# SQUAD: A Unified Framework for 3D Dyadic Scene Understanding with Applications to Visual Focus of Attention and Object Use

Berfu Karaca
Utrecht University
Utrecht, The Netherlands
b.karaca@uu.nl

Niilo V. Valtakari
Utrecht University
Utrecht, The Netherlands
n.v.valtakari@uu.nl

Jaap J. A. Denissen
Utrecht University
Utrecht, The Netherlands
j.a.denissen@uu.nl

Albert Ali Salah
Utrecht University
Utrecht, The Netherlands
a.a.salah@uu.nl

Sonja M. C. de Zwarte
Utrecht University
Utrecht, The Netherlands
s.m.c.dezwarte@uu.nl

Ronald Poppe
Utrecht University
Utrecht, The Netherlands
r.w.poppe@uu.nl

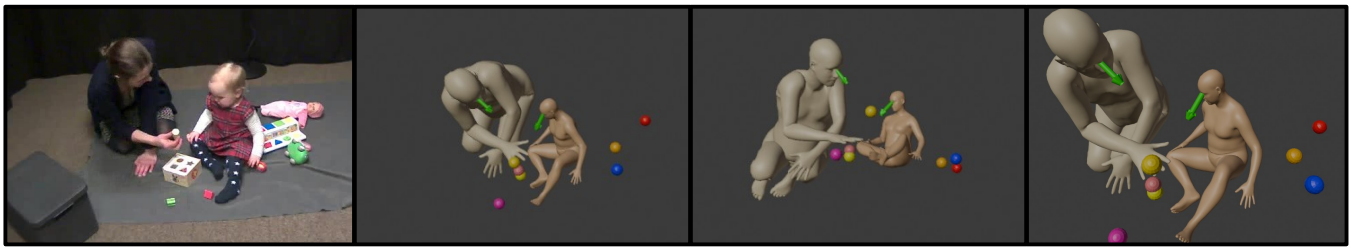

**Figure 1: Outline of our work. Multi-view camera recordings are processed with SQUAD to estimate 3D body pose and object locations. The 3D scene reconstruction can be visualized from any perspective, and facilitates quantitative, spatial analyses. We demonstrate the merits of analysis in 3D with applications of visual focus of attention (VFOA) and object use estimation.**

## Abstract

Social dyads are rich in nonverbal behavior. Particularly in parent-child interactions involving toys, the analysis of coordinated behavior can be used to assess the child's development, and the relation between the two interactants. Estimating nonverbal cues in such naturalistic settings using automated, camera-based methods remains challenging. The reliance on a single-camera view reduces the robustness and utility of measurements, as it is prone to occlusions, motion of the interactants and the camera, and insufficient resolution. To address these challenges, we focus on capturing interactions in 3D. We introduce SQUAD: Scene QUantification for the Analysis of Dyads, a 3D scene reconstruction tool focusing on naturalistic dyadic interactions recorded from multiple views. SQUAD estimates the 3D positions of body keypoints and object centers from continuous interactions without assuming calibrated cameras. We demonstrate the merits of 3D analysis using SQUAD with the automated analysis of visual focus of attention and object use, for which we develop simple modules. We evaluate our approach on challenging, free-play parent-child interactions involving 10-month-old infants and age-appropriate toys. Our experiments reveal that 3D analysis of dyads can be robust and that SQUAD can flexibly accommodate multimodal analyses and allows for the extension to a range of other nonverbal, individual or coordinated, behaviors.

## CCS Concepts

• **Applied computing** → **Psychology**; • **Computing methodologies** → **Activity recognition and understanding**.

## Keywords

visual focus of attention, gaze, touch, object interaction, parent-child interaction, free-play

**ACM Reference Format:**
Berfu Karaca, Niilo V. Valtakari, Jaap J. A. Denissen, Albert Ali Salah, Sonja M. C. de Zwarte, and Ronald Poppe. 2026. SQUAD: A Unified Framework for 3D Dyadic Scene Understanding with Applications to Visual Focus of Attention and Object Use. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3776574.3831168

## 1 Introduction

Parent-child interactions play a central role in developmental theories [16, 18] and video-recorded parent-child interactions are widely used to track child development [36, 40].

Traditionally, researchers manually code these interactions, which is time-consuming and requires extensive training, hindering large-scale quantitative analysis [12, 36]. There is a need for automated methods to reduce the manual effort, as well as the subjectivity of


*ICMI '26, Napoli, Italy*

ACM ISBN 979-8-4007-2318-6/2026/10
https://doi.org/10.1145/3776574.3831168

the analysis [19]. Despite recent advancements, automated behavior analysis in realistic settings remains challenging. Videos are often recorded at low resolution, and they include sudden movements of the interactants as well as camera movements, and partial or full occlusions. When a single view is used, depth ambiguities complicate the analysis. These difficulties are particularly evident in close proximity interactions, such as free-play scenarios [20].

**Automated 3D interaction analysis**. Existing automated approaches often operate in 2D, missing the opportunity to robustly reveal spatial relations between people and objects. When multiple views are used, proper calibration is needed to leverage the partially complementary information in each view. Yet, cameras are often controlled to keep the interactants in view or to focus on interesting parts of an interaction. Consequently, calibration information is commonly missing.

To address these challenges, we introduce SQUAD: Scene QUantification for the Analysis of Dyads, a multi-view 3D scene reconstruction tool for naturalistic dyadic interactions. SQUAD extends work by Huang et al. [17] who introduced a framework for multi-person multi-view mesh recovery from uncalibrated cameras by combining physics-geometry consistency with a learned latent motion prior. We extend this approach in several key dimensions. First, we add a scheduling step to deal with camera movement and missing person detections. Second, we provide support for the 3D recovery of object positions, to be able to analyze person-object interactions. Third, we track interactants to allow for robust behavior analysis over time. These extensions allow SQUAD to process continuous videos and consistently link behaviors to persons. SQUAD reconstructs key features needed for higher-level analysis, including 3D body keypoints and object coordinates.

**Visual focus of attention and object use**. We demonstrate the value of SQUAD on two tasks respectively: visual focus of attention (VFOA) and object use estimation. VFOA provides information regarding the attended object or person in the scene. VFOA and object use often co-occur, as individuals tend to fixate during object interaction [22]. We also combine these two tasks to arrive at multimodal analysis of VFOA, revealing the flexibility of SQUAD for the analysis of more complex behaviors. Our contributions are:

(1) We introduce SQUAD, a tool to reconstruct dyadic interactions recorded using multiple uncalibrated cameras in 3D. SQUAD estimates both 3D body poses and object locations.
(2) We extend SQUAD for the estimation of VFOA and object use, and then combine the two tasks.
(3) We demonstrate analysis of VFOA and object use from 3D reconstructions of naturalistic free-play parent-child interactions and show the contribution of multimodal analysis.[1]

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 details the data used in our experiments. Sections 4 introduces SQUAD and our applications of VFOA and object use estimation tasks. Our evaluation is given in Section 5, followed by a discussion and conclusions.

## 2 Related Work

We discuss the role of VFOA and object use in interactions before detailing methods to automatically estimate them from video.

[1]Code is available on: https://github.com/berfukaraca/SQUAD

### 2.1 Visual Focus of Attention and Object Use

VFOA offers a lens to interpret social interactions from a functional perspective, including gathering information (i.e., tracking attention), regulating interaction dynamics (i.e., conversational turns), and expressing internal states (i.e., interpersonal intimacy) [1, 21].

In parent-child interactions, VFOA is commonly linked to early social attention cues such as joint attention and mutual gaze, which are foundational for children's early social, emotional, and cognitive development [9, 27]. In particular, individual differences in joint attention skills have been associated with later language development and social emotional outcomes such as attachment [7, 39]. Furthermore, poor joint attention skills have been considered as predictive for developmental difficulties, including Autism Spectrum Disorder [5].

Similarly, objects play a central role in shaping the context of social interactions. In parent-child interactions, their importance is also emphasized in studies on shared attention [8], as well as in developmental research, particularly in relation to difficulties such as disengaging from objects [33].

### 2.2 Automated Analysis of VFOA

While eye-based gaze estimation provides a direct cue for attention [3, 6], it requires frontal face images, which are not always available in naturalistic recordings, considering limited image resolution, significant head movements, and occlusions [6, 41]. When eyes are not clearly visible, a common alternative is to use head pose as a proxy for eye gaze, supported by the coordinated movements between the eyes and head [11]. Nevertheless, head pose alone is often insufficient, since in realistic scenes, the same head pose can correspond to different gaze targets, difficult to fix due to missing eye information [34].

Traditional approaches of VFOA estimation from cameras include probabilistic models [29], wearable eye trackers [30], and gaze-likelihood heatmaps [23]. Yet, performance remains limited in naturalistic settings with frequent occlusions and restricted direct eye observations. To improve robustness and to reduce noise in VFOA estimations, several works incorporate contextual cues such as the speaking status of the interactants [34, 35]. More recent approaches employ deep learning models, typically requiring large annotated datasets [2, 4]. Furthermore, many automated methods operate in 2D, lacking 3D reasoning due to missing depth information [24, 38]. Recent work such as ChildPlay [37] incorporates depth cues to support gaze target prediction from visual input. In this context, multi-view recordings offer a promising alternative by enabling 3D reasoning through the geometric relationships between different views [26]. However, these methods also impose strong requirements like the availability of frontal face images with direct eye observations, limiting their applicability in naturalistic experimental settings [4, 15]. In a recent work, Weng et al. [43] introduced 3D analysis method HARMONI for quantitative measurements in parent-child interactions. The tool operates on a single camera view, enabling egocentric analysis and flexible recording configurations but limiting its applicability in settings with occlusions. Moreover, its estimations are limited to coarse gaze direction rather than explicit VFOA predictions.

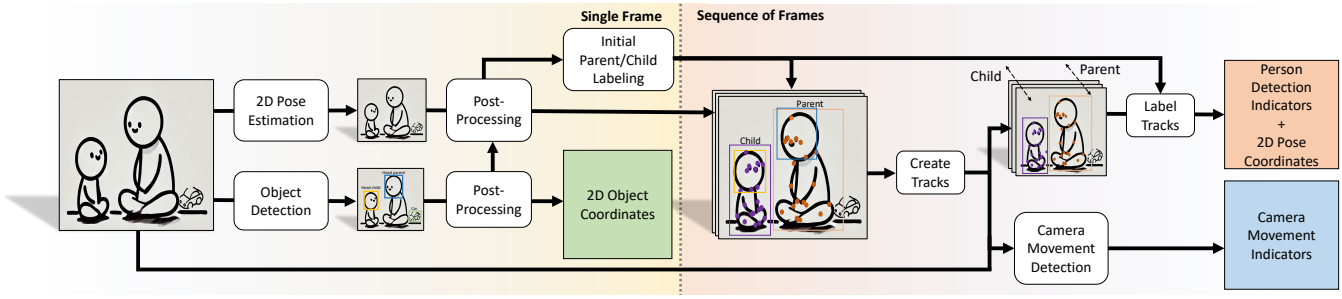


**Figure 2: SQUAD Stage 1: detecting 2D poses and objects per frame. Person detections are linked over time and assigned a parent/child label. Finally, for each frame, we estimate if each camera is moving.**

### 2.3 Automated Object Use

Human-Object Interaction Detection (HOI-DET) models represent interactions through triplets of human, object, and interaction [13]. Nonverbal Interaction Detection (NVI-DET) approach [42] adopt a similar triadic formulation, but also considering group-aware reasoning among multiple people. While these approaches focus on interaction labels, modeling hand-object contact or object use can provide more fine-grained insights into interaction dynamics, enabling more flexible analysis.

## 3 Data and Annotations

**YOUth Cohort dataset.** For our analyses, we use parent-child interaction videos from the YOUth Cohort study [28] which are available through an ethical approval process. In each interaction, an approximately 10-month-old infant and a parent engaged in free-play. Interactions lasted approximately 12 minutes each and were recorded with four cameras. An experiment assistant controlled the cameras in real-time by zooming in on a particular aspect of the interaction or, in contrast, to make sure the space was visible in its entirety. All dyads were recorded with the same experimental setup, with the same set of toys. After initial screening for video quality, we randomly selected 60 parent-child interaction videos.

**Annotations.** We sampled one frame every ten seconds in all 60 interactions. For each person, one annotator labeled which objects were attended to or used. We explicitly allowed the annotation of VFOA and object use to multiple objects per frame, which occurs regularly in practice, especially in the interactions such as object stacking. We only annotated objects that were visible in at least one view. In total 4,539 frames were annotated for both VFOA and object use. For VFOA, in total 6,773 person instances (2,916 of a parent, 3,857 of a child) were annotated, with an average of 1.50 attended objects per person. For object use, in total 5,527 person instances (2,295 of a parent, 3,232 of a child) were annotated, with an average of 1.39 touched objects per person.

To assess inter-annotator reliability, five randomly selected videos were re-annotated independently. A second annotator labeled these videos for VFOA and for object use. The macro-averaged F1-score between two sets of VFOA for parent and child was 77.32% and 78.42%, respectively, with an overall agreement of 77.87%. The macro-averaged F1-score between two sets of object use for parent and child was 87.82% and 87.28%, respectively, with an overall agreement of 87.55%.

## 4 SQUAD: Scene QUantification for the Analysis of Dyads

Given a parent-child interaction video recorded from multiple views, the SQUAD tool is applied in two stages. Stage 1 operates on a single view, while Stage 2 considers the multiple views jointly. Finally, 3D body poses, consistently linked to parent and child, are obtained along with the 3D object locations of the toys in the scene. We discuss steps of both stages in the following sections.

### 4.1 Stage 1: Single View

An overview of Stage 1 is visualized in Figure 2.

**2D pose estimation**. For each frame, multi-person 2D pose estimation is performed to detect person bounding boxes and to estimate the 2D locations of body joints. In our implementation, AlphaPose [10] is used, but SQUAD can work with a range of 2D pose estimation tools. AlphaPose follows a top-down approach, by first detecting people in the scene before estimating the 2D pose for each detection. Our motivation to not directly estimate 3D body poses from a single view is that such algorithms, in our experience, tend to produce significantly varying estimations of a person's scale and distance to camera over time. This effect occurs especially when people are seated, or have body dimensions that differ significantly from the data that these methods are typically trained on. We found that free-play data would produce very different relative positioning of the people depending on the view used. Therefore, we estimate 2D body poses first, and subsequently use multiple views to lift the body poses to 3D.

**Object detection**. In parallel, YOLOv8[2] is used to detect the toys and heads of parent and child. The toys are a car, textile book, doll, baby bottle, jump box, plush flower, and a shape sorter with lid and four colored shapes, in line with [24]. We fine-tuned a pre-trained YOLOv8 model on annotated frames from YOUth Cohort videos to enable automated toy and head detection. The supplementary material provides details. Other object detectors can be easily used with SQUAD, for example to operate on different sets of toys. The 2D object detections are post-processed to enhance robustness. First, since each toy appears at most once in a frame, duplicate detections with lower confidence are removed. Next, pose estimations corresponding to overlapping person detections are

[2]https://docs.ultralytics.com/models/yolov8/

removed. Occasionally, the doll was detected as a person. To mitigate this, person detections overlapping with detected objects were removed, with an intersection over union (IoU) above 0.8.

**Initial parent/child labeling**. To initially label person detections as parent and child, head detections from the object detection step and relevant keypoints from the 2D pose estimation are leveraged. For each detected person, SQUAD constructs a head bounding box capturing a subset of the keypoints (i.e., nose, eyes, head, shoulders), and observes the overlap with head bounding boxes. Based on the maximum overlap between two head bounding boxes, the tool labels each person detection as parent, child, or no match.

**Create tracks**. Next, sequences of frames are processed to obtain temporally consistent person tracks for each view. Unstable keypoints such as wrists, heels, and toes are excluded because these are often occluded. Because initial person labels are not sufficiently accurate, detections cannot directly be linked. Instead, tracklet pairs are constructed over time by jointly considering parent and child. For each frame, a person detection is compared to the two tracklets in the pair by computing the median displacement between the current keypoints and the last available keypoints stored in each tracklet. A detection is then assigned to the tracklet with the smallest median displacement. After this initial assignment, if any tracklets remain unmatched, a second pass is performed to match them against remaining unassigned detections. For each proposed match, two consistency checks are applied: the median displacement must be below a predefined displacement threshold, and the temporal gap between the current frame and last frame of the tracklet must not exceed a frame-gap threshold. If both checks are satisfied, the detection is included. Otherwise, the tracklets are terminated and a new tracklet pair is initiated.

After forming tracklet pairs, they are merged into person tracks using a matching function based on median within-keypoint spread. Tracklet pairs are retained only if their duration is at least one second and the difference in within-keypoint spread exceeds 10 px. Finally, gaps in the resulting tracks are filled by interpolating missing bounding boxes and keypoints using nearest-neighbor interpolation, with the maximum interpolation gap set to one second.

**Camera movement detection**. Camera movement is detected for each view using two steps. First, between consecutive frames, feature points are tracked and mean background pixel displacement is computed. To avoid camera movement detection caused by movements of people in the scene or the objects interacted with, regions covered by tracked people and detected objects are masked out. A low displacement threshold is used to reduce false negatives. In a second step, each camera's movement interval is validated by sampling pairs of frames from before and after the interval and the average displacement across pairs is recomputed. Intervals with mean displacement below a less sensitive threshold are discarded. They might correspond to incorrect estimates, potentially due to lighting changes.

**Label tracks:** Tracks are automatically linked to the parent or child by comparing their median within-keypoint spread. It is assumed that the parent exhibits a larger within-keypoint spread than the child, reflecting body scale. Match ratios are then computed for each track based on the initial parent and child labels and tracks are assigned to parent/child.

Finally, tracking results are combined with camera movement detection and per-frame person detection and camera movement binary indicators are derived.

## 4.2 Stage 2: Multi-View

In the second stage, SQUAD combines the per-view 2D information to produce 3D outputs. The construction and scheduling of the sequences is described in Section 4.3. Steps are outlined below, and depicted in Figure 3.

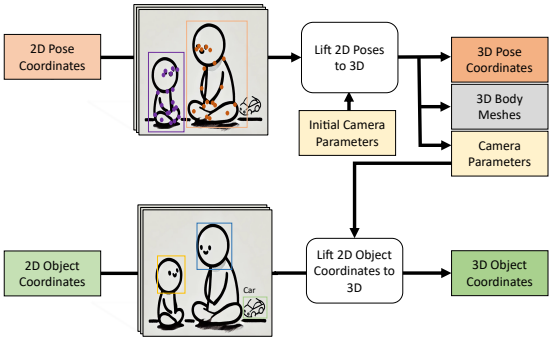


**Figure 3: SQUAD Stage 2: lifting 2D poses and objects to 3D.**

**Lift 2D poses to 3D**. The Dynamic Multi-Mesh Recovery (DMMR) method from Huang et al. [17] is used to estimate 3D poses and camera position, orientation, and zoom level. DMMR jointly optimizes these by iteratively fitting a parametric Skinned Multi-Person Linear (SMPL) body model [25] to sets of 2D keypoints. SMPL is not designed for the body proportions of 10-month-old infants. While SMIL would be a potential alternative [14], this model was developed for a younger population that is predominantly in supine position. In contrast, infants in our data move around freely and some can walk with support. Therefore, we decided to scale a SMPL model for the infants to account for the difference in body size. Given average lengths of parents and infants in the target population of 176 and 80 centimeters, respectively, our scaling factor is 45.45%.

While DMMR can deal with camera movement, we found that DMMR would compensate inaccurately estimated 2D joint coordinates across views by repositioning the cameras. Given that 2D pose estimations of especially the lower body were noisy, we decided to apply DMMR in two modes, depending on whether optimized camera parameters are available. If the cameras are not calibrated, DMMR is applied to optimize the camera parameters of all views together with the 3D pose coordinates and body meshes. Default initial parameters that are approximated on a random video in the YOUth Cohort dataset are used. Since DMMR is applied on sequences of frames without camera motion, a single set of camera parameters can be optimized over all frames instead of per frame. Once optimized camera parameters are available, DMMR is applied without optimizing the views. Also in this case, the input sequences do not contain camera motion. This is achieved by the scheduler that is discussed in Section 4.3.

**Lift 2D object coordinates to 3D**. For each view, detections of the same object are grouped across frames to form temporal

object tracks, and missing 2D detections are linearly interpolated for gaps up to 30 frames. To estimate the 3D coordinates of each toy and head in a frame, all detections from multiple views are considered. Since partial occlusions are common and could shift the detected object center, detections with confidence scores below 0.5 are filtered out. When detections from at least two views are available, the 3D object center is estimated via triangulation using the corresponding camera parameters. Finally, Gaussian smoothing is applied to the resulting 3D trajectories to obtain temporally smoother object motion.

## 4.3 Sequence Scheduling

For robust estimation of 3D body poses and camera parameters, interactions must be segmented into sequences without camera movement. To this end, the camera movement indicators are employed for temporal segmentation. Due to occlusions, low image quality, or when one of the interactants left the play area, we might end up with fewer than two person detections. For each period without camera movement, the longest subsequence is selected in which both interactants are detected in all views, as indicated by the person detection indicators. This reference subsequence is processed to optimize camera parameters, which are propagated to all remaining subsequences within the same stable period. The tool then proceeds with Stage 2, but without optimizing the camera parameters.

Our scheduling strategy enables 3D reconstruction when one interactant is temporarily missing in some views, while still relying on robustly estimated camera parameters from neighboring subsequences. This procedure is applied whenever there is a reference subsequence of at least two seconds. Our approach skips the periods in which camera movement occurs.

## 4.4 VFOA Estimation Module

We extended SQUAD to make visual focus of attention predictions. Because the 3D body pose information does not provide the direction of the eyes, eye orientation estimation is based on the orientation of the head and the pose of the upper body. VFOA is determined by checking which objects are approximately in the line of sight. We now detail these steps, see also Figure 4.

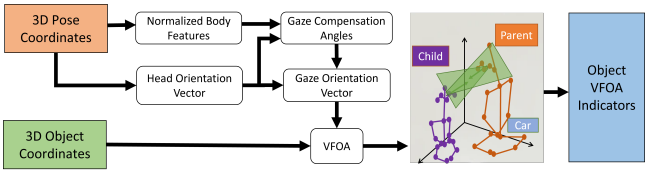


**Figure 4: VFOA based on 3D body poses and 3D object centers, estimated with SQUAD.**

**Gaze compensation**. At the core of our approach is a neural network that is trained to estimate eye gaze from the 3D body poses obtained with SQUAD. The rationale is that eye, head, and body orientation are correlated to some degree [32]. So, we trained a neural network to learn the relation between eye gaze as a function of head and body orientation. The network takes as input a hip-centered normalized representation of the upper body, and an additional head pitch value. The hip center is used as the origin [31]. The x-axis is defined by the direction between the hip joints, while y-axis is obtained by projecting a fixed vertical direction onto the plane orthogonal to this axis. All joint coordinates are translated to the hip center, and the transformed coordinates of the nose, eyes, and shoulders are then used as input to the network. This representation contains both body and head orientation, but is mainly distinctive in horizontal direction because the eyes and nose are approximately at the same height. As an additional input to the network, the head pitch is therefore included. The head orientation vector is first calculated as the vector from the point halfway the top of the head and the neck joints to the midpoint between the eyes. The pitch is then calculated as the angle between this head orientation vector and the global y-axis.

The neural network consists of one hidden layer to allow for more complex interactions between the input features. The parameters of the network are determined using the training folds. For each frame and person in the training set with an annotated VFOA target, the vector towards the estimated 3D center of the annotated objects is used as the target. The gaze compensation angles are then calculated as the horizontal and vertical angles between the target vector and the head orientation vector. To reach the corrected gaze vector, the predicted gaze compensation angles added to the head orientation vector where the horizontal component is in the direction between the two eyes, and the vertical component in the direction of the vector neck-head top.

**VFOA prediction**. We evaluate the corrected gaze vector using two approaches. The cone-based approach models attention as a spatial region in 3D space, whereas the feature-based approach addresses the same problem by learning object-wise gaze-to-center relationships.

First, we predict VFOA with a cone-based approach. Following the gaze compensation, the angle between the gaze orientation vector and the estimated center of each detected toy is checked. If this angle is below a cone angle threshold, the target is considered to be in the line of sight. Essentially, this approach considers object centers within a gaze cone to be attended to. This approach reduces reliance on the assumption that gaze is always directed at the estimated center of an object, which is typically not the case.

To avoid predicting occluded objects, the cone distance is limited to the closest object, plus an additional margin. This margin allows the model to predict multiple objects that are used simultaneously in close proximity. This occasionally happens, for example, when a shape is put in the shape sorter, or when the baby bottle is placed in the doll's mouth. The cone angle threshold and the distance margin are determined for each fold by maximizing the macro-averaged F1-score on each training fold. The angle threshold is selected from a range of 20 to 40 degrees, while the distance margin is chosen from a range of 0.05 to 0.30 meters. Then a feature-based approach is employed for VFOA prediction. For each object, the cosine similarity between the corrected gaze vector and direction from the midpoint between eyes to the 3D object coordinate is computed. This cue is used as an input feature to train object-specific XGBoost classifiers for VFOA, with separate models for each role and object class. The model outputs a probability of VFOA, which is converted into a binary prediction by applying a decision threshold. The decision

threshold is selected on the training folds from the precision-recall curve by maximizing the F1-score.

### 4.5 Object Use Module

We further extend SQUAD to predict object use, leveraging 3D object centers and wrist coordinates obtained from 3D pose coordinates. For each object, we first identify the closest wrist and then compute the distance between the object center and this wrist. Object use is determined based on this distance, evaluated separately for each object. This approach accounts for differences in object dimensions. An overview of the module is shown in Figure 5.

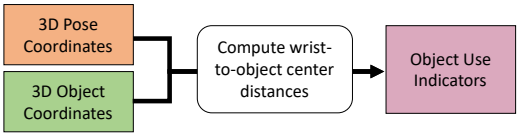


**Figure 5: Object use estimation based on 3D body poses and 3D object centers, estimated with SQUAD.**

**Object use prediction.** We use a feature-based approach for object use prediction. For each object, the distances between the object center and the two wrist joints are computed, and the closest wrist distance is used as a feature for binary object use classification. Defining the feature relative to the closest wrist, rather than a fixed one, ensures applicability across different interaction patterns. Evaluating the distance separately for each object also helps account for differences in object size and placement.

Based on this feature, object- and role-specific XGBoost classifiers are trained. The model outputs a probability of object use, which is converted into a binary prediction by applying a decision threshold. For each fold, this threshold was selected on the training set from the precision-recall curve by maximizing the F1-score.

## 5 Experimental Results

Using the VFOA and object use modules, we conducted three sets of experiments: i) unimodal VFOA prediction, ii) unimodal prediction of object use, and iii) a combination of those as the multimodal prediction of VFOA.

We processed videos of 60 selected interactions with SQUAD. Some frames could not be processed, mostly due to poor visibility and consequently missing pose detections, and partly due to camera movements. The interactions were divided into five folds. No participants in the test folds appear in the training folds. An overview is provided in the supplementary material.

For hyper-parameter tuning, we used the training folds, where we included only the frames with VFOA annotations and a valid 3D pose estimated by SQUAD, for each role. For testing, all frames with valid 3D pose estimations were included. Across the folds, the training set consisted of 2,029 frames (2,218 of a parent, 1,940 of a child) on average, and the test set consisted of 630 frames (654 of a parent, 606 of a child) on average.

For unimodal VFOA prediction we first consider the cone-based approach and then compare it with the feature-based approach that leverages gaze features for object-level VFOA prediction. For object use prediction, we adopt the feature-based approach based on wrist-object proximity. Finally, we integrate cues from both VFOA and object use to assess the contribution of object use to VFOA. The implementation of our methods is discussed subsequently.

For SQUAD, we used the default parameters described in Section 4. We trained separate gaze compensation neural networks for the parent and the child to account for differences in body dimensions. Each network included a hidden layer with 20 neurons. In the SQUAD cone-based VFOA approach, the best combination of cone angle threshold and distance margin was determined on the training folds in the 5-fold cross-validation. Hyperparameter tuning on the training folds revealed that the optimal distance margin was 0.1 meters across all folds and roles. The best angle threshold across folds ranged between 26° to 30° for the parent, and between 34° to 38° for the child. For the feature-based approaches using XGBoost, both for VFOA and object use prediction, a maximum tree depth of 3 and 20 boosting iterations were used, with early stopping applied on the training set. To address the imbalance between positive and negative samples, mainly due to the lower frequency of interactions with certain objects (e.g., baby bottle), we applied positive class weighting. Across all experiments, we report prediction performance using precision, recall, and F1-score. Macro-averaged metrics were calculated by first averaging over object classes within each fold and then across folds. Table 1 provides a comparison of unimodal VFOA and object use prediction, as well as multimodal VFOA prediction. We also detail the results for each experiment, followed by a summary of key observations.

### 5.1 Unimodal VFOA Prediction

When using the SQUAD cone-based approach for VFOA prediction, a macro-averaged F1-score of 40.25% was obtained. This score was largely affected by the reduced performance on smaller objects (such as shape sorter shapes and the bottle), which were either not detected reliably or were rarely interacted with, resulting in a very low or zero F1-scores which in turn has a strong effect on the macro-average. When they were excluded, the score increased by more than 15%. Similarly, when undetected objects were excluded from the evaluation, the macro-averaged F1-score increased to 53.75%.

To further analyze the performance of the cone-based VFOA approach, we compared SQUAD cone-based model with several alternative and baseline models. As a first baseline model, we considered a setting without gaze compensation, where the head orientation vector was directly used as the gaze vector. This baseline yielded a substantially lower F1-score of 25.55%, mainly due to poorer performance in child VFOA predictions. This shows that compensating for the missing eye orientation is beneficial. Additional baseline comparisons are provided in the supplementary materials.

Using SQUAD feature-based VFOA model, we obtained a macro-averaged score of 41.53%. As the computation of the input features required the 3D object center, the evaluation was restricted to only the detected objects. Consistent with the cone-based VFOA prediction, when the smaller objects were excluded, the macro-averaged F1-score increased by more than 10%.

Compared to the cone-based approach evaluated with detected objects, the feature-based model yielded lower performance. This can be attributed to the fundamental difference in how attention

| Setting | Person | Precision | Recall | F1 |
|---|---|---|---|---|
| **SQUAD: Cone-based VFOA** | Parent | 40.34% | 41.59% | 36.86% |
| | Child | 53.13% | 42.66% | 43.64% |
| **SQUAD: Cone-based VFOA (only detected)** | Parent | 41.53% | 63.24% | 47.20% |
| | Child | 54.11% | 73.14% | 60.29% |
| **SQUAD: Cone-based VFOA (no compensation)** | Parent | 39.16% | 36.33% | 33.57% |
| | Child | 32.08% | 15.11% | 17.53% |
| **SQUAD: Feature-based VFOA** | Parent | 30.63% | 50.67% | 36.11% |
| | Child | 40.08% | 63.08% | 46.96% |
| **Sharingan [38]** | Parent | 43.99% | 18.72% | 23.39% |
| | Child | 49.23% | 29.96% | 32.91% |
| **SQUAD: Feature-based object use** | Parent | 35.23% | 56.79% | 40.37% |
| | Child | 53.95% | 65.09% | 56.87% |
| **SQUAD: Multimodal VFOA** | Parent | 38.58% | 51.23% | 42.15% |
| | Child | 52.59% | 58.85% | 53.62% |

**Table 1: VFOA (first five rows), object use, and multimodal prediction results for SQUAD.**

was modeled. The cone-based approach defines VFOA as objects included in a spatial region, allowing multiple nearby objects to be considered simultaneously, whereas the feature-based model reduces the problem to a relationship between gaze direction and object centers. This simplification makes the model more sensitive to inaccuracies in object coordinates and gaze estimation. Still, the feature-based approach allowed the model to capture differences in how strongly attention correlated with gaze-object alignment for different objects and for roles. This demonstrates that, despite its informativeness, VFOA prediction requires richer representations beyond simple gaze-object center alignment.

**Comparison with Sharigan.** We also compared the performance of our unimodal VFOA models with Sharingan [38], where the evaluation relied on 2D information from single camera views. To associate VFOA to either person, we assigned parent and child roles to the Sharingan head detections based on their IoU with SQUAD automated 2D head detections. Furthermore, we derived VFOA predictions by identifying the object bounding boxes which the predicted gaze target fell, similar to the processing of SQUAD.

While SQUAD was not able to process without valid joints on average for 253/308 frames of the parent/child, Sharingan did not require valid joints to predict VFOA, and processed on average 130/135 parent/child frames that SQUAD was not able to process due to missing body pose. When excluded frames with missing body pose, SQUAD processed on average 237/250 parent/child frames more than Sharingan and on average 107/114 when not filtered out. Considering all annotated frames, Sharingan had gaze output but could not assign the requested role for an average of 146/208 parent/child frames per fold. It was unable to produce any gaze output for an average of 215/215 parent/child frames per fold. When restricted to Sharingan's valid-joints subset, Sharingan had gaze output but could not assign the requested role for an average of 103/127 parent/child frames per fold, and was unable to produce any gaze output for an average of 135/122 parent/child frames per fold. Overall, Sharingan produced final role-assigned VFOA predictions for an average of 547/485 parent/child frames per fold.

We compared our results with Sharingan for the subset of frames with valid body pose. Overall, Sharingan and SQUAD had an overlap of 417/350 parent/child frames evaluated. To evaluate Sharingan, we applied Sharingan on all the available views and selected the view predictions with the highest activity, calculated by the peak activity divided by the total heatmap activity to select the best predictions. Both of our unimodal VFOA models outperformed Sharingan by more than 10%. This suggests that incorporating 3D information can be beneficial for VFOA in dyadic interactions. In the absence of depth information, predictions are constrained to approximate 2D targets rather than being grounded in scene geometry, which is especially problematic in naturalistic settings with close-range objects and frequent occlusions.

## 5.2 Unimodal Object Use Prediction

With the feature-based object-use model, we obtained a macro-averaged score of 48.62%. The overall performance was substantially affected by objects with fewer touch instances, such as shape sorter objects, the flower, and the baby bottle. Despite this imbalance, more than 60% of the positive touch events were correctly detected. Furthermore, excluding small objects led to an increase of approximately 15% in the macro-averaged F1-score. These results indicate that wrist-object proximity is a reliable cue for object use in free-play settings.

## 5.3 Multimodal VFOA Prediction

The SQUAD multimodal model achieved a macro-averaged F1-score of 47.89%, an improvement of more than 5% over the feature-based VFOA model. This suggests that the inclusion of wrist-object proximity provides complimentary cues for VFOA beyond the gaze-object center alignment. In particular, multimodal integration leads to a noticeable increase in precision, indicating that object use information helps to reduce false positive predictions. This suggests that gaze alone may produce ambiguous predictions in cases where multiple objects are spatially close, whereas wrist-object proximity provides an additional constraint by indicating which object is being actively interacted with. Moreover, combining gaze and object use allows the model to better reflect the VFOA. While gaze captures where attention is directed, object use provides evidence about the underlying intention, filtering out false VFOA selections. This complementary relationship is especially beneficial in naturalistic scenarios with occlusions and close positioned targets.

**Performance per fold**. For the unimodal VFOA models, F1-scores per fold differed by approximately 8%. This difference was

considerably smaller for object use, with a difference of 2.69%. When incorporating object use cues into the feature set, the variation in VFOA F1-scores across folds decreased to 5.68%.

The observed variability across folds suggests that some interactions are inherently more complex, potentially due to the positioning of interactants and toys. This complexity affects VFOA prediction more strongly then object use estimation. We observed that some interactants appeared to rely more on eye movements rather than head movements. This was particularly noticeable for parents lying on the floor, who tend to exhibit minimal head motion while primarily shifting gaze through eye movements. A qualitative analysis of interactions is provided in the supplementary material.

**Parent vs. child**. The F1-scores for the child were higher than those for the parent. The child was typically more centered between the toys, whereas the parent was usually a bit more to the side, leading to larger angles between the child's head and the different toys, compared to the parent. Consequently, gaze compensation had a weaker positive effect on parent VFOA prediction (Table 1), driven mainly by the parent's peripheral position.

## 6 Discussion

**2D vs. 3D interaction analysis**. SQUAD allows for straightforward analysis of spatial aspects of dyadic interactions, such as the object use and the estimation of visual focus of attention. 3D reconstruction overcomes issues with having to rely on priors in the direction of gaze, as in [24, 38]. The 3D reconstructions produced by SQUAD can be well interpreted because we can observe the interaction from all perspectives. Consequently, we can verify how persons in the interaction might have observed the scene.

**Reconstruction and tracking**. One limitation of our approach is that not all frames can be processed. While camera movement is one cause for missing frames, a more important factor is the lack of robust person detection. As a result, reliable 3D pose estimation and camera parameter optimization are not always feasible. A 2D pose estimation algorithm more suitable for seated poses and child proportions would provide better results, and thus increase the ratio of valid frames. Since SQUAD is agnostic to the implementation of the 2D body pose estimation algorithm, advances in pose estimation can easily be incorporated into the tool.

**Unimodal VFOA and object use estimation**. We have demonstrated the merits of 3D scene reconstruction for VFOA and object use. Both the cone-based and feature-based models provide practical solutions for naturalistic settings. Extending these modalities over time might reveal temporal dynamics such as sustained attention, and object manipulation. A shared limitation of these approaches is the representation of targets using solely 3D centers. In our interactions, object sizes vary substantially, and attending or touching an object does not necessarily occur at its center. This becomes particularly problematic for large objects, where the object center may be farther from the actual point of interaction, which can result in incorrect selection of nearby objects. Using the volume of the object, rather than just the center, could solve this issue.

**Multimodal VFOA estimation.** The multimodal VFOA results demonstrate that combining gaze-based features with wrist-object proximity improves feature-based prediction performance. The observed improvement is primarily driven by an increase in precision, indicating that the model produces fewer false positive predictions and becomes more selective in assigning attention targets.

**Feature-based models.** Feature-based model performances were highly affected by the skewed distribution of positive and negative samples, limiting the model's ability to learn reliable patterns. In uncontrolled settings such as free-play, object use varies substantially across interactions, and the cone-based approach may provide more robust performance than feature-based models.

**Annotation biases**. We observed that it is sometimes difficult to identify whether the parent is looking at the child's face or at the object the child is manipulating. Social cues such as the parent's smile can lead to a bias toward interpreting the parent as looking at the child. Similarly, with multiple objects in an attended region, annotators tended to favor the object being manipulated as the attended object. A similar issue arises for object use annotations, which do not strictly reflect direct physical contact but may also capture broader interaction-related behaviors, such as feeding the doll with the baby bottle. In addition, hand positions are not always visible across views due to occlusions. In such cases, annotators may rely on VFOA or the child's object use cues, leading to a bias toward labeling object use. These findings suggest that both VFOA and object use annotations are influenced by context.

## 7 Conclusion

We have introduced SQUAD, a tool to reconstruct dyadic behavior from multi-view camera recordings. It estimates and tracks the 3D body poses of both interactants along with the 3D locations of objects in the scene. We first extended SQUAD with a gaze compensation network and then demonstrated its potential for quantitative 3D analysis of interactions by addressing VFOA estimation with unimodal and multimodal inputs, as well as object use estimation. Evaluations on challenging free-play parent-child interactions reveal the merits of interaction analysis in 3D. We show improved VFOA estimations compared to a recent 2D baseline, and additional benefits of incorporating multimodal cues.

Future work should explore the measurement of other interactive behaviors, such as physical contact between the interactants. We also advocate a temporal analysis of interactive behaviors, to reveal how parent and child behavior mutually influence each other.

## Safe and Responsible Innovation Statement

The YOUth Cohort data were accessed following an ethical screening procedure and are available for GPDR-compliant research through controlled access via the YOUth data request system. To protect participant privacy, we only include a video still with explicit consent, which is not used for training or evaluation. Materials to reproduce our research will be made available with the paper under a non-commercial license. We introduce a tool for analyzing parent–child interactions, supporting research on child development and parental roles. In the long term, such video-based analyses may complement existing methods for assessing social behavior.

## Acknowledgments

This work was supported by a National Growth Fund AiNed Fellowship awarded to SMCdZ (NGF.1607.22.006). The authors wish to thank Vladyslav Kalyuzhnyy for initial work on SQUAD.

## References


[1] Andrea Abele. 1986. Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior* 10, 2 (1986), 83–101.

[2] Sadia Afroze, Md Rajib Hossain, and Mohammed Moshiul Hoque. 2022. DeepFocus: A visual focus of attention detection framework using deep learning in multi-object scenarios. *Journal of King Saud University-Computer and Information Sciences* 34, 10 (2022), 10109–10124.

[3] Stylianos Asteriadis, Kostas Karpouzis, and Stefanos Kollias. 2014. Visual focus of attention in non-calibrated environments using gaze estimation. *International Journal of Computer Vision* 107, 3 (2014), 293–316.

[4] Yiwei Bao and Feng Lu. 2024. Unsupervised gaze representation learning from multi-view face images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1419–1428.

[5] Simon Baron-Cohen. 1989. Joint-attention deficits in autism: Towards a cognitive analysis. *Development and Psychopathology* 1, 3 (1989), 185–189.

[6] Haibin Cai, Bangli Liu, Jianhua Zhang, Shengyong Chen, and Honghai Liu. 2015. Visual focus of attention estimation using eye center localization. *IEEE Systems Journal* 11, 3 (2015), 1320–1325.

[7] Lisa Capps, Marian Sigman, and Peter Mundy. 1994. Attachment security in children with autism. *Development and psychopathology* 6, 2 (1994), 249–261.

[8] Nevena Dimitrova. 2020. The role of common ground on object use in shaping the function of infants' social gaze. *Frontiers in psychology* 11 (2020), 619.

[9] Philip J Dunham and Chris Moore. 2014. Current themes in research on joint attention. In *Joint attention*. Psychology Press, 15–28.

[10] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2022), 7157–7173.

[11] Yu Fang, Ryoichi Nakashima, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. 2015. Eye-head coordination for visual cognitive processing. *PloS one* 10, 3 (2015), e0121035.

[12] K Fujiwara, QS Bernhold, NE Dunbar, CD Otmar, and M Hansia. 2020. Comparing manual and automated coding methods of nonverbal synchrony. *Communication Methods and Measures* 15, 2 (2020), 103–120.

[13] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8359–8367.

[14] Nikolas Hesse, Sergi Pujades, Michael J. Black, Michael Arens, Ulrich G. Hofmann, and A. Sebastian Schroeder. 2020. Learning and Tracking the 3D Body Shape of Freely Moving Infants from RGB-D sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2020), 2540–2551.

[15] Yoichiro Hisadome, Tianyi Wu, Jiawei Qin, and Yusuke Sugano. 2024. Rotation-constrained cross-view feature fusion for multi-view appearance-based gaze estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5985–5994.

[16] Haley Kranstuber Horstman, Alexie Hays, and Ryan Maliski. 2016. Parent–Child Interaction. In *Oxford Research Encyclopedia of Communication*. Oxford University Press.

[17] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. 2021. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 710–720.

[18] Kathryn L Humphreys and Julia Garon-Bissonnette. 2025. Parent–Child Interactions in Infancy and Early Childhood. *Annual Review of Developmental Psychology* 7 (2025), 117–140.

[19] Chiara Jongerius, Timothy Callemein, Toon Goedemé, Kristof Van Beeck, Johannes A Romijn, EMA Smets, and MA Hillen. 2021. Eye-tracking glasses in face-to-face interactions: Manual versus automated assessment of areas-of-interest. *Behavior Research Methods* 53, 5 (2021), 2037–2048.

[20] Berfu Karaca, Albert Ali Salah, Jaap Denissen, Ronald Poppe, and Sonja M.C. de Zwarte. 2024. Survey of Automated Methods for Nonverbal Behavior Analysis in Parent-Child Interactions. In *Proceedings of the International Conference on Face and Gesture Recognition*. 1–11.

[21] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (1967), 22–63.

[22] Michael F Land. 2006. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research* 25, 3 (2006), 296–324.

[23] Peitong Li, Hui Lu, Ronald Poppe, and Albert Ali Salah. 2023. Automated detection of joint attention and mutual gaze in free play parent-child interactions. In *Companion Publication of the International Conference on Multimodal Interaction*. 374–382.

[24] Peitong Li, Albert Ali Salah, Joyce J Endendijk, and Ronald Poppe. 2025. Camera-based Assessment of Gendered Toy Preference in Free-Play Parent-Child Interactions. In *Proceedings of the IEEE International Conference on Development and Learning*. IEEE, 1–8.

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (2015), 248.

[26] Qiaomu Miao, Vivek Raju Golani, Jingyi Xu, Progga Paromita Dutta, Minh Hoai, and Dimitris Samaras. 2025. Multi-view Gaze Target Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5371–5381.

[27] Peter Mundy and Lisa Newell. 2007. Attention, joint attention, and social cognition. *Current Directions in Psychological Science* 16, 5 (2007), 269–274.

[28] N Charlotte Onland-Moret, Jacobine E Buizer-Voskamp, Maria EWA Albers, Rachel M Brouwer, Elizabeth EL Buimer, Roy S Hessels, Roel de Heus, Jorg Huijding, Caroline MM Junge, René CW Mandl, Pascal Pas, Matthijs Vink, Juliëtte J M Van der Wal, Hilleke E Hulshoff Pol, and Chantal Kemner. 2020. The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience* 46 (2020), 100868.

[29] Kazuhiro Otsuka, Yoshinao Takemae, and Junji Yamato. 2005. A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the International Conference on Multimodal Interfaces*. 191–198.

[30] Lorenzo Piccardi, Basilio Noris, Olivier Barbey, Aude Billard, Giuseppina Schiavone, Flavio Keller, and Claes von Hofsten. 2007. WearCam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 594–598.

[31] Ronald Poppe, Sophie Van Der Zee, Dirk KJ Heylen, and Paul J Taylor. 2014. AMAB: Automated measurement and analysis of body motion. *Behavior research methods* 46, 3 (2014), 625–633.

[32] P Radau, D Tweed, and T Vilis. 1994. Three-dimensional eye, head, and chest orientations after large gaze shifts and the underlying neural strategies. *Journal of Neurophysiology* 72, 6 (1994), 2840–2852.

[33] Lori-Ann R Sacrey, Susan E Bryson, and Lonnie Zwaigenbaum. 2013. Prospective examination of visual attention during play in infants at high-risk for autism spectrum disorder: A longitudinal study from 6 to 36 months of age. *Behavioural brain research* 256 (2013), 441–450.

[34] Samira Sheikhi and Jean-Marc Odobez. 2015. Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognition Letters* 66 (2015), 81–90.

[35] Rémy Siegfried and Jean-Marc Odobez. 2021. Visual focus of attention estimation in 3d scene with an arbitrary number of targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3153–3161.

[36] Anja Sommer, Claudia Hachul, and Hans-Günther Roßbach. 2016. Video-based assessment and rating of parent-child interaction within the national educational panel study. In *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study*. Springer, 151–167.

[37] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. 2023. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF international conference on computer vision*. 20935–20946.

[38] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. 2024. Sharingan: A Transformer Architecture for Multi-Person Gaze Following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2008–2017.

[39] Michael Tomasello and Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child development* 57, 6 (1986), 1454–1463.

[40] Ines Van Keer, Eva Ceulemans, Nadja Bodner, Sien Vandesande, Karla Van Leeuwen, and Bea Maes. 2019. Parent-child interaction: A micro-level sequential approach in children with a significant cognitive and motor developmental delay. *Research in Developmental Disabilities* 85 (2019), 172–186.

[41] Pierre Vuillecard and Jean-Marc Odobez. 2025. Enhancing 3D Gaze Estimation in the Wild using Weak Supervision with Gaze Following Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. 2024. Nonverbal interaction detection. In *European Conference on Computer Vision*. Springer, 277–295.

[43] Zhenzhen Weng, Laura Bravo-Sánchez, Zeyu Wang, Christopher Howard, Maria Xenochristou, Nicole Meister, Angjoo Kanazawa, Arnold Milstein, Elika Bergelson, Kathryn L Humphreys, Lee M. Sanders, and Serena Yeung-Levy. 2025. Artificial intelligence–powered 3D analysis of video-based caregiver-child interactions. *Science Advances* 11, 8 (2025), eadp4422.