IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 June 2024, accepted 5 August 2024, date of publication 9 August 2024, date of current version 26 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441233

RESEARCH ARTICLE

Toward AI-Mediated Avatar-Based Telecommunication: Investigating Visual Impression of Switching Between User- and AI-Controlled Avatars in Video Chat

SHIGEO YOSHIDA[1], YUKI KOYAMA[2], AND YOSHITAKA USHIKU[1], (Member, IEEE)
[1]OMRON SINIC X Corporation, Tokyo 113-0033, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan

Corresponding author: Yuki Koyama (koyama.y@aist.go.jp)

This work was supported by JST Moonshot R&D Program Grant Number JPMJMS2236.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the OMRON Corporation Technology and Intellectual Property H.Q. Research Ethics Review Committee under Application No. RD-ECE200024.

ABSTRACT Telecommunications technology has evolved rapidly, creating opportunities to diversify the communication culture through AI mediation. We anticipate that people will interchangeably use both user-controlled (controlled by transferring user movements) and AI-controlled avatars (controlled autonomously) in everyday communication. For example, users may temporarily use the AI-controlled mode during distractions. This paper argues the importance of investigating the auto-switching between user- and AI-controlled modes to improve the user experience in upcoming AI-mediated telecommunications. As a first step, we conducted a crowdsourced user experiment in a video chat context, focusing on the visual impressions of the displayed avatars. The result shows that impression improved when an appropriate switch setting is used, underscoring the value of this research direction. To identify the appropriate switch setting for our experiment, we developed a general-purpose *adaptive experimental design* tool based on Bayesian optimization, which we plan to release publicly.

INDEX TERMS Avatar, adaptive experimental design, Bayesian optimization, human–computer interaction, telecommunication.

I. INTRODUCTION

A. MOTIVATION

The use of avatars in communication has become increasingly prevalent. Video chat platforms such as Zoom[1] and Teams[2] began offering a feature that displays avatars instead of actual user appearances. In social virtual reality (VR) environments, avatars are a common communication medium. Avatars provide a means of privacy protection and reduce the psychological burden of communication [1] while offering an opportunity for self-expression beyond the limitations of one's own physical body [2].

The use of avatars is not limited to user control. They also serve as embodied representations of artificial intelligence (AI) and can be autonomously controlled [3], such as virtual assistants and non-player characters in video games and VR social networking services. Recent advances in AI learning data and expanded input/output modalities have enabled more natural communication with AI-controlled autonomous avatars, if not in the same manner as with humans.

In today's world, wherein diverse work styles and lifestyles are prerequisites, we anticipate that people will utilize both *user-* and *AI-controlled avatars* interchangeably in everyday

The associate editor coordinating the review of this manuscript and approving it for publication was Jose Saldana.

[1] https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0059415

[2] https://insider.microsoft365.com/en-us/blog/avatars-for-microsoft-teams

communication. For example, a remote worker may use a user-controlled avatar during a conversation with their colleague but automatically switch to the AI-controlled avatar when suddenly spoken to by their child. They may also use an AI-controlled avatar while traveling and auto-switch to the user-controlled avatar upon their availability. We foresee scenarios wherein the user- and AI-controlled avatars switch automatically. For example, this could occur in instances when the user's attention is diverted elsewhere or when the user cannot respond using a user-controlled avatar. However, several challenges are associated with auto-switching between avatars, including potential disruptions in communication if the auto-switching method is not properly designed.

B. RESEARCH QUESTIONS AND OUR CONTRIBUTIONS

In this work, we pose the following research question (RQ):

RQ *Does optimizing the auto-switching method between user- and AI-controlled avatars result in a superior user experience and enhanced communication?*

Further, if this is the case, it subsequently prompts two additional research questions:

RQ1 *What aspects of user experience can be enhanced through modifications in the avatar auto-switching behavior?*

RQ2 *What strategies can be implemented to optimize these auto-switching behaviors and thereby enhance user experience?*

These two topics would deserve in-depth explorations in the field of human-computer interaction (HCI), extending beyond a single paper.

As a first step towards this direction, we deliberately keep our focus to be as simple as possible. We focus on a two-person video chat, where one person uses an auto-switch mechanism and the other observes the switch without knowing that the AI-controlled mode is being used interchangeably during the chat (Figure 1). This represents a foundational scenario in avatar-based communication, which is in contrast to more complex scenarios such as multi-person chats, one-to-many videoconferences, and VR social networking with full-body avatars. We believe that this scenario is well-suited as an initial step. Our evaluation primarily examines the visual impression of the auto-switch felt by the observer. We reserve the investigation of other aspects of user experience, such as the sense of agency and performance, for future work. In summary, the hypothesis of our experiment is that well-optimized auto-switching behavior can create a better visual impression on the talking partner than those adjusted non-systematically.

To validate this hypothesis, we conducted a crowdsourced user experiment. For this experiment, we implemented a system that converted real-person videos into avatar videos and a simple mechanism to detect auto-switching timings based on the speaker's head pose. We generated various avatar videos with different auto-switch parameters using the pre-recorded chat videos. As making a stable *absolute* assessment of a given video is difficult for participants for several reasons [4], [5], [6], [7] (as detailed in subsection IV-A), the task was designed as a set of pairwise comparisons to facilitate *relative* assessment. The result shows that impression improves when an appropriate switch setting is used, thus underscoring the value of this research direction.

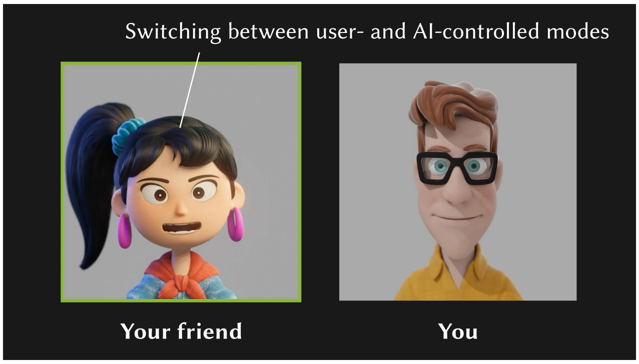

FIGURE 1. Our target scenario: video chat between two individuals using avatars. One person uses the auto-switch mechanism to switch the user- and AI-controlled modes. As the first step, we evaluate this scenario in terms of the visual impression. The character models are provided by Blender Studio under the CC BY license.

A parameter set that ensures optimal switching behaviors is required to conduct the experiment. However, identifying the correct switch setting is challenging because it is related to user perception and necessitates user experiments. The use of traditional experimental design methods (*e.g.*, Latin hypercube sampling) requires the evaluation of densely pre-sampled parameter sets, which is impractical owing to immense human effort. To address this challenge, we take an approach called *adaptive experimental design* [8]. This computational framework allows the experimental conditions (or the parameter sets to be examined) to be adaptively determined *during* the experiment based on the accumulated data. Consequently, it efficiently identifies the most suitable parameter set within a small number of tests. To implement this strategy, we developed a computational tool underpinned by *Bayesian optimization* (BO) [9] techniques tailored to our requirements. While BO drives the tool's performance, its users need not understand its intricacies. This tool was used in a preliminary experiment for parameter identification. Although we initially developed the tool for our study, its potential extends beyond our use case. This can address parameter identification problems in other avatar environments, including those with different communication styles (*e.g.*, VR social networking) and those using more sophisticated AI technologies. Furthermore, the tool can be used for broader interaction research, helping to create and fine-tune interaction techniques based on user experiment data. Given its versatility, we plan to make this tool accessible to researchers working specifically with human subjects.

Our primary contributions are summarized below.

- We highlight the significance of investigating the interaction design challenges associated with auto-switching between user- and AI-controlled avatars to enhance user experience.

- A first-step experiment centered on video chat scenarios, emphasizing visual impressions, was conducted to reinforce our assertion of the importance of this topic.
- We introduce the concept of adaptive experimental design in the context of HCI research. We also offer an adaptive experimental design tool to streamline the development of future avatar-switching systems and broader interactions.

II. RELATED WORK

A. AVATAR-BASED COMMUNICATION

Avatar serves as a visual representation of an individual in a digital environment that mediates interactions both with the digital space and among people within it [3]. Avatars are of various forms, from realistic human figures to animated characters, animals, and other imaginative entities. The methods for controlling these avatars range from full body tracking to basic movements limited to the mouth and eyes.

Avatars are primarily used for remote interactions, particularly in video chats and VR environments. As mentioned in section I, certain major video chat platforms offer avatar communication features, and this avatar-based communication is common on social VR platforms such as VRChat[3] and Horizon Worlds.[4] The use of avatars facilitates the conveyance of verbal and nonverbal information while ensuring a level of anonymity and privacy [10], [11]. Extensive research has been conducted on the use of avatars in communication, such as the psychological impact of avatar use [12], human augmentation [13], long-term usage effects [14] (with older adults [15]), and the ethical considerations regarding avatar use [16].

This work explored avatar use in communication, with a focus on techniques for switching avatars between user- and AI-controlled modes. Previous research has examined technologies and platforms [17], [18] that facilitate avatar switching in contexts such as customer support or job interview, wherein the control can switch from an operated avatar to a human operator when required. In addition, researchers have investigated several approaches for the time-division control of multiple avatars by a single person, mainly considering embodiment factors such as body ownership and sense of agency [19], [20], [21]. In contrast, our work specifically addressed the user experience of switching avatars between user- and AI-controlled modes during communication. As detailed in section III, we envisioned a situation wherein users talked via video chat through their avatars; however, if one could not respond, the system automatically switched the user-controlled avatar to an AI-controlled avatar to maintain communication.

B. ADAPTIVE EXPERIMENTAL DESIGN

Researchers often conduct experiments to collect data to understand their parameter space and search for the best parameter configuration. Designing the experimental conditions requires care due to the associated costs. Ideally, fewer conditions should be tested.

Adaptive experimental design [8] is a variant of the *Design of Experiments* (DOE) framework. Here, the values of the independent variables (*i.e.*, parameters) are adaptively sampled *while the experiment is ongoing*. This adaptivity enhances sample efficiency, allowing researchers to satisfy their goals with fewer tests. This efficiency is achieved using computational techniques. A notable technique used here is BO [9], a black-box optimization method. During the experiment, BO progressively builds a model (referred to as a *surrogate* model) from the data collected thus far and then selects the next samples based on this model. In contrast, traditional experimental design methods (using, *e.g.*, Latin hypercube sampling) mandate that all test values be predetermined *before the experiment begins*, and no analysis occurs until the experiment concludes. Figure 2 illustrates the difference between traditional and adaptive experimental designs.

The concept of adaptive experimental design has been used in various fields where experiments are expensive, such as drug and materials discovery [22], [23]. We argue that this approach also suits HCI research given the significant costs associated with human participant involvement. Notably, existing HCI literature rarely mentions the concept of adaptive experimental design. There is one exception where the concept was mentioned [24]; however, the context was not to minimize participant effort. Our work is the first to explicitly introduce this concept in the context of HCI research with a case study on interaction development based on user experimentation.

Nonetheless, several HCI studies have employed adaptive experimental design, although they do not explicitly mention the concept. Khajah et al. [25] and Dudley et al. [26] conducted iterative user experiments, asking crowd workers to test their systems to run BO and identify optimal system parameters. Chan et al. [27] investigated the experiences of interaction designers while conducting parameter search tasks using BO. Koyama et al. [28] proposed a crowd-in-the-loop BO technique for visual design that is partially aligned with the adaptive experimental design principles. BO has also been used interactive frameworks such as user-in-the-loop design optimization [29], [30], [31] and suggestive interaction [32].

III. TARGET SCENARIO

This section describes our target scenario, which we intentionally make as simple as possible.

A. TARGET SCENARIO

We focus on two-person video-chat scenarios wherein both individuals use 3D avatars. One person uses the user- and AI-controlled modes interchangeably while talking. These two modes are auto-switched by the system. The other person uses the user-controlled avatar mode and is unaware whether the talking partner uses the auto-switching functionality. In the

[3] https://hello.vrchat.com

[4] https://www.meta.com/experiences/2532035600194083/

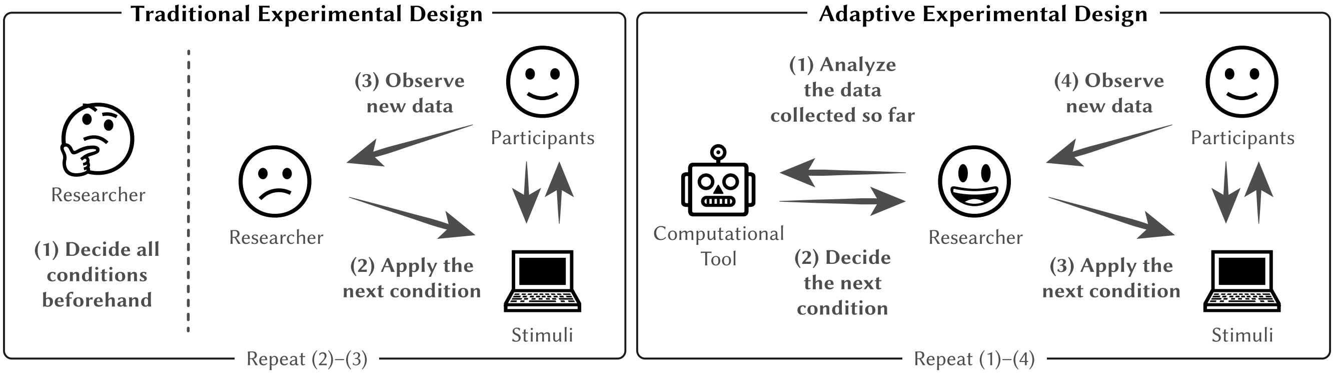
Traditional Experimental Design
Researcher
(1) Decide all conditions beforehand
Researcher
(3) Observe new data
Participants
(2) Apply the next condition
Stimuli
Repeat (2)–(3)
Adaptive Experimental Design
(1) Analyze the data collected so far
Computational Tool
(2) Decide the next condition
Researcher
(4) Observe new data
Participants
(3) Apply the next condition
Stimuli
Repeat (1)–(4)

FIGURE 2. **Difference between traditional and adaptive experimental design. (Left) The researcher determines all conditions beforehand (*e.g.*, using Latin hypercube sampling). (Right) The researcher adaptively determines the next condition at each step, utilizing computational tools and data accumulated thus far.**

user-controlled mode, the system tracks the speaker's facial expressions and head poses and transfers them to the 3D avatar. In AI-controlled mode, the system automatically operates the same 3D avatar based on predefined control programs. We do not consider manipulating the audio channels of the video chats.

B. TARGET USER EXPERIENCE ASPECTS

Although many aspects of user experience may be affected by the auto-switch of avatars, including conversation smoothness and the user's sense of agency, we focus on the visual impression that the user of the auto-switch provides to another person observing the switching. This choice simplifies our experimental setting and is suitable as the first step in investigating RQ1 (subsection I-B). In particular, we selected the following two visual impression criteria considering their orthogonality and also informed by a previous study's questionnaire [33].

- Naturalness: This indicates how human-like an avatar's motion appears. This perception can be influenced by the quality of the auto-switching and that of AI-controlled avatars. If these factors are inadequate, the avatar may appear unnatural. This may cause the viewer to notice that it is not operated by a real person, and thus feel awkward.
- Attention: This indicates to what extent the talking partner perceives the attention from the user of the auto-switching during talking. This perception is shaped by the genuine attention the talker provides and the perceived attentiveness of the AI-controlled avatar. Given that attention is vital to establishing and maintaining communication [34], we consider it important for user experience in avatar-based communication.

C. IMPLEMENTATION

We implemented our prototype avatar environment for our experiment. Our system uses male[5] and female[6] 3D avatars, as shown in Figure 1. Our prototype system is integrated with Blender [35] to utilize its rigging, animating, and rendering capabilities. For user control, the system tracks the user's facial expressions and head pose using MediaPipe [36] and then transfers them to the 3D avatar through scripting. The system simulates the AI control using a pre-recorded video capturing an "idling" (non-talking) state where the face is directed forward. Because idling videos include blinking and natural body movements, we consider it reasonable to prototype a simple AI-controlled mode for the first investigation. When a speaker talks, the system highlights them with green borders. We manually annotated the start and end times of these speaking periods rather than using automated algorithms in our prototype.

D. AUTO-SWITCH

Our system employs a simple strategy to determine the switch timing based on head pose. This approach involves the use of thresholds for the head direction. When a face is oriented directly forward, the system employs the user-controlled mode. However, if the face turns beyond a specified threshold, the AI-controlled avatar mode is activated to maintain the impression of attentiveness. We base these thresholds on the yaw and pitch angles of the head because, within our scenario (*e.g.*, looking away from the talking partner), the head pose appears to be the dominant parameter. Note that the roll angle is not considered here because human faces rarely rotate significantly in this direction. In addition, the proposed system incorporates a *tolerance duration* parameter. Instead of instantly switching the avatar mode upon crossing the threshold, the switch occurs only after the head exceeds the threshold for this specified duration. This prevents frequent switching owing to false detections. In essence, our system depends on three parameters: the yaw threshold, pitch threshold, and tolerance duration. We represent these parameters as $\mathbf{x} = [x_1, x_2, x_3] \in \mathbb{R}^3$, where we set the ranges as $x_1, x_2 \in [0, 1]$ (in radians) and $x_3 \in [0, 3]$ (in seconds). Figure 3 shows the entire pipeline of producing avatar videos.

[5] https://studio.blender.org/characters/5718a967c379cf04929a4247/v1/
[6] https://studio.blender.org/characters/ellie/v1/

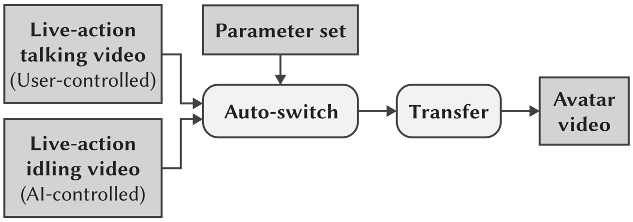
Live-action talking video (User-controlled)
Parameter set
Live-action idling video (AI-controlled)
Auto-switch
Transfer
Avatar video

FIGURE 3. Our implementation of avatar video production with auto-switching between the user- and AI-controlled modes. The "Auto-switch" module takes as input two videos (a talking video used for user control and an idling video used for AI control) and a parameter set to adjust the behavior of switching. The selected video is then converted into a 3D avatar video through the "Transfer" module.

E. CHALLENGES

Determining the right parameter values is challenging because of several factors: (1) these parameters collectively influence the switching behavior; (2) a thorough evaluation of a parameter set requires the reviewing of diverse chat scenarios, particularly those wherein the user demonstrates inattention, to prevent overfitting; and (3) the evaluation of the produced videos must be multifaceted. These complexities underscore the requirement for systematic user experiments and computational tools to identify the best parameter set.

IV. ADAPTIVE EXPERIMENTAL DESIGN TOOL

A. REQUIREMENTS

We employ adaptive experimental design to search for the optimal parameter set that provides the best user experience within our experimental setting (subsection V-B). The found optimal parameter set is then used in our main experiment (subsection V-C). For the target scenario, we identified the following requirements.

- Relative assessment: The *absolute assessment* approach (*i.e.*, direct rating of a single stimulus) has several limitations. These include inconsistent scaling across participants, time-varying scales within individual participants, and a tendency toward inflated or conservative ratings [4], [5], [6], [7]. Overcoming these limitations necessitates a thorough training of the participants to understand the appropriate values for each condition [7]. However, such training is impractical in our crowdsourcing scenario, wherein we aim to gather many participants instantly and repeatedly. Training is particularly necessary in our case because participants cannot be familiar with the design space of avatar behaviors beforehand. In contrast, the *relative assessment* approach, exemplified by pairwise comparison, is easier [37], faster [38], and more accurate [38], [39]; thus, it is often used in psychological studies [37], [40]. Moreover, Wolfert et al. [41] reported that pairwise comparison was faster and more accurate in scenarios similar to ours. Therefore, it is reasonable to use the relative assessment approach rather than the absolute assessment approach.
- Multiple objectives: We aim to optimize several criteria, making it essential to support multiple objectives. In particular, our objectives encompass attention and naturalness, as elaborated in III.
- Batch sampling: Our intention to use crowdsourcing for participant recruitment (as described in V) necessitates batch sampling (*i.e.*, sampling multiple options simultaneously at each step). Batch sampling facilitates the inclusion of more pairwise comparisons in each crowdsourcing task, thereby ensuring that individual tasks are not too short for crowdsourcing.

1) NO EXISTING TOOL SATISFIES ALL SUCH REQUIREMENTS SIMULTANEOUSLY

For example, Ax [42], a notable tool powered by BoTorch [43], does not easily incorporate relative evaluation. Furthermore, regarding algorithms, while BO algorithms that support either multiple objectives [44] or relative assessment [45] exist, none of them supports both features. This necessitated the development of a custom BO algorithm and our own tool.

B. OUR TOOL DEVELOPMENT

We devised a simple (yet non-trivial) BO algorithm that satisfied all the above requirements. First, our BO algorithm employs the recently proposed *qEUBO* technique [46] as its backbone, which supports both relative assessment and batch sampling. Second, our algorithm simplifies the multi-objective optimization task by providing a single *Pareto efficient* solution rather than multiple solutions (*Pareto front* [47]). This is in contrast to most multi-objective optimization algorithms. We enabled this simplified approach by employing the *weighted sum method* [48]. More rigorous computation of the Pareto front is an exciting topic for future research. Appendix A describes the details of the algorithm, including the rationale for the algorithmic choices, the way of mathematically incorporating the weighted sum method into the BO setting, implementation notes necessary for reproduction by others, and performance tests validating its behavior.

We implemented our algorithm and tool on top of BoTorch [43] as described in Appendix A. We designed the tool such that users do not need to be familiar with BO (and do not even need to know the algorithm being used). We plan to release the tool to HCI researchers to develop broader interactions.

C. MATHEMATICAL NOTATION

Consider the aim of maximizing n objectives by searching for the optimal parameter set. We represent the objective functions as $f_1, \ldots, f_n$. Given d parameters to optimize, they are collectively expressed as a d-dimensional vector $\mathbf{x} = [x_1, \ldots, x_d]$. Our goal is to identify an optimal parameter set, $\mathbf{x}^*$, through iterative experiments (refer to Appendix A for a mathematically precise meaning of "optimal" in this multi-objective scenario; our algorithm determines a single optimal solution directly, instead of determining multiple solutions (*i.e.*, Pareto front) and letting the user choose one of them). During iteration, our tool adaptively determines the parameter sets (samples) to be compared in the next iteration step. The number of samples at each step is denoted by $q \geq 2$, which is a user-specified hyperparameter. A lower q value results in a more cost-effective step (demanding fewer comparisons) but may necessitate more iteration steps. We set $q = 4$, as explained later.

V. EXPERIMENTS

We conducted two crowdsourced experiments, preliminary and main, to assess the visual impression when a person switches between user- and AI-controlled avatars. The preliminary experiment aimed to obtain the best parameter set for auto-switching, whereas the main experiment evaluated the visual impression of the talking partner who auto-switched the avatar modes using the obtained parameter set. The experiments were approved by the Ethical Review Board of our institute. We obtained Informed Consent from the participants.

A. DESIGN AND IMPLEMENTATION

This subsection details the common design and implementation used in both the preliminary and main experiments.

1) VISUAL STIMULI

We recorded four different videos, each showing two individuals engaged in video chat. The conversation script in each video was generated via an LLM-based chatbot to imitate a short dialogue between two people (Appendix B). Each video spans approximately 10 seconds and highlights a distinct attention-direction transition during the person corresponding to the participant's speech (approximately 3 seconds after the start of the video). The authors determined these time durations by testing several variations to identify sufficient durations necessary to grasp the dialogue context while lowering the participants' burden. The videos include the following attention transitions.

- From face forward to face sideways
- From face forward to face down
- From face sideways to face forward
- From face down to face forward

These transitions mimic common scenarios wherein a talking partner either diverts attention away from or towards the participant. Based on the approach detailed in sectionIII, we generated visual stimuli of people talking through their avatars with specified parameter sets, as shown in Figure 1.

2) CROWDSOURCING SETTINGS

The participants for the experiments were recruited from a local Japanese recruiting platform.[7] All the instructions and questions were provided in Japanese. Each participant engaged in an online task lasting approximately 30 minutes. No duplicate responses were obtained from participants. The compensation for participation was 540 Japanese Yen (approximately 3.7 USD), reflecting the standard hourly wage in Japan.

3) QUESTION FORMAT

A pairwise comparison approach was adopted for both experiments. This decision was essential, because asking participants to provide concrete rating values is impractical (see subsection IV-A). Participants are more likely to reliably provide relative judgments (*i.e.*, indicate a preference among options). Among the relative judgment methods, pairwise comparison is the simplest and most direct. Considering that subtle nuances are challenging to recall, we believe that simultaneously comparing more than two options would be ineffective.

The options for pairwise comparison were: *non-forced choice* (NFC), where ties are allowed, and *two alternative-forced choice* (2AFC), where ties are not allowed. Through initial testing before the preliminary experiment, forcing participants to select a choice that was almost indistinguishable from the others was found to be burdensome. In addition, information on the tie can be incorporated into the optimization (see Appendix A); therefore, NFC was used in the preliminary experiment. In contrast, the main experiment preferred simplicity of the analysis and its interpretation. Thus, 2AFC was used in the main experiment.

4) TASK

The participants compared two videos with the same video content but generated with different parameter sets. They were allowed to play the videos multiple times for thorough comparison. They were then asked to answer two questions: first, to select the video that they thought appeared more natural, and second, to select the one where they felt more attention was provided.

Four different sets of parameters were compared during the experiments (for each iteration step in the preliminary experiment). Six ($= \binom{4}{2}$) pairs of parameter sets were compared in the four different videos; therefore, the participants were asked to perform the comparison task 24 times.

B. PRELIMINARY EXPERIMENT: PARAMETER IDENTIFICATION

The goal of the preliminary experiment is to identify the parameters that provid the best visual impression, which is necessary to conduct the main experiment. This task would require unreasonably large amounts of data in case a traditional experimental design strategy (*i.e.*, determine all conditions beforehand). To mitigate this problem, we adopt an adaptive experimental design in the preliminary experiment, which enables us to achieve our goals at a reasonable cost.

In the preliminary experiment, 10 crowd workers were recruited per step and asked to complete the NFC questions allowing ties. We ran this experiment for 10 iterations (see Appendix A for the rationale). Therefore, a total of 100 workers were recruited.

After completing 10 iteration steps, we obtained an optimal parameter set: $\mathbf{x}^* = [0.22, 0.33, 0.73]$. Thus, the system

[7] https://www.lancers.jp/

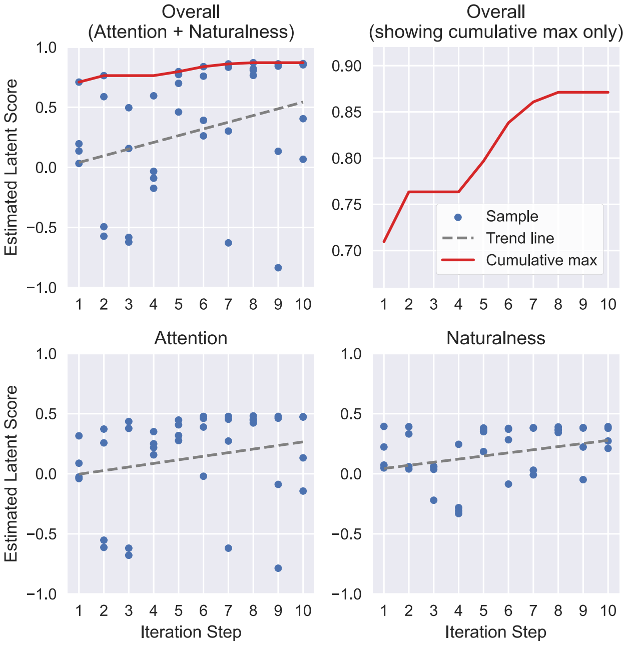

FIGURE 4. Estimated latent scores of the samples (inferred using all data available after 10 steps) over iteration steps.

would trigger a switch from user-controlled to AI-controlled mode if the head pitch deviated by more than 0.22 radians or the head yaw deviated by more than 0.33 radians from the forward direction, and one of these conditions persisted for more than 0.73 seconds. Please refer to the accompanying video figures for videos generated using the identified parameters. Figure 4 shows the scores across the iteration steps. These scores are *latent*, indicating that they were not directly observed but inferred from the pairwise comparison data gathered throughout all 10 steps.

C. MAIN EXPERIMENT: EVALUATION

The main experiment aims to validate our hypothesis that optimal auto-switching behavior creates a better visual impression than those adjusted non-systematically. The obtained parameter set was evaluated by comparing it with randomly generated parameter sets.

In the main experiment, we recruited 24 crowd workers and asked them to complete the 2AFC questions. The parameter sets compared are as follows:

- **Optimal**: The optimal parameter set obtained in the preliminary experiment: $\mathbf{x}^*$.
- **Random-*i*** ($i = 1, 2, 3$): Three randomly sampled parameter sets: $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$.

Human-adjusted parameters can be used for comparison; however, their reproducibility and subjectivity (*e.g.*, adjustment time) pose issues. Therefore, we selected random samples to simulate non-systematic (human) adjustments. Because we focus on whether the auto-switching quality impacted visual impression and not on comparing the optimal design with human-adjusted ones, random samples suffice as baselines.

We performed a statistical analysis of the experimental results using the Bayesian Bradley–Terry model with random-effects [49]. This model was used to obtain the distribution of the preference scores for each parameter set. The multiple sampling from the same participant was treated as a random effect.

Figure 5 shows the posterior distributions of the latent preference scores for the four parameter sets. Table 1 reports the estimated values of the preference score, standard deviation (SD), and 95% highest density interval (HDI) for each parameter set on attention and naturalness. The results indicate a higher base preference for Optimal in terms of both attention and naturalness.

Table 2 and Table 3 present the probabilities of the differences in the preference scores of the two parameter sets. In terms of attention, the overall probability that the preference score of Optimal was greater than that of the other random parameter sets was 99.2% ($0.9920 = 0.9998 \times 1.0000 \times 0.9994$). For naturalness, the overall probability was 86.3% ($0.8632 = 0.8826 \times 0.9986 \times 0.9794$).

VI. DISCUSSION

A. IMPLICATIONS

Overall, the results indicate that (1) the parameter is crucial in terms of visual impression (and thus is expected to affect the user experience as well), and (2) a computationally identified parameter set works better than random sampling, which simulates naïve (non-systematic) adjustment. As hypothesized, we found that the optimal parameter set significantly improved the visual impressions of the talking partner who auto-switched the user- and AI-controlled avatars compared with the non-systematically adjusted parameter sets. Despite the simplicity of this experiment, the findings highlight the necessity for well-designed auto-switching quality. We believe that implementing an auto-switch as an extension of our proposed method has the potential to enhance the user experience in AI-mediated communication.

In the main experiment, the overall comparison of preference scores for naturalness (86.3%) was generally lower than that for attention (99.2%). As depicted in Figure 4 from the preliminary experiment, the narrower variation in naturalness scores may have contributed to this outcome. In addition, this result indicates that changes in the parameter set influence attention more sensitively. However, these observations are specific to the implementation of our avatar environment. The advantage of employing computational tools is evident in facilitating these discussions and enhancing understanding of the target avatar environment. Researchers and practitioners can adopt this analytical process for their respective environments when needed.

B. FUTURE CHALLENGES

1) UNDERSTANDING VARIOUS ASPECTS OF USER EXPERIENCE

We posed a research question regarding the relation between avatar auto-switching behavior and user experience aspects (RQ1). While we leave full answers for this question in future

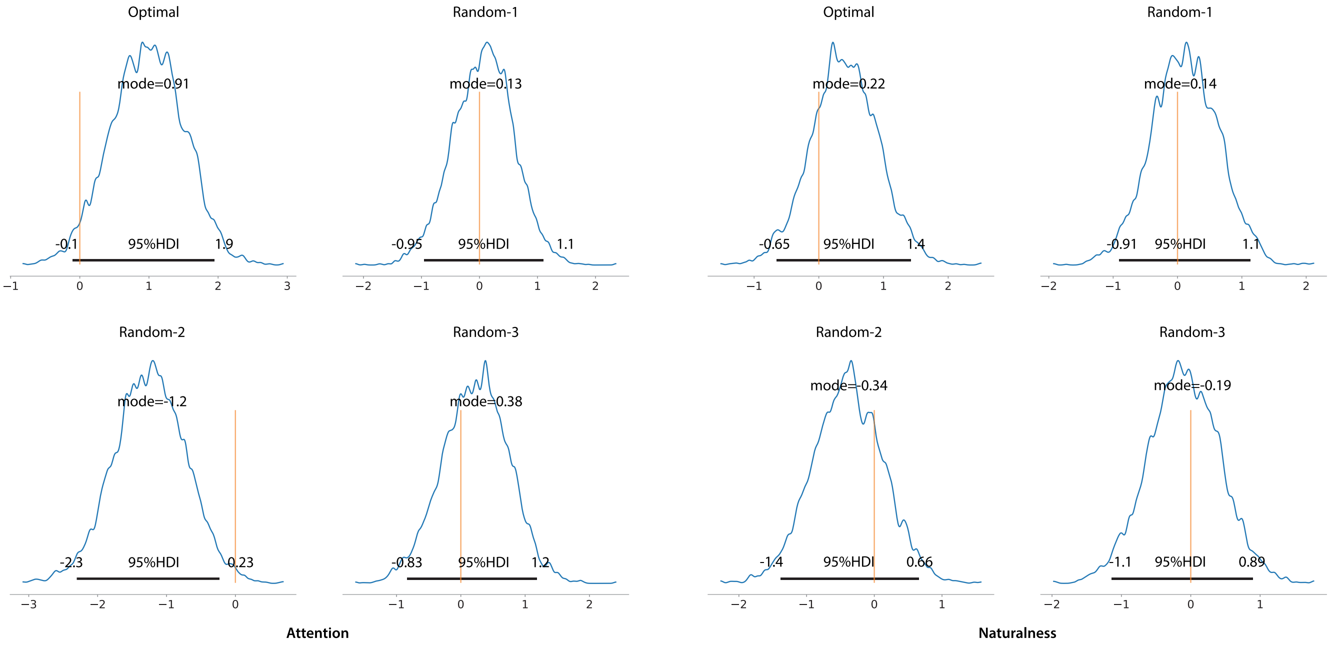

FIGURE 5. Posterior distributions of the latent preference scores and the 95% HDI (highest density interval) in each parameter set regarding attention (left) and naturalness (right). The vertical bar indicates the position of 0, as noted as a reference.

TABLE 1. Summary of point estimates of preference score on attention.

	Attention				Naturalness			
	Mean	SD	Lower 95% HDI	Upper 95% HDI	Mean	SD	Lower 95% HDI	Upper 95% HDI
Optimal	0.98	0.53	−0.10	1.95	0.40	0.52	−0.65	1.42
Random-1	0.06	0.53	−0.95	1.10	0.11	0.52	−0.91	1.13
Random-2	−1.24	0.53	−2.30	−0.23	−0.39	0.52	−1.38	0.66
Random-3	0.23	0.52	−0.83	1.19	−0.10	0.52	−1.14	0.89

TABLE 2. Probability of superiority in attention. Each cell shows the probability that the parameter set in the row is superior to the parameter set in the column.

	Optimal	Random-1	Random-2	Random-3
Optimal	-	0.9998	1.0000	0.9994
Random-1	0.0002	-	1.0000	0.2472
Random-2	0.0000	0.0000	-	0.0000
Random-3	0.0006	0.7528	1.0000	-

TABLE 3. Probability of superiority in naturalness. Each cell shows the probability that the parameter set in the row is superior to the parameter set in the column.

	Optimal	Random-1	Random-2	Random-3
Optimal	-	0.8826	0.9986	0.9794
Random-1	0.1174	-	0.9814	0.8184
Random-2	0.0014	0.0186	-	0.1240
Random-3	0.0206	0.1816	0.8760	-

work, our work focused on the user experience on the observer side, particularly the visual impression (in terms of naturalness and attention) in a video chat as the first step. This choice simplified our experimental setting, where crowdworkers did not need to actually engage in video chat.

Exploring and understanding other aspects of user experience is the primary challenge for our community. For example, further evaluation of impressions other than naturalness and attention is possible. The user experience of those using the auto-switching avatar has not yet been evaluated; we expect that the quality of auto-switching affects the sense of agency and body ownership. Also, evaluating behavioral changes (*e.g.*, how smoothly can the conversation proceed when a good auto-switch mechanism is used? Does switching affect conversation quality?) is important.

In addition, the awareness of AI use may affect experience. While we assumed a scenario wherein the observer of the auto-switching was unaware of their talking partner's use of the auto-switch function and AI-controlled avatar, it may be desirable to disclose the use of AI from the transparency perspective [50]. Future research should analyze the impact of the awareness of AI use.

2) SWITCHING CRITERIA

To simplify the experimental setup, we focused on head poses. However, relying solely on the head pose may be too simplistic for accurately detecting auto-switching timing. For example, situations may arise where the face is turned sideways, yet the individual's attention remains focused (or the opposite scenario). The next challenge involves integrating eye tracking and building a neural network model for more accurate attention detection for avatar control switching. In addition, incorporating speech content analysis and other activity-recognition technologies is a promising direction.

3) AI CONTROL FOR AUTONOMOUS AVATARS

For simplicity, our implementation uses idling videos in the AI-controlled mode. However, this naïve AI only generates idling avatar motions and cannot generate motions adaptively (*e.g.*, based on speech content). Seeking AI techniques to generate more sophisticated avatar motions will be a challenge in the future. For example, motions that attract more attention by performing appropriate reactions based on the conversation content can be generated. Machine learning can be useful in achieving such an AI; one can collect and analyze individual conversation records and train an agent to learn the rules for generating such motions based on the conversation content. Importantly, future experiments can be conducted with new AI technologies using our proposed experimental procedure.

4) SMOOTH MOTION TRANSITION

Our implementation uses a simple approach of instantly transitioning between the user- and AI-controlled modes upon detection of the switch. However, blending the avatar motions from these two modes over a certain duration may provide a smoother and more natural transition. This motion transition problem has been explored in the computer graphics community [51], [52], and incorporating these existing techniques may be beneficial. Concurrently, it is important to develop motion transition techniques specifically for avatar control switching, with an emphasis on enhancing the user experience.

5) GENERALIZABLE INSIGHTS

Our insights, as discussed in subsection VI-A, were derived based on one-on-one chat scenarios. Exploring other scenarios, such as one-to-many or group chats, presents a future challenge owing to the increased complexity of the experimental design. The one-to-many scenario, where the ''many'' are observers, could directly benefit from our findings as their experience is closely aligned with our experimental setup. This suggests the importance of the auto-switching quality in such contexts. Conversely, the impact of auto-switching quality may vary in group chats where attention shifts dynamically. Addressing a range of telecommunication scenarios in future research is vital for acquiring generalizable insights.

C. CONCLUSION

We investigated a novel interaction design problem of auto-switching between user- and AI-controlled avatars to enhance telecommunications. Our initial findings highlight the potential of this research direction. We identified and listed several challenges for future exploration, and we believe that this work will spur further research, collectively advancing the future of AI-mediated communication.

Our work was facilitated by recent advances in computational techniques. In particular, we adopted the concept of adaptive experimental design, which has not been popular in the HCI community. We anticipate that our work will serve as an example of how computational techniques could be used to drive new interaction research. With our released tool, we desire to inspire the broader community.

APPENDIX A
TOOL IMPLEMENTATION

A. ACQUISITION FUNCTION AND SURROGATE MODEL

Our tool implements a BO framework [9] as follows. We use qEUBO [46] (more specifically, the authors' public implementation[8]) for our acquisition function because it is one of the state-of-the-art techniques that can be used with pairwise comparison data and supports batch sampling. According to the experimental results reported by Astudillo et al. [46], qEUBO outperforms other candidate acquisition functions such as qEI, qTS, and MPES. We use the preferential Gaussian process model proposed by Chu et al. [53] (more specifically, the BoTorch implementation[9] with its default settings) as our surrogate model. This model is suited to our needs as it can directly learn from pairwise comparison data. For NFC pairwise comparisons (*i.e.*, when ties are allowed) and if a tie is observed, we generate two opposing responses (indicating a 50% chance of choosing either option), which are then input into the model. We model each objective function separately as a single surrogate model; each model is independently trained using pairwise comparison data on the target objective.

B. HANDLING MULTIPLE OBJECTIVES

Our target problem involves multiple objectives; that is, we want to solve the multi-objective optimization problem:

$$\max_{\mathbf{x}\in\mathcal{X}} \left[f_1(\mathbf{x}), \cdots, f_n(\mathbf{x}) \right], \tag{1}$$

where $f_i : \mathcal{X} \rightarrow \mathbb{R}$ is the i-th objective function. Ideally, we want to find a set of Pareto optimal solutions. However, no BO methods exist that support multi-objective and preferential problem settings. In this study, we take a simplified approach called the *weighted sum method* [48], where we compute the weighted sum of the multiple objectives to formulate a single objective and solve a single-objective

[8] https://github.com/facebookresearch/qEUBO

[9] https://botorch.org/api/_modules/botorch/models/pairwise_gp.html

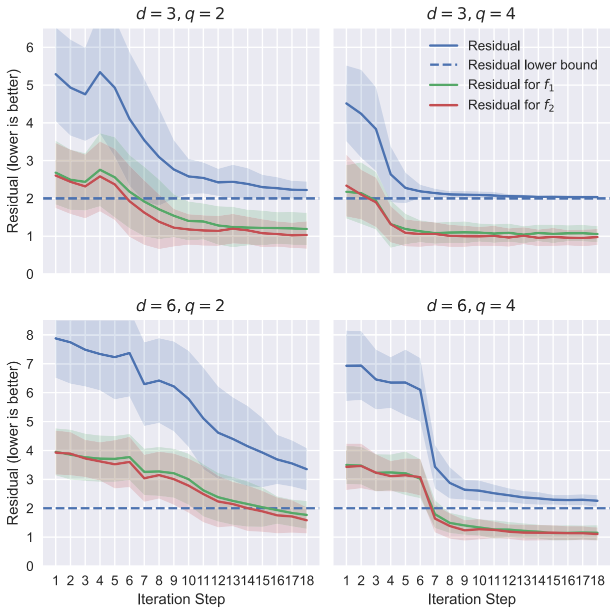

FIGURE 6. Results of performance tests using two synthetic objective functions. Semi-transparent regions indicate the range of one standard deviation from the mean. "Residual" means the sum of the residuals for the maximizers of the two objectives, f_1 and f_2.

optimization problem. In particular, we solve

$$\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \;\; f(\mathbf{x}) := \sum_{i=1}^{n} w_i f_i(\mathbf{x}), \tag{2}$$

where $w_i > 0$ $(i = 1, \ldots, n)$ are the weights. Note that solving the single-objective optimization provides a sufficient condition for Pareto optimality. As each function is represented as a Gaussian process model, each posterior predictive distribution at any data point $\mathbf{x}^*$ is a Gaussian; that is,

$$f_i(\mathbf{x}^*) \sim \mathcal{N}\left(\mu_i(\mathbf{x}^*), \Sigma_i(\mathbf{x}^*)\right). \tag{3}$$

See [53] for the detailed formula for mean and variance. It is known that the sum of Gaussian distributions is also a Gaussian distribution whose mean and variance are equal to the sum of those of the original Gaussian distributions [54]. Using this fact, we obtain

$$f(\mathbf{x}^*) \sim \mathcal{N}\left(\sum_{i=1}^{n} w_i \mu_i(\mathbf{x}^*), \sum_{i=1}^{n} w_i \Sigma_i(\mathbf{x}^*)\right). \tag{4}$$

We set $w_i = 1$ $(i = 1, \ldots, n)$ by default because the output scales of $f_1, \ldots, f_n$ are roughly aligned as they learn from pairwise comparison data using identical model settings. We packaged the above formula as a custom model that can be used with BoTorch. Users of our tool need not be familiar with the above formula.

C. PERFORMANCE TEST

To better understand how the tool works, we ran a simulation experiment where we defined and optimized synthetic test functions. Our test functions were

$$f_1(\mathbf{x}) = -\|\mathbf{x} - \mathbf{x}_1\|, \tag{5}$$

$$f_2(\mathbf{x}) = -\|\mathbf{x} - \mathbf{x}_2\|, \tag{6}$$

where we set $\mathbf{x}_1 = [1.0, 0.0, \cdots, 0.0]$ and $\mathbf{x}_2 = [-1.0, 0.0, \cdots, 0.0]$, and we set the parameter bound as $\mathcal{X} = [-3.0, 3.0]^d$. The maximizers for the objective functions f_1 and f_2 are $\mathbf{x}_1^* = \mathbf{x}_1$ and $\mathbf{x}_2^* = \mathbf{x}_2$, respectively. In this test, we measured the residuals to the maximizers of these objectives. Note that these two functions have a trade-off relationship, so we cannot realize the maximizers of these two functions simultaneously. Therefore, we expect the solutions to converge somewhere between these two maximizers. The lower bound of the total residual is 2.0. We varied the number of dimensions ($d = 3$ and $d = 6$) and the number of samples in each batch ($q = 2$ and $q = 4$). At each step, we generated synthetic data of $\binom{q}{2}$ pairwise comparisons. We ran the optimization sequences 50 times for each condition and recorded the mean and standard deviation at each iteration step. By default, the tool samples data points uniformly at random over the search space $\mathcal{X}$ (instead of using qEUBO) during the first d iteration steps.

TABLE 4. Dialogue scripts used in our experiments.

Dialogue 1
A: How's your day going? B: It's been busy, but good! I've been working on a new project. A: That sounds interesting! I'd love to hear more about it.
Dialogue 2
A: Have you watched any good movies or shows lately? B: Absolutely! I just binge-watched a new series on Netflix. It was amazing! A: Oh, really? I'm looking for some recommendations. Let's chat about it later.
Dialogue 3
A: Have you heard any exciting news recently? B: Actually, I got accepted into that coding bootcamp! A: Congratulations! That's fantastic news. I'm proud of you!
Dialogue 4
A: Do you have any fun plans for the weekend? B: Yes, I'm going hiking with friends. It should be a great adventure! A: Sounds like a lot of fun! Let's catch up later and you can share your stories.

D. RESULT

Figure 6 shows the performance. As expected, the convergence is faster when $q = 4$ than when $q = 2$. It suggests that the residual converges mostly after about 10 steps in the case of $q = 4$ and $d = 3$, based on which we decided the number of steps in our preliminary experiment (subsection V-B).

APPENDIX B
CONVERSATION CONTENT

Table 4 shows the dialogue scripts used in the experiments, which were generated using ChatGPT,[10] an LLM-based chatbot. *Person A* uses auto-switching between the user- and AI-controlled modes. *Person B* represented the participants

[10] https://chatgpt.com/

in this study. In our experiment setting, attention transitions always occurred while Person B was talking.

ACKNOWLEDGMENT

The authors appreciate Sosuke Ichihashi and Manami Nakagawa for their cooperation and participation in the conversation videos used in their experiments. The conversation scripts used in their experiments (Appendix B) were generated using ChatGPT, an LLM-based chatbot.

(Shigeo Yoshida and Yuki Koyama contributed equally to this work.)

REFERENCES

[1] J. Lee, J. Heo, H. Kim, and S. Jeong, "Fostering empathy and privacy: The effect of using expressive avatars for remote communication," in *Proc. 23rd Int. Conf. Hum.-Comput. Interact., Des. User Exper. Case Stud.* Cham, Switzerland: Springer, 2021, pp. 566–583. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-78468-3_39#citeas

[2] A. Cheymol, A. Lécuyer, J.-M. Normand, F. Argelaguet, and F. Argelaguet, "Beyond my real body: Characterization, impacts, applications and perspectives of 'dissimilar' avatars in virtual reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 11, pp. 4426–4437, Nov. 2023.

[3] K. L. Nowak and J. Fox, "Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations on communication," *Rev. Commun. Res.*, vol. 6, pp. 30–53, Jan. 2018. [Online]. Available: https://www.rcommunicationr.org/index.php/rcr/article/view/47/47

[4] E. Brochu, T. Brochu, and N. de Freitas, "A Bayesian interactive optimization approach to procedural animation design," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animation (SCA)*, Goslar, Germany: Eurographics Association, 2010, pp. 103–112.

[5] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," Dept. Elect. Eng., Univ. Washington, Seattle, WA, USA, Tech. Rep. UWEETR-2011-0004, 2011. [Online]. Available: https://vannevar.ece.uw.edu/techsite/papers/refer/UWEETR-2011-0004.html

[6] S. Kiritchenko and S. Mohammad, "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2017, pp. 465–470.

[7] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," 2017, *arXiv:1712.03686*.

[8] S. Greenhill, S. Rana, S. Gupta, P. Vellanki, and S. Venkatesh, "Bayesian optimization for adaptive experimental design: A review," *IEEE Access*, vol. 8, pp. 13937–13948, 2020.

[9] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.

[10] P. Siriaraya, C. S. Ang, and A. Bobrowicz, "Exploring the potential of virtual worlds in engaging older people and supporting healthy aging," *Behav. Inf. Technol.*, vol. 33, no. 3, pp. 283–294, Mar. 2014.

[11] E. A. Konijn, S. Utz, M. Tanis, and S. B. Barnes, *Mediated Interpersonal Communication*. London, U.K.: Routledge, 2008.

[12] N. Yee and J. Bailenson, "The proteus effect: The effect of transformed self-representation on behavior," *Hum. Commun. Res.*, vol. 33, no. 3, pp. 271–290, Jul. 2007.

[13] M. Inami, D. Uriu, Z. Kashino, S. Yoshida, H. Saito, A. Maekawa, and M. Kitazaki, "Cyborgs, human augmentation, cybernetics, and JIZAI body," in *Proc. Augmented Hum. Int. Conf. (AHs)*. New York, NY, USA: Association for Computing Machinery, Mar. 2022, pp. 230–242, doi: 10.1145/3519391.3519401.

[14] A. Oyanagi, T. Narumi, K. Aoyama, K. Ito, T. Amemiya, and M. Hirose, "Impact of long-term use of an avatar to IVBO in the social VR," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2021, pp. 322–336. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-78321-1_25#citeas

[15] S. Baker, J. Waycott, R. Carrasco, R. M. Kelly, A. J. Jones, J. Lilley, B. Dow, F. Batchelor, T. Hoang, and F. Vetere, "Avatar-mediated communication in social VR: An in-depth exploration of older adult interaction in an emerging communication platform," in *Proc. Conf. Human Factors Comput. Syst. (CHI)*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–21, doi: 10.1145/3411764.3445752.

[16] J. Wolfendale, "My avatar, my self: Virtual harm and attachment," *Ethics Inf. Technol.*, vol. 9, no. 2, pp. 111–119, Sep. 2007.

[17] T. Kawahara, N. Muramatsu, K. Yamamoto, D. Lala, and K. Inoue, "Semi-autonomous avatar enabling unconstrained parallel conversations –seamless hybrid of WOZ and autonomous dialogue systems–," *Adv. Robot.*, vol. 35, no. 11, pp. 657–663, Jun. 2021.

[18] H. Kawai, Y. Muraki, K. Yamamoto, D. Lala, K. Inoue, and T. Kawahara, "Simultaneous job interview system using multiple semi-autonomous agents," in *Proc. 23rd Annu. Meeting Special Interest Group Discourse Dialogue*, 2022, pp. 107–110.

[19] K. Takada, M. Kawaguchi, A. Uehara, Y. Nakanishi, M. Armstrong, A. Verhulst, K. Minamizawa, and S. Kasahara, "Parallel ping-pong: Exploring parallel embodiment through multiple bodies by a single user," in *Proc. Augmented Hum. Int. Conf. (AHs)*. New York, NY, USA: Association for Computing Machinery, Mar. 2022, pp. 121–130, doi: 10.1145/3519391.3519408.

[20] Y. Nakanishi, M. Fukuoka, S. Kasahara, and M. Sugimoto, "Synchronous and asynchronous manipulation switching of multiple robotic embodiment using EMG and eye gaze," in *Proc. Augmented Hum. Int. Conf. (AHs)*. New York, NY, USA: Association for Computing Machinery, Mar. 2022, pp. 94–103, doi: 10.1145/3519391.3522753.

[21] S. Kishore, X. N. Muncunill, P. Bourdin, K. Or-Berkers, D. Friedman, and M. Slater, "Multi-destination beaming: Apparently being in three places at once through robotic and virtual embodiment," *Frontiers Robot. AI*, vol. 3, p. 65, Nov. 2016.

[22] P. Pallmann, A. W. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight, L. V. Hampson, J. Holmes, A. P. Mander, L. Odondi, M. R. Sydes, S. S. Villar, J. M. S. Wason, C. J. Weir, G. M. Wheeler, C. Yap, and T. Jaki, "Adaptive designs in clinical trials: Why use them, and how to run and report them," *BMC Med.*, vol. 16, no. 1, p. 29, Dec. 2018.

[23] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, "Accelerated search for materials with targeted properties by adaptive design," *Nature Commun.*, vol. 7, no. 1, Apr. 2016, Art. no. 11241.

[24] A. Keurulainen, I. R. Westerlund, O. Keurulainen, and A. Howes, "Amortised experimental design and parameter estimation for user models of pointing," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, vol. 21. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–17, doi: 10.1145/3544548.3581483.

[25] M. M. Khajah, B. D. Roads, R. V. Lindsey, Y.-E. Liu, and M. C. Mozer, "Designing engaging games using Bayesian optimization," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, May 2016, pp. 5571–5582, doi: 10.1145/2858036.2858253.

[26] J. J. Dudley, J. T. Jacques, and P. O. Kristensson, "Crowdsourcing interface feature design with Bayesian optimization," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12, doi: 10.1145/3290605.3300482.

[27] L. Chan, Y.-C. Liao, G. B. Mo, J. J. Dudley, C.-L. Cheng, P. O. Kristensson, and A. Oulasvirta, "Investigating positive and negative qualities of human-in-the-loop optimization for designing interaction techniques," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–14, doi: 10.1145/3491102.3501850.

[28] Y. Koyama, I. Sato, D. Sakamoto, and T. Igarashi, "Sequential line search for efficient visual design optimization by crowds," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017, doi: 10.1145/3072959.3073598.

[29] Y. Koyama, I. Sato, and M. Goto, "Sequential gallery for interactive visual design optimization," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 88:1–88:2, Aug. 2020, doi: 10.1145/3386569.3392444.

[30] Y. Zhou, Y. Koyama, M. Goto, and T. Igarashi, "Generative melody composition with human-in-the-loop Bayesian optimization," in *Proc. Joint Conf. AI Music Creativity (CSMC-MuMe)*, 2020, pp. 21:1–21:10.

[31] K. Yamamoto, Y. Koyama, and Y. Ochiai, "Photographic lighting design with photographer-in-the-loop Bayesian optimization," in *Proc. 35th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 1–11, doi: 10.1145/3526113.3545690.

[32] Y. Koyama and M. Goto, "BO as assistant: Using Bayesian optimization for asynchronously generating design suggestions," in *Proc. 35th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 1–14, doi: 10.1145/3526113.3545664.

[33] S. Yamashita and R. Higashinaka, ''Data collection for empirically determining the necessary information for smooth handover in dialogue,'' in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 4060–4068. [Online]. Available: https://aclanthology.org/2022.lrec-1.432

[34] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, ''A model of attention and interest using gaze behavior,'' in *Proc. Int. Workshop Intell. Virtual Agents*. Berlin, Germany: Springer, 2005, pp. 229–240. [Online]. Available: https://link.springer.com/chapter/10.1007/11550617_20#citeas

[35] Blender Found. *Blender*. Accessed: Jun. 7, 2024. [Online]. Available: https://www.blender.org/

[36] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, ''MediaPipe: A framework for perceiving and processing reality,'' in *Proc. 3rd Workshop Comput. Vis. AR/VR IEEE Comput. Vis. Pattern Recognit. (CVPR)*, 2019. [Online]. Available: https://xr.cornell.edu/workshop/2019/papers

[37] N. Stewart, G. D. A. Brown, and N. Chater, ''Absolute identification by relative judgment,'' *Psychol. Rev.*, vol. 112, no. 4, pp. 881–911, Oct. 2005.

[38] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, ''Comparison of four subjective methods for image quality assessment,'' *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2012.03188.x

[39] N. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright, ''Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence,'' in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 38. San Diego, CA, USA: PMLR, 2015, pp. 856–865. [Online]. Available: https://proceedings.mlr.press/v38/shah15.html

[40] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, ''Evaluation of tone mapping operators using a high dynamic range display,'' *ACM Trans. Graph.*, vol. 24, no. 3, pp. 640–648, Jul. 2005, doi: 10.1145/1073204.1073242.

[41] P. Wolfert, J. M. Girard, T. Kucherenko, and T. Belpaeme, ''To rate or not to rate: Investigating evaluation methods for generated co-speech gestures,'' in *Proc. Int. Conf. Multimodal Interact. (ICMI)*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 494–502, doi: 10.1145/3462244.3479889.

[42] Meta Platforms, Inc. *Ax*. Accessed: Jun. 7, 2024. [Online]. Available: https://ax.dev/

[43] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, ''BoTorch: A framework for efficient Monte-Carlo Bayesian optimization,'' in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 21524–21538.

[44] Y.-C. Liao, J. J. Dudley, G. B. Mo, C.-L. Cheng, L. Chan, A. Oulasvirta, and P. O. Kristensson, ''Interaction design with multi-objective Bayesian optimization,'' *IEEE Pervasive Comput.*, vol. 22, no. 1, pp. 29–38, Jan. 2023.

[45] E. Brochu, N. D. Freitas, and A. Ghosh, ''Active preference learning with discrete choice data,'' in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2007, pp. 409–416. [Online]. Available: http://papers.nips.cc/paper/3219-active-preferencelearning-with-discrete-choice-data

[46] R. Astudillo, Z. J. Lin, E. Bakshy, and P. I. Frazier, ''qEUBO: A decision-theoretic acquisition function for preferential Bayesian optimization,'' in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 206, 2023, pp. 1093–1114. [Online]. Available: https://proceedings.mlr.press/v206/astudillo23a.html

[47] B. A. Smith, X. Bi, and S. Zhai, ''Optimizing touchscreen keyboards for gesture typing,'' in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. (CHI)*. New York, NY, USA: Association for Computing Machinery, Apr. 2015, pp. 3365–3374, doi: 10.1145/2702123.2702357.

[48] R. T. Marler and J. S. Arora, ''The weighted sum method for multi-objective optimization: New insights,'' *Struct. Multidisciplinary Optim.*, vol. 41, no. 6, pp. 853–862, Jun. 2010.

[49] D. I. Mattos and É. M. S. Ramos, ''Bayesian paired comparison with the bpcs package,'' *Behav. Res. Methods*, vol. 54, no. 4, pp. 2025–2045, Nov. 2021.

[50] European Commission. (2023). *Proposal for a Regulation of the European Parliament and of the Council, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. Accessed: May 12, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex

[51] Y. Koyama and M. Goto, ''Precomputed optimal one-hop motion transition for responsive character animation,'' *Vis. Comput.*, vol. 35, nos. 6–8, pp. 1131–1142, Jun. 2019.

[52] D. Bollo, ''Inertialization: High-performance animation transitions in Gears of War,'' in *Proc. GDC Vault (GDC)*, 2018, pp. 1–65. [Online]. Available: https://www.gdcvault.com/play/1025331/Inertialization-High-Performance-Animation-Transitions

[53] W. Chu and Z. Ghahramani, ''Preference learning with Gaussian processes,'' in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*. New York, NY, USA: Association for Computing Machinery, 2005, pp. 137–144, doi: 10.1145/1102351.1102369.

[54] (2023). *Sum of Normally Distributed Random Variables-Wikipedia*. Accessed: Sep. 9, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

SHIGEO YOSHIDA received the bachelor's degree in engineering, the master's degree in arts and sciences, and the Ph.D. degree in information studies from The University of Tokyo, in 2012, 2014, and 2017, respectively. From 2017 to 2022, he was with The University of Tokyo. He is currently a Principal Investigator of the Integrated Interaction Group, OMRON SINIC X Corporation, Japan. He has been especially focusing on designing interactions based on the mechanisms of perception and cognition of our body. His research interest includes human–computer interaction.

YUKI KOYAMA received the Ph.D. degree from The University of Tokyo, in 2017, advised by Prof. Takeo Igarashi. He is currently a Senior Researcher with the National Institute of Advanced Industrial Science and Technology (AIST). Since 2021, he has been with Graphinica Inc., in which he is aiming to bridge art and technology in animation production. In particular, he is interested in supporting designers by using computational techniques, such as mathematical optimization. He has published papers at ACM SIGGRAPH, CHI, and UIST. His research interests include computer graphics and human–computer interaction. He was awarded the JSPS Ikushi Prize, in 2017; the Asiagraphics Young Researcher Award, in 2021; and the IPSJ/ACM Award for Early Career Contributions to Global Research, in 2024.

YOSHITAKA USHIKU (Member, IEEE) received the B.E., M.A., and Ph.D. degrees from The University of Tokyo, Japan, in 2009, 2011, and 2014, respectively. In 2014, he joined NTT CS Laboratories, Japan, where he was involved in research on image recognition. From 2016 to 2018, he was a Lecturer with The University of Tokyo. Since 2018, he has been with OMRON SINIC X Corporation, Japan, where he is currently the Vice President of Research. His research interests include cross-media understanding through machine learning, mainly for computer vision and natural language processing. He received the ACM Multimedia Grand Challenge Special Prize, in 2011; the ACM Multimedia Open Source Software Competition Honorable Mention, in 2017; and the NVIDIA Pioneering Research Awards, in 2017 and 2018.