First impression assessment of digital human applicant images generated with posture prompts and text prompts

Wataru GANAHA, Shota HIOKI, and
Masashi NISHIYAMA[0000−0002−5964−3209]

Graduate School of Engineering, Tottori University
101 Minami 4-chome, Koyama-cho, Tottori, 680-8550, Japan
d23t5112h@edu.tottori-u.ac.jp

Abstract. We investigated whether a digital human applicant generated by a combination of posture prompts and additional text prompts elicited a good or bad first impression during an interview in a virtual space. Existing analytical studies have not evaluated whether a combination of posture prompts and additional text prompts can improve an interviewer's first impression of a digital human applicant in an interview setting. To examine this issue, we generated images of digital human applicants by combining the presence or absence of posture prompts and the presence or absence of additional text prompts. We conducted subjective assessments in which participants simulating interviewers in a virtual space reported their first impressions of the images. The experimental results demonstrated that interviewers' first impressions of a digital human applicant generated by combining posture prompts and additional text prompts were better than their first impressions of a digital human applicant generated without prompts.

Keywords: Digital human · First impression · AI image generator · Text prompt · Posture prompt.

1 Introduction

The use of digital humans in virtual interviews has attracted attention in recent years [10, 2, 11, 3, 7]. In the current study, digital human refers to an entity in a virtual space that realistically imitates the appearance and behavior of a real person. The use of digital humans is designed to help interactions in virtual interviews feel natural. Like interviews in real space, interviews in virtual space generally involve an applicant and an interviewer. An application [10, 2] involving digital humans for interview training in virtual space was proposed in a previous study, and an analytical study [11, 3] examined the use of a digital human as an interviewer. These innovations could potentially enable people with disabilities to practice interviews by interacting with a digital human interviewer. A more recent study [7] examined a paradigm in which an interviewer interacted in a virtual space with a digital human applicant representing a real person but exhibiting different "physical" characteristics. As in that study [7], the current study examined a case in which the appearance of a digital human applicant differed from that of the real person it represented.

In real-world interviews, an interviewer's first impression of an applicant is often an essential factor influencing their decision-making [5]; this is also the case in virtual space interviews in which the interviewer's first impression is of a digital human applicant. The appearance of a digital human applicant, therefore, may play an important role in determining the interviewer's first impression. Various techniques have previously been proposed to generate digital humans for virtual spaces, typically using designer-created 3D models. In recent years, services such as MetaHuman[1] and Avatar Cloud Engine[2] have been launched to generate digital humans that closely resemble real-world humans.

In both virtual space and real-world interviews, the interviewer should be able to see the applicant's entire body, from the feet to the head, to facilitate a good first impression. In a virtual space interview, it is often preferable that the digital human applicant's whole-body appearance is similar to that of a real-world human. The services mentioned above emphasize the realism of the face and hair and generate digital humans focusing on the upper body.

In the current study, we considered the use of artificial intelligence (AI)-based image generation such as Stable Diffusion [9] and DALL-E [8] to generate the whole-body appearance of digital human applicants. When using AI-based image generators, text prompts are generally employed on a trial-and-error basis. Basic text prompts describing a person's characteristics, clothing, posture, and context are often used, and additional prompts describing supplementary elements are commonly used to modify the generated image results. However, properly controlling the digital human applicant's posture using only text prompts is time-consuming and difficult. We aimed to generate posture-controlled images by providing posture prompts using ControlNet [13]. To the best of our knowledge, no previous studies have investigated whether the combination of posture and text prompts can improve the first impression of a digital human applicant in an interview setting.

The current study involved subjective assessments of the first impression of an applicant's image generated by a combination of posture and text prompts, assuming an interview in a virtual space. Figure 1 shows the aim of this paper. Specifically, we investigated the following hypothesis H1 under the condition that basic text prompts are commonly used to describe a person's characteristics, clothing, posture, and context when generating images.

H1 : In the virtual space, an interviewer perceives a good first impression from a digital human applicant's image when it is generated by a combination of posture prompts that control the applicant's posture and additional text prompts that modify the generated results.

[1] https://www.unrealengine.com/metahuman
[2] https://developer.nvidia.com/ace

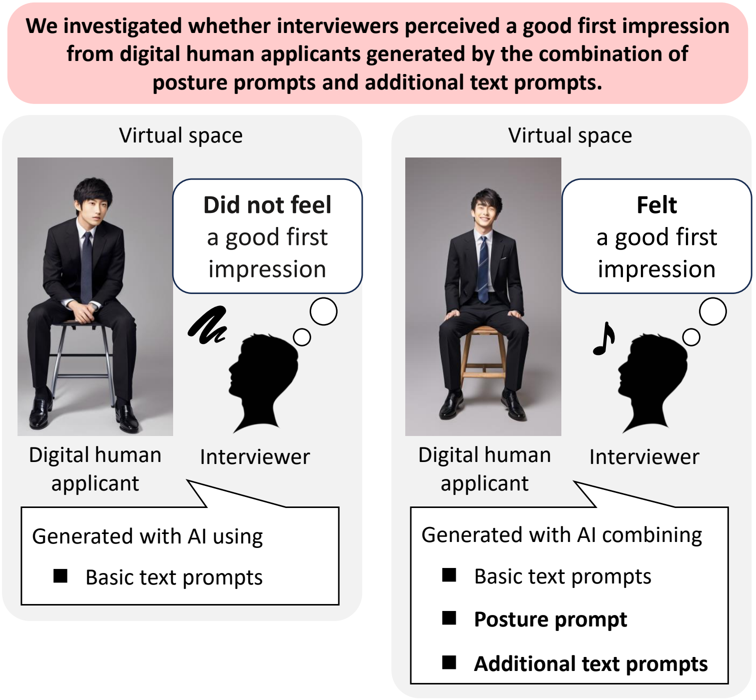

We investigated whether interviewers perceived a good first impression from digital human applicants generated by the combination of posture prompts and additional text prompts.
Virtual space
Did not feel a good first impression
Digital human applicant
Interviewer
Generated with AI using
■ Basic text prompts
Virtual space
Felt a good first impression
Digital human applicant
Interviewer
Generated with AI combining
■ Basic text prompts
■ Posture prompt
■ Additional text prompts

Fig. 1. The aim of this paper.

To investigate hypothesis H1, we generated stimulus images of digital human applicants with and without posture prompts, and with and without additional text prompts. We conducted a subjective assessment by instructing participants to simulate an interviewer in a virtual space, and to report whether their first impression of the digital human image was good or bad. The experimental results showed that the digital human applicants combined with the posture prompt and the additional text elicited a better first impression than the digital human applicants without the posture prompt.

2 Experimental design for subjective assessment

2.1 Overview

We used a subjective assessment method to investigate whether a virtual space interviewer perceived a good first impression from images of digital human applicants generated by a combination of posture prompts and additional text prompts. In this experiment, we generated digital humans representing applicants by combining the presence or absence of posture prompts and additional text prompts, while providing basic text prompts. We conducted subjective assessments, in which participants, simulating interviewers in a virtual space, evaluated their first impression of the digital human applicants as good or bad.

In preparing the stimulus images, we set the conditions for combining prompts to generate digital humans as follows:

C1: without posture prompts or additional text prompts
C2: with posture prompts and without additional text prompts
C3: without posture prompts and with additional text prompts
C4: with both posture prompts and additional text prompts

The details of the prompts used are described in Section 2.2, and the procedure for generating the stimulus images is outlined in Section 2.3.

In the subjective assessment, participants simulating interviewers assessed their first impressions of the digital human applicants. The participants were equipped with a head-mounted display (HMD) and presented with the stimulus images in a virtual space. Participants were asked to report their first impression of the digital human applicants shown in the stimulus images. The participants were instructed to assume that the digital applicants were Japanese individuals engaged in job hunting and entrance examinations. Details of the participants are described in Section 2.4, the experimental setup is outlined in Section 2.5, and the experimental procedure is detailed in Section 2.6.

2.2 Prompts for generating stimulus images

Text prompts The basic text prompts and additional text prompts used to generate the stimulus images are explained below. Table 1 shows the basic text prompts. These prompts included words that describe the characteristics, clothing, posture, and context of the person. Specifically, the digital human was described as being Japanese, wearing a business suit, and seated in a chair. Additionally, text prompts were included to ensure that the entire body of the applicant, from feet to head, was visible to the interviewer. Text prompts were also included to output images that the generative AI would evaluate as high quality. We incorporated text prompts indicating a plain background to suppress the inclusion of unspecified elements in the images. We represented the applicants in two variations (male and female) and prepared separate basic text prompts for each.

Table 2 shows the additional text prompts. The digital humans generated using only the basic text prompts exhibited variations in expression, hand position, and gaze. Specifically, the expressions of the digital humans ranged from smiling to stern, while hand positions varied between being crossed on the knees and hidden behind the body. The gaze also varied, with some looking straight ahead and others looking to the side. To improve the first impression of the digital humans, we also included words in the additional text prompts to control for a smiling expression, hands resting on the knees, and gaze directed forward.

Table 1. Basic text prompts used for generating stimulus images.

Basic text prompts (Male applicant)	full-body, best-quality, realistic, a- 20-year-old-Japanese-man, wearing-black-business-suit, derby-tie, seated-posture-on-a-folding-chair, no-background
Basic text prompts (Female applicant)	full-body, best-quality, realistic, a-20-year-old-Japanese-woman, wearing-black-business-suit, seated-posture-on-a-folding-chair, no-background

Table 2. Additional text prompts used for generating stimulus images.

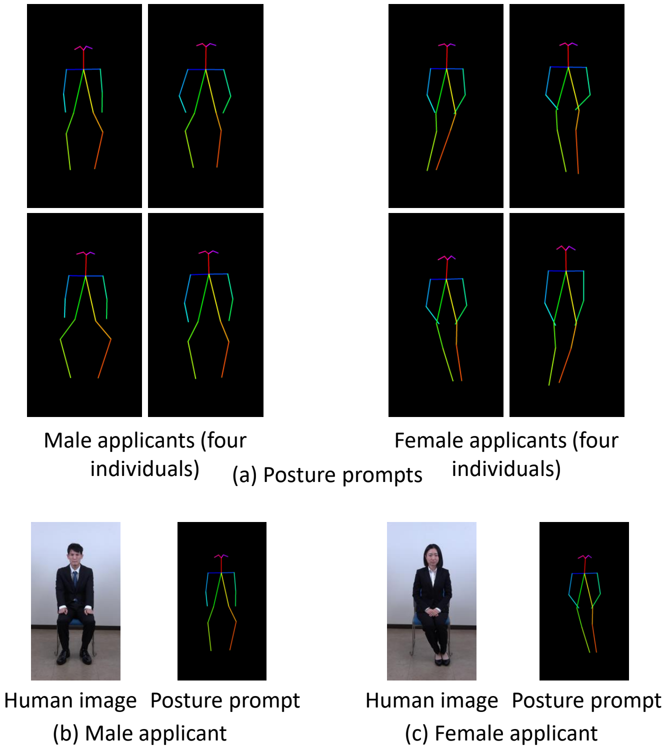
Additional text prompts	smile, hands-on-lap, look-at-viewer

Posture prompts The posture prompts used to control the posture of the digital humans in the stimulus images are explained below. First, subjects simulating applicants for the interview were photographed to obtain images of individuals. Next, the posture was estimated from the obtained images, which were then used as posture prompts. Figure 2 shows examples of the individual images and the posture prompts. When taking the photographs, the subjects were instructed to assume a posture that they thought would make a good impression on an interviewer if they were the applicants. Images of three male applicants and three female applicants were prepared respectively. Figure 2(a) shows an example of a male applicant. Figure 2(b) shows an example of a female applicant. OpenPose [4] was used to estimate the postures. Figure 2(c) shows an example of the posture prompts derived from the individual images. For males, the posture prompts often indicated sitting with a straight back and feet shoulder-width apart. For females, the posture prompts frequently indicated sitting with a straight back and feet together. We used ControlNet [13] to provide the posture prompts to the AI model.

2.3 Procedure for generating stimulus images

The procedure for generating the stimulus images is explained below. Following the conditions set in Section 2.1, images of digital human applicants were generated by combining the presence or absence of posture prompts and additional text prompts. Figure 3 shows examples of these images. We used Stable Diffusion [9] as the AI-based image generator. We used Beautiful Realistic Asians[3] for the AI model. For the stimulus images presented to the participants, 12 images were selected for each of the four conditions set in Section 2.1. The breakdown included six images of male applicants and six images of female applicants. The total number of stimulus images presented to each participant was calculated as 6 (images) × 2 (genders) × 4 (conditions) = 48 images. Images in which the

[3] https://civitai.com/models/25494/beautiful-realistic-asians



Male applicants (four individuals)

Female applicants (four individuals)

(a) Posture prompts

Human image Posture prompt

(b) Male applicant

Human image Posture prompt

(c) Female applicant

Fig. 2. Posture prompts obtained using a body keypoint detection technique from images of humans.

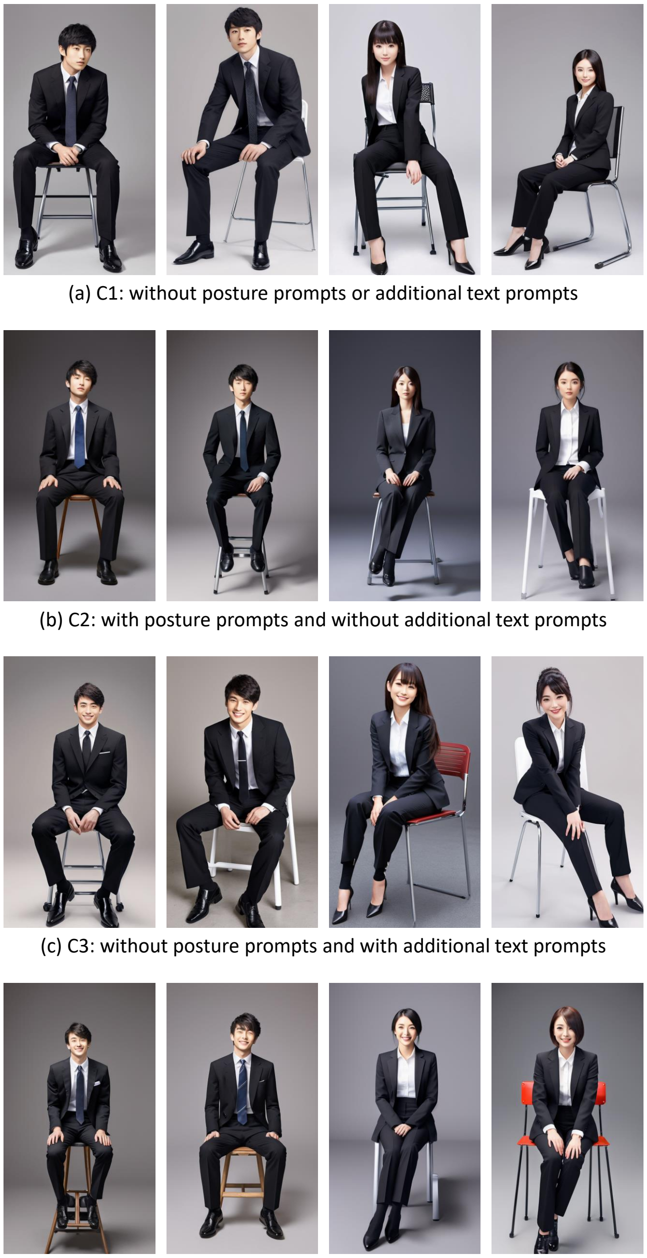
applicant's body was out of the frame or where elements specified in the text prompts were not reflected were excluded from the experiment.

2.4 Participants

Fifteen Japanese students (14 male and one female, with an average age of 23.9 ± 2.8 years) participated in the subjective assessment. On observing the digital human applicant in the stimulus image, participants were asked the following questions:

Q1: Do you feel a good first impression of the applicant?
Q2: Do you feel a bad first impression of the applicant?

Participants responded on a 5-point scale (5: yes, 4: likely yes, 3: neutral, 2: likely no, 1: no). Before conducting the subjective assessment, the participants were instructed to imagine specific interview scenarios, such as job interviews or entrance examinations, and to respond to the impressions of the digital human



(a) C1: without posture prompts or additional text prompts

(b) C2: with posture prompts and without additional text prompts

(c) C3: without posture prompts and with additional text prompts

(d) C4: with both posture prompts and additional text prompts

Fig. 3. Examples of stimulus images of digital human applicants created under conditions C1 to C4.

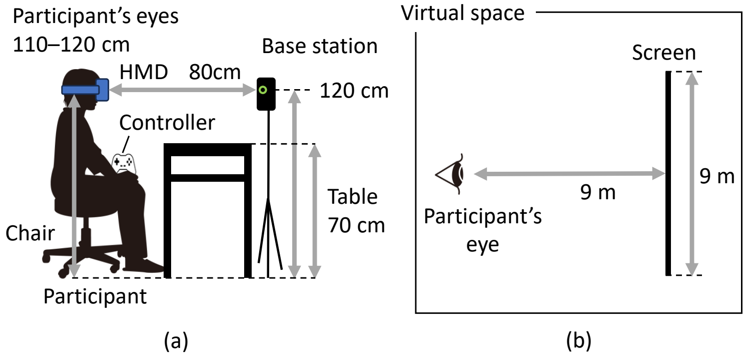


Fig. 4. Experimental setting.

applicants as if they were the interviewers themselves. We thoroughly explained the disadvantages of the subjective assessment to the participants and obtained their consent on a form before performing this assessment.

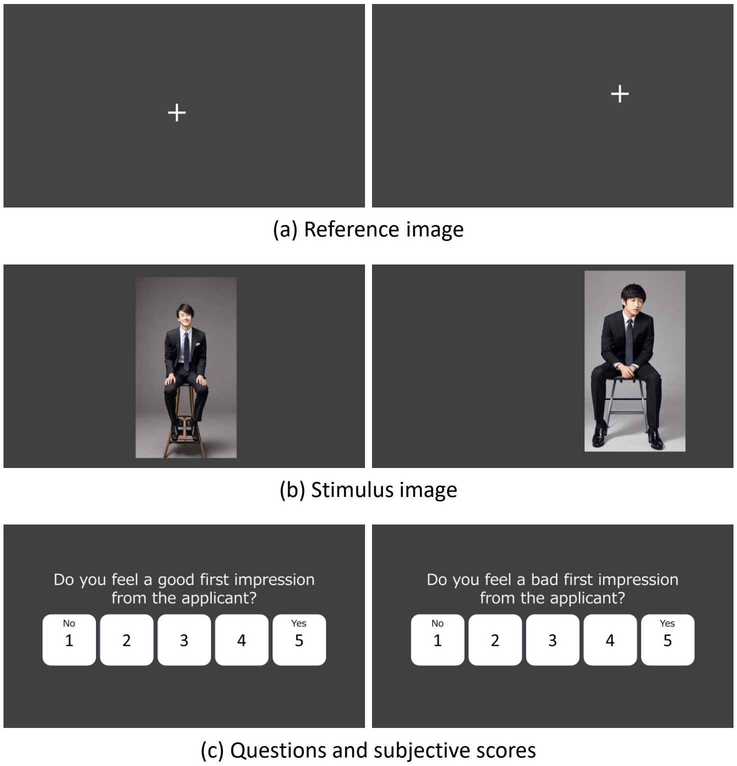
2.5 Setting

Figure 4 demonstrates the setup for displaying the stimulus images to the participants. We used an HMD (VIVE Pro Eye) to construct a virtual space simulating an interview. The participants were instructed to wear the HMD and sit 80 cm away from the base station. The stimulus images were displayed on a screen in the virtual space. To reduce subject fatigue due to visual strain, the viewing distance was secured by setting a large virtual screen. Specifically, the height of the screen was set to 9 meters and the width was set to 16 meters. The distance from the participant's viewing position to the screen was also set to 9 meters. A controller, used for answering questions, was placed in the hands of the participants.

2.6 Experimental procedures

The procedure for providing stimuli to the participants was as follows:

P1 : The questions and the method of answering were explained to the participants.
P2 : The reference image was displayed for 2 seconds.
P3 : The stimulus image was displayed for 9 seconds.
P4 : The participants were asked to answer questions Q1 and Q2.
P5 : The procedures P2 to P4 were repeated until the participants completed their responses for all 48 stimulus images.



(a) Reference image

(b) Stimulus image

Do you feel a good first impression from the applicant?

Do you feel a bad first impression from the applicant?

(c) Questions and subjective scores

Fig. 5. Examples of screens seen by participants when wearing the HMD.

Figure 5 shows an example of the screen displayed on the virtual screen. In P2, to help participants understand the timing of the stimulus image display, a reference image, as shown in Figure 5(a), was displayed beforehand, and they were instructed to focus on the center of a randomly displayed white cross on the screen. In P3, one stimulus image, as shown in Figure 5(b), was displayed for 9 seconds. A random selection was made from the 48 stimulus images without duplication. The position of the stimulus image on the display was randomized to avoid center bias [1]. In P4, questions and subjective scores, as shown in Figure 5(c), were displayed. Questions Q1 and Q2 were presented one at a time in a random order. The participants selected their responses to each question using the controller in their hands.

3 Experiments

3.1 Two-way ANOVA

We evaluated hypothesis H1, which predicted that the virtual interviewers would feel a good first impression from the images of digital human applicants generated

Table 3. The results of the analysis of variance (ANOVA) conducted on the subjective scores obtained from questions Q1 and Q2.

Question	Factor	DoF	F value	p value	Main effect	Interaction
Q1	Posture prompts	1	255.69	$< .001$	Present	-
	Additional text prompts	1	48.38	$< .001$	Present	-
	Posture prompts × Additional text prompts	1	35.48	$< .001$	-	Present
Q2	Posture prompts	1	247.25	$< .001$	Present	-
	Additional text prompts	1	15.41	$< .001$	Present	-
	Posture prompts × Additional text prompts	1	13.75	$< .001$	-	Present

by the combination of posture prompts and additional text prompts. We used the subjective scores obtained from 15 participants in the procedure P4 of Section 2.6 for hypothesis testing. We applied the aligned rank transform [12] to the subjective scores from questions Q1 and Q2, and conducted a two-way ANOVA with repeated measures. Table 3 shows the results of the ANOVA. For Q1, significant main effects were observed for the posture prompt ($F = 255.69, p < .001$) and the additional text prompt ($F = 48.38, p < .001$). Additionally, an interaction effect was found ($F = 35.48, p < 0.001$). For Q2, significant main effects were also observed for the posture prompt ($F = 247.25, p < .001$) and the additional text prompt ($F = 15.41, p < .001$), along with an interaction effect ($F = 13.75, p < .001$). Interaction effects were observed for both Q1 and Q2, indicating that the combination of posture prompts and additional text prompts significantly changed the interviewers' first impression of the digital human applicants.

We conducted the Wilcoxon signed-rank test as a test for simple main effects. Table 4 shows the results of the simple main effect tests for the posture prompt. In the table, DoF indicates the degrees of freedom. For Q1, the simple main effect of the posture prompt was observed in both the presence and absence of the additional text prompt. Similarly, for Q2, the simple main effect was also observed in both cases. Table 5 shows the results of the simple main effect tests for the additional text prompt. For Q1, a simple main effect was observed when the posture prompt was present; however, no simple main effect was found when the posture prompt was absent. In Q2, the same pattern was observed, with a simple main effect when the posture prompt was included, but not when the posture prompt was not included.

3.2 Average subjective scores

To examine how the effects of posture prompts and additional text prompts were reflected in the subjective scores obtained from the procedure P4 of Section 2.6, we divided the subjective assessment results on the basis of the presence or absence of posture prompts and additional text prompts, and calculated the average subjective scores for each category. Specifically, we calculated the average

Table 4. The simple main effects of additional text prompts.

Question	Factor	DoF	p value	Simple main effect
Q1	w/o additional text prompts	1	$< .05$	Present
	w/ additional text prompts	1	$< .05$	Present
Q2	w/o additional text prompts	1	$< .05$	Present
	w/ additional text prompts	1	$< .05$	Present

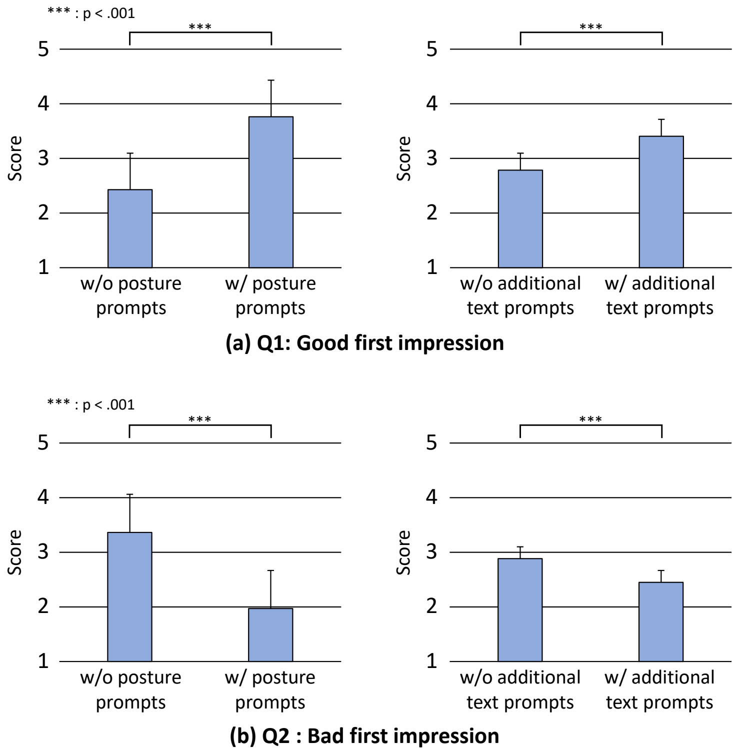
Table 5. The simple main effects of additional text prompts.

Question	Factor	DoF	p value	Simple main effect
Q1	w/o posture prompts	1	$\geq .05$	Absent
	w/ posture prompts	1	$< .05$	Present
Q2	w/o posture prompts	1	$\geq .05$	Absent
	w/ posture prompts	1	$< .05$	Present

subjective scores for the stimulus images generated under conditions C1 and C3 for the absence of posture prompts, as described in Section 2.1. For the presence of posture prompts, we calculated the average subjective scores for the stimulus images generated under conditions C2 and C4. Similarly, we calculated the average subjective scores for the stimulus images generated under conditions C1 and C2 in the absence of additional text prompts. In the presence of additional text prompts, we calculated the average subjective scores for the stimulus images generated under conditions C3 and C4. Additionally, we applied the Wilcoxon signed-rank test to the subjective scores to determine whether there were significant differences in the average scores on the basis of the presence or absence of prompts.

Figure 6(a) shows the average subjective scores obtained from question Q1, in which higher scores indicate that the interviewers felt a better first impression. A significant difference in the average subjective scores was observed on the basis of the presence or absence of posture prompts ($p < .001$). Similarly, a significant difference in the average subjective scores was also found on the basis of the presence or absence of additional text prompts ($p < .001$). In cases where posture prompts were present, the average subjective scores were higher compared with those when they were absent. Likewise, when additional text prompts were present, the average subjective scores were higher compared with those when they were absent.

Figure 6(b) shows the average subjective scores obtained from question Q2, in which higher scores indicate that the interviewers had a worse first impression. As with (a), a significant difference in the average subjective scores was observed on the basis of the presence or absence of posture prompts ($p < .001$). Additionally, a significant difference in the average subjective scores was found on the basis of the presence or absence of additional text prompts ($p < .001$). In this case, the average subjective scores were higher when posture prompts were absent compared with those when they were present. Similarly, the average subjective



(a) Q1: Good first impression

(b) Q2 : Bad first impression

Fig. 6. Average subjective scores for Q1 and Q2.

scores were higher when additional text prompts were absent compared with those when they were present.

The observation that the average subjective score for Q1 increased and the average subjective score for Q2 decreased when prompts were used compared with the scores when they were not suggests that using posture prompts and additional text prompts improved interviewers' first impressions of the digital human applicants.

3.3 Discussion

The results described in Section 3.1 and 3.2 indicated that digital human applicants with combined posture prompts and additional text prompts elicited better first impressions than digital applicants without the prompt combinations. It cannot be concluded that providing an additional text prompt improves the first impression when a posture prompt is absent, which highlights a negative finding. However, the current findings indicated that combining posture prompts and text prompts leads to a better first impression. Thus, when there is no opportunity for adjustment through additional text prompts, prioritizing posture control through posture prompts may be more effective for enhancing the first impression of a digital human applicant.

4 Additional experiment: Evaluation when changing impression words in questions

4.1 Overview

To deepen the exploration of first impressions, we conducted an additional experiment. The experiment described in Section 3 indicated that combining posture prompts with additional text prompts improved the first impression of the digital human applicants. In this additional experiment, we investigated what other impressions the interviewer might feel regarding the digital human applicants, in addition to their overall impression. To this end, we evaluated the following hypothesis (H2) regarding the digital human generated using a combination of posture prompts and additional text prompts:

H2 : The subjective scores given by participants simulating interviewers differ on the basis of the impression words included in the questions.

4.2 Experimental conditions

We used images of the digital human applicant generated under condition C4 (with posture prompts and additional text prompts) as stimulus images. To inquire about impressions other than the overall impression, we prepared the following five additional questions:

Q3: Do you feel that the applicant is curious?
Q4: Do you feel that the applicant is sincere?
Q5: Do you feel that the applicant is sociable?
Q6: Do you feel that the applicant is cooperative?
Q7: Do you feel that the applicant is nervous?

On the basis of the Big Five personality traits [6], we included impression words in the questions that could serve as decision-making criteria for the interviewer's evaluation, considering aspects such as personality, attitude, and communication skills. Ten participants (nine males and one female, with an average age of 23.7±3.1 years, all Japanese) participated in this additional experiment. All other experimental conditions were kept the same as in Section 2.

4.3 Results

We calculated the average subjective scores for the stimulus images for each question. Additionally, we performed the Steel-Dwass test as a multiple comparison method. The results are illustrated in Figure 7. Comparing the average

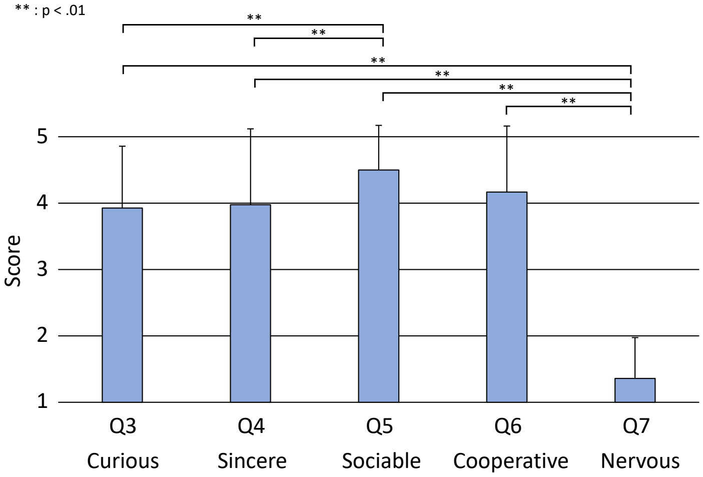


Fig. 7. Average subjective scores for each question from Q3 to Q7.

subjective scores among Q3 to Q7, Q5 (sociable) had the highest score, while Q7 (nervous) had the lowest. A significant difference was observed between the average subjective scores of Q5 and Q3 ($p < .01$). Similarly, a significant difference was found between Q5 and Q4 ($p < .01$). Significant differences were also found between Q3 and Q6, as well as Q7 ($p < .01$). However, no significant differences in average subjective scores were found between Q3 and Q4, or between Q5 and Q6. These results suggest that a digital human generated using a combination of posture prompts and additional text prompts is more likely to convey an impression of sociability rather than curiosity, sincerity, or cooperativeness. Additionally, it may be possible to avoid giving an impression of nervousness.

5 Conclusions

In the current study, we performed a subjective assessment of first impressions of a digital human applicant generated by a combination of posture prompts and text prompts, assuming a virtual interview setting. The results indicated that first impressions of digital human applicants with combined posture prompts and additional text prompts were better than first impressions of digital applicants without such combinations.

In future work, we will investigate which posture and word prompts influence first impressions most strongly, and will consider methods for selecting effective posture and text prompts. Additionally, assuming a virtual interview, we will expand subjective assessments to include both interviewers' first impressions and their overall impressions throughout the interview.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant No. JP23K11145. We appreciate Dr. Michiko Inoue, and Dr. Shunsuke Yoneda for his valuable suggestions during this research.

References

1. Bindemann, M.: Scene and screen center bias early eye movements in scene viewing. Vision research **50**(23), 2577–2587 (2010)
2. Burke, S.L., Bresnahan, T., Li, T., Epnere, K., Rizzo, A., Partin, M., Ahlness, R.M., Trimmer, M.: Using virtual interactive training agents (vita) with adults with autism and other developmental disabilities. Journal of autism and developmental disorders **48**, 905–912 (2018)
3. Burke, S.L., Li, T., Grudzien, A., Garcia, S.: Brief report: Improving employment interview self-efficacy among adults with autism and other developmental disabilities using virtual interactive training agents (vita). Journal of Autism and Developmental Disorders **51**, 741–748 (2021)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 172–186 (2021)
5. Dougherty, T.W., Turban, D.B., Callender, J.C.: Confirming first impressions in the employment interview: A field study of interviewer behavior. Journal of applied psychology **79**(5), 659–665 (1994)
6. Goldberg, L.R.: The development of markers for the big-five factor structure. Psychological assessment **4**(1), 26–42 (1992)
7. Kim, J., Sandhaus, H., Fussell, S.R.: VR job interview using a gender-swapped avatar. In: Companion Publication of the Conference on Computer Supported Cooperative Work and Social Computing. pp. 154–159 (2023)
8. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the International conference on machine learning. pp. 8821–8831 (2021)
9. Robin, R., Andreas, B., Dominik, L., Patrick, E., Bjorn, O.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
10. Smith, M.J., Ginger, E.J., Wright, K., Wright, M.A., Taylor, J.L., Humm, L.B., Olsen, D.E., Bell, M.D., Fleming, M.F.: Virtual reality job interview training in adults with autism spectrum disorder. Journal of autism and developmental disorders **44**, 2450–2463 (2014)
11. Smith, M.J., Pinto, R.M., Dawalt, L., Smith, J., Sherwood, K., Miles, R., Taylor, J., Hume, K., Dawkins, T., Baker-Ericzén, M., et al.: Using community-engaged methods to adapt virtual reality job-interview training for transition-age youth on the autism spectrum. Research in Autism Spectrum Disorders **71**, 101498 (2020)
12. Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J.: The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 143–146 (2011)

13. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)