Accepted manuscript to appear in:

Behavior Research Methods

Long-form recording of infant body position in the home using wearable inertial sensors

John M. Franchak[1], Maximilian Tang[1], Hailey Rousey[1], & Chuan Luo[1]

Author Note

We are grateful to Vanessa Scott, Tasnia Haider, and Ishapreet Kaur for their help in collecting data for the present study and to the research assistants of the UCR Perception, Action, and Development Lab for annotating videos.

This work was funded by National Science Foundation Grant BCS #1941449 to the first author. The authors declare no conflict of interest.

Correspondence concerning this article should be addressed to John M. Franchak, UC Riverside Department of Psychology, 900 University Avenue, Riverside, CA 92521. E-mail: franchak@ucr.edu

Abstract

Long-form audio recordings have had a transformational effect on the study of infant language acquisition by using mobile, unobtrusive devices to gather full-day, real-time data that can be automatically scored. How can we produce similar data in service of measuring infants' everyday motor behaviors, such as body position? The aim of the current study was to validate long-form recordings of infant position (supine, prone, sitting, upright, held by caregiver) based on machine learning classification of data from inertial sensors worn on infants' ankles and thighs. Using over 100 hours of video recordings synchronized with inertial sensor data from infants in their homes, we demonstrate that body position classifications are sufficiently accurate to measure infant behavior. Moreover, classification remained accurate when predicting behavior later in the session when infants and caregivers were unsupervised and went about their normal activities, showing that the method can handle the challenge of measuring unconstrained, natural activity. Next, we show that the inertial sensing method has convergent validity by replicating age differences in body position found using other methods with full-day data captured from inertial sensors. We end the paper with a discussion of the novel opportunities that long-form motor recordings afford for understanding infant learning and development.

Keywords: body position, motor development, everyday experiences, sitting, machine learning

Word count: 11,995

Long-form recording of infant body position in the home using wearable inertial sensors

Infants' movements facilitate and constrain how they can interact with their surroundings. Changes in *body position*—whether infants are supine on their backs, prone on their bellies, sitting, upright, or held by a caregiver—have in-the-moment consequences for vision, object exploration, and social interaction. When sitting and upright, infants have a better view of faces and distant objects compared to their view while prone (Franchak et al., 2018; Kretch et al., 2014; Luo & Franchak, 2020). While walking upright, infants move farther away from caregivers and share toys in different ways compared to infants crawling in a prone position (Chen et al., 2022; Karasik et al., 2011). As infants grow older and acquire new abilities, such as independent sitting and walking, they spend more time sitting and upright and less time held, supine, and prone (Adolph & Tamis-LeMonda, 2014; Franchak, 2019; Franchak et al., 2018; Thurman & Corbetta, 2017). Thus, characterizing individual differences in the day-to-day accumulation of body position experiences informs developmental theory by revealing differential opportunities for learning (Franchak, 2020).

In this paper, we present an inertial sensing method to classify infants' full-day, real-time body position. Our method takes inspiration from a more mature technology: Long-form audio recordings of infants' language experiences. We begin by identifying the key features of wearable audio recorders that should be replicated in long-form recordings of motor behavior. Next, we review the current state-of-the-art in measuring infant motor behavior—video and survey data—and their limitations in capturing real-time, full-day behavior. Finally, we discuss the advantages of using inertial sensing to classify motor behavior. Despite promising past results in brief, supervised sessions (Airaksinen et al., 2020, 2022; Franchak et al., 2021; Greenspan et al., 2021), the current investigation takes a needed step forward by testing accuracy over long, unsupervised recordings in what we term a *distal comparison*.

Inspiration from Long-Form Audio Methods

The LENA® recorder is a commercial device worn in a custom shirt pocket; the recorder has sufficient battery life and storage to record for an entire day. Closed-source LENA® algorithms analyze the audio recordings to provide automatic counts of useful metrics, such as the number of words spoken by adults in the vicinity of the infant. Long-form audio recordings have had a transformational impact on language development research by allowing researchers to characterize opportunities for learning in daily life. For example, measuring the amount of speech heard by infants in the home (Weisleder & Fernald, 2013) or in a daycare setting (Perry et al., 2018) revealed how individual differences in speech input predict later vocabulary. Full-day language recording synchronized with other data sources allows researchers to identify how auditory input and vocal production interact with other processes. Beyond individual differences in aggregated data, long-form recordings reveal the temporal schedule of experiences. For example, infants' daily experiences hearing music are clustered in time, with "bursty" episodes of music presence separated by relatively long periods during which music is absent (Mendoza & Fausey, 2022).

We identified five key features of long-form audio methods that should be replicated in analogous studies of motor behavior. First, wearable audio recorders are *mobile*. Measurement is not limited to a particular room because the recording device travels with the participant. Data are recorded to onboard device memory, so participants do not need to be in range of a receiver. Second, wearable audio recording is more *unobtrusive*. Participants' reactivity to observation, such as from a video camera, may influence behavior more compared with a sensor that records only motion or audio data. For example, caregivers spoke more frequently to infants during a video-recorded portion of a home recording compared with audio-only segments captured by a LENA® device (Bergelson et al., 2019). Third and fourth, recordings capture *real-time data* over a *full* *day*. Real-time data are vital for identifying processes that unfold over minutes or seconds within an individual as opposed to comparisons of aggregated data between infants. Synchronizing real-time data to other data streams helps to reveal sources of variability within an individual (e.g., Malachowski et al., 2023; Wass et al., 2022). Full-day recordings are essential for capturing experiences across the heterogeneity of daily routines that moderate behavior (e.g., play, feeding, errands) (Kadooka et al., 2021, April; Tamis-LeMonda et al., 2018). "Burstiness" of behavior means that long recordings are needed to capture clusters of events amid long periods in which they may be absent (de Barbaro & Fausey, 2022; Warlaumont et al., 2021). Fifth, *automatic classification* means that the approach can scale to analyze large numbers of participants over long recordings without the bottleneck of manual annotation/transcription.

However, automatic classification can only replace human annotation if it is sufficiently accurate. An independent assessment of the LENA® algorithms found mixed results about classification accuracy. Correlations between human transcribed counts of adult words and child vocalizations against LENA®'s automatic counts were strong, $r =$ .698 and $r = .649$, respectively (Cristia et al., 2020). For other metrics, such as the number of "conversational turns" between the child and communicative partners, agreement was poor ($r = .364$). Thus, for some use cases (and for some metrics), long-form audio recordings provide a mobile, unobtrusive way to automatically score real-time data over a full day.

Limitations of Video and Survey Methods

Video and survey methods are the current state-of-the-art in assessing infants' gross motor behavior in naturalistic tasks. Although each method has complementary advantages and disadvantages for characterizing infants' everyday motor experiences, neither method fulfills all five key features of long-form audio recordings reviewed in the previous section.

Video observation is the most common way of measuring infant motor behavior in home recordings. Most often, an experimenter with a handheld camera follows infants from room to room to ensure that their movements are visible throughout the recording session (Chen et al., 2022; Herzberg et al., 2021; Karasik et al., 2011). The primary advantage of video recording is that it captures real-time behavior. Standard 30 Hz video recording is adequate to capture changes in infant body position that occur on the timescale of seconds. However, requiring an experimenter to operate a camera is obtrusive, whereas relying on a stationary camera means that infants will be absent from view as they move from place to place. Moreover, video observation cannot easily scale to long durations or large numbers of participants. An experimenter cannot follow behind infants to record their behavior from morning to night; typical video recording sessions last 45-120 minutes (Chen et al., 2022; Herzberg et al., 2021; Karasik et al., 2011), far short of capturing the variety of activities across the full daily routine. Even if full-day videos were available, the lack of suitable automatic classification tools means that the human cost of annotation would be immense. Our annotation of body position takes approximately 2-5 hours to complete for every hour of video (depending on how often infants switch positions), meaning that a full "waking day" of approximately 11 hours for a 12-month-old (Galland et al., 2012) could take 22-55 hours of labor to annotate.

In contrast, survey methods such as daily diaries/inventories or ecological momentary assessment (EMA) are mobile, unobtrusive, can be applied across an entire day, and do not need laborious annotation. Diary studies provide caregivers with logs or structured interviews to record activities (Karasik et al., 2022; Majnemer & Barr, 2005). Ecological momentary assessment uses smartphone notifications to prompt caregivers to make repeated estimates of behaviors throughout the day (Franchak, 2019; Kadooka et al., 2021, April). Although such survey responses are valuable in aggregate, they lack the real-time temporal resolution to describe moment-to-moment changes in behavior. At best, EMA surveys prompt caregivers to make hourly observations; increasing the number of surveys per day would be too burdensome for the respondent. Thus, despite being a useful tool for estimating broad developmental changes and individual differences in infants' motor experiences, survey methods are not suited for capturing within-participant temporal dynamics.

Promise of Inertial Sensing Methods

Measuring infant movement with inertial movement units (IMUs) is a promising avenue for long-form recordings of motor behavior in the home (Bruijns et al., 2020; Cliff et al., 2009; de Barbaro, 2019; Lobo et al., 2019). Lightweight sensors (10-30 g) can be embedded in garments to make recordings fully *mobile*, and they are *unobtrusive* because they do not require a researcher to follow with a camcorder. Many commercially-available IMUs have > 12 hour battery life with onboard storage to record *real-time*, *full-day* motion data at a high sampling rate (e.g., 50-100 Hz).

The open question is whether *automatic classification* is sufficiently accurate to measure movement categories that are relevant to developmental and clinical research, and whether measurements continue to be accurate over long recording periods. Data processing algorithms are needed to classify the raw sensor data (i.e., linear and angular acceleration time series) into meaningful categories. Categorizing body position—supine on the back, prone on the belly, sitting, upright, or held off the ground by a caregiver—is complex because movement can vary greatly *within* a body position. An upright infant can be standing still or can be walking briskly across the room. A prone infant can be stationary in "tummy time", or they can crawl in a myriad of ways (Adolph et al., 1998). Moreover, the configuration of the arms, legs, and torso within a body position can vary greatly in everyday contexts. Infants can sit on the floor in a tripod position with support from an arm, in a "V" position with legs fully extended, or in a "W" position with knees bent. Sitting on a caregiver's lap without the need to maintain balance means that the legs can dangle and the torso can lean in different directions. Sitting in a high chair or car seat likely restricts the range of torso orientations compared with sitting independently or on a caregiver's lap. Finally, caregivers frequently pick up and transport infants, creating motion signals that need to be differentiated from independent activity (Kwon et al., 2019; Patel et al., 2019).

Modern approaches to human activity recognition have used machine learning to classify activity categories based on features derived from IMU data in adults (Arif & Kattan, 2015; Preece et al., 2009), children (Nam & Park, 2013; Ren et al., 2016; Stewart et al., 2018), and infants (Airaksinen et al., 2020; Franchak et al., 2021; Yao et al., 2019). Three prior investigations have used different machine learning techniques to categorize infant body position from multiple IMUs towards the goal of collecting full-day data. Airaksinen et al. (2020) tested 4- to 8-month-olds in a laboratory visit with a 4-sensor array (one on each thigh and one below each shoulder), and found 95% accuracy in distinguishing between body position categories that crawling infants could perform on the floor (excluding times that infants were held by caregivers). Using a wider age range of 6-18 months, Franchak et al. (2021) found 98% accuracy (*kappa* = 95%) in a laboratory validation study with a 3-sensor array (ankle, knee, thigh on a single leg) in categorizing body position that included infants who could both crawl and walk and also included a category for caregiver holding. Most recently, Airaksinen et al. (2022) conducted a validation study of body position classification in either a home or clinic testing 4- to 19-month-olds with a 4-sensor system, refining their previous method to detect moments that infants were carried by caregivers. Classification accuracy did not vary between lab and home settings, and was generally high (95%, *kappa* = .93). In contrast to the three machine learning studies that used 3-4 sensors (Airaksinen et al., 2020, 2022; Franchak et al., 2021), Greenspan et al. (2021) used orientation from a single hip-worn sensor to measure body position with a high degree of accuracy (*kappa* = .84). Although all four studies yielded promising classification accuracy, accuracy was assessed in brief (15-60 minute) sessions supervised by a researcher, leaving the open question of how well body position classification will scale to testing across an entire day of natural home life.

Goals of the Current Study

Accordingly, the overarching goal of the current study is to test the validity of long-form body position recording in the home during unsupervised, everyday behavior. Supervised recordings from past work (Airaksinen et al., 2020, 2022; Franchak et al., 2021), whether in the home or in the lab, let researchers set up the situation to encourage or restrict certain behaviors. Prior work focused on “free play”, in which caregivers were asked to play with the infant without restraining the infant or otherwise shaping how they could move. However, in a real day, non-play activities (e.g., eating lunch in a high chair) create challenging situations for applying automated classification of body position. For example different types of sitting—independently on the floor, supported on a caregiver’s lap, or restrained in a high chair—all need to be scored as sitting. Moreover, classifiers must be able to detect new variants of behaviors that might arise over the course of a real day; it is impossible for researchers to gather training data for every possible variation that might occur. Thus, our central question is whether models trained on video-recorded observations at the beginning of the day generalize to predict behavior at a later time. Assessing the validity of temporally *distal* periods—when infants and caregivers are unsupervised and free to follow their everyday routines—is a crucial step to establish whether automatic classification can be used to measure body position across a day.

In the current study, we report the feasibility and validity of body position classification over the full day in the home based on 34 testing sessions from 22 infants aged 4-14 months. Participants received a custom pair of infant leggings embedded with 4 IMUs (one on each ankle and one on each thigh) and a video camera to collect ground truth data about infant body position. A *proximal comparison* period began when participants received the equipment and completed a guided phone call during which caregivers were asked to elicit different body positions based on prompts from the experimenter. This "semi-supervised" period was most similar to previous recordings because it occurred during a convenient time for the infant and caregiver to play while they received instructions from the experimenter. The **first goal** of the current study was to determine the accuracy of body position classification during the proximal comparison period using this novel, semi-supervised procedure in participants' homes. Past work found better performance using "individual models"—models that were trained on one participant's data to predict their later behavior—compared with a "group model" that aggregated data from all infants to create a single body position classifier (Franchak et al., 2021), so we compared both modeling approaches in the current investigation.

The crucial test was how well models predicted later behavior over longer recordings of everyday activities. A second, *distal comparison* period followed the proximal comparison period and captured approximately 90 minutes of home behavior that was completely unsupervised. Caregivers and infants could (and did) do whatever they wished, and no researcher was present. Because this recording happened a considerable amount of time after the initial setup and instructions from the experimenter, accuracy could decline if caregivers or infants moved the garment or sensors. Moreover, increasing variation in everyday activities during the distal comparison creates a greater challenge, testing whether body position classification models can generalize to novel test cases. Thus, the **second goal** of our study was to assess accuracy during the distal comparison.

After the distal comparison period when video recording ceased, we asked caregivers to have infants wear the IMUs for the rest of the day until their regular bedtime, creating the **first real-time, full-day dataset of infant body position**. Interpreting such data required caregivers to log when infants napped, when they removed the sensor garment for diaper changes or other reasons, and when infants went to bed at the end of the day. Thus, the **third goal** of the study was to examine the quality of the full-day data. Could infants wear the sensor garment throughout the desired period? If full-day classifications of infant behavior are accurate, they should demonstrate convergent validity with other methods. Thus, we determined whether full-day body position measurements conformed to expected age differences in body position. Based on past results (Franchak, 2019), infants should spend increasingly more time sitting and upright but less time supine over the age range tested (4 to 14 months).

Methods

Participants and Design

Infants were recruited in one of two age groups: *Younger* infants were between 4 and 7 months and *older* infants were between 11 and 14 months. There were 8 infants in the 4-7 month group and 14 in the 11-14 month group. Ten infants were female and 12 were male. Families were recruited through social media advertisements and from community events in Southern California. Most infants were reported to be either Hispanic and White (n = 8) or Non-Hispanic and White (n = 7). Families were compensated $30 for every home recording session they completed. The University of California, Riverside Institutional Review Board reviewed and approved all procedures associated with the study. All caregivers gave their informed consent before the start of the study.

Most participants were tested in a single session (n = 15), but 7 participants contributed between 2-4 sessions as part of an ongoing longitudinal study (3 from the younger group and 4 from the older group). Although including more data from some participants could over-represent their characteristics in the model, we reasoned that this drawback was outweighed by having more available data to use for training. Only 1 session was excluded due to a technical error—one of the four IMU sensors failed to record, resulting in an unusable set of data for classification. Across the two age groups, we report data on a total of 34 sessions, with 14 sessions from younger infants and 20 sessions from older infants. Across sessions, younger infants' age ranged from 3.8 to 7.2 months (M = 5.2) and older infants' age ranged from 10.7 to 14.2 months (M = 11.7). The total number of recording sessions (34) exceeded the number of sessions employed in comparable past work: 10 in Nam and Park (2013), 15 in Franchak et al. (2021), 22 in Airaksinen et al. (2020), 23 in Greenspan et al. (2021), and 33 in Yao et al. (2019).

Apparatus

Four inertial movement units (IMUs) were used to record infant movement across the day (MC10 Biostamp). A custom garment was designed to hold the IMUs (Figure 1). Internal pockets were sewn into snug-fitting infant leggings so that IMUs would stay close to the body (reducing vibration). A pocket over the thigh and a pocket just above the ankle were sewn on the lateral surface of the right and left legs of the garment. Multiple sizes of the garment were created (modeled on US 3-6 mo, 6-9 mo, 9-12 mo, and 12-18 mo sizing). Caregivers indicated in advance which size would fit their infant, and to "size down" if between sizes to ensure a snug fit and minimize sensor movement. Each garment had a distinct pattern on the seat of the pants so that caregivers could identify front versus back and place the garment in the correct orientation.

Each sensor had sufficient battery and onboard storage to record accelerometer and gyroscope data for approximately 12 hours. We chose a sampling rate of 62.5 Hz (one of the available presets) based on prior work that used rates of 50-64 Hz (Airaksinen et al., 2020; Franchak et al., 2021; Yao et al., 2019). Infants also wore a LENA® recorder throughout the day in the front pocket of a LENA® shirt, located near the infant's chest, to determine whether data could be simultaneously recorded from the LENA® and IMU sensors (LENA® data were not analyzed in the current study).

Videos were captured at 30 Hz using an action camera on a miniature tripod (Insta360 ONE R) that caregivers placed in the same room as the infant. Although recordings lasted 3 hours, they were divided into two video files of approximately 90 minutes temporally separated by a gap of 40-45 s. Caregivers received a log sheet to record times that infants napped and times that the sensor garment was removed from the infant (e.g., baths, diaper changes, errands).

Procedure

Figure 2 shows an exemplar timeline of the entire procedure and recording periods for a single participant. A researcher arrived at the participant's home in the morning and set each device to record while at the doorstep. To create a recognizable synchronization point between the video recording and IMU data, the researcher dropped the sensor garment containing the IMUs on a surface in view of the camera, as in Franchak et al. (2021). All the equipment—once recording and with synchronization information recorded—was placed inside a large bucket and left outside the family's front door. The researcher then called the caregiver on the phone and walked them through a set of procedures needed to properly set up the equipment and record video for ground truth human annotation of body position. At the start of this "guided call", the caregiver was instructed to place the camera in an area that captured the majority of the room. Next, they were asked to put the pair of leggings and shirt on their infant, with the researcher providing guidance about how to correctly orient the garments.

Afterwards, the researcher asked the caregiver to complete a number of guided activities with their infant. Within view of the camera, the caregiver was asked to place their infant in several different positions: lying supine, lying prone, sitting on the floor, standing upright, held by the caregiver while the caregiver walked back and forth, crawling, walking, and sitting in a restrained seat (e.g., high chair). Depending on the infants' age and motor skill level, the positions could be done independently or were completed with assistance from the caregiver. The researcher kept time to ensure at least 1 minute of behavior for each activity. Once completed, the caregiver was then instructed to play with their infant for 10 minutes within view of the camera to collect additional ground truth data.

Afterwards, they were asked to go about their day as usual with the infant wearing the sensor garment until their bedtime, only taking off the sensor for naps, baths, diaper changes, and trips out of the house. The caregiver logged the times the sensors were removed (blank areas in the timeline in Figure 2) or the child took a nap (gray areas in the timeline in Figure 2) so that those times could be excluded from analysis. The following day a researcher picked up the bucket of equipment.

Because the camera only had the battery life to record for ~3 hours (automatically split into two 90-minute video files), this divided the day into different periods for analysis. As seen in the bottom of Figure 2, the *video period* comprised the first three hours of recording starting from the researcher's arrival when they turned on the camera. The first 90-minute video file, termed the *proximal comparison*, contained the activities during the guided call followed by a period during which infants and caregivers resumed their normal activities. Because this video contained the synchronization point, the data in this period had high temporal synchrony between IMU and video data. Synchronization errors were estimated to be less than 30-60 ms (1-2 video frames). The second video file comprised the *distal comparison*. This video recorded the next 90 minutes of natural activity. However, because of a limitation in the camera, there was a variable gap of ~40 s between the two videos, so synchronization in the distal comparison video was coarser, with estimated temporal offsets of ~5 s in either direction.

Body Position Annotation

The proximal and distal comparison videos for each participant were annotated by trained human coders to classify infant *body position* into one of 5 mutually-exclusive categories following the definitions in prior work (Franchak et al., 2021): supine, prone, sitting, upright, or held by caregiver. All coding was done using Datavyu software (datavyu.org). A Databrary repository contains the entire video recording, coding files, and raw IMU data for a single participant (https://nyu.databrary.org/volume/1580).

Supine was coded when the infant was lying on their back, on their side, or was reclined up to a 45 degree angle. Prone was coded when the infant was lying on their stomach, was stationary supported by the hands/knees or the hands/feet, or was crawling. We scored sitting to include any form of the following seated positions: 1) infants sat with their buttocks on a surface, such as on the floor or a caregiver's lap, 2) infant was in a kneeling-sit position, in which their knees were on the ground with their legs tucked underneath the buttocks, and 3) infant was in a seating device, such as a high chair, that kept the torso oriented perpendicular to the ground (a reclined position, such as in a young infant's car seat, would be counted as supine). Upright was coded when the infant was standing or squatting on the ground with two feet or walking (regardless of whether infants' balance was assisted by a caregiver or with their hands holding onto something for support). Our goal in creating a category for "held by caregiver" was to separate times when infants were in control of their body position from times when they were suspended in the air (rather than resting on furniture or a surface). Held was coded when infants were carried off of the ground. However, when the caregiver was sitting with the infant in their lap the infant's body position was coded as if the caregiver was a surface (e.g., if the infant was sitting on the caregiver's lap this was coded as sitting). Times during the video when the infant was out of view were excluded. Periods when the sensor garment was adjusted or taken off the infant were also excluded, as were transitions between body positions.

A primary coder completed annotation for the full length of the video, while an independent reliability coder completed annotation for the first thirty minutes of each video. Interrater reliability was based on the proportion of video frames that the two coders chose the same body position code. Overall agreement averaged 90.9% across video files, ranging from 68.4%-100% for individual video files. Cohen's kappa averaged 86.1% across video files, ranging from 31.0%-100% for individual video files.

Body Position Classification

The same machine learning classification process was used as in prior work (Franchak et al., 2021). Using the synchronization point, human-coded body annotations from video were linked to the corresponding times in the IMU time series data. A single, merged dataset was created with synchronized accelerometer signals (in three orientations: X, Y, and Z) and gyroscope signals (in three orientations: roll, pitch, and yaw) for each of the four sensors (left thigh, right thigh, left ankle, right ankle) with the corresponding timestamp and body position code using the *timetk* package (Dancho & Vaughan, 2023) and the *lubridate* package (Grolemund & Wickham, 2011) in R version 4.1.2 (R Core Team, 2021).

Classification training and prediction was conducted on a windowed dataset that summarized the raw, 62.5 Hz motion signals within 4-s windows. Data were reduced in time by creating overlapping moving windows (4-s long, comprising 250 samples) starting each second, which is a common unit of analysis in prior studies of human activity classification (Airaksinen et al., 2020; Franchak et al., 2021; Nam & Park, 2013). For each 4-s window, we aggregated the 250 samples to create single scores for a variety of motion features—summary statistics that could be fed into the machine learning model. The minimum, maximum, 25th percentile, 75th percentile, mean, median, skew, kurtosis, standard deviation, and sum were computed for each signal (e.g., right thigh linear acceleration along the X-axis, left ankle pitch angular acceleration). The 10 summary statistics and 24 sensor signals generated 240 columns of motion features that described movement within each window. Furthermore, a series of cross-sensor and cross-orientation summaries (such as the correlation, magnitude, and difference between pairs of sensors) added an additional 196 columns of motion features. The 436 total motion features corresponded to a single body annotation code for each 4-s window. Windows were only used for training/testing if they contained a single body position for $> 75\%$ (3 s) of time within the window to ensure that motion signals could be linked to an unambiguous example of each behavior.

The resulting windowed dataset was used for machine learning classification and validation. For each analysis reported in the results, a subset of data were defined as a “training” set and another, independent portion of the data were defined as a “testing” set (further described below). Random forest models (Breiman, 2001) used the training set to learn the body position label for each window from the set of 436 motion features using the *randomForest* package (Liaw, Wiener, et al., 2002). The resulting random forest model could later be applied to a set of testing data with the *predict* function.

Data Sharing and Transparency

Three online repositories contain openly shared data, materials, and analysis code. A Databrary repository (https://nyu.databrary.org/volume/1580) includes an exemplar participant’s recording session, with the raw video data files, the Datavyu annotations of those video files, a log file with machine-readable synchronization points and nap/diaper change times, and accelerometer and gyroscope data for each of the 4 sensors. A GitHub repository (https://github.com/JohnFranchak/body_position_classification_example) contains the exemplar participant’s data and source code to: 1) synchronize IMU and video annotations, 2) calculate windowed motion features for their data, and 3) train and test the body position classifier using an “individual model”. Because of the overall size of the full dataset and the computational power/time required to synchronize and create windowed datasets for each session, it would not be feasible to reproduce the calculations for all 34 sessions. However, in a second Github repository (https://github.com/JohnFranchak/body_position_classification_ms) we share the full results of those computations: The dataset of windowed motion features with corresponding body position codes used to validate the method. This reproducible manuscript created in RMarkdown and *papaja* (Aust & Barth, 2022) can be regenerated from those data files for full computational transparency.

Results

We report three sets of results based on 34 full-day testing sessions resulting in a total of 302 hours of movement recording.

Assess the Proximal Accuracy of Body Position Classification Models

The first set of analyses use data from the proximal comparison to determine the "best case" accuracy of the models, training and testing on similar types of data. The high degree of temporal synchronization between video and motion data during this period makes it possible to link human-coded body position annotations to each 4-s window of motion data, providing ground truth data for model training and testing. As in past work (Franchak et al., 2021), we compared two types of models: *group models* and *individual models*. To assess the accuracy of body position classification for each recording session, we reserved the (temporally) last 25% of a session's proximal comparison data as the testing set. The testing set was never used as training data, and was the same for both modeling approaches to facilitate direct comparisons between the models. We generated two different training datasets relative to each testing set. For the group model training set, we aggregated the first 75% of all **other** sessions' proximal comparison data. This leave-one-out cross-validation tested the generalization of the model to a recording session that was not used at all in the training set. The individual model training set used the first 75% of data from the testing set's session. The individual model tests whether earlier training data generalize to later testing data within an individual participant's recording. Figure 3, Table 1, and Table 2 summarize the performance of group and individual models using standard metrics for classification.

Overall Accuracy. Overall accuracy (Figure 3A) represents the proportion of 4-s windows in the testing set in which the model prediction matched the human annotation of body position. Overall accuracy for group models ($M = 0.85$) was slightly lower than accuracy for individual models (M = 0.92). Although overall accuracy from our semi-supervised, in-home data collection did not match the near-perfect accuracy (.95-.98) found in prior in-lab studies (Airaksinen et al., 2020; Franchak et al., 2021), both models approached the level of agreement found between two human coders (M = .906). Most likely, lower accuracy in the current study results from the more variable and complex behavior observed in a semi-supervised setting rather than from a difference in the quality of the classification model; although the first part of the proximal period was guided by the experimenter during a brief phone call, the remainder of the proximal period included natural behavior. Visual inspection of Figure 3A shows that accuracy values were heavily skewed, with many approaching perfect accuracy but a few sessions with very poor accuracy. Looking at the median performance suggests that the difference between models was not considerable for the typical participant (group median accuracy = 0.89; individual median accuracy = 0.93).

As in Airaksinen et al. (2020), overall accuracy decreased when fewer sensors were used. Table 3 compared overall accuracy for all four sensors (top row), pairs of 2 sensors (rows 2-5), and single sensors (rows 6-9). The highest accuracy was observed when using all four sensors (group model M = 0.846; individual model M = 0.916), and the lowest when using only a single sensor (left ankle group model M = 0.667; right ankle individual model M = 0.819). Accuracy for some pairs approached 4-sensor accuracy: Left ankle and thigh had group model accuracy of M = 0.824 and individual model accuracy of M = 0.901, only 2.2% and 1.5% worse than using all four sensors. Other pairings were less accurate; notably, using both ankles resulted in group model accuracy of M = 0.719 and individual model accuracy of M = 0.844, 12.7% and 7.2% worse than using all four sensors.

Cohen's Kappa. Strong overall accuracy can be misleading when the relative frequency of different classes is unbalanced. Accordingly, we report Cohen's kappa, a commonly-used metric that penalizes missing rare events (Figure 3B), and we provide classification metrics for each individual body position (Table 2) to account for imbalance in body position rates within and between individuals. Similar to overall accuracy, kappa values were strong for both model types with group kappas ($M = 0.75$) somewhat worse compared with individual kappas ($M = 0.82$). Guidelines for interpreting kappa statistics (Landis & Koch, 1977) consider 0.81–1.00 “Almost Perfect,” 0.61–0.80 “Substantial,” 0.41–0.60 “Moderate,” 0.21– 0.40 “Fair,” and 0–0.20 “Slight to Poor”, indicating that agreement for most group and individual model predictions fell in the Substantial to Almost Perfect range.

As in past work (Airaksinen et al., 2020; Franchak et al., 2021), all body positions were accurately classified even though performance varied somewhat between positions. As Table 2 shows, mean kappa statistics were strongest for prone (group $M = 0.860$, individual $M = 0.841$) and supine (group $M = 0.764$, individual $M = 0.912$). Sitting performance fell in the middle, and was considerably worse for group models than individual models (group $M = 0.702$, individual $M = 0.887$). Held (group $M = 0.726$, individual $M = 0.727$) and upright (group $M = 0.673$, individual $M = 0.741$) performance was the least accurate, however, average performance was still within the “Substantial” range.

Sensitivity and Positive Predictive Value. Sensitivity refers to the proportion of events of a given position that were correctly identified (e.g., out of 100 human-coded sitting windows, how many of those windows did the model correctly classify as sitting?). High sensitivity means that events are unlikely to be missed. In contrast, positive predictive value (PPV) refers to the proportion of events classified for a given position that actually belonged to that position (e.g., if the model said a baby was upright during 100 windows, how many of those windows were indeed human-coded upright events?). High PPV means that we can be confident in the event label. Table 2 shows the sensitivity and PPV by body position class for group and individual models. For group models, sensitivity and PPV were similar: They were highest for supine, prone, and sitting (the most accurately identified class) and lowest for upright and held. Results for individual models were similar to group models, with the exception of a somewhat lower sensitivity score for held. Overall, the results suggest a reasonable balance between sensitivity and PPV among different body position classes for both model types.

Measure the Distal Accuracy of Body Position Classification Models

The first set of results showed that group and individual models trained from data during the proximal comparison period were accurate during the proximal period. This suggests that the immediate accuracy of body position predictions early in the recording session was strong, but does not address how accurately predictions will be later in the recording. In the next analysis, we examine long-term performance by testing how accurately models trained from the proximal period could predict body position during the distal comparison period. A single group model was created using all sessions' proximal period training data (rather than group models leaving out a single session); the same individual models were used. Distal videos had only coarse temporal synchrony with motion recordings which precluded calculating accuracy based on the proportion of matching events. Instead, we summed the amount of time infants were predicted to be in each of the 5 body position categories from the model and compared that to the summed time for the body positions based on human coding (Franchak et al., 2021; Yao et al., 2019).

Whereas the proximal analyses used all 34 sessions, this was not possible in the distal comparison. Because the start of the visit was scheduled during a time when the infant was awake, it was common for a nap to follow the proximal period. Nine sessions were excluded because the infant was either napping or otherwise not on camera during the entire 90-minute distal recording. Three additional sessions were excluded because a caregiver accidentally turned off the video camera ($n = 1$) or (purposefully) left the house ($n = 2$). This left 22 sessions with usable distal comparison data.

Overall Agreement During the Distal Comparison. Figure 4 and Table 4 summarize the overall agreement during the distal comparison period. For each session, we calculated the actual time spent in each body position (out of times the infant was visible on camera and awake) using human annotated body position (x-axis on Figure 4). Predicted time was calculated the same way for group and individual model predictions, omitting the off-camera and nap periods to make a direct comparison. Agreement was strong across participants and across body position classes: The correlation between group model predictions and human-coded time was $r = 0.80$, and the correlation between individual model predictions and human-coded time was $r = 0.91$. As in the proximal comparison, agreement varied somewhat between body positions; in particular, agreement for held was poor. Unlike in the proximal period, some body positions were better predicted by group models and others by individual models.

Visual inspection of Figure 4 indicated two extreme outliers, which we marked by a gray square and a gray diamond. The "gray square" outlier had significant confusion between sitting and supine classification. Reviewing the video indicated that this participant spent a long period of time in a seating device that was reclined almost exactly at 45 degrees, making it difficult to determine if the infant was sitting or supine. The infant also spent a long time in the mother's arms in an ambiguous supine/sitting position. This participant's proximal accuracy was also poor because similar ambiguities appeared during the initial recording. In contrast, the "gray diamond" outlier had strong proximal accuracy, with confusion only arising in the distal period between upright and held categories. Reviewing the video showed that all disagreements occurred when the infant was in a baby walker; human coders scored this as "upright" but the models predicted it as "held". Most likely, the infant's movements in the baby walker were more similar to how a baby moved while carried, and unlike how most infants moved while walking upright.

What is notable about both outliers is that disagreements were restricted to a particular border case (supine vs. sitting; upright vs. held); accuracy for other classes remained strong. This suggests that their poor performance came as a result of spending a long time in an ambiguous position, not the result of the entire model failing to generalize to the later time period (or an error in sensor placement, such as if the parent removed the leggings and put them on backwards after a diaper change). To better capture the typical level of agreement, we report all correlations in Table 4 excluding the two outliers. Overall agreement among the non-outlier sessions was excellent for both group models ($r = 0.95$) and individual models ($r = 0.96$).

Short-Timescale Agreement during the Distal Comparison. Although overall aggregate agreement in the distal comparison was strong, it is important to show that similarly strong agreement is found within a shorter timescale. We repeated the agreement analysis after dividing the distal comparison period into nine 10-minute bins (marked by vertical dashed lines in Figure 2). Infants had varying numbers of 10-minute bins depending on how much time they were awake and on camera. Bins were included only if there was $>$ 7 minutes of usable data. Table 5 shows the agreement correlation coefficients for group and individual models, including and excluding the two outliers identified in the previous section. Performance at a short timescale was similar to performance overall: Overall agreement after excluding outliers was excellent for both group ($r = 0.92$) and individual models ($r = 0.94$). Within-position correlations were weakest for held and strongest for upright regardless of the model type. Agreement for prone was better for group models, whereas sitting and supine were better predicted by individual models.

To describe the observed amount of prediction error in 10-minute bins, we subtracted the predicted duration (in minutes) for each body position in each bin from the human coded duration in that bin to create a prediction difference score. A score of 0 would indicate no error; positive differences indicate that the model overestimated the amount of time in a position, whereas negative differences indicate underestimation. Figure 5 plots the mean prediction difference for each session for each body position. The gray shaded area marks ± 1 minute of prediction error. Most session-averaged predictions fall within 1 minute of error without a clear bias towards overestimation or underestimation. For group models, we calculated the percentage of 10-min bins across participants that had errors < 1 minute: 94.62% for held, 80.65% for supine, 94.62% for prone, 77.42% for sitting, and 94.62% for upright. For individual models, the percent of 10-min bins with < 1 minute of error was: 92.86% for held, 88.10% for supine, 88.10% for prone, 83.33% for sitting, and 96.43% for upright.

Examine the Data Quality of Full-Day Home Recordings

After the distal comparison video ended, caregivers were instructed to keep the sensors on their infants for the remainder of the day until infants went to bed, removing the sensors for naps, diaper changes, and trips out of the house. The algorithm was not designed to classify behavior during transportation (e.g., strollers, automobiles) so data collection was restricted to in-home behavior. The first two sets of results show that accuracy was consistently high across the proximal and distal recordings, providing confidence that predictions over the remainder of the day would continue to be accurate after the video recordings stopped. This leads to two final questions—how successfully did recordings capture infants' entire day, and how well do findings from full-day classification converge with findings that use other methodologies?

How Well Did Recordings Capture Infants' Entire Day? Figure 6 depicts body position timelines across the day for each session, divided into younger (A) and older infants (B). Predictions from the group model were used to ensure that motion data were classified consistently across all sessions. Moreover, for infants who did not display all 5 body positions in the training period, group models were necessary to predict those behaviors across the entire day. Session start times ranged from 09:15 to 13:20 with a median of 10:25. Sometimes infants were unexpectedly asleep at the scheduled time, leading to a few sessions in which infants began wearing the sensors later than intended (such as #11 in the younger group). With two exceptions, recordings lasted until the infants' bedtime. Older participant #1 had the equipment picked up on the same day rather than the next day, so the recording ended at 17:00. Older participant #11 wore the equipment in the morning, but the family left the house at 10:15 and remained out for the rest of the day, choosing not to put the equipment back on when they returned home in the late afternoon. Among those participants who wore the equipment until bedtime, recordings ended between 17:25 and 22:15 with a median of 19:20.

Among participants who wore the sensors from morning to bedtime, all infants wore the sensor garment during 100% of the intended recording period, excluding the times caregivers were asked to remove the garment (naps, diaper changes, trips out of the house), based on caregiver logs of wear time. No caregiver reported removing the garment for any other reason (such as infant discomfort), meaning that the majority of the time during the day either resulted in usable body position data or was excluded due to caregiver-reported naps (gray periods in Figure 6). The total length of the recording period ranged from 6.92-11.75 hours with a median of 9.04 hours. Nap times reported during the recording period ranged from 0.00-4.99 hours with a median of 2.25 hours. Body position data were available to describe 37.3-100.0% of the awake portion of the recording period (median = 100.0%). Younger infant #14 had the least portion of the waking day accounted for; the family left the house to run errands during the majority of the recording period, and the infant napped during much of the remaining time at home. Overall, the recordings produced a median of 6.14 hours of motion data, with the entire dataset totaling 206.66 hours. Even the shortest full recording (3.06 hours) exceeded the longest observational sessions in past work in which an experimenter operated a camera.

Considering that body position annotation takes 2-5 hours of labor per 1 hour of behavior, the 207 hours of data we recorded would have taken 413-1033 hours of labor to annotate from video. However, only an estimated 64-160 hours of labor was needed to annotate the initial hour of each session's video that was used to train the machine learning models—an immense savings in the human labor cost of annotation that allows the method to scale to larger sample sizes and recording durations.

Can Full-Day Estimates of Body Position Reveal Age Differences? In order to show the convergent validity of the classification method, we demonstrate how the full-day recordings can reveal individual differences in body position according to age, replicating age differences revealed in past work (Franchak, 2019). Table 6 summarizes the percent of time that younger and older infants spent in each body position out of their awake samples, as predicted by group and individual models. EMA surveys (Franchak, 2019) found that from 3-12 months, infants spent less time held, reclined, and supine but spent increasingly more time sitting and upright. The most straightforward comparisons between the two investigations are upright and prone, because they were defined identically. In Franchak (2019), upright time was 0.6-5.5% of the time for infants 3-6 months, increasing to 22.0% at 12 months. Results from the current study were similar: younger infants' upright time estimated from full-day group models was 5.41% compared with 18.24% for the older group. In Franchak (2019), prone time was 2.9-9.2% of the time for infants 3-6 months, and was 7.2% at 12 months. In the current study, prone time was 14.03% for younger infants and 15.09% for older infants. Furthermore, our results are similar in showing that held and supine time was less for older infants than younger infants (Table 6). Sitting was greater for older infants (44.39%) compared with younger infants (28.91%). Moreover, sitting was the most frequent body position followed by upright in 12-month-olds (Franchak, 2019), just as we observed in the current study.

Figure 7 shows that both group and individual model predictions captured age-related differences in body position. Age in months and upright time were significantly, positively correlated using group (r = 0.65, p < .001) and individual models (r = 0.52, p = .002). Similarly, sitting time increased with age based on predictions from both group (r = 0.57, p < .001) and individual models (r = 0.72, p < .001). In contrast, supine time decreased with age using predictions from both group (r = -0.61, p < .001) and individual models (r = -0.61, p < .001). No clear age trends were found for prone time. Age was not significantly related to held time for group models (r = -0.25, p = .148), but individual models showed a significant negative correlation (r = -0.37, p = .039) resulting from an outlier in the younger group.

Discussion

Here, we demonstrated the validity of long-form recordings of infant body position using wearable inertial sensors. Models trained during the proximal comparison period performed well on testing data collected from the same recording period. Most important, accuracy was consistently strong later in the distal recording period when behavior was completely unsupervised: Human-coded and model-predicted durations of body positions during the distal comparison period were highly correlated, even when narrowing to the scale of 10 minutes. Examining full recordings showed that the new method allows us to capture more data (median = 6.14 hours of awake behavior) than is typical with video methods (1-2 hours) while simultaneously reducing the human labor needed to annotate it. Ultimately, age differences in body position mirrored past findings that employed other methods, suggesting that the outcome of full-day body position recordings is suitable for describing developmental changes in motor behavior.

Accurate Results in Challenging Circumstances

How did classification in the current investigation compare to prior work (Airaksinen et al., 2020, 2022; Franchak et al., 2021)? Across body positions, median accuracy was 89.40% for group models (Kappa = 0.77) and 93.30% for individual models (Kappa = 0.85) during the proximal comparison. Although these accuracy and kappa values are slightly lower than in past work (accuracy 95-98%, kappas .93-.95), we note that those past values were obtained under ideal circumstances. A researcher applied the sensors in the lab (Airaksinen et al., 2020; Franchak et al., 2021) or home (Airaksinen et al., 2022) and set the stage for how infants and caregivers should interact. For example, both Franchak et al.

(2021) and Airaksinen et al. (2022) instructed the caregivers that the recorded sessions should mimic “playtime”, which would encourage more common activities that would be easier to classify (e.g., crawling and sitting on the floor) and discourage idiosyncratic and potentially more challenging activities (e.g., positioning in infant-specific furniture like exersaucers and walkers, sitting or lying for long periods while eating or nursing). Examining the timelines in Figure 6 shows how much variability there is between and within sessions in the earlier parts of the recording that were used for training and testing.

The challenge increased in the distal comparison period. Since sessions were scheduled at convenient times for the infant and caregiver, the transition from the proximal to distal period increased the odds that infants might need a nap, eat a snack/meal, or engage in a less typical activity (e.g., watching TV). Indeed, the distal comparison in Figure 2 contains a long period (almost 50 minutes) where the infant sat in a high chair eating lunch. Regardless of these challenges, accuracy in the distal comparison continued to be strong. Barring two outliers (which we will discuss further), agreement was high for group ($r = 0.92$) and individual models ($r = 0.94$). Even at a finer timescale, most errors within 10-minute bins were less than 1 minute in total for all five body positions. For comparison, the reliability of human coders on the body position code was ~90%, putting model-predicted accuracy on a similar level to human coders.

Idiosyncrasies in activities and device use make it challenging to decide *a priori* whether “odd” classifications (from the expectation of the researcher) are possible. For example, younger infant #8 on Figure 6 spent an unusual amount of time upright at an age where infants cannot yet stand. Inspecting the video revealed that the infant spent long periods of time suspended in a jumper that supported their body in an upright position. Although it is not feasible to collect large amounts of video data to check model predictions, interviewing caregivers about common activities and devices may provide a way to understand unexpected predictions. Counterintuitively, the two outlier participants increased our confidence that the body position classification method works as intended. At first glance, seeing low agreement rates for two sessions would suggest that the models perform poorly at predicting some *infants'* behavior. Instead, we found that for those infants, errors were restricted to two positions during a single, long-lasting event, while the other three body positions continued to be classified correctly. In other words, the models failed to predict a particular *event* for each of the two infants.

Benefits and Drawbacks of Different Modeling Approaches

Throughout the paper, we compared results from two modeling approaches: group models that included all participants' data in the training set versus individual models that used only one participant's data. When considering different metrics, testing periods, body positions, as well as the logistical benefits of each, there is no clear winner. Below, we discuss the pros and cons to each approach so that researchers might decide what works best for their intended application.

Overall accuracy and kappa values were better in individual models compared with group models when collapsed across body positions in the proximal period. However, within-class performance did not always favor individual models—prone predictions were better for group models, and held predictions were almost identical. In the distal comparison (after removing outliers), overall correlations were nearly identical. Within classes, group models had an advantage for prone, upright, and held, whereas individual models had an advantage for sitting and supine. Possibly, individual models are better suited for capturing unique aspects of the devices used for sitting and reclining that lead to better performance in those classes. Finally, age effects (either by group or by age correlation) were nearly identical across the two methods, suggesting that either method would lead to the correct conclusion regarding developmental changes in body position. Outliers were present in both models' predictions: group models produced three sessions with the worst overall accuracy, but individual models produced the single worst kappa.

The outliers for the distal comparison period appeared in both models' predictions. For aggregated full-day body position, the most blatant outlier (a young infant held $>$ 60% of the day) resulted from an individual model. The choice of model might depend on which behavior is most important to capture for a given research question.

Aside from differences in validity, researchers may favor one modeling approach based on logistical concerns. Group models have several major advantages. First, by definition they are trained on more data, which might make predictions more consistent. Second, group models remove the need to get optimal training data from each participant. It can be difficult to elicit every behavior of interest in each infant; individual models require that every behavior that will be later predicted was displayed by each individual infant. In fact, group models remove the need to get *any* training data for a particular infant. Finally, applying a group model means that individual differences between infants are due to their behavior being classified in a consistent way, rather than having a separate set of rules for predicting each individual. However, given the variety in infant behavior, individual models are more agile in capturing the unique and unexpected. Likewise, individual models can be an excellent choice for rapid prototyping and pilot testing. A researcher can get proof-of-concept data from a single participant or a new kind of classification scheme (i.e., locomotion, restraint in device) without annotating an entire sample. Individual models may also be applied in clinical cases where infants vary widely in their motor abilities and/or use of assistive devices.

As in past work (Airaksinen et al., 2020), performance was best when using all four sensors and degraded when using any subset of two (i.e., thighs only, left leg only). Small decrements and accuracy were seen for some pairs (particularly when using the left ankle and thigh), but decrements for other pairs (both ankles) or individual sensors were substantial. Using a thigh and ankle pair instead of four sensors would be more cost effective without a large decrease in accuracy, but in some applications the increased accuracy might be worth the cost. Beyond group and individual models and sensor pairings, there were many degrees of freedom in our choice of how best to model body position. For the sake of brevity, we chose to report the best approach, not every possible variation in modeling. However, our openly shared data and code allow researcher to experiment with different sets of motion features, different sets of training data, different machine learning algorithms, and different hyperparameter values. We tested but did not report models that used either only accelerometer data and only gyroscope data; in both cases, performance was degraded compared with models that use both types of motion. We also tried different machine learning algorithms (XGBoost, Tabnet), but found that random forest models performed the best. Finally, we tried modeling each age group separately, but found no benefit to performance by tailoring to the particular age group.

Novel Research Possibilities

Long-form recordings of motor behavior bring about new research possibilities. Although past work using video (Chen et al., 2022; Herzberg et al., 2021; Karasik et al., 2011) produced real-time data, such data were limited to a relatively small part of the day. And although survey methods (Franchak, 2019) could capture moments scattered throughout the day, they do not produce real-time data. The combination of real-time, full-day data about infant motor behavior is unprecedented and offers new opportunities for understanding infants' everyday experiences. Collecting dense data over the entire variety of daily experiences helps to more accurately measure infants' experiences in aggregate (as in Figure 7) without biasing results from a particular type of activity (e.g., play). Long-form recordings also have the potential to measure clinically-relevant outcomes for infants with motor delay or other pediatric concerns.

In the past decade, developmental scientists have discovered that distributional information about infants' experiences matter for how infants learn (Clerkin et al., 2017; Kachergis et al., 2017; Raz et al., 2019). Skewed distributions—those that favor the repetition of a small subset of experiences—facilitate learning. For example, recordings from wearable head cameras in the home indicate that infants see a small subset of objects with high frequency (Clerkin et al., 2017), but most objects are seen infrequently. Heterogeneity is also found in how experiences are distributed in time—burstiness and clustering are seen in infants' daily experiences seeing objects and hearing music (Casillas & Elliott, 2021; Mendoza & Fausey, 2022). Long-form motor recordings provide a novel opportunity to measure the temporal structure in the sequences of body positions seen in Figure 6. How skewed are infants' experiences with different motor behaviors, and how are they clustered in time? The distributional structure of a particular motor experience in a real day, such as time spent sitting, might be predicted by concurrent sitting skill and/or might predict future sitting skill. Moreover, real-time recording of motor behavior provides a way to measure how infant motor behavior links to other types of experiences in the moment in daily life. Malachowski et al. (2023) used a combination of LENA® and EMA surveys to find that infants heard less adult speech when restrained in seating devices. Combining LENA® and long-form motor recordings can take this a step further by measuring more precisely how speech and body position co-vary within a day.

What underpins these novel research applications is that body position classifications can be applied automatically at scale. Annotating a video corpus of infant behavior from a moderately sized sample—such as 40 infants recorded for 2 hours each in 2 separate visits (Herzberg et al., 2021)—is incredibly labor intensive. Annotating a larger, more representative sample of *hundreds of infants* while simultaneously scoring *full-day data* would be prohibitive. With sufficiently accurate group models, researchers could annotate a moderately-sized video corpus to then apply the model to hundreds of full-day recordings. Unlike with video, the added cost of more IMU sensor recording time is low, meaning that future studies could sample across multiple days of behavior to better understand intra- versus inter-day variability. Prior diary methods show that infants inconsistently display new motor skills on a day-to-day basis (Adolph & Robinson, 2011;

Adolph et al., 2008). Multi-day recordings could further uncover how infants' experiences vary over timescales from seconds to days.

Limitations

We acknowledge several limitations in the current approach. First, despite the large amount of data collected per infant, the sample of infants was small (22 unique participants). Since the study combined data from participants who completed single sessions and participants who completed multiple sessions, the seven infants with multiple sessions are over-represented in the dataset. A larger, more representative sample would be needed to determine whether the method would generalize to the broader population.

Second, a stronger test of full-day accuracy would be to compare accuracy at the start of the recording period to accuracy at the end, rather than using the distal comparison video that followed the first 90 minutes of the study. As we described in the procedure section, synchronizing video cameras and inertial sensors required capturing a synchronization point on video. We completed this procedure prior to giving equipment to caregivers so that they were not responsible for synchronization; in fact, caregivers never (purposefully) touched a button on the video camera. An end-of-day video would require a lot of effort and compliance on the part of caregivers, and any mistakes in the procedure would lead to misaligned and unusable data. Given that most recordings ended at infants' bedtimes, having an experimenter visit families in the evening to video record would be intrusive.

Third, we relied on caregiver logs of infants' naps and times when sensors were removed to separate usable data (times when infants wore the sensors while awake) from unusable data. Future work should aim to further assess caregivers' perceptions about the usability and comfort of the garment. Although infants wore the sensor garment throughout all of the desired times, we did not collect independent data (aside from caregivers' logs) to verify wear time. Most caregivers were diligent about completing logs, however, there were a few cases that caregivers may have failed to report naps (younger infants #1 and #4). In the future, algorithms can be developed to automatically identify periods of sleep and times when sensors are off the body. Such algorithms exist for adults using wrist-worn sensors, however, we did not use them because they have not been validated for infant participants wearing sensors on thighs/ankles. Differences in infants' positioning while asleep (held in caregivers arms, laying in cribs, reclined in strollers and carseats) might make classifying sleep more difficult compared with adults based on movement.

Fourth, although the goal was to collect fully naturalistic behavior, our recording protocol led to some changes that might have affected (or missed) some behaviors. We opted to restrict recording to times when participants were in the home. Inertial sensors can travel with participants, and there is no doubt that infants could have worn the sensors out of the home on errands. However, we do not yet have training data to detect periods of transportation (e.g., moving in a car or in a stroller) that might add noise and/or lead to misclassification of body positions. It is also unknown whether participants might have behaved differently had they not worn the sensors. The added bulk in the garment from the sensors (including the LENA® recorder) might have made some positions more uncomfortable, such as lying prone with the LENA® recorder worn on the chest, and/or influenced how caregivers chose to position infants throughout the day.

Conclusions

In summary, the current study demonstrates the validity of long-form recordings of infant motor behavior in the home. Our analyses show that body position classifications—whether infants are supine, prone, sitting, upright, or held—are accurate immediately and even following a substantial delay. Most important, we find substantial agreement between human-coded and model-predicted body position for data collected during truly everyday, unsupervised activities that create the most challenging cases for automatic classification. In most cases, model prediction accuracy approached human reliability, suggesting that model predictions can be confidently used in analyses of full-day infant behavior. Examining the resulting corpus of $>$ 200 hours of real-time infant body position showed the feasibility of capturing data covering the majority of infants' awake time and the variety of activities contained therein. A rudimentary analysis of aggregated data showed that full-day position estimates conformed to expectations about age differences in motor behavior. Future work employing the method can go beyond aggregated measures of behavior to uncover the temporal structure in infants' daily motor experiences.

Declarations

Funding

This work was funded by National Science Foundation Grant BCS #1941449.

Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics Approval

The study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki. The study procedures were approved by the Institutional Review Board of the University of California, Riverside, Protocol HS-15-050.

Consent

Consent to participate: All caregivers provided written informed consent prior to the start of the study. Consent to publish: Additional written consent was obtained by caregivers for data sharing of audio and video data.

Data, Material and Code availability

A Databrary repository (https://nyu.databrary.org/volume/1580) includes an exemplar participant's recording session, with the raw video data files, the Datavyu annotations of those video files, a log file with machine-readable synchronization points and nap/diaper change times, and accelerometer and gyroscope data for each of the 4 sensors. A GitHub repository (https://github.com/JohnFranchak/body_position_classification_example) contains the exemplar participant's data and source code to: 1) synchronize IMU and video annotations, 2) calculate windowed motion features for their data, and 3) train and test the body position classifier using an “individual model”. Because of the overall size of the full dataset and the computational power/time required to synchronize and create windowed datasets for each session, it would not be feasible to reproduce the calculations for all 34 sessions. However, in a second Github repository (https://github.com/JohnFranchak/body_position_classification_ms) we share the full results of those computations: The dataset of windowed motion features with corresponding body position codes used to validate the method.

References

Adolph, K. E., & Robinson, S. R. (2011). Sampling development. *Journal of Cognition and Development*, *12*, 411–423. https://doi.org/10.1080/15248372.2011.608190

Adolph, K. E., Robinson, S. R., Young, J. W., & Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychological Review*, *115*, 527–543. https://doi.org/10.1037/0033-295X.115.3.527

Adolph, K. E., & Tamis-LeMonda, C. S. (2014). The costs and benefits of development: The transition from crawling to walking. *Child Development Perspectives*, *8*, 187–192. https://doi.org/10.1111/cdep.12085

Adolph, K. E., Vereijken, B., & Denny, M. A. (1998). Learning to crawl. *Child Development*, *69*, 1299–1312. https://doi.org/10.1111/j.1467-8624.1998.tb06213.x

Airaksinen, M., Gallen, A., Kivi, A., Vijayakrishnan, P., Häyrinen, T., Ilén, E., Räsänen, O., Haataja, L. M., & Vanhatalo, S. (2022). Intelligent wearable allows out-of-the-lab tracking of developing motor abilities in infants. *Communications Medicine*, *2*(1). https://doi.org/10.1038/s43856-022-00131-6

Airaksinen, M., Räsänen, O., Ilén, E., Häyrinen, T., Kivi, A., Marchi, V., Gallen, A., Blom, S., Varhe, A., Kaartinen, N., et al. (2020). Automatic posture and movement tracking of infants with wearable movement sensors. *Scientific Reports*, *10*(1), 1–13.

Arif, M., & Kattan, A. (2015). Physical activities monitoring using wearable acceleration sensors attached to the body. *PLoS ONE*, *10*, e0130851.

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*, e12715.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bruijns, B. A., Truelove, S., Johnson, A. M., Gilliland, J., & Tucker, P. (2020). Infants' and toddlers' physical activity and sedentary time as measured by accelerometry: A systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, *17*(1), 14.

Casillas, M., & Elliott, M. (2021). Cross-cultural differences in children's object handling at home. *PsyArXiv*.

Chen, Q., Schneider, J. L., West, K. L., & Iverson, J. M. (2022). Infant locomotion shapes proximity to adults during everyday play in the u.s. *Infancy*. https://doi.org/10.1111/infa.12503

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B*, *372*, 20160055.

Cliff, D. P., Reilly, J. J., & Okely, A. D. (2009). Methodological considerations in using accelerometers to assess habitual physical activity in children aged 0–5 years. *Journal of Science and Medicine in Sport*, *12*(5), 557–567.

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & Bergelson, E. (2020). A thorough evaluation of the language environment analysis (LENA) system. *Behavior Research Methods*, *53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

Dancho, M., & Vaughan, D. (2023). *Timetk: A tool kit for working with time series in R* [https://github.com/business-science/timetk, https://business-science.github.io/timetk/].

de Barbaro, K. (2019). Automated sensing of daily activity: A new lens into development. *Developmental Psychobiology*, *61*(3), 444–464.

de Barbaro, K., & Fausey, C. M. (2022). Ten lessons about infants' everyday experiences. *Current Directions in Psychological Science*, *31*(1), 28–33. https://doi.org/10.1177/09637214211059536

Franchak, J. M. (2019). Changing opportunities for learning in everyday life: Infant body position over the first year. *Infancy*, *24*, 187–209.

Franchak, J. M. (2020). The ecology of infants' perceptual-motor exploration. *Current Opinion in Psychology*, *32*, 110–114.

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2018). See and be seen: Infant-caregiver social looking during locomotor free play. *Developmental Science*, *21*, e12626.

Franchak, J. M., Scott, V., & Luo, C. (2021). A contactless method for measuring full-day, naturalistic motor behavior using wearable inertial sensors. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.701343

Galland, B. C., Taylor, B. J., Elder, D. E., & Herbison, P. (2012). Normal sleep patterns in infants and children: A systematic review of observational studies. *Sleep Medicine Reviews*, *16*(3), 213–222. https://doi.org/10.1016/j.smrv.2011.06.001

Greenspan, B., Cunha, A. B., & Lobo, M. A. (2021). Design and validation of a smart garment to measure positioning practices of parents with young infants. *Infant Behavior and Development*, *62*, 101530.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. https://www.jstatsoft.org/v40/i03/

Herzberg, O., Fletcher, K. K., Schatz, J. L., Adolph, K. E., & Tamis-LeMonda, C. S. (2021). Infant exuberant object play at home: Immense amounts of time-distributed, variable practice. *Child Development*, *93*(1), 150–164.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive science*, *41*(3), 590–622.

Kadooka, K., Caufield, M., Fausey, C. M., & Franchak, J. M. (2021, April). Visuomotor learning opportunities are nested within everyday activities. *Paper presented at the biennial meeting of the Society for Research in Child Development.*

Karasik, L. B., Kuchirko, Y., Dodojonova, R. M., & Elison, J. T. (2022). Comparison of U.S. and Tajik infants' time in containment devices. *Infant and Child Development*, *31*(4). https://doi.org/10.1002/icd.2340

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2011). Transition from crawling to walking and infants' actions with objects and people. *Child Development*, *82*, 1199–1209. https://doi.org/10.1111/j.1467-8624.2011.01595.x

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development*, *85*, 1503–1518. https://doi.org/10.1111/cdev.12206

Kwon, S., Zavos, P., Nickele, K., Sugianto, A., & Albert, M. V. (2019). Hip and wrist-worn accelerometer data analysis for toddler activities. *International Journal of Environmental Research and Public Health*, *16*(14), 2598.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

Lobo, M. A., Hall, M. L., Greenspan, B., Rohloff, P., Prosser, L. A., & Smith, B. A. (2019). Wearables for pediatric rehabilitation: How to optimally design and use products to meet the needs of users. *Physical Therapy*, *99*(6), 647–657.

Luo, C., & Franchak, J. M. (2020). Head and body structure infants' visual experiences during mobile, naturalistic play. *PLoS ONE*, *15*, e0242009.

Majnemer, A., & Barr, R. G. (2005). Influence of supine sleep positioning on early motor milestone acquisition. *Developmental Medicine and Child Neurology*, *47*, 370–376.

Malachowski, L. G., Salo, V. C., Needham, A. W., & Humphreys, K. L. (2023). Infant placement and language exposure in daily life. *Infant and Child Development*. https://doi.org/10.1002/icd.2405

Mendoza, J. K., & Fausey, C. M. (2022). Everyday parameters for episode-to-episode dynamics in the daily music of infancy. *Cognitive Science*, *46*(8). https://doi.org/10.1111/cogs.13178

Nam, Y., & Park, J. W. (2013). Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor. *IEEE Journal of Biomedical and Health Informatics*, *17*, 420–426.

Patel, P., Shi, Y., Hajiaghajani, F., Biswas, S., & Lee, M.-H. (2019). A novel two-body sensor system to study spontaneous movements in infants during caregiver physical contact. *Infant Behavior and Development*, *57*, 101383. https://doi.org/10.1016/j.infbeh.2019.101383

Perry, L. K., Prince, E. B., Valtierra, A. M., Rivero-Fernandez, C., Ullery, M. A., Katz, L. F., Laursen, B., & Messinger, D. S. (2018). A year in words: The dynamics and consequences of language experiences in an intervention classroom (N. O. Schiller, Ed.). *PLOS ONE*, *13*(7), e0199893. https://doi.org/10.1371/journal.pone.0199893

Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., & Howard, D. (2009). A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Transactions on Biomedical Engineering*, *56*, 871–879.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Raz, H. K., Abney, D. H., Crandall, D., Yu, C., & Smith, L. B. (2019). How do infants start learning object names in a sea of clutter? *Annual Conference of the Cognitive Science Society*, 521–526.

Ren, X., Ding, W., Crouter, S. E., Mu, Y., & Xie, R. (2016). Activity recognition and intensity estimation in youth from accelerometer data aided by machine learning. *Applied Intelligence*, *45*(2), 512–529.

Stewart, T., Narayanan, A., Hedayatrad, L., Neville, J., Mackay, L., & Duncan, S. (2018). A dual-accelerometer system for classifying physical activity in children and adults. *Medicine and Science in Sports and Exercise*, *50*(12), 2595–2602.

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2018). Routine language: Speech directed to infants during home activities. *Child Development*, *90*(6), 2135–2152. https://doi.org/10.1111/cdev.13089

Thurman, S. L., & Corbetta, D. (2017). Spatial exploration and changes in infant-mother dyads around transitions in infant locomotion. *Developmental Psychology*, *53*, 1207–1221.

Warlaumont, A. S., Sobowale, K., & Fausey, C. M. (2021). Daylong mobile audio recordings reveal multitimescale dynamics in infants' vocal productions and auditory experiences. *Current Directions in Psychological Science*, *31*(1), 12–19. https://doi.org/10.1177/09637214211058166

Wass, S., Phillips, E., Smith, C., Fatimehin, E. O., & Goupil, L. (2022). Vocal communication is tied to interpersonal arousal coupling in caregiver-infant dyads. *eLife*, *11*. https://doi.org/10.7554/elife.77399

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152.

Yao, X., Plötz, T., Johnson, M., & Barbaro, K. d. (2019). Automated detection of infant holding using wearable sensing: Implications for developmental science and intervention. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(2), 1–17.

Table 1

Model performance metrics (overall accuracy, Cohen's Kappa, sensitivity, positive predictive value) from the proximal comparison period. Descriptive statistics are shown separately for group and individual models.

	Group			Individual		
Metric	Median	Mean	SD	Median	Mean	SD
Overall Accuracy	0.894	0.846	0.131	0.933	0.916	0.072
Kappa	0.768	0.746	0.161	0.849	0.821	0.143
Sensitivity	0.856	0.825	0.127	0.847	0.841	0.119
Pos Pred Value	0.826	0.810	0.125	0.928	0.899	0.107

Table 2

Model performance metrics for each body position (supine, prone, sitting, upright, and held) during the proximal comparison period, shown separately for group and individual models.

		Group			Individual		
Metric	Position	Median	Mean	SD	Median	Mean	SD
Kappa	Supine	0.907	0.764	0.295	0.983	0.912	0.166
	Prone	0.968	0.860	0.259	0.942	0.841	0.246
	Sitting	0.816	0.702	0.297	0.915	0.887	0.127
	Upright	0.707	0.673	0.281	0.822	0.741	0.236
	Held	0.732	0.726	0.209	0.826	0.727	0.277
Sensitivity	Supine	1.000	0.905	0.180	1.000	0.954	0.129
	Prone	1.000	0.894	0.234	0.974	0.849	0.272
	Sitting	0.910	0.811	0.257	0.964	0.915	0.136
	Upright	0.837	0.730	0.297	0.891	0.786	0.253
	Held	0.852	0.772	0.228	0.773	0.702	0.312
Pos Pred Value	Supine	0.995	0.828	0.292	1.000	0.932	0.134
	Prone	0.987	0.877	0.236	1.000	0.892	0.210
	Sitting	0.896	0.802	0.261	0.972	0.945	0.079
	Upright	0.839	0.739	0.283	0.923	0.825	0.228
	Held	0.852	0.794	0.240	0.943	0.899	0.134

Table 3

Median, mean, and SD of overall accuracy calculated with different sets of sensor features. The top row shows performance using all features calculated from the four sensors. Rows 2-5 show accuracy using pairs of sensors (left thigh and ankle, right thigh and ankle, left and right thigh, left and right ankle), and rows 6-9 show accuracy using each individual sensor.

	Group			Individual		
Sensors	Median	Mean	SD	Median	Mean	SD
All	0.894	0.846	0.131	0.933	0.916	0.072
Left Thigh/Ankle	0.891	0.824	0.177	0.915	0.901	0.071
Right Thigh/Ankle	0.842	0.800	0.167	0.897	0.898	0.086
Both Thighs	0.859	0.791	0.195	0.907	0.898	0.072
Both Ankles	0.797	0.719	0.196	0.864	0.844	0.108
Left Thigh	0.842	0.762	0.208	0.889	0.872	0.077
Left Ankle	0.717	0.667	0.208	0.851	0.825	0.117
Right Thigh	0.846	0.772	0.178	0.896	0.876	0.092
Right Ankle	0.704	0.683	0.160	0.830	0.819	0.114

Table 4

Correlations between human-coded and model-predicted body position durations across the entire distal comparison period. Correlations are provided within each body position and overall. Correlations are presented separately for group and individual models with and without the two outlier participants.

	With Outliers		Without Outliers	
Position	Group	Individual	Group	Individual
Supine	0.88	0.98	0.94	0.97
Prone	0.97	0.86	0.97	0.84
Sitting	0.79	0.97	0.91	0.95
Upright	0.63	0.83	0.99	0.95
Held	0.02	0.04	0.73	0.60
Overall	0.80	0.91	0.95	0.96

Table 5

Correlations between human-coded and model-predicted body position durations using 10-minute bins during the distal comparison period. Correlations are provided within each posture and overall, and computed separately using group and individual models with and without outlier participants.

	With Outliers		Without Outliers	
Position	Group	Individual	Group	Individual
Supine	0.76	0.96	0.88	0.93
Prone	0.96	0.90	0.96	0.89
Sitting	0.72	0.93	0.89	0.92
Upright	0.91	0.93	0.98	0.96
Held	0.51	0.46	0.67	0.63
Overall	0.80	0.94	0.92	0.94

Table 6

Summary of age differences in full-day body position for younger (4- to 7-month) and older (11- to 14-month) infants. Values shown are the mean percent of time for each body position averaged across infants in each group. Standard deviations are shown in parentheses. Descriptive statistics are shown separately for group and individual models.

	Group		Individual	
Position	Younger	Older	Younger	Older
Supine	38.6% (24.1)	14.0% (8.1)	43.1% (32.0)	10.5% (9.6)
Prone	14.0% (14.0)	15.1% (6.3)	11.1% (11.4)	16.6% (10.2)
Sitting	28.9% (15.1)	44.4% (9.4)	20.3% (14.8)	46.6% (12.4)
Upright	5.4% (7.1)	18.2% (7.4)	9.5% (13.3)	18.4% (7.8)
Held	13.0% (7.7)	8.3% (5.2)	16.0% (20.1)	8.0% (7.6)

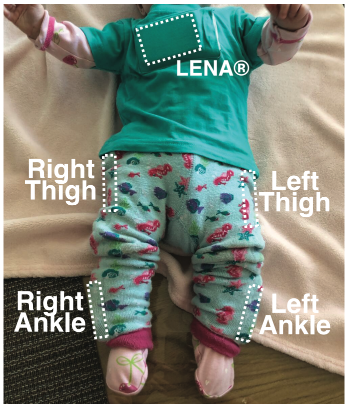

Figure 1. Sensor garment worn by infant participant. Four IMUs were placed in interior pockets sewn into a tightly-fitting pair of infant leggings. White dashed rectangles mark the approximate locations of each sensor pocket (above the left and right ankles and on the left and right thighs, just below the thighs). A white dashed rectangle also marks the LENA audio recorder worn in a pocket on the infant's shirt.

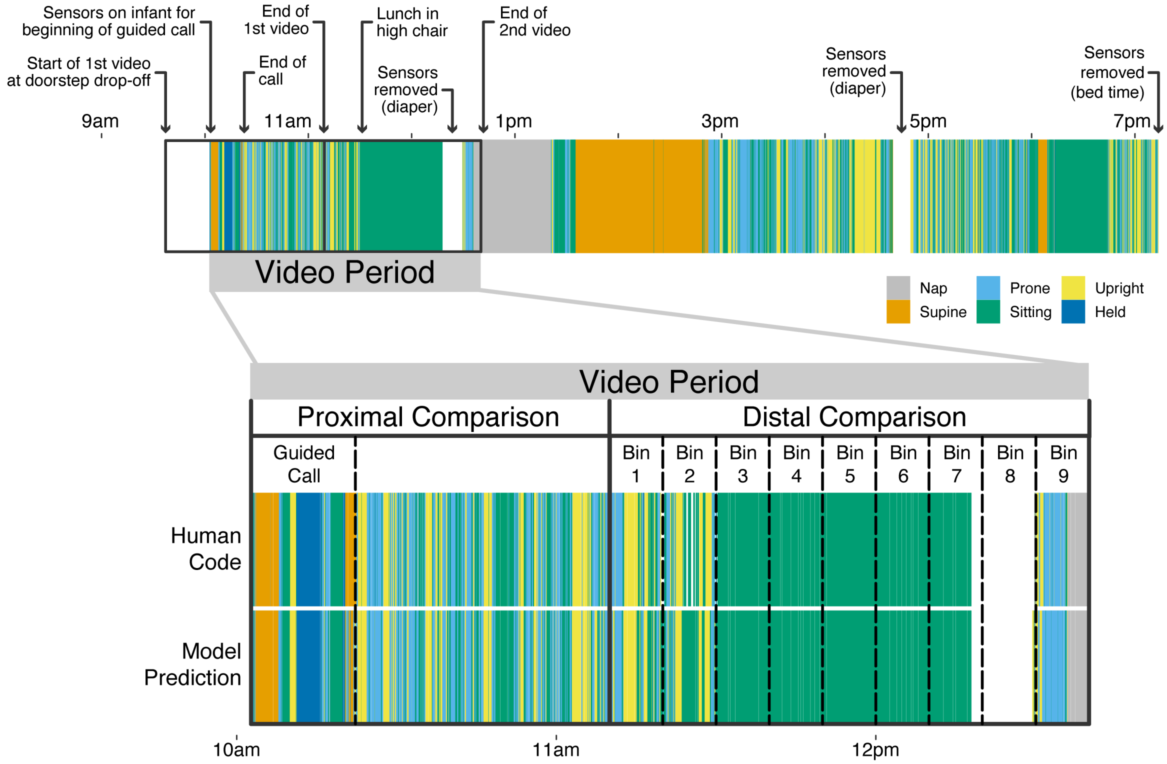

Figure 2. Timeline from an exemplar participant (older infant 15). The top row shows the model-predicted body position across the entire recording period. Annotations indicate when the video camera was turned on by the experimenter when arriving at the house, when the sensors were first placed on the infant, when the guided call took place, and when the video files were recorded. Gray areas on the timeline indicate naps, and white areas indicate times when the sensors were removed. The bottom row shows a zoomed-in view of the video period during which ground truth data were available. The top timeline shows human-coded body position and the bottom row shows model-predicted body position; these were the data used for validation. The first part of the video period was the proximal comparison, when video and motion data were highly synchronized. The second part of the video recording was the distal comparison that had coarser synchronization. Accuracy data for the distal comparison are provided overall and during 10-minute bins, marked by vertical dashed lines.

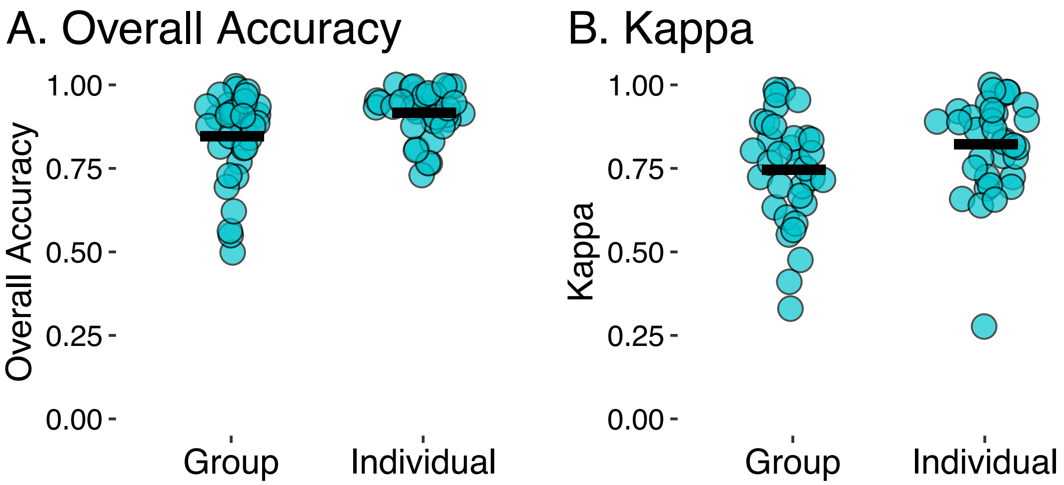

Figure 3. Metrics of agreement between human-annotated body position and model predictions of body position from the proximal comparison period. Overall accuracy (A) and Cohen's Kappa (B) are plotted separately for group models and individual models. Each blue circle represents the accuracy for each recording session. Horizontal black bars indicate the mean across sessions.

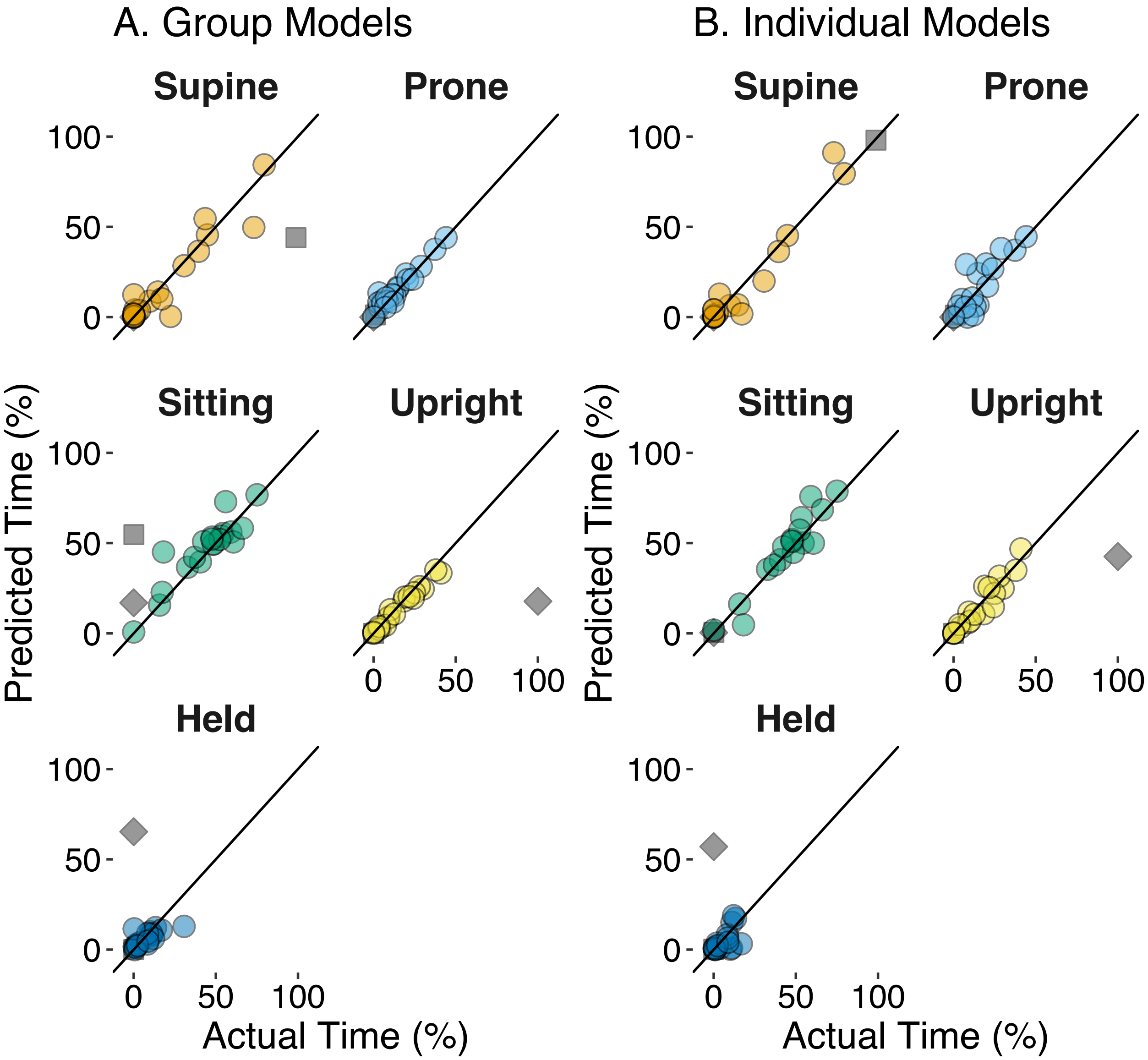

Figure 4. Overall agreement between human-coded body position and model-predicted body position in the distal comparison. Agreement for group models is shown in (A) and agreement for individual models is shown in (B). Plots are shown separately for each body position with a reference line that indicates perfect agreement; each point in a plot represents data for a single session. The two outlier participants are plotted in dark gray, with a different shape marking each individual.

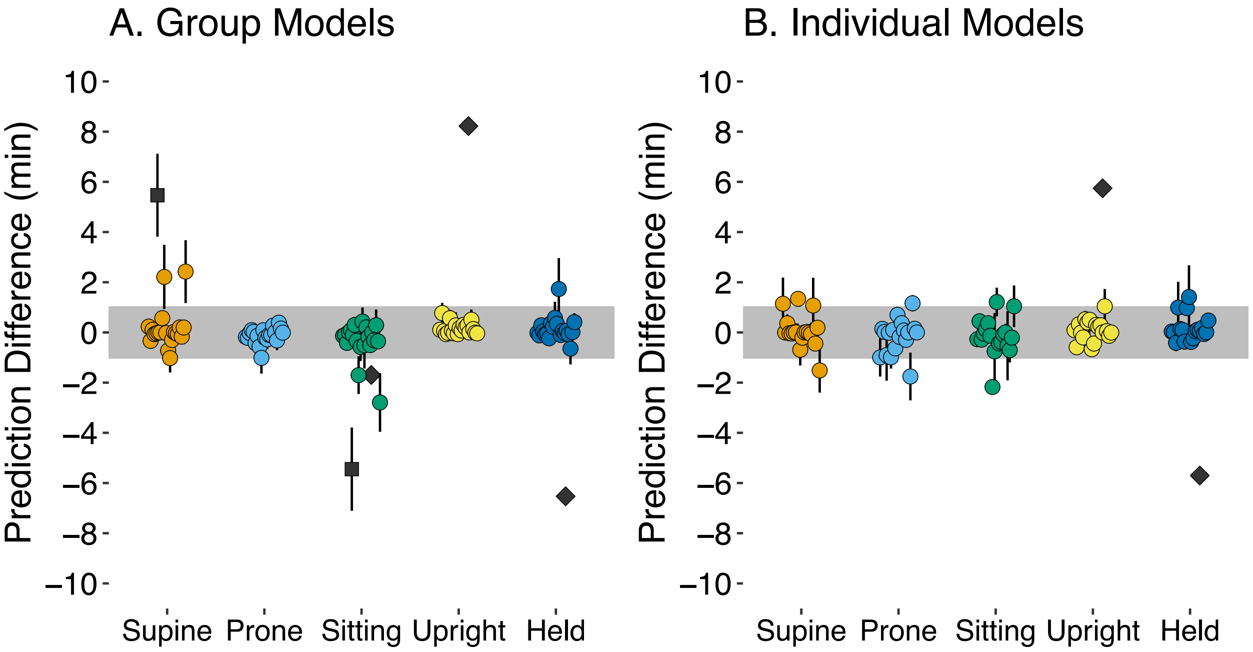

Figure 5. Prediction difference (difference in minutes between human-coded and model-predicted body position) for 10-minute bins in the distal comparison period. Each point shows the mean and SE for a single recording session for each body position, summarizing the prediction difference for each of their 10-minute bins. Points falling within the gray shaded region indicate that average prediction errors were less than 1 minute. Performance is plotted separately for (A) group models and (B) individual models. The two outlier participants are plotted in dark gray, with a different shape marking each individual.

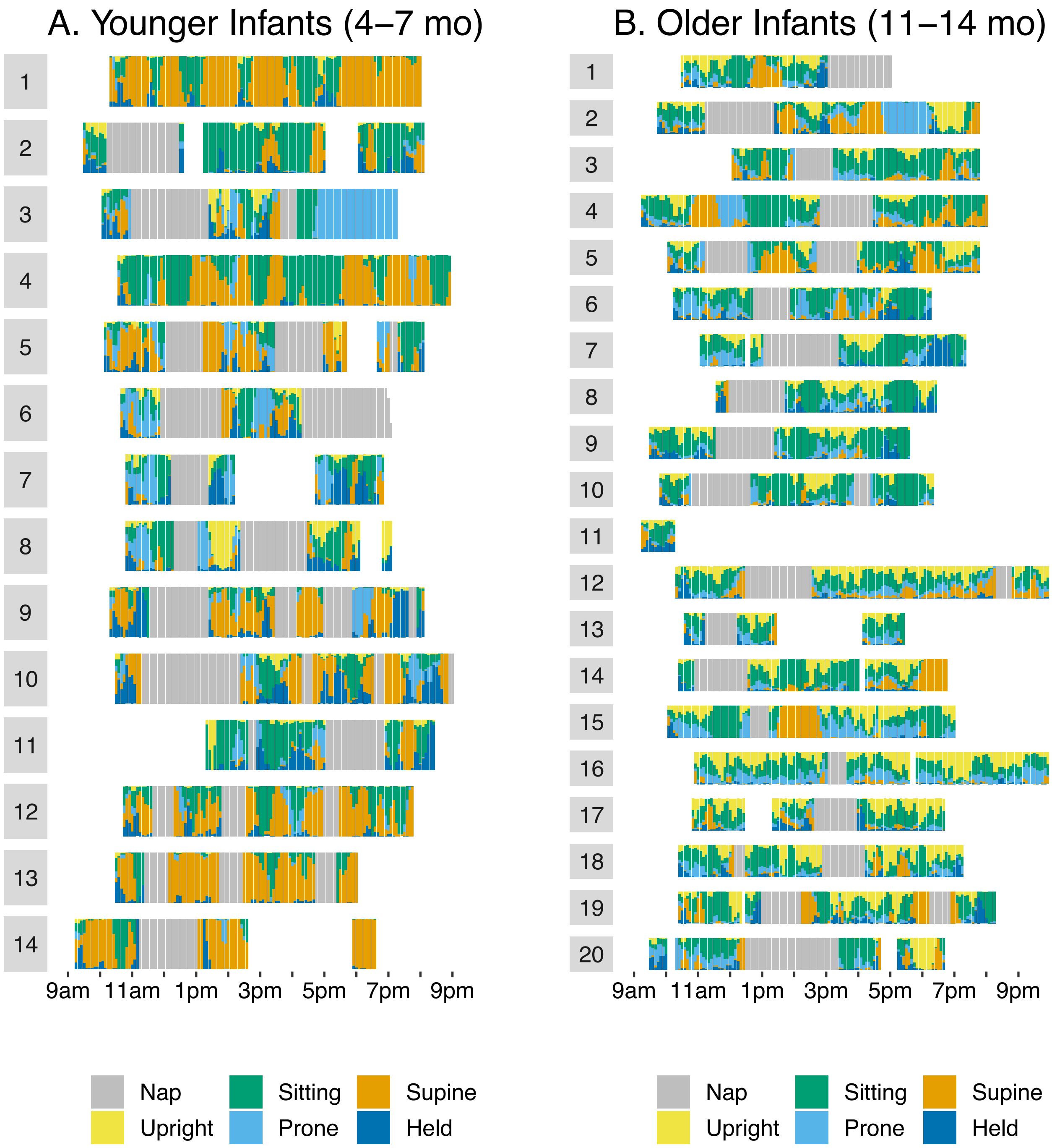

Figure 6. Full-day timelines for each individual recording session, split by (A) younger infants and (B) older infants. Each participant's timeline shows a stacked bar graph with the proportion of time spent in each of the body positions for every 5-minute period throughout the day, based on group model predictions of body position. The x-axis shows time of day. Caregiver-reported naps are marked by gray bars; blank gaps indicate caregiver-reported times that the sensor garment was removed for diaper changes, baths, or trips out of the house.

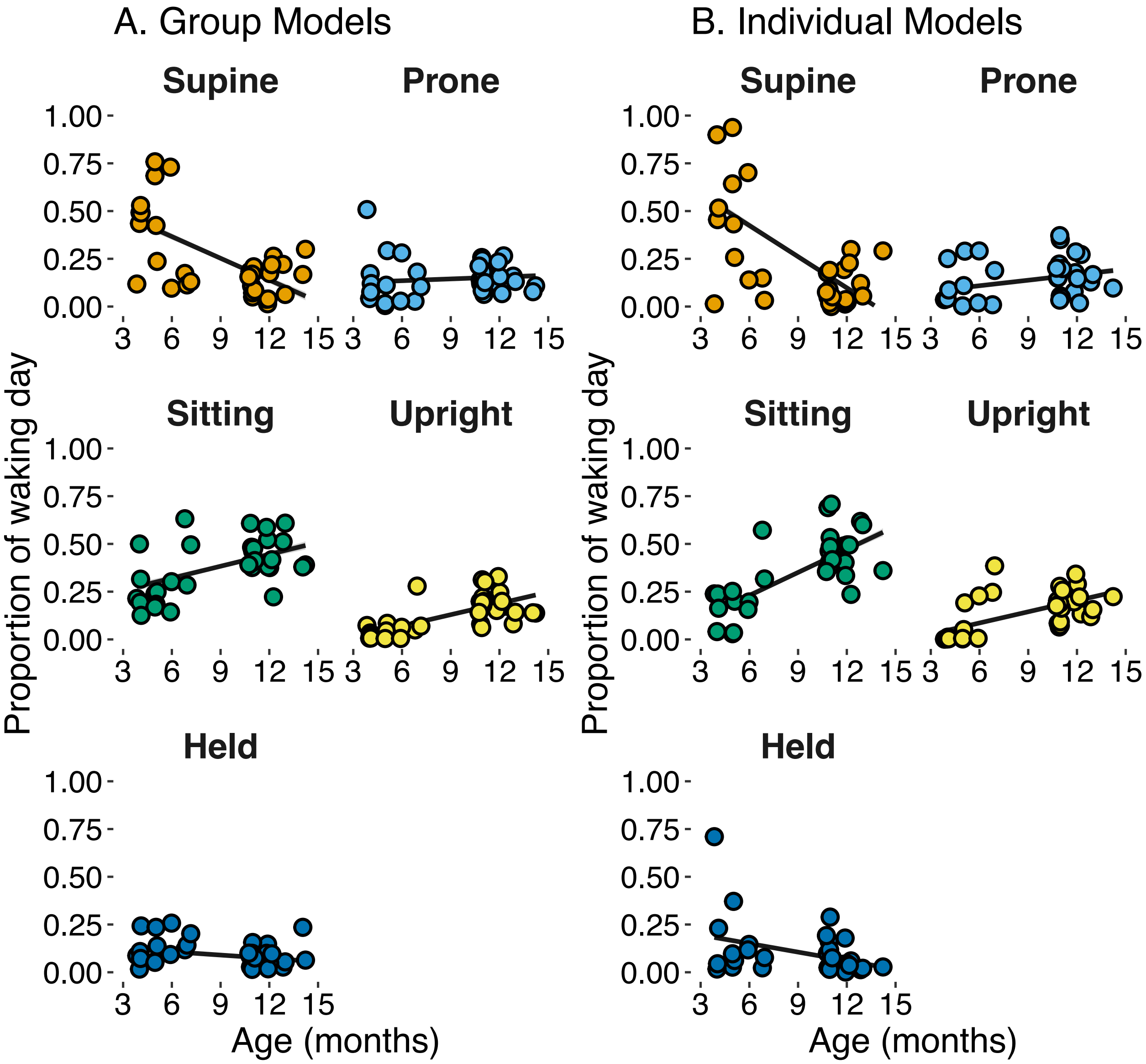

Figure 7. Age differences in daily body position predicted from (A) group models and (B) individual models. Each circle represents one full-day recording session's proportion of time in each body position (y-axis) against age in months (x-axis). Black lines indicate best fit regression lines, which show decreases in supine time and increases in sitting and upright time with age.