Title
Artificial intelligence-powered 3D analysis of video-based caregiver-child interactions.

Journal
Science Advances, 11(7)

Authors
Weng, Zhenzhen
Bravo-Sánchez, Laura
Wang, Zeyu
et al.

Publication Date
2025-02-14

DOI
10.1126/sciadv.adp4422

Peer reviewed

NEUROSCIENCE

Artificial intelligence–powered 3D analysis of video-based caregiver-child interactions

Zhenzhen Weng[1]†, Laura Bravo-Sánchez[2], Zeyu Wang[2], Christopher Howard[3]‡, Maria Xenochristou[2]§, Nicole Meister[4], Angjoo Kanazawa[5], Arnold Milstein[6], Elika Bergelson[7], Kathryn L. Humphreys[8], Lee M. Sanders[9]*, Serena Yeung-Levy[2,3,4,6]*

We introduce HARMONI, a three-dimensional (3D) computer vision and audio processing method for analyzing caregiver-child behavior and interaction from observational videos. HARMONI operates at subsecond resolution, estimating 3D mesh representations and spatial interactions of humans, and adapts to challenging natural environments using an environment-targeted synthetic data generation module. Deployed on 500 hours from the SEEDLingS dataset, HARMONI generates detailed quantitative measurements of 3D human behavior previously unattainable through manual efforts or 2D methods. HARMONI identifies longitudinal trends in child-caregiver interaction, including child movement, body pose, dyadic touch, visibility, and conversational turns. The integrated visual and audio analysis further reveals multimodal trends, including associations between child conversational turns and movement. Open-sourced for large-scale analysis, HARMONI facilitates advancements in human development research. HARMONI achieves 63 to 80% consistency on key attributes with human annotators on SEEDLingS and 84 to 93% consistency on videos taken from a laboratory setting while achieving >100 times savings in time.

INTRODUCTION

During the initial stages of human life, the brain undergoes notable developmental changes, marked by a pronounced plasticity that allows for it to be molded by experience, most often via child interactions with their caregivers (*1*). Examining the behaviors of infants and children in the context of caregiver interactions provides valuable insight into how these experiences may shape development (*2*). Developmental scientists have used the best available methods to document children's developing competencies (*3*), and now advancements in video technology have facilitated a proliferation of child observations recorded for analysis. Despite progress in data capture, the review and quantification of data properties often encounter constraints due to the manual coding of extensive video footage, which demands considerable time and expense. This limitation hinders the capacity to conduct large-scale quantitative analysis, ultimately restricting the depth and scope of scientific investigation achievable through video observation. Manual coding and smaller-scale studies also reduce the heterogeneity of samples, potentially impairing the insights to be gained from a more diverse set of children, caregivers, and families. As a result, it becomes more difficult to translate basic science to real-world settings, such as pediatric care, early childhood education, and the home.

In light of the challenges associated with manual video coding, innovative techniques for extracting and quantifying analyzable properties of human behavior in video recordings may offer alternative avenues for exploring scientific hypotheses. One such approach involves use of auxiliary sensors and devices, generating quantitative streams of values, such as from wearable accelerometers or head-mounted cameras (*4*). However, these sensors can prove unwieldy and unfeasible for implementation across extensive study populations. More recent approaches (*5*–*8*) attempt to directly estimate parameters from video data using artificial intelligence (AI) techniques. Applications of natural language processing have yielded phoneme, syllable, and word count estimates from audio recordings (*9*, *10*), while computer vision has facilitated predictions of 2D object and body keypoint locations in images (*11*–*13*). Nevertheless, such video analysis efforts remain rudimentary and brittle, failing to capture the intricacy and sophistication of human interaction within a three-dimensional (3D) context.

Here, we introduce HARMONI (Holistic 3D Analysis of Responsive Human Movements from Observational Video of Natural Interactions), a 3D computer vision and audio method for the extraction and quantification of intricate properties of 3D human behavior and interaction from single-view videos featuring children and caregivers. Our approach is capable of operating at a granular temporal resolution in an automated manner, estimating detailed 3D mesh representations (i.e., 3D digital reconstructions) of humans in video frames over time. HARMONI demonstrates robust adaptability to videos collected from challenging natural environments (e.g., home settings) through an environment-targeted synthetic training data generation module, eliminating the need for collecting extensive labeled training datasets in individual environments. The 3D human mesh estimates produced by HARMONI facilitate the computation and interpretation of features such as human movement, visibility, touch, and body pose at subsecond resolution.

In developing HARMONI, we specifically selected features grounded in well-established developmental science research. These features are intended to capture essential aspects of child-caregiver interactions that substantially influence child development. For

[1]Institute for Computational & Mathematical Engineering, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [5]Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA, USA. [6]Clinical Excellence Research Center, Stanford University, Stanford, CA, USA. [7]Department of Psychology, Harvard University, Cambridge, MA, USA. [8]Department of Psychology and Human Development, Vanderbilt University, Nashville, TN, USA. [9]Pediatrics and Health Policy, Stanford University, Stanford, CA, USA.
*Corresponding author. Email: lsanders@stanford.edu (L.M.S.); syyeung@stanford.edu (S.Y.-L.)
†Present address: Waymo LLC, Mountain View, CA, USA.
‡Present address: Ekho Inc., New York, NY, USA.
§Present address: Fujitsu Research of America, Santa Clara, CA, USA.

example, 3D spatial understanding facilitates the measurement of key metrics such as location, distance traversed (*14*), caregiver-child proximity (*15*), touch (*16*), and field of view (*17*). Research has shown that these physical dynamics—such as the proximity of a child to their caregiver, the child's orientation, and their movement patterns—are critical indicators of developmental progress and have been linked to various developmental outcomes. Additionally, child pose (*18*) relates to motor development milestones (*19*), while audio features, including child-initiated conversations (*20*), conversational turn counts (CTCs) (*21*), and adult speech counts (*22*), are crucial for assessing shared attention and language learning. By incorporating these features into the tool, we ensure that it provides outcomes that are relevant and valuable to both developmental scientists and social scientists. While a subset of these outputs has been previously studied for child development (e.g., caregiver-child distance and adult word counts), HARMONI augments previous work by extracting multiple multimodal features with a single method. Moreover, HARMONI's approach to feature definition based on mesh representations presents an opportunity for researchers to redefine or augment features for their use case (e.g., subtypes of behaviors). There are further possibilities that have not been studied in this work but could be facilitated by HARMONI. Similarly to (*23*), HARMONI's mesh-based representation could be used to anonymize videos depicting sensitive data.

We validate HARMONI using two datasets comprising video of caregiver-child interactions, SEEDLingS (*24*) from natural home environments, and CMU Panoptic-Toddlers (*25*) from a laboratory setting. The SEEDLingS dataset encompasses >500 hours of longitudinal interaction videos from 44 children, and their caregivers, recorded monthly between child age 6 to 17 months. Here, we explore whether and to what extent child movement, body pose, caregiver-child touch (i.e., any body part of one person being in contact with any body part of the other), and caregiver-child visibility (i.e., whether the other party is in their view) change across development, given that we expect there may be change, although not previously tested, in each of these domains in relation to chronological age. Our approach not only analyzes this substantial volume of video data without incurring human labor costs, but also generates detailed quantitative measurements of 3D human behavior and interaction, previously unattainable through manual efforts or existing methods. Previous works that quantify human interactions have relied on self-reported measures (*26*–*29*), sensor data from electronic devices (*30*, *31*), and 2D or semiautomatic 3D outputs from computer vision algorithms (*11*, *32*). HARMONI stands apart by introducing a fully automated approach that goes beyond these conventional methods. By harnessing the power of 3D computer vision and state-of-the-art audio processing methods, HARMONI enables the extraction and analysis of highly detailed quantitative indicators from video recordings of natural human interactions. We include CMU Panoptic-Toddlers as an additional validation set to demonstrate that HARMONI's performance can be enhanced with optimal video capture settings, where the ideal camera position offers a clear, unobstructed view of both individuals.

We demonstrate how HARMONI's visual analysis can be integrated with automated audio analysis to explore multimodal, audiovisual trends, such as the interaction between increased child CTCs and child movement. HARMONI can be an effective instrument for producing scientific insights and verifying hypotheses across diverse populations, developmental phases, and cultural backgrounds. We make HARMONI publicly available as a tool for facilitating large-scale quantitative analysis of human behavior in video recordings, paving the way for innovative advancements in human development research.

RESULTS

AI-based extraction of 3D mesh representations of children and caregivers from video

HARMONI, an AI-based method, automates the extraction of 3D mesh representations of humans from video, enabling the computation of objective measurements that describe human behavior and interaction. Figure 1 shows a visualization of a video sequence with child and caregiver human meshes extracted from the visual data and audio features extracted from the audio data, and example caregiver-child interaction measurements that can be extracted. Specifically, given a video as input, HARMONI produces a set of 3D human meshes for every video frame, corresponding to detected children and caregivers. First, it performs body tracking and body type classification to detect children and caregivers in video frames, using only a few user-provided bounding boxes of each individual as guides. Subsequently, using the SMIL (*33*) body model for children and the SMPL (*34*) body model for adult caregivers, it employs a deep neural network to estimate the parameters corresponding to the respective body model for each detected human. These parameters include a 10D vector for body shape, a 72D vector for 3D body joint locations in axis-angle representation, and a 6D vector for camera parameters. Combined, the parameters specify a 3D human mesh through a forward generative process.

The deep neural network used by HARMONI for estimating 3D human mesh parameters is pretrained on a set of widely used public computer vision datasets (*35*, *36*), not from human development studies. These datasets include those collected with 3D ground truth using motion capture, as well as naturalistic image datasets of humans manually annotated with 2D keypoints. However, training on these datasets alone may not transfer to previously unseen video data distributions, such as those in human development research. Therefore, HARMONI includes an environment-targeted synthetic training data generation module, which, given raw, unannotated video from a new environment (e.g., an unseen home setting), automatically generates synthetic training data representative of that environment to fine-tune the 3D mesh estimation model. Since SEEDLingS, like many in-the-wild video captures, lacks the annotated labels required for supervised training of a human pose estimator, this kind of training paradigm helps us create synthetic training data that mimic the real data distribution, allowing us to train an effective human pose estimator. Last, HARMONI uses an optimization algorithm to refine estimated body parameters based on physical plausibility constraints, such as adjusting the body to be above the ground plane and smoothing predictions across frames.

Quantification of individual child and caregiver behaviors

The sequence of 3D child and caregiver human meshes output by HARMONI for each video frame enables the computation of objective measurements that describe human behavior and interaction. For instance, 3D distance and movement across a physical space can be computed. The relative positioning of body parts can be filtered or clustered to quantify the presence of body poses and behaviors. Head orientation can also be used to determine

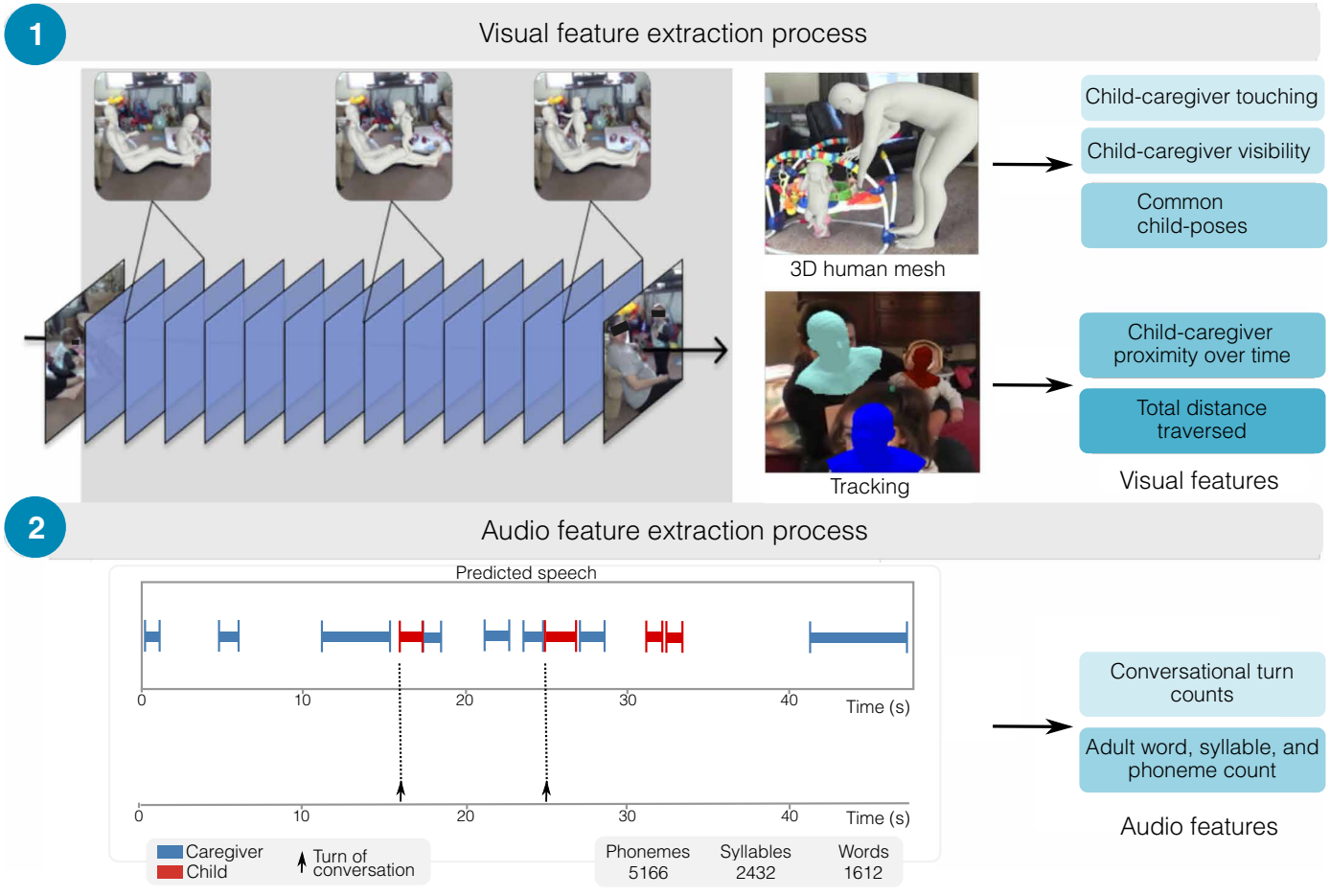

Fig. 1. Overview of key capabilities of HARMONI. Visualization of a video sequence with extracted caregiver-child human meshes, and example measurements that can be extracted with HARMONI.

field-of-view and e, and distance between child and caregiver meshes can be used to determine touching and proximity over time. Additional measurements may be composed using the human mesh as the starting point.

As HARMONI is a fully automated approach, it can extract measures of interest from large amounts of video without requiring human labor. We deployed HARMONI on 500 hours of video from SEEDLingS. This dataset comprises hour-long recordings each month for 44 caregiver-child dyads, from ages 6 to 17 months. Figure 2 shows the longitudinal distribution of measures characterizing individual child and caregiver behaviors, computed over these months. Specifically, Fig. 2 (A and B) displays child and caregiver distance traversed (in meters), respectively. Figure 2C shows an example of one caregiver-child pair movement trajectory in 3D, during a recording period. Metrics obtained from HARMONI reveal an increase in child distance traversed across development ($\beta = 0.30$ [95% confidence interval (CI) 0.21, 0.40], $P = 1.417 \times 10^{-12}$) and a decrease in caregiver distance traversed across child development ($\beta = -0.14$ [95% CI −0.22, −0.06], $P = 4.454 \times 10^{-4}$).

Figure 2D presents changes in child pose over time. We clustered the poses into four poses of interest for children based on the orientation of the reconstructed 3D human bodies and plotted the percentage of time the child spent in each pose, after filtering out frames where no child is detected as well as frames where the child pose is ambiguous (only the upper half of the body is detected). We observe that across development, children spend increasing amounts of time in an upright pose ($\beta = 0.34$ [95% CI 0.26, −0.42], $P = 4.686 \times 10^{-16}$). Note that SEEDLingS contains complex scenarios where the child may be in an upright position with assistance (such as being in a chair or being carried by a caregiver), and to distinguish between upright with and without assistance, we would need to detect and localize the objects in the scene in addition to the HARMONI-produced poses, as a child could be assisted by a baby bouncer or walker. Therefore, we leave that for future work. However, we conducted a sensitivity analysis by filtering out frames where the child is not on the ground plane and found no notable difference in the trend compared to Fig. 2D. Across SEEDLingS, we did not detect statistically significant trends in the proportion of child time spent in the other considered poses (i.e., seated, prone, and supine), as a function of age.

Quantification of dyadic child and caregiver behaviors

Next, we quantified dyadic child and caregiver behaviors. Figure 3A displays the proportion of time over hour-long recording periods that the child and caregiver touched, over all families and months of development. Touching decreased across child development ($\beta = -0.27$

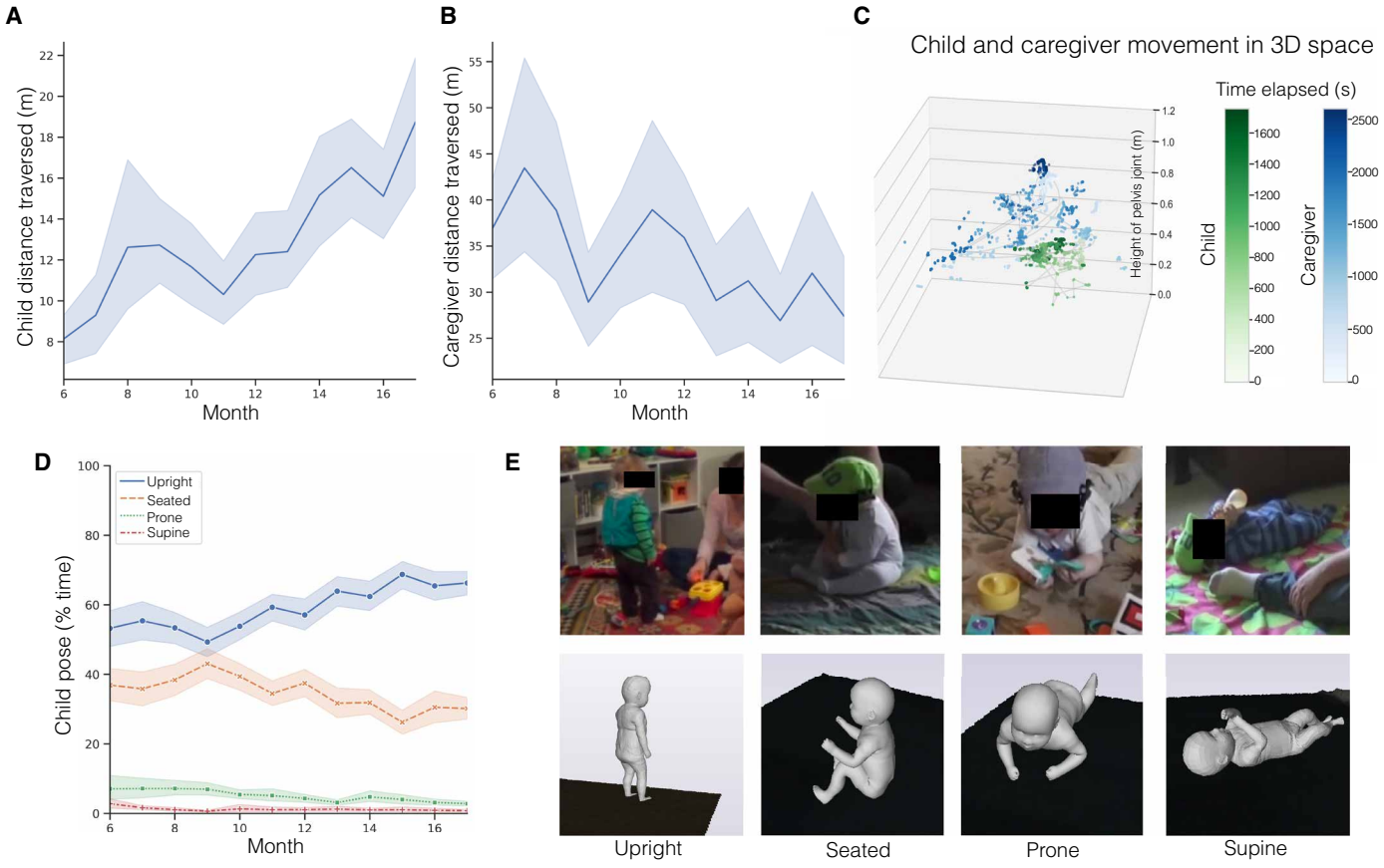

Fig. 2. Quantification of individual child and caregiver behaviors. (**A**) Total child distance traversed in the room during an hour-long recording period, over all families and months of development. (**B**) Total caregiver distance traversed in the room across time. (**C**) Example of child and caregiver movement across a room over the length of a recording, for one family. (**D**) Proportion of time that child exhibits body poses corresponding to "upright," "sitting," "prone," and "supine" over all families and months of development. (**E**) Examples of video frames and corresponding HARMONI-extracted 3D child meshes, which are automatically categorized under each of the pose categories. On the basis of the orientation of the reconstructed 3D human bodies, poses are categorized into prone, supine, and either upright or seated. Upright and seated are further differentiated based on the angle between torso and thighs. Plots show central tendency and 95% CIs as shaded areas.

[95% CI −0.37, −0.17], $P = 4.945 \times 10^{-10}$). Figure 3B demonstrates touching computed from 3D mesh representations. Given the 3D estimation, the same caregiver-child dyad can be visualized from different perspectives for additional clarity (e.g., camera view and view from above). Whether two humans are touching can be computed based on the distance between the corresponding 3D meshes. Some of the deviations from the overall longitudinal trend in touching are partially due to special cases of behavior observed during some months. For instance, month 8 and 9 recordings for one participating family consisted of long play sessions, which led to spikes in touching.

Figure 3C presents the change in dyadic visibility over months of development. In this case, visibility is defined as "in the visual range (i.e., field of view)," with a 180-degree field of view based on head orientation. Figure 3D illustrates how this visibility property can be computed using the head orientation of a 3D human mesh. Here, no statistically significant change is observed for (i) caregiver being in the visual range of the child, (ii) child being in the visual range of the caregiver, or (iii) both in each other's visual range, as a function of child age. Finally, Fig. 3E displays change in relative caregiver-child distance over recording periods, building on the individual room traversal measures in Fig. 2. Again, no statistically significant change is observed. See Fig. 3F for an example of how the relative caregiver-child distance can be computed from a video.

Quantification of multimodal audiovisual interactions

In a final analysis, we used HARMONI to quantify multimodal audiovisual interactions within the 500 hours of SEEDLingS data. Here, HARMONI's visual analysis is integrated with an existing audio analysis pipeline that extracts audio from video recordings and outputs speaker diarization (i.e., who is talking when) along with associated word, phoneme, and syllable count estimates. Further details on our audio analysis implementation can be found in Materials and Methods. Figure 4A shows an example of the multimodal, interleaved visual and audio outputs produced by HARMONI. Figure 4B illustrates the relationship between the amount of time (seconds) the child is engaged in a self-initiated conversational turn and child movement, revealing a correlation between increased verbal initiation and a concurrent rise in independent movement, as gauged by the distance traversed in the room. Child-initiated conversation duration was associated with distance traversed, $r = 0.126$ [95% CI 0.040, 0.210], $P = 4.186 \times 10^{-3}$. The association remained, though was attenuated, after co-varying for child age, $r_{partial} = 0.089$, $P = 4.351 \times 10^{-2}$. Correlation and partial correlation are computed

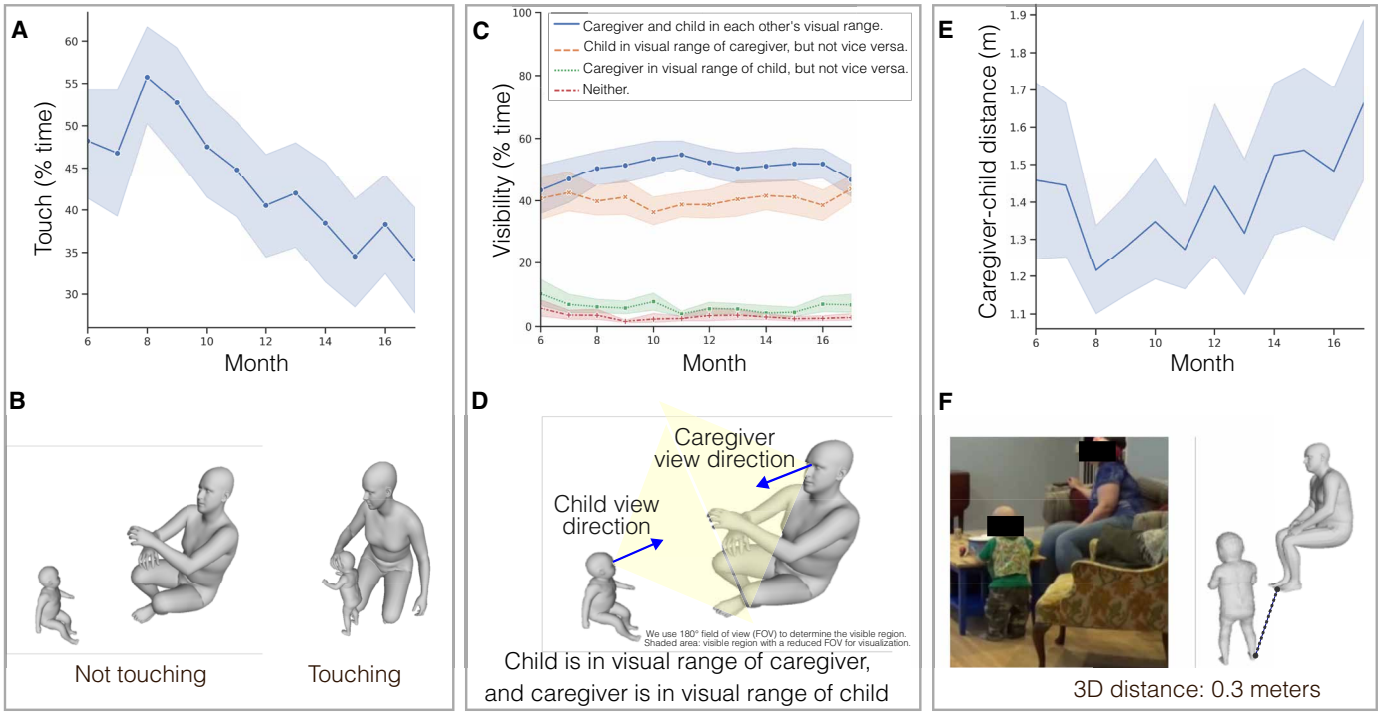

Fig. 3. Quantification of dyadic child and caregiver behaviors. (**A**) Caregiver-child touching, across families by child age. (**B**) Example of how touching can be computed from 3D mesh representations. (**C**) Child and caregiver touching. (**D**) Example of how visibility can be computed using the head orientation of a 3D human mesh. (**E**) Total caregiver-child relative distance, across families by child age. (**F**) Example of how 3D distance between child and caregiver can be computed.

over all videos. This trend is particularly pronounced among 3 of the 44 families, where the correlation between child-initiated conversation duration and distance traversed is greater than 0.5 (Fig. 4C). Nevertheless, no such pattern is observed for caregiver distance traversed or caregiver-child visibility measures.

HARMONI consistency with human annotation

The sequences of 3D meshes generated by HARMONI, representing children and caregivers, exhibit strong consistency with independent human annotation of both 2D body keypoint locations and measures of visibility, touch, and body pose.

SEEDLingS dataset

To evaluate the accuracy of the reconstructed 3D meshes, their 2D projections were compared with independent human annotation of 2D keypoint locations for 23 major body joints per person. The evaluation set consists of 1400 video frames, where 3 frames were randomly sampled from each video. On the entire evaluation set, the model achieves 63.8% and 80.0% consistency with manual annotation (treated as the "ground truth"), as measured by percentage of correct keypoints (PCK), for children and caregivers, respectively. Figure 5A shows examples of 3D mesh estimations corresponding to different measured PCK consistency values. While a comparison with 3D annotations is not feasible due to the impracticality of human annotation in 3D, the consistency of 2D projections with 2D annotations provides a useful assessment, as the 3D representation is also constrained by the known structure and articulation of human bodies modeled in the SMPL and SMIL body representations.

Using the generated 3D meshes to extract example objective measures of interest, we observe high consistency with human annotations of visibility, touch, and body pose. On the 784 images from the evaluation set where there exist both caregiver and child, the determination of whether touching occurs is consistent between the computer vision model and the human 66.6% of the time, and the determination of visibility is consistent 66.4% of the time. Precision/recall for touch and visibility is 67%/68% and 66%/67%, respectively. In this context, visibility is defined as one of four categories: Both child and caregiver are in each other's field of view, only one is in the field of view of the other (either caregiver or child), and neither are in the field of view of each other. The task was more challenging for younger, smaller children; for ages 14 to 17 months, touch consistency rises to 75.3% and visibility to 71.2%.

We note that our accuracies are calculated on frames where both annotators agree with each other and that the model detects both adult and child. Even for human coders, accurately determining whether a dyad is in contact or visible to each other solely from 2D images can be difficult (annotators agree with each other 92% of the time). Figure 5B shows examples of manual annotation versus HARMONI outputs. It includes instances with discrepancies between HARMONI and the annotators' assessment that illustrate failure modes. However, HARMONI often produces relatively reasonable outputs even when different from the manual annotation (Fig. 5B). We also highlight that SEEDLingS contains nonstandardized, unconstrained home environments with substantial occlusions, and video capture was not optimized for AI-based video analysis. Higher consistencies would likely be observed in more controlled environments, or when video capture settings are optimized for AI analysis. This is as demonstrated through our second validation set, CMU Panoptic-Toddlers (*25*).

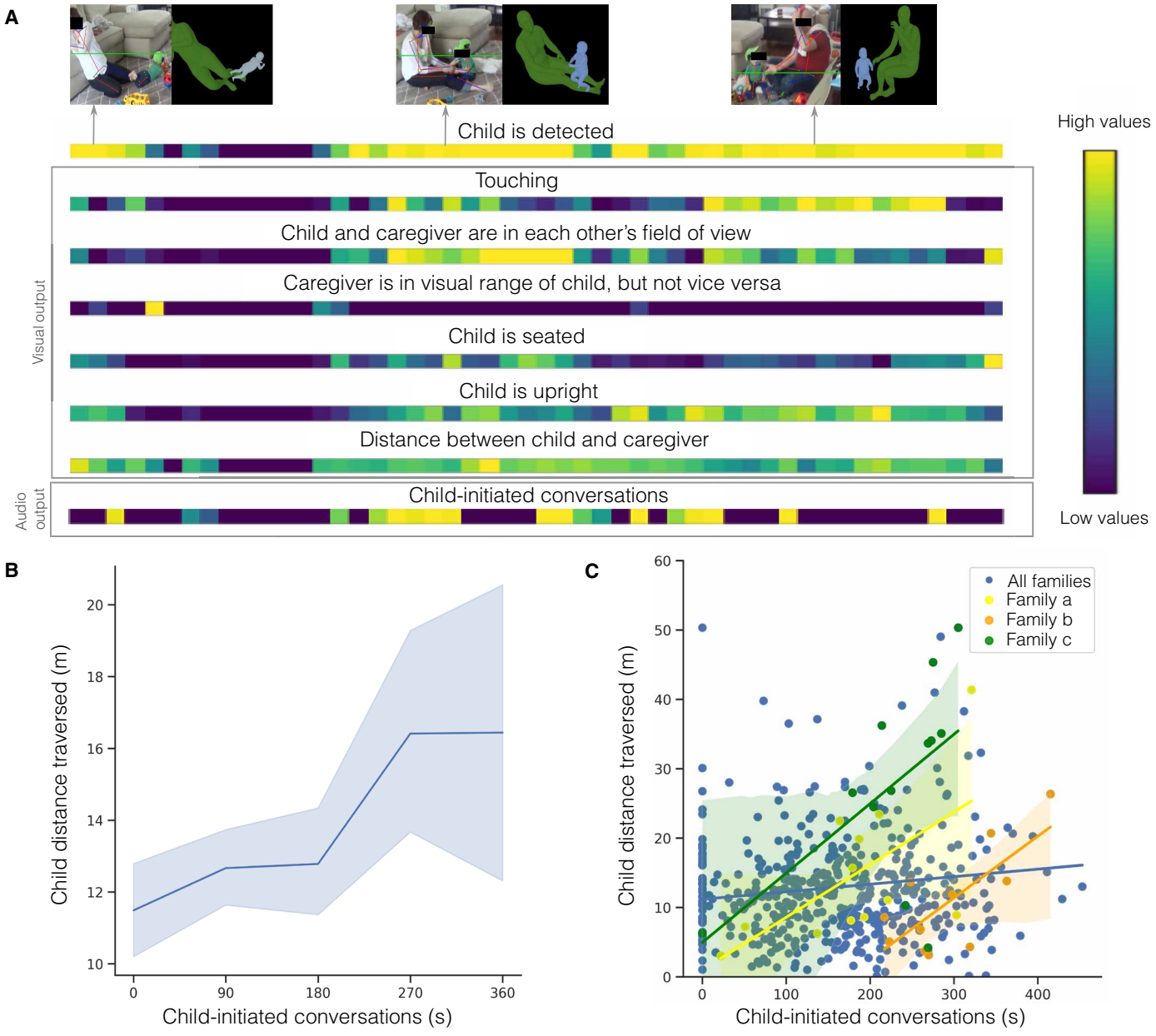

Fig. 4. Quantification of multimodal audiovisual interactions. (**A**) Timeline showing an example of interleaving visual and audio outputs produced by HARMONI on a single hour-long video. We color-code the per-second predictions aggregated by minute, so a high value means that the predictions within that minute are mostly positive. (**B**) Child distance traversed over recording periods, binned by child-initiated conversations in seconds. (**C**) The trend is particularly pronounced for certain families (family id is anonymized).

CMU Panoptic-Toddlers dataset

We evaluated HARMONI on the toddlers subset of the CMU Panoptic dataset (*25*). This dataset was captured in Pittsburgh, USA. It features a 2-year-old Asian toddler and is captured in a motion capture laboratory with high-resolution cameras. This dataset contains videos from dense viewpoints—given that some of these viewpoints contain extremely cropped humans, we retained those where both caregiver and child are fully visible. Our final test set contained 30,000 frames from a total of 40 videos (10 distinct scenes, with four videos per scene corresponding to different viewpoints). For five of the scenes, the original dataset included ground truth 2D keypoints that we used to evaluate our model output. For the other five scenes, we randomly sampled two frames from each video and human annotators manually labeled ground truth 2D keypoints in these frames.

On this evaluation dataset, HARMONI achieves 97% PCK for caregivers and 74.3% PCK for children. HARMONI achieves 84% and 93% accuracy on touch and visibility, with corresponding precision and recall of 93% and 93% for touch, and 96% and 90% for visibility.

DISCUSSION

Here, we have demonstrated the efficacy of HARMONI's 3D computer vision and audio analysis to automatically extract fine-grained, objective, and multimodal measures of human behavior from extensive

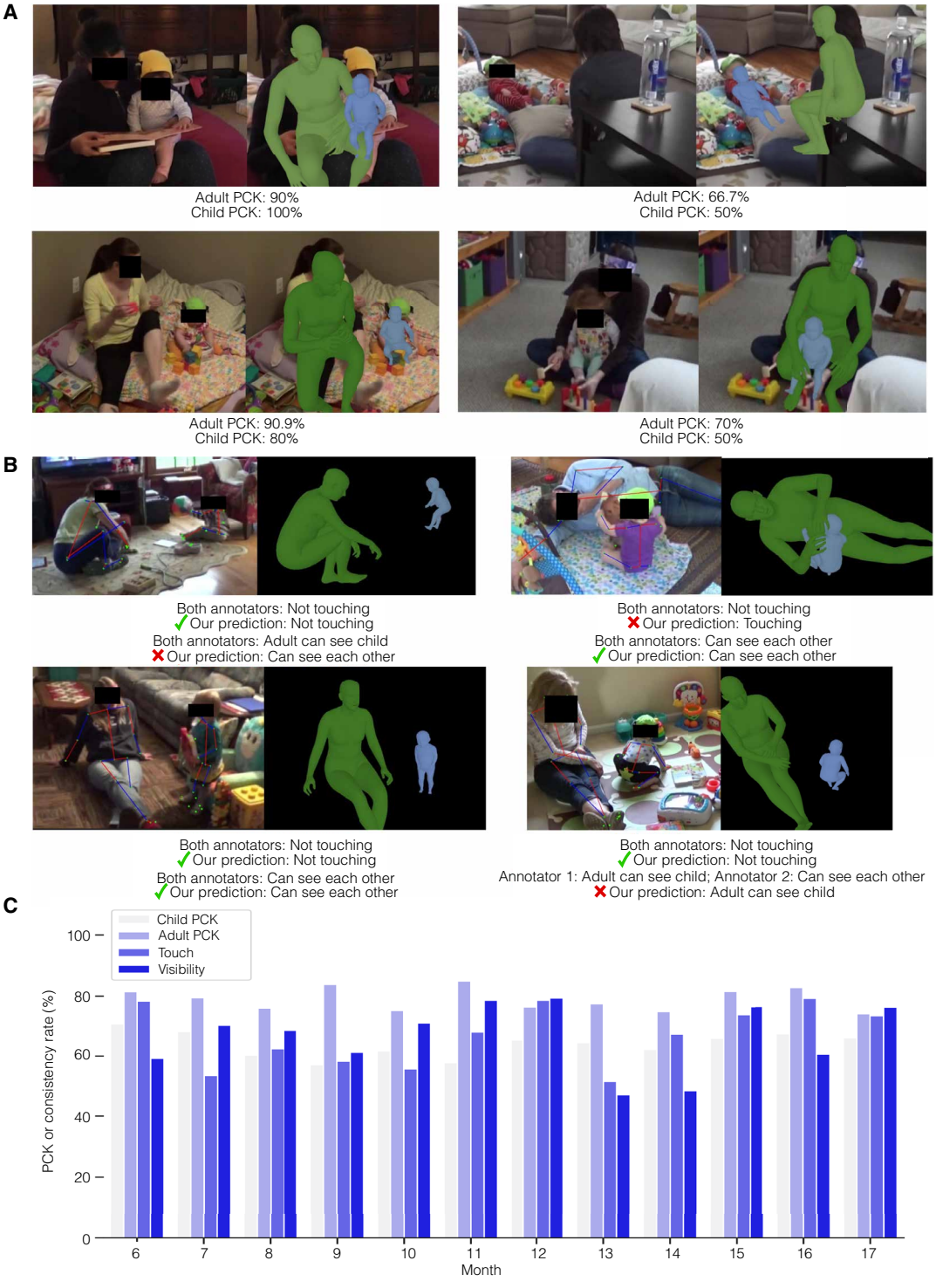

Fig. 5. Example output and consistency with human annotation. (**A**) Examples of 3D mesh estimations corresponding to different measured PCK values. (**B**) Examples of manual annotation versus HARMONI outputs, including failure modes. HARMONI often produces relatively reasonable outputs even when different from the manual annotation. (**C**) PCK and touch/visibility consistencies by developmental month. Months 13 and 14 in this dataset tend to contain video captures with a lot of occlusions, for example, where the caregiver stays in one corner of the room and is partially out of the frame. Therefore, the accuracy is slightly worse compared to other months.

single-view video data with minimal human intervention. By using HARMONI to analyze 500 hours of video from the SEEDLingS dataset, we showcased the capability to acquire subsecond resolution measures of touch, visibility, and pose, and to integrate these with audio analysis to quantify multimodal audiovisual interactions, using a server equipped with a consumer-grade graphical processing unit (GPU). Leveraging eight such GPUs in parallel enabled us to complete the analysis in 132 hours of server computation time, substantially reducing human labor costs. Assuming that highly trained human annotators can annotate a frame in 5 s, while HARMONI takes 0.25 s per frame, parallelizing HARMONI over eight GPUS can lead to 160:1 savings in time. This innovative approach has the potential to greatly expand the scope and scale of quantitative child development research by enabling the examination of a broad range of detailed objective measures and substantially increasing the number of caregiver-child dyads under study.

While many existing datasets of caregiver-child interactions, including SEEDLingS, are limited in terms of population diversity as well as sample size, the scalability of HARMONI's analysis capabilities facilitates the expansion of child development research across a larger and more diverse range of subjects, incorporating typically underrepresented populations. The model requires only standard video capture as input, easily obtainable from consumer-grade cameras or mobile phones, eliminating the need for additional instrumentation. Consequently, HARMONI can serve as a robust tool for generating scientific knowledge and testing hypotheses on a wider population across developmental stages, cultures, and other factors. This approach may enhance our understanding of patterns of human interaction and promote equity in child development research by incorporating greater diversity and enabling the assessment of human interaction as an outcome in experimental and pragmatic clinical trials.

Furthermore, our approach has the potential to catalyze new research endeavors across a range of fields where human behavior analysis is of value. Beyond the realm of caregiver-child dyads, HARMONI can be readily adapted to various settings involving different numbers of caregivers and children. This versatility renders it a valuable asset for investigating scientific hypotheses in domains such as behavioral pediatrics, other subfields of behavioral science, population health, and clinical care. In conclusion, the widespread adoption of HARMONI and similar methods could lead to a paradigm shift in the study of human behavior, expanding the boundaries of research and enabling the discovery of novel insights into the complexities of human interaction.

MATERIALS AND METHODS

SEEDLingS dataset

We used the SEEDLingS longitudinal dataset (*24*) to conduct developmental comparisons between children as approved by the Stanford University Institutional Review Board (IRB No. 56740) adhering to the data use guidelines from the data repository Databrary (*37*). The sample consisted of predominantly white, middle-class children ($N = 44$), with two identifying as mixed-race, one as Latino, and the remainder as white. Participants represented a range of incomes, with generally above-average maternal education levels. All children were exposed to English (>75%) in their homes and reported no visual or auditory impairments at birth. Videos were collected at a time of day that was convenient for parents within a week of their birth date each month (e.g., the 7-month visit was within a week of the child turning 7 months old). This was typically between 9 a.m. and 6 p.m. Monday to Friday, but within that range was up to the parents to schedule in a way that did not interfere with naptime. This included playtime, meals, etc. at the parents' discretion. The videos featured one to four camera viewpoints stitched together, with the number of camera angles dependent on each child's willingness to wear the camera headset and human-related error (Fig. 6).

Preprocessing the SEEDLingS videos involved downsampling to one frame per second and cropping to display only the third-person viewpoint. Downsampling aimed to expedite inference while maintaining granularity, while cropping sought to minimize noise and complexity. To accurately assess the model's generalizability to in-the-wild videos without multiple viewpoints, videos were cropped accordingly. The audio was stripped from the video, converted to a mono channel, and resampled at 16 kHz.

HARMONI visual analysis model

The components of the HARMONI visual analysis model are shown in Fig. 7. HARMONI first preprocesses the videos to extract tracklets of humans in the videos. Then, for each tracklet, a deep neural network predicts the human meshes from the tracklets, and postprocessing is applied to refine and smooth the meshes. Last, the caregiver-child features are extracted from the predicted meshes using predefined rules.

As SEEDLingS videos were obtained in a less structured setting and had more camera movement, we performed multiple preprocessing steps. Note that the ideal camera position provides a clear, unobstructed view of both individuals with minimal pitch angle (angle between the camera view and the floor). To preprocess the videos, PySceneDetect (*38*) is first used to segment videos into distinct scenes (i.e., shot detection). The transition of a scene is identified by thresholding the amount of change in image-level features (RGB histograms) across two frames. Scenes shorter than 60 s are discarded, as these typically result from the photographer moving the camera. For the SEEDLingS dataset, this eliminated 5.5% of the total frames. We subsequently use OpenPose (*39*) on the downsampled frames to acquire the 2D keypoints of each person in each frame. Tracklets are identified using PHALP (*40*), which is a tracking algorithm that is robust on closely interacting humans. This is because it takes not only the location of the person in the image but also the appearance of the person. Therefore, even when two persons are overlapping, tracks can be robustly differentiated based on the appearance of the person (derived from the masked region of the person). Note that although OpenPose has its limitations as a 2D keypoint detector, here we use it primarily as a human detector for tracklet association at this step. We noticed that OpenPose can reliably detect the existence of the child if the child is not being occluded; thus, the positive detections are solid for tracking.

As many SEEDLingS videos featured multiple household members, we annotated the dyad of interest with 10 bounding boxes and used a pretrained person reidentification network [TransReID (*41*)] to filter out irrelevant individuals such as other people in the family other than the main caregivers and child enrolled in the study. Note that child of interest in SEEDLingS wears a hat, which helps to identify the target child during this filtering stage. In general, the person of interest needs a marker only when easily confused with others. Filtering occurred in two stages, with cleaner tracklets and fewer identity switches resulting from the application of a lenient threshold during tracking. After obtaining tracklets, a second round of tracklet-level filtering is conducted using majority voting on individual detections in

Fig. 6. Example snapshot from one of the caregiver-child interaction videos in the 6-month partition. The three different camera angles stitched together in the video include two baby egocentric cameras and one third-person camera. Since HARMONI is designed to analyze third-person video, video feeds were cropped accordingly to keep only the third-person camera view.

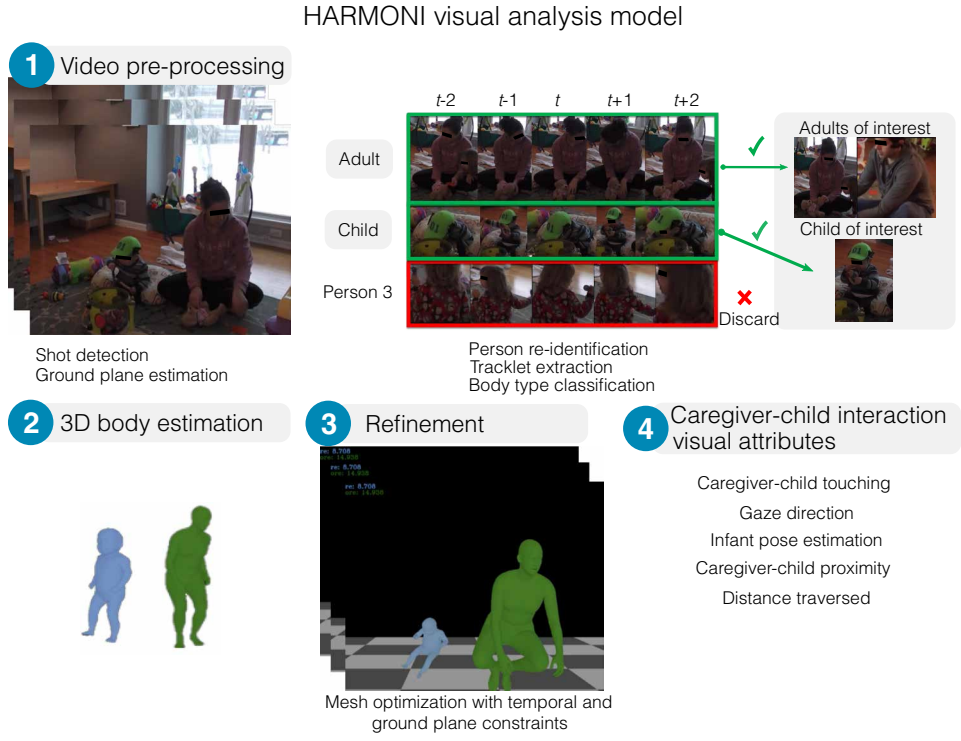

Fig. 7. System figure of the HARMONI visual analysis model. HARMONI processes the video in four steps: (1) preprocessing to identify the tracklets of the dyads of interest. (2) Estimation of 3D human meshes for each person of interest. (3) Refinement of 3D meshes via constrained optimization to improve the accuracy of the predictions. (4) Determining visual attributes from the estimated 3D meshes.

the tracklet. Finally, we performed a last round manual tracklet correction, filtering out noisy tracklets with numerous identity switches and those not involving the target dyad. This step required approximately 100 s per video. For CMU Panoptic-Toddler, we did not need these preprocessing steps as it is captured in a clean motion capture room with static cameras and there were no extra people in the room. In general, if the dataset contains only the dyads of interest, no additional marking (e.g., hats) or excessive filtering will be needed.

After obtaining clean tracklets of the dyads, we classify each tracklet to be either child or adult, which is necessary due to the different body models used for reconstructing children [SMIL (*33*)] and adults [SMPL (*34*)]. Children in SEEDLingS all wear a hat with

head-mounted cameras, so it is easy to train a customized body type classifier to determine whether the tracklet represents a child or an adult (i.e., caregiver). The body type classifier is trained using a semisupervised method [MixMatch (*42*)] with 200 labeled child images from SyRIP (*43*), 200 adult images randomly sampled from MPII (*36*), and 5000 unlabeled images from SEEDLingS. The classifier achieves 88% accuracy on a held-out test set with manually annotated bounding boxes and ground truth body type labels. For CMU Panoptic-Toddlers, we use Grounding DINO (*44*), which is capable of open-set object detection.

We then use human mesh recovery models to predict the human meshes. As many caregivers from the SEEDLingS videos were kneeling on the floor, we used DAPA (*45*), a deep neural network that we train adaptively to generalize well to SEEDLingS, through the use of its environment-targeted synthetic training data generation module, to produce initial estimates of each person. For CMU Panoptic-Toddlers, we use CLIFF (*46*) and modify it to handle child body models. The initial estimates comprise body model parameters and camera parameters, enabling the reconstruction of the 3D body mesh. Next, the body parameters are refined using an optimization routine (*47*), with several objectives: (i) ensuring consistency between the reconstructed 3D human and the 2D keypoint predictions from OpenPose; (ii) maintaining smoothness of the 3D keypoints and their 2D projections within a moving time window; and (iii) constraining child's position in 3D space to be above the ground plane, as estimated using a panoptic segmentation model (*48*) and depth estimation model (*49*) from five randomly sampled frames per scene. Note that when we refine the meshes using OpenPose's 2D keypoint detections, we weigh the per-keypoint loss terms by the predicted confidence scores to minimize the effect of spurious keypoint predictions, as we observe that high-confidence predicted keypoints tend to be reliable. Finally, the One Euro Filter (*50*) is applied to the predictions on each tracklet to remove jittering across frames.

Speed

To improve efficiency of video processing, we group preprocessed data into batches (16 persons within a track) and run HARMONI on batched data. On a single NVIDIA TITAN V GPU, it takes 0.2 s per batch to perform 3D body estimation, and 4 s per batch to run the refinement optimization routine for 10 iterations. When visualizing 3D meshes is desired, instead of simply computing quantitative features, rendering 3D humans takes approximately 1 s per video frame.

Determining dyadic visibility and touch

HARMONI determines caregiver-child visibility and touch labels a rule-based way from the estimated 2D and 3D positions of caregivers and children. Specifically, person A's visibility to person B was defined by whether person A's head was within person B's 180-degree field of view. HARMONI approximates the field of view by the volume separated by the plane passing through person B's left/right ear and neck joints.

Touching is defined as the state in which any part of the caregiver's body comes into contact with the child's body. When HARMONI detects both caregiver and child in a frame, it determines touching using two thresholds. If the minimum 2D distance (ignoring the z axis) between the sets of keypoints falls below a predefined 2D threshold (3% of image height), we compute the minimum 3D distance between keypoints from caregiver and child. If the 3D distance is also below the established 3D threshold (25 cm), we consider the dyad as touching. If either check fails, we determine that the dyad is not touching. The 2D and 3D tiered approach is used to mitigate reduced reliability in depth calculation for keypoint prediction.

To assess consistency of HARMONI visibility and touch outputs with manual annotation, two annotators are asked to annotate each frame in the validation set (1500 frames) individually, and consistency between HARMONI and human annotators (Fig. 5) is reported on frames where two annotators agree with each other, which is 92% of the total frames. Annotators followed the same definition for visibility and labeled each image as "child is in the visual range of caregiver, but not vice versa," "caregiver is in the visual range of child, but not vice versa," "caregiver and child are in each other's visual range," "caregiver and child are not in each other's visual range," or "N/A" (when the image did not contain both child and caregiver). For touching, annotators annotated each image as "touching," "not touching," or "N/A" (when the image did not contain both child and caregiver).

Audio model

The audio model, working with mono-channel, 16-kHz audio tracks, comprises two open-source models: VBHMM x-vectors Diarization (VBx) (*51*) and Automatic LInguistic Unit Count Estimator (ALICE) (*9*), with ALICE using Voice Type Classifier (VTC) (*52*) for broad-class speaker diarization (Fig. 8). The speaker diarization label indicates when speakers of the following classes are speaking: adult male, adult female, key child, or other child.

The VTC model, an open-source alternative to LENA (*10*), shows a notable performance improvement of 10.6 F-measure averaged across the five classes on the ACLEW-Random test set (*52*), as compared to LENA. As depicted in Fig. 8, the VTC model pushes fixed-length, overlapping subsequences of preprocessed audio through SincNet (*53*) to generate descriptive, low-level audio representations. These representations are fed through a stack of bidirectional Long Short-Term Memory (LSTM) models (*54*) and fully connected layers to produce diarization output with voice type classification. The VTC model is pretrained on a large corpora of child-centered audio recordings covering various environments, conditions, and languages. We use the speech labels of the diarization output file from the VTC model to create a speech activity prediction file, which is fed into the VBx model along with the original preprocessed audio file.

The VBx model, winner of the DIHARD II challenge (*51*), first extracts X-vectors (*55*, *56*) using the Kaldi toolkit (*57*), followed by agglomerative hierarchical clustering (AHC) with two distinct interpolated PLDA (*58*) models trained on separate datasets [VoxCeleb (*59*) and the DIHARD development set]. These AHC clusterings are used for initial assignment of X-vectors to speaker clusters. Iterative variational inference on a Bayesian Hidden Markov Model (BHMM) (*55*) is then applied to the X-vectors to generate final clusterings and outputs. Information from both diarization outputs is leveraged to generate an individual speaker diarization with broad-class classifications. Subsequently, we calculate CTCs by counting the turns for each speaker from the diarization results. Finally, we use ALICE to extract the child vocalization count, and adult word, syllable, and phoneme counts. The open-source audio model (ALICE) we are using was optimized for a multilingual corpus and was validated

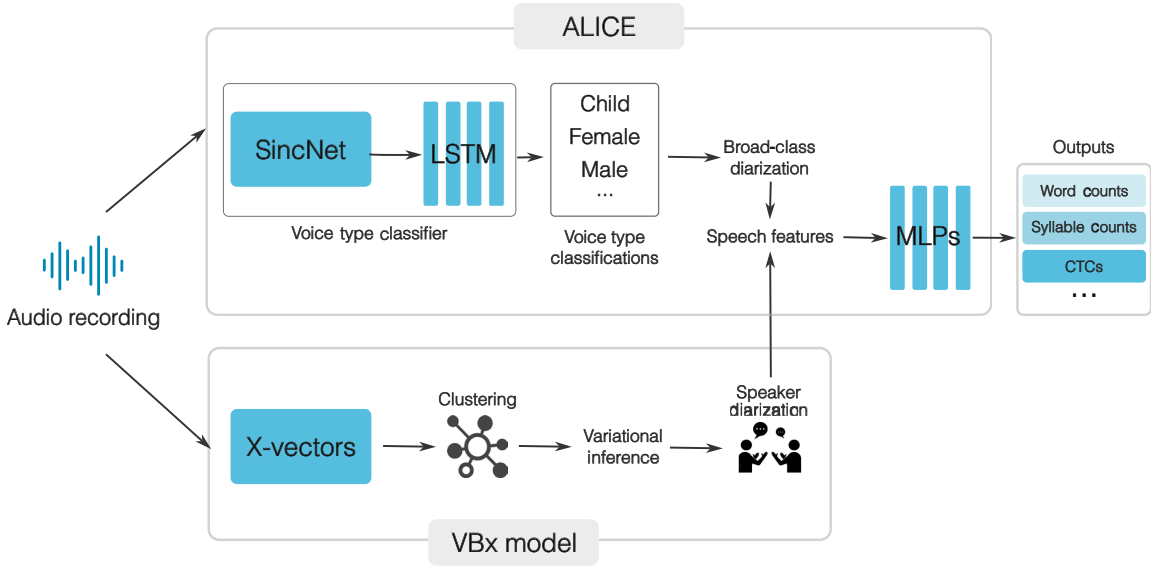

Fig. 8. System diagram of the audio analysis model. The audio processing side of the tool is composed of two open-source models, VBx and ALICE. ALICE uses VTC for broad-class speaker diarization.

through cross-language generalization experiments. On the other hand, LENA is optimized for only American English.

To validate the accuracy of the speaker diarization model in the domain of at-home child speech, we randomly sampled 309-min-long audio clips from the SEEDLingS videos. The audio clips were transcribed by a paid, accredited third-party transcription service (https://www.transcribeme.com/). We report diarization error rate (DER), the standard metric to assess the accuracy of a speaker diarization system. DER is computed by comparing the reference speaker labels (ground truth) with the diarization system's output labels and calculating the percentage of errors. These errors can include speaker false alarms (when a speaker is incorrectly detected) and missed detections (when a speaker is not detected). A lower DER indicates better performance, as it signifies a closer alignment between the diarization output and the reference speaker labels.

The DER on the sampled audio clips is 32.3%. Albeit slightly higher than the reported 27.11% of the same model on the DIHARD evaluation set, it is important to note that the evaluation set of DIHARD includes multiple domains much easier than an in-the-wild child speech setting. Error analysis suggested that most misclassified speech comes from challenging environments with music or television playing in the background. On the CTC predictions, our model achieves a Pearson coefficient of 22.3% as compared to LENA's CTC coefficient of 36% (*60*). Pearson's coefficient for adult word count predictions is 72.8%. Last, we see a 56.2% median absolute relative error on the adult word count predictions, which is comparable to LENA's performance on other English language datasets.

Statistical analysis

We assessed changes in each outcome variable across development through growth model analyses. First, to test whether age was a statistically significant predictor of each outcome variable, we used mixed-effects models to test the statistical significance of the change in each variable over time. Then, we added quadratic and cubic terms of age to test whether any change over time was nonlinear. We have also provided spline models in the Supplementary Materials. We elect to keep the mixed-effects models in the main text due to their interpretability. However, we direct those interested in exploring additional models to the Supplementary Materials. These models offer a wider range of options when considering the balance between interpretability and the degree of freedom.

We consider three types of mixed-effects models: those with random intercepts, random slopes, or both. (Formulas for the models are in the Supplementary Materials.) Mixed-effects model with random intercepts allows each family to have its own baseline value of the dependent variable. In other words, it accounts for variability in the starting points of the variable across families. This is the simplest random effect structure and assumes that the effect of time (slope) is constant across families. Mixed-effects model with random slopes allows the effect of time on the dependent variable to vary across families, which means that each family can have a different rate of change over time. Mixed-effects model with both random intercepts and random slopes allows for the most flexibility in the model. It accounts for variability in both the baseline value of the dependent variable and the rate of change over time across families.

We used AIC (Akaike information criterion) and BIC (Bayesian information criterion) for model selection. For three outcome variables (touch, proportion of time the child spent in upright position, and child distance traversed) that showed statistically significant trend across child development, the best models were with random slopes. For caregiver distance traversed, the best model was with random intercepts. In addition, for all outcome variables, the linear model provides the best fit for the data according to AIC and BIC, meaning the relationship between these outcome variables and child development can be best described as linear. For each model, we reported the standardized betas along with their corresponding CIs, as well as *P* values, which provide information about the statistical significance of the predictor variables in our selected models. Last, to measure the relationship between child distance traversed

and child-initiated conversation duration, we calculated the correlation between these two variables. In addition, we calculated partial correlation between child distance traversed and child-initiated conversation duration, holding child age constant.

Supplementary Materials

This PDF file includes:

Supplementary Text
Fig. S1
Tables S1 to S5
References

REFERENCES AND NOTES

1. C. A. Nelson, K. M. Thomas, M. D. H. de Haan, *Neuroscience of Cognitive Development: The Role of Experience and the Developing Brain* (John Wiley & Sons, 2012).
2. D. G. Gee, Caregiving influences on emotional learning and regulation: Applying a sensitive period model. *Curr. Opin. Behav. Sci.* **36**, 177–184 (2020).
3. K. de Barbaro, C. M. Fausey, Ten lessons about infants' everyday experiences. *Curr. Dir. Psychol. Sci.* **31**, 28–33 (2022).
4. B. Long, S. Goodin, G. Kachergis, V. A. Marchman, S. F. Radwan, R. Z. Sparks, V. Xiang, C. Zhuang, O. Hsu, B. Newman, D. L. K. Yamins, M. C. Frank, The BabyView camera: Designing a new head-mounted camera to capture children's early social and visual environments. *Behav. Res. Methods* **56**, 3523–3534 (2024).
5. N. Kojovic, S. Natraj, S. P. Mohanty, T. Maillart, M. Schaer, Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Sci. Rep.* **11**, 15069 (2021).
6. V. Marchi, A. Hakala, A. Knight, F. D'Acunto, M. L. Scattoni, A. Guzzetta, S. Vanhatalo, Automated pose estimation captures key aspects of General Movements at eight to 17 weeks from conventional videos. *Acta Paediatr.* **108**, 1817–1824 (2019).
7. J. Li, A. Bhat, R. Barmaki, Dyadic movement synchrony estimation under privacy-preserving conditions, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 762–769.
8. S. Tafasca, A. Gupta, J. Odobez, ChildPlay: A new benchmark for understanding children's gaze behaviour, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2023), pp. 20878–20889.
9. O. Räsänen, S. Seshadri, M. Lavechin, A. Cristia, M. Casillas, ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behav. Res. Methods* **53**, 818–835 (2021).
10. D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, J. Hanse, Signal processing for young child speech language development, in *First Workshop on Child, Computer and Interaction. WOCCI.* (2008).
11. C. Jongerius, T. Callemein, T. Goedemé, K. Van Beeck, J. A. Romijn, E. M. A. Smets, M. A. Hillen, Eye-tracking glasses in face-to-face interactions: Manual versus automated assessment of areas-of-interest. *Behav. Res. Methods* **53**, 2037–2048 (2021).
12. T. Callemein, K. Van Beeck, G. Brône, T. Goedemé, Automated analysis of eye-tracker-based human-human interaction studies, in *Information Science and Applications 2018* (Springer Singapore, 2019), pp. 499–509.
13. D. Y. Isaev, M. Sabatos-DeVito, J. M. Di Martino, K. Carpenter, R. Aiello, S. Compton, N. Davis, L. Franz, C. Sullivan, G. Dawson, G. Sapiro, Computer vision analysis of caregiver–child interactions in children with neurodevelopmental disorders: A preliminary report. *J. Autism Dev. Disord.* **54**, 2286–2297 (2024).
14. K. E. Adolph, B. I. Bertenthal, S. M. Boker, E. C. Goldfield, E. J. Gibson, Learning in the development of infant locomotion. *Monogr. Soc. Res. Child Dev.* **62**, 1–140 (1997).
15. W. Barnett, C. L. Hansen, L. G. Bailes, K. L. Humphreys, Caregiver-child proximity as a dimension of early experience. *Dev. Psychopathol.* **34**, 647–665 (2022).
16. P. L. Blackwell, The influence of touch on child development. *Infants Young Child.* **13**, 25–39 (2000).
17. G. Pusiol, L. Soriano, M. C. Frank, L. Fei-Fei, Discovering the signatures of joint attention in child-caregiver interaction, in *Proceedings of the Annual Meeting of the Cognitive Science Society. CogSci. (Vol. 36, No. 36)* (2014).
18. K. E. Adolph, C. S. Tamis-LeMonda, S. Ishak, L. B. Karasik, S. A. Lobo, Locomotor experience and use of social information are posture specific. *Dev. Psychol.* **44**, 1705–1714 (2008).
19. R. J. Gerber, T. Wilks, C. Erdie-Lalena, Developmental milestones: Motor development. *Pediatr. Rev.* **31**, 267–277 (2010).
20. V. C. Salo, L. S. King, I. H. Gotlib, K. L. Humphreys, Infants who experience more adult-initiated conversations have better expressive language in toddlerhood. *Infancy* **27**, 916–936 (2022).
21. L. Bloom, C. Margulis, E. Tinker, N. Fujita, Early conversations and word learning: Contributions from child and adult. *Child Dev.* **67**, 3154 (1996).
22. R. R. Romeo, J. A. Leonard, S. T. Robinson, M. R. West, A. P. Mackey, M. L. Rowe, J. D. E. Gabrieli, Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychol. Sci.* **29**, 700–710 (2018).
23. J. Li, V. Chheang, P. Kullu, E. Brignac, Z. Guo, K. E. Barner, A. Bhat, R. L. Barmaki, MMASD: A multimodal dataset for autism intervention analysis. arXiv:2306.08243 [cs.CV] (2023).
24. E. Bergelson, A. Amatuni, S. Dailey, S. Koorathota, S. Tor, Day by day, hour by hour: Naturalistic language input to infants. *Dev. Sci.* **22**, e12715 (2019).
25. H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, Y. Sheikh, Panoptic Studio: A massively multiview system for social interaction capture. arXiv:1612.03153 [cs.CV] (2015).
26. A. C. Huston, S. Rosenkrantz Aronson, Mothers' time with infant and time in employment as predictors of mother-child relationships and children's early development. *Child Dev.* **76**, 467–482 (2005).
27. J. T. Suvilehto, E. Glerean, R. I. M. Dunbar, R. Hari, L. Nummenmaa, Topography of social touching depends on emotional bonds between humans. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13811–13816 (2015).
28. A. Schmidt, A. C. Kramer, A. Brose, F. Schmiedek, A. B. Neubauer, Distance learning, parent-child interactions, and affective well-being of parents and children during the COVID-19 pandemic: A daily diary study. *Dev. Psychol.* **57**, 1719–1734 (2021).
29. H. Uzun, N. H. Karaca, Ş. Metin, Assesment of parent-child relationship in Covid-19 pandemic. *Child. Youth Serv. Rev.* **120**, 105748 (2021).
30. R. M. Fasano, L. K. Perry, Y. Zhang, L. Vitale, J. Wang, C. Song, D. S. Messinger, A granular perspective on inclusion: Objectively measured interactions of preschoolers with and without autism. *Autism Res.* **14**, 1658–1669 (2021).
31. V. C. Salo, P. Pannuto, W. Hedgecock, A. Biri, D. A. Russo, H. A. Piersiak, K. L. Humphreys, Measuring naturalistic proximity as a window into caregiver-child interaction patterns. *Behav. Res. Methods* **54**, 1580–1594 (2022).
32. D. Han, N. Aziere, T. Wang, O. Ossmy, A. Krishna, H. Wang, R. Shen, S. Todorovic, K. Adolph, Infants' developing environment: Integration of computer vision and human annotation to quantify where infants go, what they touch, and what they see, in *2024 IEEE International Conference on Development and Learning (ICDL)* (IEEE, 2024), vol. 25, pp. 1–8.
33. N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Müller-Felber, A. Sebastian Schroeder, Learning an infant body model from RGB-D data for accurate full body motion analysis, in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018* (Springer International Publishing, 2018), pp. 792–800.
34. F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. J. Black, Keep It SMPL: Automatic estimation of 3D human pose and shape from a single image, in *Computer Vision—ECCV 2016* (Springer International Publishing, 2016), pp. 561–578.
35. C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1325–1339 (2014).
36. M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: New benchmark and state of the art analysis, in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2014), pp. 3686–3693.
37. D. A. Simon, A. S. Gordon, L. Steiger, R. O. Gilmore, Databrary: Enabling sharing and reuse of research video, in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (Association for Computing Machinery, 2015), pp. 279–280.
38. B. Castellano, Pyscenedetect (2020).
39. Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 7291–7299.
40. J. Rajasegaran, G. Pavlakos, A. Kanazawa, J. Malik, Tracking people by predicting 3D appearance, location & pose. arXiv:2112.04477 [cs.CV] (2021).
41. S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, TransReID: Transformer-based object re-identification, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), pp. 15013–15022.
42. D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C. Raffel, MixMatch: A holistic approach to semi-supervised learning. arXiv:1905.02249 [cs.LG] (2019).
43. X. Huang, N. Fu, S. Liu, S. Ostadabbas, Invariant representation learning for infant pose estimation with small data, in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (IEEE, 2021), pp. 1–8.
44. S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, L. Zhang, Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv:2303.05499 [cs.CV] (2023).
45. Z. Weng, K.-C. Wang, A. Kanazawa, S. Yeung, Domain adaptive 3D pose augmentation for in-the-wild human mesh recovery, in *2022 International Conference on 3D Vision (3DV)* (IEEE, 2022), pp. 261–270.

46. Z. Li, J. Liu, Z. Zhang, S. Xu, Y. Yan, CLIFF: Carrying location information in full frames into human pose and shape estimation, in *Computer Vision—ECCV 2022* (Springer Nature Switzerland, 2022), pp. 590–606.
47. Z. Weng, S. Yeung, Holistic 3d human and scene mesh estimation from single view images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2021), pp. 334–343.
48. B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, L.-C. Chen, Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. arXiv:1911.10194 [cs.CV] (2019).
49. R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1623–1637 (2022).
50. G. Casiez, N. Roussel, D. Vogel, 1 € filter: A simple speed-based low-pass filter for noisy input in interactive systems, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2012), pp. 2527–2530.
51. F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotný, H. Zeinali, J. Rohdin, But system for the second dihard speech diarization challenge, in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), pp. 6529–6533.
52. M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, A. Cristia, An open-source voice type classifier for child-centered daylong recordings. arXiv:2005.12656 [eess.AS] (2020).
53. M. Ravanelli, Y. Bengio, Speaker recognition from raw waveform with SincNet. arXiv:1808.00158 [eess.AS] (2018).
54. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
55. M. Diez, L. Burget, S. Wang, J. Rohdin, J. Černocký, Bayesian HMM based x-vector clustering for speaker diarization. *Interspeech*, 346–350 (2019).
56. G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, S. Khudanpur, Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge, *Interspeech* (2018).
57. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, K. Veselý, The Kaldi speech recognition toolkit, in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (IEEE Signal Processing Society, 2011).
58. P. Kenny, Bayesian speaker verification with heavy tailed priors. *Proc. Odyssey*, 14. (2010).
59. A. Nagrani, J. S. Chung, A. Zisserman, VoxCeleb: A large-scale speaker identification dataset. arXiv:1706.08612 [cs.SD] (2017).
60. A. Cristia, M. Lavechin, C. Scaff, M. Soderstrom, C. Rowland, O. Räsänen, J. Bunce, E. Bergelson, A thorough evaluation of the Language Environment Analysis (LENA) system. *Behav. Res. Methods* **53**, 467–486 (2021).
61. D. Han, N. Aziere, T. Wang, O. Ossmy, A. Krishna, H. Wang, R. Shen, S. Todorovic, K. Adolph, Infants' developing environment: Integration of computer vision and human annotation to quantify where Infants go, what they touch, and what they see, in *IEEE International Conference on Development and Learning (ICDL)* (IEEE, 2024), pp. 1–8.
62. C. Suarez-Rivera, J. L. Schatz, O. Herzberg, C. S. Tamis-LeMonda, Joint engagement in the home environment is frequent, multimodal, timely, and structured. *Inf. Dent.* **27**, 232–254 (2022).
63. K. S. Kretch, J. M. Franchak, K. E. Adolph, Crawling and walking infants see the world differently. *Child Dev.* **85**, 1503–1518 (2014).

Acknowledgments: We thank K. Ochoa and K. Ramirez for their tireless efforts in manually coding videos for child and caregiver poses and touch and visibility labels for SEEDLingS. We thank A. Brown and N. Hassan for coding the labels for CMU Panoptic. We are grateful to L. Malachowski for statistical consultation and support. **Funding:** We appreciate the generous support of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). S.Y.-L. is a Chan Zuckerberg Biohub–San Francisco Investigator. K.L.H. is supported by the National Science Foundation (2042285), National Institute of Mental Health (R01MH129634), and Jacobs Foundation (2017-1261-05; 2016-1251-07). L.M.S. is supported by the National Institutes of Health (1R01MD013844-01A1, 5R01CA20458504, 5R01AG04979103) and Patient-Centered Research Institute (AD-2018C1-11238). SEEDLingS data were collected by E.B. with support of the National Institutes of Health (DP5-OD019812). L.B.-S. is supported by the Fulbright U.S. Student Program, sponsored by the U.S. Department of State and Fulbright Colombia. **Author contributions:** Conceptualization: K.L.H., L.M.S., S.Y.-L., C.H., A.K., E.B., A.M., and M.X. Methodology: Z.We., L.B.-S., Z.Wa., C.H., M.X., S.Y.-L., K.L.H., L.M.S., A.K., E.B., A.M., and N.M. Investigation: Z.We., L.B.-S., M.X., E.B., A.M., and N.M. Software: Z.We., L.B.-S., C.H., M.X., Z.Wa., and N.M. Visualization: M.X., L.B.-S., Z.We., and L.M.S. Validation: C.H., Z.We., L.M.S., E.B., and N.M. Formal analysis: M.X., C.H., Z.We., K.L.H., N.M., and L.B.-S. Resources: E.B., L.M.S., S.Y.-L., and A.M. Supervision: K.L.H., L.M.S., S.Y.-L., and A.M. Writing—original draft: M.X., Z.We., C.H., L.M.S., and A.M. Writing—review and editing: A.K., L.M.S., E.B., A.M., Z.We., M.X., L.B.-S., and S.Y.-L. Data curation: C.H., L.M.S., E.B., and M.X. Funding acquisition: L.M.S., A.M., and S.Y.-L. Project administration: L.M.S., K.L.H., and S.Y.-L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The data used in the paper are publicly available. The code and annotations needed to reproduce the results are available at https://zenodo.org/records/14630497. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 8 May 2024
Accepted 15 January 2025
Published 19 February 2025
10.1126/sciadv.adp4422