



Predicting and Understanding Turn-Taking Behavior in Open-Ended Group Activities in Virtual Reality

PORTIA WANG, Stanford University, USA

EUGY HAN, Stanford University, USA

ANNA C.M. QUEIROZ, Stanford University, USA

CYAN DEVEAUX, Stanford University, USA

JEREMY N. BAIENSON, Stanford University, USA

In networked virtual reality (VR), user behaviors, individual differences, and group dynamics can serve as important signals for future speech behaviors, such as who the next speaker will be and the timing of turn-taking behaviors. The ability to predict and understand these behaviors offers opportunities to provide adaptive and personalized assistance, for example helping users with varying sensory abilities navigate complex social scenes and instantiating virtual moderators with natural behaviors. In this work, we predict turn-taking behaviors using features extracted based on social dynamics literature. We discuss results from a large-scale VR classroom dataset consisting of 77 sessions and 1660 minutes of small-group social interactions collected over four weeks. In our evaluation, gradient boosting classifiers achieved the best performance, with accuracies of 0.71–0.78 AUC (area under the ROC curve) across three tasks concerning the “what”, “who”, and “when” of turn-taking behaviors. In interpreting these models, we found that group size, listener personality, speech-related behavior (e.g., time elapsed since the listener’s last speech event), group visual attention (e.g., the group’s head orientation towards the speaker), and the listener and previous speaker’s head pitch, head y-axis position, and left hand y-axis position more saliently influenced predictions. Results suggested that these features remain reliable indicators in novel social VR settings, as prediction performance is robust over time and with groups and activities not used in the training dataset. We discuss theoretical and practical implications of the work.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Computer-Mediated Communication, Social Interaction, Turn-Taking Behavior, Virtual Reality

ACM Reference Format:

Portia Wang, Eugy Han, Anna C.M. Queiroz, Cyan DeVeaux, and Jeremy N. Bailenson. 2025. Predicting and Understanding Turn-Taking Behavior in Open-Ended Group Activities in Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW317 (November 2025), 40 pages. <https://doi.org/10.1145/3757498>

Authors’ Contact Information: Portia Wang, Stanford University, Stanford, California, USA, portia@stanford.edu; Eugy Han, Stanford University, Stanford, California, USA; Anna C.M. Queiroz, Stanford University, Stanford, California, USA; Cyan DeVeaux, Stanford University, Stanford, California, USA; Jeremy N. Bailenson, Stanford University, Stanford, California, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2573-0142/2025/11-ARTCSCW317

<https://doi.org/10.1145/3757498>

1 Introduction

Effective and organized social interactions require its social actors to accurately infer speaking intention and take turns accordingly. In most platforms for facilitating computer mediated communication (CMC), users retain a high level of autonomy over when and how they speak, utilizing features such as the muting and unmuting button during videoconferencing. However, along with the benefits of user control comes the responsibility for individuals to correctly infer social group dynamics. Such inferences can be difficult to navigate as users utilize digital platforms such as video conferencing and social virtual reality (VR), where the level of immersion and representation of users (e.g., nonverbal behavior, avatar appearance) can differ from one another [1, 103]. In immersive social interactions, differences in avatar representation and characteristics of social VR platforms can yield adverse outcomes such as unwanted overlap in speech [9, 115], lack of control for users with physical disabilities [68], misperceived and absence of nonverbal cues [68, 76], and difficulty gauging the end of speaking turns [76, 115]. For users with visual and audio disabilities, it can also be challenging for them to engage in seamless interactions based solely off of perceptual and auditory feedback [4, 56]. Paradoxically, as immersive communication evolves to be hardware-agnostic, with users interacting across traditional screen devices and mixed-reality headsets [87], the mismatch in social cues delivered and perceived by users across different devices present further complications for users to navigate socially.

Being able to predict user intentions for actions such as turn taking can alleviate these challenges through real-time interventions and training. Social VR platforms can implement personalized assistance for users who are struggling to interpret nonverbal cues under technological constraints (e.g., joining large meetings though audio feed) or are different in physical abilities (e.g., auditory and visual abilities). This assistance can also benefit populations with medical conditions such as autism and ADHD that inhibit their ability to accurately interpret nonverbal behaviors [26, 33]. Relatedly, understanding *what* behavior precedes turn taking can be useful in training practitioners such as educators and healthcare workers to better navigating social settings. For example, by training early-career practitioners to recognize nonverbal cues and individually characteristics influencing speaking intentions, they can be better equipped to navigate social situations.

In more structured settings such as classrooms and focus groups, knowing who will likely speak next and what nonverbal cues precede speech can improve social dynamics. For one, practitioners who need to monitor multiple virtual sessions simultaneously, inevitably missing nonverbal and verbal cues, can leverage turn-taking predictions to remain informed of the social dynamics and better facilitate group conversations. These practitioners can also instantiate virtual agents to moderate social VR scenes with natural nonverbal behaviors that do not cut off current speakers and more broadly design mechanisms for smoothly guiding the conversations. Failure to accurately infer social dynamics can yield unwanted interruptions and lower the perceived ease of use of these intelligent systems [67]. As such, we argue that it is important to study whether we can *robustly* predict virtual turn-taking behaviors and how nonverbal and verbal behaviors as well as individual and group characteristics influence turn-taking predictions.

In this work, we investigated turn-taking behaviors in VR and leveraged the medium's fine-grain tracking data on motion and verbal behaviors. Using the VR motion data, researchers have implemented pipelines for deriving features such as head movement [70] and gaze [1], both of which have been previously useful for predicting speech behaviors in face-to-face social interactions [44, 45]. As such, our first research question considers whether it is possible to predict turn taking.

- Research Question 1 (**RQ1**): Can we predict turn-taking behaviors in VR open-ended group activities from features extracted from motion tracking data, speech-related behavior, and individual and group differences?

Building on this, we leveraged VR to place individuals into varying virtual spaces and studied social dynamics across groups, social activities, and spatial contexts over time. Studying diverse social interactions at scale is important as building models that are robust to unseen groups, activities, and time suggests higher practicality and reliability. Concretely, our second research question concerns generalizability.

- Research Question 2 (**RQ2**): How is turn-taking behavior prediction performance affected when evaluated on groups, activities, and time not seen in training?

Finally, we interpreted the predictive models. Analyzing how features predict turn taking can deepen our understanding of virtual social dynamics, allowing future researchers to benchmark against face-to-face interactions. This understanding can also help practitioners build more efficient models with curated features and compose social groups exhibiting certain turn-taking behaviors.

- Research Question 3 (**RQ3**): How are the extracted features related to turn-taking predictions and performance? Specifically, what features are strongly related to performance, and how are they associated with turn-taking predictions?

To answer these questions, we built models predicting speech dynamics, focusing on three tasks that capture the “what”, “who”, and “when” of turn-taking behaviors. We formulated features based on tracking data and individual and group characteristics based on social dynamics literature. We evaluated performance across 77 sessions and 1660 minutes of open-ended VR group discussions spanning four weeks, each with three to four university students. Notably, we found that gradient boosting classifiers achieved the best accuracies, predicting turn transitions and identifying the new speaker with 0.75–0.78 AUC (i.e., area under the ROC curve), and differentiating between moments immediately preceding turn transitions and those sampled at prior moments with accuracies of 0.71–0.72 AUC. These results demonstrated that we were able to predict turn-taking behaviors with accuracies considerably higher than prediction by chance (i.e., 0.50 AUC) and highlighted the potential non-linear relationships and interactions between extracted features.

Additional feature analyses revealed the importance of listener personality, group size, speech-related behavior (e.g., preceding speaker sequence), group visual attention, head pitch and y-axis position, and left hand y-axis position. Our models showed comparable results when evaluated on time, activities, and groups not seen during training, which demonstrated the reliability of these turn-taking indicators in novel social settings. From these results, we highlight their theoretical relevance and outline practical implications for how practitioners can leverage behavioral predictions to facilitate effective VR social interactions.

In summary, we make the following contributions. First, we formulated features based on social dynamics literature and leveraged both the VR tracking data and individual and group characteristics. Using a large-scale VR dataset with 1660 minutes of open-ended activities collected over four weeks, we then demonstrated the feasibility and robustness of predicting VR turn-taking behaviors. Finally, through interpreting how features influence model performance and prediction probabilities, we present theoretical and practical insights and highlight the potential of our work for intervention, support, and training.

2 Related Work

2.1 Nonverbal Behavior in Social Interactions

Nonverbal behavior offers insights into social dynamics. Gaze, for example, is related to action patterns [53] and attention [57, 97], and provides important signals to understanding dyadic collaborations [5, 48], user intention and coordination [3], learning outcomes [86], conversational attention [108], and social engagement [79]. Gesture and body orientation can influence the sense

of feeling addressed by conversation partners [77], while proxemics, body orientation, and gaze, can be effective in predicting one's intention of joining social groups [12].

It is then unsurprising that nonverbal communication makes up a key component of virtual interactions, for which social VR offers additional benefits compared to video conferencing. By tracking the users' eyes, heads, and hands, VR systems render avatars that accurately convey nonverbal cues such as interpersonal distance and gaze, fostering social interactions that are perceived positively and comparable to face-to-face interactions [68, 100]. Communication through nonverbal behaviors such as bodily and facial gestures induce more positive dyadic interactions in VR, with these cues also being predictive of interpersonal attraction [83]. Using motion data, VR social interactions can also be augmented, for example through increasing mimicry [8, 89, 106] and transforming gaze and interpersonal distance [91, 92, 111, 113]. With VR placing users in controllable virtual environments, it is also possible to examine nonverbal behavior longitudinally [70, 71]. Works highlighted how nonverbal behavior changes over time – users looked at others more and their interpersonal distance increased [70], and user identifiability lowered with greater temporal delay between training and testing sessions [71].

VR tracking data also enables fine-grain analyses on synchrony [73, 105], self-efficacy and learning [88], physiological responses [66], classroom discourse [102], interpersonal distance [21, 70], context [35], design behaviors [112], and user identification [71, 72, 74, 75, 78]. Particularly relevant to our work is the research of DeVeaux et al. [21]. The authors extracted linguistic patterns from transcripts and notably found a positive correlation between the use of impersonal pronouns and the median interpersonal distance. This research highlighted the importance of considering language use to study VR affordances and the potential of leveraging nonverbal behavior to uncover psychological nuances that surveys fail to capture. Key distinctions between their work and ours are their focus on the nonverbal behavior of interpersonal distance, linguistic styles, and session-level analyses. In contrast, we studied turn-taking behaviors and focused on predicting them at moments within sessions. Beyond interpersonal distance, we also included other nonverbal features (e.g., egocentric motion). One other distinction lies in the VR activities analyzed: while DeVeaux et al. [21] studied instructor-led discussions collected during a university course in Fall 2021, we focused on open-ended activities recorded a year later through the same course.

We contribute to past works by predict turn-taking behaviors through VR motion data. Findings on changes in nonverbal behaviors over time [70, 71] emphasized the need to study social behaviors through a longitudinal lens and across multiple sessions. These insights motivated our examination of turn-taking predictions across unseen weeks, groups, and activities.

2.2 Modeling Human Behavior and Individual- and Group- Level Differences using Tracking Data

Besides using tracking data to understand human behaviors, monitoring user motion, often unobtrusively, can further enable interventions and assessments. For example, as smartphone data such as touch and typing behaviors [55, 109] and audio, textual, and video data [52] are predictive of personality traits and affective states, one can construct personalized experiences based on the user's current behaviors. Another example involves monitoring gaze to assess cognitive load, and using this information to facilitate adaptations of mixed reality interfaces [62]. Recently, researchers also built models for predicting team viability using online team text conversations and demonstrated the potential of using automated features for assessments and intervention [16]. Others proposed using individual and social behaviors to model purchase decisions for e-commerce recommendations [118] and job burnouts for early-stage interventions [116].

Similarly, we see opportunities in predicting VR speech behaviors for intervention and assistance. Fortunately, though humans can struggle to infer speaker behavior [39], research that leverages

tracking data for speech behavior predictions has shown greater promise. Works looked to speech sequences [85], mouth and head motion [43, 44], respiration [46], gaze behavior [45, 58], and more broadly user motion [17] as predictors when modeling speech behavior. Notably, these works have found that markov models were effective in predicting the next speaker based on the two preceding speakers [85] and that support vector machines leveraging features on gaze transition patterns are predictive of turn-changing characteristics [45].

Unlike the social scenes examined in previous studies, where users completed a single or series of similar tasks [17, 45] and either stood in fixed positions [58] or remained seated [17, 45], VR social interactions are diverse in activities and typically involve less constraints on movement in the physical and virtual spaces. This raises the question of whether turn-taking predictions are possible when there are greater variations in activities and nonverbal cues. Failure to make robust predictions in these scenarios limits the prospects of intervention and assistance in VR. To address this, we investigated VR turn-taking behaviors from four weeks of open-ended activities with little restriction on virtual motion. We contribute by also interpreting how features are related to turn-taking predictions, highlighting their theoretical and practical significance.

2.3 Social Interactions across Individuals and Groups

Tracking data's ability to predict individual differences such as personalities and affective states [47, 52, 55, 109] suggests that individual differences can be used to predict user behaviors. Individual differences such as gender, age, and personalities are related to how people take up physical and virtual spaces [27, 41, 42, 80], while personality, public speaking anxiety, and immersive tendencies of individuals are predictive of user experience and perceived quality of interactive systems [18].

Personality traits, in particular, are related to speech behaviors. For example, conscientiousness and extraversion are related to basic speech features such as pitch estimate [47]. Extraversion and neuroticism are also associated with different speech patterns, notably with introverts exhibiting longer silences between utterances [90]. Introverts and extroverts were also found to speak at different levels of language abstraction [10] and differ in reaction time when verbally responding to prompts and questionnaires [59, 84]. Broadly speaking, extraversion, agreeableness, and neuroticism also correlate with features related to speaking turns, speaking length, and average speaking turn duration [6, 60].

Broadly speaking, social interactions can vary depending on the characteristics of social groups. For example, the association between trait dominance—measured as expressed control and prosocial interpersonal power and influence—and speaking time is influenced by the composition of social groups, namely whether the group was composed randomly or of extremes (i.e., pairing together the most and least dominant individuals) [69]. Comparing conversations between dyad and triads, the presence of an additional listener led to individuals speaking louder in situations with high noise levels, and listeners orienting their heads more optimally (i.e., rotating their heads to approximately 30 degrees from the speaker) [32].

We rely on past findings to identify features related to personality and group compositions for modeling turn taking. By studying how these features are related to social behaviors such as speaking intentions, we explore how individual and group characteristics can help build immersive tools for facilitating social interactions.

3 Methods

3.1 Recordings of Open-ended Group Social Interactions in VR

We studied the longitudinal VR classroom dataset collected by Han et al. [34], and focused on the subset collected in Fall 2022 as part of a university course on VR. In it, 146 university students, out of



Fig. 1. Screenshots of the Open-Ended Group Activities. Students donned VR gear in remote physical locations but joined together in different virtual environments for weekly group discussions and participated in activities related to topics on accessibility, avatars, medical, and education.

whom 117 consented, engaged in weekly discussions and responded to open-ended design prompts in groups of two to four for four weeks. Given our interest in turn taking, we only analyzed VR sessions with all consenting participants¹. The topics of discussion and activities changed weekly, with themes such as accessibility and education. We detail session activities in Appendix A. Each week, students gathered in the same groups and used the social VR platform ENGAGE while being physically remote in their own private spaces. Each student used a Meta Quest 2 headset and two hand-held controllers to partake in the sessions and was allowed to move around virtually using smooth translation and teleportation. No teaching staff was present in these sessions, so student groups recorded the discussion sections using the platform's recording feature. Figure 1 shows screenshots from the recorded group activities.

The authors varied the virtual environments (i.e., ceiling height, amount of visible space) to examine how context influences attitude, nonverbal behavior, and design behavior [35, 112]. This variation in context allows us to model interactions occurring in diverse virtual spaces, setting it apart from literature that examined speech behavior in the same physical or virtual environment [17, 58]. This particular dataset was also well-suited for our purpose given its open-ended activities, which fostered more natural interactions compared to the instructor-led sections in the rest of the dataset. While previous research have examined language use using the data collected in Fall 2021 [21, 22], we are the first to use the Fall 2022 data to examine turn-taking behaviors of open-ended group activities.

Given our interest in group interactions, we filtered out moments and sessions with two students. For predicting turn-taking behaviors, looking at groups larger than two also makes the problem nontrivial. Our final dataset consisted of 77 VR sessions collected from 26 groups and 100 unique students. Out of the 77 sessions, 35 were three-person discussions, and 42 were of four people.

¹A research personnel not part of the teaching staff randomly assigned students into groups and maximized the number of groups with all consenting students. This ensured that students received comparable learning experience as the teaching staff was blinded to the consenting status, while also allowed researchers to examine fully-consenting group interactions.

During data collection, some participants missed discussions, leading to a high number of three-person discussions. Some sessions were dropped due to technical difficulties such as software updates and incomplete recordings. The dataset aggregated a total of 1660.60 minutes of tracking data, with each session taking on average 21.57 minutes ($SD=6.44$).

The recording files collected several forms of user and session data, namely that of the scene, user motion, and audio. Since we wanted to extrapolate turn-taking behavior insights that are generalizable across social platforms and social settings, we focused on encoding information related to user motion and audio, which is logged at 30 Hz. Motion data consisted of the position (i.e., x, y, z) and orientation (i.e., roll, pitch, yaw) of the user's headset, two controllers, and the "root". The "root" tracks the user's global position and orientation within the virtual environment, while the headset and controllers are tracked within the coordinate system dictated by the "root". From the tracked motion, we can derive information regarding a user's egocentric behavior as well as those related to other users. The audio information is recorded both through audio files and a floating-point value between 0 and 1 representing volume.

3.2 Categories of Turn-Transition Behaviors

3.2.1 Pre-processing Speech Events. We began by identifying all speech events by labeling Inter-Pausal-Units (IPU) [54], defined as the stretches of speech activity by a single speaker. Using the audio tracking data, we determined users as actively speaking when their speech volume is greater than 0.1. To reduce noise from short pauses between speaking activities, we joined adjacent speech events of the same user when the gap between them is within 0.5s. The threshold of 0.5s follows prior literature that found the average pause duration during read speech, interviews and public presentations to be between 0.38 to 0.53 seconds [15, 63].

For each session recording, we proceeded to determine the main speaker by first labeling users who are the sole speaker during speech events as the main speaker. When there are multiple speakers, we implemented the following labeling scheme. To start, we eliminated speech events that are completely overlapped by another speech event. Then, we assign the main speaker as the speaker who ends their speech event last. The start time of the new main speaker is marked as the moment the previous main speaker finished speaking. We used Python for extracting IPUs and assigning main speakers.

3.2.2 Defining and Labeling Turn-transition Behaviors. From the labeled main speakers, we extracted four types of turn-transition behaviors following closely the categorizations used by Jokinen et al. [49]: clean turn taking, overlap turn taking, backchanneling, and continuing speech. These categories provide a formal framework for us to examine turn-taking behaviors. Upon obtaining the labels, we removed data points associated with speech events shorter than 323 milliseconds (i.e., the average duration for enunciating fast words found in [101]) to filter out noise. We used Python for processing the audio input. Figure 2 shows categorization of each of the categories based on an example audio input. Table 1 shows the categories' summary statistics.²

- **Clean Turn Taking.** Clean turn taking occurs whenever the main speaker has changed and that there was no overlapping between the previous and new speaker turns. The start of a clean turn taking event was marked as the beginning of the new speaker's speech event. Figure 2b-1 shows examples of these behaviors.
- **Overlap Turn Taking.** As shown in Figure 2b-2, overlap turn taking occurs whenever the main speaker changes and the two speech events overlap. In other words, the start of the

²A repository with code used for labeling the four categories is available at github.com/pwang1230/turntaking-wang-2025.

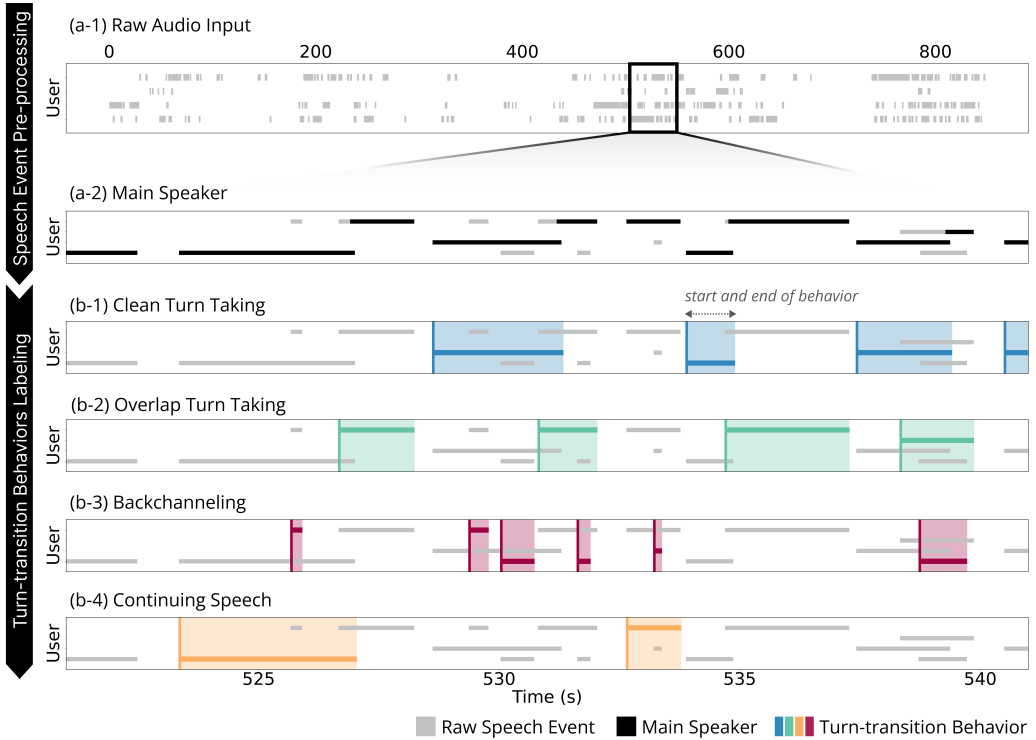


Fig. 2. Categorization of Turn-Transition Behaviors based on Example Raw Audio Input. Each row represents a unique user in the VR group discussion session. (a) shows example of speech event pre-processing and (b) shows subsequent behavioral labeling. In (a-2), the main speaker is labeled with black-colored segments. In (b-1)–(b-4), the start of each turn-transition behavior is denoted by a vertical line and its duration shaded in with the color associated with the turn-transition behavior. The speech utterances plotted are based on a recorded VR session and are further edited to improve visual clarity (e.g., adjusting speech events in close proximity).

Table 1. Summary Statistics of Turn-Transition Categories. The table presents the count of labels for each category by week and their sum across all four weeks. In parenthesis, we report the frequency of categories, where the unit is occurrences per minute across all sessions. Notably, the frequencies and counts of clean turn taking and continuing speech exceed that of overlap turn taking and backchanneling.

	Week 1	Week 2	Week 3	Week 4	Total
Clean Turn Taking	1675 (4.42)	2386 (4.71)	1712 (5.12)	2149 (4.87)	7922 (4.77)
Overlap Turn Taking	622 (1.64)	939 (1.86)	771 (2.30)	904 (2.05)	3236 (1.95)
Backchanneling	571 (1.51)	947 (1.87)	754 (2.25)	771 (1.75)	3043 (1.83)
Continuing Speech	1866 (1.12)	2446 (4.83)	1600 (4.78)	2292 (5.20)	8204 (4.94)

speech event of the new speaker precedes the end of the previous speech event. We again marked the start of a turn taking event as the beginning of the new speaker's speech event.

- **Backchanneling.** We determined backchanneling by checking for instances when a non-main speaker began and ended their speech event without ever being the main speaker. The

start of a backchanneling event is labeled as the start of the speech event of the non-main speaker. We show examples of backchanneling in Figure 2b-3. One noteworthy scenario to this labeling scheme is when the non-main speaker's speech event covers a change in the main speaker. As shown in Figure 2b-3, specifically the final backchanneling event by the bottom user, the main speaker changed during the speech event. Since the bottom user was never labeled the main speaker during their speech event, we labeled this instance as backchanneling.

- **Continuing Speech.** Continuing speech is similar to the scattered contribution category proposed by Jokinen et al. [49]. In their formulation, scattered contributions are defined as two adjacent speech instances of the same speaker being spaced less than 200ms apart, while those that are further apart are considered sequences of turns by the same speaker. Since we are interested in understanding how and when turn transitions take place (i.e., changes in speakers), we combined these two scenarios under the broader category of continuing speech. More concretely, we label continuing speech by first locating instances of adjacent speech events of the same user and then marking the start of the later speech event as the beginning of the turn-transition behavior. Figure 2b-4 shows instances of continuing speech behaviors.

3.3 Feature Selection

We formulated features based on prior literature. As reviewed in Section 2, user motion, individual and group differences, and verbal behaviors are predictive of user intentions and social dynamics. Past works have also shown that information relating to the current speaker [17] and their relationship with listeners [44, 45] offer insights to speech behaviors. Importantly, although our dataset varied the virtual spatial context, we chose to not encode these differences because they were manipulated at two distinct levels (i.e., high and low for ceiling height, large and small for amount of visible space). Encoding this study-driven parametrization into predictive models will not generalize to new spatial contexts. Instead, we treat these spatial variations as stimulus sampling to enhance external validity [51, 114].

Formally, we extracted features related to listeners, previous speakers, the dyadic relationship between listeners and speakers, their relationship with the group, and verbal behaviors. We further encoded features related to two users: a user of interest which we refer to as the *main user*, and the previous speaker, which we refer to as the *reference user*. When extracting motion-related features, we examined the tracking data of 1-second windows at moments prior to the start of turn-transition behaviors. The 1-second window size is determined based on piloting and aligns with past literature predicting speaking intentions [17]. Table 2 presents a summary of extracted features.

3.3.1 Speech-Related Features. The first feature group is related to speech behaviors, as they were shown to be predictive of who the next speaker is [85]. Following Parker [85], we encoded speech sequences leading up to the start of the examined window. For this, we begin at the timestamp at the start of the examined window, and trace backwards each time the main speaker changes, at which point we increment the turn index by 1. If the main speaker for a given turn is the main user, we assign the feature value at the corresponding turn index “u”. If we encounter a new speaker at a preceding turn, we assign the feature value a new symbol representing that user (e.g., “a”, “b”, “c”). For turns with a speaker who we had already created a symbol for, we assigned their symbol to the feature value. In our setup, we encoded 10 preceding speaker indices, assigning “NA” when there were no more changes in main speakers preceding the last encoded turn. In practice, we processed the speech sequence features at each turn index using one-hot encoding.

Table 2. Summary of Features Selected for Prediction Tasks. We present features by their groups, namely those related to speech, individual and group differences, egocentric motion, dyadic relationship, and group relationship.

Feature Group	Description of Features	Number of Features ¹
Speech-Related [85]	Prior speech sequences	10 preceding speaker indices
	Number of speaking turns before the main user's last speech event	1
	Time before the end of the main user's last speech event	1
	Whether the main user has spoken before	1
Individual- and Group-Level Differences [6, 32, 47, 60, 69, 90]	Main user big-5 personality	5
	Reference user big-5 personality	5
	Big-5 personality averaged across all users	5
	Group size	1
Egocentric Motion ² [6, 17, 44, 70, 72]	Position and orientation of a user's headset and two hand-held controller, where the user is either the main user or reference user	3 tracked points $\times$ 6 DOFs $\times$ 3 summary stats. $\times$ 2 meas. $\times$ 2 users = 216
Dyadic Relationship [17, 45, 49, 58]	Direct visual attention from the main user to the reference user	3 summary stats. $\times$ 2 meas. = 6
	Direct visual attention from the reference user to the main user	3 summary stats. $\times$ 2 meas. = 6
	Interpersonal distance between the main user and reference user	3 summary stats. $\times$ 2 meas. = 6
	Visual shared space calculated between the main user and reference user	3 summary stats. $\times$ 3 dists. $\times$ 2 meas. = 18
Group Relationship [17, 45, 49, 58]	Direct visual attention from a user to all remaining users, where the user is either the main user or reference user	3 summary stats. $\times$ 2 meas. $\times$ 2 users = 12
	Direct visual attention from all remaining users to a user, where the user is either the main user or reference user	3 summary stats. $\times$ 2 meas. $\times$ 2 users = 12
	Interpersonal distance between a user and all remaining users, where the user is either the main user or reference user	3 summary stats. $\times$ 2 meas. $\times$ 2 users = 12
	Visual shared space calculated between a user and all remaining users, where the user is either the main user or reference user	3 summary stats. $\times$ 3 dists. $\times$ 2 meas. $\times$ 2 users = 36

¹ 3 Tracked points refer to the headset and two controllers; 6 degrees of freedom refer to x, y, and z position and the roll, pitch, and yaw orientation; 3 summary statistics refer to the minimum, maximum, and average across the examined time period; 2 measurements refer to the raw value and first order derivative (i.e., velocity); 3 distances refer to 1, 5, and 10 meters. DOFs = degrees of freedom; stats. = statistics; meas. = measurements; dists. = distances.

² Due to the body-space transformation and centering operations, the total number of non-trivial features for egocentric motion is 198.

We introduced three additional features to describe the main user's verbal behaviors: (1) the number of speaking turns before the main speaker's last speech event, (2) the duration of time before the end of the main user's last speech event, and (3) whether the main user has spoken before during the current session. Deriving these features for the reference user was unnecessary since that user is typically set to the previous speaker.

3.3.2 Individual- and Group- Level Features. Drawing from works showing correlations between speech behaviors and personalities and group compositions [6, 32, 47, 69, 90], we formulated features based on individual and group characteristics. Specifically, we encoded the Big-5 personality traits [30] of the main and reference users, and group average. The authors of the dataset measured Big-5 personality using the ten item personality measure (TIPI) [30], with each personality trait ranging between 1–7 and calculated as the mean of two items.

3.3.3 Egocentric Motion Features. Egocentric motion features capture the pose and motion of individuals. For this, we extracted egocentric motion features from both the main and reference users. For each user, we extracted the position (i.e., x, y, z) and orientation (i.e., pitch, yaw, roll) of a user's headset and two controllers by both their raw values and velocities. From the raw values and velocities across the 1-second window, we summarized each of the 6 degrees of freedom by their average, minimum, and maximum. We transformed the features corresponding to the raw values into the body-space coordinate system proposed in Miller et al. [71] and centered the horizontal coordinates (i.e., x, z) such that the individual's head position is at the origin of the horizontal plane, as opposed to their global coordinates in the virtual space. The rationale behind this transformation is to encode the egocentric poses by mapping the raw coordinates and orientations to a coordinate system based on the individual's forward head direction. This allows the features to encode poses even when individuals move around physical or virtual environments.

When calculating the velocity of positions and orientations, we derived velocities from raw values after transforming the original values into the body-space coordinates without the centering operation. Not centering the coordinates based on the head position allows the features to retain information of the user's head position velocities in the horizontal plane. When encoding head yaw, we did not calculate its velocity based on the transformed values but instead on the yaw angles prior to any transformations. This procedure preserves information on how a user rotates their head in the yaw axis, which body-space transformation loses by reorienting based on head yaw.

3.3.4 Dyadic Relationship Features. The next feature group describes the dynamics between the main and reference users. Similar to prior work [17, 45, 49], we extracted the direct visual attention (i.e., gaze direction approximated using head orientation) from the main user to the reference user, and vice versa. Specifically, we determined user A's direct visual attention towards user B by calculating the angle between user A's forward head orientation in the yaw axis and the vector pointing from user A's head position to user B's head position in the horizontal plane. Similar to the egocentric motion features, we derived the raw values and velocities across the 1-second window and summarized them by their average, minimum, and maximum. Since users moved around virtual environments in the dataset, we also encoded interpersonal distance as the distance between two users' head position in the horizontal plane.

One characteristic that direct visual attention fails to capture is whether two users are facing the same direction. For example, if both users' head orientations are at 90 degrees from the other user, they could be facing the same direction or have their backs against one another. These two scenarios can entail different dynamics as users facing the same direction are likely looking at similar parts of the environment. We therefore introduced a measurement of visual shared space quantifying how much users' visual field of views overlap. Concretely, we first drew isosceles triangles with the vertex connecting the two equal sides of length vs_l located at the user's head and oriented in the head's forward direction in the horizontal plane. The angle between the two sides is set to the horizontal field of view of the Meta Quest 2 headset (104 degrees). The raw value of the visual shared space between two users at distance l is defined as the amount of spatial overlap in m^2 of the two triangles drawn with $vs_l = l$. In simpler terms, this measurement estimates how

much virtual space the two users share in their field of views, assuming that their visual attention stretches vs_l meters. For our setup, we extracted values for vs_l at 1, 5, and 10 meters.

Using the raw values calculated for direct visual attention, interpersonal distance and visual shared space, we extracted the raw values and their changes (i.e., velocity) over the sampled windows and summarized the two measurements into their average, minimum, and maximum.

3.3.5 Group Relationship Features. Finally, we encoded features between a user of interest (i.e., main or reference user) with the group. Specifically, we extracted (1) the dyadic direct visual attention angle from each user of interest to each group member and (2) those from each group member to that user. Similar to the dyadic relationship features, we calculated the interpersonal distance between each of the two users with the rest of the group. Finally, we derived the visual shared space between each dyadic pair between a user of interest and the rest of the group at $vs_l = 1, 5, 10$ meters. From the raw values extracted for direct visual attention, interpersonal distance, and visual shared space across the sampled window, we derived the average value and velocity for each dyadic pair, grouped them by whether they are related to the main or reference user, and finally summarized them by their average, minimum, and maximum.

3.4 Turn-Taking Behavior Prediction Tasks

Drawing from past literature on predicting speech behaviors [17, 44, 45, 58, 85], we focused on three prediction tasks, which capture the “what”, “who”, and “when” of turn-taking behaviors.

- **Turn Taking vs. Continuing Speech.** The “what” poses the question of whether we can predict the type of turn-transition behavior. In particular, in line with past research [45, 49, 58], can we distinguish whether a turn-transition behavior will be a turn-transition to a new speaker (i.e., turn taking), or a continuing speech from the previous speaker (i.e., turn keeping)?
- **Next Speaker Prediction.** The “who” focuses on predicting who the new speaker is prior to the start of a turn-taking behavior (i.e., clean turn taking, overlap turn taking). Next speaker prediction is another common task used in modeling speech behavior [43, 45, 46, 85].
- **Timing of Turn-Taking Behaviors.** To investigate the “when” of turn-taking behaviors, we investigate whether we can predict when the next speaker will speak, a task that frequented past literature [44–46]. Specifically, can we differentiate between moments associated with a new speaker right before a turn taking event and those sampled before these moments?

3.5 Machine Learning Models

To study RQ1, we predicted the three tasks outlined in Section 3.4 by first extracting the features in the 1-second windows prior to the start of the turn-transition behaviors. For each task, we built and compared performance across four predictive models commonly used to predict individual characteristics and group dynamics [16, 71, 75]: logistic regression, multi-level perceptron (MLP) classifier, random forest classifier, and gradient boosting classifier. We formulated each task as a binary classification task and measured prediction performance using the area under the curve (AUC) of the receiver operating characteristics curve. We implemented all models using Python’s scikit-learn library [24].

For evaluation, we employed cross-validation similar to [16, 28]³. To start, we partitioned the data into training and testing and standardized all continuous features. For measuring accuracy, we reported the performance of models trained on 90% of the data and evaluated on the remaining 10% (i.e., k-fold ≈ 10). Specifically, we calculated the averages and standard errors of the AUC evaluated

³We conducted an additional analysis that focused on evaluating model performance across sessions, groups, and weeks. Since the results largely align with our cross-validation evaluations, we include this analysis in Appendix D.

Table 3. Model Performance (measured as the AUC of the ROC curve) on Predicting Turn-Taking Behavior vs. Continuing Speech. For metrics using cross validation, which we denote using the subscript cv, we report the average and standard error across all folds. An AUC of 0.50 means that the model's ability to distinguish between positive and negative samples is no better than random chance. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metrics			
	Session _{cv}	Group _{cv}	Week _{cv}	Week 4
Logistic Regression	0.71 (0.01)	0.69 (0.01)	0.70 (0.00)	0.69
MLP Classifier	0.72 (0.01)	0.71 (0.01)	0.71 (0.00)	0.71
Random Forest Classifier	0.77 (0.00)	0.75 (0.01)	0.76 (0.00)	0.76
Gradient Boosting Classifier	0.78 (0.00)	0.77 (0.01)	0.78 (0.00)	0.78

on the testing data across all folds. As a benchmark measurement, we partitioned the data such that we trained on data from 90% of the sessions and tested was on the remaining 10% (Session_{cv}).

We report two additional sets of results evaluating the robustness of models on unseen groups, activities, and weeks (RQ2). First, we measured performance on unseen groups by partitioning the data such that we trained on data from 90% of groups and tested on the remaining 10% (Group_{cv}). We then quantified performance across unseen weeks using models trained on all but one week's data and tested on the remaining week (Week_{cv}). Finally, we report performance on the final week after training models on data from the first three weeks. As group activities differed weekly, our evaluation metric for unseen activities is the same as that for unseen weeks.

4 Results

In this section, we report on model performance (RQ1) and their robustness (RQ2) for the prediction tasks outlined in Section 3.4. Then, we present analyses investigating feature importance and how the features are related to model predictions (RQ3).

4.1 Predicting Turn-Taking Behaviors

4.1.1 Differentiating between Turn Taking and Continuing Speech. We built models for predicting whether a speech event was going to be a turn transition or one where the previous speaker will continue to speak. We compared clean turn taking, and not overlap turn taking, to continuing speech as there are by definition pauses before both types of turn-transition behaviors. We defined positive samples as those corresponding to the beginning of turn-taking behaviors and negative samples as those corresponding to the start of continuing speech behaviors. We chose the upcoming speaker as a main user for positive samples and a randomly selected user who is not the previous speaker for negative samples. For both positive and negative samples, the reference user was the previous speaker. We randomly sampled the dataset to maintain an equal number of positive and negative samples, which yielded 15844 samples.

We present model performance in Table 3. Notably, gradient boosting classifiers outperformed other models across all performance metrics, achieving a benchmark accuracy (i.e., Session_{cv}) of 0.78 AUC, and 0.77–0.78 AUC across the three remaining metrics. Random forest classifiers achieved the second highest accuracy on the benchmark metric with a 0.77 AUC, followed by the MLP classifier at 0.72 AUC, and the logistic regression at 0.71 AUC. When evaluating on unseen groups, performance accuracies were slightly lower but generally robust across all four models. Compared to the benchmark metric, models achieve similar accuracies on unseen weeks and activities.

Table 4. Model Performance (measured as the AUC of the ROC curve) on Next Speaker Prediction. For metrics using cross validation, which we denote using the subscript $_{cv}$, we report the average and standard error across all folds. An AUC of 0.50 means that the model's ability to distinguish between positive and negative samples is no better than random chance. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metrics			
	Session $_{cv}$	Group $_{cv}$	Week $_{cv}$	Week 4
Logistic Regression	0.73 (0.01)	0.72 (0.01)	0.73 (0.01)	0.75
MLP Classifier	0.72 (0.01)	0.71 (0.01)	0.72 (0.01)	0.74
Random Forest Classifier	0.75 (0.01)	0.74 (0.01)	0.74 (0.00)	0.75
Gradient Boosting Classifier	0.77 (0.01)	0.75 (0.01)	0.77 (0.01)	0.78

Table 5. Model Performance (measured as the AUC of the ROC curve) on the Timing of Turn Taking. For metrics using cross validation, which we denote using the subscript $_{cv}$, we report the average and standard error across all folds. An AUC of 0.50 means that the model's ability to distinguish between positive and negative samples is no better than random chance. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metrics			
	Session $_{cv}$	Group $_{cv}$	Week $_{cv}$	Week 4
Logistic Regression	0.61 (0.00)	0.61 (0.00)	0.61 (0.00)	0.60
MLP Classifier	0.63 (0.00)	0.63 (0.01)	0.63 (0.00)	0.61
Random Forest Classifier	0.68 (0.00)	0.67 (0.01)	0.67 (0.01)	0.66
Gradient Boosting Classifier	0.72 (0.00)	0.71 (0.01)	0.71 (0.00)	0.71

4.1.2 Predicting the Next Speaker. We predicted the upcoming speaker of a turn-taking event. Specifically, we aggregated positive samples by extracting those corresponding to moments at the beginning of turn-taking events (i.e., clean turn taking, overlap turn taking), where the main user is set to the upcoming speaker. We then collected negative samples at the same moments, but with the main speaker set to a user who is neither the upcoming nor the previous speaker. The reference user is set to the previous speaker for all samples. We randomly resampled negative labels to balance positive and negative samples, yielding 22316 samples.

Table 4 summarizes our results. Gradient boosting classifier achieved the highest accuracies across all four performance metrics, with an AUC of 0.77 on the benchmark metric, followed by the random forest classifier at 0.75 AUC, the logistic regression at 0.73 AUC, and finally the MLP classifier at 0.72 AUC. All four prediction models were robust when tested on unseen groups and weeks, with the performance on unseen groups again being lower but comparable to the benchmark evaluations.

4.1.3 Predicting the Timing of Turn Taking. Finally, we aimed to understand whether we can differentiate between moments associated with an upcoming speaker right before a turn taking event and those sampled prior to these moments. We first collected positive samples by extracting samples corresponding to moments before the start of turn-taking events (i.e., clean turn taking, overlap turn taking), with the main user set to the upcoming speaker. For negative samples, we aggregated samples corresponding to moments that were prior to the 1-second window immediately

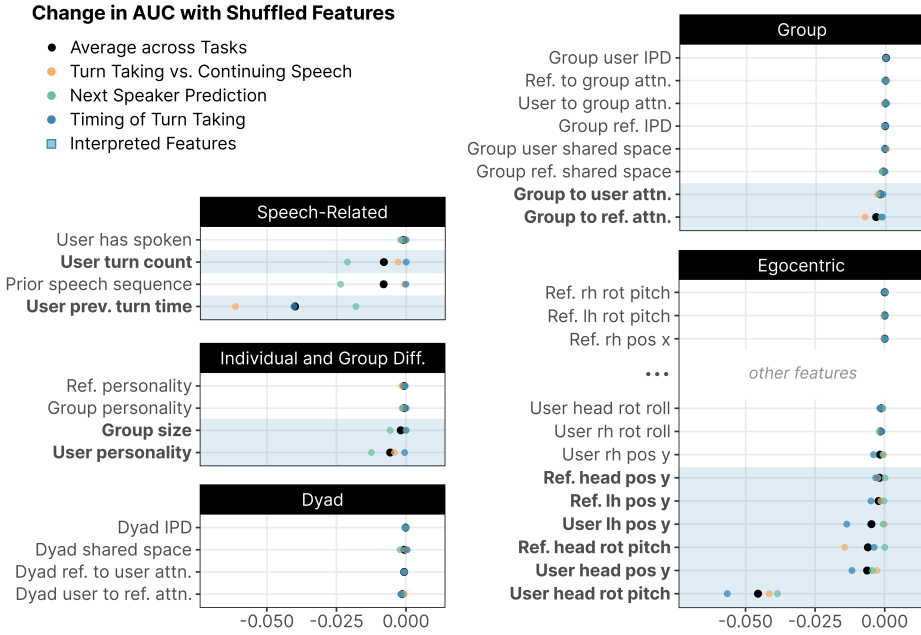


Fig. 3. Feature Importance Analysis. Each panel reports the change in AUC of the ROC after shuffling features across the three prediction tasks for a given feature group. Black dots denote the average drop in accuracy across three tasks. Features within each panel are ordered in increasing average drop in accuracy from top to bottom. Lower values (i.e., greater decrease in accuracy) represents higher MDA. For clearer presentation, we did not plot egocentric features with moderate importance. Bolded features and their corresponding rows with light blue backgrounds indicate features that are further interpreted in Section 4.3. Diff. = differences; Ref. = reference user; prev. = previous; IPD = interpersonal distance; attn. = visual attention; lh = left hand; rh = right hand.

before the start of turn-taking event. For this, we began by sampling moments at windows starting 2, 4, 6, 8, 10, and 12 seconds prior to the start of turn transition events. We then filtered out data samples for which any speech event had occurred between the earlier moment and the start of the turn transition event for the upcoming speaker. In other words, the negative samples consisted only of moments when the upcoming speaker is not preparing to speak. This sampling technique is similar to Chen et al. [17] as we sampled through a sliding window. For both positive and negative samples, the reference user is set to the previous speaker at the sampled moment. We randomly resampled the aggregated data to ensure an equal number of positive and negative samples, resulting in 22316 samples.

We detail performance results in Table 5, which were lower than those from previous tasks. The best performing model, the gradient boosting classifier, achieved an accuracy of 0.72 AUC on the benchmark metric, while the logistic regression model recorded the lowest accuracy of 0.61 AUC on the same metric. The random forest and MLP classifiers had benchmark accuracies of 0.68 and 0.63 AUC, respectively. Performance on unseen groups, activities, and weeks were robust, with models reporting slightly lower but similarly accuracies on other cross validation evaluations.

4.2 Feature Importance Analysis

4.2.1 Mean Decrease in Accuracy Based on Gradient Boosting Classifiers. To understand how features are related to model performance and predictions (**RQ3**), we first computed the mean decrease in accuracy (MDA) for the top performing model (i.e., gradient boosting classifier), a technique used for probing feature importance in machine learning models [37, 72, 82]. Specifically, we organized features based on their feature groups and definition, for example splitting up egocentric motion features by the user, tracked points, and axis of motion. For each task, we built models used to compute the benchmark cross validation performance (i.e., Session_{cv}) and evaluated them on each testing set with a particular set of features randomly shuffled. This procedure breaks up any relationship the set of features may have with the model prediction. We then calculated the average change in the testing AUC across all folds and present them in Figure 3. We omitted plotting egocentric features with moderate importance, but included all features in Appendix B.

Analysis of speech-related features showed the highest importance on the user's previous turn time, followed by prior speech sequence, user turn count, and whether the user has spoken earlier in the session. For the main user's previous turn time feature, feature importance for the turn taking vs. continuing speech task was the highest, followed by the timing prediction and next speaker prediction tasks. For features on the main user's turn count and prior speech sequence, they exhibited the highest importance for models built to predict the next speaker, followed by the turn taking vs. continuing speech task, and lastly the timing prediction task. Our results on the individual and group differences features revealed a comparably high importance of the main user personality, followed by group size, group personality, and lastly the reference user personality.

For egocentric motion features, the top six features were related to the main and reference users' head yaw rotation, head y-axis position, and left hand y-axis position. Within the top six features, feature importance for those related to the main user were in most cases higher for the timing prediction task compared to other tasks. For dyad relationship features, those related to direct visual attention had greater importance than those related to the visual shared space and interpersonal distance, though their change in AUC near zero implies low importance. In terms of group relationship features, features related to the group's direct visual attention towards the reference and main users exhibited greater importance than the remaining features.

4.2.2 Feature Significance Based on Logistic Regression Models. We investigated predictor significance of feature groups using logistic regressions. Benefits of studying linear models include their high interpretability and simplified independence assumption affording well-defined hypothesis tests for evaluating feature significance [19]. In contrast, recent methods for interpreting more complex models such as the random forest and gradient boosting classifiers have not addressed evaluations of feature importance through statistical tests [31, 64, 65]. For this, we grouped features based on their constructs. Then, for each trained logistic regression, we conducted Wald tests to evaluate joint linear hypotheses, specifically testing whether all features related to a given construct significantly contribute to the model. We present the full results in Appendix C but highlight general trends for the remainder of this section. We evaluated significance at $\alpha=.05$.

Specifically, all but one group of speech-related features significantly contributed to model prediction across all three tasks ($ps<.046$). For individual and group differences features (i.e., personality, group size), most significantly impacted the tasks of predicting turn taking vs. continuing speech and next speaker prediction ($ps<.020$). In contrast, for individual and group differences features, only features related to the previous speaker's personality ($p<.001$) significantly predicted turn-taking timing. For dyad-related features, visual attention features between the reference and main users significantly predicted the next speaker ($ps<.001$) while the interpersonal distance between the reference and main users was significant in distinguishing between turn taking and continuing speech behaviors ($p=.004$). For features related to groups dynamics, all constructs significantly

predicted turn taking vs. continuing speech ($ps < .017$). Group features related to visual attention were also significant in predicting at least one of the two remaining tasks. Finally, for egocentric features, 18 groups of constructs were significant in predicting turn taking vs. continuing speech ($ps < .035$), while only 6 were significant in predicting the next speaker ($ps < .027$). Notably, all 6 were related to the main user's egocentric behavior. For predicting the timing of turn taking, 12 groups of constructs were significant ($ps < .043$).

4.3 Feature Interpretation Analysis

To study **RQ3**, we analyzed how features with high importance are related to predictions of the best performing models (i.e., gradient boosting classifiers). We selected features within each feature group with high feature importance based on MDA. Notably, we included features deemed insignificant in the Wald-tests in Section 4.2.2 for two reasons. First, we interpreted the gradient boosting classifiers and not the logistic regressions. Using MDA from the gradient boosting classifiers allowed us to capture insights from features that may lack significance under the assumptions of a linear model but still contribute non-linearly. Additionally, logistic regressions are sensitive to collinearity and can complicate assessments of feature significance [23]. Another issue is an overemphasis on features with small effects given a large sample [104]. By leveraging MDA instead, we selected features that best reflected their contribution to the best performing models.

In aggregate, we analyzed 56 constructs. For speech-related features, we examined 2 constructs: the user turn count and user previous turn time. There were 6 constructs related to individual and group differences: main user personalities and group size. 12 were related to the group relationship, specifically those capturing the group's direct visual attention towards the main user and reference user. The remaining 36 were egocentric motion constructs, namely those describing the head position in the y-axis, rotation in the yaw axis, and left hand position in the y-axis for the main and reference users. We did not interpret individual prior speech sequence features as research has suggested the need to consider the interactions between prior speakers [85].

We interpreted features using partial dependence, a common technique used to interpret machine learning models such as random forests [7, 29] and gradient boosting classifiers [11, 81]. By definition, partial dependence measures the marginal effect on model prediction through varying the values to a feature and calculating the average probability estimates after this procedure. In our setup, we chose the gradient boosting classifiers trained on the entire dataset and varied each feature between its [0.05, 0.95] percentile.

4.3.1 Speech-Related Features. Seen in Figure 4a, varying the amount of time and number of turns to the main user's previous turn revealed large effects on the probability estimates. Specifically, the results suggested that listeners who had spoken more recently are more likely to engage in turn-taking behaviors and be the next speaker (i.e., predicting turn taking vs. continuing speech and the next speaker). The models also predicted that listeners are more likely to speak at a time closer to their previous speech activity. Our results on the user turn count feature revealed for the next speaker prediction task that listeners whose last speaking event occurred at earlier turns are less likely to become the next speaker.

4.3.2 Individual- and Group- Level Differences. Shown in Figure 4b, for predicting the next speaker and differentiating between turn taking and continuing speech, listeners with higher values of extraversion are more likely to be the next speaker. There was little variation in probabilities estimates when varying main user personalities for predicting the timing of turn-taking. This suggests that listeners with different levels of Big-5 personalities do not exhibit different patterns in *when* they decide to speak. There were weaker but noticeable relationships between the probabilities estimates and the listener's conscientiousness, neuroticism, and openness. For example, for the next

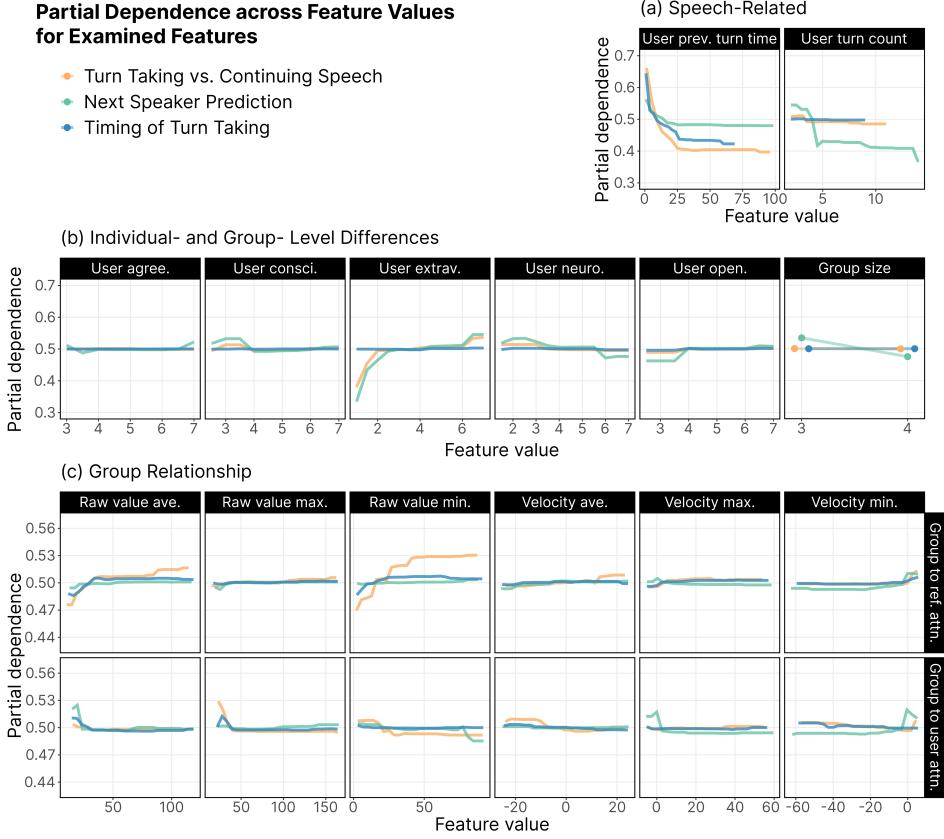


Fig. 4. Partial Dependence for Selected Features Related to Speech, Individual- and Group- Level Difference, and Group Relationship. Higher y values denote greater probability estimates for the main user's speaking intentions. Raw values are quantified in degrees, and velocities in degrees per second. (a–c) show the partial dependence of the selected features, with colors denoting the prediction task. Features related to speech, the main user's extraversion, and raw values of direct visual attention from the group to the main and reference users revealed noticeable changes in the average probability estimates across feature values. Ref. = reference user; ave. = average; max. = maximum; min. = minimum; agree. = agreeableness; consci. = conscientiousness; extrav. = extraversion; neuro. = neuroticism; open. = openness; attn. = visual attention.

speaker prediction task, the model predicted that listeners reporting a lower level of openness were less likely to be the next speaker.

4.3.3 Group Relationship. The most salient relationship in the partial dependence plots regarding group relationships (Figure 4c) were those related to the group's direct visual attention towards the previous speaker. Notably, the models predicted that a new speaker is more likely to take over a previous turn when listeners are looking away from the previous speaker (i.e., larger minimum and average value for the direct visual attention angle from the group to the reference user). There is a similar but weaker trend when predicting the timing of turn taking.

The partial dependence plots for the group's direct visual attention towards a listener revealed the opposite directionality between prediction estimates and feature values, where a listener is predicted to be more likely to take over a turn when the remaining group is looking more directly

Partial Dependence across Feature Values for Examined Features (cont'd)

- Turn Taking vs. Continuing Speech
- Next Speaker Prediction
- Timing of Turn Taking

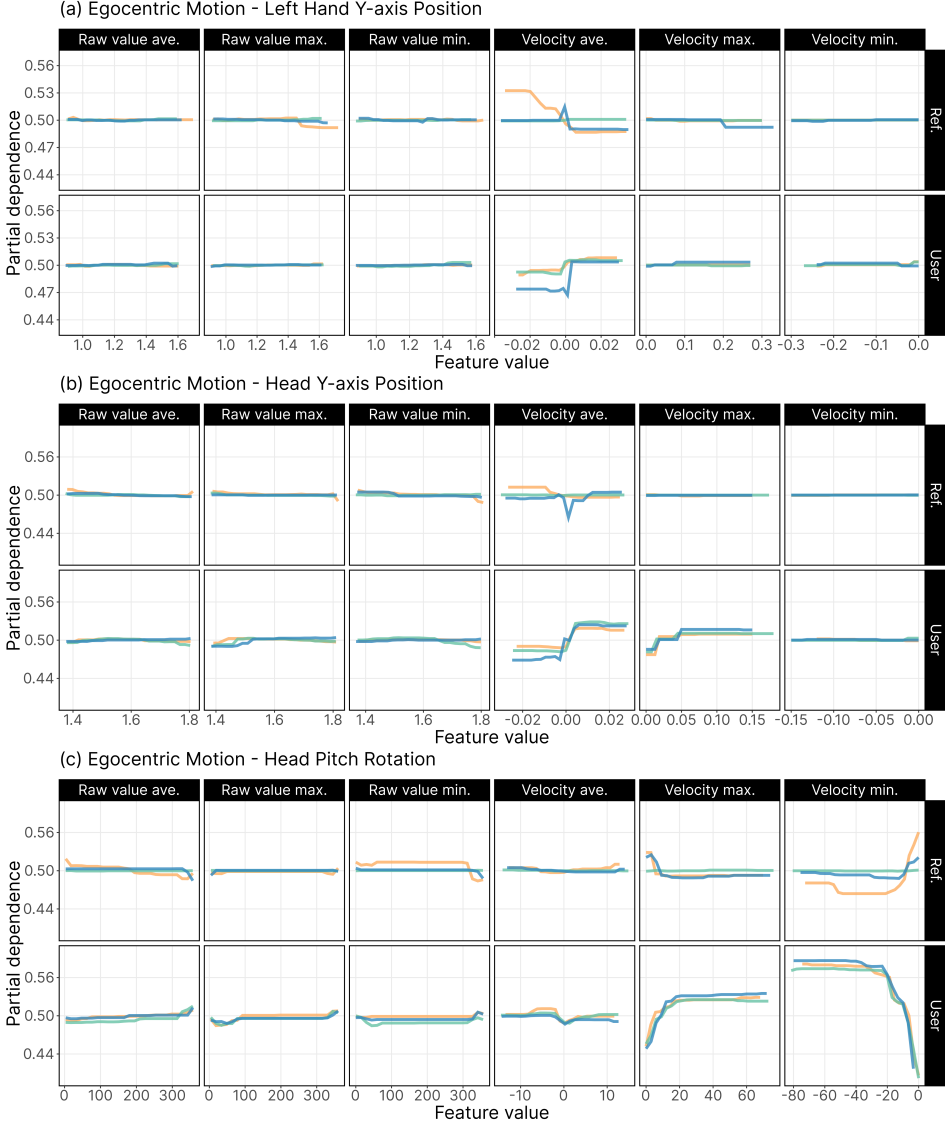


Fig. 5. Partial Dependence for Selected Features Related to Egocentric Motion. Higher y values denote greater probability estimates for the main user's speaking intentions. Raw values are quantified in degrees, and velocities in degrees per second. (a–c) show the partial dependence of the selected features, with colors denoting the prediction task. Notably, results showed the largest visual changes in average probability estimates when varying (1) the average velocity of the users' left hand y-axis position, (2) the maximum and minimum velocities of the users' head pitch rotation, and (3) the average velocity of the users' head y-axis position. Ref. = reference user; ave. = average; max. = maximum; min. = minimum.

at them (i.e., small maximum and average value for the direct visual attention angle from the group to the main user). This pattern is, however, less pronounced than the greater variability in probability estimates exhibited for direct visual attention measurements from the group to the previous speaker.

4.3.4 Egocentric Motion. Figure 5 presents our findings on egocentric motion features. Shown in Figure 5a, the feature with the greatest variation in partial dependence for the **left hand y-axis position** was the average velocities for both the listener and previous speaker. Notably, the model used for differentiating between turn taking and continuing speech predicted that previous speakers who move their left hand upward in the vertical axis (i.e., large positive values in the average y-axis velocity) are more likely to continue speaking. For listeners, models for all three tasks predicted that listeners are more likely to speak in moments right after they move their left hand upward (i.e., large positive values in the average y-axis velocity).

For egocentric features related to **head y-axis position**, the salient trends were related to velocities (Figure 5b). To start, the model for the turn-taking timing task predicted that a listener is more likely to begin speaking if the previous speaker's head had on average not moved a substantial amount in either direction in the vertical axis. This was not the case for models for the other two tasks. On the contrary, all three models predicted that a listener is more likely to begin speaking if their head had on average moved upward in the y-axis (i.e., large positive values for the average y-axis velocity). Relatedly, a listener is predicted to more likely speak if their head had at any point in the sampled window moved quickly upward in the vertical axis (i.e., large positive values for the maximum y-axis velocity).

Finally, for features related to **head pitch rotation**, we found distinctive patterns in maximum and minimum velocities of the pitch rotations for the main and reference users (Figure 5c). We interpret this as the models finding the extremes of an individual's head pitch rotation velocities across the 1-second window more informative than their averages. There were no noticeable trends in partial dependence for the previous speaker's head pitch rotation on the next speaker prediction task. Models for the remaining two tasks predicted that a listener is less likely to speak if the previous speaker has at any point extensively rotated their head upward, which corresponds to small negative values for the minimum pitch velocity. The models also predicted that a listener is less likely to speak if the previous speaker has at any moment within the 1-second window extensively rotated their head downward (i.e., large positive values for the maximum pitch velocity). The trends in partial dependence on the listener's head pitch rotation were opposite to that of the previous speaker. Namely, all three models predicted a listener is more likely to speak when they have engaged in substantial downward head rotation (i.e., large positive values for the maximum pitch velocity) or upward head rotation (i.e., small negative values for the minimum pitch velocity).

5 Discussion

5.1 Practical Implications of Turn-Taking Prediction

5.1.1 Model Selection and Performance. With regards to **RQ1**, we found that the gradient boosting classifiers achieved consistently the best performance on all tasks across performance metrics. This suggests that the relationship between the features and outcome variable is non-linear and that interactions between features are likely relevant for prediction. Specifically, the models obtained accuracies of 0.75–0.78 AUC on the tasks of identifying the new speaker and distinguishing between turn transitions and continuing speech, and had poorer performance on predicting when a new user will begin speaking with accuracies of 0.71–0.72 AUC. These accuracies are considerably higher than prediction by chance, are considered to have excellent discrimination between positive and negative samples [119], and comparable to prior research [17].

As AUC represents the probability that a random positive sample is rated higher than a random negative sample [13], there remains room for prediction improvements, albeit limited by the complexity inherent to social dynamics and VR [20, 68]. To further explore the practical implications of turn-taking predictions, we present an additional analysis in Appendix E. In it, we consider the task of predicting continuing speech and turn-taking events and extended the binary classification problem into a multiclass one. This analysis yielded comparable findings and found that three-person groups yielded higher prediction accuracy than four-person ones. One possible explanation for the lower performance in larger groups is the greater unpredictability in larger-group social dynamics. While there were more four-person discussions than three-person ones in our dataset, greater complexity in social dynamics may warrant more data for larger group sizes. When modeling and deploying comparable models in real-world settings, our findings here suggest that practitioners should sample diverse group sizes during data collection, exercise caution when predicting turn-taking behaviors for groups whose sizes differ from the training data, and not assume higher prediction accuracies in smaller groups will translate directly to larger groups.

Additionally, our model accuracies suggest that near-term uses cases should prioritize applications that either 1) are minimally impacted by potential false labels or 2) build off insights from feature importance and interpretation. Given the lower performance in predicting turn taking timing, practitioners should build tools that leverage predictions for who the next speaker is, as opposed to when they will begin speaking.

Importantly, overall accuracy alone does not capture the full picture to model performance, for generalizability is critical for practical deployments. In our work, we showed that predictive performance is robust over time and across unseen groups and activities (RQ2), with accuracies for evaluations based on unseen groups being slightly lower but comparable to benchmark metrics. Considering also the diverse group activities and virtual mobility allowed in our dataset, these results suggest that practitioners can expect robustness across time and unseen activities and groups when training and deploying predictive models using similar approaches. One possible explanation for the robustness is the high number of groups and wide range of activities seen during training allowing models to extrapolate generalizable insights.

5.1.2 Application of Findings. Our findings offer insights for practical applications. For one, for educators and organizers facilitating group discussions, having a tool for predicting next speakers can help with moderation (e.g., smoother transition between speakers). As these facilitators may need to multi-task and monitor multiple sessions [94], such tools can reduce the cognitive load for interpreting the nonverbal behavior of multiple social scenes concurrently. Another use case is in facilitating interactions with users joining at different levels of immersion (e.g., audio only, full body tracking). For example, model that predict and notify audio-only users of the speaking intentions of immersive users can help them better navigate social scenarios. Conversely, systems can model nonverbal behaviors from audio-only users and notify fully-immersed users of possible speaking intentions of audio-only users, thereby addressing the asymmetry in immersion.

Another application is in training user awareness in detecting turn-taking behaviors. One option is to visually guide users to notice predictive nonverbal features before speaking turns, similar to immersive training systems [14, 40]. Such training can be beneficial to populations who struggle socially [26, 33] and professionals in domains such as education and healthcare. Our features analysis on individual and group differences also offers implications on group composition. Namely, aligning with works on social dynamics and individual differences [6, 10, 60, 90], practitioners interested in shaping turn-taking behaviors can vary group size and leverage individuals characteristics to form the “ideal social group.”

Finally, our work sheds light on virtual agents instantiation. Conversational agents struggle with interrupting at appropriate times [67], and we foresee future systems leveraging turn-taking predictions to decide whether and when to interject in natural and non-intrusive ways. Possible application of virtual agents include teachers in classrooms, moderators in focus groups and conferences, and facilitators in support groups.

5.2 Theoretical Implications of Turn-Taking Behaviors in VR

Understanding how features predict turn-taking behaviors offer practical benefits (e.g., smaller models with curated features) and theoretical implications (RQ3). Here, we expand on the latter, drawing parallels to past literature and supplementing them with post-hoc analysis.

The results on individual and group differences extend works demonstrating that personalities such as extraversion, agreeableness, and neuroticism are correlated with user's verbal behaviors [6, 60]. As noted in Appendix C, features related to personalities and group sizes significantly contributed to logistic regression models in at least one of the three tasks. Additionally, we found that more extroverted listeners are predicted to more likely be the next speaker, corroborating past findings that extraversion is related to speaking time [60] and speaking turns [6]. Our analysis on group size suggested that listeners in three-person groups are more likely to be the next speaker compared to listeners in four-person groups. Though we focused on individual feature interpretation, it is also possible that models used group size in combination with motion-related features for their predictions, as group size can be related to head orientations [32]. Another related theoretical thread to consider regarding group size is how differences in group composition (e.g., personality, demographics) may affect social dynamics and psychological processes differently. For example, larger groups, compared to smaller ones, tend to report lower levels of group identity and greater role differentiation in performing group tasks, and consequently require more control from leadership to coordinate efforts across group members [38]. As such, with many of the collaborative activities participants engaged in within our dataset (see Appendix A for the list of discussion and group activities), we imagine more drastic and fundamental differences in group behaviors, both verbally and nonverbally, across group that greatly differ in size. While we studied three to four person social groups, just how comparable and similar are turn-taking behaviors between groups examined in this paper and those from much larger groups (e.g., over 10 people)? These findings and considerations highlight the need to consider individual and group characteristics when predicting social dynamics and instantiating virtual agents with distinct personas and natural behaviors. One possible approach is to continue leveraging VR to study large and diverse demographics with high variances in individual differences, and doing so in controlled settings [110].

Speech-related features exhibited high importance and salient effects on probability estimates for the gradient boosting classifiers, and were also significant predictors for the logistic regressions. Furthermore, our models predicted that those who more recently spoke, both in terms of speaking turns and time, are more likely to be the next speaker. While we did not analyze speech sequence features given the likely interactions between turn indices [85], it provided important signals to predicting speech behaviors, in particular to next speaker prediction.

Our results revealed that group visual attention behavior is useful in distinguishing between turn transition and continuing speech, with the individual that the group is looking more directly towards predicted to be more likely the upcoming speaker. These findings extend past insights that visual attention measurements are useful in modeling speech behaviors [49, 50] and classroom discourse [102] by demonstrating that the social group's direct visual attention towards individuals is still predictive of the next speaker even when VR groups are not constrained to fixed virtual positions. While features related to direct visual attention between the listener and previous speaker were significant in predicting the next speaker in logistic regressions, they did not exhibit high

importance for gradient boosting models. Nevertheless, these features could still offer insights on speech behaviors when sampled at other moments, for example during speech events or immediately after speaking turns.

Our models predicted that listeners who engaged in substantial rotation in head pitch (e.g., nodding, abruptly looking up or down) and moved their heads and left hands vertically upward exhibited more speaking intentions. Our feature significance results largely echoed the importance of these features. An additional analysis detailed in Appendix F on head pitch revealed that the combination of high instantaneous speeds in both the upward and downward directions can be further indicative of speaking intentions. We also found that the previous speaker's left hand y-axis position and the head y-axis position and pitch rotation influenced performance, though their feature significance and effects on probability estimates are less consistent across tasks. Of note, features related to the previous speaker's left hand y-axis position were not significant in the logistic regressions. Possible explanations include their potential interactions with other features and non-linear contributions to the predictions. The inconsistency across tasks could be due to the different roles the previous speaker plays in the tasks. For example, while the previous speaker's egocentric motion could help predict whether they will continue to speak, it is not informative for predicting which listener will speak next. One explanation for the finding on left-hand motion is its association with users raising their hands to access the menu, tablet, and audio button. These findings corroborate past research demonstrating the predictive capabilities of tracking data for modeling speech behavior [17, 44], and contribute insights on how egocentric motion is related to the predicted behaviors.

Importantly, though our findings align with prior literature, feature interpretation can differ across virtual context. Similar to how social factors such as group size and personality influenced our model predictions, virtual context such as room size can also impact nonverbal behaviors [35, 36]. Platform-specific characteristics such as the location of virtual menus could influence body motion and visual attention preceding speech events. Though our dataset varied group size and spatial context, model performance and interpretation may still differ when evaluated on novel social settings.

6 Limitations and Future Work

Our work has several limitations. To start, we did not interpret features related to prior speech sequences or incorporated features related to verbal transcripts. Works should examine how virtual speech sequences differ from face-to-face ones and investigate how they predict turn-taking dynamics. Researchers should also extract additional verbal features using transcripts, for example features using Linguistic Inquiry and Word Count [107]. Relatedly, though we did not encode virtual context characteristics, works should sample turn-taking behaviors in drastically-different virtual spaces and curate approaches for parameterizing them. Researchers should also explore whether incorporating gaze-related features enhances model performance since our dataset did not contain them. To balance model performance, complexity, and interpretability, we used summary statistics and standard machine learning models. We imagine works exploring other deep learning architectures [117, 120] to capture additional nuances of group dynamics. Another avenue of future work lies in benchmarking the real-time performance of these models and leveraging predictive models to reduce miscommunication [2] and mitigate verbal harassment [25, 96].

In our dataset, students gathered in fixed groups using the same social VR platform and embodied avatars that lacked nonverbal cues such as facial expressions. To fully evaluate model robustness, researchers should examine different social platforms with varying levels of avatar representation [100] and immersion [1], change group membership over time [93], and vary group sizes [36]. As noted in Section 5.1.1, differences in group sizes beyond the scope of our paper could yield drastic

differences in social dynamics and lower the applicability of our approach. Different activity types, for example those varying in virtual mobility (e.g., scavenger hunt vs. discussion), could elicit similar concerns. Future research should therefore probe the limits to improving performance through naively scaling up data and sampling across a more diverse set of group interactions (e.g., group size, activity type). It may also be necessary to consider different machine learning architectures for modeling certain group activities, for example training expert sub-models to handle different social scenarios before integrating them into for making predictions [61]. Doing so will not only suggest technical improvements to the applicability of predicting turn-taking behaviors but also built toward a better understanding of how VR contexts influence social dynamics.

One other limitation pertains to the dataset's convenience sample of university students. To investigate the generalizability of our insights, researcher should study more representative demographics. Additionally, discrepancies between virtual and physical motion (e.g., physically seated users with standing virtual avatars) warrant scrutiny for their impact on turn-taking behaviors. Finally, while we examined three turn-transition categories, we excluded backchanneling. As backchanneling and overlap speech make up a key component to social dynamics [95] and have unique challenges in CMC due to latencies and diminished nonverbal behaviors [98–100], works should investigate overlapping speech, for example through predicting unwanted interruptions.

7 Conclusions

Being able to predict turn taking in VR affords opportunities for understanding immersive social interactions and enables systems to administer support and intervention. In this work, we studied turn-taking behaviors of student engaging in open-ended group activities in VR over four weeks. We predicted turn-taking behaviors using features describing individual and group characteristics and extracted features concerning speech-related behaviors, egocentric motion, and dyadic and group relationships. We found that gradient boosting classifiers achieved the best performance, considerably better than prediction by chance. Additional analysis revealed that listener personality, group size, group visual attention, and listener and previous speaker's head pitch, head y-axis position, and left hand y-axis position were key features affecting performance and predictions. Our results suggest that these features are reliable indicators, as models were robust when evaluated on unseen activities, weeks, and groups. Taken together, our work contributed a better understanding of how tracking data and individual and group characteristics can predict VR turn taking. We believe our insights will motivate research on modeling social dynamics, and support practitioners to use behavioral predictions to deliver assistance and training in facilitating immersive social interactions.

Acknowledgments

We thank Cindy Chen, Mark Roman Miller, Monique Santoso, and Yujie Tao for their feedback. This work was supported by the Sozo Graduate Fellowship.

References

- [1] Ahsan Abdullah, Jan Kolkmeier, Vivian Lo, and Michael Neff. 2021. Videoconference and Embodied VR: Communication Patterns Across Task and Medium. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 453 (oct 2021), 29 pages. <https://doi.org/10.1145/3479597>
- [2] Daniel Akselrad, Cyan DeVeaux, Eugy Han, Mark Roman Miller, and Jeremy N. Bailenson. 2023. Body Crumple, Sound Intrusion, and Embodiment Violation: Toward a Framework for Miscommunication in VR. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) (CSCW '23 Companion). Association for Computing Machinery, New York, NY, USA, 122–125. <https://doi.org/10.1145/3584931.3606968>

- [3] Franco Amati and Susan E Brennan. 2018. Eye gaze as a cue for recognizing intention and coordinating joint action. *Eye-tracking in interaction* (2018), 21–46.
- [4] Nicole Anderson. 2021. Accessibility challenges of video conferencing technology. In *Design, Operation and Evaluation of Mobile Communications: Second International Conference, MOBILE 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings* 23. Springer, 185–194.
- [5] Sean Andrist, Wesley Collier, Michael Gleicher, Bilge Mutlu, and David Shaffer. 2015. Look together: Analyzing gaze coordination with epistemic network analysis. *Frontiers in psychology* 6 (2015), 144911.
- [6] Oya Aran and Daniel Gatica-Perez. 2013. One of a kind: inferring personality impressions in meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (Sydney, Australia) (ICMI '13). Association for Computing Machinery, New York, NY, USA, 11–18. <https://doi.org/10.1145/2522848.2522859>
- [7] Lidia Auret and Chris Aldrich. 2012. Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering* 35 (2012), 27–42.
- [8] Jeremy N Bailenson and Nick Yee. 2005. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science* 16, 10 (2005), 814–819.
- [9] Miguel Barreda-Ángeles, Sophie Horneber, and Tilo Hartmann. 2023. Easily applicable social virtual reality and social presence in online higher education during the covid-19 pandemic: A qualitative study. *Computers & Education: X Reality* 2 (2023), 100024. <https://doi.org/10.1016/j.cexr.2023.100024>
- [10] Camiel J. Beukeboom, Martin Tanis, and Ivar E. Vermeulen. 2013. The Language of Extraversion: Extraverted People Talk More Abstractly, Introverts Are More Concrete. *Journal of Language and Social Psychology* 32, 2 (2013), 191–201. <https://doi.org/10.1177/0261927X12460844>
- [11] Rok Blagus and Lara Lusa. 2017. Gradient boosting for high-dimensional prediction of rare events. *Computational Statistics & Data Analysis* 113 (2017), 19–37. <https://doi.org/10.1016/j.csda.2016.07.016>
- [12] Andrea Bönsch, Alexander R. Bluhm, Jonathan Ehret, and Torsten W. Kuhlen. 2020. Inferring a User's Intent on Joining or Passing by Social Groups. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) (IVA '20). Association for Computing Machinery, New York, NY, USA, Article 10, 8 pages. <https://doi.org/10.1145/3383652.3423862>
- [13] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.
- [14] John E Brough, Maxim Schwartz, Satyandra K Gupta, Davinder K Anand, Robert Kavetsky, and Ralph Pettersen. 2007. Towards the development of a virtual environment-based training system for mechanical assembly operations. *Virtual reality* 11 (2007), 189–206.
- [15] Estelle Campione and Jean Véronis. 2002. A large-scale multilingual study of silent pause duration. In *Proc. Speech Prosody 2002*. 199–202. <https://doi.org/10.21437/SpeechProsody.2002-35>
- [16] Hancheng Cao, Vivian Yang, Victor Chen, Yu Jin Lee, Lydia Stone, N'godjigui Junior Diarrassouba, Mark E. Whiting, and Michael S. Bernstein. 2021. My Team Will Go On: Differentiating High and Low Viability Teams through Team Interaction. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 230 (jan 2021), 27 pages. <https://doi.org/10.1145/3432929>
- [17] Jiadong Chen, Chenghao Gu, Jiayi Zhang, Zhankun Liu, and Shin 'ichi Konomi. 2024. Sensing the Intentions to Speak in VR Group Discussions. *Sensors* 24, 2 (2024), 362.
- [18] Mathieu Chollet, Pranav Ghate, Catherine Neubauer, and Stefan Scherer. 2018. Influence of Individual Differences when Training Public Speaking with Virtual Audiences. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) (IVA '18). Association for Computing Machinery, New York, NY, USA, 1–7. <https://doi.org/10.1145/3267851.3267874>
- [19] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics* 19 (2018), 1–14.
- [20] Ziedune Degutyte and Arlene Astell. 2021. The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings. *Frontiers in Psychology* 12 (2021), 616471.
- [21] Cyan DeVeaux, David M. Markowitz, Eugy Han, Mark Roman Miller, Jeffrey T. Hancock, and Jeremy N. Bailenson. 0. Presence and Pronouns: An Exploratory Investigation into the Language of Social VR. *Journal of Language and Social Psychology* 0, 0 (0), 0261927X241248646. <https://doi.org/10.1177/0261927X241248646>
- [22] Cyan DeVeaux, David M. Markowitz, Eugy Han, Mark Roman Miller, Jeffrey T. Hancock, and Jeremy N. Bailenson. 2023. Descriptive Linguistic Patterns of Group Conversations in VR. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 785–786. <https://doi.org/10.1109/VRW58643.2023.00235>
- [23] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 1 (2013), 27–46.

- [24] Pedregosa Fabian. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12 (2011), 2825.
- [25] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 85 (apr 2022), 30 pages. <https://doi.org/10.1145/3512932>
- [26] Nancy L Freeman, Adrienne Perry, and James M Bebko. 2002. Behaviour is communication: Nonverbal communicative behaviour in students with autism and instructors' responsivity. *Journal on Developmental Disabilities* 9, 2 (2002), 145–155.
- [27] Guido Friebe, Marie Lalanne, Bernard Richter, Peter Schwarzmann, and Paul Seabright. 2021. Gender differences in social interactions. *Journal of Economic Behavior & Organization* 186 (2021), 33–45.
- [28] Julian Frommel, Valentin Sagl, Ansgar E. Depping, Colby Johanson, Matthew K. Miller, and Regan L. Mandryk. 2020. Recognizing Affiliation: Using Behavioural Traces to Predict the Quality of Social Interactions in Online Games. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3313831.3376446>
- [29] Cesare Furlanello, Markus Neteler, Stefano Merler, Stefano Menegon, Steno Fontanari, Angela Donini, Annapaola Rizzoli, and C Chemini. 2003. GIS and the random forest predictor: Integration in R for tick-borne disease risk assessment. In *Proceedings of DSC*. Citeseer, 2.
- [30] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.
- [31] Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. 2018. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755* (2018).
- [32] Lauren V Hadley, William M Whitmer, W Owen Brimijoin, and Graham Naylor. 2021. Conversation in small groups: Speaking and listening strategies depend on the complexities of the environment and group. *Psychonomic bulletin & review* 28 (2021), 632–640.
- [33] Cathy W Hall, Andrea D Peterson, Raymond E Webster, Larry M Bolen, and Michael B Brown. 1999. Perception of nonverbal social cues by regular education, ADHD, and ADHD/LD students. *Psychology in the Schools* 36, 6 (1999), 505–514.
- [34] Eugy Han and Jeremy N. Bailenson. 2024. Lessons for/in virtual classrooms: designing a model for classrooms inside virtual reality. *Communication Education* 73, 2 (2024), 234–243. <https://doi.org/10.1080/03634523.2024.2312879>
- [35] Eugy Han, Cyan DeVeaux, Jeffrey T. Hancock, Nilam Ram, Gabriella M. Harari, and Jeremy N. Bailenson. 2024. The influence of spatial dimensions of virtual environments on attitudes and nonverbal behaviors during social interactions. *Journal of Environmental Psychology* 95 (2024), 102269. <https://doi.org/10.1016/j.jenvp.2024.102269>
- [36] Eugy Han, Mark R Miller, Cyan DeVeaux, Hanseul Jun, Kristine L Nowak, Jeffrey T Hancock, Nilam Ram, and Jeremy N Bailenson. 2023. People, places, and time: a large-scale, longitudinal study of transformed avatars and environmental context in group interaction in the metaverse. *Journal of Computer-Mediated Communication* 28, 2 (01 2023), zmac031. <https://doi.org/10.1093/jcmc/zmac031> arXiv:<https://academic.oup.com/jcmc/article-pdf/28/2/zmac031/48520441/zmac031.pdf>
- [37] Hong Han, Xiaoling Guo, and Hua Yu. 2016. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. 219–224. <https://doi.org/10.1109/ICSESS.2016.7883053>
- [38] A Paul Hare. 1981. Group size. *American Behavioral Scientist* 24, 5 (1981), 695–708.
- [39] Peter A. Heeman and Rebeca Lunsford. 2015. Can overhearers predict who will speak next?. In *Turn-Taking and Coordination in Human-Machine Interaction - Papers from the AAAI Spring Symposium, Technical Report (AAAI Spring Symposium - Technical Report)*. AI Access Foundation, 30–35. Publisher Copyright: Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 2015 AAAI Spring Symposium ; Conference date: 23-03-2015 Through 25-03-2015.
- [40] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J. Quinn. 2021. AdapTutAR: An Adaptive Tutoring System for Machine Tasks in Augmented Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 417, 15 pages. <https://doi.org/10.1145/3411764.3445283>
- [41] Tina Iachini, Yann Coello, Francesca Frassinetti, and Gennaro Ruggiero. 2014. Body space in social interactions: a comparison of reaching and comfort distance in immersive virtual reality. *PloS one* 9, 11 (2014), e111511.
- [42] Tina Iachini, Yann Coello, Francesca Frassinetti, Vincenzo Paolo Senese, Francesco Galante, and Gennaro Ruggiero. 2016. Peripersonal and interpersonal space in virtual and real environments: Effects of gender and age. *Journal of Environmental Psychology* 45 (2016), 154–164.
- [43] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2016. Analyzing mouth-opening transition pattern for predicting next speaker in multi-party meetings. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*

- (Tokyo, Japan) (*ICMI '16*). Association for Computing Machinery, New York, NY, USA, 209–216. <https://doi.org/10.1145/2993148.2993189>
- [44] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2017. Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings. In *Proceedings of the 5th International Conference on Human Agent Interaction* (Bielefeld, Germany) (*HAI '17*). Association for Computing Machinery, New York, NY, USA, 181–187. <https://doi.org/10.1145/3125739.3125765>
- [45] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM Trans. Interact. Intell. Syst.* 6, 1, Article 4 (may 2016), 31 pages. <https://doi.org/10.1145/2757284>
- [46] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Using Respiration to Predict Who Will Speak Next and When in Multiparty Meetings. *ACM Trans. Interact. Intell. Syst.* 6, 2, Article 20 (aug 2016), 20 pages. <https://doi.org/10.1145/2946838>
- [47] Alexei V Ivanov, Giuseppe Riccardi, Adam J Sporka, and Jakub Franc. 2011. Recognition of personality traits from human spoken conversations. In *Twelfth annual conference of the international speech communication association*.
- [48] Allison Jing, Kieran May, Brandon Matthews, Gun Lee, and Mark Billingham. 2022. The Impact of Sharing Gaze Behaviours in Collaborative Mixed Reality. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 463 (nov 2022), 27 pages. <https://doi.org/10.1145/3555564>
- [49] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Trans. Interact. Intell. Syst.* 3, 2, Article 12 (aug 2013), 30 pages. <https://doi.org/10.1145/2499474.2499481>
- [50] Kristiina Jokinen, Masafumi Nishida, and Seiichi Yamamoto. 2009. Eye-gaze experiments for conversation monitoring. In *Proceedings of the 3rd International Universal Communication Symposium* (Tokyo, Japan) (*IUCS '09*). Association for Computing Machinery, New York, NY, USA, 303–308. <https://doi.org/10.1145/1667780.1667843>
- [51] Hanseul Jun, Mark Roman Miller, Fernanda Herrera, Byron Reeves, and Jeremy N. Bailenson. 2022. Stimulus Sampling With 360-Videos: Examining Head Movements, Arousal, Presence, Simulator Sickness, and Preference on a Large Sample of Participants and Videos. *IEEE Transactions on Affective Computing* 13, 3 (2022), 1416–1425. <https://doi.org/10.1109/TAFFC.2020.3004617>
- [52] Onno P. Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction. In *Annual Meeting of the Association for Computational Linguistics*.
- [53] Adam Kendon and Mark Cook. 1969. The consistency of gaze patterns in social interaction. *British Journal of Psychology* 60, 4 (1969), 481–494.
- [54] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and speech* 41, 3-4 (1998), 295–321.
- [55] Nikola Kovačević, Christian Holz, Tobias Günther, Markus Gross, and Rafael Wampfler. 2023. Personality Trait Recognition Based on Smartphone Typing Characteristics in the Wild. *IEEE Transactions on Affective Computing* (2023).
- [56] Raja S. Kushalnagar and Christian Vogler. 2020. Teleconference Accessibility and Guidelines for Deaf and Hard of Hearing Users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 9, 6 pages. <https://doi.org/10.1145/3373625.3417299>
- [57] Grete Helena Kütt, Teerapaun Tanprasert, Jay Rodolitz, Bernardo Moyza, Samuel So, Georgia Kenderova, and Alexandra Papoutsaki. 2020. Effects of Shared Gaze on Audio- Versus Text-Based Remote Collaborations. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 136 (oct 2020), 25 pages. <https://doi.org/10.1145/3415207>
- [58] Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal Turn Analysis and Prediction for Multi-party Conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 436–444. <https://doi.org/10.1145/3577190.3614139>
- [59] Sinae Lee, Jangwoon Park, and Dugan Um. 2021. Speech Characteristics as Indicators of Personality Traits. *Applied Sciences* 11, 18 (2021). <https://doi.org/10.3390/app11188776>
- [60] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. 2010. Employing social gaze and speaking activity for automatic determination of the Extraversion trait. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (Beijing, China) (*ICMI-MLMI '10*). Association for Computing Machinery, New York, NY, USA, Article 7, 8 pages. <https://doi.org/10.1145/1891903.1891913>
- [61] Bin Lin, Zhenyu Tang, Yang Ye, Jiayi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947* (2024).

- [62] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) (UIST '19). Association for Computing Machinery, New York, NY, USA, 147–160. <https://doi.org/10.1145/3332165.3347945>
- [63] Shimeng Liu, Yoshitaka Nakajima, Lihan Chen, Sophia Arndt, Maki Kakizoe, Mark A Elliott, and Gerard B Remijn. 2022. How pause duration influences impressions of English speech: Comparison between native and non-native speakers. *Frontiers in psychology* 13 (2022), 778018.
- [64] Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).
- [65] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).
- [66] Tiffany Luong and Christian Holz. 2022. Characterizing physiological responses to fear, frustration, and insight in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3917–3927.
- [67] Torsten Maier, Nicolas F Soria Zurita, Elizabeth Starkey, Daniel Spillane, Christopher McComb, and Jessica Menold. 2022. Comparing human and cognitive assistant facilitated brainstorming sessions. *Journal of Engineering Design* 33, 4 (2022), 259–283.
- [68] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. "Talking without a Voice": Understanding Non-verbal Communication in Social Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 175 (oct 2020), 25 pages. <https://doi.org/10.1145/3415246>
- [69] Marianne Schmid Mast. 2002. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research* 28, 3 (2002), 420–450.
- [70] Mark Roman Miller, Cyan DeVeaux, Eugy Han, Nilam Ram, and Jeremy N. Bailenson. 2023. A Large-Scale Study of Proxemics and Gaze in Groups. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 409–417. <https://doi.org/10.1109/VR55154.2023.00056>
- [71] Mark Roman Miller, Eugy Han, Cyan DeVeaux, Eliot Jones, Ryan Chen, and Jeremy N Bailenson. 2023. A large-scale study of personal identifiability of virtual reality motion over time. *arXiv preprint arXiv:2303.01430* (2023).
- [72] Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A Landay, and Jeremy N Bailenson. 2020. Personal identifiability of user tracking data during observation of 360-degree VR video. *Scientific Reports* 10, 1 (2020), 17404.
- [73] Mark Roman Miller, Neeraj Sonalkar, Ade Mabogunje, Larry Leifer, and Jeremy Bailenson. 2021. Synchrony within Triads using Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 400 (oct 2021), 27 pages. <https://doi.org/10.1145/3479544>
- [74] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. 2022. Temporal effects in motion behavior for virtual reality (vr) biometrics. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 563–572.
- [75] Alec G. Moore, Ryan P. McMahan, Hailiang Dong, and Nicholas Ruozzi. 2021. Personal Identifiability and Obfuscation of User Tracking Data From VR Training Sessions. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 221–228. <https://doi.org/10.1109/ISMAR52148.2021.00037>
- [76] Fares Moustafa and Anthony Steed. 2018. A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (Tokyo, Japan) (VRST '18). Association for Computing Machinery, New York, NY, USA, Article 22, 10 pages. <https://doi.org/10.1145/3281505.3281527>
- [77] Arne Nagels, Tilo Kircher, Miriam Steines, and Benjamin Straube. 2015. Feeling addressed! The role of body orientation and co-speech gesture in social communication. *Human brain mapping* 36, 5 (2015), 1925–1936.
- [78] Vivek Nair, Wenbo Guo, Justus Mattern, Rui Wang, James F O'Brien, Louis Rosenberg, and Dawn Song. 2023. Unique identification of 50,000+ virtual reality users from head & hand motion data. In *32nd USENIX Security Symposium (USENIX Security 23)*. 895–910.
- [79] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) (IUI '10). Association for Computing Machinery, New York, NY, USA, 139–148. <https://doi.org/10.1145/1719970.1719990>
- [80] Nasser Nassiri, Norman Powell, and David Moore. 2010. Human interactions and personal space in collaborative virtual environments. *Virtual reality* 14 (2010), 229–240.
- [81] Alexey Natekin and Alois Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* 7 (2013), 21.
- [82] Kristin K Nicodemus. 2011. On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics* 12, 4 (2011), 369–373.
- [83] Catherine Oh Kruzic, David Kruzic, Fernanda Herrera, and Jeremy Bailenson. 2020. Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments. *Scientific reports* 10, 1 (2020), 20626.
- [84] Jangwoon Park, Sinae Lee, Kimberly Brotherton, Dugan Um, and Jaehyun Park. 2020. Identification of Speech Characteristics to Distinguish Human Personality of Introversive and Extroversive Male Groups. *International Journal*

- of *Environmental Research and Public Health* 17, 6 (2020). <https://doi.org/10.3390/ijerph17062125>
- [85] Kevin C Parker. 1988. Speaking turns in small group interaction: A context-sensitive event sequence model. *Journal of Personality and Social Psychology* 54, 6 (1988), 965.
 - [86] Zhongling Pi, Ke Xu, Caixia Liu, and Jiumin Yang. 2020. Instructor presence in video lectures: Eye gaze matters, but not body orientation. *Computers & Education* 144 (2020), 103713.
 - [87] Huajian Qiu, Paul Strel, Tiffany Luong, Christoph Gebhardt, and Christian Holz. 2023. ViGather: Inclusive Virtual Conferencing with a Joint Experience Across Traditional Screen Devices and Mixed Reality Headsets. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 232 (sep 2023), 27 pages. <https://doi.org/10.1145/3604279>
 - [88] Anna C. M. Queiroz, Géraldine Fauville, Adina T. Abeles, Aaron Levett, and Jeremy N. Bailenson. 2023. The Efficacy of Virtual Reality in Climate Change Education Increases with Amount of Body Movement and Message Specificity. *Sustainability* 15, 7 (2023). <https://doi.org/10.3390/su15075814>
 - [89] Stéphane Raffard, Robin N Salesse, Catherine Bortolon, Benoît G Bardy, José Henriques, Ludovic Marin, Didier Stricker, and Delphine Capdevielle. 2018. Using mimicry of body movements by a virtual agent to increase synchronization behavior and rapport in individuals with schizophrenia. *Scientific reports* 8, 1 (2018), 17356.
 - [90] Ronald W Ramsay. 1968. Speech patterns and personality. *Language and Speech* 11, 1 (1968), 54–63.
 - [91] Daniel Roth, Gary Bente, Peter Kullmann, David Mal, Chris Felix Purps, Kai Vogeley, and Marc Erich Latoschik. 2019. Technologies for Social Augmentations in User-Embodied Virtual Reality. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology* (Parramatta, NSW, Australia) (VRST '19). Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. <https://doi.org/10.1145/3359996.3364269>
 - [92] Daniel Roth, Constantin Klehnbeck, Tobias Feigl, Christopher Mutschler, and Marc Erich Latoschik. 2018. Beyond Replication: Augmenting Social Behaviors in Multi-User Virtual Realities. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 215–222. <https://doi.org/10.1109/VR.2018.8447550>
 - [93] Niloufar Salehi and Michael S. Bernstein. 2018. Hive: Collective Design Through Network Rotation. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 151 (nov 2018), 26 pages. <https://doi.org/10.1145/3274420>
 - [94] Arissa J. Sato, Zefan Sramek, and Koji Yatani. 2023. Groupnamics: Designing an Interface for Overviewing and Managing Parallel Group Discussions in an Online Classroom. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 701, 18 pages. <https://doi.org/10.1145/3544548.3581322>
 - [95] Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society* 29, 1 (2000), 1–63.
 - [96] Kelsea Schulenberg, Guo Freeman, Lingyuan Li, and Catherine Barwulor. 2023. "Creepy Towards My Avatar Body, Creepy Towards My Body": How Women Experience and Manage Harassment Risks in Social Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 236 (oct 2023), 29 pages. <https://doi.org/10.1145/3610027>
 - [97] Azin Semsar, Hannah McGowan, Yuanyuan Feng, H. Reza Zahiri, Adrian Park, Andrea Kleinsmith, and Helena M. Mentis. 2020. Quality of and Attention to Instructions in Telementoring. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 165 (oct 2020), 21 pages. <https://doi.org/10.1145/3415236>
 - [98] Lucas M. Seuren, Joseph Wherton, Trisha Greenhalgh, and Sara E. Shaw. 2021. Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics* 172 (2021), 63–78. <https://doi.org/10.1016/j.pragma.2020.11.005>
 - [99] Yu Sheng. 2021. How the function of video-conference software, Zoom, interfere with turn-taking in the online classroom?. In *2021 International Conference on Education, Information Management and Service Science (EIMSS)*. 209–215. <https://doi.org/10.1109/EIMSS53851.2021.00052>
 - [100] Harrison Jesse Smith and Michael Neff. 2018. Communication Behavior in Embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173863>
 - [101] Mitchell S Sommers and Joe Barcroft. 2006. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America* 119, 4 (2006), 2406–2416.
 - [102] Philipp Stark, Alexander J. Jung, Jens-Uwe Hahn, Enkelejda Kasneci, and Richard Göllner. 2024. Using Gaze Transition Entropy to Detect Classroom Discourse in a Virtual Reality Classroom. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications (ETRA '24)*. Association for Computing Machinery, New York, NY, USA, Article 23, 11 pages. <https://doi.org/10.1145/3649902.3653335>
 - [103] Frank Steinicke, Nale Lehmann-Willenbrock, and Annika Luisa Meinecke. 2020. A First Pilot Study to Compare Virtual Group Meetings using Video Conferences and (Immersive) Virtual Reality. In *Proceedings of the 2020 ACM Symposium on Spatial User Interaction* (Virtual Event, Canada) (SUI '20). Association for Computing Machinery, New York, NY, USA, Article 19, 2 pages. <https://doi.org/10.1145/3385959.3422699>

- [104] Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the P value is not enough. *Journal of graduate medical education* 4, 3 (2012), 279–282.
- [105] Yilu Sun, Omar Shaikh, and Andrea Stevenson Won. 2019. Nonverbal synchrony in virtual reality. *PloS one* 14, 9 (2019), e0221803.
- [106] Bronwyn Tarr, Mel Slater, and Emma Cohen. 2018. Synchrony and social connection in immersive virtual reality. *Scientific reports* 8, 1 (2018), 3693.
- [107] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
- [108] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) (CHI '01). Association for Computing Machinery, New York, NY, USA, 301–308. <https://doi.org/10.1145/365024.365119>
- [109] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R Schinazi, Markus Gross, and Christian Holz. 2022. Affective state prediction from smartphone touch and sensor data in the wild. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [110] Portia Wang and Jeremy Bailenson. [n. d.]. Virtual reality as a research tool. In *Routledge Handbook of Communication and Social Cognition*. Routledge/Taylor and Francis.
- [111] Portia Wang, Mark R. Miller, and Jeremy N. Bailenson. 2023. The Belated Guest: Exploring the Design Space for Transforming Asynchronous Social Interactions in Virtual Reality. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 617–618. <https://doi.org/10.1109/VRW58643.2023.00151>
- [112] Portia Wang, Mark R Miller, Eugy Han, Cyan DeVeaux, and Jeremy N Bailenson. 2024. Understanding virtual design behaviors: A large-scale analysis of the design process in Virtual Reality. *Design Studies* 90 (2024), 101237.
- [113] Portia Wang, Mark R. Miller, Anna C.M. Queiroz, and Jeremy N. Bailenson. 2024. Socially Late, Virtually Present: The Effects of Transforming Asynchronous Social Interactions in Virtual Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3613904.3642244>
- [114] Gary L Wells and Paul D Windschitl. 1999. Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin* 25, 9 (1999), 1115–1125.
- [115] Julie Williamson, Jie Li, Vinoba Vinayagamoorthy, David A. Shamma, and Pablo Cesar. 2021. Proxemics and Social Interactions in an Instrumented Virtual Reality Workshop. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 253, 13 pages. <https://doi.org/10.1145/3411764.3445729>
- [116] Jue Wu, Junyi Ma, Yasha Wang, and Jiangtao Wang. 2021. Understanding and Predicting the Burst of Burnout via Social Media. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 265 (jan 2021), 27 pages. <https://doi.org/10.1145/3434174>
- [117] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2021), 4–24. <https://doi.org/10.1109/TNNLS.2020.2978386>
- [118] Fengli Xu, Zhenyu Han, Jinghua Piao, and Yong Li. 2019. "I Think You'll Like It": Modelling the Online Purchase Behavior in Social E-commerce. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 65 (nov 2019), 23 pages. <https://doi.org/10.1145/3359167>
- [119] Shengping Yang and Gilbert Berdine. 2017. The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles* 5, 19 (2017), 34–36.
- [120] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.

A Descriptions of Weekly Group Activities

Table 6. Descriptions of Weekly Group Discussion and Activities.

Week	Topic	Discussion Agenda Items & Prompts
1	Accessibility	• Introduce yourself (i.e., name, year, major, what are you most excited about learning in this class, favorite thing you have done in VR) • Talk through the preliminary ideas for your “Built VR world/scene” project • Discuss accessibility within the context of ENGAGE (e.g., what are the constraints?) • List things that ENGAGE does well vs. does not do well (e.g., using sticky notes)
2	Avatars	• Consider the templates of storyboards we’ve provided for your storyboard assignment. What are some elements you are considering including in your storyboard? How do you plan on using the affordances unique to VR, such as presence, the ability to move around in 3D space, spatialized sound, etc.? Are you planning on showcasing this in your storyboard? • Reimagine what your avatar would look like. Either draw an avatar that you wish represents you or an avatar you would like to embody. This can, but doesn’t have to, be a human avatar. • Show-and-tell for created designs
3	Medical	• Consider the medical applications we learned in the readings, class, and meditation AltspaceVR journey. What was the most surprising, promising, or concerning? How does this class change your perception of using VR for medical purposes? • Collaboratively work with your group members to create a meditation room or a safe space using any of the ENGAGE tools (e.g., 3D pen, IFX, sticky notes)
4	Education	• Consider a target audience/population (e.g., students of a certain age group, students with a certain learning disability, older students). • Consider a goal (e.g., retaining factual information, having students experience something) • Consider a topic of interest (e.g., language, STEM, social skills) • Empathize, define, ideate, and prototype an application tailored to your audience, goal, and topic. Have a member of your group test out/role-play a student using the application

B Feature Importance for All Features

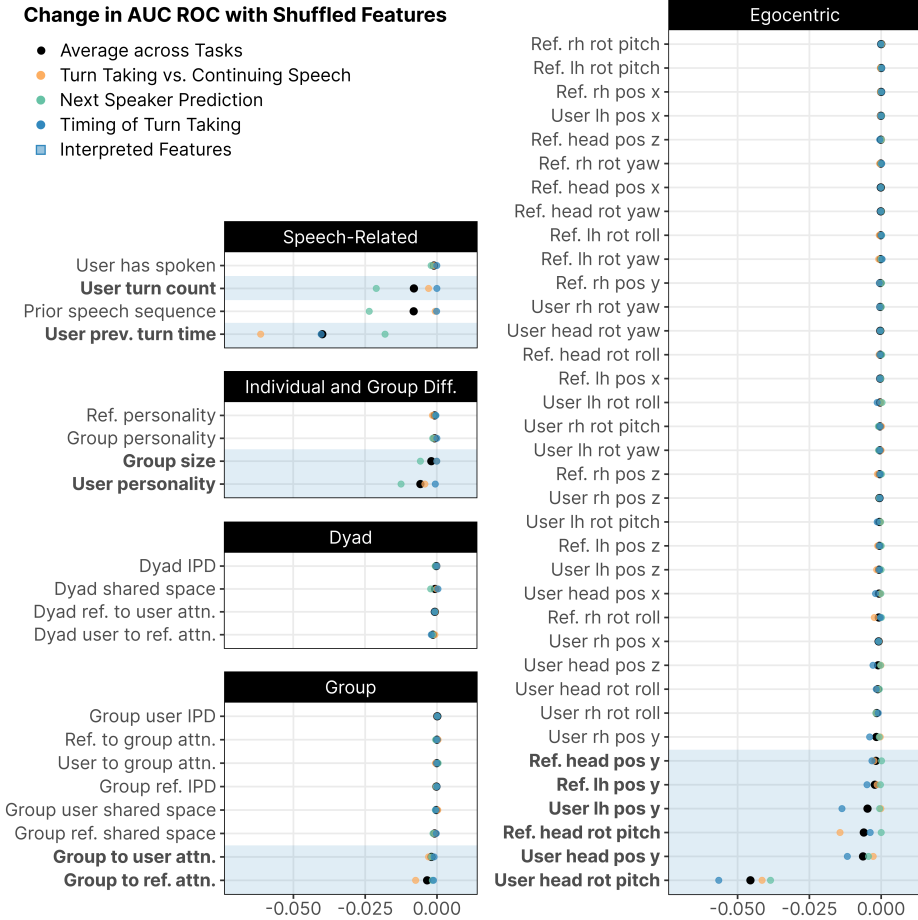


Fig. 6. Feature Importance Analysis. Each panel reports the change in AUC of the ROC after shuffling features across the three prediction tasks for a given feature group. Black dots denote the average drop in accuracy across three tasks. Features within each panel are ordered in increasing average drop in accuracy from top to bottom. Bolded features and their corresponding rows with light blue backgrounds indicate features that are further interpreted in Section 4.3. Diff. = differences; Ref. = reference user; prev. = previous; IPD = interpersonal distance; attn. = visual attention; lh = left hand; rh = right hand.

C Wald Tests for Logistic Regression Models

We present results from joint Wald-tests ran on the trained logistic regression models, specifically the p values from the joint hypothesis tests that all features related to a specific construct do not significantly contribute to the predictive model. The orders of the feature groups and constructs follow that of Appendix B. Features with statistical significance, evaluated at $\alpha = .05$, are bolded.

Table 7. Results from Wald-tests on Speech-Related Features. We report the p values from the joint hypothesis tests. prev. = previous.

	Turn Taking vs. Continuing Speech	Next Speaker Prediction	Timing of Turn Taking
User has spoken	<.001	<.001	.264
User turn count	.046	<.001	.004
Prior speech sequence	<.001	<.001	<.001
User prev. turn time	<.001	.046	<.001

Table 8. Results from Wald-tests on Individual and Group Differences Features. We report the p values from the joint hypothesis tests. Ref. = reference user.

	Turn Taking vs. Continuing Speech	Next Speaker Prediction	Timing of Turn Taking
Ref. personality	.020	<.001	<.001
Group personality	.189	<.001	.191
Group size	.001	<.001	.646
User personality	.001	<.001	.189

Table 9. Results from Wald-tests on Dyad Features. We report the p values from the joint hypothesis tests. Ref. = reference user; IPD = interpersonal distance; attn. = visual attention.

	Turn Taking vs. Continuing Speech	Next Speaker Prediction	Timing of Turn Taking
Dyad IPD	.004	.066	.095
Dyad shared space	.177	.096	.628
Dyad ref. to user attn.	.846	<.001	.057
Dyad user to ref. attn.	.246	<.001	.087

Table 10. Results from Wald-tests on Group Features. We report the *p* values from the joint hypothesis tests. Ref. = reference user; IPD = interpersonal distance; attn. = visual attention.

	Turn Taking vs. Continuing Speech	Next Speaker Prediction	Timing of Turn Taking
Group user IPD	<.001	.588	.276
Ref. to group attn.	<.001	<.001	.016
User to group attn.	<.001	.298	.002
Group ref. IPD	.009	.347	.056
Group user shared space	.017	.982	.332
Group ref. shared space	<.001	.784	.292
Group to user attn.	<.001	.080	.007
Group to ref. attn.	<.001	<.001	.070

Table 11. Results from Wald-tests on Egocentric Features. We report the p values from the joint hypothesis tests. Ref. = reference user; prev. = previous; IPD = inter-personal distance; attn. = visual attention; lh = left hand; rh = right hand.

	Turn Taking vs. Continuing Speech	Next Speaker Prediction	Timing of Turn Taking
Ref. rh rot pitch	.028	.964	.297
Ref. lh rot pitch	.195	.993	.014
Ref. rh pos x	.143	.997	.007
User lh pos x	.015	.181	.566
Ref. head pos z	.018	.588	.505
Ref. rh rot yaw	<.001	.984	.187
Ref. head pos x	.154	.541	.190
Ref. head rot yaw	.202	.823	.277
Ref. lh rot roll	.252	.991	.766
Ref. lh rot yaw	.470	.967	.197
Ref. rh pos y	.493	.992	.103
User rh rot yaw	.002	.017	.173
User head rot yaw	.001	.590	.343
Ref. head rot roll	.058	.947	.191
Ref. lh pos x	.007	.994	.027
User lh rot roll	.961	.316	.043
User rh rot pitch	.564	.225	.665
User lh rot yaw	.012	.052	.035
Ref. rh pos z	.646	.999	.149
User rh pos z	.080	.816	.595
User lh rot pitch	.110	.161	.440
Ref. lh pos z	.198	.977	.057
User lh pos z	.185	.246	.021
User head pos x	.111	.086	.054
Ref. rh rot roll	<.001	.971	.090
User rh pos x	.048	.215	.558
User head pos z	.719	.052	.219
User head rot roll	.003	<.001	<.001
User rh rot roll	.002	<.001	.027
User rh pos y	.035	.082	.230
Ref. head pos y	<.001	.605	<.001
Ref. lh pos y	.102	.942	.213
User lh pos y	.002	.027	.087
Ref. head rot pitch	<.001	.640	.040
User head pos y	<.001	.015	<.001
User head rot pitch	<.001	<.001	<.001

D Additional Results on Model Predictions

In this section, we present additional analyses on prediction performance across sessions, groups, and weeks. To do this, we first randomly split each of the three datasets, one for each task, into a training set with 90% of the data and a testing set containing the remaining 10%. Then, we built models on the training data and evaluated them on the testing set in three ways: across sessions, groups, and weeks. Specifically, when evaluating across weeks, we calculated the average AUC and standard error given the trained models’ performances across four test sets, each containing the testing data from a specific week. We repeated this procedure for extracting the AUC and standard errors of the models across sessions and groups. Note that the reported average AUC here gives equal weight to each session, group, and week, which differs from the cross-validation evaluations presented in Section 4.1.

Table 12. Model Performance (measured as the AUC of the ROC curve) on Predicting Turn-Taking Behavior vs. Continuing Speech. We report the average AUC and standard error on the testing set partitioned based on sessions, groups, and weeks. An AUC of 0.50 means that the model’s ability to distinguish between positive and negative samples is no better than random chance. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metrics		
	By Session	By Group	By Week
Logistic Regression	0.70 (0.02)	0.71 (0.02)	0.73 (0.01)
MLP Classifier	0.71 (0.02)	0.70 (0.02)	0.72 (0.02)
Random Forest Classifier	0.76 (0.01)	0.77 (0.01)	0.79 (0.01)
Gradient Boosting Classifier	0.80 (0.01)	0.81 (0.01)	0.82 (0.01)

Table 13. Model Performance (measured as the AUC of the ROC curve) on Next Speaker Prediction. We report the average AUC and standard error on the testing set partitioned based on sessions, groups, and weeks. An AUC of 0.50 means that the model’s ability to distinguish between positive and negative samples is no better than random chance. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metrics		
	By Session	By Group	By Week
Logistic Regression	0.72 (0.02)	0.72 (0.02)	0.72 (0.01)
MLP Classifier	0.72 (0.02)	0.72 (0.02)	0.72 (0.01)
Random Forest Classifier	0.75 (0.02)	0.75 (0.02)	0.75 (0.00)
Gradient Boosting Classifier	0.75 (0.02)	0.76 (0.02)	0.77 (0.00)

Table 14. Model Performance (measured as the AUC of the ROC curve) on the Timing of Turn Taking. We report the average AUC and standard error on the testing set partitioned based on sessions, groups, and weeks. An AUC of 0.50 means that the model’s ability to distinguish between positive and negative samples is no better than random chance. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metrics		
	By Session	By Group	By Week
Logistic Regression	0.61 (0.02)	0.62 (0.02)	0.62 (0.01)
MLP Classifier	0.64 (0.01)	0.64 (0.02)	0.64 (0.021)
Random Forest Classifier	0.69 (0.01)	0.68 (0.02)	0.69 (0.01)
Gradient Boosting Classifier	0.72 (0.02)	0.71 (0.02)	0.73 (0.01)

E Additional Results on Predicting Continuing Speech and Next Speaker

While our formulation of turn-taking behavior prediction tasks regarded turn-taking behavior vs. continuing speech and next speaker prediction as two individual tasks, we further present an alternative task that aims to predict speaking intentions more generally. This task holds practical importance as it addresses the broader question of forecasting who will speak next, be it speaker continuation or a switch to a new speaker. Importantly, this task differs from that of turn-taking behavior vs. continuing as the former seeks to distinguish between upcoming speakers and users who will not speak, whereas the latter focuses on distinguishing between two types of upcoming speakers.

Specifically, we defined positive samples as those sampled immediately preceding a continuing speech or clean turn taking speech event. For clean turn-taking behaviors, we chose the main user as the upcoming speaker and the previous speaker as the reference user. For samples corresponding to continuing speech, we chose the upcoming speaker as both the main and reference user. Negative samples were selected also at moments immediately preceding either a clean turn taking or continuing speech event. For the negative samples, we set the reference user as the previous speaker and the main user as one of the users who will not be the upcoming speaker. Across all samples, when the main user and the reference user are chosen to be the same user, we set the dyad-related features to zero after standardization. We down-sampled the training data such that there is an equal number of positive and negative data. Similar to Section 4.1, we report in Table 15 the AUC for the four machine learning models across the performance metrics introduced in Section 3.5.

We extended our binary classification models to multiclass prediction. Specifically, we evaluated each instance before a turn-taking or continuing speech event to predict which group member, including the current speaker, would speak next. Unlike our previous individual-based predictions, this formulation focuses on group-level predictions. Using the trained binary models from the previous step, we evaluated each user in the scene (i.e., chosen user as the main user, previous speaker as the reference user) and selected the user with the highest probability of a positive label as the predicted next speaker. Comparing the predictions with the true next speakers, we calculated the testing data accuracy and further reported separately the accuracies of the testing data with three-person groups and four-person groups. When presenting accuracies for cross-validation evaluations, we also report their standard errors across all folds. Tables 16–18 summarize our results.

Table 15. Binary Classification Model Performance (measured as the AUC of the ROC curve) on Predicting Continuing and Speech Next Speaker. For metrics using cross validation, which we denote using the subscript *cv*, we report the average and standard error across all folds. An AUC of 0.50 means that the model’s ability to distinguish between positive and negative samples is no better than random chance. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metrics			
	Session _{cv}	Group _{cv}	Week _{cv}	Week 4
Logistic Regression	0.74 (0.01)	0.73 (0.01)	0.74 (0.01)	0.73
MLP Classifier	0.75 (0.00)	0.74 (0.01)	0.74 (0.01)	0.74
Random Forest Classifier	0.79 (0.00)	0.79 (0.01)	0.79 (0.01)	0.79
Gradient Boosting Classifier	0.82 (0.00)	0.81 (0.01)	0.82 (0.00)	0.81

Table 16. Multiclass Classification on Continuing Speech and Next Speaker Prediction using Session-Based Cross-Validation. We report accuracy across testing samples and, in parenthesis, the standard errors of the accuracies across all folds. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metric: Session _{cv}		
	All	Three-Person	Four-Person
Logistic Regression	0.51 (0.01)	0.54 (0.01)	0.48 (0.01)
MLP Classifier	0.54 (0.01)	0.57 (0.01)	0.51 (0.01)
Random Forest Classifier	0.57 (0.01)	0.60 (0.01)	0.54 (0.01)
Gradient Boosting Classifier	0.62 (0.01)	0.65 (0.01)	0.59 (0.01)

Table 17. Multiclass Classification on Continuing Speech and Next Speaker Prediction using Group-Based Cross-Validation. We report accuracy across testing samples and, in parenthesis, the standard errors of the accuracies across all folds. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metric: Group _{cv}		
	All	Three-Person	Four-Person
Logistic Regression	0.51 (0.01)	0.53 (0.02)	0.49 (0.02)
MLP Classifier	0.52 (0.01)	0.55 (0.01)	0.50 (0.02)
Random Forest Classifier	0.57 (0.01)	0.58 (0.02)	0.55 (0.02)
Gradient Boosting Classifier	0.61 (0.01)	0.63 (0.01)	0.59 (0.01)

Table 18. Multiclass Classification on Continuing Speech and Next Speaker Prediction using Week-Based Cross-Validation. We report accuracy across testing samples and, in parenthesis, the standard errors of the accuracies across all folds. Bolded numbers denote best performance by metric.

Prediction Model	Performance Metric: Week _{cv}		
	All	Three-Person	Four-Person
Logistic Regression	0.51 (0.01)	0.54 (0.01)	0.48 (0.01)
MLP Classifier	0.54 (0.01)	0.57 (0.01)	0.49 (0.00)
Random Forest Classifier	0.57 (0.01)	0.60 (0.01)	0.54 (0.00)
Gradient Boosting Classifier	0.62 (0.01)	0.65 (0.01)	0.58 (0.00)

As shown in Table 15, gradient boosting classifiers yielded the highest performance with an AUC of 0.81–0.82 across the four performance metrics. Similar to results presented in Section 4.1, the random forest classifiers yielded the second best performance with an AUC of 0.79, followed by the MLP classifiers and the logistic regressions at 0.74–0.75 AUC and 0.73–0.74 AUC. Tables 16–18 revealed similar trends, namely that the gradient boosting classifiers yielded the highest accuracy across all evaluation metrics (i.e., 59–65%). We also found that the accuracies among three-person groups are higher than that for four-person groups. For reference, random chance predictions would result in an accuracy of 33.33% for three-person groups and 25.00% for four-person groups.

F Two-Variable Partial Dependence Plots for Maximum Upward and Downward Speed for Head Pitch

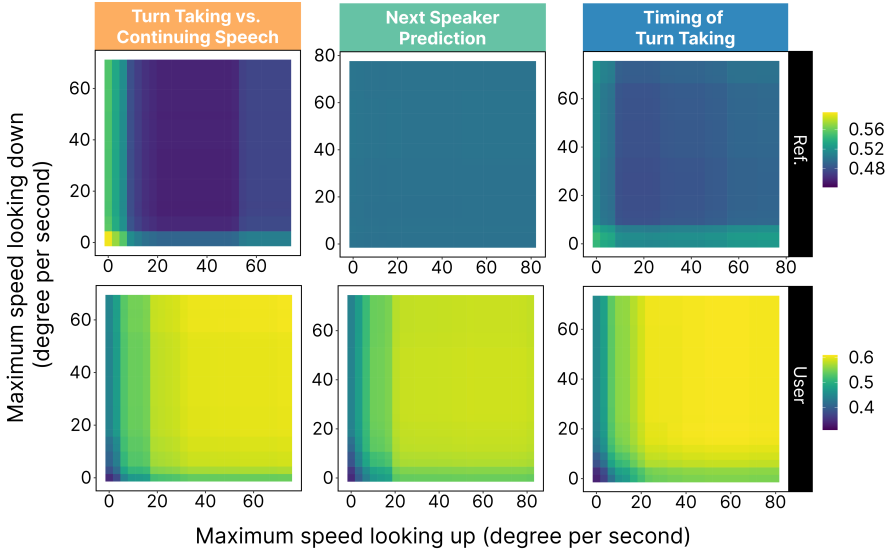


Fig. 7. Two-Variable Partial Dependence Plots for Maximum Speed for Head Pitch. The y- and x- axes quantify the maximum speed of the user looking downward and upward, respectively. The y-axis represents values for the maximum head pitch velocity and the x-axis represents values for the minimum head pitch velocity. We negated the original velocity values in the x-axis so that greater values in both axes represent greater instantaneous rotation. The plotted values denote the average probability estimates after setting the two features to their corresponding values based on their coordinates. Ref. = reference user.

Our results indicated that the extent to which an individual rotates their head in the pitch axis (i.e., velocity upward, velocity downward) is helpful in predicting speaking intentions. To understand how the features for upward head rotation (i.e., minimum head pitch velocity) and downward head rotation (i.e., maximum head pitch velocity) are related, we created two-variable partial dependence plots. Similar to the partial dependence plots we presented in Figures 4 and 5, two-variable partial dependence is derived by first varying the values to the two features and then calculating the average probability estimates of the trained model after this procedure.

Figure 7 shows the probability estimates after varying the extent of head pitch rotation for the reference and main users across the three tasks. Notably, we see that for all tasks, the combination of greater maximum speed in looking both upward and downward for the main user yielded higher probability of main user speaking intentions. Visually, it also appears that the probability estimates are noticeably higher when the maximum speeds for looking upward and downward are both greater than ≈ 10 degrees per second. This suggests that motions that yield “high enough” values of upward and downward pitch velocities (e.g., nodding) can be indicative of greater speaking intentions. The reverse is true for the tasks of continuing speech vs. turn transition and timing of turn transition regarding the previous speaker, with greater amount of upward and downward head rotation for the previous speaker being associated with smaller probability estimates for the main user’s speaking intentions. There was no distinct pattern observed for the previous speaker’s head pitch velocities in predicting the next speaker.

Received July 2024; revised December 2024; accepted March 2025