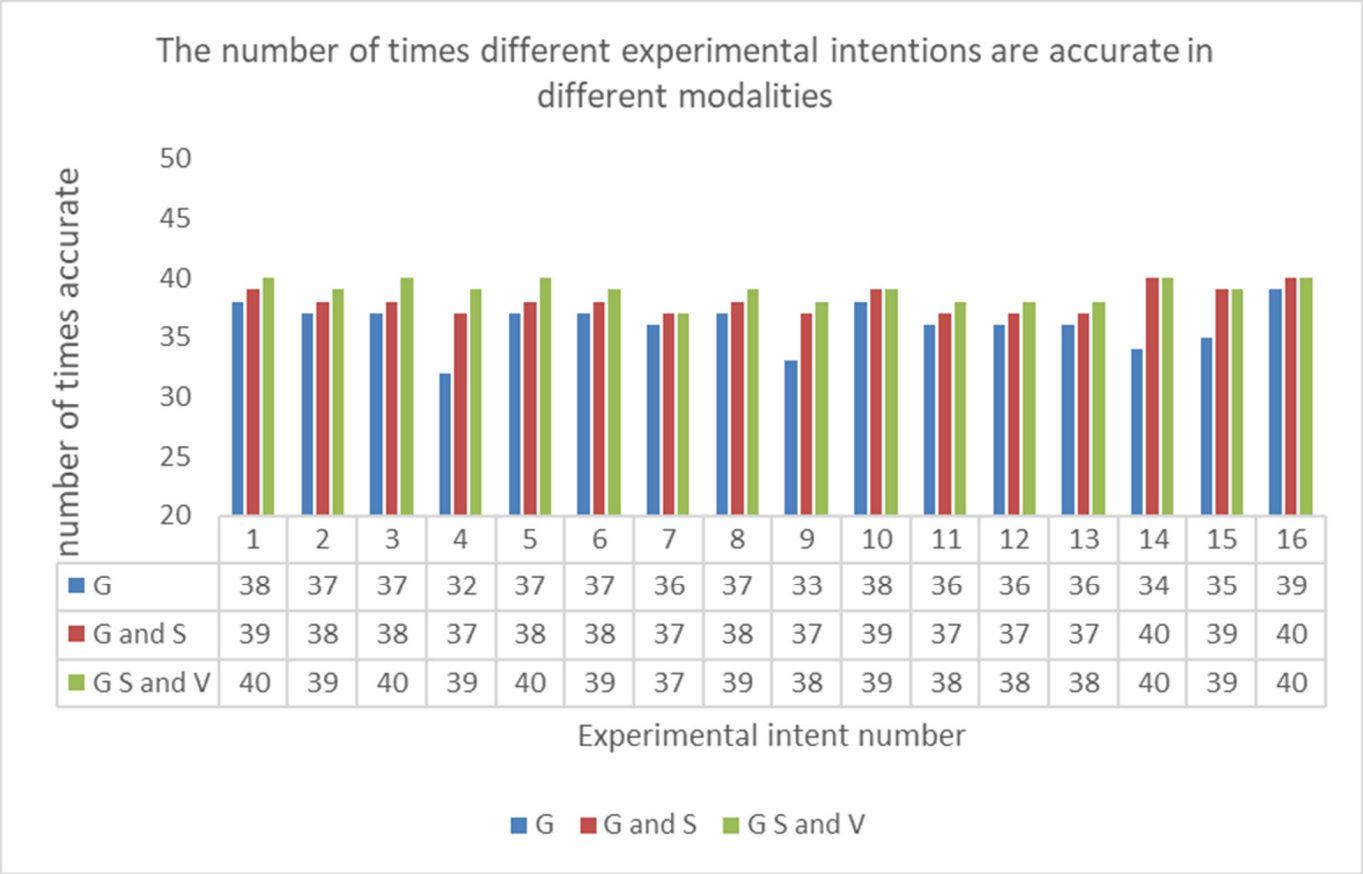


Figure 13. The accuracy of intent understanding in different modalities.

In the actual test, the MFIRA algorithm exhibited a range of 70–80 accurate results. Within that range, the average number of correct inferences is 72.25, and the average accuracy was 90.03% for the group that solely used gesture information. The group that used gesture and speech information had an average of 76.125 correct MFIRA inferences and an average accuracy of 95.2%. Additionally, in the group that used gesture, speech, and visual information, the average number of MFIRA inferences was 77.875, with an average accuracy rate of 97.3%. The bar chart visually demonstrates that the algorithm's accuracy gradually increased as the modality increased, while keeping the operation constant. Analyzing the above line graph, adding speech information significantly improved the MFIRA algorithm's average accuracy, by 4.9%. Moreover, the average accuracy further improved by 2.1% after adding visual information, resulting in an overall accuracy of 97.3%, which had a positive effect on the test results.

In order to analyze the specific influence of multi-channel information fusion on the accuracy of intention understanding, we analyzed 16 operational intentions for this experimental test. In total, each operational intention was tested 40 times by eight students. The visualization of the test results is shown in Figure 14.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
G	38	37	37	32	37	37	36	37	33	38	36	36	36	34	35	39
G and S	39	38	38	37	38	38	37	38	37	39	37	37	37	40	39	40
G S and V	40	39	40	39	40	39	37	39	38	39	38	38	38	40	39	40

Figure 14. The number of times different experimental intentions were accurate in different modalities.

In the figure, it should be noted that the recognition effect was significantly improved in STEP4, STEP9, STEP14, and STEP15 by adding speech modal information. For example, in STEP4, the operation intention was to pour water into a beaker. In this step, the user needed to select a container with water from a longer distance and pour it into the beaker. The operation process was relatively lengthy, and since the user gestured with or shook the device while selecting the container, if only the gesture recognition model was used, it was challenging to handle the user's misoperation, which was able to cause a failure of intention inference. However, the speech features could be used as a supplement to the user's behavioral information. This also explains the difference in the table, where the unimodal gestures had higher recognition rates compared to the fusion model for the three operational actions, but did not work well in the final intent inference.

5.2.3. Teaching Guidance Method Setting Based on MFIRA Algorithm

We stored the steps of the "sodium-water reaction" in the system's database. When the user performed the operation, we used the MFIRA algorithm to obtain the user's experimental intention by combining their completed steps and checking whether they conformed to the specifications. If the steps did not conform to the specifications, the system reminded the user to redo the activity. Conversely, when the steps were correct, the system updated the completion progress. If the user did not operate for a prolonged duration, the system prompted the user, using speech, to complete the next operation based on their progress and the "sodium-water reaction" steps. The flowchart for the teaching guidance method is shown in Figure 15.

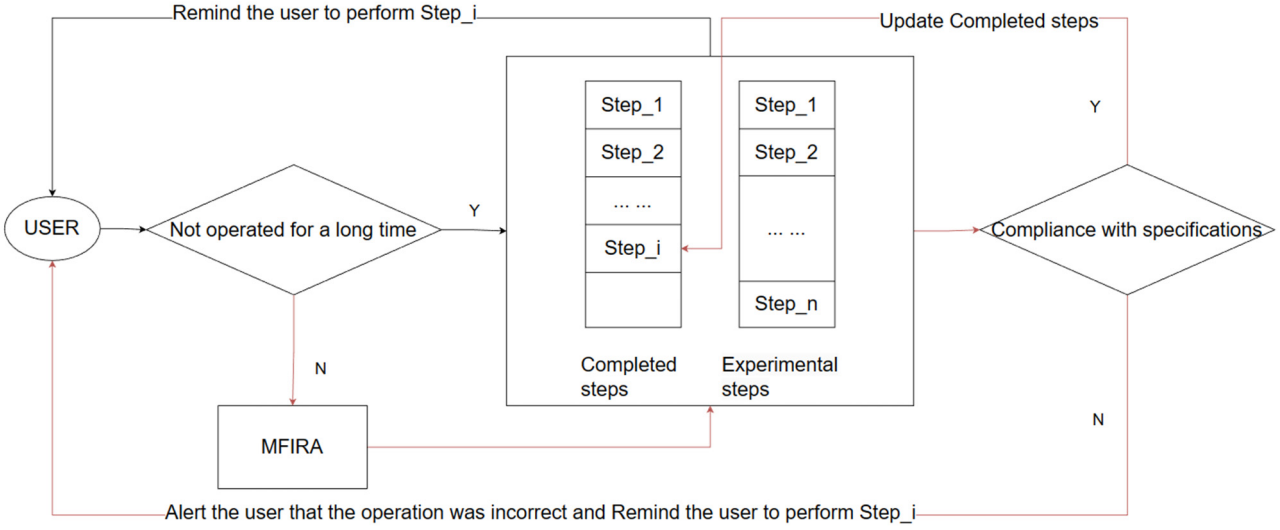
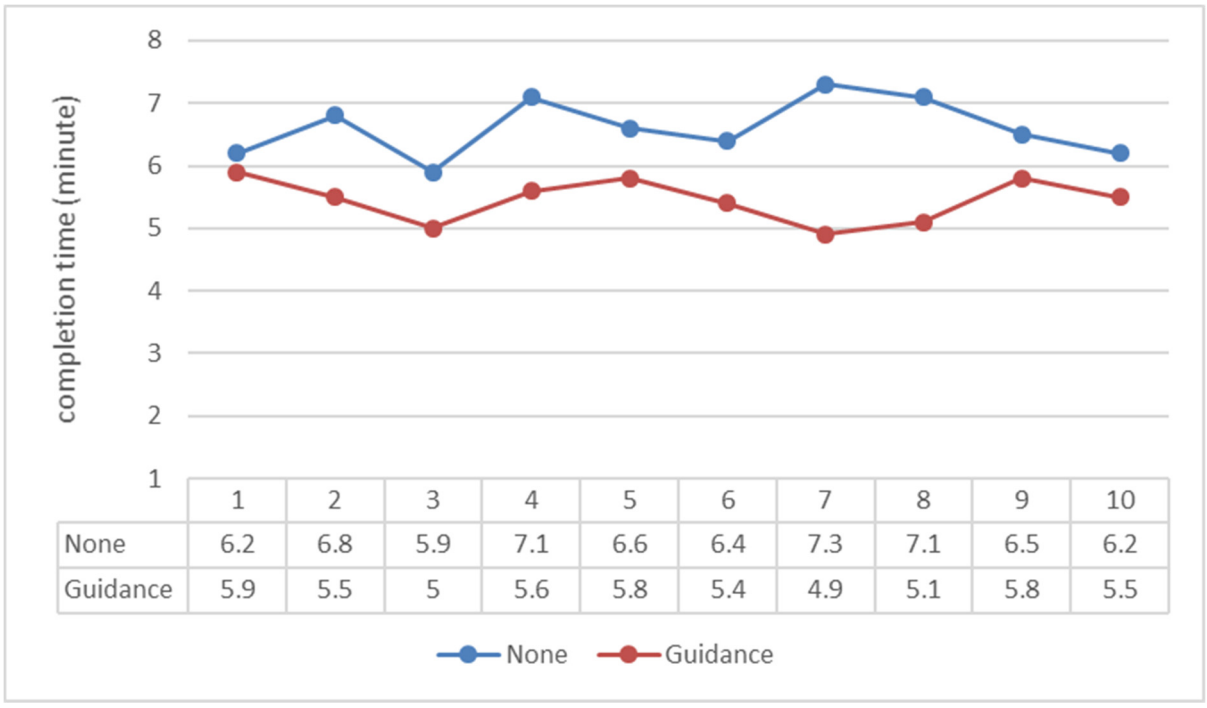


Figure 15. The flowchart for the teaching guidance method.

To verify the method's effectiveness, we invited 20 experimenters and divided them into two groups for the experimental operation. The first group conducted the experiment without the use of any teaching guidance or error correction. The second group utilized the MFIRA-based teaching guidance scheme, which provided automated reminders and guidance, along with error correction. If an error occurred in either group, the system undid the mistake and returned to the previous operation interface. We recorded the completion of the experiments of the two groups separately, and the results are shown in Figure 16.



	1	2	3	4	5	6	7	8	9	10
None	6.2	6.8	5.9	7.1	6.6	6.4	7.3	7.1	6.5	6.2
Guidance	5.9	5.5	5	5.6	5.8	5.4	4.9	5.1	5.8	5.5

Figure 16. Completion times of students with and without guidance in AR experiments.

The first group consisted of 10 testers, whose average completion time was 6.61 min. The second group consisted of 10 testers who utilized the instructional guidance program; their average completion time decreased to 5.45 min. The guidance program resulted in a 17.55% improvement in the average completion time. We analyzed the operations of

the second group of testers, and recorded the number of operation errors, as well as the number of errors that were corrected by the system. The results are presented in Figure 17.

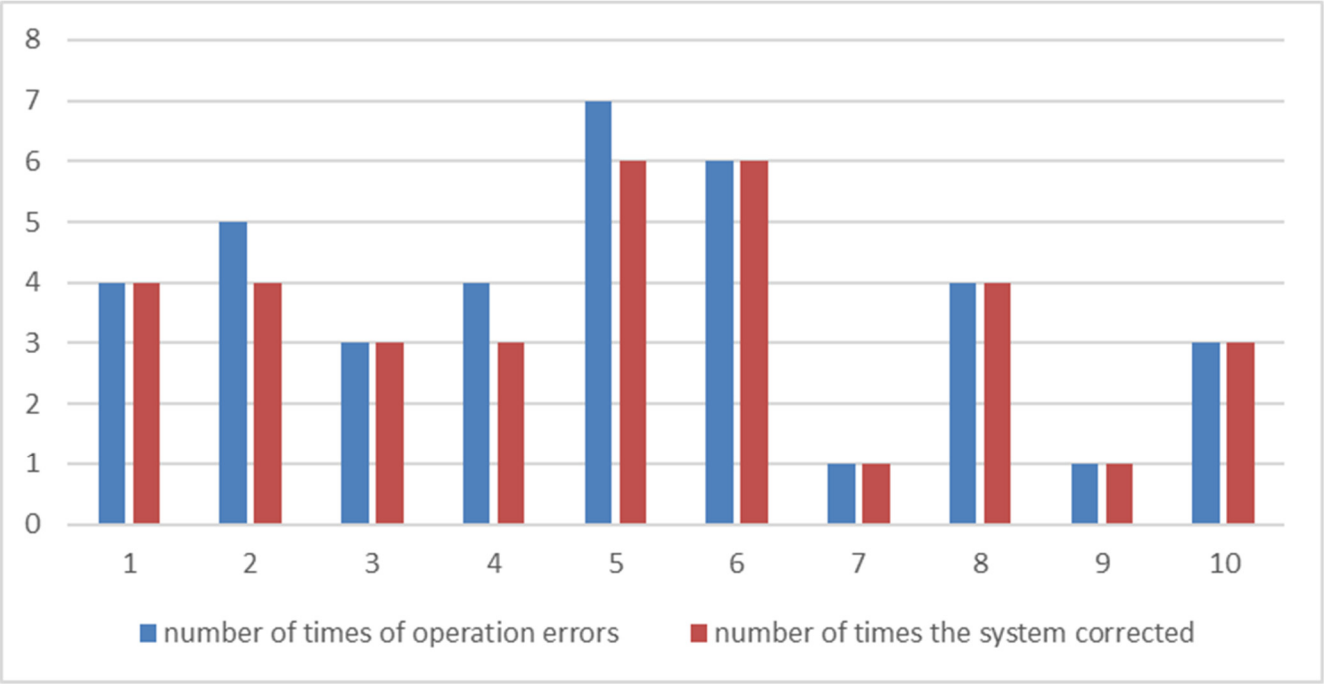


Figure 17. Comparison of the number of errors in the experiments and error corrections in the system.

In total, 160 experimental operations were conducted among ten testers, resulting in 38 errors being made. The system corrected 35 of these errors, resulting in an error correction rate of 92.11%. Therefore, the guidance scheme effectively improved the operation efficiency among users, guiding students to complete experimental operations. This has the potential to enhance the teaching effectiveness in actual secondary school education.

5.2.4. Comparison Experiments

Several scholars have utilized multimodal fusion algorithms to discern the experimental intentions of users in virtual laboratory settings for high school chemistry instruction. Xiao et al. [33] developed a virtual teaching scenario based on the Kinect platform and proposed a multimodal fusion intention recognition algorithm (TMFA), where users operated in front of a KinectV2 camera utilizing voice commands to facilitate their experimentation. Despite using multi-channel data information during the experimentation process, the TMFA algorithm essentially fuses serially in regard to the multimodal information. This means that only one information channel is used during each intention recognition, such as using voice commands to recognize intention A or gestures to recognize intention B. In contrast to the TMFA, this paper used the MFIRA algorithm, which fuses parallel information for several channels simultaneously to extract the experimental intentions of users. In order to confirm whether the MFIRA algorithm has superior intention recognition capabilities to the TMFA algorithm, we invited eight experimenters to complete the "Sodium Water Reaction" experiment on both the Kinect and Hololens 2 platforms. We compared and investigated the results of both the TMFA algorithm and the MFIRA algorithm in the comparative study, which included the average intention recognition rates and completion times of the experiment, as shown in Figure 18.

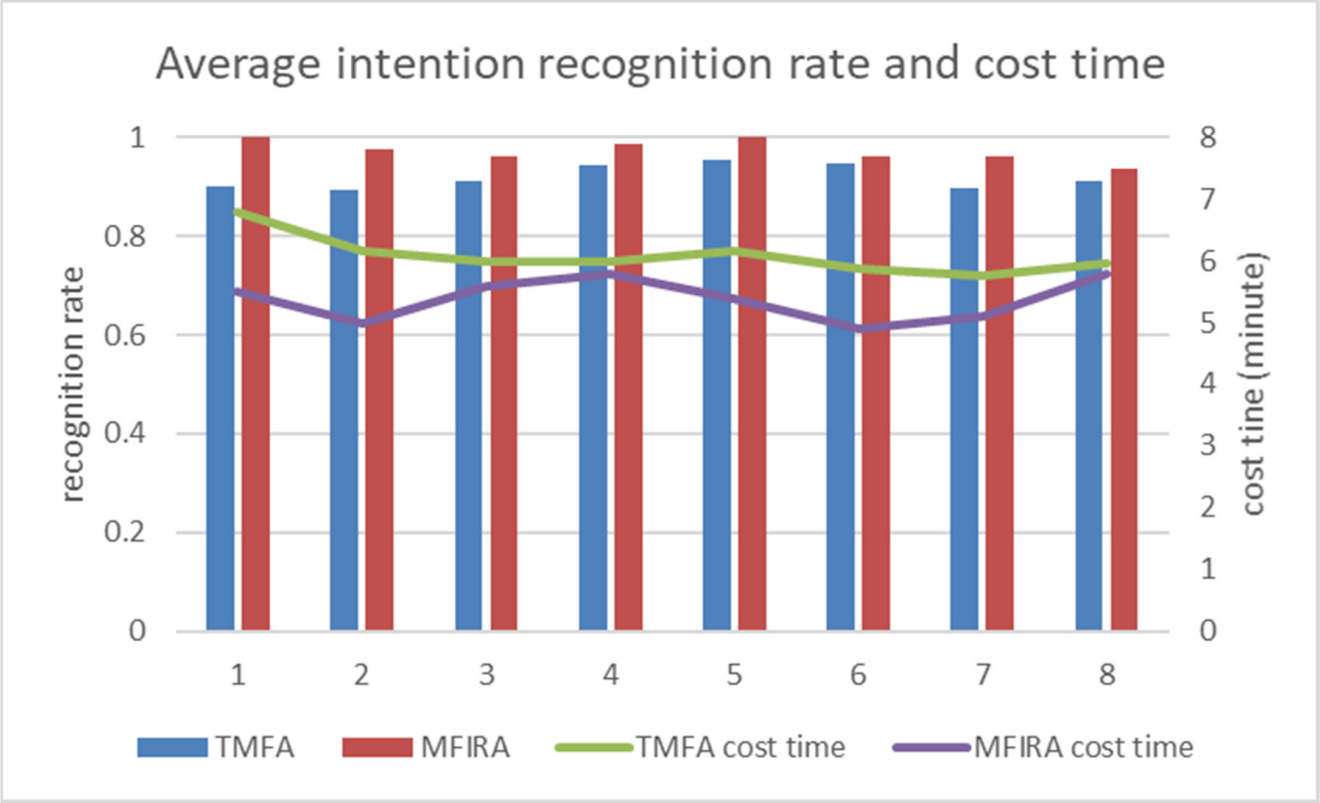


Figure 18. Average intention recognition rate and cost time.

In Figure 18, we can observe that the mean intention recognition rates for both algorithms exceeded 90%, with the accuracy of the TMFA algorithm being 92.04% and that of the MFIRA algorithm being 97.34%. This indicates that both algorithms can accurately recognize the experimental intentions of users. However, in comparison to the TMFA algorithm, the MFIRA algorithm had a recognition rate 5.31% higher. Additionally, it is evident that under the guidance of the MFIRA algorithm, the completion time for the experiment was significantly reduced. During the testing phase, the average completion time for the eight testers based on the TMFA algorithm was approximately 6.1 min, while the average completion time based on the MFIRA algorithm was 5.38 min. These results show that the MFIRA algorithm can aid users in completing experiments more accurately and efficiently.

5.2.5. Cognitive Load and User Evaluation

To evaluate whether the AR experiment system designed in this paper reduces users' cognitive load, a set of controlled experiments was conducted. The testers used four experimental platforms in a single day, including the NOBOOK platform (NOBOOK), which is mainly operated by keyboard and mouse; the traditional virtual experiment platform (Zeng et al. [3]); the virtual platform with the help of a Kinect device (Xiao et al. [33]); and the Hololens 2. After completing each experiment, NASA evaluations were conducted based on six user evaluation metrics: mental demand (MD), physical demand (PD), time demand (TD), performance (P), effort (E), and frustration (F). The NASA evaluation metrics [34] were based on a 5-point scale, where 0-1 indicated a low cognitive load, 1–2 indicated a relatively low cognitive load, 2–3 indicated an overall cognitive load, 3–4 indicated a relatively high cognitive load, and 4–5 indicated a very high cognitive load. The results are illustrated in Figure 19.

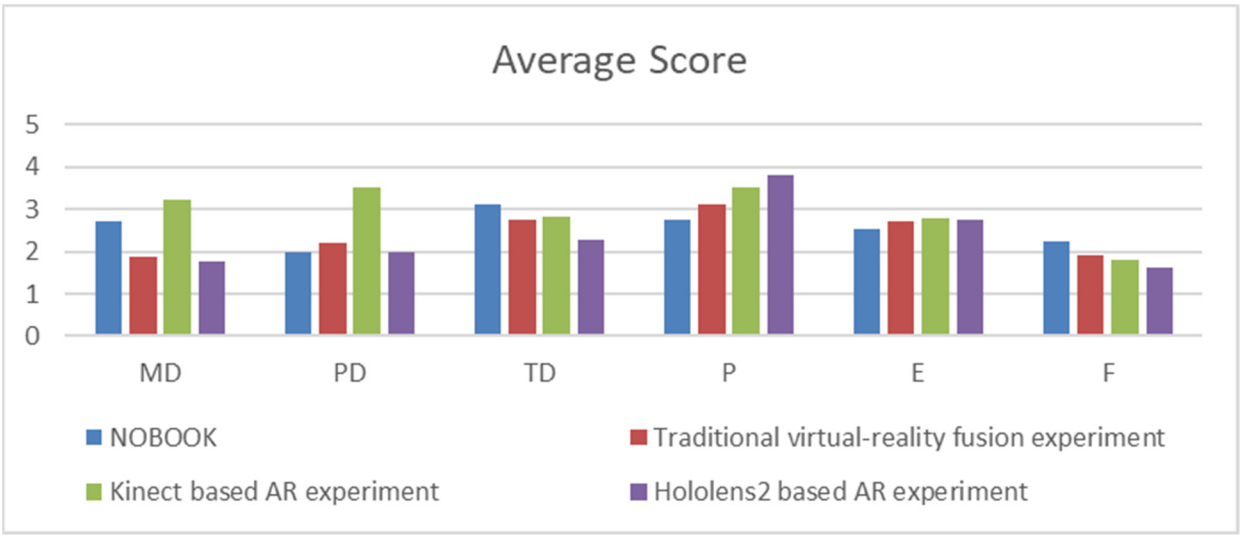


Figure 19. Comparison of operational loads.

The graph indicates that the AR experiment system using Hololens 2 had lower MD and TD scores compared to other platforms, indicating that users found our experimental process simpler to use. This is because using other platforms requires volunteers to understand various functions of each platform beforehand, such as the NOBOOK experiment platform and the operation process based on a Kinect system. Moreover, the Hololens and Kinect systems scored higher on the P-indicators. Volunteers mentioned that operating the experiment on other platforms required more effort in terms of understanding the platform itself; however, when using the Hololens system, volunteers were better able to focus on the experimental phenomenon and results. By observing the phenomenon through the screen and the system's explanation of the experimental mechanism, the experimenter was also able to deepen their understanding of the experimental phenomenon. Additionally, the intelligent experimental system corrected irregularities in the experimental process, helping volunteers to better understand the key points of the experimental operation.

In summary, compared to other experimental platforms, the AR system designed in this paper enables users to conduct experiments in a more intelligent and natural way, while also improving their experimental immersion and operational ability more effectively.

6. Summary and Outlook

This paper presents the design and implementation of an AR chemistry experiment system based on Hololens 2 and proposes a multimodal fusion intent recognition algorithm (MFIRA). The algorithm features: (1) the fusion of gesture and speech information through machine learning feature-layer fusion to identify user actions and avoid the misidentification of user gestures, which can lead to failure of intent inference during the operation process; (2) concrete modeling of the user's gestures, speech, and visual attention information gathered during the experimental process to obtain the target object probability sequences for each channel and finally fuse the three types of modality information via an objective weighting method; (3) analysis of the conflicts between modality information through the Cartesian product of the identified user action results and target object probability sequences. Finally, the algorithm extracted the users' operational intents in order to guide and correct user operations.

This paper primarily addresses two problems. (1) The memory load problem caused by the lack of perception of user intent in traditional virtual experiment systems due to a single interaction modality was resolved by constructing an AR chemistry experiment platform based on Hololens 2 with multimodal perceptual capabilities. (2) An MFIRA algorithm was proposed for multimodal fusion in AR chemistry experiments, and a novel fusion strategy was designed that would resolve the difficulty of parallel analysis of mul-

tiple modal intention stemming from previous multimodal fusion algorithms, emphasize the correlation between each modality, and improve the accuracy of intent recognition.

The performance of the MFIRA algorithm was verified experimentally and reached an accuracy of 97.3% in terms of inferring user experimental intent, with a correction rate of 92.11% for user errors during operation, based on the MFIRA algorithm. The Hololens 2-based AR chemistry experiment system was evaluated by NASA and was shown to effectively reduce the user's cognitive load during operation compared to other experimental platforms, receiving better feedback.

Author Contributions: Conceptualization, Z.F.; Methodology, Z.X.; Software, Z.X.; Formal analysis, Z.X., D.K. and H.C.; Resources, Z.F.; Writing—original draft, Z.X.; Writing—review & editing, Z.F. and X.Y.; Supervision, Z.F. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

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

Data Availability Statement: Not applicable.

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

References

1. Li, N.; Xiong, Z.; Mo, Z. Research on the Assessment of Cognitive Ability in High School Chemistry Experiments in Northwest China. *Chem. Educ. Teach.* **2020**, *4*, 7–13.
2. NOBOOK Virtual Lab. Available online: https://school.nobook.com.cn/site (accessed on 28 May 2023).
3. Zeng, B.; Feng, Z.; Xu, T.; Xiao, M.; Han, R. Research on intelligent experimental equipment and key algorithms based on multimodal fusion perception. *IEEE Access* **2020**, *8*, 142507–142520. [CrossRef]
4. Aljuhani, K.; Sonbul, M.; Althabiti, M.; Meccawy, M. Creating a Virtual Science Lab (VSL): The adoption of virtual labs in Saudi schools. *Smart Learn. Environ.* **2018**, *5*, 16. [CrossRef]
5. Morozov, M.; Tanakov, A.; Gerasimov, A.; Bystrov, D.; Cvirco, E. Virtual chemistry laboratory for school education. In Proceedings of the IEEE International Conference on Advanced Learning Technologies, 2004. Proceedings, Joensuu, Finland, 30 August–1 September 2004; IEEE: Piscataway, NJ, USA, 2004; pp. 605–608.
6. Tingfu, M.; Ming, G.; Lily, Q.; Gang, Z.; Yong, P. Three-dimensional virtual chemical laboratory based on virtual reality modeling language. In Proceedings of the 2008 IEEE International Symposium on IT in Medicine and Education, Xiamen, China, 12–14 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 491–496.
7. Bogusevschi, D.; Muntean, C.; Muntean, G.M. Teaching and learning physics using 3D virtual learning environment: A case study of combined virtual reality and virtual laboratory in secondary school. *J. Comput. Math. Sci. Teach.* **2020**, *39*, 5–18.
8. Salinas, P.; Pulido, R. Visualization of conics through augmented reality. *Procedia Comput. Sci.* **2015**, *75*, 147–150. [CrossRef]
9. De Castro Rodrigues, D.; de Siqueira, V.S.; da Costa, R.M.; Barbosa, R.M. Artificial Intelligence applied to smart interfaces for children's educational games. *Displays* **2022**, *74*, 102217. [CrossRef]
10. Lenz, L.; Janssen, D.; Stehling, V. Mixed reality voice training for lecturers. In Proceedings of the 2017 4th Experiment@ International Conference (Exp. at'17), Faro, Portugal, 6–8 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 107–108.
11. Wörner, S.; Kuhn, J.; Scheiter, K. The best of two worlds: A systematic review on combining real and virtual experiments in science education. *Rev. Educ. Res.* **2022**, *92*, 911–952. [CrossRef]
12. Chhabria, S.A.; Dharaskar, R.V.; Thakare, V.M. Survey of fusion techniques for design of efficient multimodal systems. In Proceedings of the 2013 International Conference on Machine Intelligence and Research Advancement, Katra, India, 21–23 December 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 486–492.
13. Holzapfel, H.; Nickel, K.; Stiefelhagen, R. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA, 13–15 October 2004; pp. 175–182.
14. Corradini, A.; Mehta, M.; Bernsen, N.O.; Martin, J.; Abrilian, S. Multimodal input fusion in human-computer interaction. In *NATO Science Series Sub Series III Computer and Systems Sciences*; IOS Press: Yerevan, Armenia, 2005; Volume 198, p. 223.
15. Mollaret, C.; Mekonnen, A.A.; Ferrané, I.; Pinquier, J.; Lerasle, F. Perceiving user's intention-for-interaction: A probabilistic multimodal data fusion scheme. In Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 29 June–3 July 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–6.
16. Ge, W.; Cheng, C.; Zhang, T.; Zhang, J.; Zhu, H. User intent for virtual environments. In *Recent Developments in Intelligent Systems and Interactive Applications: Proceedings of the International Conference on Intelligent and Interactive Systems and Applications (IISA2016)*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 296–302.

17. Mounir, S.; Cheng, C. Complex event processing for intent understanding in virtual environments. *Int. J. Comput. Theory Eng.* **2017**, *9*, 185–191. [CrossRef]
18. Yang, M.; Tao, J. Intelligence methods of multi-modal information fusion in human-computer interaction. *Sci. Sin. Informationis* **2018**, *48*, 433–448. [CrossRef]
19. Jiang, R.M.; Sadka, A.H.; Crookes, D. Multimodal biometric human recognition for perceptual human–computer interaction. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2010**, *40*, 676–681. [CrossRef]
20. Hui, P.Y.; Meng, H. Latent semantic analysis for multimodal user input with speech and gestures. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2013**, *22*, 417–429. [CrossRef]
21. Alameda-Pineda, X.; Yan, Y.; Ricci, E.; Lanz, O.; Sebe, N. Analyzing free-standing conversational groups: A multimodal approach. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 5–14.
22. Liu, H.; Fang, T.; Zhou, T.; Wang, L. Towards robust human-robot collaborative manufacturing: Multimodal fusion. *IEEE Access* **2018**, *6*, 74762–74771. [CrossRef]
23. Vu, H.A.; Yamazaki, Y.; Dong, F.; Hirota, K. Emotion recognition based on human gesture and speech information using RT middleware. In Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Taipei, Taiwan, 27–30 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 787–791.
24. Wang, Z.; Fang, Y. Multimodal fusion of spatial-temporal features for emotion recognition in the wild. In *Proceedings of the Advances in Multimedia Information Processing–PCM 2017: 18th Pacific-Rim Conference on Multimedia, Harbin, China, 28–29 September 2017*; Revised Selected Papers, Part I 18; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 205–214.
25. Zhao, R.; Wang, K.; Divekar, R.; Rouhani, R.; Su, H.; Ji, Q. An immersive system with multi-modal human-computer interaction. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 517–524.
26. Pan, Z.; Luo, T.; Zhang, M.; Cai, N.; Li, Y.; Miao, J.; Li, Z.; Shen, Y.; Lu, J. MagicChem: A MR system based on needs theory for chemical experiments. *Virtual Real.* **2022**, *26*, 279–294. [CrossRef] [PubMed]
27. Wang, H.; Feng, Z.; Tian, J.; Fan, X. MFA: A Smart Glove with Multimodal Intent Sensing Capability. *Comput. Intell. Neurosci.* **2022**, *2022*, 3545850. [CrossRef] [PubMed]
28. Pérez-Marín, D.; Paredes-Velasco, M.; Pizarro, C. Multi-mode Digital Teaching and Learning of Human-Computer Interaction (HCI) using the VARK Model during COVID-19. *Educ. Technol. Soc.* **2022**, *25*, 78–91.
29. Oramas, S.; Nieto, O.; Barbieri, F.; Serra, X. Multi-label music genre classification from audio, text, and images using deep features. *arXiv* **2017**, arXiv:1707.04916.
30. Che, W.; Feng, Y.; Qin, L.; Liu, T. N-LTP: An open-source neural language technology platform for Chinese. *arXiv* **2020**, arXiv:2009.11616.
31. Ludwig, C.J.; Gilchrist, I.D. Stimulus-driven and goal-driven control over visual selection. *J. Exp. Psychol. Hum. Percept. Perform.* **2002**, *28*, 902. [CrossRef] [PubMed]
32. Gezeck, S.; Fischer, B.; Timmer, J. Saccadic reaction times: A statistical analysis of multimodal distributions. *Vis. Res.* **1997**, *37*, 2119–2131. [CrossRef] [PubMed]
33. Xiao, M.; Feng, Z.; Yang, X.; Xu, T.; Guo, Q. Multimodal interaction design and application in augmented reality for chemical experiment. *Virtual Real. Intell. Hardw.* **2020**, *2*, 291–304. [CrossRef]
34. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; North-Holland: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.