Within-Session Reliability of fNIRS in Robot-Assisted Upper-Limb Training

Yi-Chuan Jiang, *Graduate Student Member, IEEE*, Chen Zheng, Rui Ma, Yifeng Chen, *Member, IEEE*, Sheng Ge, *Member, IEEE*, Chenyang Sun, Jianjun Long, Peng Fang, and Mingming Zhang, *Senior Member, IEEE*

***Abstract*— Functional near-infrared spectroscopy (fNIRS) seems opportune for neurofeedback in robot-assisted rehabilitation training due to its noninvasive, less physical restriction, and no electromagnetic disturbance. Previous research has proved the cross-session reliability of fNIRS responses to non-motor tasks (e.g., visual stimuli) and fine-motor tasks (e.g., finger tapping). However, it is still unknown whether fNIRS responses remain reliable 1) in gross-motor tasks, 2) within a training session, and 3) for different training parameters. Hence, this study aimed to investigate the within-session reliability of fNIRS responses to gross-motor tasks for different training parameters. Ten healthy participants were recruited to conduct right elbow extension-flexion in three robot-assisted modes. The *Passive* mode was fully motor-actuated, while *Active1* and *Active2* modes involved active engagement with different resistance levels. FNIRS data of three identical runs were used to assess the within-session reliability in terms of the map-(R^2) and cluster-wise ($R_{overlap}$) spatial reproducibility and the intraclass correlation (ICC) of temporal features. The results revealed good spatial reliability (R^2 up to 0.69, $R_{overlap}$ up to 0.68) at the subject level. Besides, the within-session temporal reliabilities of Slope, Max/Min, and Mean were between good and excellent ($0.60 < ICC < 0.86$). We also found that the within-session reliability was positively correlated with the intensity of the training mode, except for the temporal reliability of HbO in *Active2* mode. Overall, our results demonstrated good within-session reliability of fNIRS responses, suggesting fNIRS as reliable neurofeedback for constructing closed-loop robot-assisted rehabilitation systems.**

***Index Terms*— Functional near-infrared spectroscopy (fNIRS), within-session reliability, robotics, upper-limb training.**

Manuscript received 29 August 2023; revised 6 February 2024; accepted 13 March 2024. Date of publication 19 March 2024; date of current version 25 March 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFF1205200, Grant 2022YFF1202500, and Grant 2022YFF1202502; in part by the National Natural Science Foundation of China under Grant 62303211; in part by the Shenzhen Science and Technology Program under Grant JCYJ20220530113811027, Grant JCYJ20220818103602004, and Grant JCYJ20210324104203010; and in part by the Guangdong Provincial Key Laboratory of Advanced Biomaterials under Grant 2022B1212010003. *(Corresponding author: Mingming Zhang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Southern University of Science and Technology under Application Nos. 20220161, and performed in line with the Declaration of Helsinki.

Yi-Chuan Jiang, Chen Zheng, Rui Ma, Yifeng Chen, Chenyang Sun, and Mingming Zhang are with Shenzhen Key Laboratory of Smart Healthcare Engineering, Guangdong Provincial Key Laboratory of Advanced Biomaterials, and the Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: zhangmm@sustech.edu.cn).

Sheng Ge is with the Key Laboratory of Child Development and Learning Science, School of Biological Science and Medical Engineering, Ministry of Education, Southeast University, Nanjing 210096, China.

Jianjun Long is with The First Affiliated Hospital of Shenzhen University, Shenzhen University School of Medicine, Shenzhen Second People's Hospital, Shenzhen 518000, China.

Peng Fang is with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, and Shenzhen Engineering Laboratory of Neural Rehabilitation Technology, Shenzhen Institute of Advanced Technology, Shenzhen 518055, China.

Digital Object Identifier 10.1109/TNSRE.2024.3378467

I. Introduction

NONINVASIVE neuroimaging has great potential to reveal spatial and temporal changes in neural activity underlying stroke rehabilitation [1]. As an established noninvasive neuroimaging technique, functional near-infrared spectroscopy (fNIRS) can reflect neural activation by measuring the concentration changes of hemoglobin in target brain areas. Compared with other modalities, such as functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and electroencephalogram (EEG), fNIRS has the advantages of no electromagnetic disturbance, less physical restriction, and ease of use [2], [3]. These unique properties make fNIRS an excellent choice to characterize neural activation in rehabilitation training [4].

Robot-assisted rehabilitation enables repetitive and intensive practice at a relatively low cost [5], [6], and has been proven beneficial in improving motor performance in stroke patients [7]. Closing the robot-assisted training loop with fNIRS-based neurofeedback can help refine training protocols in a timely manner, potentially enhancing patient engagement and improving recovery efficiency [8], [9]. However, there are three prerequisites for utilizing fNIRS as neurofeedback. Firstly, fNIRS responses need to be demonstrated as reliable in gross-motor tasks, such as elbow extension-flexion being common in rehabilitation training. Secondly, reliable responses should be captured even within a training session. This allows

for rapid tuning of robotic parameters, which can be accomplished in just a few minutes at the run level. It is worth noting that a session generally consists of a few runs, thus "within-session" is essentially "between-runs". Thirdly, it is crucial to ensure the reliable of fNIRS responses for different robotic parameters (e.g., resistance, speed, and trajectory, etc.), so that these robotic parameters can be adjusted to elicited desired brain activation during closed-loop training.

Previous studies have extensively investigated the reliability of fNIRS responses to visual and auditory stimuli. In Plichta et al.'s study [10], participants were asked to gaze at the visual stimuli while hemodynamic responses in visual cortex were measured using fNIRS. Their results showed good reliability of fNIRS responses in terms of intraclass correlation coefficient (ICC) of up to 0.84. Blasi et al. [11] conducted an experiment to examine the reliability of fNIRS response to social stimuli, both auditory and visual, in infants. Good reliability was found at the group level (spatial overlap of 0.94). In addition to external stimuli, researchers also investigated the reliability of fNIRS responses to fine-motor tasks, including finger opposing [1], finger tapping [2], [12], [13], and hand grasping [14], [15]. Their results showed acceptable reliability of fNIRS responses in fine-motor tasks. In brief, previous research examining the reliability of fNIRS responses has mainly focused on non-motor or fine-motor tasks, with little investigation into gross-motor tasks.

Most existing studies have focused on the cross-session reliability of fNIRS responses [1], [2], [10], [11], [12], [14], [16]. For example, Tian et al. [16] conducted a test-retest experiment to evaluate the cross-session reliability of fNIRS responses induced by repetitive transcranial magnetic stimulation (rTMS). They found moderate-to-high reliability of both fNIRS amplitudes and spatial activation patterns between two scan sessions with a two to three-day interval. Broscheid et al. [17] evaluated the cross-session reliability of the mean, slope, and area under the curve for hemodynamic response function (HRF) derived from oxyhemoglobin (HbO) and deoxyhemoglobin (HbR), reporting fair-to-good cross-session reliability of these temporal features. Zhang et al. [13] reported a good within-session reliability of fNIRS responses, as evidenced by a good consistency (Pearson's $r = 0.77$) in spatial activation patterns from two half sessions. Bae et al. [15] investigated the reliability of fNIRS spatial activation patterns from two runs that were 15 minutes apart and found very poor within-session reliability. There is currently a lack of research on the within-session reliability of fNIRS temporal features.

While previous studies have assessed the reliability of fNIRS responses for different task types, there is limited research on how task parameters affect the reliability of fNIRS responses. To the best of our knowledge, only one study [15] has investigated the reliability of fNIRS responses for different training parameters. The velocity of robot-assisted passive grasping was set to three levels (slow at 0.25 Hz, moderate at 0.5 Hz, and fast at 0.75 Hz), and the reliability of fNIRS responses was evaluated in terms of ICC. Their results showed that there was almost no reliability of fNIRS responses (ICC = 0.002) for each of the tested training parameters. However, the results may be worth further investigation due to the low ICC value. This may be attributed to the limited number of subjects involved in the study or the significant amount of random error in the experiment.

In this study, we focused on investigating the within-session reliability of fNIRS responses during robot-assisted upper-limb training, which is essential for developing fNIRS-based neurofeedback that target specific brain regions (including the contralesional area) and operate at relatively frequent update rates (within several minutes at the run level). Prior to implementing fNIRS-based neurofeedback to construct a closed-loop robot-assisted rehabilitation system, the following three problems need to be thoroughly explored and validated: whether fNIRS responses remain reliable 1) in gross-motor tasks, 2) within a training session, and 3) for different training parameters. To achieve this goal, we used a customized rehabilitation robotic system to precisely control the training parameters, resulting in three types of robot-assisted mode with varying resistance levels (*Passive*, *Active1*, and *Active2*). The experiment consisted of three identical runs, where participants were instructed to perform robot-assisted right elbow extension-flexion based on visual cues. fNIRS was used to measure hemodynamic changes in the left motor cortex. Both spatial and temporal reliabilities were evaluated for each robot-assisted mode and each hemoglobin species in terms of the following reliability indices: 1) the coefficient of determination R^2, 2) the degree of spatial overlap $R_{overlap}$, and 3) the ICC. In this context, "spatial reliability" refers to the reproducibility of spatial activation patterns, while "temporal reliability" refers to the consistency of temporal features. On the basis of findings, practical guidance was provided for future fNIRS-based closed-loop rehabilitation research.

II. Materials and Methods

A. Participants

Ten healthy adults (five males and five females, mean age = 23.4 ± 6.7 years, range 19 – 28 years) participated in this study. All participants were confirmed to be right-handed by the Edinburgh Handedness Inventory and reported no history of neurological or psychiatric disorders. This study was approved by the Ethics Committee of Southern University of Science and Technology (20220161), and conducted in adherence to the declaration of Helsinki. All participants provided written informed consent after a detailed explanation of the experiment and the fNIRS technique.

B. Robotic System

Details of robotic system design were described in our previous publications [18], [19]. As illustrated in Fig. 1, the robotic handle was connected to the motion module via a three-axis forces sensor, which can measure the human-robot interaction force at a 1000 Hz sampling rate. In addition, the robotic system was equipped with 18 optoelectronic switches to determine the coordinate origin and limiting positions. This system supports upper-limb rehabilitation training in both passive and active modes.

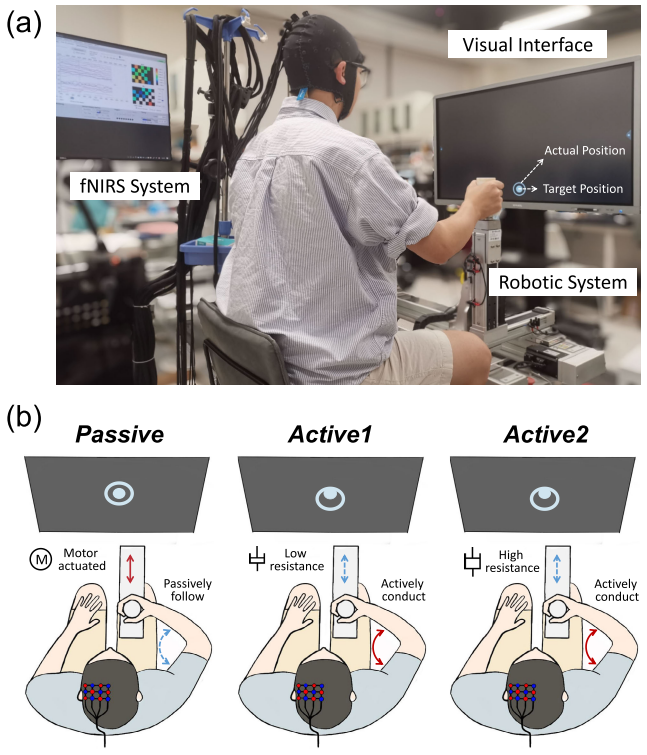

Fig. 1. Experimental setup and task. (a) Overall setup consisted of a visual interface, a custom-made rehabilitation robot system, and a fNIRS neuroimaging system. (b) Experimental task was robot-assisted right elbow extension-flexion in *Passive*, *Active1*, and *Active2* modes. In *Active1* and *Active2* modes, the participants were required to minimize the difference between target (solid circle) and actual (hollow circle) handle positions to maintain constant motion.

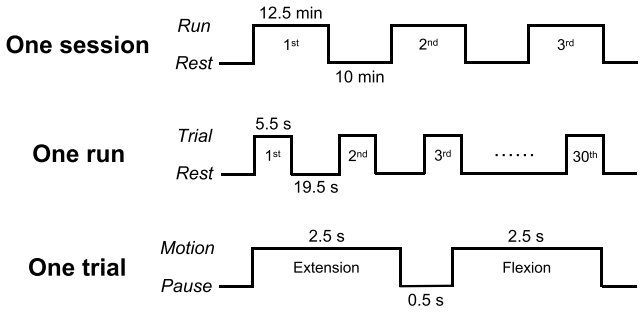

Fig. 2. Experimental procedures. The experimental session was divided into three identical runs. Each run consisted of 30 trials, during which each mode was randomly performed 10 times. A single trial included an extension-flexion stage for 5.5 s and followed with a resting stage for 19.5 s.

C. Experimental Tasks and Procedures

In the current study, we aimed to investigate the within-session reliability of hemodynamic responses evoked by robot-assisted upper-limb training. The participants sat in front of the robot with their right hands holding the right handle naturally (see Fig.1.a). The height and position of the seat were adjusted to make each of them feel comfortable. During the experiment, participants were asked to perform robot-assisted right elbow extension-flexion in the following three robot-assisted modes:

Passive Mode: The extension-flexion movement was entirely motor-actuated and guided by the robot.

Active1 Mode: The resistance value was set to 1.0 N·s/cm. Participants actively conducted extension-flexion movements.

Active2 Mode: The resistance value was set to 3.3 N·s/cm. Participants actively conducted extension-flexion movements.

Notably, the resistance of the robotic system remained constant in both *Active1* and *Active2* modes throughout the entire process of elbow extension-flexion. Thus, the intensity levels of *Passive*, *Active1*, and *Active2* can be classified as low, moderate, and high, respectively (see Fig.1.b).

All extension-flexion movements were performed along the vertical axis, and the distance from the proximal point to the distal end was 20 cm. We selected a medium movement speed of 8 cm/s and set a response delay of 0.5 s between extension and flexion, resulting in an entire front-and-back straight-line movement lasting 5.5 s. During the elbow extension-flexion movements, both the target and actual trajectories of the robotic handle were displayed in real-time on the screen. In *Passive* mode, the target and actual trajectories were identical. However, in *Active1* and *Active2* modes, participants were instructed to minimize the discrepancy between the target and actual handle positions to ensure a consistent motion. We used the PsychoPy software [20] for controlling the presentation of visual cues. The robot's motion control and position feedback were achieved via user datagram protocol (UDP) communication.

As depicted in Fig. 2, the experiment session consisted of three identical runs. Each run comprised 30 trials, where each elbow extension-flexion mode (*Passive*, *Active1*, and *Active2*) was randomly executed 10 times. Each trial encompassed a task stage lasting 5.5 s (2.5 s of extension, 0.5 s of pause, and 2.5 s of flexion), followed by a rest period of 19.5 s between trials. To ensure physical and mental relaxation, a 10 min break was provided between runs. Prior to the experiment, all participants were thoroughly briefed on experimental tasks and procedures. Additionally, a 5 min pre-training session was conducted to familiarize them with the operation of the robotic system and training modes.

D. fNIRS Acquisition

fNIRS signal acquisition was conducted using a continuous-wave fNIRS system (NIRScout, NIRx Medizintechnik GmbH, Germany), equipped with 24 laser sources and 24 detectors operating at wavelengths of 785 nm and 830 nm (see Fig.1.a). To determine the optimal probe arrangement, covering the pre-motor cortex (PMC), supplementary motor area (SMA), and primary motor cortex (M1) of the left hemisphere [3], [21], we employed the fNIRS optodes location decider (fOLD) toolbox [22]. In order to maximize the utilization of fNIRS probes, a sensitivity threshold of 15% was set, and optode locations with coverage below this threshold were excluded. As a result, a configuration of 6 sources and 6 detectors was obtained, conforming to the 10-5 international system (Source 1 was positioned at FFT7h). This setup allowed for 17 fNIRS measurement channels with an inter-optode distance of 3 cm (as shown in Fig.3.a), enabling a sampling rate of 10.4 Hz. Furthermore, photon transport simulations were performed by

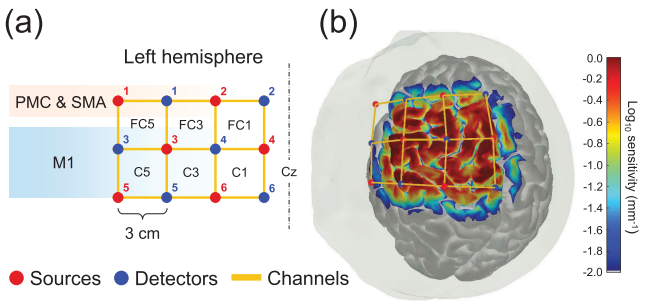

Fig. 3. Arrangement and sensitivity of fNIRS optodes and channels. Optode array set up with 6 sources and 6 detectors resulting in 17 channels over the left PMC, SMA and M1 with 3 cm separation. Estimated measurement sensitivity to brain regions.

Monte Carlo Extreme module embedded in the AtlasViewer package [23], to evaluate the migration of photons within the head tissues and identify the cortical regions that can potentially be measured by the fNIRS channels. The sensitivity profile (see Fig.3.b) demonstrates the capability of the designed fNIRS channel arrangement to detect hemodynamic responses in the aforementioned cortical regions.

Prior to the experiment, we conducted measurements of the participant's head circumference, nasion-inion distance, and the distance between the left and right periauricles. An easy cap with fNIRS optodes inserted at the 10-5 international system was placed in the middle between nasion to inion and left to right periauricles (reference point Cz). The PsychoPy software was used to synchronize fNIRS recording via event triggers.

The recorded data can be downloaded from the following website: https://doi.org/10.6084/m9.figshare.22178630.v2.

E. Raw Data Quality Evaluation

In this study, we employed two commonly used metrics to evaluate the quality of raw data: 1) Scalp Coupling Index [24], and 2) Peak Power [25]. Both of metrics examine the signal for the presence of heart beat signal, which indicates that fNIRS optodes were in contact with the scalp.

1) Scalp Coupling Index (SCI): A signal with good scalp-optode coupling is characterized by a strong pulsation of optical signals at both wavelengths, and defined SCI as the normalized cross-correlation between the raw data at each wavelength. The SCI value of 0.75 is a commonly used threshold for identifying good scalp-optode coupling.

2) Peak Power (PP): The spectral power of the cross-correlated signal can be used as an estimator of the strength of the cardiac signal. In contrast to the SCI, raw signal containing movement artifacts yield PP value close to zero. A threshold value of 0.1 for the PP is typically associated to a good quality signal.

Both the SCI and PP values were calculated for each channel, each run, and each subject using the MNE-NIRS toolbox (v.0.6) [26].

F. fNIRS Preprocessing

FNIRS signals were preprocessed using the Brain AnalyzIR toolbox (v.2022.4.26) [27] implemented in MATLAB 2021a (MathWorks Inc., MA, USA). First, 20 s of raw light intensity signals before the first stimulus and 25 s after the last stimulus were trimmed to remove unwanted data. Next, the signals were resampled to 10 Hz. Afterwards, the resampled light intensity signals were converted into optical density. Finally, the optical density signals were converted into the concentration changes of oxygenated (ΔHbO) and deoxygenated hemoglobin (ΔHbR) based on the modified Beer-Lambert law [28]. The differential path length factor was adjusted according to the age of each participant [29].

G. HRF Estimation With AR-IRLS Model

The HRFs of HbO and HbR were estimated for each channel, each run, and each subject using the autoregressive, iterative robust least-squares (AR-IRLS) [30]. Previous studies [31], [32] have demonstrated that the use of AR-IRLS for solving general linear model (GLM) can remove serially correlated errors and motion artifacts, thereby improving HRF estimation. We used the AR-IRLS model embed in the Brain AnalyzIR toolbox [27], and for the solution, please refer to [30].

Specifically, the preprocessed single-run fNIRS data, which includes all 30 trials (10 trials for each robot-assisted mode), was used to construct the GLM model. The HRF was modeled using finite impulse responses (FIRs), distributed over a range from 0 to 20 s, with 0 s representing the onset of the task. The weights β of the regressors were solved by the AR-IRLS for each robot-assisted mode. Consequently, the single-run HRF for each robot-assisted mode can be obtained by convolving the corresponding β coefficients with the FIR model. By repeating this process, the HRFs were extracted for each channel, each run, and each subject.

H. Within-Session Reliability Evaluation

To comprehensively assess the within-session reliability, we focused on two different aspects. First, we examined the within-session spatial reliability. The coefficient of determination (R^2) and the degree of spatial overlap ($R_{overlap}$) were calculated to assess the map- and cluster-wise spatial reliability, respectively. Second, we examined the within-session temporal reliability. The intraclass correlation coefficients (ICC) of five commonly used temporal features were calculated to quantify the temporal reliability of HRF over three runs.

1) Spatial Reliability: In this study, we evaluated within-session spatial reliability based on activation patterns from different runs. Subject-level activation pattern was constructed by generating individual t-map. Specifically, we conducted one-sample t-test on AR-IRLS produced β coefficients for each channel and used the resulting t-value as activation intensity of that channel (see details in Sec.II-I). The activation pattern (t-map) was composed of activation intensities from different channels. By repeating this process for three runs (Run1, Run2, and Run3), three robot-assisted modes (*Passive*, *Active1*, and *Active2*), and two hemoglobin species (HbO and HbR), a total of 18 activation patterns can be obtained. We utilized a linear mixed effects model to estimate the group-level t-values and employed a similar approach to construct

group-level activation patterns as we did for subject-level activation patterns. With the estimated activation patterns, we assessed within-session spatial reliability at both subject- and group- levels using two metrics that have been utilized in previous fNIRS research [2], [10].

First, the map-wise assessment was performed to investigate the within-session spatial reliability of a global activation map by computing the coefficient of determination (R^2) of t-values between two runs. A high R^2 (close to 1) represents that the variance in spatial activation of one run can largely be explained by another run, indicating high map-wise spatial reliability.

Second, the cluster-wise assessment was utilized to evaluate the reproducibility of activated channels between two runs. The activated channels were inspected based on the fixed number of channels strategy [33]. In this study, we explored a Top 20% channel quantity threshold, which means that the four channels with the highest t-values (for HbO) or lowest t-values (for HbR) in the activation pattern were considered as activated channels. The degree of spatial overlap ($R_{overlap}$) can be calculated by the following formula [10]:

$$R_{overlap} = 2 \times \frac{C_{overlap}}{C_j + C_k} \tag{1}$$

where C_j and C_k are the numbers of activated channels in two runs, and $C_{overlap}$ is the quantity of identical activated channels in both runs.

2) Temporal Reliability: We employed the following five temporal features to depict the time course of hemodynamic response [34], [35]:

TTP/TTN: The time to peak (or nadir) for HbO (or HbR).

Slope: The slope of a linear least squares fit to HRF between 0 and 4 s.

Max/Min: The maximum (or minimum) HbO (or HbR) amplitude between 2 and 8 s.

Mean: The mean amplitude between 2 and 8 s.

Std: The standard deviation of amplitude between 2 and 8 s.

The determination of the time window is based on whether significant change in feature values (see Sec.III-D).

The within-session reliability of these temporal features was assessed using ICC, based on an absolute agreement, two-way random effects model with repeated measures [14], [36]:

$$ICC\,(2,1) = \frac{MS_s - MS_e}{MS_s + (k-1)\,MS_e + \frac{k}{n}\,(MS_r - MS_e)} \tag{2}$$

$$MS_s = \frac{k}{n-1} \sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}\right)^2 \tag{3}$$

$$MS_r = \frac{n}{k-1} \sum_{i=1}^{k} \left(\overline{y}^i - \overline{y}\right)^2 \tag{4}$$

$$MS_e = \frac{1}{(n-1)\,(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_j^i - \overline{y}_j - \overline{y}^i + \overline{y}\right) \tag{5}$$

where k and n represent the number of runs and subjects ($k = 3$ and $n = 10$ in this study); MS_s, MS_r, and MS_e denote the between-subjects mean squares, between-runs mean squares, and error mean squares, respectively; y_j^i denotes the feature derived from the averaged HRF for the $i'th$ run and the $j'th$ subject, $\overline{y}_j$ is the mean of feature for the $j'th$ subject, $\overline{y}^i$ is the mean of features for the $i'th$ run, and $\overline{y}$ is the mean of all runs across all subjects.

A high ICC (close to 1) represents low total (between-runs and error) variability relative to between-subjects variability, indicating high within-session reliability of the temporal feature. Conversely, a low ICC (close to 0) represents that temporal features of different runs lack consistency [37].

Firstly, the temporal features were derived from the average HRF across the four activated channels of Run1, as described in Sec.II-H.1. The ICC values were then calculated for each temporal feature to assess the cluster-wise temporal reliability. Secondly, we compared ICC values across different numbers of activated channel, ranging from 1 to 17 with a step size of 1.

All the reliability indexes, including the coefficient of determination R^2, the degree of spatial overlap $R_{overlap}$, and the ICC, were evaluated according to the criteria proposed by Cicchetti and Sparrow [38]. Index values were considered as 'excellent' above 0.75, 'good' between 0.59 and 0.75, 'fair' between 0.40 and 0.58, and 'poor' for values lower than 0.40.

I. Statistical Analysis

With the obtained regression coefficients β and the covariance matrices Cov_β and σ^2 (see Sec.II-G), statistical inference was performed to estimate the activation level for each channel by testing the null hypothesis, that the regression coefficients β over the defined period (from 2 to 8 s) were not significantly different from zero. The formula for such a one-sample t test is given by [27]:

$$t = \frac{c \cdot \beta}{\sqrt{c \cdot Cov_\beta \cdot c^T}} \tag{6}$$

where t is the t-value of that channel and c is the contrast vector. We set the elements of the contrast vector within the predefined time window (from 2 to 8 s) to 1, while the remaining elements are set to 0.

The quality of raw data from three runs was analyzed using one-way ANOVAs. Our hypothesis was that there would be no differences in the quality of the raw data across the three runs.

The within-session spatial reliabilities (R^2 and $R_{overlap}$) were analyzed using three-way ANOVAs (or three Kruskal-Wallis tests, if spatial reliabilities were non-normally distributed) with three factors: robot-assisted modes (*Passive*, *Active1*, and *Active2*), pairwise runs (Run 1-2, Run 2-3, and Run 1-3) and hemoglobin species (HbO and HbR). If there is a significant effect of the first factor, it would support our hypothesis that within-session spatial reliability may vary between robot-assisted modes. Additionally, by examining the impact of the second and third factors, we can determine how within-session spatial reliability changes over time and identify which type of hemoglobin species, HbO or HbR, demonstrates greater spatial reliability.

The effects of robot-assisted modes and runs on temporal features were analyzed with two-way ANOVA or two Kruskal-Wallis tests, depending on feature distribution. Our hypothesis

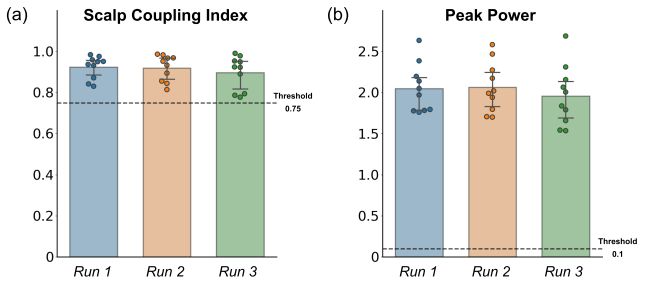

Fig. 4. The averaged (a) SCI and (b) PP values across all channels and all participants. The dashed line indicates the threshold.

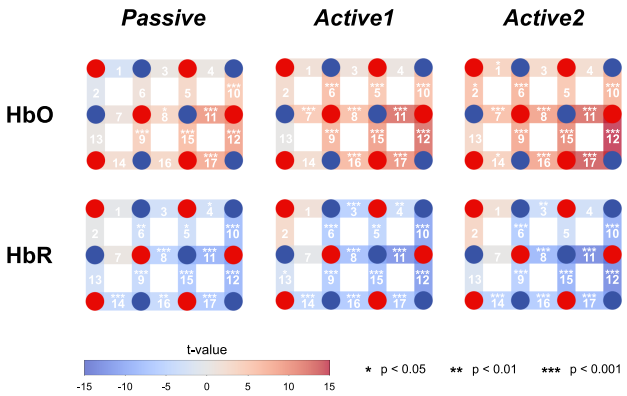

Fig. 5. Group-level t-maps of brain activation specific for three different modes of elbow extension-flexion. The white number indicate the channel number and the asterisk indicates the corresponding range of the p-value.

was that the temporal features of the *Active* mode are more significant than those of the *Passive* mode and that the features change over time.

Statistical analyses were performed using SPSS (IBM SPSS Statistics 26.0, IBM Corporation, USA). The normality of the data was evaluated using Shapiro-Wilk tests. A post-hoc t-test was applied when significant main effects and interactions were found. Statistical significance level was set to 0.05 (confidence intervals are 95%) for all statistical tests. Bonferroni correction was used for multiple comparisons.

III. Results

A. Raw Data Quality

Fig. 4 illustrates the SCI and PP for three runs. The averaged SCI values across all channels and all participants for Run1, Run2, and Run3 were 0.92, 0.92, and 0.90, respectively. All participants had SCI values above the threshold of 0.75. One-way ANOVA revealed there was no significant difference of SCI value between three runs ($F_{(2,29)} = 0.444$, $p = 0.646$). Similarly, the averaged PP values across all channels and all participants for Run1, Run2, and Run3 were 2.05, 2.07, and 1.96, respectively. All participants had PP values above the threshold of 0.1. One-way ANOVA indicated there was no significant difference of PP value between three runs ($F_{(2,29)} = 0.314$, $p = 0.733$). Since all channels had SCI values and PP values exceeding the threshold, no channels were pruned during the preprocessing.

B. Spatial Activation Patterns

Fig. 5 depicts the group-level t-maps for three robot-assisted modes, organized into two rows corresponding to HbO and HbR. In *Passive* mode, group-level analysis of HbO identified channels #11, #12, #17, and #15 with the top four highest t-values. Regarding HbR, channels with the top four lowest t-values were #11, #12, #10, #17. In *Active1* mode, #11, #12, #17, #15 for HbO and #11, #12, #10, #8, for HbR were channels with the top four highest or lowest t-values. In *Active2* mode, the four most activated channels for HbO were #12, #17, #11, #15, and for HbR they were #11, #12, #10, #15. All of these channels were significantly activated ($p < 0.001$). Overall, the dorsal aspect of the left M1 revealed consistent channel activation across three robot-assisted modes and two hemoglobin species, with the most activated channels concentrated in this region. Furthermore, there were notable differences in activation level among the three robot-assisted modes. The highest (or lowest) t-values for HbO (or HbR) were observed in *Active2* mode with values of 15.14 (-13.31), followed by *Active1* mode with values of 12.78 (-12.37), and *Passive* mode with values of 9.75 (-9.48). These results indicate that active movements elicited stronger activation compared to passive movements. Additionally, the intensity of fNIRS responses was positively correlated with the intensity of robot-assisted mode (*Passive* < *Active1* < *Active2*).

C. Within-Session Spatial Reliability

Fig. 6 illustrates the scatterplots of group-level t-maps, which distributed across three planes in a 3D coordinate system. Each point represents t-value at a single channel derived from two runs (e.g., the point on the bottom plane corresponds to the t-value from Run 1 and Run 2). The solid line in each plane represents the linear regression line ($y = ax + b$, where x and y are t-values from two different runs) that best fits the data points. Channels exhibiting the top four highest or lowest t-values in both runs were highlighted in red (for HbO) or blue (for HbR) and labeled with their respective channel numbers. As revealed in Fig. 6, the data points were tightly clustered around the regression line, with coefficients of determination (R^2) greater than 0.75. This indicates that more than 75% of the variability in t-values of one run can be explained by t-values of another run. In other words, these high R^2 values prove excellent consistency in channel activation between runs. Additionally, the degree of overlap ($R_{overlap}$) ranged from 0.75 to 1.00 for HbO, and from 0.50 to 0.75 for HbR, indicating excellent cluster-wise spatial reliability in HbO, and fair-to-good reliability for HbR.

In comparison to the spatial reliability observed at the group level, the within-session spatial reliability at the subject level was somewhat lower. As listed in Table I, the R^2 for *Passive*, *Active1*, and *Active2* modes were fair ($0.44 < R^2 < 0.52$), fair ($0.48 < R^2 < 0.56$), and fair-to-good ($0.52 < R^2 < 0.62$) for HbO. For HbR, the R^2 were fair-to-good ($0.57 < R^2 < 0.61$), fair-to-good ($0.57 < R^2 < 0.64$), and good ($0.62 < R^2 < 0.69$) for *Passive*, *Active1*, and *Active2* modes. The $R_{overlap}$ for all three modes ranged between fair and good

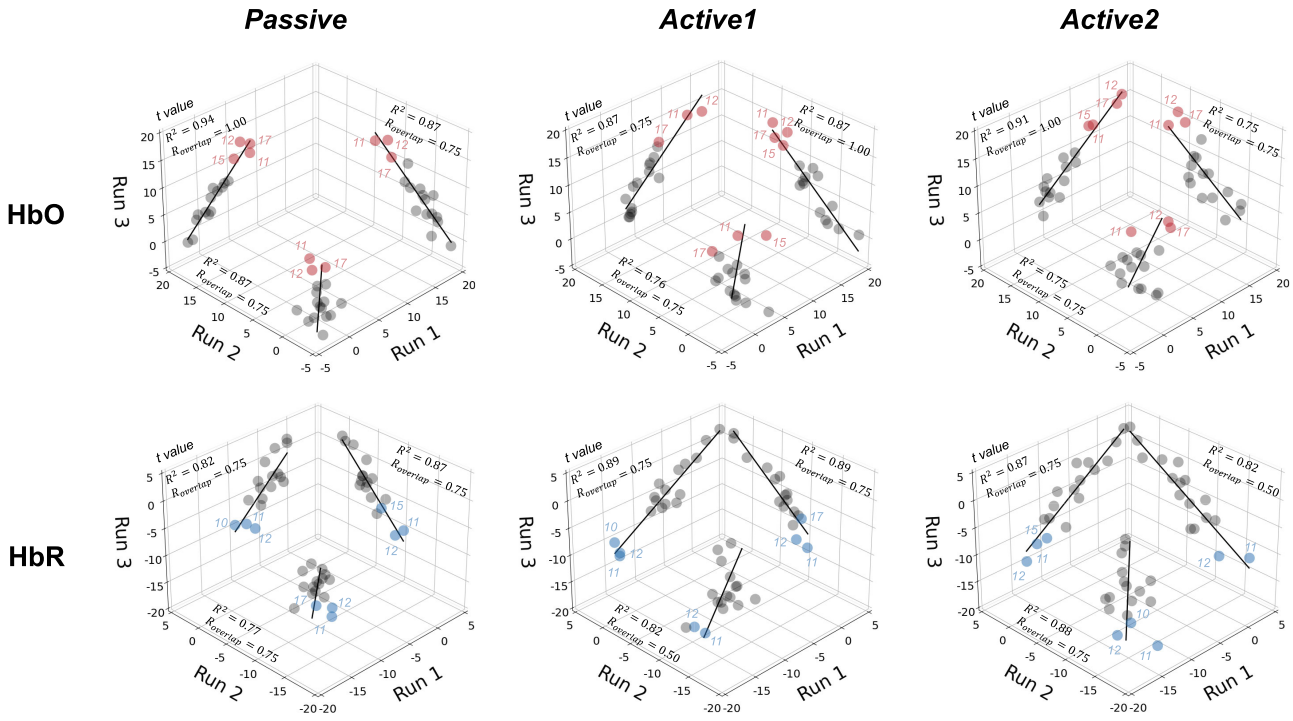

Fig. 6. Scatter plots of group-level t-values for spatial reliability assessment. Each data point represents the t-value at a single channel derived from two runs. The solid lines are linear regression lines that best fitted the scatter plot of data points. The red or blue points (labeled with the channel number) indicate the channels with the top four highest (for HbO) or lowest (for HbO) t-values in both runs. R^2 and $R_{overlap}$ denote the determination coefficient of the regression and the degree of spatial overlap, respectively.

TABLE I

SPATIAL RELIABILITY EXPRESSED WITH THE DETERMINATION COEFFICIENT OF THE REGRESSION (R^2) AND THE DEGREE OF SPATIAL OVERLAP ($R_{overlap}$) BASED ON THE OBTAINED INDIVIDUAL T-VALUES

Species	Mode	R^2			$R_{overlap}$		
		Run 1-2	Run 2-3	Run 1-3	Run 1-2	Run 2-3	Run 1-3
HbO	Passive	0.52 (0.08-0.93)	0.53 (0.01-0.86)	0.44 (0.01-0.70)	0.58 (0.25-1.00)	0.63 (0.25-1.00)	0.48 (0.25-0.75)
	Active1	0.56 (0.01-0.81)	0.54 (0.05-0.84)	0.48 (0.07-0.91)	0.65 (0.25-1.00)	0.65 (0.25-1.00)	0.60 (0.25-0.75)
	Active2	**0.62 (0.51-0.79)**	**0.57 (0.10-0.93)**	**0.52 (0.01-0.88)**	**0.68 (0.50-1.00)**	**0.68 (0.25-1.00)**	**0.63 (0.50-0.75)**
HbR	Passive	0.60 (0.26-0.92)	0.61 (0.16-0.97)	0.57 (0.03-0.90)	0.63 (0.25-1.00)	0.65 (0.25-1.00)	0.58 (0.25-1.00)
	Active1	0.61 (0.20-0.86)	0.64 (0.10-0.87)	0.57 (0.01-0.96)	0.68 (0.25-1.00)	0.65 (0.25-1.00)	0.63 (0.25-1.00)
	Active2	**0.69 (0.47-0.96)**	**0.65 (0.09-0.94)**	**0.62 (0.40-0.93)**	**0.68 (0.50-1.00)**	**0.68 (0.25-1.00)**	**0.63 (0.25-0.75)**

First value is the average of all individuals, and values in parenthesis are the minimum and maximum values.

($0.44 < R_{overlap} < 0.68$) for both HbO and HbR. Although the results of the three-way ANOVA did not reveal any significant main effects on robot-assisted modes, pairwise runs, or hemoglobin species, we observed: 1) the spatial reliability exhibited a positive correlation with the intensity of robot-assisted mode (*Passive* < *Active1* < *Active2*); 2) the spatial reliability was higher between adjacent runs (Run 1-2 & Run 2-3) compared to non-adjacent runs (Run 1-3); 3) the spatial reliability of HbR was higher than that of HbR.

D. Temporal Activation Patterns

Fig. 7 shows the average fNIRS responses across all channels and participants in *Passive*, *Active1*, and *Active2* modes. A noticeable increase (or decrease) in HbO (or HbR) can be observed during the elbow extension-flexion task (from 0 to 5.5 s). The hemodynamic changes exhibit a relatively steep slope between 0 and 4 s, with the peak (or trough) of HbO (or HbR) typically occurring around 5 s. In *Passive* mode, the mean peak (or trough) amplitudes for three runs were 1.41 (-0.81), 1.01 (-0.66), and 1.30 (-0.75) $\mu mol/L$. For *Active1* mode, the mean peak (or trough) amplitudes for three runs were 3.14 (-1.08), 2.40 (-0.85), and 2.00 (-1.02) $\mu mol/L$. In the *Active2* mode, the mean peak (or trough) amplitudes for three runs were 4.02 (-1.22), 2.51 (-0.87), and 2.50 (-1.04) $\mu mol/L$. Active movements elicited higher activation compared to passive movements, with the most significant difference observed between 2 and 8 s. The time taken to return to baseline (recovery time) for *Passive*, *Active1*, and *Active2* modes was approximately 15 s, 12 s, and

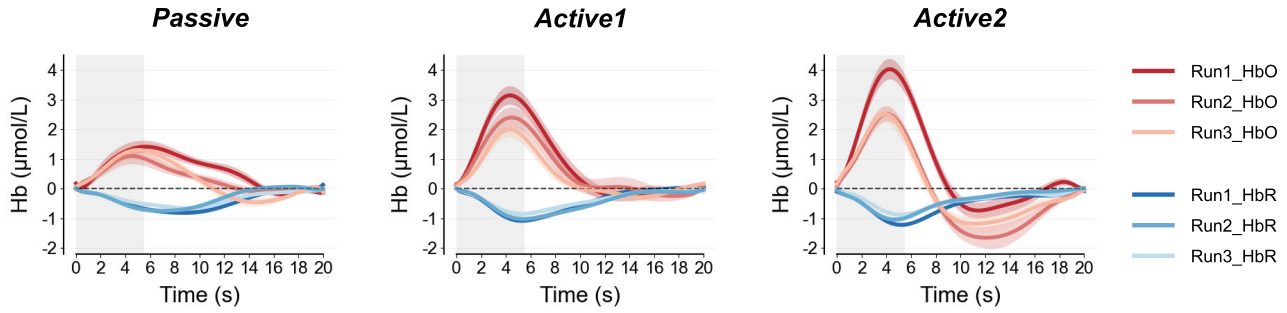

Fig. 7. Grand-averaged time-course of HbO and HbR across all channels and all participants for *Passive*, *Active1*, and *Active2* modes. Grey areas indicate the extension-flexion stage (from 0 to 5.5 s).

TABLE II
TEMPORAL FEATURES BASED ON THE OBTAINED INDIVIDUAL HRF

Species	Mode	TTP/TTN (s)			Slope (μ)			Max/Min ($\mu mol/L$)			Mean ($\mu mol/L$)			Std ($\mu mol/L$)		
		Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
HbO	Passive	5.96	5.17	5.14	0.35	0.27	0.31	1.93	1.51	1.90	1.21	0.88	1.09	0.57	0.44	0.58
	Active1	4.91	4.54	4.36	0.87	0.63	0.54	3.43	2.64	2.21	2.50	**1.91**	1.46	0.79	0.70	0.64
	Active2	4.45	4.32	3.87	**1.11**	**0.68**	**0.66**	**4.22**	**2.68**	**2.64**	**3.02**	1.60	**1.55**	**1.03**	**0.93**	**0.97**
HbR	Passive	7.20	6.86	6.40	-0.11	-0.09	-0.13	-0.83	-0.72	-0.85	-0.58	-0.45	-0.59	0.20	0.17	0.21
	Active1	5.69	5.66	5.79	-0.24	-0.19	-0.24	-1.14	-0.90	-1.05	-0.89	**-0.71**	**-0.85**	**0.26**	0.18	0.20
	Active2	5.39	5.46	5.14	**-0.26**	**-0.20**	**-0.25**	**-1.25**	**-0.95**	**-1.13**	**-0.99**	-0.70	-0.83	0.24	**0.22**	**0.26**

The bold values in the table are the maximum (or minimum) feature values of HbO (or HbR) among three robot-assisted modes.

10 s, respectively. Furthermore, a trend of decreasing activation over runs was observed in *Active1* and *Active2* modes.

Table II lists the averaged temporal features for each mode and run, based on HRFs collected from ten participants. Two-way ANOVA with Bonferroni corrected post-hoc multiple comparison was used to analyze the effect of mode and runs on each feature. Results showed that the mode had a significant main effect on TTP ($F_{(2,81)} = 5.08$, $p = 0.004$), Slope ($F_{(2,81)} = 7.98$, $p = 0.001$), Max ($F_{(2,81)} = 4.92$, $p = 0.010$), Mean ($F_{(2,81)} = 3.62$, $p = 0.031$), and Std ($F_{(2,81)} = 11.03$, $p < 0.001$) of HbO, and TTN ($F_{(2,81)} = 13.17$, $p < 0.001$), Slope ($F_{(2,81)} = 9.85$, $p < 0.001$), Min ($F_{(2,81)} = 3.69$, $p = 0.029$), Mean ($F_{(2,81)} = 4.84$, $p = 0.010$) of HbR. Post-hoc t-tests revealed that the TTP and TTN were significantly earlier in *Active1* and *Active2* modes compared to *Passive* mode ($p < 0.05$). The Slope of HbO, and the Slope and Mean of HbR were significantly larger/smaller in *Active1* and *Active2* modes than in *Passive* mode. In addition, the Max and Mean of HbO, and the Min of HbR were significantly larger/smaller in *Active2* mode than in *Passive* mode ($p < 0.05$). No significant difference was observed between features in *Active1* and *Active2* modes. Regarding the runs, the two-way ANOVA revealed near significant main effects of runs on Slope ($F_{(2,81)} = 2.62$, $p = 0.079$), Max ($F_{(2,81)} = 2.78$, $p = 0.068$), and Mean ($F_{(2,81)} = 2.80$, $p = 0.066$) of HbO. Although the temporal feature difference between runs was not statistically significant, a decreasing trend in mean HbO (or increasing trend in HbR) features was still observed over runs, especially between Run1 and Run2 in *Active1* and *Active2* modes.

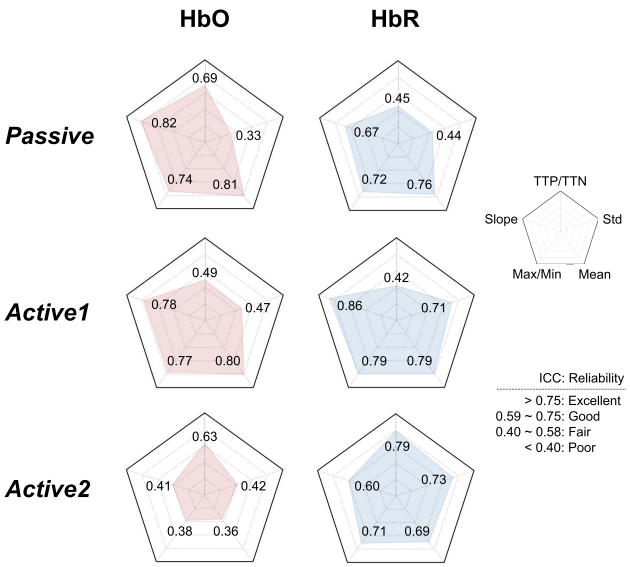

Fig. 8. Radar charts of averaged intraclass correlation coefficients (ICC) for temporal features across all participants.The scale of radar charts ranges from 0 to 1 with an interval of 0.2.

E. Within-Session Temporal Reliability

Fig. 8 displays the within-session temporal reliabilities of various features, as measured by ICC. The results show that in *Passive* mode, the ICC values demonstrate good-to-excellent reliability for Slope, Max/Min, and Mean for both HbO and HbR, while TTP/TTN and Std exhibit poor-to-good reliability. In *Active1* mode, Slope, Max/Min, and Mean show excellent reliability, and TTP and Slope have fair reliability, except for

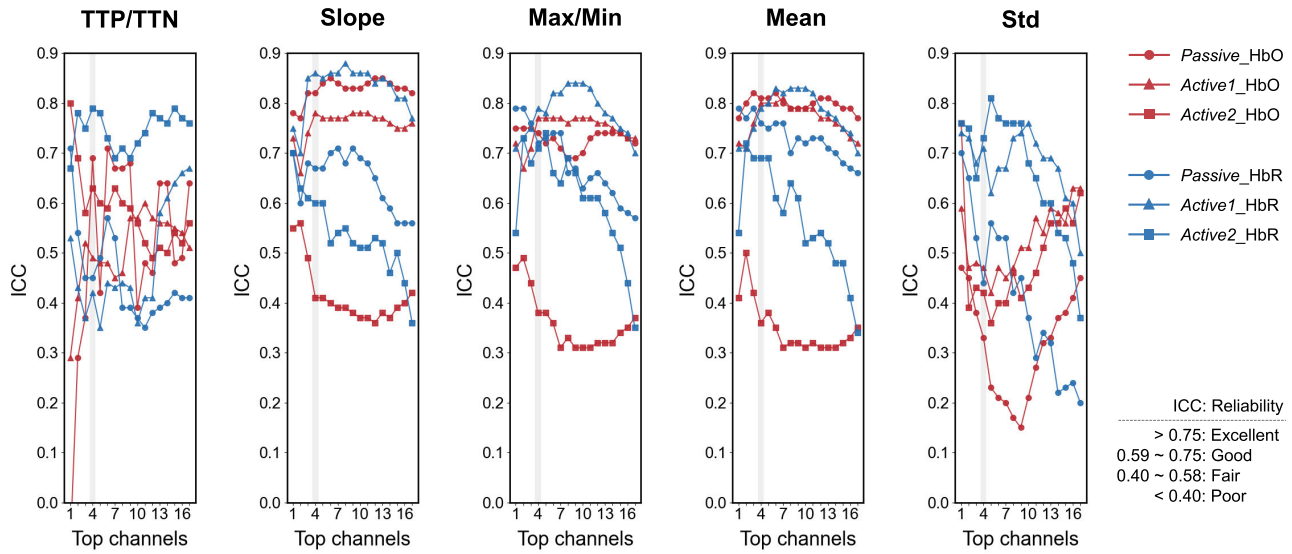

Fig. 9. Intraclass correlation coefficients (ICC) of temporal features for different number of activated channels. Channel activations were sorted according to the t-values of the first run.

TTN in HbR. In *Active2* mode, most features show poor-to-fair reliability in HbO, except for TTP. In contrast, all features exhibit good-to-excellent reliability in HbR. Overall, the ICCs for Slope, Max/Min, and Mean are higher than those for TTP/TTN and Std, especially in *Passive* and *Active1* modes. Additionally, most ICCs in *Active1* mode are higher than those in *Passive* and *Active2* modes. The weak reliabilities in *Passive* mode may be attributed to relatively weak and irregular fNIRS responses. On the other hand, the poor reliability of HbO in *Active2* mode may be due to decreased activation over runs. Notably, the TTP/TTN in *Active2* mode is more reliable than those in *Passive* and *Active1* modes, which could be attributed to the fact that the time to peak (or nadir) is not influenced by the decrease in activation.

We further investigated the effect of channel number on the within-session temporal reliability, and the results are depicted in Fig. 9. In both *Passive* and *Active1* modes, for both HbO and HbR, the within-session temporal reliabilities of Slope, Max/Min, and Mean were consistently good-to-excellent across nearly all channel numbers. However, in *Active2* mode, for both HbO and HbR, the within-session temporal reliabilities of Slope, Max/Min, and Mean ranged between poor and fair across all channel numbers. Compared with the aforementioned features, TTP/TTN and Std displayed less reliability across all channel numbers. In addition, the results revealed that neither using the most activated channel (Top 1) nor all channels (Top 1-17) resulted in the highest ICC. Instead, the average temporal features of the four most activated channels (Top 1-4) exhibited the most reliable performance. This suggests that employing too few or too many channels may compromise the within-session temporal reliability.

IV. Discussion

The within-session reliability of fNIRS responses is the prerequisite for building robot-assisted rehabilitation systems with fNIRS-based neurofeedback. Previous research has proved the cross-session reliability of fNIRS responses to non-motor tasks non-motor tasks [10], [11] and fine-motor tasks [1], [2], [12], [13], [14]. However, it is still unknown whether fNIRS responses remain reliable 1) in gross-motor tasks, 2) within a training session, and 3) for different training parameters. Therefore, this study focused on investigating the within-session reliability of fNIRS responses in robot-assisted upper-limb training. The results revealed that 1) the raw data quality was acceptable in robot-assisted upper-limb training, 2) the within-session spatial reliability was good-to-excellent at the group level and fair-to-good at the individual level, 3) the temporal features, including Slope, Max/Min, and Mean, had good-to-excellent within-session reliability in most cases, 4) the within-session reliability was positively correlated with the intensity of robot-assisted mode (*Passive* < *Active1* < *Active2*), except for the temporal reliability in *Active2* mode. These findings are discussed in greater detail below.

Previous studies have extensively investigated the reliability of fNIRS responses to non-motor tasks and fine-motor tasks [1], [2], [10], [11], [12], [13], [14], [15], with little investigation into gross-motor tasks that may induce large motion artifacts (e.g., upper-limb training). During gross-motor tasks, head or skin movements can lead to decoupling between optodes and the scalp, potentially reducing the within-session reliability. In this study, we utilized the scalp coupling index (SCI) and the peak power (PP) to evaluate the quality of raw data. The obtained results revealed that both the values of SCI and PP exceeded the threshold, indicating that the raw data quality was acceptable in gross motor tasks. In addition, the quality of raw data did not change significantly over time.

The existing studies on examining the reliability of fNIRS responses have relied mainly on assessing the reproducibility of spatial activation patterns [2], [10], [11], [14], [15], [16], [39], [40]. Thus, we conducted the same procedure to evaluate the within-session spatial reliability. As illustrated in Fig. 6,

almost all points were located near the linear regression lines, indicating that the map-wise spatial reliability was excellent at the group level. In most cases, the top four most activated channels were reproducible in both runs, resulting in goo-to-excellent cluster-wise spatial reliability. Compared with the group level, the within-session spatial reliability at the subject level was lower (see Table I). The map- and cluster-wise spatial reliabilities were fair-to-good ($0.44 < R^2 < 0.69$, $0.44 < R_{overlap} < 0.68$) at the subject level. The results revealed acceptable within-session spatial reliability of fNIRS responses.

Given that the majority of fNIRS studies employed temporal features to characterize the changes in hemodynamic responses over time [34] and form neurofeedback with temporal features from multiple channels [41], investigating the within-session reliability of these features has great significance. In this study, we examined the within-session reliability of five commonly used temporal features, including TTN/TTP, Slope, Max/Min, Mean, and Std in terms of ICC. As illustrated in Fig. 8, the reliabilities of Slope, Max/Min, and Mean were higher than those of TTN/TTP and Std in *Passive* and *Active1* mode. However, the within-session reliabilities of Slope, Max/Min, and Mean for HbO in *Active2* mode were poor-to-fair ($0.36 < \text{ICC} < 0.41$), which was significantly lower than those in *Passive* ($0.74 < \text{ICC} < 0.82$) and *Active1* ($0.77 < \text{ICC} < 0.80$) modes.

Closed-loop robot-assisted rehabilitation has the potential to enhance patient engagement and improve recovery efficiency by allowing for timely adjustments of training parameters based on neurofeedback [8], [42]. The question then arises whether fNIRS responses remain reliable under different training parameters. To the best of our knowledge, only one study [15] has examined the reliability of fNIRS response across various training parameters. The velocity of robot-assisted passive grasping was manipulated at three levels (slow at 0.25 Hz, moderate at 0.5 Hz, and fast at 0.75 Hz), and the reliability of fNIRS responses was assessed using the ICC. Their findings revealed that there was almost no reliability of fNIRS responses ($\text{ICC} = 0.002$) for the tested training parameters. The extremely low ICC value may be due to the limited number of subjects (only 6 in their study) or the significant amount of random error in the experiment (residual error variance was several hundred times larger than between-session variance). The small sample size likely resulted in low between-subjects mean squares, while the substantial random error contributed to high error mean squares. According to Eq. 2, both of these factors may lead to a very low ICC. Nevertheless, they attributed the poor reliability to weak fNIRS responses to passive movements [43]. In the current study, we examined the reliability of fNIRS responses during both passive and active movements. Our results revealed that the within-session reliability was positively correlated with the intensity of robot-assisted mode (*Passive* < *Active1* < *Active2*),except for the temporal reliability of HbO in *Active2* mode. Even in Passive mode, the fNIRS responses had excellent within-session spatial ($0.77 < R^2 < 0.94$, $0.75 < R_{overlap} < 1.00$) at the group level and temporal reliabilities (ICC up to 0.81 and 0.82 for Slope and Mean of HbO). This results strongly countered their conclusion that there was no reliability in the fNIRS responses to robot-assisted passive training.

Intriguingly, the within-session temporal reliability of Slope, Max/Min, and Mean for HbO in *Active2* mode was poor-to-fair ($0.36 < \text{ICC} < 0.41$), which was lower than those in *Passive* ($0.74 < \text{ICC} < 0.82$) and *Active1* ($0.77 < \text{ICC} < 0.80$) modes. The low within-session temporal reliability was due to the significant decrease in fNIRS responses during Run2 and Run3, when compared to those recorded during Run1 (see Fig. 7 and Table II). During motor training, motor skill acquisition [44], [45] and fatigue [46], [47] could result in decreased activation. For healthy adults, robot-assisted elbow flexion-extension is not a skilled motor task. Thus, the decreased fNIRS responses observed during high-intensity training in *Active2* mode may be attributed to fatigue, which is consistent with Shibuya et al.'s study [48]. The short rest period in our experimental design could have contributed to fatigue accumulation.

Plichta et al. [49] conducted a study showing that changes in fNIRS responses depend predictably on task paradigm and channel location. Specifically, for simple motor tasks, the strongest fNIRS responses were observed in channels located at the center of the region of interest, with responses attenuated in peripheral areas [34]. Our results revealed strong hemodynamic responses in the dorsal aspect of the left M1 (see Fig. 5), whose activation has been proven to correlate with right elbow movements [50]. In addition, both the within-session spatial and temporal reliabilities of fNIRS responses were higher in this region (see Figs. 6 & 9), suggesting that fNIRS responses can be reliably measured by placing a small number of optodes over the target brain area. Such a few-channel arrangement is more practical in clinical rehabilitation [51].

Several limitations need to be noted in this study. First, the reliability of fNIRS responses needs to be further verified in patients with upper-limb dysfunction. Second, the low within-session reliability of fNIRS responses in *Active2* mode warrants further investigation, which could be achieved by incorporating a longer between-runs rest. Third, short-distance measurement and short-channel regression method [14] should be employed to minimize scalp effect and systemic noise for a more accurate reliability assessment. Additionally, future work should also explore the inter-session reliability to facilitate the application of fNIRS in longitudinally assessing rehabilitation outcomes.

V. Conclusion

In this study, we investigated the within-session reliability of fNIRS responses in robot-assisted upper-limb training. The obtained results revealed fair-to-good spatial reliability at the individual level and good-to-excellent temporal reliability of Slope, Max/Min, and Mean. Besides, the within-session reliability was positively correlated with the intensity of the training mode, especially for the within-session spatial reliability. We also found that fNIRS responses within the most activated brain area had higher spatial and temporal reliabilities. These results indicate that fNIRS

can be used as reliable neurofeedback for constructing closed-loop robot-assisted rehabilitation systems, which will pave the way for the application of fNIRS in clinical neurorehabilitation.

REFERENCES

[1] G. Strangman, R. Goldstein, S. L. Rauch, and J. Stein, "Near-infrared spectroscopy and imaging for investigating stroke rehabilitation: Test-retest reliability and review of the literature," *Arch. Phys. Med. Rehabil.*, vol. 87, no. 12, pp. 12–19, Dec. 2006.

[2] M. M. Plichta et al., "Event-related functional near-infrared spectroscopy (fNIRS) based on craniocerebral correlations: Reproducibility of activation?" *Human Brain Mapping*, vol. 28, no. 8, pp. 733–741, Aug. 2007.

[3] K. Saita et al., "Biofeedback effect of hybrid assistive limb in stroke rehabilitation: A proof of concept study using functional near infrared spectroscopy," *PLoS ONE*, vol. 13, no. 1, Jan. 2018, Art. no. e0191361.

[4] P. Shi, A. Li, and H. Yu, "Response of the cerebral cortex to resistance and non-resistance exercise under different trajectories: A functional near-infrared spectroscopy study," *Frontiers Neurosci.*, vol. 15, p. 1328, Oct. 2021.

[5] C. Duret, A.-G. Grosmaire, and H. I. Krebs, "Robot-assisted therapy in upper extremity hemiparesis: Overview of an evidence-based approach," *Frontiers Neurol.*, vol. 10, p. 412, Apr. 2019.

[6] Q. Miao et al., "Performance-based iterative learning control for task-oriented rehabilitation: A pilot study in robot-assisted bilateral training," *IEEE Trans. Cognit. Develop. Syst.*, vol. 15, no. 4, pp. 2031–2040, Apr. 2021.

[7] R. Bertani, C. Melegari, M. C. De Cola, A. Bramanti, P. Bramanti, and R. S. Calabrò, "Effects of robot-assisted upper limb rehabilitation in stroke patients: A systematic review with meta-analysis," *Neurol. Sci.*, vol. 38, no. 9, pp. 1561–1569, Feb. 2017.

[8] R. Sitaram et al., "Closed-loop brain training: The science of neurofeedback," *Nature Rev. Neurosci.*, vol. 18, no. 2, pp. 86–100, Feb. 2017.

[9] Y. Jiang et al., "Characterization of bimanual cyclical tasks from single-trial EEG-fNIRS measurements," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 146–156, 2022.

[10] M. M. Plichta et al., "Event-related functional near-infrared spectroscopy (fNIRS): Are the measurements reliable?" *NeuroImage*, vol. 31, no. 1, pp. 116–124, May 2006.

[11] A. Blasi, S. Lloyd-Fox, M. H. Johnson, and C. Elwell, "Test–retest reliability of functional near infrared spectroscopy in infants," *Neurophotonics*, vol. 1, no. 2, Sep. 2014, Art. no. 025005.

[12] H. Zhang, Y.-J. Zhang, L. Duan, S.-Y. Ma, C.-M. Lu, and C.-Z. Zhu, "Is resting-state functional connectivity revealed by functional near-infrared spectroscopy test-retest reliable?" *J. Biomed. Opt.*, vol. 16, no. 6, 2011, Art. no. 067008.

[13] H. Zhang, L. Duan, Y.-J. Zhang, C.-M. Lu, H. Liu, and C.-Z. Zhu, "Test–retest assessment of independent component analysis-derived resting-state functional connectivity based on functional near-infrared spectroscopy," *NeuroImage*, vol. 55, no. 2, pp. 607–615, Mar. 2011.

[14] D. G. Wyser et al., "Characterizing reproducibility of cerebral hemodynamic responses when applying short-channel regression in functional near-infrared spectroscopy," *Neurophotonics*, vol. 9, no. 1, Mar. 2022, Art. no. 015004.

[15] S. Bae, Y. Lee, and P.-H. Chang, "There is no test–retest reliability of brain activation induced by robotic passive hand movement: A functional nirs study," *Brain Behav.*, vol. 10, no. 10, 2020, Art. no. e01788.

[16] F. Tian et al., "Test-retest assessment of cortical activation induced by repetitive transcranial magnetic stimulation with brain atlas-guided optical topography," *J. Biomed. Opt.*, vol. 17, no. 11, Nov. 2012, Art. no. 116020.

[17] K.-C. Broscheid, D. Hamacher, J. Lamprecht, M. Sailer, and L. Schega, "Inter-session reliability of functional near-infrared spectroscopy at the prefrontal cortex while walking in multiple sclerosis," *Brain Sci.*, vol. 10, no. 9, p. 643, Sep. 2020.

[18] M. Zhang, C. Sun, Y. Liu, and X. Wu, "A robotic system to deliver multiple physically bimanual tasks via varying force fields," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 688–698, 2022.

[19] C. Sun et al., "Bilateral asymmetry of hand force production in dynamic physically-coupled tasks," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1826–1834, Apr. 2022.

[20] J. W. Peirce, "PsychoPy—Psychophysics software in Python," *J. Neurosci. Methods*, vol. 162, nos. 1–2, pp. 8–13, May 2007.

[21] J. Zheng, P. Shi, M. Fan, S. Liang, S. Li, and H. Yu, "Effects of passive and active training modes of upper-limb rehabilitation robot on cortical activation: A functional near-infrared spectroscopy study," *NeuroReport*, vol. 32, no. 6, pp. 479–488, 2021.

[22] G. A. Zimeo Morais, J. B. Balardin, and J. R. Sato, "FNIRS Optodes' location decider (fOLD): A toolbox for probe arrangement guided by brain regions-of-interest," *Sci. Rep.*, vol. 8, no. 1, p. 3341, Feb. 2018.

[23] C. M. Aasted et al., "Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial," *Neurophotonics*, vol. 2, no. 2, May 2015, Art. no. 020801.

[24] L. Pollonini, C. Olds, H. Abaya, H. Bortfeld, M. S. Beauchamp, and J. S. Oghalai, "Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy," *Hearing Res.*, vol. 309, pp. 84–93, Mar. 2014.

[25] L. Pollonini, H. Bortfeld, and J. S. Oghalai, "PHOEBE: A method for real time mapping of optodes-scalp coupling in functional near-infrared spectroscopy," *Biomed. Opt. Exp.*, vol. 7, no. 12, p. 5104, 2016.

[26] R. Luke et al., "Analysis methods for measuring passive auditory fNIRS responses generated by a block-design paradigm," *Neurophotonics*, vol. 8, no. 2, May 2021, Art. no. 025008.

[27] H. Santosa, X. Zhai, F. Fishburn, and T. Huppert, "The NIRS brain AnalyzIR toolbox," *Algorithms*, vol. 11, no. 5, p. 73, May 2018.

[28] L. Kocsis, P. Herman, and A. Eke, "The modified Beer–Lambert law revisited," *Phys. Med. Biol.*, vol. 51, no. 5, pp. N91–N98, Mar. 2006.

[29] F. Scholkmann and M. Wolf, "General equation for the differential pathlength factor of the frontal human head depending on wavelength and age," *J. Biomed. Opt.*, vol. 18, no. 10, Oct. 2013, Art. no. 105004.

[30] J. W. Barker, A. Aarabi, and T. J. Huppert, "Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS," *Biomed. Opt. Exp.*, vol. 4, no. 8, pp. 1366–1379, Aug. 2013.

[31] H. Santosa, X. Zhai, F. Fishburn, P. J. Sparto, and T. J. Huppert, "Quantitative comparison of correction techniques for removing systemic physiological signal in functional near-infrared spectroscopy studies," *Neurophotonics*, vol. 7, no. 3, Sep. 2020, Art. no. 035009.

[32] J. W. Barker, A. L. Rosso, P. J. Sparto, and T. J. Huppert, "Correction of motion artifacts and serial correlations for real-time functional near-infrared spectroscopy," *Neurophotonics*, vol. 3, no. 3, May 2016, Art. no. 031410.

[33] C. Tegeler, S. C. Strother, J. R. Anderson, and S.-G. Kim, "Reproducibility of bold-based functional MRI obtained at 4 T," *Hum. Brain Mapping*, vol. 7, no. 4, pp. 267–283, 1999.

[34] A. von Lühmann, A. Ortega-Martinez, D. A. Boas, and M. A. Yücel, "Using the general linear model to improve performance in fNIRS single trial analysis and classification: A perspective," *Frontiers Hum. Neurosci.*, vol. 14, p. 30, Feb. 2020.

[35] D. R. Leff et al., "Assessment of the cerebral cortex during motor task behaviours in adults: A systematic review of functional near infrared spectroscopy (fNIRS) studies," *NeuroImage*, vol. 54, no. 4, pp. 2922–2936, Feb. 2011.

[36] L. Li, L. Zeng, Z.-J. Lin, M. Cazzell, and H. Liu, "Tutorial on use of intraclass correlation coefficients for assessing intertest reliability and its application in functional near-infrared spectroscopy–based brain imaging," *J. Biomed. Opt.*, vol. 20, no. 5, May 2015, Art. no. 050801.

[37] T. Johnstone et al., "Stability of amygdala bold response to fearful faces over multiple scan sessions," *NeuroImage*, vol. 25, no. 4, pp. 1112–1123, May 2005.

[38] D. V. Cicchetti and S. A. Sparrow, "Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior," *Amer. J. Mental Deficiency*, vol. 86, no. 2, pp. 127–137, 1981.

[39] M. Schecklmann, A.-C. Ehlis, M. M. Plichta, and A. J. Fallgatter, "Functional near-infrared spectroscopy: A long-term reliable tool for measuring brain activity during verbal fluency," *NeuroImage*, vol. 43, no. 1, pp. 147–155, Oct. 2008.

[40] S. L. Novi et al., "Integration of spatial information increases reproducibility in functional near-infrared spectroscopy," *Frontiers Neurosci.*, vol. 14, p. 746, Jul. 2020.

[41] M. Lührs and R. Goebel, "Turbo-satori: A neurofeedback and brain–computer interface toolbox for real-time functional near-infrared spectroscopy," *Neurophotonics*, vol. 4, no. 4, 2017, Art. no. 041504.

[42] M. Chiappalone and M. Semprini, "Using robots to advance clinical translation in neurorehabilitation," *Sci. Robot.*, vol. 7, no. 64, Mar. 2022, Art. no. eabo1966.

[43] I. Loubinoux et al., "Within-session and between-session reproducibility of cerebral sensorimotor activation: A Test–Retest effect evidenced with functional magnetic resonance imaging," *J. Cerebral Blood Flow Metabolism*, vol. 21, no. 5, pp. 592–607, May 2001.

[44] M. Hatakenaka, I. Miyai, M. Mihara, S. Sakoda, and K. Kubota, "Frontal regions involved in learning of motor skill—A functional NIRS study," *NeuroImage*, vol. 34, no. 1, pp. 109–116, Jan. 2007.

[45] T. Ikegami and G. Taga, "Decrease in cortical activation during learning of a multi-joint discrete motor task," *Exp. Brain Res.*, vol. 191, no. 2, pp. 221–236, Nov. 2008.

[46] H. van Duinen, R. Renken, N. Maurits, and I. Zijdewind, "Effects of motor fatigue on human brain activity, an fMRI study," *NeuroImage*, vol. 35, no. 4, pp. 1438–1449, May 2007.

[47] J. L. Taylor, M. Amann, J. Duchateau, R. Meeusen, and C. L. Rice, "Neural contributions to muscle fatigue: From the brain to the muscle and back again," *Med. Sci. Sports Exercise*, vol. 48, no. 11, pp. 2294–2306, 2016.

[48] K. Shibuya and N. Kuboyama, "Decreased activation in the primary motor cortex area during middle-intensity hand grip exercise to exhaustion in athlete and nonathlete participants," *Perceptual Motor Skills*, vol. 111, no. 1, pp. 19–30, Aug. 2010.

[49] M. M. Plichta, M. J. Herrmann, A.-C. Ehlis, C. G. Baehne, M. M. Richter, and A. J. Fallgatter, "Event-related visual versus blocked motor task: Detection of specific cortical activation patterns with functional near-infrared spectroscopy," *Neuropsychobiology*, vol. 53, no. 2, pp. 77–82, 2006.

[50] J. D. Meier, T. N. Aflalo, S. Kastner, and M. S. A. Graziano, "Complex organization of human primary motor cortex: A high-resolution fMRI study," *J. Neurophysiol.*, vol. 100, no. 4, pp. 1800–1812, Oct. 2008.

[51] S. Ge et al., "A brain–computer interface based on a few-channel EEG-fNIRS bimodal system," *IEEE Access*, vol. 5, pp. 208–218, 2017.