

Facilitator training system for interactive art appreciation using Large Language Models and Mixed Reality

Ryota Fukuda¹ and Ryo Kurazume²

Abstract—Interactive art appreciation is a method in which multiple participants engage in repeated discussions to interpret artworks. In this approach, a facilitator plays a crucial role by asking questions to the viewers. However, there are currently limited opportunities for facilitator training. To address this, this paper proposes a Mixed Reality (MR) AI system that utilizes MR headsets and a Large Language Model to train facilitators. The system features five virtual viewers, each with distinct personalities and whose facial expressions change through dialogue, allowing users to practice facilitation through interactive dialogue. Contents of what the virtual viewers say are generated by the Large Language Model, GPT-4o. Moreover, in addition to GPT-4o, conversations are conducted using the speech recognition system, Whisper, and the speech synthesis system, Azure Text-To-Speech.

I. INTRODUCTION

Interactive art appreciation [1][2] is a method in which a group of people engage in discussions to interpret artworks together. An image of interactive art appreciation is shown in Fig.1.



Fig. 1. An image of interactive art appreciation (Generated by GPT-4o)

In interactive art appreciation, there are both viewers and a facilitator. In Fig.1, the person on the left serves as the facilitator, while the others are viewers. The facilitator in interactive art appreciation plays the role of asking questions and guiding the discussion. A skilled facilitator is essential for ensuring a smooth and effective interactive art appreciation experience. However, the current training and development of facilitators rely primarily on repeated practice in real settings,

such as art appreciation sessions held in museums. As a result, there is a challenge in securing sufficient opportunities for facilitator training and development.

In this study, we focus on developing a facilitator training system as a solution to this problem. By utilizing mixed reality (MR) headsets and a Large Language Model to virtually simulate interactive art appreciation, we have developed a mixed reality (MR) AI system that enables individuals to train as facilitators without the spatial constraints of museums or other physical locations.

II. RELATED WORKS

Training systems utilizing extended reality (XR) technologies such as augmented reality (AR), virtual reality (VR), and MR have been proposed in various forms. Saufnay et al. [3] proposed a VR training system dedicated to the improvement of public speaking skills. It allows users to practice their speeches in a realistic virtual environment with an interactive audience. Furuno et al. [4] proposed a system that allows users to acquire complaint handling skills in an environment similar to actual work by reproducing situations for handling complaint with interactive customer actors in VR space. Their experiment showed that the training effect on customer service skills such as empathy and honorific expression, and the psychological resistance to handling complaints decreased. Oyanagi et al. [5] proposed VR training simulator for ground staff in airport. Under the scenario of a customer inquiring about a purchased ticket and an airport ground staff member responding, a user takes on the role of airport ground staff member and can repeatedly train in a short period of time on the basis of theory into practice. Nakazawa et al. [6] proposed a training system using AR to support caregivers of people with dementia in acquiring the practical skills needed to communicate effectively with patients. In addition, Miyawaki et al. [7] have proposed a training system for care of dementia patients that combines the system in [6] with a Large Language Model. This system allows users to train their nursing skills while conversing with a virtual dementia patient. The training system proposed in this paper also uses MR and Large Language Models. However, the purpose of this system is to train facilitators for interactive art appreciation, and the training progresses through conversations with multiple avatars.

III. SYSTEM CONFIGURATION

A. Setting the situation for interactive art appreciation

In interactive art appreciation virtually recreated by this system, the user, wearing a headset (Quest 3, Meta), is

¹Ryota Fukuda is with Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan. r-fukuda@irvs.ait.kyushu-u.ac.jp

²Ryo Kurazume is with Faculty of Information Science and Electrical Engineering, Kyushu University, Japan. kurazume@ait.kyushu-u.ac.jp

required to act as a facilitator. There are five virtual viewers, and the discussion takes place around artworks positioned both in front of and behind the user.

In actual interactive art appreciation, the facilitator typically faces away from the artwork, occasionally turning to look at it while engaging with the viewers. To better replicate this real-world scenario, artworks are placed not only in front of the user but also behind them. Additionally, the artwork positioned behind the user is displayed larger than the one in front.

Fig.2 illustrates the virtual space as seen by the user in this system, while Fig.3 shows the artwork positioned behind the user, captured from behind the virtual viewers.



Fig. 2. Virtual interactive art appreciation



Fig. 3. A work of art on the back (View from behind the virtual viewers)

B. Five virtual viewers

1) *Setting the characteristics of virtual viewers:* In interactive art appreciation recreated by this system, there are five virtual viewers, each with a unique name and personality. Below are the names and personalities of each virtual viewer.

- Satomi Hamaguchi : Cheerful, active, kind and sensitive
- Masahiro Sanada : Calm and smart
- Kai Matsubayashi : Very cheerful, positive and hasty
- Tsubasa Ozora : Positive, active and optimistic
- Ume Yasumori : Very negative and aloof

As shown in Fig.4, a prompt containing details such as the name and personality of each virtual viewer is prepared individually and fed into the Large Language Model GPT-4o (OpenAI) to define their characteristics. This setup ensures

that each virtual viewer speaks differently based on their characteristics.

```

### SETTING OF A CHATBOT (YOU) ###
Please refer to the following for character settings.
Character Settings
Name: Satomi Hamaguchi
Number: 1
Gender: Female
Background: Artist
Age: 56 years old
Personality: Cheerful, active, kind and sensitive
Favorite food: Orange
Hobbies: Taking a picture
Hometown: Hokkaido
Art knowledge: A lot of knowledge

```

Fig. 4. Setting the characteristics of the virtual viewer (Part of the prompt)

2) *Virtual viewer's emotions:* Each virtual viewer in this system has an emotion value ranging between 0 and 1. This value dynamically changes throughout the dialogue, increasing when the conversation is positive and decreasing when it is negative, accumulating over time. The changing emotion values are calculated by feeding the prompt shown in Fig.5 into GPT-4o.

```

### SPECIFICATION ###
* The chatbot has virtual emotion parameters in its memory.
* The chatbot's conversation reflects its emotional parameters.
* The chatbot's emotional parameters will fluctuate throughout the conversation.

### EMOTION PARAMETERS ###
* Joy, Sadness, Anticipation, Surprise, Fear, Disgust, Trust
* Each parameter ranges from 0-10.
* For all outputs, first calculate the current chatbot's emotional parameters. However, do not display this emotional parameter.
* Then, output role-played conversation that reflects the chatbot's emotional parameters.

* In addition, in the following conversation, you will receive words, and return an emotion score named "emotion" between 0 and 1 to determine whether the conversation is negative or positive.
* If the conversation is negative, the emotion score is small. If the conversation is positive, the emotion score is high. The emotion score should fluctuate throughout the conversation. The initial emotion score is 0.5.
* In addition, you will return a property score of the words you just received named "message" between 0 and 1 to determine whether the words are negative or positive. If the words are negative, the property score is small. If the words are positive, the property score is high.

* In addition, the tone of your response will change to reflect the current emotion parameter value.

```

Fig. 5. Calculating emotion values (Part of the prompt)

Additionally, the virtual viewers' facial expressions change accordingly: when the emotion value is close to 0, their expressions become gloomy, and when it is close to 1, they appear cheerful. Fig.6 illustrates these facial expression changes.

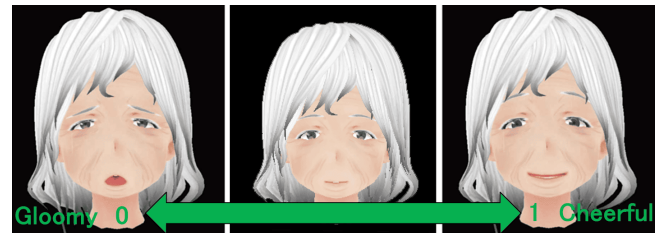


Fig. 6. Changes in the virtual viewer's facial expression

C. Dialogue with virtual viewers

We will explain how the dialogue between the user and the virtual viewers is implemented. Fig.7 illustrates the processing flow of the dialogue between the user and the virtual viewers.

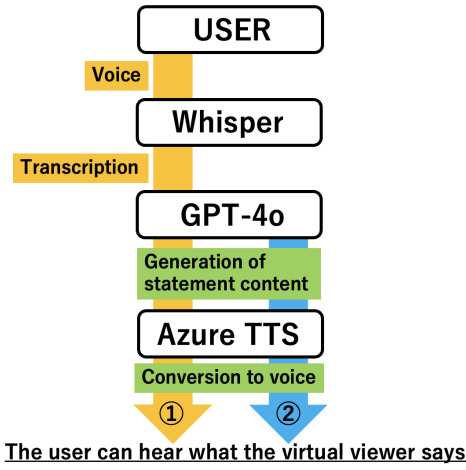


Fig. 7. Processing flow of the dialogue between the user and the virtual viewers

There are two types of dialogue processing flows between the user and the virtual viewers. One occurs when the virtual viewers respond to the user's statement (① in Fig.7), and the other occurs when they speak spontaneously (② in Fig.7).

In the processing flow of ① in Fig.7, the user's voice is first converted into text data using the speech recognition model, Whisper (OpenAI). Next, the text data is input into GPT-4o, which generates the virtual viewer's statement content. Finally, the generated text is converted into voice using the speech synthesis system, Azure Text-To-Speech (Microsoft). In interactive art appreciation, it is important for facilitators to maintain eye contact while conversing with viewers. Therefore, this system is designed so that the virtual viewer who is making eye contact with the user at the moment the user finishes speaking will be the one to respond, meaning that eye contact determines which virtual viewer interacts with the user. However, in real interactive art appreciation, facilitators do not always maintain eye contact while speaking; they may speak while looking at artworks or address the group as a whole without directing their words to a specific individual. To account for this, if the user does not make eye contact with any virtual viewer while speaking, one of the five virtual viewers is randomly selected to respond. Additionally, if the user explicitly addresses one of the virtual viewers by name, that specific viewer will be the one to reply, even if there was no eye contact.

In the processing flow of ② in Fig.7, instead of using text data obtained from Whisper, instructions such as "Speak in accordance with the flow of the conversation" are input into GPT-4o to generate the virtual viewer's statement content. This process occurs when the virtual viewers speak spontaneously, which happens in the following three cases.

- If there is 10 seconds of silence (1 in 2 chance of spontaneous speech)
- After the virtual viewers speak (1 in 5 chance of spontaneous speech)
- If the virtual viewers nominate another person

First, if neither the user nor the virtual viewers speak for

10 seconds, one of the five virtual viewers will speak with a 50% probability. Second, after the virtual viewers speak, there is a 20% probability that one of them will respond to that statement. Third, if a specific virtual viewer is addressed by name, that viewer will speak.

1) *Impression data:* When generating the virtual viewers' statements with GPT-4o, impression data is also provided as input. This impression data was collected from approximately 500 people, who were asked about their thoughts and impressions after viewing the artwork. Specifically, the data is organized into the following three categories.

- What do you think it was trying to express?
- Where did you think that from?
- How did this work of art make you feel?

The reason for feeding this impression data into GPT-4o is to ensure that the virtual viewers' responses are as close as possible to what real people actually think and feel when viewing the artwork.

The impression data includes both positive and negative content. Therefore, the data is categorized into positive and negative groups. Since each virtual viewer in this system has a predefined personality, positive virtual viewers are assigned positive impression data, while negative virtual viewers are assigned negative impression data. By assigning different impression data based on the virtual viewers' personalities, this system ensures that their responses are more consistent with their character traits. Fig.8 shows an example of positive impression data, and Fig.9 shows an example of negative impression data.

• What do you think it was trying to express?	Hope and the bright future.
• Where did you think that from?	Because the stars and moonlight shine brightly in the night sky.
• How did this work of art make you feel?	I felt hope.

Fig. 8. An example of positive impression data

• What do you think it was trying to express?	Anxious Night.
• Where did you think that from?	Swirling clouds.
• How did this work of art make you feel?	I felt turbulent and restless.

Fig. 9. An example of negative impression data

Here, we conducted an evaluation to verify the effectiveness of using the impression data in order to make the virtual viewer's responses more similar to human impressions. The evaluation used the CLAIR [8], which quantifies the similarity between sentences on a scale from 0 to 100 using a Large Language Model, and it has been shown to correlate more highly with human judgment compared to

traditional methods. The specific evaluation method involved selecting one virtual viewer and asking 20 different questions under two conditions: with impression data and without impression data provided to GPT-4o. The similarity between each response and the impression data was then quantified. Table I shows the similarity scores for both conditions, with the responses to the questions omitted, based on the results obtained from CLAIR. In addition, box plots of the obtained scores are shown in Fig.10.

TABLE I
SIMILARITY SCORE WITH IMPRESSION DATA

Question	Use	Not use
What do you think about this painting?	95	90
Tell us your impression of this painting.	95	85
Tell us how you felt when you saw this painting.	90	85
Which part are you interested in?	95	90
Is there anything you like about this painting?	90	85
What was your first impression of this painting?	95	85
Where did you think that from?	95	85
Is there anything else?	85	85
What do you think from that?	85	50
Tell us what you think when you look at this painting.	90	85
Tell us whatever you see.	85	85
Tell us anything about this painting.	85	85
Is there anything you are interested in?	85	70
Why did you think that?	90	45
What do you think the artist of this painting focused on?	95	85
What do you think the artist intended when he drew this painting?	95	85
Think about what the artist was feeling when he drew this painting.	85	70
What do you think the artist was trying to express?	85	85
Where did you get that idea from?	95	85
How did you feel when you saw this painting?	80	75
AVERAGE	89.8 ±5.0	79.8 ±12.3

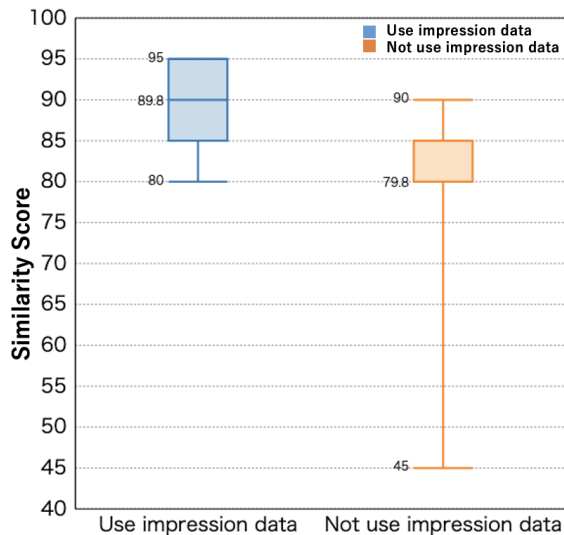


Fig. 10. Box plots of similarity score with impression data

The similarity score with the impression data was an average of 89.8 with a standard deviation of 5.0 when the impression data was used, and an average of 79.8 with a standard deviation of 12.3 when it was not used. Therefore, when the impression data was used, the responses were more consistently similar to the impression data, indicating that feeding impression data into GPT-4o is effective.

2) *Saying prepared phrases:* Since the process of generating a response from the virtual viewers involves Whisper, GPT-4o, and Azure Text-To-Speech, as shown in Fig.6 ①, there is an unavoidable silent waiting time after the user speaks. This resulted in a problem where the silent period felt too long, making it unclear whether the virtual viewers were reacting. To address this, we implemented a mechanism where pre-prepared phrases are spoken by the virtual viewers during the waiting time before generating a full response. Specifically, we prepared 10 different filler phrases, such as "Yes, yes..." or "Uh, let me see...", which do not disrupt the natural flow of the conversation. One of these phrases is randomly selected and spoken while the response is being generated.

Here, to evaluate the effectiveness of the pre-prepared phrases, we measured 10 times the silent waiting time occurring before the virtual viewer responded to the question, "What do you think about this painting?", in both conditions: when the pre-prepared phrases were spoken and when they were not. By comparing the average silent time between the user's speech and the virtual viewer's response in both conditions, we assessed the effectiveness of the pre-prepared phrases. The results are shown in Table II.

TABLE II
MEASUREMENT RESULTS OF SILENCE TIME BEFORE RESPONSE

	With prepared phrases	Without prepared phrases
1	2.41s	6.61s
2	3.33s	4.55s
3	4.43s	5.43s
4	2.00s	4.26s
5	2.71s	6.53s
6	2.40s	4.73s
7	2.10s	4.72s
8	2.45s	5.56s
9	2.57s	5.36s
10	1.34s	4.91s
AVERAGE	2.57s	5.27s

The silent waiting time between the user's speech and the virtual viewer's response was 2.57 seconds on average when pre-prepared phrases were used, compared to 5.27 seconds on average when they were not used. This means that the silent time was reduced by approximately half when pre-prepared phrases were used. Therefore, using pre-prepared phrases effectively reduces silent waiting time, making it easier for the user to recognize that the virtual viewers are responding, proving the effectiveness of this approach.

3) *Examples of dialogue with the virtual viewers:* Examples of dialogue with the virtual viewers are shown in Fig.11.

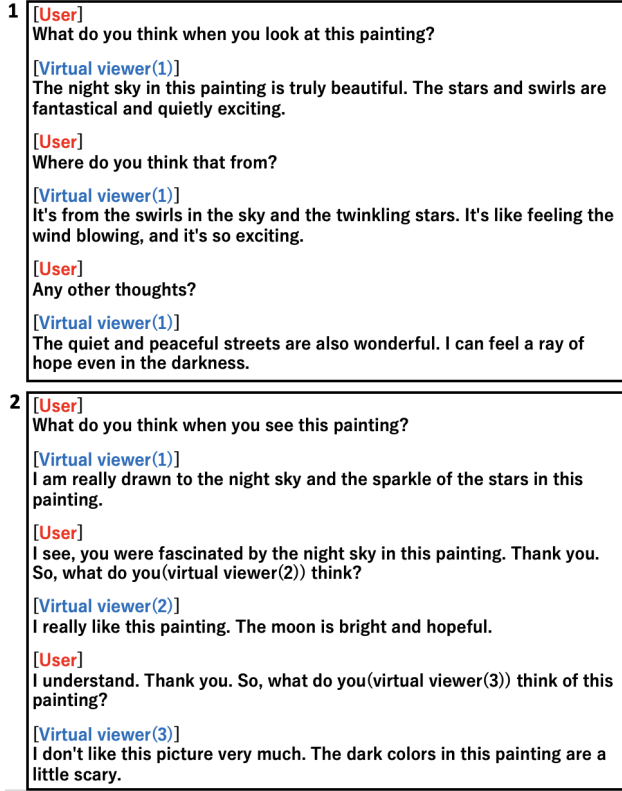


Fig. 11. Examples of dialogue with virtual viewers

D. Instruction Mode

In interactive art appreciation, facilitators utilize techniques such as paraphrasing, summarizing, and focusing, as described below. It is essential to effectively apply these techniques.

- Paraphrasing : Accurately rephrasing the viewer's statement
- Summarizing : Linking multiple statements and providing a summary
- Focusing : Shifting the conversation to a different topic

In this system, there is a feature called "Instruction Mode" that allows the user to train in these three techniques. In Instruction Mode, the user is guided on which technique to apply next. The instructions are displayed to the user after the virtual viewers have spoken, as shown in Fig.12.



Fig. 12. Displayed instructions (Paraphrasing)

The displayed instructions change progressively, as shown

in Fig.13. In the first stage, a paraphrasing instruction appears. However, this paraphrasing instruction is not always displayed; it appears with a 50% probability. After a total of three paraphrasing instructions have been given, the system transitions to the next stage, where a summarizing instruction is displayed. Once a summarizing instruction has been displayed once, the system moves to the final stage, where a focusing instruction appears. After one focusing instruction, the system returns to the first stage, where a paraphrasing instruction is displayed again.

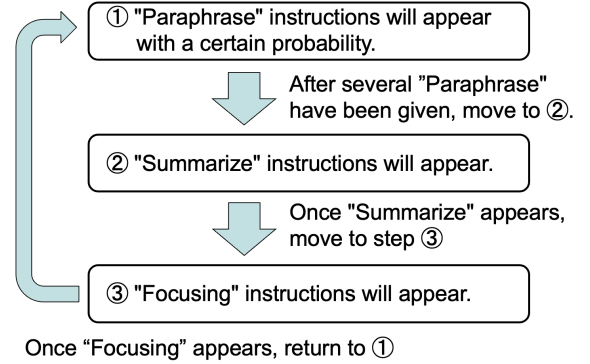


Fig. 13. Flow of displayed instructions

In Instruction Mode, the system not only provides instructions but also evaluates whether the user was able to follow them correctly on a five-level scale and gives appropriate advice based on the evaluation. The evaluation results are displayed in five levels: "1 Worst," "2 Bad," "3 Good," "4 Great," and "5 Perfect," along with corresponding advice. Fig.14 shows an example of the evaluation and advice display.

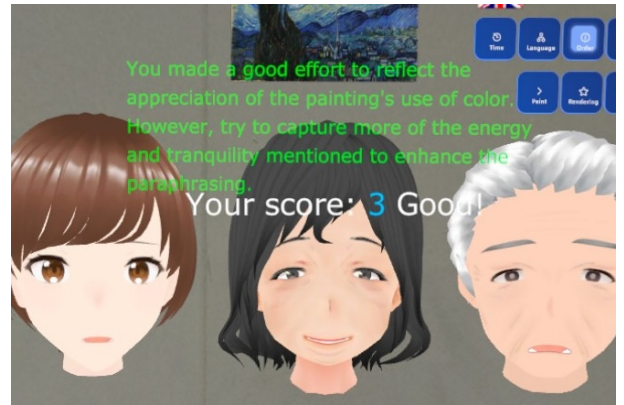


Fig. 14. Evaluation and advice in Instruction Mode

The evaluation and advice are generated using GPT-4o/Claude by inputting the past statements of the virtual viewers, the user's statements, the artwork image, and prompts related to evaluation and advice.

IV. EVALUATION EXPERIMENT

To evaluate this system, identify areas for improvement, and contribute to future development, an evaluation ex-

periment was conducted. In this experiment, 12 facilitator trainees participating in the "Koto! Koto? Kawasaki" project in Kawasaki City, Kanagawa Prefecture, Japan, experienced this system and completed a questionnaire with a five-point rating scale (1 to 5). Fig.15 shows the scene during the experiment. Additionally, Table III presents the questionnaire items and the average score for each item.



Fig. 15. The scene during the experiment

TABLE III
AVERAGE SCORE FOR EACH QUESTIONNAIRE ITEM

Questionnaire	Score
A:Do the virtual viewers' facial expressions help you learn?	3.83
B:Is the displayed time useful for learning?	3.91
C:Do conversations with the virtual viewers help you learn?	3.92
D:Are the displayed faces friendly?	2.42
E:Are the virtual viewers close to the natural behavior of real humans?	2.33
F:Do you think instructions about what to say next would help you learn?	3.36
G:Does this system help you learn interactive art appreciation?	3.83
H:Are you interested in this system?	4.33

Items A, B, C, G, and H received relatively high ratings, with an average score of 3.8 or higher. This indicates that dialogue with the virtual viewers whose facial expressions change was evaluated as having a certain level of effectiveness in learning interactive art appreciation. Additionally, many participants showed interest in this system after experiencing it.

V. FUTURE PLANS

Currently, the dialogue between the user and the virtual viewers utilizes Whisper, GPT-4o, and Azure Text-To-Speech, which results in a long processing time and slow response speed. Therefore, we are exploring a method to achieve dialogue between the user and the virtual viewers using only GPT-4o, without relying on Whisper and Azure Text-To-Speech.

In actual interactive art appreciation, as the dialogue progresses, people not only deepen their understanding of the artwork but also get to know each other's personalities and characteristics. To bring interactions with the virtual viewers

closer to this kind of human interaction, we plan to assign backgrounds and personal histories to each virtual viewer, allowing their personalities and traits to be reflected in their speech.

VI. CONCLUSIONS

We have developed a system that utilizes a Large Language Model and MR headsets to train and cultivate facilitators for interactive art appreciation. In this system, users can engage in dialogue with virtual viewers about artworks displayed in a mixed reality space, allowing them to train alone regardless of location. Additionally, by utilizing features such as Instruction Mode, users can refine their facilitation techniques.

In the future, we aim to make the behavior of the virtual viewers more similar to that of real humans, enabling this system to provide more practical training, similar to training conducted between real people.

ACKNOWLEDGMENT

This study was supported by the JST Co-Creation Field Formation Support Program, "Co-Creation Hub for Art Communication to Build an Inclusive Society" (JPMJPF2105). We would also like to express our gratitude to Professor Hideaki Kawabata of Keio University for providing the impression data, as well as to facilitators Ms. Noriko Kondo and Ms. Miyuki Kasuga for their cooperation in the development of this system.

REFERENCES

- [1] P. Yenewine, Art Communication Research Center of Kyoto University of the Arts, "[Art Appreciation to Improve Academic Ability : Visual Thinking Strategies — Where does that idea come from?] Gakuryoku wo nobasu bijyutukannsyou : Visual Thinking Strategies — Doko kara sou omou? (in Japanese)," Tankosha Publishing Co., Ltd., 2015.
- [2] S. Inaniwa, T. Ito, Y. Kono, C. Suzuki, and Y. Watanabe, "[Museum Thinking for Children and Adults] Kodomo to otona no tame no museum sikou (in Japanese)," Sayusha Co., Ltd., 2022.
- [3] S. Saufnay, E. Etienne, and M. Schyns, "Improvement of Public Speaking Skills using Virtual Reality: Development of a Training System," in Proceedings of 2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp.1-3, IEEE, 2024.
- [4] T. Furuno, W. Donghao, S. Fujita, Y. Omi, K. Okawara, K. Shiratori, H. Nishizaki, T. Utsuro, and J. Hoshino, "Scenario-Based Customer Service Training System With Conversational Customer Actors," The Journal of the Society for Art and Science, Vol.21, No.2, pp.46-54, 2022.
- [5] A. Oyanagi, K. Aoyama, R. Ohmura, T. Tanikawa, and M. Hirose, "Training simulator for service industry using virtual reality environment," Transactions of the Virtual Reality Society of Japan, Vol.25, No.1, pp.78-85, 2020.
- [6] A. Nakazawa, M. Iwamoto, R. Kurazume, M. Nunoi, M. Kobayashi, and M. Honda, "Augmented reality-based affective training for improving care communication skill and empathy," PLOS ONE, Vol.18, No.7, e0288175, 2023, DOI:10.1371/journal.pone.0288175.
- [7] T. Miyawaki, Y. Nishiura, R. Fukuda, K. Nakashima, and R. Kurazume, "Development of Dementia Care Training System using AR and Large Language Model," in Proceedings of 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp.903-908, 2024, DOI:10.1109/SMC54092.2024.10831644.
- [8] D. M. Chan, S. Petryk, J. E. Gonzalez, T. Darrell, and J. Canny, "CLAIR : Evaluating Image Captions with Large Language Models," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.13638-13646, 2023.