May I Help You? Predicting Help-seeking for Robot Assistance in Challenging Environments

Tianqi Liu[1], Wei-Che (Harry) Lin[1], Yejoon Yoo[1], Saleh Kalantari*†[12] and Andrea Stevenson Won*†[13]

***Abstract*— In challenging environments, AI-powered robots have potential to be helpful partners when humans become incapacitated. Wayfinding is a key task in which human guides often rely on nonverbal cues to decide to offer their partners help, yet current robot assistants lack this ability. As a first step to address this gap, we created an immersive virtual environment (IVE) designed to elicit help-seeking behaviors in a diving navigation task under environmental stressors. Seventy-two participants completed goal-directed wayfinding tasks while we recorded multimodal behavioral signals, including locomotion speed, gaze, head and body movement, and electrodermal activity (EDA). We then trained predictive models to infer when participants would seek help from a robot partner. We found that interpretable features such as increased gaze scanning and locomotion deceleration emerging as robust predictors of help-seeking in the virtual environment. This work offers preliminary design insights for proactive agent assistance in high-stakes scenarios.**

I. INTRODUCTION

Artificial intelligence (AI) is increasingly envisioned as a source of proactive assistance rather than a passive tool [1], [2]. To offer effective assistance, AI-powered robots must recognize when humans need help and respond appropriately [3]. This capability is particularly critical in high-stakes conditions, such as divers operating underwater [4], firefighters navigating hazardous spaces [5], and low-visibility industrial operations [6]. In these scenarios, timely support can be lifesaving, yet providing explicit, interpretable commands is often difficult, unreliable, or even impossible [7], [8]. A long-term vision for robots as intelligent assistants is to proactively assist in natural, seamless, and context-sensitive ways, without depending solely on explicit commands [9]. This requires systems that can infer human needs—particularly help-seeking—from implicit behavioral signals in real time, and intervene at the right moment [10]. Prior research has demonstrated that behavioral cues from facial expressions, gaze, posture, or physiological activity can serve as predictors for cognitive states [11]. However, little work has explored how such signals can be leveraged to model help-seeking as an overt decision under demanding conditions.

We created a VR simulation representing a challenging environment inspired by underwater wayfinding based on the following reasons. First, divers already operate alongside AI-powered underwater autonomous vehicles (UAV) which primarily communicate through full-body nonverbal cues rather than natural language [12]. Second, cognitive degradation underwater can occur suddenly (e.g., disorientation), creating clear moments where the human needs assistance. Third, challenges related to diving; for example, limited visibility, unpredictable forces, and constrained verbal communication, can be generalized to other high-demand contexts such as firefighting, search-and-rescue, and low-visibility industrial operations [6].

Current AUV–diver interaction strategies remain limited. Some approaches rely on remote human operators to supervise or control the robot, which is labor-intensive and impractical for many missions [13]. Others have the AUV continuously lead, which reduces diver autonomy and can increase fatigue [14]. Mode-switching methods, where divers explicitly signal for help through gestures, are also problematic since underwater command gestures are physically taxing, error-prone, and difficult to detect reliably [15]. These constraints highlight a critical gap: the lack of proactive strategies that allow robots to recognize when divers need assistance without requiring explicit commands.

As a first step to address this gap, we developed a ML model that predicts help-seeking during an underwater wayfinding simulation based on multimodal signals, including locomotion speed, gaze, head, body, and electrodermal activity (EDA). Our VR-based approach enabled precise tracking of behavioral signals while providing a realistic yet repeatable testbed for embodied human–AI interaction [16], [17]. In this paper, we present:

1) A machine learning approach that leveraged multimodal behavioral features to predict help-seeking during underwater wayfinding in a VR simulation.
2) An analysis of key behavioral patterns underlying help-seeking in the current task, offering insight into how humans navigate and how they request assistance.
3) A novel VR-based underwater simulation platform that replicates challenging environmental conditions, supporting controlled studies of user behavior in simulated high-difficulty contexts.

Overall, our findings demonstrated the potential of behavioral signal-based human modeling to enable adaptive, intuitive, and proactive AI-powered robot assistance—an important step toward robot systems that can operate more intelligently alongside humans in environments where explicit communication is constrained.

*These authors contributed equally to this work.
†Corresponding authors' email: {sk3268, asw248}@cornell.edu
[1]Department of Information Science, Cornell University, Ithaca, USA
[2]Department of Human Ecology, Cornell University, Ithaca, USA
[3]Department of Communication, Cornell University, Ithaca, USA

II. RELATED WORK

To understand how humans request and receive assistance in challenging environments, we analyzed prior research on (i) defining help-seeking, (ii) how human cognitive states have been modeled and inferred by AI and robots, and (iii) how VR testbeds provide controlled yet immersive platforms for such investigations.

A. Defining Help-Seeking

Help-seeking is commonly defined as a self-regulatory process in which individuals recognize difficulty, evaluate the need for support, and signal for assistance [18]. In human-robot interaction (HRI), the concept of help-seeking has been broadened to include the social and contextual appropriateness of assistance [3].

In our study, we defined help-seeking as the explicit act of requesting directional support from a robot assistant, recognizing that help-seeking needs could be anticipated before an overt request was made.

B. AI for Modeling Cognitive States and User Needs

A longstanding goal in human–agent interaction is to enable AI systems to infer users' cognitive states from implicit signals, allowing them to better understand user needs. Prior work has used multimodal behavioral and physiological signals to train machine learning models that classify cognitive states such as fatigue, attention, and cognitive load in both real world or simulated contexts [19], [20]. Stiber et al. [21] modeled user confusion—defined as periods when users require instruction but have not yet resolved the problem—using behavioral signals. Their results showed that machine learning models based on lightweight behavioral features achieved performance comparable to deep neural networks. In navigation contexts, Zhu et al. [22] detected navigational uncertainty from EEG signals during VR wayfinding; while Kattenbeck et al. [23] used eye tracking and inertial data to predict spatial familiarity.

While prior work demonstrates that cognitive states like confusion and uncertainty can be predicted from behavioral and physiological signals, these internal states do not specify the *appropriate timing* for assistance. Moreover, predicting help-seeking differs from predicting internal cognitive states such as confusion or uncertainty. Confusion and uncertainty are often modeled as continuous states that can be labeled at each moment, commonly through EEG, frequent self-reports, or researcher annotations of recordings [22]. By contrast, help-seeking is an *overt*, socially situated decision: at each moment, the user either requests assistance or does not. This binary decision is shaped not only by confusion or workload, but also by time pressure, task rules, perceived cost of asking, and individual help-seeking tendencies. This distinction introduces unique modeling challenges, including rare positive events, gradual emergence of need, and individual differences in when users choose to ask for help. Additionally, EEG and self-report labeling approaches can be difficult to deploy depending on the quality of challenging environments, while behavior signals such as gestures and gaze are increasingly used as input methods [14], making them more feasible candidates for real-time assistance timing [6]. We took a first step to address this gap by modeling help-seeking as a high-level decision that can nonetheless be anticipated through multimodal behavioral cues. Predicting help-seeking directly identifies when assistance is needed without requiring ongoing self-reports, offering a more actionable basis for proactive robot assistance in high-pressure settings.

C. Virtual Reality as a Testbed for Human Modeling

VR has been widely used to study cognition and behavior because it offers controlled, reproducible, and immersive conditions while supporting continuous capture of multimodal signals. In underwater contexts, Xia et al. [24] improved robot remote control with VR-augmented visual/haptic feedback under forces and visibility constraints; while cyber-physical platforms such as UnrealTHASC [17] and HoloOcean [16] integrated robot simulation. Beyond underwater simulation, VR has also been widely adopted as a research tool for simulating challenging conditions and modeling human behavior without high physical risks. For instance, Moussaïd et al. [25] used VR to examine evacuation behavior under emergency stress; while Rokooei et al. [26] applied VR to construction safety training.

These studies established VR as a reliable testbed for capturing behavioral and physiological signals to model human states. Building on this foundation, our work leveraged VR as a testbed to simulate a realistic and challenging underwater wayfinding task, and modeled help-seeking within this setting.

III. METHODS

In a within-participants design, participants completed two wayfinding tasks in a virtual "diving" environment with the assistance of an agent "robot".

A. Participants

We recruited 72 participants (38 female, 34 male), aged 19–51 years (M = 26.01, SD = 6.89) from a university community through flyers and email, and compensated with $20 USD. Two were excluded for misunderstanding directions. Participants self-reported their race and ethnicity by selecting all applicable categories: five "Biracial/Mixed/Multicultural/Multi-racial", one "Black American", twelve "Caucasian/White", twenty-nine "East Asian" two "Hispanic/Latino/Chicano/Puerto Rican", seven "Middle Eastern", nine "South Asian", six "Southeast Asian", and one "Other/I prefer not to answer this question". Participants' self-reported sense of direction was calculated from the Santa Barbara Sense of Direction Scale [27] (M = 4.33, SD = 0.95). All experimental procedures were approved by the Institutional Review Board, and all participants signed informed consent.

B. Materials

The virtual environment was developed in Unreal Engine 5.3. Participants wore a Meta Quest Pro VR headset with

integrated eye tracking. Electrodermal activity (EDA) was recorded using a BIOPAC MP160 system with two BioNomadix wireless sensors attached to the participant's left palm.

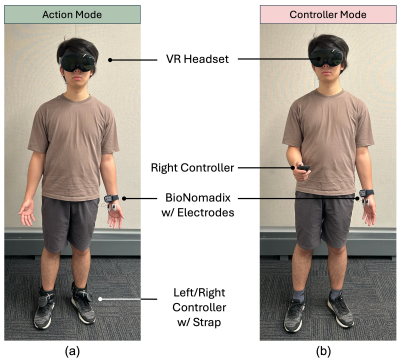

Fig. 1. Experimental setup for the two locomotion conditions. (a) Action mode: controllers strapped to both ankles captured toe-rising locomotion. (b) Controller mode: joystick navigation using the right-hand controller.

Each participant completed two wayfinding tasks, one in each locomotion mode. The locomotion setup and equipment are shown in Figure 1.

- **Action mode**: Participants alternately rose onto their toes. Two VR controllers strapped to their ankles captured the frequency and amplitude of the toe-rise movements, which determined forward locomotion. Participants physically rotated their bodies to change direction.
- **Controller mode**: Participants remained standing in place and navigated using the thumbstick on the right-hand controller. Forward motion was controlled by the thumbstick, while rotation was discretized into 15° increments to reduce simulator sickness.

In both conditions, forward speed was limited to a maximum of 1.6 m/s, and head orientation defined the forward axis. The action mode introduced richer, more naturalistic body movement, which better reflects real-world behaviors but adds variability compared to the controller mode.

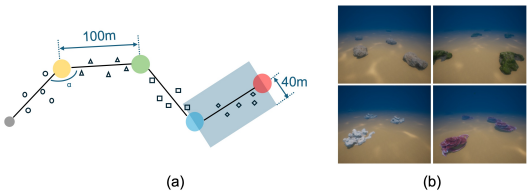

Fig. 2. Task design and experimental conditions. (a) Schematic of a sample navigation route. As an example, the blue shaded box on the fourth route indicates one area where landmarks (shown as small geometric icons) could be generated. (b) Example screenshots of the four routes.

In each wayfinding task, participants navigated to four color-coded targets (yellow, green, blue, red), each placed 100 meters apart. Targets were invisible beyond 10 meters, requiring participants to maintain orientation and backtrack when lost. Each inter-segment turning angle (α) was randomly selected from 90 degrees to 270 degrees (see Figure 2 (a)). Ten visual landmarks randomly distributed along each route segment supported navigation. Landmark appearance for each segment varied to sustain attention and reduce reliance on memory (Figure 2 (b)).

To simulate real-world diving conditions and create challenges that would elicit help-seeking, we introduced two environmental stressors: reduced visibility and currents. **Visibility** was manipulated by dynamically adjusting brightness and water clarity every 10–30 seconds under two conditions: high and low [28]. **Currents** were implemented as randomized simulated forces applied to the avatar, accompanied by spatialized audio cues [29]. Two conditions were used: currents and still-water. In the current condition, disturbances occurred randomly every 10–30 seconds. Currents consisted of *(1) push*, a horizontal translational force that randomly displaced the avatar by 0–5 meters, and *(2) torque*, a rotational force around the vertical axis that randomly rotated the avatar by 0–180°.

The four route segments in each wayfinding task were arranged in a 2×2 design, with visibility (high vs. low) and currents (present vs. absent) as environmental challenges. These factors were presented in a random order across participants.

C. Procedure

Participants completed informed consent and a pre-survey followed by task introduction and equipment setup. The two wayfinding tasks (action mode and controller mode) were completed in random order. Each was preceded by a short training, and followed by a post-condition survey. Finally, participants completed a post-VR survey.

D. Help-Seeking Interaction

Participants were not informed of the study's underlying goal. At the beginning of the task, they were told that an AUV robot would follow them and could be summoned for directional help. When called, the robot indicated the correct path with a light for five seconds. Participants were reminded that faster completion would result in a higher score. To reduce underuse of assistance, they were prompted in the task script: "Please ask the robot for help if you feel help is needed. We recommend asking for help around 10 times during the task." In practice, participants requested help an average of 8.31 times (SD = 3.67).

Help was requested using a full back-and-forth wave of the right forearm, inspired by real-world underwater signaling. A researcher monitored participants in real time and manually triggered the robot guide upon detecting the gesture, while participants were unaware that the robot was operated manually.

E. Measures

We continuously logged multimodal behavioral signals and help-seeking events.

Head, body, and speed signals. Unreal Engine logs captured head position and rotation, controller position and rotation, avatar rotation, and locomotion speed. Signals were recorded at irregular intervals with a mean sampling rate of approximately 100 Hz and millisecond timestamps.

Gaze signals. We recorded gaze origin and direction at 50 Hz using the Meta Quest Pro eye-tracking API.

EDA signal. Electrodermal activity was recorded with the BIOPAC system at 2000 Hz using AcqKnowledge 5.0.

Help-seeking events. Robot help requests were logged with millisecond precision.

Surveys. The *pre-VR survey* included the Santa Barbara Sense of Direction Scale (SBSOD) [27]. After each locomotion condition, participants reported whether they had felt uncertain but chose not to request help. The *post-VR survey* collected demographic information.

F. Data Preprocessing and Feature Extraction

The behavioral data streams, including head, speed, body and gaze, were resampled to 20 Hz and synchronized across modalities [30]. Short gaps (all <200 ms) were linearly interpolated in each dimension.

1) Segmentation and Labeling: For each help request, the preceding data was segmented into consecutive, non-overlapping 5-second windows moving backward until the previous help event or the start of the recording (Figure 3). Segments shorter than 5 seconds at the boundary were excluded. The window immediately before the help request (0–5s) was labeled as 1 (help-seeking). To account for individual variability in the onset of help-seeking predictors, the next three windows, encompassing 5 to 20 seconds prior to the help request, were regarded as an indeterminate buffer zone. Participants often exhibited early behavioral cues of help-seeking, but the onset timing varied widely across individuals. To account for this variability, the buffer period was considered an ambiguous region that could not be reliably classified as either help-seeking or non-help-seeking and were excluded from model training and evaluation. This buffer region is further analyzed in Section IV-C to examine individual differences in help-seeking thresholds. All remaining windows (>20 s before a help request) were labeled as 0 (non-help-seeking).

We evaluated multiple window and buffer configurations and selected a 5-second window with a three-windows buffer based on a comparative analysis reported in the supplementary materials (Appendix I, `https://osf.io/26wsu/`).

To further ensure that our labeling strategy does not inadvertently rely on gesture-related artifacts, Appendix II in the supplementary materials reports a complementary robustness analysis showing that model performance remains stable even after aggressively removing the seconds immediately preceding each help request.

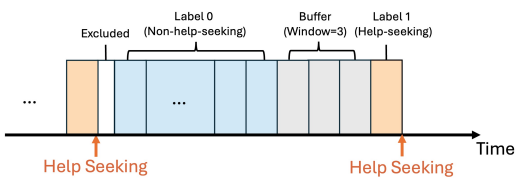

Fig. 3. Window labels: help-seeking (class 1, orange), non-help-seeking (class 0, blue), buffer (gray), and excluded windows (white).

After segmentation and labeling, the dataset contained 8,954 samples in Class 0 (non–help-seeking, 90.55%) and 935 samples in Class 1 (help-seeking, 9.45%). Under action mode, Class 0 included 4,230 samples (89.79%) and Class 1 included 481 samples (10.21%). Under controller mode, Class 0 included 4,724 samples (91.23%) and Class 1 included 454 samples (8.77%).

2) Feature Extraction: From each 5-second segment, we extracted a set of 419 features, including basic (n=133), sequence-based (n=266) and deviation-based (n=19) features and participant identifier (n=1) spanning all signal modalities. Table I summarizes the basic features and their counts.

Locomotion Speed. Raw speed signals were smoothed with a 0.3 Hz low-pass filter to remove artifacts from the walking algorithm. From smoothed speed traces, we computed seven basic statistics, including mean, variance (var), maximum, minimum, movement intensity (MI), entropy, and peak-to-peak amplitude (Amplitude_ptp). Acceleration and deceleration was derived from smoothed speed, from which we extracted mean, variance, maximum, and minimum. A speed below 0.6 m/s was recorded as a stop. To characterize stop-and-move patterns, we quantified counts and durations of short stops ($\geq$0.2 s) and long stops ($\geq$1.0 s), including total and maximum stop durations. Finally, we computed overall movement duration ($t_{\text{move}} = t_{\text{total}} - t_{\text{stop}}$), movement ratio ($t_{\text{move}}/t_{\text{total}}$), and stop ratio ($t_{\text{stop}}/t_{\text{total}}$), where t_{total} is the segment length and t_{stop} is the total stop duration.

Eye Gaze. We denoised the gaze data using a 3-sample median filter to remove spikes followed by a light 4 Hz Butterworth low-pass filter to reduce jitter without smearing saccades. Fixations were identified with the dispersion-threshold algorithm (IDT; dispersion threshold = 0.1 normalized frame units; duration threshold = 200 ms), in line with prior work [31]. For fixation features, we computed the mean, minimum, maximum, and variance of fixation duration, dispersion, and dispersion in X and Z dimensions, and fixation frequency. For saccades (rapid eye movement between multiple fixations), we extracted the mean, minimum, maximum, and variance of amplitude and duration, as well as skewness of amplitude and the long/short gaze-length ratio (GL ratio).

Head Movement. Yaw angles were unwrapped to remove the $\pm 180^\circ$ wrap–around discontinuity. To reduce frame-to-frame jitter, we applied a second-order Butterworth low-pass filter (cutoff 6 Hz). From head orientation (yaw, pitch, roll, and head yaw (θ_{head})), we computed time-domain features: mean, minimum, maximum, variance, absolute maximum, root mean square (RMS), standard deviation (std), energy, peak-to-peak amplitude, mobility index (MI), entropy, and ZCR [23]. For HMD position, we computed variance, std, and Amplitude_ptp. From head yaw, we extracted frequency-domain features: dominant frequency, fundamental period, spectral centroid, and spectral roll-off, the first three cepstral coefficients to capture low-frequency structure.

Body Orientation. In the controller mode, body yaw corresponded to avatar yaw. In the action mode, body yaw was estimated from the orthogonal direction of the

TABLE I

SUMMARY OF EXTRACTED FEATURES FOR HELP-SEEKING PREDICTION. FEATURES ARE COMPUTED BY APPLYING THE STATISTICAL MEASURES IN THE "EXTRACTED FEATURES" COLUMN TO THE SIGNALS IN THE SECOND COLUMN.

Modality	Signals	Extracted Features	Count
Speed	Speed	Mean, var, max, min, MI, entropy, Amplitude_ptp	7
	Speed: acceleration, deceleration	Mean, var, max, min	8
	Short stop, long stop	Count, duration_total, duration_max	6
	Movement	Duration_total, move ratio, stop ratio	3
Gaze	Fixation: duration, dispersion (disp), disp X, disp Z	Mean, var, min, max	16
	Fixation frequency	–	1
	Saccade: amplitude, duration	Mean, var, min, max	8
	Saccade amplitude	Skewness, long/short GL ratio	2
Head	HMD: yaw, pitch, roll; head yaw	Mean, var, min, max, absmax, RMS, std, energy, Amplitude_ptp, MI, entropy, ZCR	48
	HMD: X, Y, Z	Var, std, ptp	9
	Head yaw	CepstralCoeff_1–3, dominant frequency, fundamental period, spectral centroid, spectral rolloff	7
Body	Yaw	Net displacement, std	2
EDA	EDA	Mean, std, max, min, range	5
	Tonic, phasic	Mean, std	4
	SCR	Count, amplitude_mean, amplitude_max, height_mean, rise time_mean, recovery time_mean, presence	7

line connecting the two ankle-mounted controllers in the horizontal plane, and head yaw relative to the body (θ_{head}) was computed as the difference between HMD yaw and body yaw. Body yaw angles were first unwrapped and then smoothed with a second-order Butterworth low-pass filter (6 Hz cutoff). For each segment, we computed the net angular displacement and the standard deviation of body yaw.

Electrodermal Activity (EDA). We used NeuroKit2 [32] python package to clean EDA data. From the cleaned EDA data, we computed mean, standard deviation, maximum, minimum, and range. After decomposing the data into tonic and phasic components, we also extracted their mean and standard deviation. From skin conductance responses (SCRs) detected in the phasic component, we extracted count, mean and maximum amplitude, mean peak height, average rise time, average recovery time, and a binary indicator of whether or not any SCR occurred.

Sequence-Based and Deviation-Based Features. We derived temporal features to capture short-term dynamics. For each feature x, we computed sequence-based statistics using a sliding window (w=3). At time t, the rolling mean and standard deviation were defined as $\mu_t = \frac{1}{w}\sum_{i=t-w+1}^{t} x_i$ and $\sigma_t = \sqrt{\frac{1}{w}\sum_{i=t-w+1}^{t}(x_i-\mu_t)^2}$, capturing local trends and variability.

We further computed deviation-based features to measure departures from a sliding baseline: $\text{Dev}_t = x_t - \frac{1}{w}\sum_{i=1}^{w} x_{t-i}$. We applied the deviation-based features to a subset of features (N=19) for which short-term fluctuations are theoretically meaningful, including locomotion speed and acceleration (speed mean, variance, acceleration variance, acceleration mean, speed ZCR), stop–move patterns (number of short stops, movement ratio, stop ratio), gaze dynamics (fixation duration, dispersion, dispersion X, dispersion Z, saccade amplitude and duration, gaze-length ratio), and head movement (head yaw mean, min, max, var).

IV. RESULTS

A. Classification Model

We trained and evaluated both user-dependent and user-independent classifiers using the full set of 419 features. Five machine-learning models were considered: Logistic Regression, Balanced Random Forest, LightGBM, XGBoost, and CatBoost in Python 3.11. Hyperparameters were optimized using *grid search*, with the tested parameter ranges and the comparison across classification methods reported in Appendix III.

Among these models, CatBoost achieved the best trade-off between precision and recall for help-seeking detection and was selected as the final model (max_iterations = 150, depth = 6, learning_rate = 0.1, class_weight = [1,10]).

1) User-dependent Model: For user-dependent models, features were standardized per participant using z-score to remove inter-individual baseline differences and highlight within-participant variability. We applied five-fold cross-validation, each trained on four folds and tested on the fifth fold. As shown in Table II, when trained and tested on controller mode (Ctrl) data, the classifier reached an accuracy of 95.69%, AUC of 85.50%, and an F1 score of 76.21% for the help-seeking class. In the action mode (Act), performance was slightly lower (accuracy = 94.10%, AUC = 79.09%, F1 = 70.27%). Training on the combined dataset yielded comparable results (accuracy = 94.38%, AUC = 80.72%, F1 = 71.71%). Across all settings, the classifier achieved high performance for the non–help-seeking class

TABLE II

PERFORMANCE OF THE USER-DEPENDENT AND USER-INDEPENDENT MODELS. METRICS INCLUDE ACCURACY (ACC), AREA UNDER THE ROC CURVE (AUC), PRECISION (PREC), RECALL (REC), AND F1 SCORE FOR BOTH HELP-SEEKING (1) AND NON–HELP-SEEKING (0) CLASSES.

	User Dependent			**User Independent**		
	Ctrl	Act	All	Ctrl	Act	All
Acc	95.69	94.10	94.38	95.45	91.48	93.35
AUC	85.50	79.09	80.72	96.94	92.26	94.47
Prec (1)	74.22	72.42	68.38	71.22	67.77	69.40
Rec (1)	78.63	68.39	75.51	84.47	72.80	78.30
F1 (1)	76.21	70.27	71.71	73.58	64.00	68.51
Prec (0)	97.94	96.43	97.42	98.25	95.59	96.84
Rec (0)	97.33	97.02	96.35	96.73	94.88	95.75
F1 (0)	97.63	96.72	96.88	97.43	95.00	96.14

TABLE III

TOP TEN FEATURES RANKED BY SPEARMAN CORRELATION WITH HELP-SEEKING. THE TABLE REPORTS CORRELATION COEFFICIENT (ρ) AND SIGNIFICANCE LEVEL (p).

Feature	ρ	p
Saccade amplitude_mean	+0.220	<0.001
Head yaw_std	+0.183	<0.001
Saccade amplitude_mean (sequence)	+0.135	<0.001
Speed deceleration_max	+0.122	<0.001
Speed_max	-0.100	<0.001
Speed acceleration_mean	-0.097	<0.001
Speed acceleration_mean (sequence)	-0.090	<0.001
Fixation disp. X_mean (sequence)	-0.070	<0.001
Body yaw_std	+0.050	<0.001
Phasic_mean	-0.022	0.011

(F1 = 96.72–97.63%). Overall, controller mode slightly outperformed action mode, likely because thumbstick-based navigation is more controlled and less noisy than physical body rotations, leading to clearer signal patterns.

2) User-independent Model: Since user-dependent models could rely on individual behavioral patterns, we also tested user-independent classification models to gain a more robust and generalized classifier. We evaluated the classifier using the leave-one-user-out (LOUO) method, where we trained the model on the data of 69 participants and evaluated it on the data of the 1 left-out participant. As per-participant standardization could introduce information leakage, raw features were used. Statistical analyses were conducted on the average evaluation result per participant.

The results demonstrate robust but slightly lower performance compared to user-dependent models. As shown in Table II, controller mode again provided the highest performance, with accuracy of 95.45%, AUC of 96.94%, and an F1 score of 73.58% for the help-seeking class. Training on action mode data led to weaker results (accuracy = 91.48%, AUC = 92.26%, F1 = 64.00%). The combined dataset produced intermediate outcomes (accuracy = 93.35%, AUC = 94.47%, F1 = 68.51%). Again, non–help-seeking classification remained strong across all settings (F1 = 95.00–97.43%).

Taken together, both user-dependent and user-independent models demonstrated robust performance in predicting help-seeking. User-dependent models achieved higher overall accuracy and balanced F1 scores, while user-independent models demonstrated strong cross-participant generalization.

B. Most Important Behavioral Features and Interpretation

To better understand which behavioral signals drove the model's performance, we analyzed feature importance from the user-dependent model. We then computed the Spearman correlations for the top ten features with the binary help-seeking label (0 = non-help-seeking, 1 = help-seeking). Table III shows these features with correlation coefficients (ρ) and significance (p).

The analysis revealed several consistent patterns. Gaze and head movement features were the strongest positive correlates of help-seeking. Broader saccades (Saccade amplitude_mean, ρ = 0.22, $p < 0.001$; Saccade amplitude_mean (sequence), ρ = 0.14, $p < 0.001$) and greater horizontal head scanning (Head yaw_std, ρ = 0.18, $p < 0.001$) were robustly associated with a higher likelihood of requesting help, indicating that participants often engaged in active visual search and scanning behaviors when needed help. Body yaw variability (Body yaw_std) also contributed positively, further indicating that broader body rotation signal an increased probability of help-seeking.

Locomotion features showed consistent negative associations: lower maximum speed (Speed_max, ρ = –0.10, $p <$ 0.001) and reduced acceleration (Speed acceleration_mean, ρ = –0.097, $p < 0.001$; Speed acceleration_mean (sequence), ρ = –0.09, $p < 0.001$) predicted increased help-seeking, reflecting hesitation and reduced movement vigor. Conversely, abrupt halts (Speed deceleration_max, ρ = 0.12, $p < 0.001$) were positively associated with help-seeking.

Horizontal gaze dispersion (Fixation disp. X_mean) showed a negative association with help-seeking. This can be explained by the IDT algorithm we used, where dispersion is defined as the spatial spread of gaze points within a fixation cluster. During non–help-seeking periods, participants often maintained longer fixations, allowing greater within-cluster drift and yielding higher dispersion values. In contrast, during confusion, participants produced shorter fixations interspersed with broader saccades.

Finally, physiological features contributed little to the prediction model. Phasic EDA (Phasic_mean, ρ = –0.022, p = 0.011) showed only a very small negative effect.

A feature ablation analysis is provided in Appendix IV.

C. Gradual Emergence of Help-Seeking Predictors

Our predictive models deliberately excluded the potentially ambiguous *buffer period* from 5–20s prior to the help request, as this period is variable in regard to whether or not it would include clear help-seeking needs. To analyze this period, we applied the model in a LOUO manner (n=70) to estimate, for each held-out participant, the average probability of help-seeking at five timepoints, including the non-help-seeking baseline along with each of the four windows

leading up to the help-seeking action (buf0: 20–15s, buf1: 15–10s, buf2: 10–5s, help: 5–0s before the request).

When aggregated across participants (Figure 4 (a)), the mean predicted probability increased steadily as the request moment approached, although individual trajectories varied. This pattern suggests that behavioral indicators of help-seeking tend to increase gradually in the seconds preceding an explicit request. To further examine these patterns, we clustered the five-point trajectories using k-means (k=3, silhouette = 0.314). The resulting clusters indicated three tendencies (Figure 4 (b)): some participants showed early increases but delayed the request ("early-signs, late-act", n=18, 25.7%), some exhibited a gradual rise with a sharp increase shortly before requesting help (n=23, 32.9%), and others showed only modest changes prior to the request (n=29, 41.4%).

To relate these trajectories to subjective preferences, participants answered a question after each condition asking whether they had felt uncertain but chose not to ask the robot for help. Participants who answered "no" in both conditions were labeled Threshold-low, indicating they tended to request help quickly once uncertain, while those who answered "yes" at least once were labeled Threshold-high. For each trajectory we computed two indices: the slope across the five time points and the change from baseline to the help window. Spearman correlations showed that Threshold was significantly associated with the magnitude of change (ρ=-.29, p=.014, 95% CI $[-.52, -.05]$), but not with slope (ρ=-.17, p=.15). When trajectories were grouped by Threshold (Figure 4 (c)), Threshold-low participants tended to show sharper increases close to the request, whereas Threshold-high participants showed earlier increases but delayed the request.

Overall, these analyses suggest that behavioral indicators of help-seeking often develop gradually before an explicit request, and that individuals differ in how quickly they act once uncertainty arises.

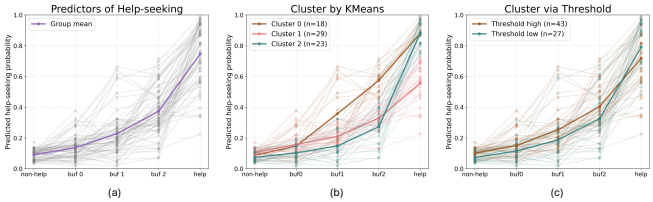

Fig. 4. Predicted help-seeking trajectories in the 20 seconds before a request. (a) Across all participants, trajectories (gray lines) and group mean (purple) show a consistent rise in help-seeking predictors as the help-seeking moment approached. (b) Clustering of individual trajectories. (c) Grouping by self-reported help-seeking threshold.

V. Discussion

Our results show that help-seeking from a robot partner can be predicted from observable behavioral signals before an explicit request occurs. Both user-dependent (71.71% F1 for help-seeking) and user-independent (68.51% F1 for help-seeking) models achieved stable performance. This suggests that although help-seeking is influenced by multiple factors such as internal uncertainty or time pressure, it produces consistent external patterns. Gaze scanning, head movement, and slowdown behaviors were the strongest predictors under the current setting. Prior work has linked these markers to spatial uncertainty and navigational confusion [23], [22]; our findings suggest that these signals may also function as anticipatory indicators of overt help-seeking decision-making. Together, these results provide initial insights into how AI-powered robot systems operating in high-stakes environments might anticipate help-seeking from observable behavioral signals to offer assistance proactively.

Another important finding is that help-seeking predictors emerged gradually rather than suddenly, beginning approximately 20 seconds prior to the explicit help-seeking request under the current task. This provides empirical support for an evidence-accumulation view of help-seeking, aligning with established human decision-making theories [33]. Although most prior navigation research treats help-seeking as a discrete event [18], our results suggest that help-seeking needs are a process that unfolds over time.

Additionally, individuals differed in when they chose to request help. This implies that proactive systems should consider personal thresholds, since the same observable behavior may not reflect the same level of need across users. In repeated interactions, short calibration phases or online adaptation could help align assistance timing with individual preferences. This may reduce over-assistance for confident users and under-assistance for those who need earlier support.

VI. Limitations and Future Work

This study has several limitations that should be considered in future work. First, our simulated challenging condition was inspired by diving wayfinding. While this allowed controlled data collection with rich multimodal signals, it was not completely realistic; for example, participants navigated in a vertical body posture. Although we argue that the interpretable predictive patterns are still useful, validation in real environments is obviously necessary.

Second, our task focused on one specific type of help-seeking for wayfinding. Help-seeking behavior may differ in other tasks, such as collaborative problem solving, emergency response, or skill training, where the costs, social meanings, and timing of asking for assistance may vary.

Finally, we evaluated predictive performance offline but did not test the model in a real-time interactive setting. The practical value of proactive assistance depends not only on prediction accuracy but also on how well the system supports real-time collaboration. Future studies should integrate the model into a live assistive robot system that continuously monitors users' behavioral states and provides proactive guidance when help-seeking is predicted. Such systems could move beyond responding only to explicit requests and support more adaptive human-robot collaboration, where robot partners better infer when human partners may need assistance and provide timely support.

VII. Conclusion

This study developed a predictive model of human help-seeking from multimodal behavioral signals collected during wayfinding in a VR simulation of a challenging environment. We found that observable behaviors—particularly gaze and head scanning and locomotion slowdown—preceded explicit help requests. Beyond predictive performance, our findings suggest that help-seeking unfolds gradually and that individuals thresholds for assistance request differ. These findings provide initial insights for the design of AI-powered robot assistants that not only respond to explicit requests but may also anticipate when support is needed. Such capabilities could contribute to more natural and timely assistance in challenging environments.

References

[1] C. Meurisch, C. A. Mihale-Wilson, A. Hawlitschek, F. Giger, F. Müller, O. Hinz, and M. Mühlhäuser, "Exploring user expectations of proactive ai systems," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–22, 2020.

[2] V. Chen, A. Zhu, S. Zhao, H. Mozannar, D. Sontag, and A. Talwalkar, "Need help? designing proactive ai assistants for programming," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–18.

[3] R. Ramnauth, D. Brščić, and B. Scassellati, "Should i help?: A skill-based framework for deciding socially appropriate assistance in human-robot interactions," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2024, pp. 2051–2058.

[4] L. Christensen, J. de Gea Fernández, M. Hildebrandt, C. E. S. Koch, and B. Wehbe, "Recent advances in ai for navigation and control of underwater robots," *Current Robotics Reports*, vol. 3, no. 4, pp. 165–175, 2022.

[5] A. Hassanein, M. Elhawary, N. Jaber, and M. El-Abd, "An autonomous firefighting robot," in *2015 International Conference on Advanced Robotics (ICAR)*, 2015, pp. 530–535.

[6] S. Jain and B. Argall, "Probabilistic human intent recognition for shared autonomy in assistive robotics," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–23, 2019.

[7] A. Birk, "A survey of underwater human-robot interaction (u-hri)," *Current Robotics Reports*, vol. 3, no. 4, pp. 199–211, 2022.

[8] A. Bonarini, "Communication in human-robot interaction," *Current Robotics Reports*, vol. 1, no. 4, pp. 279–285, 2020.

[9] Y. T.-Y. Hou, W.-Y. Lee, and M. Jung, ""should i follow the human, or follow the robot?"—robots in power can have more influence than humans on decision-making," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–13.

[10] H. Lyu, X. Wang, N. Zhang, S. Ma, Q. Zhu, Y. Luo, F. Tsung, and X. Ma, "Signaling human intentions to service robots: Understanding the use of social cues during in-person conversations," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–21.

[11] J. A. Duncan, F. Alambeigi, and M. W. Pryor, "A survey of multimodal perception methods for human–robot interaction in social environments," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 4, pp. 1–50, 2024.

[12] J. R. Pelletier, B. W. O'Neill, J. J. Leonard, L. Freitag, and E. Gallimore, "Auv-assisted diver navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 208–10 215, 2022.

[13] G. Marani, S. K. Choi, and J. Yuh, "Underwater autonomous manipulation for intervention missions auvs," *Ocean Engineering*, vol. 36, no. 1, pp. 15–23, 2009.

[14] N. Mišković, M. Bibuli, A. Birk, M. Caccia, M. Egi, K. Grammer, A. Marroni, J. Neasham, A. Pascoal, A. Vasilijević, *et al.*, "Caddy—cognitive autonomous diving buddy: Two years of underwater human-robot interaction," *Marine technology society journal*, vol. 50, no. 4, pp. 54–66, 2016.

[15] M. J. Islam, M. Fulton, and J. Sattar, "Toward a generic diver-following algorithm: Balancing robustness and efficiency in deep visual detection," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 113–120, 2018.

[16] E. Potokar, S. Ashford, M. Kaess, and J. G. Mangelson, "Holoocean: An underwater robotics simulator," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 3040–3046.

[17] S. Surve, J. Guo, J. C. Menezes, C. Tate, Y. Jin, J. Walker, and S. Ferrari, "Unrealthasc–a cyber-physical xr testbed for underwater real-time human autonomous systems collaboration," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2024, pp. 196–203.

[18] R. P. Darken and J. L. Sibert, "Wayfinding strategies and behaviors in large virtual worlds," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1996, pp. 142–149.

[19] S. K. Sardar, N. Kumar, and S. C. Lee, "A systematic literature review on machine learning algorithms for human status detection," *IEEE Access*, vol. 10, pp. 74 366–74 382, 2022.

[20] E. Hwang and J. Lee, "Looking but not focusing: Defining gaze-based indices of attention lapses and classifying attentional states," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1–14. [Online]. Available: https://dl.acm.org/doi/10.1145/3706598.3714269

[21] M. Stiber, D. Bohus, and S. Andrist, ""uh, this one?": Leveraging behavioral signals for detecting confusion during physical tasks," in *Proceedings of the 26th International Conference on Multimodal Interaction*, ser. ICMI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 194–203. [Online]. Available: https://doi.org/10.1145/3678957.3685727

[22] B. Zhu, J. G. Cruz-Garza, Q. Yang, M. Shoaran, and S. Kalantari, "Identifying uncertainty states during wayfinding in indoor environments: An eeg classification study," *Advanced Engineering Informatics*, vol. 54, p. 101718, 2022.

[23] M. Kattenbeck, I. Giannopoulos, N. Alinaghi, A. Golab, and D. R. Montello, "Predicting spatial familiarity by exploiting head and eye movements during pedestrian navigation in the real world," vol. 15, no. 1, p. 7970, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-025-92274-4

[24] P. Xia, H. You, and J. Du, "Visual-haptic feedback for rov subsea navigation control," *Automation in Construction*, vol. 154, p. 104987, 2023.

[25] M. Moussaïd, M. Kapadia, T. Thrash, R. W. Sumner, M. Gross, D. Helbing, and C. Hölscher, "Crowd behaviour during high-stress evacuations in an immersive virtual environment," *Journal of The Royal Society Interface*, vol. 13, no. 122, p. 20160414, 2016.

[26] S. Rokooei, A. Shojaei, A. Alvanchi, R. Azad, and N. Didehvar, "Virtual reality application for construction safety training," *Safety science*, vol. 157, p. 105925, 2023.

[27] M. Hegarty, A. E. Richardson, D. R. Montello, K. Lovelace, and I. Subbiah, "Development of a self-report measure of environmental spatial ability," *Intelligence*, vol. 30, no. 5, pp. 425–447, 2002.

[28] K. J. DeMarco, M. E. West, and A. M. Howard, "Underwater human-robot communication: A case study with human divers," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014, pp. 3738–3743.

[29] G. Fauville, A. Voşki, M. Mado, J. N. Bailenson, and A. Lantz-Andersson, "Underwater virtual reality for marine education and ocean literacy: technological and psychological potentials," *Environmental Education Research*, pp. 1–25, 2024.

[30] A. Khan, N. Hammerla, S. Mellor, and T. Plötz, "Optimising sampling rates for accelerometer-based human activity recognition," *Pattern Recognition Letters*, vol. 73, pp. 33–40, 2016.

[31] R. Islam, K. Desai, and J. Quarles, "Cybersickness prediction from integrated hmd's sensors: A multimodal deep fusion approach using eye-tracking and head-tracking data," in *2021 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 2021, pp. 31–40.

[32] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior research methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[33] J. R. Busemeyer and J. T. Townsend, "Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment," *Psychological review*, vol. 100, no. 3, p. 432, 1993.