2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Where and When Should the Teleoperated Avatar Look: Gaze Instruction Dataset for Enhanced Teleoperated Avatar Communication*

Kenya Hoshimure[12], Jun Baba[12], Junya Nakanishi[1], Yuichiro Yoshikawa[1] and Hiroshi Ishiguro[1]

Abstract— **Effective teleoperated avatar communication requires expressing social behaviors. Gaze behavior is one of the crucial social behaviors and includes reflexive reactions to the avatar's surroundings and intentional responses to the operator's speech and actions. Teleoperated avatars must have their gaze behavior controlled according to situational changes in both the avatar's and operator's contexts. However, it is not clear how to adjust the avatar's gaze in response to changes in both situations. In this paper, we collect a dataset of gazing positions that the avatar is instructed to face, taking into account both avatar and operator situations, and annotation labels that represent both situations in detail. We then exploratorily analyze the ratio of gazing positions per situation through dynamic area-of-interest (AOI) analysis. Our analysis provides insights into determining the gaze behavior of teleoperated avatars.**

I. INTRODUCTION

Teleoperated avatars have attracted significant attention since they enable service provision through remote interactions via humanoid robots or computer-generated (CG) avatars. These systems are particularly anticipated to be applied in fields where interpersonal communication with users is essential, such as customer service [1], sales promotion [2], education [3], and healthcare [4]. Utilizing teleoperated avatars for employment enables work from distant locations and offers opportunities for those limited by physical or personal circumstances to engage in desired work.

Teleoperated avatars must adjust their social behaviors, such as eye gaze, gestures, and facial expressions, to fit the conversational context and situational changes. In particular, eye gaze is a primary non-verbal cue integrated into most social robots and CG avatars [5] and plays a critical role in enhancing dialogue experiences. For example, in service encounters, directing the avatar's gaze toward a specific individual to engage in eye contact, focusing the avatar on recommended products to induce shoppers' attention to the products, and indicating a specific direction to guide people can enhance the experience for interlocutors. These avatar behaviors are important to gain wide social acceptance.

An important aspect of realizing natural avatar gaze behavior is considering both the operator's situation and the avatar's surroundings. Conventional avatar gaze control methods are mainly divided into three forms: 1. avatar-side autonomous, 2. operator-side imitation, and 3. operator-side manual control. In avatar-side autonomous methods, they focus on generating autonomous gaze behavior using only information surrounding the avatar. This includes a gaze model based on the saliency map of the avatar's vision [6], adjusting gaze behavior based on distance from the interaction partner [7], and gazing at conversation partners who speak less [8]. Operator-side imitation methods use a motion capture or eye tracking system to directly reflect the operator's head or eye movements in the avatar in real-time [9], [10], allowing operators to express their gaze intuitively and unconsciously. Operator-side manual control methods allow operators to consciously decide the avatar's gaze position by some interfaces, and they have been adopted in many studies [1], [11] because of their ease of implementation. However, these methods do not hybridize information from both the avatar and operator sides for creating avatar gaze behavior and are not sufficient to control appropriate avatar gaze behaviors. For example, if a customer is in front of the avatar and the operator mentions a recommended product, avatar-side autonomous methods cannot shift the avatar's gaze from the customer to the product according to the operator's utterance. If the operator looks down to read a manual while providing information, the avatar with operator-side imitation methods might also look down. Manipulating the manual interface increases the operator's cognitive load, and the operator may forget to control the avatar gaze during product explanations to customers. These mismatches interfere with speech understanding and diminish the customer's impression. Determining the avatar's gaze positions based on contexts from both the avatar's side and the operator's side is necessary for more natural gaze behavior in teleoperated avatars. However, it is not clear how to adjust the avatar's gaze in response to changes in both contexts.

In this study, we aim to obtain insights for supporting gaze control in teleoperated avatars by collecting a gaze instruction dataset. To obtain appropriate gaze positions matched for both the avatar and operator, we need data on operators' and customers' dialogues and behaviors from both sides, as well as data on suitable gaze positions for these situations. In order to collect such suitable gaze positions, it is necessary for 1) the participants (subjects who agreed to provide data acting as operators) instructing gaze positions to understand the avatar's situation and the operator's intention, 2) the participants to carry out only gaze position instructions, and 3) multiple participants to instruct gaze positions in the same scenes. We first divided the participant's task into two phases, the avatar operation task and the gaze instruction task, to create a situation in which the participant can only direct

*This work was supported by JST Moonshot R & D Grant Number JPMJMS2011.

[1]Graduate School of Engineering Science, Osaka University, 1-3, Machikaneyama, Toyonaka, Osaka 560-0043 Japan

[2]AI Lab, CyberAgent Inc., 19F-23F Shibuya Scramble Square 2-24-12 Shibuya Shibuya-Ku, Tokyo 150-6121, Japan

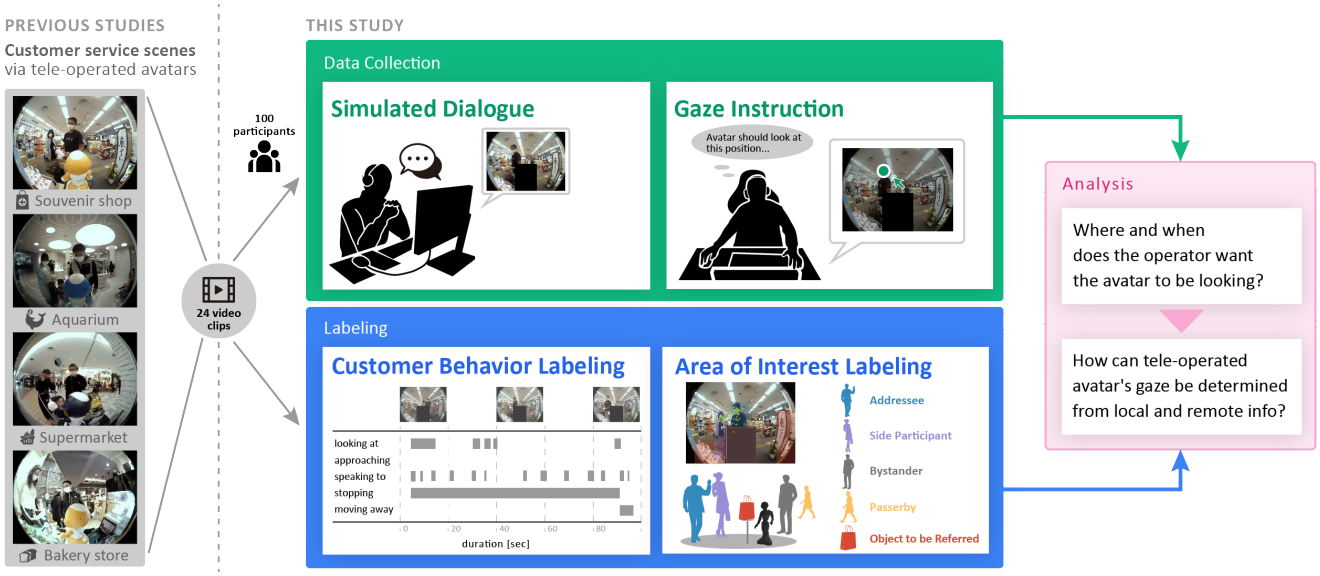

Fig. 1: **Study overview.** We used 24 video clips of interactions between a teleoperated avatar and customers in a commercial facility, which were collected in previous studies. A total of 100 participants, who agreed to provide data, took part in the experiment. They acted as operators and engaged in simulated dialogues with the video clips. Afterward, the participants used a mouse cursor to annotate the areas they wanted the avatar to focus on. This allowed us to collect both the operator's behavioral data and gaze instruction data for the avatar. In parallel, we manually annotated the video clips to create data that included customer behaviors and types of areas of interest (AOIs). We then analyzed the collected data to understand when and where the operator wanted the avatar to focus its gaze.

the gaze position while watching the situation in which they was operating the avatar. Furthermore, by employing a simulated dialogue task toward the video as the avatar operation task, we made it possible for all participants to put gazing instructions into the video of the same scene. Using the data collected in this way, we employed dynamic areas of interest (dynamic AOI) analysis [12], an extension of the AOI analysis for eye movement analysis, applied to dynamic visual scenes. Traditional eye movement analysis techniques include AOI analysis, heatmap analysis to visualize gaze distributions, and scanpath analysis that represents gaze behavior as sequences of fixations and saccades. However, heatmap analysis fails to retain temporal information, and scanpath analysis struggles to identify patterns from multiple scanpaths for dynamic visual stimuli [13]. These limitations make them unsuitable for our research. For dynamic AOI analysis, we label the AOIs, including the changing positions of customers and objects, to capture temporal variations. We also annotate customers' behaviors around the avatar, including their interactions with the avatar and products, along with operators' utterances and gazes toward the screen. From the identified insights, we propose a hypothesis for a gaze control model of teleoperated avatars that considers both the avatar and operator sides. The contributions of this paper are the construction of a dataset of gaze intentions linked to operational data of teleoperated avatars in near-real environments and the provision of insights into the gaze control model for avatars.

II. METHODS

Our goal was to develop hypotheses and insights regarding the gaze behavior control of teleoperated avatars, based on information from both the avatar and operator sides. Fig. 1 illustrates the overview of our study. We have constructed a dataset for gaze instructions, which includes video data (with audio) of participants conducting simulated dialogues while viewing videos from the operator's perspective. The dialogue scenes viewed by participants focus on service encounters, a representative example of service provision via avatars. These scenes utilize a total of 24 videos, captured in various environments such as aquariums, bakery stores, supermarkets, and airport souvenir shops. We record the participants' facial videos and speech audio synchronized to their viewing videos. In the subsequent gaze instruction task, participants keep indicating where they want the avatar to look on the video using a mouse cursor while watching the same video and listening to their speech audio. By having one participant perform these two separate tasks, we enable them to consciously be aware of the internal state of the operators and focus on only gaze instructions. This dataset encompasses gaze instruction data that records the points on the video where the avatar should look, according to the operator's viewpoint, as well as behavioral labels for customers and operators, and labels for AOI.

A. Scene & Clip Selection

To reveal the characteristics of operators in service provision using avatars, we focused on customer service dialogue

scenes. Customer service dialogues, as a representative example of service provision, have been extensively addressed in previous research due to their clear objectives such as product sales, recommendations, and information guidance, while also necessitating the building of good relationships through small talk and everyday conversation, thereby encompassing a wide range of dialogue content and contexts. The dialogue partner may not always be a single customer; situations may include multiple people speaking simultaneously or changes in the dialogue partner, making it a suitable dialogue situation for collecting more complex gaze instruction patterns. A total of 24 video clips of teleoperated avatar scenes were used, employing videos collected in prior research [14]–[17]. These videos were recorded in various settings, such as aquariums, bakery stores, supermarkets, and airport souvenir shops, capturing avatars engaging in customer service with actual clients. Each video clip includes moments from when a customer approaches the avatar to start a conversation to when they leave upon its conclusion. To observe distinctive gaze instruction behaviors, video clips were selected to minimize overlap among the following attributes.

- **Number of Dialogue Partners**: This refers to the number of dialogue partners included in the video, with selections made for both `singular` and `multiple` cases. The selected range of dialogue partner numbers was from a minimum of 1 to a maximum of 5.
- **Main Task**: This represents the primary customer service tasks performed by the teleoperated avatar, including two types: `question-answering tasks` such as exhibit explanations, and `recommendation tasks` such as promoting products or flyers.
- **Number of Objects to be Referred**: This counts the number of objects the operator might refer to in the video, including recommended products or objects near the avatar and the dialogue partner.

B. Participants

This dataset was collected from 100 participants (50 males and 50 females) aged between 20 and 49 years (average age 34.18, with 34 in their 20s, 32 in their 30s, and 34 in their 40s). Participants were recruited through a staffing agency with compensation offered for their participation. Recruitment criteria required participants to be native Japanese speakers, as the customer service scenes were in Japanese, and to be comfortable speaking to people on a video without concern for being watched by others. Before data collection, participants were provided with a consent form detailing the significance and purpose of the research, how the collected data would be handled, and the rights of the participants. Data was collected only from those who consented to participate. It is noted that there were no participants who refused to participate in this experiment. This experiment was approved by the Research Ethics Committee of Osaka University (Reference number: R-1-5-4).

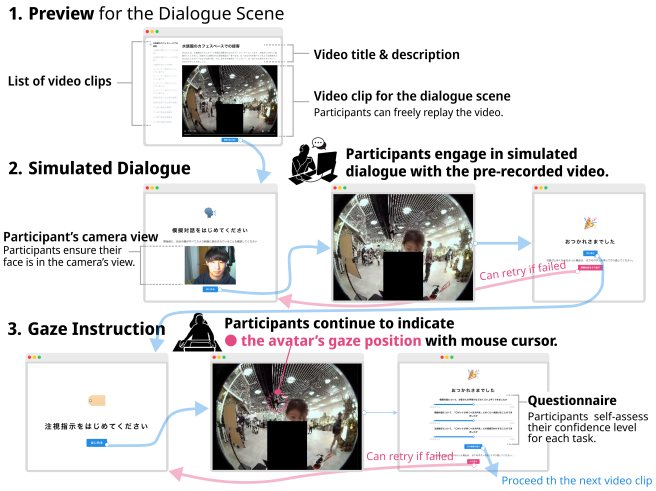

Fig. 2: The custom web browser-based tool we developed to play customer service videos and record gaze instruction.

C. Operator-Side Data collection procedure

In anticipation of actual remote operation scenarios of avatars, participants were seated in soundproof booths facing a desk with a laptop, where they listened to the customer service scene audio and recorded their dialogue responses using a stereo headset (Jabra Engage 50, manufactured by Jabra). A standard wireless mouse was used to measure gaze instruction locations.

For recording gaze instructions, a custom web browser-based tool (Fig. 2) was used. To explain how to use the tool, a data collection administrator first demonstrated the entire data collection flow to the participants and then allowed the participants to experience the data collection flow themselves while being observed by the administrator. If participants made mistakes in using the tool or did not understand the task, the administrator provided immediate clarification.

Initially, participants watched a 1 to 2-minute video from the perspective of a teleoperated avatar operator, engaging in customer service scenes. During this, participants could freely repeat the video to review the conversation content between the customer and the avatar as often as needed.

In the simulated dialogue task, participants engaged in a simulated conversation while watching a full-screen video on the laptop, and their actions were recorded using the camera at the top of the laptop. The videos watched by the participants during the simulated dialogues had all original operator voices removed, and the avatar's position in the video was obscured with a black rectangle to prevent participants from seeing the avatar's orientation or gestures. Participants were generally free to interact with the customer, but for 12 out of the 24 videos, scripts with three responses per video were provided for reading. And participants were allowed to refer to them freely during the simulated dialogues. If participants felt they had not performed well in the dialogue, they were allowed to try again as many times as necessary.

Subsequently, for the videos used in the simulated dialogues, participants were instructed to continuously indicate gaze instruction points with the mouse cursor, recording the

cursor coordinates for each frame of the video. Participants were generally instructed to freely click and move the mouse cursor to show where they wanted the avatar to look, and in cases where there was no specific point they wanted the avatar to gaze at, they were instructed not to click. Examples of desired gaze locations included points of interest such as the customer's face or hands. Additionally, participants were verbally requested to aim for behavior appropriate for customer service when giving instructions.

D. Postprocess for Operator-Side Data

The participants' speech was transcribed using the Google Cloud Speech-To-Text API and stored in JSON format. The transcription data included the recognized words as well as the start and end timestamps for each word. Additionally, to assess the divergence between the participants' gaze positions and the target positions indicated by mouse interaction with the avatar, the participants' estimated gaze positions were extracted from video data using a gaze estimation model. Among various models, we employed the ETH-XGaze [18] deep learning model, as it offers a wide range of maximum gaze angles and head angles based on the dataset used during training. ETH-XGaze outputs the estimated head position vector $\boldsymbol{h} = [h_x, h_y, h_z]$ and the estimated gaze direction vector $\boldsymbol{g} = [g_x, g_y, g_z]$ in the camera coordinate system $\Sigma_c = (x_c, y_c, z_c)$. The estimated gaze point $\boldsymbol{p}$ on the screen in the screen coordinate system $\Sigma_s = (x_s, y_s, z_s)$ was computed using the vectors $\boldsymbol{h}$ and $\boldsymbol{g}$ output by the model, as $\boldsymbol{p} = [p_x, p_y, 0] = \boldsymbol{h} + \boldsymbol{g} \times \left| \frac{h_z}{g_z} \right|$. The estimated gaze positions fell within the screen boundaries 63.52% of the time, while 36.48% were outside the screen.

E. Additional Labeling for Avatar-Side Video Data

1) Customer Behavior Labeling: To capture how operators changed gaze instruction locations in response to various customer situations and contexts, key situational changes in the dialogue scenes were labeled as customer behavior attributes. Table I shows the labels for customer behavior designed in this study. Major situational changes in this study include the dialogue partner, i.e., the customer's engagement with and actions toward the avatar or objects around the avatar. Referring to the user behavior phases in prior research [19], actions toward the avatar range from looking at the avatar, approaching, stopping, speaking, to leaving. Additionally, as a frequently occurring action in the videos used, the action of customers bringing their faces closer to the avatar was also included as a label. Actions toward objects around the avatar that frequently appear in the videos were also labeled as mentioning and showing interest in the object. Each frame of the video was labeled with customer behavior attributes, assigning a value of 0 if the action was not occurring and 1 if it was. If multiple customers were included in the same frame, a value of 1 was assigned if any one of the customers was determined to be acting.

1	Whether the customer is **looking at the avatar**
2	Whether the customer is **approaching the avatar**
3	Whether the customer is **stopping in front of the avatar**
4	Whether the customer is **speaking to the avatar**
5	Whether the customer is **moving away from the avatar**
6	Whether the customer is **bringing their face closer to the avatar**
7	Whether the customer is **mentioning objects**
8	Whether the customer is **showing interest in objects**

TABLE I: Labels for annotating customer behavior.

2) Areas of Interest (AOIs) Labeling: For the dialogue scenes addressed in this study, which primarily involve customer service interactions, potential main gaze targets for the avatar include regions occupied by people present in the videos and areas near the avatar where recommended products or objects of mention are placed. Thus, this research defines the regions in the videos where people appear and the areas of objects that the avatar should be mentioned as AOIs, including their temporal variations in labeling. Additionally, considering that the same person can have different levels of engagement in the dialogue, ranging from actively participating to not participating, a customer attribute was introduced to distinguish between individuals in the person regions. The classification of participation structures in dialogue [20] was referenced to define the customer attributes as follows:

1) ***Addressee***: An individual is directly involved in the dialogue with the avatar and the primary recipient of the conversation.
2) ***Side Participant***: An individual is directly involved in the dialogue with the avatar but not the primary recipient. In this study, individuals located near the *Addressee* and involved in the dialogue are defined as *Side Participant*.
3) ***Bystander***: An individual who is near the location where the dialogue with the avatar is taking place but is not directly involved in the dialogue. In this study, individuals near the *Addressee* but not involved in the dialogue are defined as *Bystander*.
4) ***Passerby***: An individual who is not near the location of the dialogue or is near but not looking toward the avatar or in any way involved in the dialogue. All customers in the video are initially classified as *Passerby*.

The above four customer attributes change according to the behavior of the customers. Fig. 3 shows the rules of these changes as state transitions.

The positions of individuals appearing in the videos were detected using the object detection model YOLO v7 [21], and the attributes of these individuals as dialogue participants were labeled across all frames of the video using the CVAT [22]. The authors established annotation rules based on the state transitions of the aforementioned dialogue participant attributes, and annotators carried out their work under these guidelines. Additionally, the locations of referred objects were manually annotated by the authors with the label *Object to be Referred*. In cases where a region did not correspond to either dialogue participant attributes or referenced objects,

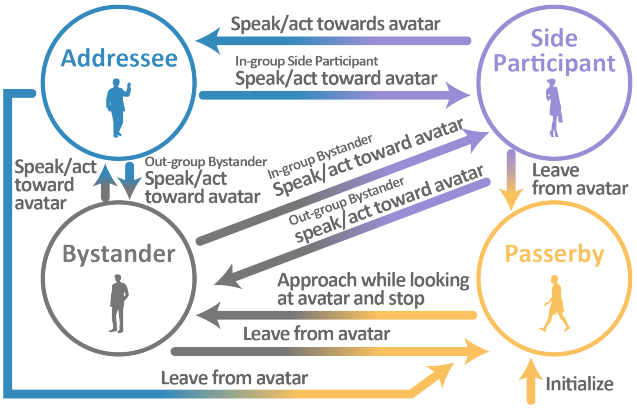

Fig. 3: **State transition model of customer attributes annotation.** Annotators classified the customers in the scenes based on this rule. Gaze instruction data were aggregated for each classified customer attribute.

the label *Others* was applied.

F. Analysis

Based on the labeling of AOIs, we conducted a detailed observation of their distribution across various attributes. These attributes include customer behavior attributes, video attributes, and participant attributes. The analysis aims to elucidate how these attributes influence the labeling of AOIs. We used the Wasserstein distance to calculate the differences in distribution compared to the baseline.

III. RESULTS

A. Overall Trends

The distribution of gaze instruction percentages across all AOIs from all videos is shown at the top of Fig. 4. The first row in the graph shows the distribution across all data, used as a baseline for comparison with other distributions. Approximately 60% of gaze instruction behaviors are directed toward the *Addressee*, indicating that gaze is frequently instructed toward the current dialogue partner in many scenes. The gaze directed toward other customer attributes (*Side Participant*, *Bystander*, *Passerby*) accounted for roughly 10% each. The proportion of gaze instructions toward Objects to be Referred was observed to be very low, at only 0.4%.

B. Notable Changes from Baseline

The second to sixth bars in Fig. 4 represent the distribution of conditions with notable differences compared to the baseline. From the conditions with a remarkable distance from the baseline, we extracted the top four where the proportion of gaze instructions to the *Addressee* increased. Compared to the baseline *Addressee* proportion of 62.9%, the increase was +24.9 points when customers were bringing their faces closer to the avatar ($W_p = 8.30$), +13.9 points when customers were showing interest in objects ($W_p = 5.50$), +15.1 points when customers were mentioning objects ($W_p = 5.16$), +14.8 points when operator's main task is a question answering ($W_p = 4.95$), and +9.0 points when customers were speaking to the avatar ($W_p = 3.91$), where W_p is the Wasserstein distance with the baseline distribution. In these four conditions, the proportion of gaze instructions toward *Bystander*, *Passerby*, and *Others* decreased.

C. Gaze Instructions Toward Mentioned Objects

Next, we examine the gaze instruction positions when operators refer to objects situated on the avatar's side. It is posited that the majority of gaze instructions directed at objects to be mentioned occur either immediately before or after the participant mentions the object. We investigated the proportion of gaze instructions directed at the *Object to be Referred* area within 5 seconds before and after the participant's mention. Words thought to be referring to the object (proper nouns and pronouns) and their times of occurrence were extracted through annotation from the participant's transcript data. The proportion of cases in which operators provided gaze instructions to the object after mentioning it was 5.33%, whereas cases without gaze instructions accounted for 94.6%. This indicates that the frequency of gaze instructions directed at the object area immediately before or after mentioning the object is exceedingly low.

D. Gaze Instructions with Off-Screen Operator Gaze

Subsequently, we present the outcomes for gaze instruction locations during periods when the operator's gaze was away from the screen. It is important to note, given that the simulated dialogue task and the gaze instruction task were conducted separately, that instances when the operator looked away from the screen and when gaze instructions were issued, although at different times, pertained to the same individual. The seventh and eighth bars in Fig. 4 show the proportion of gaze instructions directed at various AOIs when the participant's gaze was either on the PC screen or elsewhere during the dialogue. No significant variation was observed in the distribution of gaze instruction proportions between instances when the participant's gaze was within the screen and when it was outside, indicating that the presence or absence of visual attention to the screen did not influence the pattern of gaze instructions.

E. Gaze Instructions in Avatar-Led Dialogues

The last four bars in Fig. 4 present the distributions of gaze instruction proportions by task and the presence or absence of a script. As noted earlier, when the main task involves question and answer, there is an increase in the proportion of gaze instructions directed at the *Addressee*. In contrast, when the avatar's main task is to provide recommendations, the proportion directed at the *Addressee* decreases by more than 5 points relative to the baseline, with a slight increase observed in the proportions for *Side Participant*, *Bystander*, and *Passerby*. Additionally, when a script is provided, there is a decrease in the proportion of gaze instructions toward the *Addressee* by approximately 4.7 points compared to the baseline, along with a slight increase in the proportions for other areas of interest.

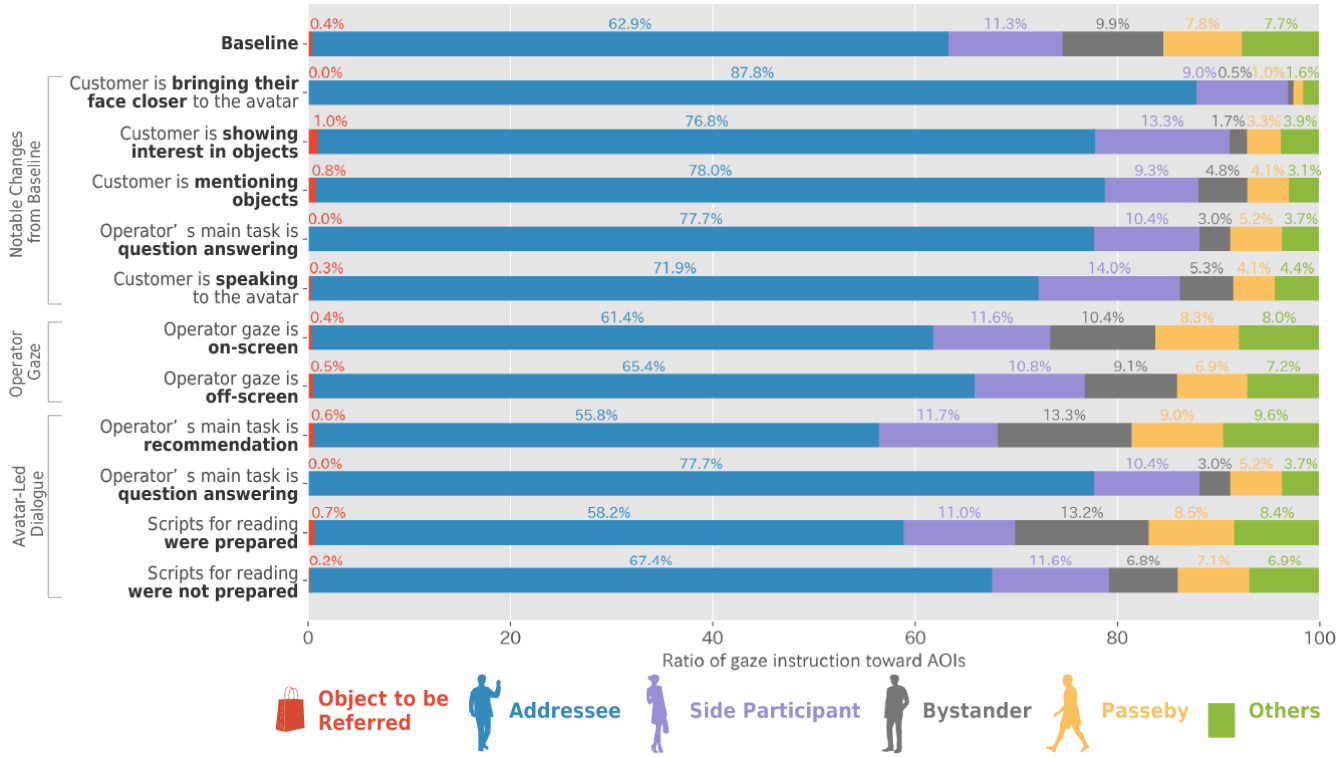

Fig. 4: The ratio of gaze instructions toward the AOIs. The vertical axis represents the labels of each condition and the horizontal axis shows the frequency of gaze instruction points contained within each AOI as a percentage, with the total aggregated to 100%.

IV. DISCUSSION

Overall, more than 60% of gaze instruction behaviors were directed toward the *Addressee* area, confirming that in most dialogue situations, gaze instructions were directed toward the person directly involved in the conversation. The gaze directed toward other customer attributes (*Side Participant*, *Bystander*, *Passerby*) was approximately 10% each. Similar to this study, previous research [23] that calculated the distribution of gaze positions based on the participation attributes of the interlocutors found that in dyadic dialogues where a *Bystander* was present, 76% of the gaze was directed toward the *Addressee*, dominating the gaze distribution, with *Bystander*s receiving almost no gaze at 8%. Unlike previous studies, this research was conducted in a remote dialogue scenario using a PC screen, and gaze instruction positions were used as the data source for gaze positions. Despite these differences, the similar outcomes suggest that the actual gaze situations in face-to-face dialogues targeted by previous studies could be somewhat replicated by the data collection method used in this research.

Notably, throughout the overall trend, the percentage of gaze instructions directed toward *Object to be Referred* was very low, at 0.4%. This indicates that moments where objects are directed to be looked at are scarce. Even when focusing on recommendation task videos where an *Object to be Referred* to was always present, it was only 0.6%. This means the average seconds directed toward the object area per video per person is only 0.52s. Additional analysis of recommendation task videos revealed that the average time for 81 participants to direct their gaze instruction to the target object areas was less than 1 second, and the average time for 11 out of 16 recommendation task videos was less than 0.5 seconds. As mentioned in section II-B, participants were requested to instruct behavior appropriate for customer service. They referred to the object many times in many video scenes and had many times to prompt customers to act toward it. Nevertheless, the frequency of gaze instructions toward object areas was low. A possible factor could be participants' lack of experience in remote customer service or remote dialogue. Most participants in this study were unfamiliar with conducting dialogues using teleoperated avatars. In particular, almost none of them had seen the avatar's behavior from the customer's side, and it is possible that there was a lack of lessons and training on avatar behavior appropriate for customer service, such as joint attention. Conversely, however, this reflects the high difficulty in providing appropriate gaze instructions identified as a challenge in this study, suggesting the need for operational support in avatar gaze behaviors.

Focusing on attributes that increase the percentage of gaze instructions toward each AOI, it was suggested that gaze instructions toward the *Addressee* increase when the customer engages in some action toward the avatar (stopping, speaking, approaching). All actions by the customer

toward objects, such as mentioning or reacting to objects, also suggested a possible increased focus on the *Addressee*. This suggests that the avatar must be controlled to view the customer according to the customer's behavior toward the avatar and the product. In human-human interaction in service encounters, it was expected that when a customer performed some action on an object, the employee might pay joint attention and look at the same object. However, the participants in this study were not highly skilled, which may have increased their gaze instructions at *Addressee* possibility. Further research is needed on participant skill levels in customer service and avatar operation. At least one implication of these results for avatar gaze support is that recognizing the object of the customer's action (avatar or surrounding objects) is very important, and improving its performance will lead to natural avatar gaze control.

Focusing on the operator's situation, the percentage of gaze toward AOIs did not vary regardless of whether the operator was looking at the screen, suggesting that operators want the avatar to maintain the same gaze behavior even when not looking at the screen. This implies that control methods that directly reflect the operator's gaze or head orientation in the avatar's gaze behavior are not desirable, and semi-autonomous gaze control support tailored to the situation is suggested. Additionally, it was observed that in most instances when operators mentioned objects, gaze instructions were not directed toward the mentioned object areas. The timing of mentioning objects in conversations varied, including mentioning product names as part of task explanations or when asking questions to customers. Therefore, it's crucial not only to identify whether objects were mentioned but also whether there was an operator's intention to direct gaze toward them when mentioned. To control appropriate gaze behavior, it's necessary to accurately recognize the operator's unexpected actions or the intentions behind their speech and adjust the avatar's gaze control accordingly.

Moreover, the percentage of gaze toward AOIs also varied depending on the main task assigned to the operator and the information provided. When the main task was question-answering, attention toward the *Addressee* increased, while in recommendation scenarios, the percentage toward the *Addressee* decreased, and the percentage toward other areas increased. In question-answering tasks, it's sufficient to direct gaze behavior toward a specific customer who asked a question. On the other hand, the recommendation tasks require explaining the product to several people around the avatar and require active eye contact with all of the group customers and with the bystanders around them. A similar consideration applies to the presence or absence of scripts. In this study, scripts were provided for half of all videos, allowing free reference during dialogues. The presence of scripts likely led to a dispersion of gaze instructions due to the increased content the avatar needed to convey. A mechanism may be needed to recognize what task the operator or avatar is about to perform and switch between focused and dispersed attention.

Finally, based on the analysis results obtained, we propose a gaze decision model to support the gaze behaviors of teleoperated avatars as a hypothesis. The focus targets for teleoperated avatars should be individuals performing any action towards the avatar, such as speaking, asking questions, directing gaze, or bringing their faces closer. Additionally, individuals targeted by actions from the dialogue partner, such as speaking, touching, exchanging glances, or approaching, as well as objects mentioned by the operator or dialogue partner, should also be considered for gaze. The duration of gaze for each action varies, necessitating adjustments in gaze duration and direction depending on the context. When there are multiple targets to focus on, the gaze should be distributed. However, if there is only one target, the gaze should concentrate on that specific target. The duration of gaze on each target follows a probability distribution, and when a new target of interest appears, the focus of gaze should shift. For example, when the avatar makes an announcement, there may be multiple targets to attend to, but when the avatar answers a question, the target of interest is singular. The gaze decision model is constructed based on the information of who is doing what to whom. Therefore, to automate the avatar's gaze behavior in response to customer actions, it is essential to accurately recognize the actions of customers and the targets of those actions. If these actions can be recognized, they can be implemented with a simple algorithm. Gaze duration is sampled from a normal distribution with a mean of one second, after which the gaze shifts to another target. If a new action is detected, the gaze shifts to that action. The primary challenge for future research is to accurately recognize who is performing what action toward whom and to determine the gaze targets based on this information.

V. CONCLUSION

In this study, we aimed to derive insights for supporting gaze control in teleoperated avatars and constructed a dataset of gaze instruction data where operators indicated where they wanted the teleoperated avatar to look. Exploratory data analysis was conducted on that dataset to analyze situations in which there were significant differences compared to the overall average and situations in which there were not. By aggregating the proportions of gaze instruction locations directed toward defined AOIs, such as people and objects, and analyzing these about both the avatar's and the operator's situations, our findings suggest that supporting gaze behavior control requires accurate recognition of the target and type of customer actions around the avatar, as well as the operator's unexpected actions and speech intentions. It is essential to adjust the avatar's gaze control based on this information.

As a future challenge, it is first necessary to verify whether the insights and hypotheses presented in this study are correct. To achieve this, additional experiments that support this study's findings across a broader range of situations are required. Since this research was limited to service dialogue scenes, it is not guaranteed that the insights presented can be universally applied to other dialogue scenes, such as

conversations with friends or acquaintances aimed at building relationships through casual talk. Moreover, the customer behaviors investigated in this study focused on actions directed toward the avatar or objects near the avatar, excluding actions toward people around the customer. Therefore, further investigation is needed for situations where such behaviors are observed. And, as the operators were exclusively Japanese, the impact of nationality or cultural differences on gaze behavior has not been clarified, necessitating careful consideration when applying the findings of this study to different environments. Also, a detailed analysis of the temporal trajectories of gaze instruction locations has not been conducted. Utilizing methods such as dynamic scanpath analysis could reveal characteristics related to the order of gaze points visited and fixation duration, potentially providing information necessary for generating more appropriate gaze behaviors.

Furthermore, there is a demand for verifying a gaze control model based on the insights provided in this study. This model must have the ability to recognize actions and their targets around the avatar from video information in real-time while simultaneously recognizing the operator's unexpected actions and speech intentions, adjusting these pieces of information to control appropriate gaze behaviors. Further exploration is needed regarding individual recognition technologies and adjustment algorithms, especially ensuring accuracy and real-time capabilities in dynamic environments. The impact of avatar gaze behaviors based on the model on the operator and the interlocutors in actual dialogue scenes must be verified through experiments involving various dialogue scenarios and participant groups.

Lastly, since this study is based on the process of exploratory data analysis, the presented analysis results cannot conclusively prove anything. In particular, as the statistical significance of differences has not been confirmed, the results should be used as a foundation for hypothesis generation. In this respect, further hypothesis-confirming research and analyses based on a broader dataset are required. With these limitations in mind, future research is expected to deepen the understanding of gaze control across various dialogue scenes and cultural backgrounds of operators, contributing to the development of more effective teleoperated avatar systems.

REFERENCES

[1] J. Baba, S. Song, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "Local vs. Avatar Robot: Performance and Perceived Workload of Service Encounters in Public Space," *Frontiers in Robotics and AI*, vol. 8, 2021.

[2] Y. Okafuji, S. Song, J. Baba, Y. Yoshikawa, and H. Ishiguro, "Influence of collaborative customer service by service robots and clerks in bakery stores," *Frontiers in Robotics and AI*, vol. 10, p. 1125308, Jul. 2023.

[3] M. Kawata, M. Maeda, Y. Yoshikawa, H. Kumazaki, H. Kamide, J. Baba, N. Matsuura, and H. Ishiguro, "Preliminary Investigation of the Acceptance of a Teleoperated Interactive Robot Participating in a Classroom by 5th Grade Students," in *Social Robotics*, ser. Lecture Notes in Computer Science, F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, and S. S. Ge, Eds. Cham: Springer Nature Switzerland, 2022, pp. 194–203.

[4] A. Yoshida, H. Kumazaki, T. Muramatsu, Y. Yoshikawa, H. Ishiguro, and M. Mimura, "Intervention with a humanoid robot avatar for individuals with social anxiety disorders comorbid with autism spectrum disorders," *Asian Journal of Psychiatry*, vol. 78, p. 103315, Dec. 2022.

[5] "A survey on the design and evolution of social robots — Past, present and future," *Robotics and Autonomous Systems*, vol. 156, p. 104193, Oct. 2022.

[6] L. Itti, N. Dhavale, and F. Pighin, "Photorealistic Attention-Based Gaze Animation," in *2006 IEEE International Conference on Multimedia and Expo*, Jul. 2006, pp. 521–524.

[7] K. Tatarian, M. Chamoux, A. K. Pandey, and M. Chetouani, "Robot Gaze Behavior and Proxemics to Coordinate Conversational Roles in Group Interactions," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, Aug. 2021, pp. 1297–1304.

[8] S. Gillet, R. Cumbal, A. Pereira, J. Lopes, O. Engwall, and I. Leite, "Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 303–311.

[9] S. L. Rogers, R. Broadbent, J. Brown, A. Fraser, and C. P. Speelman, "Realistic Motion Avatars are the Future for Social Interaction in Virtual Reality," *Frontiers in Virtual Reality*, vol. 2, 2022.

[10] D. Cazzato, C. Cimarelli, J. L. Sanchez-Lopez, M. A. Olivares-Mendez, and H. Voos, "Real-Time Human Head Imitation for Humanoid Robots," in *Proceedings of the 2019 3rd International Conference on Artificial Intelligence and Virtual Reality*, ser. AIVR 2019. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 65–69.

[11] S. Vikkelsø, T.-H. Hoang, F. Carrara, K. D. Hansen, and B. Dinesen, "The telepresence avatar robot OriHime as a communication tool for adults with acquired brain injury: An ethnographic case study," *Intelligent Service Robotics*, vol. 13, no. 4, pp. 521–537, Oct. 2020.

[12] M. Friedrich, N. Rußwinkel, and C. Möhlenbrink, "A guideline for integrating dynamic areas of interests in existing set-up for capturing eye movement: Looking at moving aircraft," *Behavior Research Methods*, vol. 49, no. 3, pp. 822–834, Jun. 2017.

[13] A. Li and Z. Chen, "Representative Scanpath Identification for Group Viewing Pattern Analysis," *Journal of Eye Movement Research*, vol. 11, no. 6, Nov. 2018.

[14] S. Song, J. Baba, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "Costume vs. Wizard of Oz vs. Telepresence: How Social Presence Forms of Tele-operated Robots Influence Customer Behavior," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2022, pp. 521–529.

[15] S. Song, B. Jun, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "Service Robots in a Bakery Shop: A Field Study," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 134–140.

[16] S. Mochizuki, S. Yamashita, K. Kawasaki, R. Yuasa, T. Kubota, K. Ogawa, J. Baba, and R. Higashinaka, "Investigating the Intervention in Parallel Conversations," in *Proceedings of the 11th International Conference on Human-Agent Interaction*, ser. HAI '23. New York, NY, USA: Association for Computing Machinery, Dec. 2023, pp. 30–38.

[17] S. Song, J. Baba, Y. Okafuji, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "Wingman-Leader Recommendation: An Empirical Study on Product Recommendation Strategy Using Two Robots," *IEEE Robotics and Automation Letters*, pp. 1–7, 2024.

[18] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation," Jul. 2020.

[19] J. Müller, F. Alt, D. Michelis, and A. Schmidt, "Requirements and design space for interactive public displays," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1285–1294.

[20] H. H. Clark and T. B. Carlson, "Hearers and Speech Acts," *Language*, vol. 58, no. 2, pp. 332–373, 1982.

[21] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Jul. 2022.

[22] "CVAT," https://www.cvat.ai/.

[23] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, pp. 12:1–12:33, Jan. 2012.