Toward User-Aware Interactive Virtual Agents: Generative Multi-Modal Agent Behaviors in VR

Bhasura S. Gunawardhana* University of Houston
Yunxiang Zhang† New York University
Qi Sun ‡ New York University
Zhigang Deng § University of Houston

Figure 1: With our method, VR users can interact with a virtual agent naturally through a combination of multiple modalities including verbal expressions, body language, and eye contact. In the right, the top row shows a VR user's multi-modal behaviors while he interacts with a virtual agent; the bottom row shows the agent's real-time reactions.

Abstract

Virtual agents serve as a vital interface within XR platforms. However, generating virtual agent behaviors typically rely on pre-coded actions or physics-based reactions. In this paper we present a learning-based multimodal agent behavior generation framework that adapts to users' in-situ behaviors, similar to how humans interact with each other in the real world. By leveraging an in-house collected, dyadic conversational behavior dataset, we trained a conditional variational autoencoder (CVAE) model to achieve user-conditioned generation of virtual agents' behaviors. Together with large language models (LLM), our approach can generate both the verbal and non-verbal reactive behaviors of virtual agents. Our comparative user study confirmed our method's superiority over conventional animation graph-based baseline techniques, particularly regarding user-centric criteria. Thorough analyses of our results underscored the authentic nature of our virtual agents' interactions and the heightened user engagement during VR interaction.

Index Terms: Virtual agents, human-VR interaction, user-conditioned motion generation, user-aware interaction.

1 Introduction

Extended Reality (XR) technologies have heralded an exciting era of immersive interaction, enabling users to see, hear, and interact with virtual content as if it were part of their physical surroundings. This paradigm shift towards lifelike virtual environments naturally extends our desire for virtual agents that emulate human-like communication and interaction. The development of such intelligent virtual agents is paramount for future XR systems, with potential benefits across diverse applications such as client services, professional work settings, and video games.

Despite long-standing aspirations for responsive virtual agents in XR platforms, these agents have not yet achieved the level of naturalness and realism required for real-time interaction with the level of naturalness and realism required, often making them easily distinguishable from actual users [30]. This limitation is primarily due to several shortcomings in previous approaches: (i) *Overreliance on scripted rules:* Many prior attempts have leaned heavily on pre-scripted rules and a fixed set of pre-coded responses, limiting the agents' adaptability and authenticity. (ii) *Lack of contextual adaptation:* These virtual agents typically fail to adapt to the user's in-situ behaviors, missing the contextual nuances essential for lifelike interactions. (iii) *Limited response diversity:* The responses generated by such agents generally lack necessary diversity, leading to predictable and repetitive interactions. Addressing these issues is crucial to achieving the goal of creating responsive and socially-aware agents within XR environments.

To overcome these limitations, we developed a machine-learning-based virtual agent system capable of natural interactions with VR users using multimodal cues in real-time. Our virtual agents engage in verbal communication while performing appropriate eye contact and body language based on the VR user's behavior. This is achieved by learning a conditional variational autoencoder (CVAE) model using a high-quality dataset of dyadic communication behaviors. Specifically, we tracked subjects' eyes, head, and body motion during natural conversations and trained the CVAE model to generate one subject's behaviors while conditioned on the other's. Consequently, our model can generate natural eye contact and responsive body movements for virtual agents based on the VR user's verbal and non-verbal actions.

Our user study and quantitative evaluations demonstrate that our model can bluegenerate user-aware communication behaviors for virtual agents, significantly facilitating natural human-agent interactions in VR. Virtual agents with model-generated behaviors are perceived as plausible and appropriate by human users, leading to more engaging XR interaction experiences. This is validated through subjective evaluations and quantifiable behavioral measures.

*e-mail: bsgunawardhana@uh.edu
†e-mail: yunxiang.zhang@nyu.edu
‡e-mail: qisun@nyu.edu
§e-mail: zdeng4@central.uh.edu

In summary, our work makes the following main contributions: (i) A multimodal and high-fidelity dataset of dyadic natural communication behaviors, including speech, eyes/head, and full-body motion data. Besides the dataset, we make our model/code available to the community (*https://cg-im.github.io/ismar24_intelligent_agent/*). (ii) A user-aware natural communication behavior generation model for interactive virtual agents. (iii) An end-to-end runtime system that generates responsive and human-like agent behaviors based on the tracked behaviors of the VR user.

2 Related Work

Motion synthesis for virtual humans. Numerous previous efforts have been made to create gaze models, from conveying emotional nuances in gaze behavior [54] to emulating turn-taking protocols [32, 60], highlighting conversation points of interest [38], and simulating agent attending behaviors across various activities and cognitive actions [26]. Some existing studies concentrated on leveraging statistical models to animate natural gaze behaviors [14, 17, 33, 37, 43], without incorporating the semantics of communicative context.

Researchers introduced various techniques to generate realistic head motion, including using dynamic programming algorithms to synthesize head motion [12] and co-speech gestures [59], and extending Hidden Markov Models (HMMs) for head motion generation [10, 11, 35, 55]. Additionally, rule-based methods have been used to facilitate virtual conversations between two parties, driven by speech and semantically annotated texts [45]. Researchers have also focused on effectively synchronizing head movement with gaze during conversations [25, 33, 43, 46] and learning rules from annotated conversation datasets to generate gestures, including head movements, for listening agents [19, 44].

Researchers learned statistical models with predefined gesture units [39, 47], or employed deep learning techniques including deterministic models [9, 41], generative models [67, 68], and diffusion models [4] for co-speech gesture generation. To ensure semantic alignment between speech and gesture, neural network systems have been specifically designed to explicitly model both gesture rhythm and content semantics [5, 6, 40].

Human-agent interaction. Digital agents have been an intuitive and promising interface that contextualizes complex AI computation to human users [42]. Examples include the Microsoft Office Clippit, Apple Siri, Google Assistant. One representative implementation is robots. An extensive line of research has been proposed to study and understand the interaction between humans and robots, such as trust [20, 66] and efficiency [13, 18]. In virtual environments, humanoid agents have been implemented as virtual agents to enable social naturalness and presence. The main advantages of using agents in XR-based communications include trustworthiness and co-presence [8], but depending on the interpretation of visual realism [22, 49] and behavioral realism [28, 53, 65]. Also, compared to visual realism, the behavioral realism of agents is generally believed to be a more important factor for social interaction and social presence [50, 64]. The behavioral realism includes verbal and non-verbal behaviors, responsiveness, and interactivity with the user [30].

Similar to motion synthesis techniques, existing immersive agents are either driven by physics-based engines [63], pre-computation [34], or mapping real users' actions at run-time (e.g., the Meta Horizon Worlds). Recent studies used machine learning models, including Generative Adversarial Networks (GAN) [24, 48] and Long Short-Term Memory (LSTM) models [15], to generate certain non-verbal facial behaviors of the agent (e.g., facial expressions and head movements) based on user states such as affective state and facial expressions. However, modeling the implicit and cognitive process to drive agents that adapt to the real users' actions in a socially realistic manner still remains an under-explored challenge. The main goal of this work is to develop an end-to-end framework with a high-fidelity dataset and a real-time deep neural network.

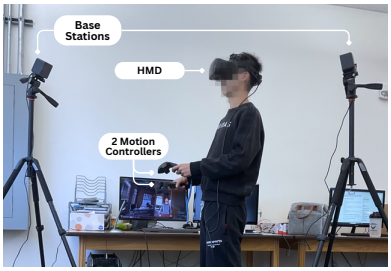

Figure 2: A runtime snapshot of our system, where a VR user wears a Varjo Aero HMD and holds two hand motion controllers.

3 System Overview

In this section, we provide an overview of our end-to-end interactive conversational agent system, developed using Unreal Engine 5 [16]. Our system uses state-of-the-art technology and hardware components, as illustrated in Figure 2.

To ensure a high-quality virtual reality experience, we choose the Varjo Aero HMD headset [3] due to its high VR fidelity and integrated eye-tracking capabilities. Additionally, we incorporate two HTC vive [2] motion controllers to capture hand movements of the VR user. The controllers are tracked within a two-meter diagonal setup consisting of two HTC base stations. We use Steam to create a user-friendly play area that measures approximately 1.5 meters by 1.5 meters, allowing immersive VR conversations and interactions.

Before a user engages with our system, we need to perform an HMD calibration process for the specific user. Specifically, foveated rendering calibration entails the user focusing on a white circle shown on the headset screen for a brief period. Foveated rendering optimizes the resolution around the fixation region, enhancing the visual experience.

During human-agent interaction, the VR user needs to push and hold a trigger button on the left-hand motion controller, while speaking naturally with gestures. Once the user completes the speech, he/she can release the trigger button on the left-hand motion controller. The virtual agent in our system responds interactively to the user's speech input, delivering not only vocal responses but also synchronized lip movements, eye contact, body language, and hand gestures. All these dynamic behaviors of the virtual agent are generated in real-time by our system, facilitating multi-round conversations.

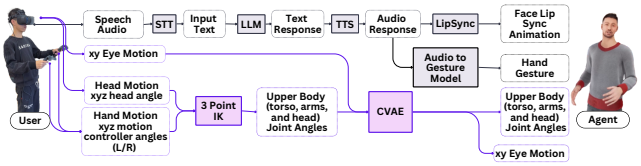

Figure 3: Pipeline overview of our system

Figure 3 shows a pipeline overview of our system and its core components. Inputs to our system can be categorized into two types: audio and motion. Audio consists of the user's live speech signals, while motion includes 11 features: three head rotation angles from the HMD, three rotation angles from two motion controllers for left/right hands, respectively, and HMD-provided eye-tracking XY positions. Our system outputs two types of responses: audio responses and motion responses. Motion responses are further divided into third-party motions and our CVAE model-generated motions, which are blended into the agent character's bone hierarchy as the final output. Third-party motions include facial lip-sync blendshapes for the MetaHuman skeleton and joint angles of both hands and

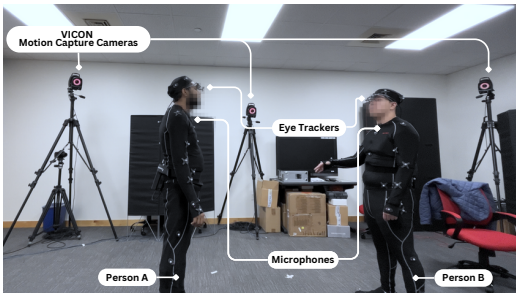

Figure 4: A snapshot of our dyadic conversation data acquisition experiments

fingers. Our CVAE model generates 35 upper body (i.e., torso, arm, head) joint angles and XY eye coordinates in the agent's viewing frustum. The internal processing between input and output is explained below.

When the VR user speaks, we first employ the OpenAI's Whisper API to transcribe the user's speech to text (STT). The text is then further fed into the ChatGPT developed by OpenAI, a sophisticated large language model (LLM), to generate a text response, ensuring the preservation of conversational context and flow in the generated text response. The LLM-generated text responses are subsequently transformed into audio using text-to-speech (TTS) technology (i.e., Google TTS with configurations speechtext, premium, en-US-Neural2-A, MALE, MP3). To further enhance the immersive experience, we utilize the MetaHuman SDK [1] to generate synchronized lip-sync animations based on the generated audio. To efficiently generate user-aware conversational behaviors for the virtual agent, including eye movements, torso movements, arms and head movements, we train a CVAE model (in Section 5). It takes input from the VR user's in-situ behavior, and outputs the eye movements, head movements, and body movements of the virtual agent. Furthermore, we integrate the method in [29] to dynamically generate hand gestures based on audio input, further improving the realism of conversational experience.

4 Data Acquisition

In this work, we used an in-house data acquisition system to record high-quality, synchronized, multi-modal behavior data from dyadic conversations [36]. The resultant dataset encompassed various behavioral cues, including eye movements, head movements, hand gestures, body movements, and audio signals. Each participant in dyadic conversations was equipped with specialized equipment to facilitate data capture, as depicted in Figure 4.

To capture head movements, hand gestures, and body movements, we outfitted each participant with a motion capture (mocap) suit in our experimental setup. These suits were placed with optical mocap markers that allowed the VICON optical motion capture system to precisely record the motions of all conversational participants.

For eye movement tracking, we chose the Ergoneers Dikablis Glass 3 eye tracker due to its high accuracy, wireless functionality, and comprehensive software suite that facilitated in-depth analysis. This eye tracker featured two cameras directed towards the user's eyes, tracking eye movements through the pupil's xy coordinates. Simultaneously, a third camera captured video footage from the participant's perspective, which we utilize as a reference for the motion data cleanup process. This setup ensured comprehensive eye movement data acquisition.

Audio data, a crucial component of our dataset, was captured using high-definition microphones to ensure clarity and high-quality recordings. Synchronization of the eye tracking data and the audio data was orchestrated using the D-Lab software provided by Ergoneers, ensuring precise alignments in our experiments.

In our data acquisition experiments, a total of three participants (two males and one female) engaged in dyadic conversations. They were formed in three pairs and were encouraged to discuss freely topics of personal interest, such as hobbies and campus life. The recorded data from these interactions totaled approximately 62 minutes, 63 minutes, and 55 minutes for the respective pairs, cumulatively amounting to 180 minutes of raw recorded data. Subsequently, we further processed and cleaned the data, resulting in 112 minutes of high-quality refined data.

5 Our Method

We formulate the task of user-aware behavior synthesis for virtual agents as a conditional generative learning problem. Specifically, we train a conditional variational autoencoder (CVAE) model [58] to synthesize the virtual agent's head, eye, and body movements based on the tracked in-situ motion of the VR user, which also encompasses eye, head, and body movements. CVAE is a variant of the base variational autoencoder (VAE) model [27]. In the following, we briefly describe the VAE model and our CVAE model.

5.1 Variational Autoencoder (VAE)

A VAE model consists of an encoder and a decoder. The encoder maps input data x to a latent space z with parameters θ and ϕ. The decoder maps the latent space back to the data space. The objective of VAE is to maximize the evidence lower bound (ELBO) on the marginal likelihood as follows.

$$\log p_\theta(x) \geq E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)), \quad (1)$$

where $p_\theta(x|z)$ denotes the likelihood function, $q_\phi(z|x)$ denotes the approximate posterior distribution of z, $p(z)$ denotes the prior distribution of z (typically assumed to follow a standard normal distribution), and D_{KL} is the Kullback-Leibler divergence.

5.2 Conditional Variational Autoencoder (CVAE)

In a CVAE model, both the encoder and decoder are influenced by additional contextual information denoted as c. This implies that the encoder and the decoder not only consider the data x and the latent variable z, but also this supplementary context c.

The objective function for CVAE becomes:

$$\log p_\theta(x|c) \geq E_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)] - D_{KL}(q_\phi(z|x,c)||p(z|c)), \quad (2)$$

where $p_\theta(x|z,c)$ is the conditional likelihood function, $q_\phi(z|x,c)$ denotes the approximate posterior distribution of z conditioned on c, and $p(z|c)$ denotes the prior distribution of z conditioned on c. In practice, this conditioning is often achieved by incorporating c as an additional input to both the encoder and decoder networks.

5.3 Our CVAE Model and Implementation

In our specific problem, we possess the ground truth data c_A, which includes upper body joint angles and eye gaze data for person A at each frame. Similarly, x_B comprises upper body joint angles and eye gaze data for person B at each frame. Notably, both c_A and x_B consist of 35 features, encompassing head movement angles, and joint angles associated with the pelvis, thorax, left and right clavicles, humeri (upper arms), radii (forearms), and hands. Each of these anatomical segments contributes three joint angles along the three axes. Furthermore, eye movement features, represented as *eye_x* and *eye_y*, denote the horizontal and vertical positions of the pupil, respectively. The primary objective of our CVAE model is to generate these motions for person B, frame by frame, based on the motion of person A. Figure 5 illustrates the architecture of our CVAE model.

Our CVAE model was implemented using PyTorch. The encoder module consists of two encoder blocks, each of which contains a

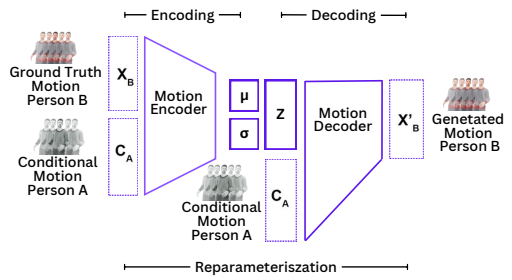

Figure 5: Architecture illustration of our CVAE model

series of layers for effectively processing input data. Within each block, the first layer is a 1-dimensional convolutional layer, followed by a batch normalization layer to ensure stable activations. The CVAE model was tailored to process inputs with 35 features, and we defined the latent space with a dimensionality of 10. The model architecture included hidden layers with 128 units, specifically designed to handle sequence data with a length of 32.

5.4 CVAE Model Training Details

The training was initialized with a random seed of 123. We set the initial learning rate to 0.001 and trained the model using mini batches, each containing 128 samples. During our model training, we optimize the parameters θ and ϕ to maximize the Evidence Lower Bound (ELBO) evaluation metric, as defined in the aforementioned objective function (Eq. 2), which combines the reconstruction loss and KL divergence. The training spanned over 50 epochs; to mitigate the risk of overfitting and regularize the model parameters, we applied a weight decay factor of 0.0005. Our model was trained on a PC with an Intel Core i7-8700K processor, 24 GB of RAM, and NVIDIA GeForce RTX 2080 Ti Graphics card.

Once our CVAE model has been trained, we can leverage the encoder and decoder to generate motion sequences for person B based on new motion data from person A. Specifically, we first input c_A for each frame into the encoder, which yields the latent variable z. After that, we use the decoder with z and c_A to produce the synthesized motion x'_B for person B.

5.5 Runtime Algorithms and System

After the above CVAE model has been trained, our runtime algorithms and system work as follows. To facilitate efficient runtime inference of our system, we pre-converted our PyTorch-based CVAE model to the Open Neural Network Exchange(ONNX) format. Subsequently, in the Unreal Engine we harnessed the Neural Network Inference (NNI) plugin to execute model inference efficiently. In order to acquire the VR user's real-time motion as input to our CVAE model, we used the transformations from the VR HMD headset and two motion controllers to generate corresponding transformations for other upper body bones using inverse kinematic techniques. Specifically, we utilized the Forward And Backward Reaching Inverse Kinematics (FABRIK) solver [7]. Meanwhile, we can obtain live eye gaze data of the VR user from the HMD headset [3]. The amalgamation of these two data at each frame serves as the input to our CVAE model during runtime.

STT	LLM Repsonse	TTS	Lip Sync*	CVAE*
0.53	0.6	0.8	5.96	2.43

Table 1: Average running time (in seconds) consumed by core components of our pipeline. Note that LipSync* and CVAE* run in parallel.

To create animations for the virtual agent, we utilize the MetaHuman SDK to generate lip-sync animations based on audio input. In this setup, we configure MetaHuman in the mapping mode to enable the mapping of MetaHuman-compatible lip-sync blend shapes. To ensure real-time efficiency, we segment the response audio, allowing us to obtain the animations in chunks as they are generated. Table 1 shows the average running time consumed by core components of our pipeline. The average time to generate multi-modal response on an off-the-shelf computer was about 6.20 seconds, and the main latency bottleneck is the lip-sync component. As a comparison, the average response delay is about 3.06 seconds according to our acquired real-life human conversation dataset. To synthesize hand gestures, we implemented an encoder-decoder model, following the approach outlined by Kucherenko et al. [29]. Here, the encoder processes the speech signal to derive a set of representational features, and then the decoder utilizes these features to generate corresponding hand gesture motion sequences.

5.6 Comparison with References

We further assessed the performance of our CVAE model by directly comparing it with the reference data in the test dataset. Figure 6 and Figure 7 show the head and eye motion trajectories produced by our CVAE model juxtaposed with the reference data. The two figures highlight that our model successfully captures the fundamental motion patterns, underscoring the capability of latent space learning within our model. Furthermore, not only does our model reflect the overall motion trends, but also the nuances of the generated trajectories also closely align with the reference data. This alignment attests to our model's ability to accurately replicate the dynamics and features of real-world conversational motion.

6 Evaluation: User Study

To evaluate the effectiveness and usefulness of our system, we conducted a comparative user study between our approach and a baseline. In addition to collecting subjective evaluations via questionnaires, we also recorded the eye tracking signals of participants when they wore the HMD headset and interacted with the virtual agent by both our method and the baseline method. Subsequently, we analyzed and compared the objective behavioral measures between the two methods.

Conditions: The following two conditions were used in our comparative study.

- *Our Method*. The proposed algorithms and system in this work were used to automatically generate the conversational behaviors of the virtual agent.
- *Baseline*. In the baseline, torso, arms, and head movements, and eye movements of the virtual agent were generated by the animation graph method in the Unreal Engine, while the lip-sync animation and hand gesture generation modules were the same as those in our approach. This baseline setup served to highlight the distinction in motion synthesis achieved through our CVAE model, thereby enabling a neat evaluation of its effectiveness in enhancing conversational dynamics.

Hypotheses: In this study, we want to statistically test the following hypotheses. (1) H_{fa}: Our method achieves a statistically significantly higher *focused attention* than the baseline method. (2) H_{pu}: Our method achieves a statistically significantly higher *perceived usability* than the baseline method. (3) H_{ae}: Our method achieves a statistically significantly higher *aesthetic appeal and naturalness* than the baseline method. (4) H_{rw}: Our method achieves a statistically significantly higher *rewarding experience* than the baseline method. (5) H_{rb}: Our method achieves a statistically significantly higher *reactive behavior* than the baseline method. (6) H_{ue}: Our method achieves a statistically significantly higher *user engagement* than the baseline method, through the quantitative analysis of objective eye movement signals.

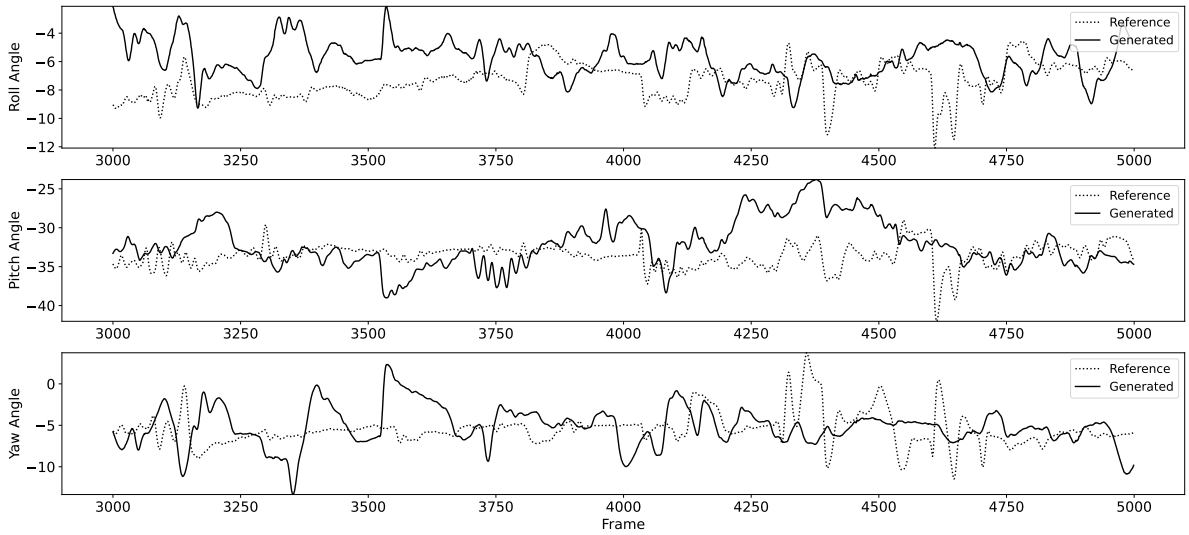

Figure 6: Comparisons between the ground-truth data and the corresponding generated head motion trajectories by our CVAE model.

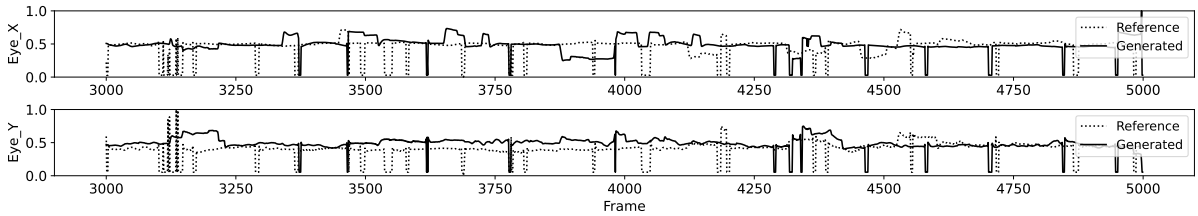

Figure 7: Comparisons between the ground-truth data and the corresponding normalized eye motion trajectories by our CVAE model.

Participants: A total of 15 volunteers were recruited from a university campus to participate in our study. The group consisted of 5 females and 10 males, with an average age of 27.1 years and a standard deviation of 3.0. These participants hailed from a range of academic disciplines, including computer science, biochemistry, physics, and mechanical engineering. Their familiarity with Virtual Reality (VR) devices varied: 8 (53.3%) of them had no prior experience with VR devices, 5 (33.3%) had used VR occasionally in the past, and 2 (13.3%) were regular or professional users of VR. Participants unfamiliar with VR technology were offered a session to acquaint themselves with the VR devices before the study commenced, serving as an introductory warm-up. None of the participants had prior experience in interacting with agents in VR. Importantly, all participants were kept unaware of the study's hypotheses and specific conditions.

Study design: In the study, participants engaged with the system over two distinct sessions, unaware of the specific system in operation for each session. For each participant, the sequence of our method and the baseline method was randomized across these sessions. Each interaction session with the virtual agent spanned approximately 10 minutes. Before starting, participants were guided to focus on the agent's non-verbal cues rather than the conversational content. To ensure uniformity in the conversation across sessions, a structured sequence of questions and potential follow-ups was provided. While they were not required to strictly adhere to the questions provided, they were encouraged to maintain the conversation within the stipulated topic, ensuring a fluid dialogue. The set of dialog questions was also displayed on the VR environment wall for reference, should participants need assistance during the interaction.

During each session, we captured multiple data points including the first-person perspective of the user, the audio of both the participant and the virtual agent, eye-tracking data from the HMD headset [3], and a third-person video recording of the VR user's interaction with the system. Upon completion of each session, participants filled out a questionnaire (the full questionnaire is enclosed in Table 2 in the supplemental document) to provide subjective evaluations and feedback on their experiences and the system. Drawing inspiration from a short form of User Engagement Scale (UES) questionnaire by O'Brien et al. [51], our questionnaire consisted of five categories: focused attention (FA), perceived usability (PA), aesthetic appeal (AE), reward experience (RW), and reactive behavior (RB). Participants responded on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). It should be noted that we did not choose the NASA-TLX questionnaire [21] in our study, because it is designed to mainly assess workload, while the main focus of our study is the user engagement aspect.

Task: Participants were instructed to attentively observe the virtual agent's head, eyes, upper torso, and hand movements, as well as its non-verbal reactive behaviors during their conversation. They were instructed to ensure that the conversational context remained within the specified topic, without any significant deviations. Furthermore, they were encouraged to ask follow-up questions based on the agent's responses adhering to the topic at hand. The primary topic was lunar exploration, encompassing its history, present status, and future possibilities. It was reiterated that the main interest of this study was not the discussion content, but the agent's movements and reactions.

Table 2: Post-experimental questionnaire used in our study

Focused Attention (FA)
FA-1. I was completely absorbed in my interaction with the virtual agent.
FA-2. The agent's non-verbal cues enhanced my engagement in the conversation.
FA-3. I lost track of time due to the realistic nature of the agent's movement
FA-4. I was so involved in the conversation that I blocked out things around me.
FA-5. The agent's body language made me forget I was in a virtual environment
Perceived Usability (PU)
PU-1. I found the multimodal interactions (e.g., speech, eye gaze, nodding, gestures) with the virtual agent easy to use
PU-2. I felt in control while navigating through the multimodal interactions with the virtual agent
PU-3. I was able to complete my tasks successfully with the virtual agent
PU-4. I found the features of the multimodal interactions with the virtual agent to be well-integrated.
PU-5. I found it easy to understand the agent's non-verbal cues (head nodding, eye movement).
PU-6. I did not feel frustrated by the lack of synchronization between my movements and the agent's reactions.
Aesthetic Appeal (AE)
AE-1. The agent's movements were fluid and natural.
AE-2. The agent's facial animations were aesthetically appealing.
AE-3. I found the agent's gestures to be visually pleasing.
AE-4. The agent's body language appealed to my visual senses.
Rewarding Experience (RW)
RW-1. Interacting with the virtual agent was worthwhile.
RW-2. I consider my interaction with the virtual agent a success.
RW-3. The interaction with the virtual agent worked out the way I had planned.
RW-4. I would recommend interacting with the virtual agent system to my family and friends.
RW-5. I continued to interact with the virtual agent out of curiosity.
Reactive Behavior (RB)
RB-1. The virtual agent's eye gaze was responsive to my actions.
RB-2. The virtual agent's head and body movements were synchronized with the conversation.
RB-3. The virtual agent's hand gestures enhanced the communication.
RB-4. The virtual agent mirrored or complimented my actions effectively.
RB-5. The virtual agent's non-verbal cues were understandable and realistic.

6.1 Subjective Measures via Questionnaires

We summarize and plot the aggregated user ratings in Figure 8 by calculating the mean response for each question category listed in Table 2. Participants gave overall higher ratings when our system was adopted as compared to the baseline method: $3.77 \pm 0.85 > 3.42 \pm 0.97$. In addition, this result held for each individual question group: $3.45 \pm 0.82 > 3.09 \pm 0.93$ (FA), $3.81 \pm 0.87 > 3.42 \pm 1.04$ (PU), $3.78 \pm 0.82 > 3.45 \pm 0.86$ (AE), $4.27 \pm 0.72 > 4.11 \pm 0.76$ (RW), and $3.53 \pm 0.77 > 3.01 \pm 0.79$ (RB). By isolating the results for each individual question, we observed that our method performed better than the baseline method across all questions, as illustrated in Figure 9.

Based on the collected user rating data, we also performed statistical tests to verify the aforementioned hypotheses: H_{fa}, H_{pu}, H_{ae}, H_{rw}, and H_{rb}. A one-way repeated measures ANOVA indicated that the agent motion generation method had a statistically significant effect on the user ratings ($F_{1,748} = 28.21, p < .001, \eta^2_{\text{partial}} = 0.036$) and thus rejected the null hypothesis. Similarly, a two-sided independent T-test confirmed the rejection of the null hypothesis ($t_{748} = 5.31, p < .001$). A one-way repeated measures ANOVA breakdown for each individual question group gave: $F_{1,148} = 6.26, p < .05, \eta^2_{\text{partial}} = 0.041$ (FA), $F_{1,178} = 7.31, p < .01, \eta^2_{\text{partial}} = 0.039$ (PU), $F_{1,118} = 4.63, p < .05, \eta^2_{\text{partial}} = 0.038$ (AE), $F_{1,148} = 1.74, p = 0.18, \eta^2_{\text{partial}} = 0.012$ (RW), $F_{1,148} = 16.37, p < .001, \eta^2_{\text{partial}} = 0.100$ (RB). Notably, the superior performance of our method over the baseline method is statistically significant for all individual question groups except for RW (rewarding experience). In other words, our hypotheses H_{fa}, H_{pu}, H_{ae}, and H_{rw} are statistically verified to be true, while for H_{rw} we failed to reject the null hypothesis.

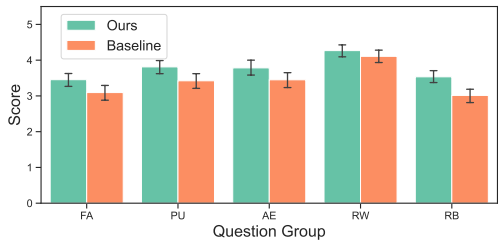

Figure 8: Aggregated user ratings for each question group (capped lines indicate 95% confidence intervals). Notably, our method enabled better user experience than the baseline method in terms of all five categories of subjective evaluations.

6.2 Objective Behavioral Measures

In our user study, a Varjo Aero HMD headset was used to record the eye-tracking data of participants, serving as a robust medium for analyzing user engagement [56]. Based on the recorded eye tracking data, we compute the following objective behavioral measures to compare our method and the baseline method: gaze stability measure, pupil diameter measure, gaze allocation, and mutual gaze.

Gaze stability measure: The Varjo Base software logged gaze data during run-time to capture the stability metric from the recent history of combined gaze ray samples [62]. The travel distances in the gaze ray angles of recent samples are summed up as $\sum \Delta\theta$, where $\Delta\theta$ represents the change in gaze angles between consecutive samples. An interpolation was performed to map this summed value to the stability range $[0, 1]$, with 0 indicating a large traveled distance (poor stability) and 1 indicating a small traveled distance (good stability). The mapping is defined by empirically determined angle constants and historical samples [62]. This stability metric serves as a reflection of the user's focus steadiness, with a score of 1.0 denoting absolute stability and thereby a high level of engagement. Similar methodologies have been used to measure user engagement levels from eye tracking data [31, 52, 57].

We used the above methodology to evaluate both our method and the baseline method in terms of user engagement, measured through the stability metric. Figure 10(a) illustrates the average stability metrics and 95% confidence intervals for a total of 15 participants using both our method and the baseline method. Figure 10(b) compares the aggregated results between our method and the baseline method. As clearly observed from these two figures, our method achieved significantly higher stability metrics, individual-wise or aggregated, than the baseline method.

A one-way ANOVA analysis revealed a significant difference between the two methods, with $F_{1,224029} = 31,591, p < 0.001, \eta^2_{\text{partial}} = 0.000141$, indicating that our method achieved statistically significantly higher stability metrics and thus user engagements than the baseline method.

Pupil diameter measure: We also computed the pupil diameter measure to compare our method and the baseline method. Specifically, we averaged the pupil diameters of both eyes for each participant. Figure 11(a) shows the comparison of the averaged pupil diameters between our method and the baseline method, while Figure 11(b) shows the aggregated comparison. A one-way ANOVA was conducted to examine statistical significance between

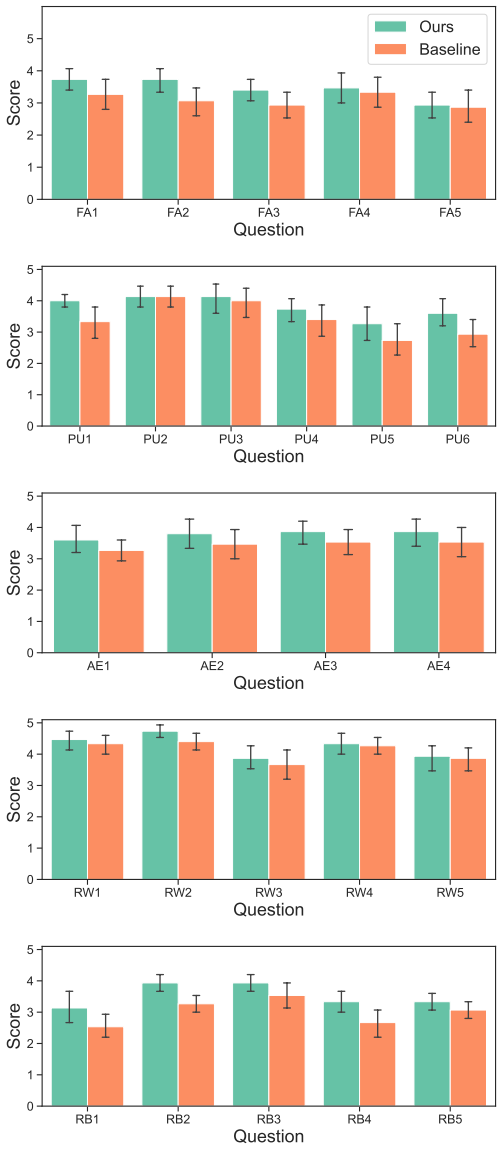

Figure 9: Question-wise user rating breakdown (from top to bottom: focused attention (FA), perceived usability (PU), aesthetic appeal (AE), rewarding experience (RW), and reactive behavior (RB)). Our method achieved higher user ratings than the baseline method for all the questions.

the two methods. This analysis yielded $F_{1,224029} = 16,394, p < 0.001, \eta^2_{\text{partial}} = 0.000073$, elucidating a statistically significant difference between the two methods.

As reported in previous literature [61, 69], the pupil diameters of humans are positively related to their attention in various tasks. For example, in the psychophysiological experimental study by van den Brink et al. [61], they found robust linear relationships between the pupil diameter and several measures of task performance, which suggested that attentional lapses tended to occur when pupil diameter was small. In another recent experimental study by Zijlstra et al. [69], they found that, for humans, a small but significant attentional bias towards dilated pupils, compared to intermediate-sized pupils and intermediate-sized pupils when compared to small pupils. Therefore, as mentioned above, the pupil diameters of participants using our

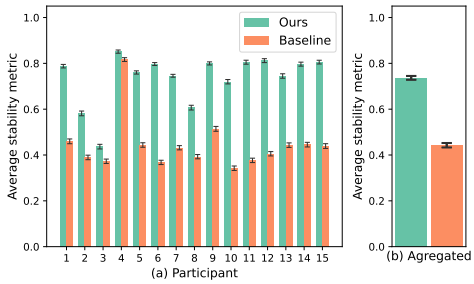

Figure 10: (a) The average stability metrics (capped lines indicate 95% confidence intervals) of individual participants in our study between our method and the baseline method. (b) Aggregated average stability metrics between our method and the baseline method

system are statistically significantly larger than the pupil diameters of the same participants using the baseline system. These results provide statistically-grounded objective measures to support that participants pay significantly more attention to the virtual agent and thus have more user engagement when using our system, compared to the baseline system.

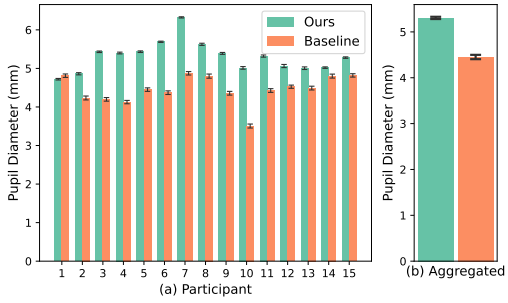

Figure 11: (a) The average pupil diameters (capped lines indicate 95% confidence intervals) of individual participants in our study between our method and the baseline method. (b) Aggregated average pupil diameters between our method and the baseline method.

Gaze allocation: We also collected gaze allocation information of participants during the user study. Specifically, according to the eye gaze at each frame (real-time tracked by the VR headset), we real-time compute which triangle in the virtual agent mesh intersects with the participant's gaze and store the triangle index. Then, we assume that the participant assigns the gaze to this specific triangle at that moment. Finally, we visualize the gaze allocation counts of the triangles in the virtual agent mesh in Figure 12. As clearly shown in this figure, when participants interacted with the virtual agent using our method, they allocated more gazes to the face of the virtual agent than using the baseline method. Based on previous research studies on gaze allocation in face-to-face human communication [23], a person P_A generally assigns more gaze to the face of the partner P_B in dyadic conversations when P_B pays more attention to P_A (e.g., eye contact) during communication. Extending this finding from face-to-face human communication to human-agent interaction in this work, the comparative result in Figure 12 implies that, during human-agent interaction, the automated virtual agent by our method pays more attention to participants than the virtual agent by the baseline method; therefore, when our method is used, the participants assign more of their gaze to the face of the virtual agent.

Eye contact (mutual gaze): Utilizing the VR user's real-time tracked gaze and the virtual agent's animated gaze, we are able to determine the proportion of eye contact (mutual gaze) shared be-

tween the VR user and the virtual agent throughout their interaction. Additionally, we calculated both the proportion of time the VR user spent looking at the virtual agent and the proportion of time the virtual agent spent focusing on the VR user. Specifically, we created a 3D bounding sphere for the head of the VR user (or the virtual agent). When the gaze of P_A intersects the bounding sphere of P_B at time t, we consider that P_A looks at P_B. Mutual gaze means that P_A and P_B look at each other at that moment. Figure 13 illustrates the comparative results between our method and the baseline method. From this figure, we can see that the virtual agent generated by our method had substantially more mutual gaze (27.0%) than the virtual agent generated by the baseline method (16.4%). In terms of the proportion of time one side spent looking at the other side, our method also achieved higher proportions than the baseline method.

In sum, the aforementioned hypothesis H_{ue} was statistically verified to be true. Our findings on objective behavioral measures provide grounded evidence that our method can improve user engagement, compared to the traditional baseline method, which also lays a promising first step for further exploration of optimizing methods for better user engagement and attention.

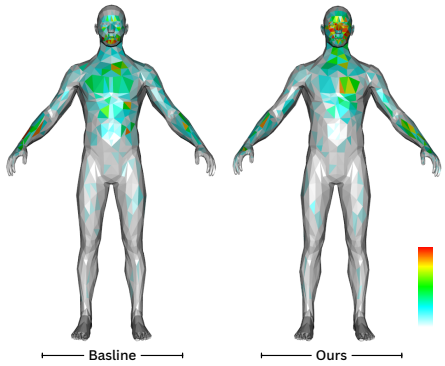

Figure 12: Gaze allocation visualization of our user study.

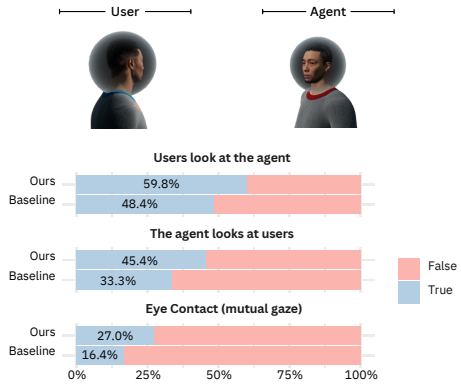

Figure 13: Comparative results of eye contact and gazes between our method and the baseline method in our user study.

7 Discussion, Limitations and Future Work

Our experimental results validate that our method consistently outperformed the baseline method in terms of selected user factors on both subjective and objective measures. Participants provided higher subjective ratings for focused attention, perceived usability, aesthetic appeal, rewarding experience, and reactive behavior when using our method compared to the baseline. Objective eye tracking data corroborated these findings, showing greater gaze stability, larger pupil diameters (indicative of greater attention), more focused gaze allocation on the virtual agent's face, and longer mutual gaze duration. These improvements suggest enhanced user engagement and interaction quality through more realistic agent behaviors. It is crucial to highlight that this superior performance was not due to aesthetically superior character assets or the introduction of more interaction modalities. Both our method and the baseline method utilized identical interaction modalities, 3D character models, and algorithms for generating text responses, hand gestures, and lip-sync animations. The sole differentiation was the method employed for dynamic motion generation of the torso, arms, head, and eyes. In essence, our approach, when juxtaposed with traditional animation graph techniques, proves its merit in generating more interactive and responsive behaviors for intelligent virtual agents.

While our user study provides compelling evidence of our method's benefits, the relatively small sample size may limit its generalizability. Future studies could benefit from a larger and more diverse pool to validate these results across different demographics. The primary focus of the current study was to investigate the capabilities of reactive motion generation in virtual agents, excluding the acoustic audio features and the semantic content derived from speech audio. Despite these omissions, we have laid a robust foundational framework that can be considerably augmented by further harnessing the available audio data. The potential integration of these audio features is promising in improving reactive motion generation, making virtual agents more lively, responsive, and engaging during interactions. Moreover, our system's motion synthesis model uses motion signals rather than language features, allowing for support of multiple languages. Additionally, the integrated large language model can generate responses in various languages.

Nevertheless, the scope of conversational motions in our current study is somewhat confined, since our dataset is centered primarily on stand-up conversations. We avoided the data capture of seated conversations due to potential occlusion issues with the seating equipment. While these are prevalent, they only capture a small segment of possible interaction scenarios. By incorporating a diverse range of dyadic conversation settings, such as seated dialogues, walking discussion, or exchanges within a vehicular context, we could achieve a more holistic grasp and modeling of reactive motion in virtual agents. Although leveraging a large, diverse datasets and latest generative models can enhance the current motion quality, solving the problem of real-time efficiency is crucial for the integration into end-to-end conversation systems. Future directions of research could involve modeling multi-party communication behaviors using this foundational framework.

8 Conclusion

We introduce a learning-based framework designed to generate multimodal agent behaviors that are acutely aware of user actions. Specifically, we employ a CVAE-based neural network model that perceives VR-tracked user activities in real-time, subsequently guiding the virtual agents' responses, spanning both verbal and non-verbal cues such as head/eye movements and body language. Our high-fidelity social communication dataset enables the model to mimic human interpersonal communications. We hope this research establishes the critical step toward human-mimetic virtual agents as a promising interface that bridges VR users with the rapidly advancing AI technologies such as LLM.

Acknowledgments

This work is supported in part by NSF grants 2005430, 2225861 and 2232817.

REFERENCES

[1] Audio To Lip sync — MetaHumanSDK — docs.metahumansdk.io. https://docs.metahumansdk.io/metahuman-sdk/reference/metahumansdk-unreal-engine-plugin/audio-to-lip-sync. [Accessed 25-05-2024]. 3

[2] Controller — VIVE United States — vive.com. https://www.vive.com/us/accessory/controller/. [Accessed 26-05-2024]. 2

[3] Varjo Aero — varjo.com. https://varjo.com/products/aero/. [Accessed 25-05-2024]. 2, 4, 5

[4] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 2

[5] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022. 2

[6] T. Ao, Z. Zhang, and L. Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv preprint arXiv:2303.14613*, 2023. 2

[7] A. Aristidou and J. Lasenby. Fabrik: A fast, iterative solver for the inverse kinematics problem. *Graphical Models*, 73(5):243–260, 2011. 4

[8] S. Aseeri and V. Interrante. The influence of avatar representation on interpersonal communication in virtual social environments. *IEEE transactions on visualization and computer graphics*, 27(5):2608–2617, 2021. 2

[9] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *Proc. of 2021 IEEE virtual reality and 3D user interfaces (VR)*, pp. 1–10. IEEE, 2021. 2

[10] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1075–1086, 2007. 2

[11] C. Busso, Z. Deng, U. Neumann, and S. Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation and Virtual Worlds*, 16(3-4):283–290, 2005. 2

[12] E. Chuang and C. Bregler. Mood swings: expressive speech animation. *ACM Transactions on Graphics (TOG)*, 24(2):331–347, 2005. 2

[13] J. W. Crandall and M. A. Goodrich. Characterizing efficiency of human robot interaction: A case study of shared-control teleoperation. In *Proc. of IEEE/RSJ international conference on intelligent robots and systems 2002*, vol. 2, pp. 1290–1295. IEEE, 2002. 2

[14] Z. Deng, J. P. Lewis, and U. Neumann. Automated eye motion using texture synthesis. *IEEE Computer Graphics and Applications*, 25(2):24–30, 2005. 2

[15] S. Dermouche and C. Pelachaud. Generative model of agent's behaviors in human-agent interaction. In *Proc. of 2019 ACM International Conference on Multimodal Interaction*, pp. 375–384, 2019. 2

[16] Epic Games. Unreal engine. 2

[17] A. Fukayama, T. Ohno, N. Mukawa, M. Sawaki, and N. Hagita. Messages embedded in gaze of interface agents — impression management with agent's gaze. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 41–48. ACM, 2002. 2

[18] R. R. Galin and R. V. Meshcheryakov. Human-robot interaction efficiency and human-robot collaboration. In *Robotics: Industry 4.0 issues & new intelligent control paradigms*, pp. 55–63. Springer, 2020. 2

[19] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency. Virtual rapport. In *Proc. of IVA'06*, vol. 6, pp. 14–27. Springer, 2006. 2

[20] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527, 2011. 2

[21] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908. Sage publications, 2006. 5

[22] P. Heidicker, E. Langbehn, and F. Steinicke. Influence of avatar appearance on presence in social VR. In *Proc. of 2017 IEEE symposium on 3D user interfaces (3DUI)*, pp. 233–234. IEEE, 2017. 2

[23] R. S. Hessels, G. A. Holleman, A. Kingstone, I. T. Hooge, and C. Kemner. Gaze allocation in face-to-face communication is affected primarily by task structure and social context, not stimulus-driven factors. *Cognition*, 184:28–43, 2019. 7

[24] Y. Huang and S. M. Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11–18, 2017. 2

[25] A. Jin, Q. Deng, Y. Zhang, and Z. Deng. A deep learning-based model for head and eye motion generation in three-party conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2(2):1–19, 2019. 2

[26] S. C. Khullar and N. I. Badler. Where to look? automating attending behaviors of virtual human characters. *Autonomous Agents and Multi-Agent Systems*, 4(1):9–23, 2001. 2

[27] D. P. Kingma, M. Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 3

[28] T. Koda and T. Ishida. Cross-cultural study of avatar expression interpretations. In *Proc. of International Symposium on Applications and the Internet (SAINT'06)*, pp. 7–pp. IEEE, 2006. 2

[29] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104, 2019. 3, 4

[30] C. Kyrlitsias and D. Michael-Grigoriou. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality*, 2:786665, 2022. 1, 2

[31] M. Lalmas, H. O'Brien, and E. Yom-Tov. *Measuring user engagement*. Springer Nature, 2022. 6

[32] B. J. Lance and S. C. Marsella. A model of gaze for the purpose of emotional expression in virtual embodied agents. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 199–206, 2008. 2

[33] B. Le, X. Ma, and Z. Deng. Live speech driven head-and-eye motion generators. *IEEE Transactions on Visualization and Computer Graphics*, 18(11):1902–1914, Nov 2012. 2

[34] J. Lee and K. H. Lee. Precomputing avatar behavior from human motion data. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 79–87, 2004. 2

[35] J. Lee and S. Marsella. Learning a model of speaker head nods using gesture corpora. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pp. 289–296, 2009. 2

[36] M.-C. Lee, M. Trinh, and Z. Deng. Multimodal turn analysis and prediction for multi-party conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 436–444, 2023. 3

[37] S. P. Lee, J. B. Badler, and N. I. Badler. Eyes alive. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pp. 637–644. ACM, 2002. 2

[38] J. C. Lester, S. G. Towns, C. B. Callaway, J. L. Voerman, and P. J. FitzGerald. Embodied conversational agents. chap. Deictic and Emotive Communication in Animated Pedagogical Agents, pp. 123–154. MIT Press, Cambridge, MA, USA, 2000. 2

[39] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. In *Acm siggraph 2010 papers*, pp. 1–11. 2010. 2

[40] Y. Liang, Q. Feng, L. Zhu, L. Hu, P. Pan, and Y. Yang. Seeg: Semantic energized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10473–10482, 2022. 2

[41] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10462–10472, 2022. 2

[42] M. Luck and R. Aylett. Applying artificial intelligence to virtual reality: Intelligent virtual environments. *Applied artificial intelligence*, 14(1):3–32, 2000. 2

[43] X. Ma and Z. Deng. Natural eye motion synthesis by modeling gaze-

head coupling. In *Proc. of 2009 IEEE Virtual Reality Conference*, pp. 143–150. IEEE, 2009. 2

[44] R. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In *International Workshop on Intelligent Virtual Agents*, pp. 25–36. Springer, 2005. 2

[45] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 25–35, 2013. 2

[46] S. Masuko and J. Hoshino. Generating head–eye movement for virtual actor. *Systems and Computers in Japan*, 37(12):33–44, 2006. 2

[47] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions On Graphics (TOG)*, 27(1):1–24, 2008. 2

[48] B. Nojavanasghari, Y. Huang, and S. Khan. Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv:1801.09092*, 2018. 2

[49] N. Ogawa, T. Narumi, and M. Hirose. Effect of avatar appearance on detection thresholds for remapped hand movements. *IEEE transactions on visualization and computer graphics*, 27(7):3182–3197, 2020. 2

[50] C. S. Oh, J. N. Bailenson, and G. F. Welch. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5:409295, 2018. 2

[51] H. L. O'Brien, P. Cairns, and M. Hall. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112:28–39, 2018. 5

[52] T. Renshaw, R. Stevens, and P. D. Denton. Towards understanding engagement in games: an eye-tracking study. *On the Horizon*, 17(4):408–420, 2009. 6

[53] D. Roth, J.-L. Lugrin, D. Galakhov, A. Hofmann, G. Bente, M. E. Latoschik, and A. Fuhrmann. Avatar realism and social interaction quality in virtual reality. In *Proc. of 2016 IEEE virtual reality (VR)*, pp. 277–278. IEEE, 2016. 2

[54] K. Ruhland, S. Andrist, J. Badler, C. Peters, N. Badler, M. Gleicher, B. Mutlu, and R. Mcdonnell. Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In *Eurographics State-of-the-Art Report*, pp. 69–91, 2014. 2

[55] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1330–1345, 2008. 2

[56] C. Shagass, R. A. Roemer, and M. Amadeo. Eye-tracking performance and engagement of attention. *Archives of General Psychiatry*, 33(1):121–125, 1976. 6

[57] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, and V. M. de Jesus Filipe. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In *Proc. of International Conference on Technology and Innovation in Learning, Teaching and Education*, pp. 52–68. Springer, 2022. 6

[58] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 3

[59] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513, 2004. 2

[60] K. R. Thórisson. *Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action*, pp. 173–207. Springer Netherlands, Dordrecht, 2002. 2

[61] R. L. Van Den Brink, P. R. Murphy, and S. Nieuwenhuis. Pupil diameter tracks lapses of attention. *PLoS One*, 11(10):e0165274, 2016. 7

[62] Varjo. Get started: Eye tracking with varjo headset, 2023. Accessed: Sep. 29, 2023. 6

[63] J. Ventrella. Avatar physics and genetics. In *Virtual Worlds: Second International Conference, VW 2000 Paris, France, July 5–7, 2000 Proceedings 2*, pp. 107–118. Springer, 2000. 2

[64] A. M. Von der Pütten, N. C. Krämer, J. Gratch, and S.-H. Kang. "it doesn't matter what you are!" explaining social effects of agents and avatars. *Computers in Human Behavior*, 26(6):1641–1650, 2010. 2

[65] A. S. Williams, J. Garcia, and F. Ortega. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3479–3489, 2020. 2

[66] R. E. Yagoda and D. J. Gillan. You want me to trust a robot? the development of a human–robot interaction trust scale. *International Journal of Social Robotics*, 4:235–248, 2012. 2

[67] S. Ye, Y.-H. Wen, Y. Sun, Y. He, Z. Zhang, Y. Wang, W. He, and Y.-J. Liu. Audio-driven stylized gesture generation with flow-based model. In *Proc. of European Conference on Computer Vision (ECCV) 2022*, pp. 712–728. Springer, 2022. 2

[68] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 469–480, 2023. 2

[69] T. Zijlstra, E. van Berlo, and M. Kret. Attention towards pupil size in humans and bonobos (pan paniscus). *Affective Science*, 3(4):761–771, 2022. 7