Reducing Robotic Upper-Limb Assessment Time While Maintaining Precision: A Time Series Foundation Model Approach

Faranak Akbarifar[1*], Nooshin Maghsoodi[1], Sean P. Dukelow[2], Stephen H. Scott[3, 4], Parvin Mousavi[1]

[1*]School of Computing, Queen's University, 25 Union St., Kingston, K7L 2N8, ON, Canada.
[2]Hotchkiss Brain Institute, University of Calgary, 3330 Hospital Dr NW, Calgary, T2N 4N1, AB, Canada.
[3]Centre for Neuroscience Studies, Queen's University, 18 Stuart St., Kingston, K7L 3N6, ON, Canada.
[4]Providence Care Hospital, 752 King St., Kingston, K7L 4X3, ON, Canada.

*Corresponding author(s). E-mail(s): f.akbarifar@queensu.ca;
Contributing authors: nooshin.maghsoodi@queensu.ca;
spdukelo@ucalgary.ca; steve.scott@queensu.ca; mousavi@queensu.ca;

Abstract

Purpose: Visually Guided Reaching (VGR) on the Kinarm robot yields sensitive kinematic biomarkers but requires 40–64 reaches, imposing time and fatigue burdens. We evaluate whether time series foundation models can replace unrecorded trials from an early subset of reaches while preserving reliability of standard Kinarm parameters.

Methods: We analyzed VGR speed signals from 461 stroke, and 599 control participants across 4- and 8-target reaching protocols. We withheld all but the first 8 or 16 reaching trials and used ARIMA, MOMENT, and Chronos models, fine-tuned on 70% of subjects, to forecast synthetic trials. We recomputed four kinematic features of reaching (reaction time, movement time, posture speed, max speed) on combined recorded plus forecasted trials and compared to full-length references using ICC(2,1).

Results: Chronos forecasts restored ICC ($\geq$ **0.90**) for all parameters with only 8 real trials plus forecasts, matching the reliability of 24–28 recorded reaches

($\boldsymbol{\Delta ICC \leq 0.07}$). MOMENT yielded intermediate gains, while ARIMA improvements were minimal. Across cohorts and protocols, synthetic trials replaced reaches without significantly compromising feature reliability.
Conclusion: Foundation-model forecasting can greatly shorten Kinarm VGR assessment time. For the most impaired stroke survivors, sessions drop from 4–5 minutes to about 1 minute while preserving kinematic precision. This forecast-augmented paradigm promises efficient robotic evaluations for assessing motor impairments following stroke.

Keywords: Kinarm assessment, Visually guided reaching, time series forecasting, foundation models

1 Introduction

Upper-limb robotic devices are increasingly used in research laboratories to assess stroke-related impairments. These devices yield precise, quantitative performance measures that complement clinical scales [1–4]. An example is the *Kinarm Exoskeleton Lab*, an interactive robotic device whose kinematic measures correlate with traditional clinical scales such as the Fugl–Meyer Assessment [5] and the Chedoke–McMaster Stroke Assessment (CMSA) [6] [7–9]. Kinarm Standard Tests (KST) [10, 11] provides a suite of behavioural tasks to assess a broad range of sensory, motor and cognitive functions and has been used to assess impairments associated with stroke [12–15] and a broad range of other neurological diseases and injuries [16–19]. During each task, the robotic system records precise upper limb location and velocity data, quantifies task-specific spatiotemporal features, combines them into a single Task Score, and leverages statistical models to compare a participant's performance against healthy individuals, considering factors such as age, sex, and handedness [1, 7, 8, 20].

Visually Guided Reaching (VGR) task [12] is commonly used to quantify upper limb motor function through spatiotemporal features that are categorized into 5 attributes of sensorimotor control: upper-limb postural control, reaction time, initial movement, movement corrections and total movement metrics, and are sensitive to identify mild impairment [9]. A practical limitation is duration: a full VGR block presently requires 40–64 reaches ($\approx$ 4–5 minutes), which constrains throughput and increases fatigue. Shorter sessions reduce subject fatigue and overall assessment time, improve throughput and scheduling, and can make robotic assessment more tolerable for individuals with severe motor and cognitive impairments. In the current paper, we ask whether a forecast-augmented VGR session can maintain measurement precision and clinical decisions while substantially reducing on-robot time.

Time reduction on Kinarm has been pursued at three complementary levels. (i) Task redundancy: Quantitative similarity between Object Hit (OH) [14] and Object Hit and Avoid (OHA) [13] has been tested by predicting the OH parameter set from OHA with Fast Orthogonal Search (FOS) [21] and then classifying stroke vs. control using the predicted OH parameters; classification performance was comparable to using the measured OH parameters, supporting removal of one member of the

OH/OHA pair in batteries where both appear [22]. (ii) Protocol flow: Using non-linear hierarchical ordering [23] on discretized (pass/fail) task outcomes, an inferred partial order lets early tasks predict downstream outcomes with high confidence, enabling conditional skipping of later tasks. In one cohort this yielded 97% confidence for a subgroup and ≥8 minutes saved [24]; generalized across five standard tasks, a hierarchical task-selection strategy that combines ordering, parameter ranking, and cross-task prediction reported large (≥50%, up to 90%) time reductions [25]. (iii) Within-task VGR: Reduced-trial VGR protocols (fewer reaches) were evaluated by reestimating the VGR parameters under each scheme and validating diagnostic utility with an SVM classifier; sensitivity, specificity, and accuracy were nearly identical to the full protocol, indicating substantial recording-time savings at minimal loss in classification performance [26].

The strategies above shorten assessment time by removing tasks, reordering tasks to skip later ones, or recording fewer trials. Here, we take a complementary route: retain the existing analytics and scoring pipeline, but replace a subset of recorded trials with model-generated forecasts conditioned on the subject's early trials. This keeps the protocol and outcome measures intact while cutting the recording time and preserving downstream interpretability (the same speed-derived parameters are computed on recorded+forecasted trials). We adopt the state-of-the-art time series foundation models because they learn transferable priors on extensive time series data and are adapted with limited domain data [27]. In our experiments we instantiate this idea with two representative forecasters, Chronos [28] and MOMENT [29], compared with a classical baseline (ARIMA) [30]. To our knowledge, these models have not been evaluated for robotic stroke assessment nor for forecasting reaching patterns in VGR. We therefore pose the unrecorded trials as conditional sequence generation given early trials as context and measure success by agreement of downstream KST parameters computed on real + forecasted trials with those from full recordings.

We introduce a forecast-augmented version of the Kinarm VGR assessment that preserves the standard task design and analytics while reducing recorded trials; formulate a time-series, direction-conditioned forecasting approach that generates subject-specific trials from a short recorded context without altering canonical VGR features or their computation [11, 12]; design a subject-level, multi-site evaluation across stroke and control cohorts that compares forecast-augmented sessions to full-length references using established reliability metrics (e.g., ICC); and quantify uncertainty by combining variability from subject sampling with variability due to repeated selection of forecasted trials. The remainder of this article describes the dataset and preprocessing, then details the forecasting setup and evaluation protocol; we next present results and ablations, and conclude with clinical implications, limitations, and future work.

2 Methods

2.1 Participants

Participants with stroke and healthy volunteers were recruited from the communities of Kingston, Ontario and Calgary, Alberta. Participants with stroke were recruited

from the inpatient acute stroke or stroke rehabilitation units at the Foothills Medical Centre, the Dr. Vernon Fanning Care Centre in Calgary, Alberta, Canada, and Providence Care, and St Mary's of the Lake Hospital, Kingston, Ontario, Canada. Participants with stroke were assessed using a variety of clinical measures, including the Chedoke-McMaster Stroke Assessment [6]. This study was reviewed and approved by the University of Calgary Conjoint Health Research Ethics Board and the Queen's University Research Ethics Board. All participants gave their written informed consent to have their data collected for research purposes before performing the assessment. The inclusion criteria for participants with stroke included a first-identified stroke within the last 35 days and being at least 18 years old. Individuals were excluded if they had other neurological disorders (e.g. Parkinson's, Multiple Sclerosis), upper limb orthopedic impairments, difficulties understanding tasks instructions or signs of apraxia [31]. Healthy control participants fulfilled the same inclusion and exclusion criteria, except had no stroke history, and did not have any known neurological or upper arm musculoskeletal conditions.

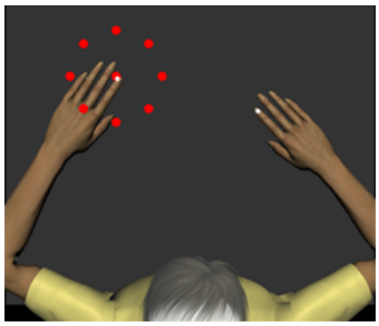
2.2 Apparatus

All data in this study were gathered using the Kinarm Exoskeleton Lab [11][7], an interactive robotic system equipped with an integrated virtual reality system designed to measure upper limb sensorimotor function. The participants were positioned in an adjustable-height chair and the robotic linkages were adjusted to align the robot's joints with the participants' shoulder and elbow joints, enabling them to both flex and extend their shoulders and elbows while supporting the arm against gravity (Figure 1a). The virtual reality display in this interactive robotic system provided visual feedback aligned with the horizontal workspace with a physical barrier obscuring direct view of the participants' arms.

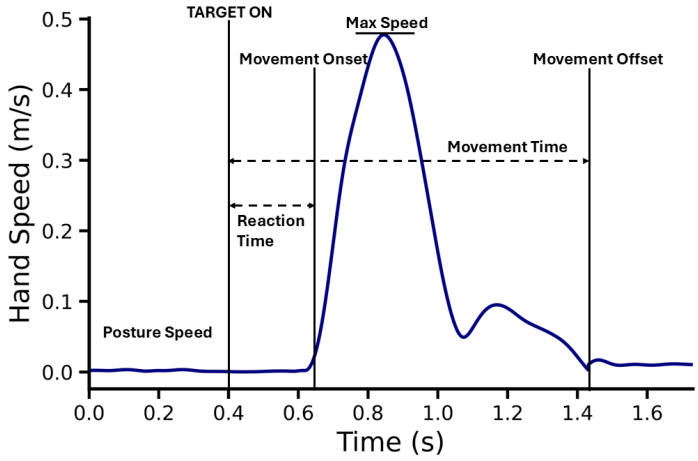
2.3 VGR task

In the VGR task, the integrated virtual reality system provides visual feedback on hand position and spatial objects aligned within the horizontal workspace (Figure 1b). Over the years, several closely related variants of the Kinarm VGR task have been developed. In each version, participants view their hand as a bright circle on the exoskeleton's co-located virtual-reality display and reach toward one of several 10 cm-distant peripheral targets (Figure 1b). In the earliest "reachout_std" 8-target protocol (P1), eight equiangular targets (45° spacing) appear once per block over eight blocks (64 trials), with two "catch" trials per block (no target) to assess postural stability; only the outward reaches are analysed [12]. A second 8-target variant (P2: "8-target 10 cm reach") preserves the same spatial layout and total trial count but implements updated hold-times and inter-trial intervals in the Kinarm software. More recently, the "4-target In&Out 10 cm" protocol (P3) reduced the workspace to four targets at 0°, 90°, 180° and 270°: here each peripheral direction is reached five times in a pseudo-random order, followed by a return to the central target (when the central target re-illuminates), and four catch trials (no target) are interspersed, yielding up to 40 outbound–inbound movements per session. Across all variants, hand speed and

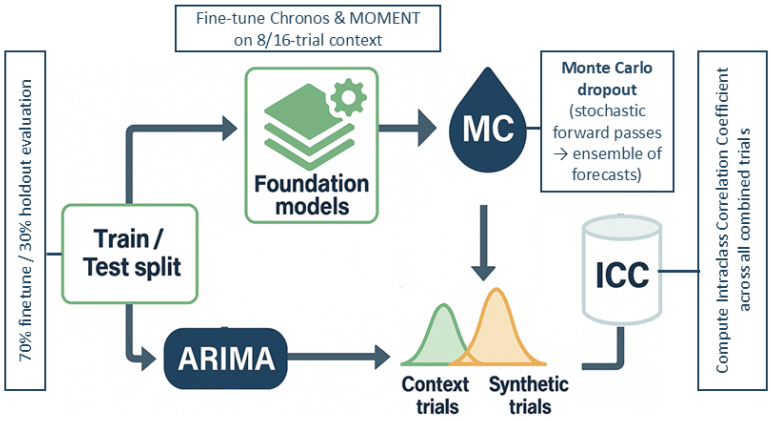
(a) Kinarm exoskeleton robot



(b) VGR virtual workspace



(c) A sample speed trial



(d) Forecast-augmented assessment pipeline

Fig. 1: Apparatus and task. (a) Participant in the Kinarm Exoskeleton Lab. (b) Virtual workspace for the Visually Guided Reaching (VGR) task: the hand cursor moves from the central start target to one of the eight peripheral targets that appear in random order. (c) Representative control-trial hand-speed trace illustrating computation of the four selected Kinarm parameters—posture speed, reaction time (TARGET ON→Movement Onset), movement time (Movement Onset→Movement Offset), and max speed— with TARGET ON, Movement Onset and Movement Offset marked. (d) Study pipeline: an 8-trial context is fed into a fine-tuned foundation model (Chronos or MOMENT) with Monte-Carlo dropout to generate additional synthetic trials; real + forecasted trials are then used to recompute KST-defined speed parameters and their intraclass correlation coefficients (ICC).

Table 1: Number of control and participants with stroke by VGR task variant.

Task variant	Control (n)	Stroke (n)	Max trials per participant
P1: reachout_std (8-target)	93	21	64
P2: 8-target 10 cm reach	45	337	64
P3: 4-target In&Out 10 cm RT	461	73	40

position are sampled at 1 kHz. Fourteen spatiotemporal features are extracted per trial (see [12] for full definitions) and form the basis for normative Task Scores. In the present study, we evaluate forecast quality using four task parameters (posture speed, reaction time, movement time, and max speed) within our reliability and forecasting analyses.

2.4 Dataset formation

For every participant we extracted the hand speed signal of the more affected limb for participants with stroke and the non-dominant limb in controls—from the raw Kinarm sensor files (1 KHz sampling).

Data were collected over several years under the VGR protocols mentioned in Table 1. To obtain a temporally consistent snippet from every file we anchored each trial to the TARGET_ON event, extracted a window beginning 200 ms *before* that event and ending at

$$t_{\text{stop}} = \text{TARGET\_ON} + \text{ReactionTime} + \text{TotalMovementTime},$$

where ReactionTime and TotalMovementTime are provided per trial in the Kinarm log. This window spans late movement preparation, and movement execution for all participants regardless of protocol year. For participants with stroke that could not reach the end target, the window ended at the end of trial.

Each extracted speed trace was linearly resampled to a fixed $L = 64$ points—first, to normalise trial length across subjects whose movements naturally vary in duration, and second, because the Chronos family of time-series foundation models shows best performance when forecast windows have lengths of up to 64 samples [28].

Following the KST manual [11], we compute four Kinarm parameters per trial and use them in all downstream reliability analyses. Panel (c) of Figure 1 annotates a representative speed signal with the event times used for each definition: posture speed — the median hand speed during the stationary hold at the start; reaction time — the interval from TARGET_ON to Movement Onset; movement time — the interval from Movement Onset to Movement Offset; and maximum speed — the peak hand speed between Movement Onset and Movement Offset. For a given set of trials, each parameter is summarized by the mean or median of its per-trial values, depending on the parameter.

2.5 Forecasting Models

We compare a statistical baseline (ARIMA) with two foundation model forecasters (MOMENT, Chronos) for conditional generation of unrecorded reaches. Each model consumes a context of $c \in \{8, 16\}$ recorded reaches, where every reach is a hand speed sequence resampled to $L = 64$ samples (Section 2.4), and outputs a 64-sample forecast.

ARIMA

We use a classical ARIMA(p, d, q) [30] baseline on each subject's concatenated context signal y_t ($t = 1, \ldots, 512$, i.e., 8 trials $\times$ 64 samples). For each subject we choose (p, d, q) by minimizing AICc over a small grid $p, q \in \{0, 1, 2, 3\}$ and $d \in \{0, 1\}$; if models tie, we prefer the lower order. After fitting the selected model, we generate multiple stochastic forecast paths by simulating future innovations $\varepsilon_t \sim \mathcal{N}(0, \hat{\sigma}^2)$ and propagating them through the fitted ARIMA state. Concretely, for a requested K synthetic trials (e.g., $K \in \{8, 16, 24, \ldots\}$), we forecast a horizon $h = 64K$ samples and draw M independent paths $\{\hat{y}^{(m)}_{512+1:512+h}\}_{m=1}^{M}$. Each length-$h$ path is then partitioned into K non-overlapping 64-sample segments in chronological order to serve as synthetic trials, which we append to the 8-trial context before recomputing the four VGR parameters. Unlike MOMENT/Chronos, the ARIMA baseline is direction-agnostic (no conditioning on target direction); its role is to provide a transparent statistical comparator with uncertainty coming from the Gaussian innovations $\varepsilon_t \sim \mathcal{N}(0, \hat{\sigma}^2)$.

MOMENT

Our goal was to forecast a subject's next reach *for a specified direction* using that same subject's prior reaches. We adapt `MOMENT-1-small` for single-trial, direction-conditioned forecasting [29]. Each subject contributes 9 hand speed trials (1 channel, 64 samples): 8 as context (one per direction) and 1 representative as the forecast target; directions are encoded as $d \in \{0, \ldots, 7\}$. A subject-level sampler yields batches of one subject so each optimization step processes context $\in \mathbb{R}^{1\times8\times64}$, context_dirs $\in \mathbb{Z}^8$, forecast $\in \mathbb{R}^{1\times64}$, *and*forecast_dir $\in \mathbb{Z}$. We freeze the backbone embedder and also train a lightweight aggregator head (dropout p=0.1) that: (i) encodes each context trial; (ii) applies a normalized mask that keeps only embeddings whose direction matches forecast_dir; (iii) concatenates the masked average with a learned embedding of forecast_dir; and (iv) projects to a 64-sample forecast. We minimize DTW loss with Adam (lr 5×10^{-4}, weight decay 10^{-5}), early stopping (patience 15), and a max of 100 epochs. Data splits are at the subject level to prevent leakage. At test time, dropout remains active in the head to draw 8 Monte Carlo [32] samples per direction.

Chronos

We fine-tuned a Chronos forecaster [28] (`chronos-t5-small`) on resampled hand-speed sequences using an 8×64 context and a 64-token future target. To condition forecasts on movement direction *without* altering Chronos's numeric (uniform-bin) tokenizer, we reserved eight additional special tokens in the vocabulary and used them as direction controls, `[DIR_0]`... `[DIR_7]`. For each training instance we prepended the matching `[DIR_d]` to the encoder input, and used the next 64 tokens as labels (teacher

forcing) with a deterministic last-window split per series. Training used HuggingFace `Trainer` with AdamW (fused), learning rate 1×10^{-4} , per-device batch size 32 with gradient accumulation 2, logging every 500 steps, evaluation/checkpointing every 2,000 steps (keeping the latest 3), and early stopping (patience=5). At inference, we prepend the requested direction token and draw 64 stochastic sample paths via top-k decoding; decoded forecasts (de-quantized using the learned scale) are then combined with the eight recorded context trials to compute KST features and ICC.

2.6 Forecast Experiments

In the context of time series forecasting, a "context" refers to a known initial portion of sequential data provided to the model, serving as historical information from which future points are predicted. For our study, the context specifically consisted of a subset of early recorded trials selected to represent every unique movement direction once (8 reaches), providing a representative initial snapshot of each participant's reaching behavior. This initial set of context trials enables the forecasting model to learn subject-specific temporal patterns and general movement dynamics, thus allowing reliable prediction of subsequent, unrecorded reaches.

To ensure the context contained at least one exemplar of each movement direction, we systematically selected the first eight trials with unique direction codes in the following way. For the 8-target protocols (P1 and P2), we selected the first occurrence of each of the eight unique target directions, yielding eight context trials. In the 4-target protocol (P3), we similarly selected the first "reach-out" trial and "return" trial for each of the four targets, again totalling eight context trials. We did a similar process for selecting 16 trials based on movements associated with the first two presentations of each target.

Trials beyond these context trials were treated as future events for forecasting. Participants with fewer than eight valid trials were excluded from the primary (8-trial) analysis. For the 16-trial sensitivity analysis, we analogously required at least 16 valid trials.

After preprocessing, the dataset is represented by

$$X^{(i)} \in \mathbb{R}^{8\times 64},$$

where the rows are context trials ordered by target direction and the columns are resampled speed samples. The accompanying labels are: stroke/control status, CMSA score, protocol identifier (P1–P3), trial-level metadata (reaction time, movement time, etc.).

We merged our stroke and control datasets and conducted a randomized 70–30 train-test split, ensuring that each participant's data appeared exclusively in one set. The foundation models, MOMENT and Chronos, were fine-tuned on the training set, with each trial uniformly resampled to a fixed length of 64 samples for compatibility with model input constraints. Forecasting began from eight context trials representing the first trials performed in each direction (two trials per direction for the 4-target variant, one trial per direction for the 8-target variants). We created multiple forecast groups using **ARIMA** (statistical baseline) [30], **MOMENT** (transformer-based foundation model) [29], **Chronos** (tokenized foundation model) [28]. Reliability of

kinematic features from combined real+forecasted trials for each method was assessed using the ICC, as detailed in the following subsection.

To assess robustness to the amount of historical context, we repeated all preprocessing, modeling, and evaluation steps with a 16-trial context, i.e., $X^{(i)} \in \mathbb{R}^{16\times 64}$. All other settings were unchanged. Comparative results for 8- and 16-trial contexts are reported in the Results.

2.7 Evaluation metrics

As stated in subsection 2.4, we compute per-trial values for four KST speed-based parameters. For each parameter and cohort, we summarize trial-level repeatability at the subject level using the ICC, and we report the mean and 95% confidence interval across subjects.

We estimate agreement between paired measurements using a two-way random-effects, absolute-agreement intraclass correlation, ICC(2,1) [33, 34]. For interpretability we follow Koo and Li (2016): ICC $<$ 0.50 = poor, 0.50–0.75 = moderate, 0.75–0.90 = good, and $>$ 0.90 = excellent [35].

Error bars in our figures reflect two sources of variability, depending on the condition. For the recorded-only baseline, we estimate uncertainty via a bootstrap over subjects (with replacement, B=1000) and report the 2.5–97.5th percentiles of ICC across bootstrap replicates. For forecast-augmented points, we quantify forecast selection variability by repeatedly sampling the required number of forecasted trials from each subject's dropout-generated forecast pool and recomputing ICC; we display the mean across repeats with $\pm$ one standard deviation as error bars.

To assess the effect of adding forecasted trials, we first construct a recorded-only benchmark curve. For each trial count X, we compute a subject-level metric (e.g., the mean reaction time over X recorded trials) and then the ICC between the X-trial metric and the complete-trial metric; uncertainty is obtained by bootstrapping subjects (B=1000), as described above.

For forecast-augmented protocols, we fix a context size (e.g., c=8 recorded trials) and add $k \in \{0, 8, 16, \ldots\}$ forecasted trials per subject. Forecasts are produced by keeping dropout active at inference to form a forecast pool per subject; for each k we draw R independent selections from this pool, recompute the metric on the (c+k) trials, and evaluate ICC against the complete-trial metric. Plotted squares show the mean ICC across the R selections with vertical error bars indicating the standard deviation (forecast-selection variability). Improvements are summarized by the absolute change ΔICC and the percent change relative to the recorded-only baseline, both averaged across subjects. For model benchmarking, we repeat the identical procedure for each forecaster (MOMENT, Chronos) and the ARIMA baseline using the same preprocessing and splits.

3 Results

3.1 Participants

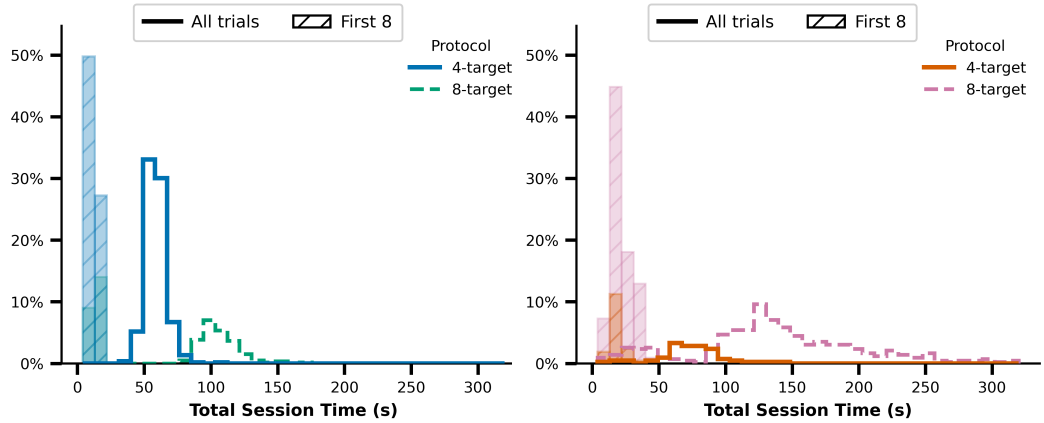
The dataset comprised robotic assessments from 461 participants with stroke and 599 neurologically healthy controls. Key demographic and clinical characteristics for both groups appear in Table 2. Right-handedness predominated (90% in controls, 92% in participants with stroke), with mixed handedness recorded in only five individuals (four controls, one participant with stroke) according to the Edinburgh Handedness Inventory [36]. Based on the Chedoke–McMaster Stroke Assessment, 73% of participants with stroke showed motor deficits in their more-affected limb. Every participant completed one of three VGR task variants, but the distribution was uneven: control participants were mainly tested on the 4-target version, whereas participants with stroke mainly undertook the 8-target protocols (Table 1). Consequently, all analyses are reported separately by *cohort* (control vs. stroke) and by *protocol* (4- vs. 8-target).

To quantify the time cost of full-length recordings, we computed each participant's total session time as the sum of trial durations, and, for comparison, a first-8 total by summing only the first eight recorded trials. Figure 2 summarizes these distributions by cohort and protocol. Panels (a–b) show that trimming to eight trials shifts the session-time histograms markedly leftward and narrows their spread for both cohorts, whereas the stroke distributions remain right-shifted relative to controls (*i.e.*, group differences are preserved rather than collapsed by truncation. Panel (c) (ECDFs) highlights the practical gain: with the first-8 design, all sessions finish within about a minute across protocols, whereas the all-trials curves extend well beyond two to five minutes, especially for participants with stroke in the 8-target task.

Table 2: Demographic and clinical data for control and participants with stroke.

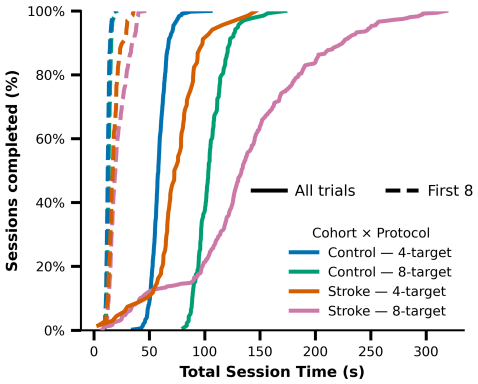
	Control (n=599)	**Stroke (n=461)**
Age	46 (18–93)	62 (18–92)
Sex	167 M, 201 F	221 M, 115 F, 1 O
Dominant hand	330 R, 34 L, 4 A	310 R, 26 L, 1 A
Days since stroke	—	16 (1–84)
Type of stroke	—	290 I, 45 H, 2 U
CMSA [1–7]		
More affected arm	—	[16, 40, 48, 23, 58, 56, 96]
Less affected arm	—	[0, 0, 0, 1, 12, 61, 263]

Notes. Data are mean (range). CMSA = Chedoke–McMaster Stroke Assessment; M = male; F = female; O = other; R = right; L = left; A = ambidextrous; I = ischemic stroke; H = hemorrhagic stroke; U = unknown.



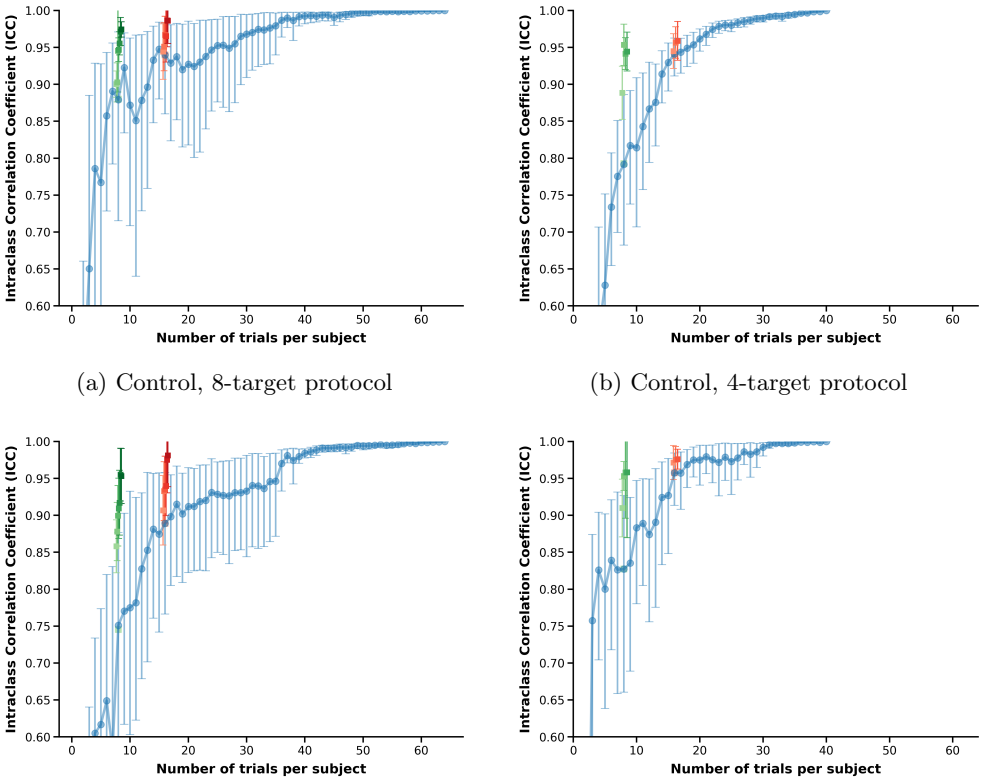
(a) Control — total session times (histogram)

(b) Stroke — total session times (histogram)



(c) ECDF of total session time by cohort and protocol

Fig. 2: Session duration with all trials vs. first eight, by cohort and protocol. (a) Control and (b) Stroke: normalized histograms of total VGR session time (sum of trial durations), overlaid by protocol (4-target vs. 8-target). Within each panel, solid outlines show all recorded trials and hatched bars show first eight trials only; binning is shared and percentages sum to 100% per cohort. (c) Empirical cumulative distribution (ECDF) of total session time by cohort × protocol. Solid curves use all trials; dashed curves use the first eight. Across cohorts and protocols, the first-8 condition shifts distributions leftward and steepens the ECDF, indicating markedly shorter sessions, while preserving the expected ordering (8-target > 4-target; stroke > control).



(a) Control, 8-target protocol

(b) Control, 4-target protocol

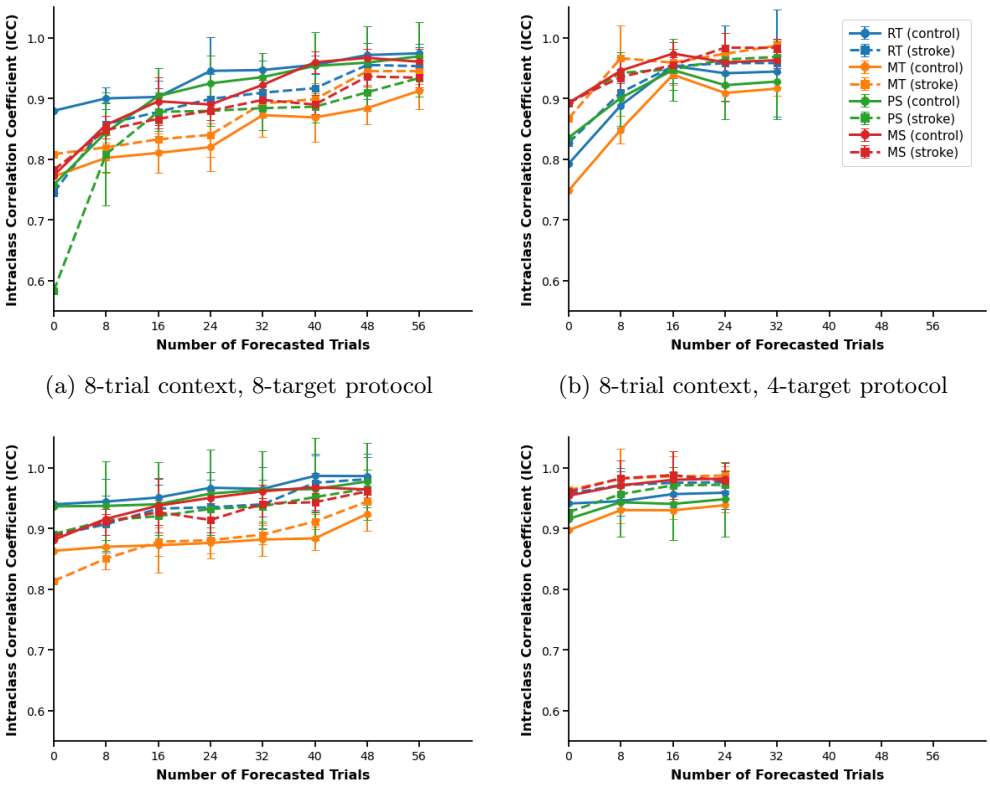
(c) Stroke, 8-target protocol

(d) Stroke, 4-target protocol

Fig. 3: Reaction time reliability (ICC) as a function of total trials for the Chronos model. (a) Control, 8-target; (b) Control, 4-target; (c) Stroke, 8-target; ; (d) Stroke, 4-target. Blue curve: ICC between X-trial RT (median of X recorded trials) and complete-trial RT; error bars from subject bootstrap (B=1000). Squares: ICC for forecast-augmented protocols (8 recorded + k forecasts); error bars incorporate subject bootstrap and forecast-selection repeats.

3.2 Baseline reliability

Figure 3 illustrates the reliability of each Kinarm parameter (Reaction Time, Movement Time, Posture Speed, and Max Speed) as a function of the number of recorded trials, using ICCs calculated through bootstrapping with 95% confidence intervals. To visualize the effect of forecast augmentation, square markers are overlaid on the same ICC–versus–trial-count curves: the lightest square denotes the ICC from the eight recorded "context" trials (no augmentation), and progressively darker squares indicate the ICC achieved after adding forecast-generated trials in blocks of eight (i.e., +8, +16, +24, . . .). Vertical error bars on these squares show the standard deviation across Monte Carlo–dropout passes used to generate forecasts. The horizontal axis



(a) 8-trial context, 8-target protocol

(b) 8-trial context, 4-target protocol

(c) 16-trial context, 8-target protocol

(d) 16-trial context, 4-target protocol

Fig. 4: Intraclass Correlation Coefficient (ICC) for reaction time (RT), movement time (MT), posture speed (PS), and max speed (MS), plotted as a function of the number of forecasted trials added to a fixed context. Each subplot corresponds to one of four task configurations: (a) 8-trial context, 8-target protocol; (b) 8-trial context, 4-target protocol; (c) 16-trial context, 8-target protocol; (d) 16-trial context, 4-target protocol. Solid lines with circular markers represent control participants; dashed lines with square markers represent participants with stroke. Colored lines indicate different kinematic parameters: blue = RT, orange = MT, green = PS, red = MS. The x-axis spans from 0 (context only) to 56 forecasted trials added in increments of 8, maintaining a consistent scale across all panels. Y-axis values denote ICC calculated from real + forecasted trials, with error bars representing standard deviation. The plots illustrate how forecasted trials affect reliability across groups, parameters, and task designs, with higher ICC generally achieved as more forecasted trials are included—though the extent of improvement varies by parameter and cohort.

Table 3: Forecasting performance for the 8-target VGR protocol.

Parameter	Cohort	Method	Context = 8			Context = 16		
			ICC@8	Best	Δ	ICC@16	Best	Δ
Reaction Time	Control	Chronos	0.88	0.97	0.09	0.94	0.99	0.05
		MOMENT	0.88	0.93	0.05	0.94	0.96	0.02
		ARIMA	0.88	0.90	0.02	0.94	0.93	0.01
	Stroke	Chronos	0.75	0.95	0.21	0.89	0.98	0.09
		MOMENT	0.75	0.90	0.15	0.89	0.95	0.06
		ARIMA	0.75	0.85	0.10	0.89	0.90	0.01
Movement Time	Control	Chronos	0.77	0.91	0.14	0.86	0.92	0.06
		MOMENT	0.77	0.89	0.12	0.86	0.90	0.04
		ARIMA	0.77	0.85	0.08	0.86	0.88	0.02
	Stroke	Chronos	0.81	0.95	0.14	0.81	0.94	0.13
		MOMENT	0.81	0.90	0.09	0.81	0.92	0.11
		ARIMA	0.81	0.88	0.07	0.81	0.90	0.09
Posture Speed	Control	Chronos	0.76	0.97	0.21	0.94	0.98	0.04
		MOMENT	0.76	0.95	0.19	0.94	0.96	0.02
		ARIMA	0.76	0.90	0.14	0.94	0.92	-0.02
	Stroke	Chronos	0.58	0.93	0.35	0.89	0.97	0.07
		MOMENT	0.58	0.88	0.30	0.89	0.94	0.05
		ARIMA	0.58	0.82	0.24	0.89	0.90	0.01
Max Speed	Control	Chronos	0.77	0.96	0.19	0.88	0.96	0.08
		MOMENT	0.77	0.94	0.17	0.88	0.94	0.06
		ARIMA	0.77	0.90	0.13	0.88	0.91	0.03
	Stroke	Chronos	0.78	0.93	0.15	0.89	0.96	0.07
		MOMENT	0.78	0.90	0.12	0.89	0.95	0.06
		ARIMA	0.78	0.86	0.08	0.89	0.92	0.03

for the 4-target protocol has been extended to align scale consistency across subplots, facilitating direct comparison of reliability metrics across task configurations. For control participants, the ICC curves demonstrate that median ICC exceeded the excellent threshold (≥ 0.90 per Koo and Li, 2016 [35]) after approximately 10–14 recorded trials across all parameters and task variants. For participants with stroke, reaching similar reliability required an additional 4–6 recorded trials. These findings establish baseline benchmarks against which forecast-augmented strategies are compared.

3.3 Reliability with forecasts

Figure 4 comprehensively displays ICC trajectories when combining real trials with Chronos forecasted data across participant groups, task variants, and kinematic parameters. Each subplot consistently ranges from 0 to 56 forecasted trials (increments

Table 4: Forecasting performance for the 4-target VGR protocol.

Parameter	**Cohort**	**Method**	Context = 8			Context = 16		
			ICC@8	Best	Δ	ICC@16	Best	Δ
Reaction Time	Control	Chronos	0.79	0.94	0.15	0.94	0.96	0.02
		MOMENT	0.79	0.90	0.11	0.94	0.94	0.00
		ARIMA	0.79	0.88	0.09	0.94	0.92	-0.02
	Stroke	Chronos	0.83	0.96	0.13	0.96	0.98	0.02
		MOMENT	0.83	0.92	0.09	0.96	0.96	0.00
		ARIMA	0.83	0.88	0.05	0.96	0.92	-0.04
Movement Time	Control	Chronos	0.75	0.92	0.17	0.90	0.94	0.04
		MOMENT	0.75	0.88	0.13	0.90	0.92	0.02
		ARIMA	0.75	0.85	0.10	0.90	0.90	0.00
	Stroke	Chronos	0.87	0.99	0.12	0.96	0.99	0.02
		MOMENT	0.87	0.95	0.08	0.96	0.98	0.02
		ARIMA	0.87	0.90	0.03	0.96	0.94	-0.02
Posture Speed	Control	Chronos	0.84	0.93	0.09	0.92	0.95	0.03
		MOMENT	0.84	0.90	0.06	0.92	0.93	0.01
		ARIMA	0.84	0.88	0.04	0.92	0.92	0.00
	Stroke	Chronos	0.89	0.96	0.07	0.93	0.97	0.04
		MOMENT	0.89	0.85	0.04	0.93	0.94	0.01
		ARIMA	0.89	0.80	-0.09	0.93	0.90	-0.03
Max Speed	Control	Chronos	0.89	0.96	0.07	0.95	0.98	0.03
		MOMENT	0.89	0.94	0.05	0.95	0.96	0.01
		ARIMA	0.89	0.90	0.01	0.95	0.94	-0.01
	Stroke	Chronos	0.89	0.98	0.09	0.96	0.98	0.02
		MOMENT	0.89	0.95	0.06	0.96	0.98	0.02
		ARIMA	0.89	0.90	0.01	0.96	0.94	-0.02

of 8), providing an intuitive visualization of reliability changes as forecasts are progressively added. Notably, Chronos forecasts substantially enhanced reliability across parameters and task variants compared to when only the actual 8 trials are sampled. For instance, panel (a) highlights that control participants, under the 8-target protocol with an initial 8-trial context with forecasted trials, achieved ICC values between 0.90–0.97 for all parameters with forecast augmentation—effectively matching reliability levels typically observed only after recording 24–28 actual trials. In participants with stroke under the same protocol (panel b), Reaction Time and Movement Time rapidly achieved ICC $\geq$ 0.90 with as few as eight forecasted trials appended to an 8-trial context. Posture Speed and Max Speed for participants with stroke also significantly benefited from forecasts, achieving comparable ICC values after a slightly expanded (16-trial) context.

3.4 Method comparison

Tables 3 and 4 summarize forecasting precision for Chronos, MOMENT, and ARIMA at the clinically relevant checkpoints of eight and sixteen recorded trials. For each method and context length, we report the following: — ICC8/16, the reliability attainable when forecasts replace the remaining recorded trials; — Best ICC, the maximum reliability observed anywhere along the forecast-augmented curve; — ΔICC, the gain relative to the baseline ICC at the same context length.

Chronos delivered the largest ΔICC in 29 of the 32 parameter–cohort–protocol cells (91 %), with typical gains of 0.10–0.21 for controls and 0.12–0.35 for stroke. MOMENT closed roughly half that gap, whereas ARIMA rarely exceeded a 0.10 ICC improvement and sometimes degraded reliability (e.g. Posture Speed *stroke*, 4-target: ΔICC = –0.02).

4 Discussion

This study is the first to demonstrate that a foundation-model approach can improve kinematic-based assessment of motor performance associated with stroke. Chronos, fine-tuned on only 70 % of our mixed stroke/control cohort and prompted with eight context trials, lifted ICC to within 5 points of the full-length reference for three of the four canonical parameters in controls and two in stroke (Figures 3, Table 3). With a 16-trial context the gap narrowed to $\Delta\text{ICC} \leq 0.07$ for *all* parameters and cohorts, effectively restoring assessment fidelity while discarding 60–75 % of the original recording time.

MOMENT inherited the same pretraining recipe as Chronos but lagged by $\sim$ 0.03–0.06 ICC points in most scenarios, mirroring the relative ranking reported in the open-data Time-Bench benchmark [29]. The classical ARIMA baseline provided a useful lower bound: although its point forecasts were accurate for trajectories with highly stereotyped speed profiles, the model did not capture trial-to-trial variability and therefore under-estimated between-subject variance, yielding markedly smaller best-case ICCs (Tables 3, 4). Together these results confirm that modern transformer-based forecasters—and not autoregressive statistics—are required to synthesise physiologically plausible reaches.

Limiting VGR to a compact recorded context and substituting model forecasts reduces the number of physically executed reaches while leaving the scoring pipeline unchanged. In practice, reducing the 8-target protocol from 64 recorded trials to eight recorded trials plus forecasts yields large session-time savings ($\geq$80%; Fig. 2). Crucially, the control–stroke separation in session-time distributions is preserved, indicating that forecast-based truncation does not bias group comparisons or reorder protocols. Together with the reliability results reported above, this supports the clinical viability of forecast-augmented assessment: it sharply cuts on-robot time without compromising interpretability or reliability.

Chronos' superior performance likely stems from three architectural choices. First, its tokenisation scheme discretises continuous signals into a "time series vocabulary" that retains fine temporal detail while enabling powerful masked-language pretraining. Second, the model's relative multi-scale attention spans hundreds of samples, letting

it capture both the bell-shaped speed pulse and the low-frequency curvature inherent in human reaches. Third, the foundation-model objective yields a rich uncertainty estimate that, when sampled via Monte-Carlo dropout, produces a diverse pool of candidate trials. Selecting subsets of that pool (green vs. orange markers in Figures 3) shows that diversity, not just mean accuracy, is critical for restoring ICC.

The concept of forecast-augmented assessment challenges the long-standing assumption that every trial must be physically executed. Once a model is fine-tuned to a dataset, subsequent patients could complete only the context block, after which the clinician would receive both standard KST metrics; in our evaluation we summarize uncertainty via error bars that reflect between-subject sampling and forecast-selection variability (See 2.7). This pipeline could increase throughput in busy clinics, reduce fatigue in highly impaired populations who struggle with lengthy assessments, and enables adaptive testing: if early forecasts indicate low reliability, the robot can request a few extra real trials and stop as soon as the ICC target is met.

Prior efforts focused on shortening Kinarm assessments by recording fewer trials or skipping predictable tasks, rather than forecasting missing trajectories. For VGR specifically, Mostafavi *et al.* showed that a reduced set of reaches can preserve SVM-based stroke/control discrimination from standard speed-derived parameters [26]. At the battery level, hierarchical task ordering and cross-task similarity were used to infer outcomes and skip redundant tasks, yielding substantial time savings [22, 24, 25]. In all cases, time is saved by selecting or reordering a fixed subset of physically recorded tasks/trials and discarding the rest. However, this approach simply considered whether a similar number of individuals were identified as impaired and not how trial reduction impacted assessment precision.

In contrast, we *forecast* the missing or noncompleted trials from as few as eight real reaches and then recompute the full suite of Kinarm parameters on the union of real + synthetic trials. This enables us to preserve the diagnostic richness of the entire task (all eight movement directions, continuous speed profiles, endpoint stability, etc.) while reducing on-robot time. Moreover, by generating multiple Monte Carlo–dropout realizations we explicitly model trial-level uncertainty, something that fixed-subset methods cannot capture. By sampling dropout-induced variations, we both reflect our confidence in each forecast and supply a richer ensemble of trajectories for more robust downstream feature estimation.

4.1 Limitations

Several caveats merit discussion. First, although we simulated prospective use by withholding 30 % of participants for testing, genuine deployment will inevitably encounter domain shifts—new robot firmware, alternate seating configurations, or even different Kinarm installations, which may degrade out-of-the-box performance; lightweight, continual finetuning or federated updating strategies will likely be needed to maintain precision. Second, while our foundation models excel at reproducing aggregate kinematic statistics, they remain black boxes with respect to the underlying biomechanical strategies; integrating them with biomechanically constrained decoders or musculoskeletal simulators could yield more interpretable, in silico trials.

4.2 Future directions

Several avenues for extending this work merit exploration. First, the same forecasting framework could be applied to other Kinarm tasks, reverse Visually Guided Reaching [37] or Ball on Bar [38], so that the entire twelve-task KST battery might be completed in less time, improving clinical throughput. Second, because our Monte Carlo–dropout approach produces not a single "best guess" trajectory but a distribution of plausible future reaches, it becomes possible to visualise each patient's individualized performance envelope. Clinicians could use these envelopes to set personalised difficulty thresholds for rehabilitation programs, selecting targets or perturbations that sit just beyond a patient's lower confidence bound, thereby delivering rehabilitation exercises optimally tuned to each individual's capacity and uncertainty profile. Finally, while our finetuning leveraged a few hundred participants, foundation models promise even broader generality when pretrained on massive movement corpora. By ingesting millions of hours of daily-life accelerometer or wearable-sensor recordings, one could endow these models with priors that capture the full spectrum of human motion. Such large-scale pretraining would not only improve forecasts for stroke rehabilitation but also likely transfer to other movement disorders such as Parkinson's disease, cerebral palsy, or multiple sclerosis.

5 Conclusion

This study targeted the assessment time burden of full-length Kinarm VGR blocks (40–64 trials), which can be fatiguing for some stroke survivors. We investigated whether time-series foundation models (Chronos, MOMENT) could forecast the missing trials from only 8–16 recorded reaches while preserving the reliability of VGR parameters. Using a forecast-augmented protocol, we showed that Chronos, fine-tuned on our mixed control–stroke cohort, recovers ICCs to within 0.05–0.07 of the full-length reference (parameter/cohort dependent) while reducing session time by 75–88% for the most impaired participants with stroke; MOMENT underperformed Chronos, and ARIMA provided only modest gains. To our knowledge, this is the first demonstration that foundation models can safely replace a large fraction of physically recorded robotic reaches in a reaching task.

Acknowledgements. We would like to sincerely thank Simone Appaqaq, Kimberly Moore, Ethan Heming, Mary-Jo Demers, Helen Bretzke, and Justin Peterson for their valuable assistance with participant recruitment and assessment, database management, and technical support.

Declarations

- **Funding** This research was funded by a Canadian Institutes of Health Research Operating Grant (MOP 106662), a Heart and Stroke Foundation of Canada Grant-in-Aid (G-13-0003029), an Alberta Innovates Health Solutions Team Grant (201500788), an Ontario Research Fund – Research Excellence grants (ORF-RE 04-47, RE-09-112), Vector AI Institute, a Canada CIFAR AI Chair, Natural Sciences and Engineering Research Council of Canada, and Queen's University.

- **Conflict of interest/Competing interests** SHS is co-founder and CSO of Kinarm (formally known as BKIN Technologies), the company that commercializes the robotic technology used in the present study. All other authors confirm no conflict of interest.
- **Ethics approval and consent to participate** This study was reviewed and approved by the Queen's University Research Ethics Board and the University of Calgary Conjoint Health Research Ethics Board. All participants gave their written informed consent to have their data collected for research purposes before performing the assessment.
- **Consent for publication** Consent for publication was obtained from all participants.
- **Data availability** The dataset generated for the present study is not accessible to the public. However, data can be obtained from SHS (steve.scott@queensu.ca) upon reasonable request.
- **Materials availability** Not applicable
- **Code availability** All Python code used in the analysis of this data is available on https://github.com/FaranakAk
- **Author contribution** FA conceived the concept and methodology, analyzed data, contributed to the interpretation of results, drafted the manuscript, and participated in the editing process.
  NM contributed to methodology, interpretation of results, and partcipated in editing the manuscript.
  SPD coordinated the collection of data, contributed to task design, participated in the editing process.
  SHS coordinated the collection of data, contributed to task design, provided the concept and methodology, contributed to the interpretation of results, participated in the editing process.
  PM conceived the concept and methodology, contributed to the interpretation of results, participated in the editing process.
  All authors reviewed and approved the final manuscript before its submission.

References

[1] Scott, S.H., Dukelow, S.P.: Potential of robots as next-generation technology for clinical assessment of neurological disorders and upper-limb therapy. Journal of rehabilitation research and development **48**(4), 335–353 (2011) https://doi.org/10.1682/jrrd.2010.04.0057

[2] Balasubramanian, S., Colombo, R., Sterpi, I., Sanguineti, V., Burdet, E.: Robotic assessment of upper limb motor function after stroke. American journal of physical medicine & rehabilitation **91**(11), 255–269 (2012)

[3] Gassert, R., Dietz, V.: Rehabilitation robots for the treatment of sensorimotor deficits: a neurophysiological perspective. Journal of NeuroEngineering and Rehabilitation **15**(1), 46 (2018) https://doi.org/10.1186/s12984-018-0383-x

[4] Park, K., Ritsma, B.R., Dukelow, S.P., Scott, S.H.: A robot-based interception task to quantify upper limb impairments in proprioceptive and visual feedback after stroke. Journal of NeuroEngineering and Rehabilitation **20**(1), 137 (2023)

[5] Fugl-Meyer, A., Jääskö, L., Leyman, I., Olsson, S., Steglind, S.: The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scandinavian journal of rehabilitation medicine **7**(1), 13–31 (1975)

[6] Gowland, C., Stratford, P., Ward, M., Moreland, J., Torresin, W., Van Hullenaar, S., Sanford, J., Barreca, S., Vanspall, B., Plews, N.: Measuring physical impairment and disability with the chedoke-mcmaster stroke assessment. Stroke **24**(1), 58–63 (1993) https://doi.org/10.1161/01.str.24.1.58

[7] Scott, S.H.: Apparatus for measuring and perturbing shoulder and elbow joint positions and torques during reaching. Journal of Neuroscience Methods **89**(2), 119–127 (1999) https://doi.org/10.1016/S0165-0270(99)00053-9

[8] Simmatis, L.E.R., Early, S., Moore, K.D., Appaqaq, S., Scott, S.H.: Statistical measures of motor, sensory and cognitive performance across repeated robot-based testing. Journal of NeuroEngineering and Rehabilitation **17**(1), 86 (2020) https://doi.org/10.1186/s12984-020-00713-2

[9] Akbarifar, F., Dukelow, S.P., Mousavi, P., Scott, S.H.: Computer-aided identification of stroke-associated motor impairments using a virtual reality augmented robotic system. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization **10**(3), 252–259 (2022)

[10] Simmatis, L., Krett, J., Scott, S.H., Jin, A.Y.: Robotic exoskeleton assessment of transient ischemic attack. PLoS ONE **12**(12), 0188786 (2017) https://doi.org/10.1371/journal.pone.0188786

[11] Kinarm: Official website. https://kinarm.com/. Accessed October 1, 2025 (2023)

[12] Coderre, A.M., Zeid, A.A., Dukelow, S.P., Demmer, M.J., Moore, K.D., Demers, M.J., Bretzke, H., Herter, T.M., Glasgow, J.I., Norman, K.E., Bagg, S.D., Scott, S.H.: Assessment of upper-limb sensorimotor function of subacute stroke patients using visually guided reaching. Neurorehabilitation and Neural Repair **24**(6), 528–541 (2010) https://doi.org/10.1177/1545968309356091 . PMID: 20233965

[13] Tyryshkin, K., Coderre, A.M., Glasgow, J.I., Herter, T.M., Bagg, S.D., Dukelow, S.P., Scott, S.H.: A robotic object hitting task to quantify sensorimotor impairments in participants with stroke. J Neuroeng Rehabil **11**, 47 (2014)

[14] Bourke, T.C., Lowrey, C.R., Dukelow, S.P., Bagg, S.D., Norman, K.E., Scott, S.H.: A robot-based behavioural task to quantify impairments in rapid motor decisions and actions after stroke. Journal of neuroengineering and rehabilitation **13**(1), 91 (2016)

[15] Mochizuki, G., Centen, A., Resnick, M., Lowrey, C., Dukelow, S.P., Scott, S.H.: Movement kinematics and proprioception in post-stroke spasticity: assessment using the kinarm robotic exoskeleton. Journal of neuroengineering and rehabilitation **16**, 1–13 (2019)

[16] Mang, C.S., Whitten, T.A., Cosh, M.S., Scott, S.H., Wiley, J.P., Debert, C.T., Dukelow, S.P., Benson, B.W.: Robotic assessment of motor, sensory, and cognitive function in acute sport-related concussion and recovery. Journal of neurotrauma **36**(2), 308–321 (2019)

[17] Simmatis, L.E., Jin, A.Y., Keiski, M., Lomax, L.B., Scott, S.H., Winston, G.P.: Assessing various sensorimotor and cognitive functions in people with epilepsy is feasible with robotics. Epilepsy & Behavior **103**, 106859 (2020)

[18] Vanderlinden, J.A., Semrau, J.S., Silver, S.A., Holden, R.M., Scott, S.H., Boyd, J.G.: Acute kidney injury is associated with subtle but quantifiable neurocognitive impairments. Nephrology Dialysis Transplantation **37**(2), 285–297 (2022)

[19] Bray, N.W., Raza, S.Z., Avila, J.R., Newell, C.J., Ploughman, M.: Robotic rigor: Validity of the kinarm end-point robot visually guided reaching test in multiple sclerosis. Archives of Rehabilitation Research and Clinical Translation **6**(4), 100382 (2024)

[20] Scott, S.H., Lowrey, C.R., Brown, I.E., Dukelow, S.P.: Assessment of neurological impairment and recovery using statistical models of neurologically healthy behavior. Neurorehabilitation and Neural Repair, 15459683221115413 (2022)

[21] Korenberg, M.J.: Identifying nonlinear difference equation and functional expansion representations: the fast orthogonal algorithm. Annals of biomedical engineering **16**(1), 123–142 (1988)

[22] Mostafavi, S.M., Dukelow, S., Scott, S., Mousavi, P.: Evaluation of similarities between robotic tasks for reduction of stroke assessment time. In: 2015 IEEE International Conference on Rehabilitation Robotics (ICORR), pp. 211–216 (2015). IEEE

[23] Bart, W.M., Krus, D.J.: An ordering-theoretic method to determine hierarchies among items. Educational and psychological measurement **33**(2), 291–300 (1973)

[24] Mostafavi, S.M., Dukelow, S.P., Scott, S.H., Mousavi, P.: Hierarchical task ordering for time reduction on kinarm assessment protocol. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2517–2520 (2014). IEEE

[25] Mostafavi, S.M., Scott, S., Dukelow, S., Mousavi, P.: Reduction of assessment time for stroke-related impairments using robotic evaluation. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(7), 945–955 (2017) https:

//doi.org/10.1109/TNSRE.2017.2669986

[26] Mostafavi, S.M., Dukelow, S.P., Glasgow, J.I., Scott, S.H., Mousavi, P.: Reduction of stroke assessment time for visually guided reaching task on kinarm exoskeleton robot. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5296–5299 (2014). https://doi.org/10.1109/EMBC.2014.6944821

[27] Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)

[28] Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D.C., Wang, H., Mahoney, M.W., Torkkola, K., Wilson, A.G., Bohlke-Schneider, M., Wang, Y.: Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815 (2024)

[29] Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., Dubrawski, A.: MOMENT: A family of open time-series foundation models. arXiv preprint arXiv:2402.03885 (2024)

[30] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, 5th edn. John Wiley & Sons, Hoboken, NJ (2015)

[31] Vanbellingen, T., Kersten, B., Winckel, A., Bellion, M., Baronti, F., Müri, R., Bohlhalter, S.: A new bedside test of gestures in stroke: the apraxia screen of tulia (ast). Journal of Neurology, Neurosurgery & Psychiatry **82**(4), 389–392 (2011)

[32] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016). PMLR

[33] Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86**(2), 420 (1979)

[34] McGraw, K.O., Wong, S.P.: Forming inferences about some intraclass correlation coefficients. Psychological methods **1**(1), 30 (1996)

[35] Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine **15**(2), 155–163 (2016)

[36] Veale, J.F.: Edinburgh handedness inventory–short form: a revised version based on confirmatory factor analysis. Laterality: Asymmetries of Body, Brain and Cognition **19**(2), 164–177 (2014)

[37] Lowrey, C.R., Dukelow, S.P., Bagg, S.D., Ritsma, B., Scott, S.H.: Impairments in

cognitive control using a reverse visually guided reaching task following stroke. Neurorehabilitation and Neural Repair **36**(7), 449–460 (2022)

[38] Lowrey, C.R., Jackson, C., Bagg, S., Dukelow, S., Scott, S.: A novel robotic task for assessing impairments in bimanual coordination post-stroke. Int J Phys Med Rehabil **3**, 002 (2014)