The AI4Autism Project: A Multimodal and Interdisciplinary Approach to Autism Diagnosis and Stratification

Samy Tafasca
stafasca@idiap.ch
Idiap Research Institute, Martigny

Anshul Gupta
agupta@idiap.ch
Idiap Research Institute, Martigny

Nada Kojovic
Nada.Kojovic@unige.ch
University of Geneva, Geneva

Mirko Gelsomini
mirko.gelsomini@supsi.ch
SUPSI, Manno

Thomas Maillart
thomas.maillart@unige.ch
University of Geneva, Geneva

Michela Papandrea
michela.papandrea@supsi.ch
SUPSI, Manno

Marie Schaer
Marie.Schaer@unige.ch
University of Geneva, Geneva

Jean-marc Odobez
odobez@idiap.ch
Idiap Research Institute, Martigny

ABSTRACT

Nowadays, 1 in 36 children is diagnosed with autism spectrum disorder (ASD) according to the Centers for Disease Control and Prevention (CDC) [52], which makes this condition one of the most prevalent neurodevelopmental disorders. For children on the autism spectrum who face substantial developmental delays, the trajectory of their cognitive growth can be markedly improved by interventions if the condition is identified early. Therefore, there is a critical need for more scalable screening and diagnostic tools, as well as the need to improve phenotyping to refine estimates of ASD symptoms in children. Here, we introduce AI4Autism: a 4-year project funded by the Swiss National Science Foundation, which aims to address the needs outlined above. In this project, we examine the potential of digital sensing to provide automated measures of the extended autism phenotype. This is accomplished using multimodal techniques based on computer vision and Internet of Things sensing, for the purpose of stratifying autism subtypes in ways that would allow for precision medicine. We present an overview of our main results so far, introducing datasets and annotations that we intend to make publicly available, as well as methods and algorithms for analyzing children's behaviors and producing an ASD diagnosis.

KEYWORDS

Autism, computer vision, internet of things, datasets, neural networks, gaze detection, behavior analysis.

ACM Reference Format:
Samy Tafasca, Anshul Gupta, Nada Kojovic, Mirko Gelsomini, Thomas Maillart, Michela Papandrea, Marie Schaer, and Jean-marc Odobez. 2023. The AI4Autism Project: A Multimodal and Interdisciplinary Approach to Autism Diagnosis and Stratification. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3610661.3616239

ICMI '23 Companion, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0321-8/23/10…$15.00
https://doi.org/10.1145/3610661.3616239

1 INTRODUCTION

Autism is a prevalent lifelong neurodevelopmental disorder characterized by deficits in communication (e.g. language delay) and reciprocal social interactions (e.g. turn-taking in conversations), and by the presence of restricted and repetitive behaviors and interests (e.g. hand-flapping) [1]. The term spectrum in ASD is used to refer to the wide range of possible symptoms and their perceived severity. The level of intellectual and adaptive functioning is highly variable, as well as the level of functioning in the different areas of life in adulthood (e.g. relationships, education, employment).

Autism often becomes evident in the first 3 years of childhood, although it can sometimes remain undiagnosed until adulthood when the symptoms are more subtle. For children with a larger developmental delay, getting a proper and early diagnosis is of critical importance for access to early intervention programs that have a tremendous impact on the child's long-term outcome [27]. It is indeed now widely recognized that early intervention has the potential to substantially transform the future of affected children [27, 59, 77, 93], and that there is a direct relationship between the age at the start of the intervention and the outcome [93]. Even in the absence of a specialized intervention program, an earlier diagnosis is associated with a significantly better outcome, as parents can learn how to optimally support the development of their child.

In addition to screening, there is also an important need to better understand and delineate the whole spectrum of ASD. Clinicians currently lack insights and methods to best stratify between different autism sub-types, which could then be associated with a different prognosis or sensitivity to treatment [6]. This motivates the need for having a more granular strategy for autism characterization, taking into account the intricacies of symptoms, hence moving towards precision medicine, i.e. tailoring intervention strategies to the specific characteristics of each individual [20].

Recent developments in digital sensing and machine learning have offered opportunities for seamless sensing of body movement, social scene capture, and measure of object manipulation. Despite considerable efforts, most studies in computational behavior coding and digital phenotyping for autism have suffered from the following

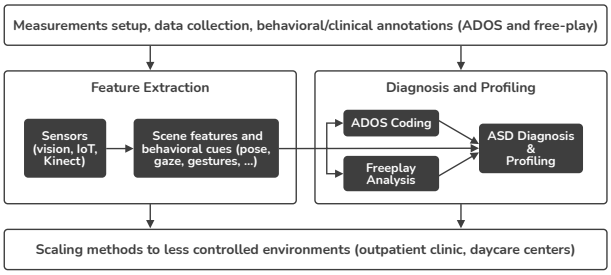

Figure 1: Key elements of the AI4Autism project.

limitations: 1. modest sample sizes, 2. mono-modal approaches, 3. focus on eliciting specific behaviors by largely controlled prompts, and 4. technical difficulties related to behavior sensing (viewpoints, children population, image resolution for gaze).

The AI4Autism project is an interdisciplinary project combining the skills of experts in clinical research, engineering, and computational social sciences in order to address these technical, scientific, and clinical limitations. An overview of its main aspects is given in figure 1. First, from a clinical research perspective, the project is grounded on the Geneva Autism Cohort consisting of more than 450 young children with ASD and their age-matched typically developing (TD) peers, extensively assessed with standardized clinical and cognitive assessments. Our aim is to design digital tools for screening, behavioral coding, and automated profiling of autism phenotypes, testing them first in a controlled setting with well-established clinical protocols, before deploying them to 'in the wild' field environments, such as daycare centers.

Digital sensing encompasses two approaches to be combined down the road: 1. With Internet of Things (IoT) sensors, investigate the monitoring of fine-grained motor skills developments of very young children, through the integration of inertial and low-cost Ultra-Wide-Band (UWB) indoor localization data. 2. Leveraging the availability of large behavioral and clinical annotation data, develop novel computer vision and multimodal machine learning methods for the analysis of motor-gaze coordination patterns, and for ASD diagnosis and profiling, with a focus on interpretable models.

In the remainder of this manuscript, we first present the main aims of the project (Section 2), before delving into our methodological approach, and how it contrasts with the current state of the art in digital phenotyping for autism (Section 3). Next, we motivate and describe the three datasets collected and their annotations (Section 4). Finally, we present the different computational models we have developed so far (Section 5), before concluding the paper with a summary and an overview of future works (Section 6).

2 PROJECT OBJECTIVES

The AI4Autism project has been designed to develop tools for automated behavioral coding and digital phenotyping of autism. The objective is to extract and comprehend the most relevant features for detecting non-verbal social and motor behaviors, which are key for diagnosing autism [2, 15, 30, 72, 91]. This approach is particularly suited for screening very young children (age 1 to 4), as the features that distinguish children with ASD from their typically developing peers are largely non-verbal at this age [26, 50, 51]. Our main aims are the following:

Aim 1: better stratification of ASD subtypes. From a clinical perspective, one goal is to study whether the fine-grained quantification of the child's behavior in a controlled setting can actually be exploited for better stratifying different autism sub-types, which is not possible with the current gold standard [35]. We expect that such stratification that goes beyond current clinical approaches can trigger the discovery of previously undetected autism sub-types, associated with a different prognosis, sensitivity to treatment, or neuro-biological mechanisms.

Aim 2: build a large-scale curated database of manual annotations of behaviors in preschoolers with autism. Intensive manual annotations still represent the only way to obtain a fine-grained quantification of specific autism symptoms. Drawing inspiration from the ADOS coding protocol, we set out to perform intensive fine-grained annotations of non-verbal behaviors, like joint attention, gestures, coordination of visual contact with other nonverbal behaviors, or play behaviors. Such annotations will be used for the study in aim 1, and serve as a basis for training computer-vision and IoT ASD profiling machine learning approaches (aim 3).

Aim 3: designing tools for automated ASD identification and digital phenotyping in young children under controlled conditions. Considering the need for robust automated and unbiased screening approaches, we investigate interpretable multimodal deep learning techniques (computer vision and IoT) for the recognition of ASD behavioral patterns related to free play activities and the coordination between gaze and posture/gestures, and obtain an automated classification between ASD and non-ASD, as well as a digital ASD profile. To build these tools, we will primarily work on data collected to analyze specific behaviors in clinical conditions, following specific assessment structures (ADOS), based on study samples that are clearly either ASD or typically developing (TD). Different methodological frameworks will be investigated here, as described in the next subsections.

Aim 4: validating and generalizing models and tools for ASD screening at large. While clinical data represents an important scientific step to obtain relevant datasets with annotations for both ASD profile and ground-truth behaviors, our aim is to design models working in more challenging conditions, from several perspectives: sensing, with more versatile recording structures such as mobile cameras or IoT set-ups and variable viewpoints or lighting conditions; different assessment protocols (less structured, shorter, etc); or less clearly defined clinical populations, e.g. children suspected of having autism seeking diagnosis in an outpatient clinic. These individuals might have autistic traits, or other comorbid conditions, departing from clear-cut research samples composed of well-separated positives and negatives. To demonstrate the scalability of our approach, we aim to validate models in the "real life" conditions of an outpatient clinic or directly in daycare centers.

3 RESEARCH DIRECTIONS

Several key elements need to be investigated to address the above aims. The first one, which we discuss in Section 4, is related to the creation of useful annotated datasets to enable clinical research, developing machine learning methods, and open science. The second one is the overall strategies for designing clinical and computational methods. In the following, we introduce the main limitations of the state-of-the-art from different perspectives and present the general approach we take to address them.

3.1 Clinical Perspective

Practices and limitations in autism screening and profiling. Autism spectrum disorder is a clinical diagnosis that relies on criteria described in the Diagnostic and Statistical Manual of Mental Disorders [1]. Early symptoms usually include altered modulation of eye contact and diminished engagement in social interactions, whereas non-social symptoms include altered quality of play, repetitive behaviors, mannerisms, and sensory issues. As the child develops language, he/she can often express himself with stereotyped language, echolalia, altered prosody, and show difficulties in turn-taking conversations. In short, ASD symptoms can present in a large variety of ways, depending on age and cognitive abilities.

Regarding screening, approaches largely rely on parent questionnaires given by primary healthcare practitioners, which are thus influenced by the subjective assessments in parents' reports [82], yielding suboptimal screening sensitivity [10]. Furthermore, their use depends on the level of training of these practitioners [25, 40]. To ensure unbiased access for each child, several authors advocate that screening should happen directly in childcare centers rather than in healthcare practices [23, 41, 46, 89]. This would, however, imply large-scale training of daycare workers. In this context, behavioral sensing tools clearly hold a promise to support autism screening [21, 75], in an automated, unbiased, and scalable manner.

Another critical challenge is to provide clinicians with relevant autism sub-types that could be associated with different prognoses [6]. Currently, the best clinical and research practice for autism phenotyping relies on an observational scale, the Autism Diagnostic Observation Schedule (the ADOS-2, [50]). It consists of a 45-60 min observational assessment completed by a highly-trained clinician that aims at creating a playful environment to elicit social behaviors, and rating behaviors on a scale of 0 to 3 following very strict and standardized procedures. While extremely useful for clinical practice standardization, the relatively coarse symptom granularity measure prevents and restricts the use of these tools to monitor the change over time. Therefore, digital tools for extended phenotyping carry the promise to gather more comprehensive sets of measures for autism symptoms on a continuous scale.

ASD digital stratification. As stated above, the current gold-standard diagnostic assessment ADOS relies on manual coding by expert clinicians, and only provides a semi-quantitative assessment of autism symptoms. Following the deep phenotyping principle, which aims to gather information about disease manifestations in a more individual and granular way [22, 87], we can achieve a more fine-grained quantification of these symptoms. This approach holds the potential for delineating autism sub-types and developing personalized medicine for autistic individuals [7]. Currently, such fine-grained descriptions of autism symptoms however largely rely on labor-intensive manual annotations by experts. A better delineation could help to provide more individualized intervention strategies for autistic children. Indeed, current intervention strategies in younger autistic children rely on play-based behavioral intervention. Despite the general positive gain associated with such methods [27, 78], their benefits remain variable from one child to another [64, 85]. There is currently little knowledge on which type of intervention better works for each child [84]. We thus critically need to identify relevant autism subgroups that better respond to each intervention type. In this endeavor, we advocate deep phenotyping, in particular the one provided by digital phenotyping, which has the potential to provide us with meaningful autism subgroups. We thus aim to develop a comprehensive stratification strategy for automated quantification of a large variety of autistic symptoms, with a particular focus on preschool age.

3.2 Computational Perspective

Computer vision and behavior recognition methodology. The use of computer vision to obtain objective and quantitative standardized observations supporting the diagnosis of ASD children is relatively young. Earlier work used to rely on high-end screen-based eye trackers [43, 44, 67, 73]. Recently, research works have emerged to analyze children's behavior in more ecological settings, including real interactions [72]. Due to the prevalence of attention and emotional deficit in ASD children, a vast majority of research has been dedicated to the measurement of these two cues [2, 16, 65, 72]. Most earlier works used head pose as a proxy for gaze [18] and were facing the sensor placement issue with moving children. To alleviate this problem, wearable sensors have been proposed either by instrumenting the child with an eye-tracking sensor [58], facing issues with the compliance of the kids, or using a sensor worn by the experimenter/clinician [16], allowing for the extraction of more accurate attentional cues (*e.g.* eye contact). Regarding motor gesture analysis for ASD, [92] is the only work we are aware of, but it relies on a very contrived diagnosis pipeline (*i.e.* voice prompt, imposed starting hand pose and position), and with school-aged children (*i.e.* 10 years old).

In terms of settings, most research targeted standard face-to-face clinician-child or parent-child interaction sessions. Other diagnosis settings include social robots, seen as simpler interaction partners that can elicit desired behaviors to train or test children's responses [2]; Kinect sensors to measure social deficits in children with ASD [11]; or as an attempt to dramatically scale the screening process of ASD for children, mobile healthcare sensors combining affective, attention, and basic name-call stimuli head rotation reactions [39]. None of the above has considered natural interactions in open settings like ADOS, as we do.

Investigated strategies. In this project, we investigate two different approaches for autism analysis, as shown in Fig 2:
(1) one-stage detection from raw inputs;
(2) a compositional approach which first identifies (or exploits as a supervision signal) low-level cues (*e.g.* gestures, gaze), which can then be used to predict ADOS-2 score items and the ASD diagnosis.

The idea behind the first method is to operate directly on image sequences from ADOS-2 sessions or on standard modalities extracted from them (*e.g.* pose, depth) using off-the-shelf pre-trained models, and to classify whether the patient is ASD or TD (see Section 5.1 and [45]). While it offers the benefit of avoiding the need for behavioral annotations and discovering specific patterns, this approach sacrifices interpretability. There's also a risk that models trained this way might rely on contextual elements from the session environment (*e.g.* the clinician's attitude toward the child) that could indicate a diagnosis, rather than solely on the child's observable behaviors. The second approach adopts a two-step process where we first model relevant motor coordination behavioral patterns related to posture, gestures, gaze, attention and

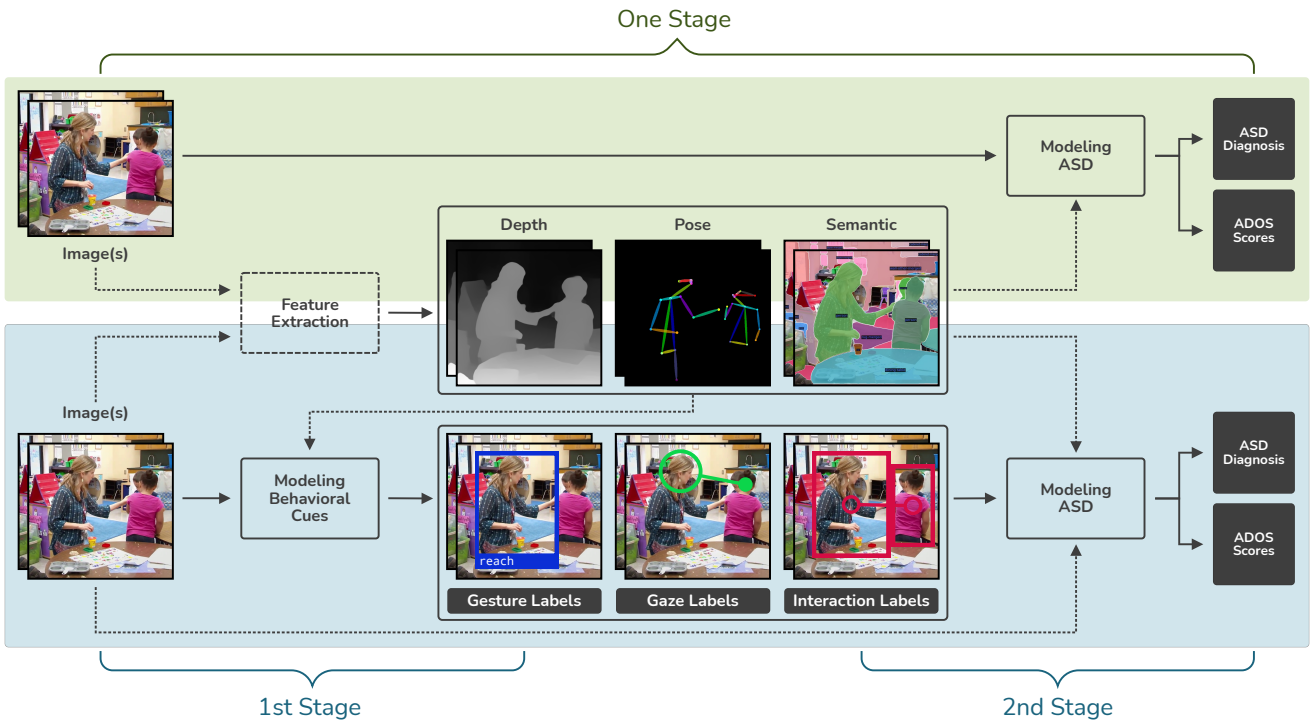

Figure 2: Two different approaches for ASD analysis, which can both be trained end-to-end. In green (top), directly inferring the ASD diagnosis and scores from multiple features extracted from the raw video sequences. In blue (bottom), the network architecture includes intermediate modeling of several behavioral cues, which can lead to more interpretable diagnoses.

communication, interactions or play activities, which are used in a second stage to infer the ASD binary label or build a relevant ASD profile from the ADOS-2 items. A major benefit is to enable the quantification of these behaviors, which in turn can help uncover semantically meaningful ASD profiles.

IoT sensing. One of the methodologies we investigated involves the analysis of play behaviors. Play, in fact, is an instinctive need for humans. It correlates with a healthy development process. Playing development and types of play – functional and symbolic – are strictly related to children's abilities and development. Many models exist to describe the relationship between children's play and their development [56, 61]. However, the literature focuses mostly on populations older than two years old [5] with a strong focus on the child's social, emotional, and cognitive development [14, 86], with little attention devoted to the sensory-motor aspect of play [80].

In this project, we focus on studying children's play behaviors from the perspective of sensory-motor developments: we target very small children, exploit a non-invasive innovative approach based on toys' inertial measurements, UWB-based localization, and video analysis, and collect statistically significant data to provide meaningful results directly applicable in clinical practice. This approach allows for a more fine-grained, objective, and automated analysis of play behavior, converging in the recognition of play patterns relevant to the analysis of children's neurodevelopmental disorders. Our goal is to exploit this IoT-based play analysis methodology as a support for traditional social and health research. It provides the technological tools allowing quantitative answers to questions like: "How and for how long does a child interact with a toy?", and "How do the child-toy interactions evolve with time?", hence deriving precise children play behavioral models and measures that can be related to clinical data.

4 DATASETS

Datasets are the cornerstone of our research. In this section, we introduce the different datasets that we have collected, along with their purposes and annotation protocols. Given our aims, on the one hand, we need clinical data in sufficient amounts to assess methods as well as research novel ASD stratifications. To this end, we mainly rely on the Geneva Autism cohort (see description in the introduction and Section 4.1). On the other hand, we need data with behavioral annotations in order to train machine learning models. To address the latter, we rely on both publicly available benchmarks (ChildPlay dataset, Section 4.3) which can be disseminated for research, as well as autism-relevant data, as behaviors exhibited in such cases can be pathological (see sections 4.1).

4.1 UniGe ADOS Dataset

Motivation. ADOS, the current gold-standard diagnostic assessment for autism, relies on manual expert coding and provides a semi-quantitative assessment of autism symptoms. As our goal is to investigate the development of a fully automated methodology to better quantify autism symptoms following the principles of deep phenotyping, we aim to more densely annotate such ADOS sessions to investigate our research questions, from both the clinical

perspective (see ASD stratification in aim 1 and Section 3.1) and behavior computational perspective (see aim 3 and Section 3.2).

Data Collection. Children included in this project are assessed within the context of the Geneva Autism Cohort, [31, 47, 68]. which employs a comprehensive phenotyping strategy combining standardized clinical tools with neuroimaging and digital phenotyping tools. Within this framework, autistic symptoms are evaluated using the ADOS protocol [26, 50, 51]. The ADOS comprises different modules specifically designed to elicit and observe autism symptoms at various developmental and language levels of individuals. These modules assess approximately 30 specific behaviors and assign scores on a scale ranging from 0 (no evident symptom) to 3 (very strong evidence of symptomatology). The coding process is closely intertwined with the assessment, as clinicians take notes during the evaluation, and the coding progression actively shapes the session flow. The ultimate objective is to obtain a fully scored session based solely on live interaction.

Annotation Protocol. Taking inspiration from the ADOS coding scheme, we designed an annotation protocol focusing on behaviors primarily characterized by nonverbal components, which is well suited to be used with computer vision tools. In the ADOS, coding thresholds for each behavior are determined based on their frequency, with more stringent thresholds applied to behaviors that are less prevalent in the general population and vice versa. For instance, for high-frequency behaviors like the initiation of joint attention, the absence of at least one clear example of a three-step joint attention behavior (involving visual contact and/or vocalization or gesture) within a single ADOS session is considered indicative of a symptom in this area. Once a child demonstrates such behavior or multiple repeated instances, the ADOS does not differentiate further in its scoring. With our adapted annotation protocol, our aim is to enhance the measurement granularity by documenting all instances of a given behavior, along with their duration. A video annotation sheet sample is provided in the appendix (see Fig. 9). The annotation is carried out using the tool BORIS [32].

4.2 The Geneva Pose for Autism dataset

Motivation. To demonstrate the capacity of computer vision tools to support automated identification, we built a dataset using solely the information contained in the body pose of the participants in the social interaction. Note that in this way, the data are anonymized, thus overcoming the sensitive question of personal data protection.

Data. For this purpose, we used available ADOS videos collected between 2013 and 2020 in the Geneva Autism Cohort. Sixty-eight children with autism (2.8 ± 0.93 years) and 68 typically developing children (2.55 ± 0.97 years) were included, divided into two equal and balanced training and test sets. We used the multi-person 2D pose estimation OpenPose technology [13] to extract the pose of all people present in the room (child, parent, clinician) and generate videos of only the skeletal keypoints, which formed our dataset (see Fig. 4A and B), which constitutes the data. We refer the interested reader to our previously published work [45] for more details.

Annotations. As these videos are coming from the Geneva Autism Cohort, following an ADOS protocol, each video comes from the corresponding behavioral annotations in a 0-3 range, as well as a global ASD vs TD score.

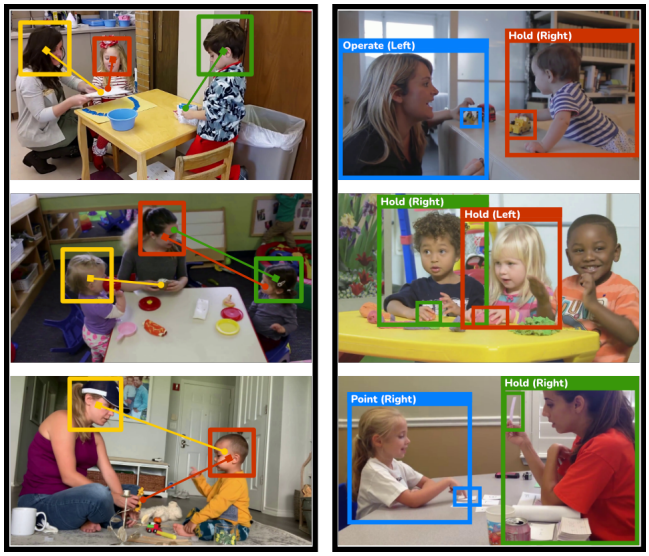

Figure 3: Samples from the ChildPlay dataset overlaid with annotations. Left: head box and gaze point annotations. Right: human-object interaction annotations. ChildPlay is the first public benchmark for analyzing children's gaze and interaction behaviors in free-play environments.

4.3 ChildPlay Dataset

Motivation. The ChildPlay dataset (see Fig. 3, and [81] for details) is a set of videos featuring children in free-play environments interacting with their surroundings. The dataset is rich in unprompted social behaviors, communicative gestures, and interactions. It is publicly available[1] and features high-quality dense gaze annotations, including a gaze class to account for special scenarios that arise in 2D gaze following. They can also be used to model other attention-related behaviors like shared attention, gaze shifts, eye contact, and fixations. To the best of our knowledge, this is the first representative gaze dataset aiming to cover children.

Data Collection. We relied on the YouTube video search engine with queries like "children playing toys", "childcare center", or "kids observation" to retrieve videos matching our aims. The scene context of our videos ranged from childcare facilities and schools to homes and therapy centers. We obtained a dataset of 401 clips, mainly restricted to indoor environments, showing 1 or 2 adults and multiple children. The age group varies from toddlers to preteenagers. The dominant activity of children is "playing with toys", but the dataset also includes a few clips containing other interactions such as behavioral therapy exercises.

Annotation Protocol. In every clip, we selected up to 3 people and for each of them, in every frame, we annotated the head bounding box, a 2D gaze point, and a gaze label. We also provide the person class label (adult vs. child). The gaze label addresses an important limitation with existing datasets, in which annotating a 2D gaze point is enforced in every frame, with only a standard inside vs. outside label to denote if the person looks within the frame or not [19]. However, there are many situations where annotating 2D gaze points is highly challenging, if not impossible. To avoid this, our gaze label was defined to include 7 non-overlapping classes

[1] Childplay is available at https://www.idiap.ch/en/dataset/childplay-gaze

to account for special scenarios: inside-frame, outside-frame, gaze-shift, occluded, eyes-closed, uncertain, not-annotated.

We are also exploring extra possible layers of annotation related to gestures and interactions with objects such as holding, operating, and pointing (see right column of Fig. 3). This will allow the joint modeling of gaze and interactions, which is instrumental in analyzing motor-gaze coordination patterns.

Comparison with the literature. The Multimodal Dyadic Behavior (MMDB) dataset [72], the Self-Stimulatory Behaviors dataset (SSBD) [69], DREAM [8] and 3D-AD [74] are all datasets meant to tackle different aspects of autism, be it stimming behaviors (arm flapping, head banging), speech and vocalizations, communicative gestures (e.g. pointing, reaching, etc.), or gaze patterns (e.g. shared attention). However, they are either anonymized, limited in terms of behaviors, or restricted to lab environments (e.g. screening or therapy sessions). In contrast, ChildPlay boasts a higher diversity of scenes, people, gestures, viewing angles, and lighting conditions.

5 METHODS AND ANALYSIS

5.1 End-to-end ASD Detection from Raw Inputs

This approach relates to the first category described in Fig. 2. That is, given only videos of the pose of people, can we directly identify ASD vs TD children without identifying atomic behavioral cues? The method is described in more detail in [45].

Model architecture. To this end, we relied on the Geneva pose dataset (see Sec. 4.2). We split the videos of this dataset into 5-second segments to train a neural network. This network relied on a VGG16 convolutional neural network pre-trained on Imagenet to extract high-dimensional semantic features from individual frames, and use them as input to an LSTM temporal model. The trained model was applied to make predictions on 5s video segments of the test set, which were further aggregated over the entire duration of the video for each subject to make a global prediction.

Results. Using the above network, we reached a prediction accuracy of 80.9% and a F1 score of 0.82 on the unseen test set (sensitivity: 85.3%, specificity: 76.5%). Further tests on an independent and unbalanced set of 106 preschoolers (105 with ASD) led to a similar level of accuracy (0.81) demonstrating the robustness of our approach. We further observed that the confidence probability for ASD obtained was highly correlated with many clinical parameters (severity of autism symptoms measured with the ADOS, $p<0.001$; measures of level of daily functioning, $p<0.001$; cognitive functioning as measured with best estimate IQ, $p=0.012$). Strikingly, the ASD probability was also highly correlated with many individual items from the ADOS, particularly those related to the coordination between gaze and gestures (see Fig.4C). We further tested how much the prediction accuracy was influenced by the video length, and the results demonstrate that an average 0.7 accuracy is already obtained with 10 min video segments (see Fig.4D). The prediction consistency is also very high across the video of a single individual even with relatively short video segments where for instance, for half of the ASD samples, our method achieves a 100% consistency in prediction based on randomly selected segments summing up to only 10 minutes. These important results demonstrate the feasibility of video-based automated identification of autism symptoms.

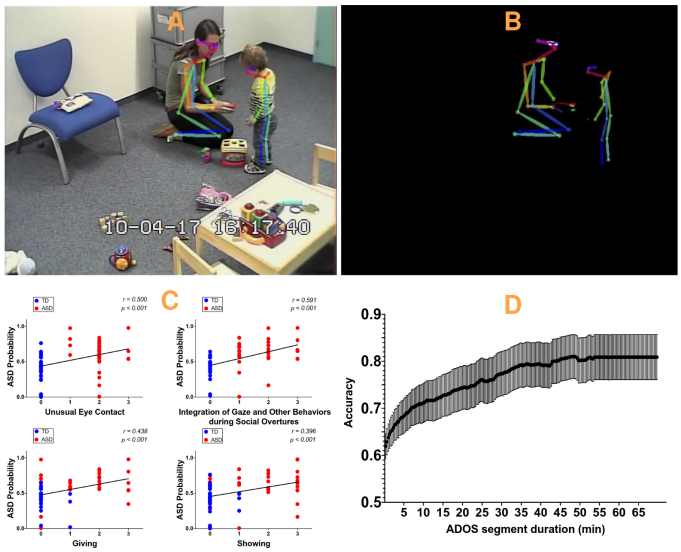

Figure 4: Using OpenPose Reconstruction during the ADOS scene to generate body pose images (A, B). ASD Probability obtained from a neural network based on the Pose Estimation highly correlates with specific behaviors as coded in the ADOS set-up by trained clinicians (C). The accuracy in the prediction increases with the duration of videos, with the final accuracy being 81% in our sample. Assessment of the stability in the predictions (D).

In the context of this Sinergia proposal, we aim to push the boundaries of this type of analysis, by increasing the accuracy in detecting autism over even shorter video duration, and by using the technology to provide deep phenotyping of autism signs.

5.2 Attention Modeling

Motivation. Understanding attention behaviors is a key component for autism diagnosis, and we have investigated this topic. Prior works have focused on estimating gaze directions or proxies for it [34, 49], and from there, people's Visual Focus of Attention (VFOA), defined as looking at a specific person or object [3, 33, 60, 76], but such methods required access to frontal views of people and/or knowledge of the 3D scene structure and were contrived to specific face-to-face interaction settings. To analyze gaze in more general scenes, we have followed [71], which introduced the Gaze Following task, defined as predicting the gaze target of a person (*i.e.* defined as 2D coordinates in the image). Predictions of all people can then be further processed to infer their social gaze patterns. Below we detail prior works in this direction, as well as ours.

Gaze Following from Images. Typical gaze following methods have a two-branch architecture (such as our approach, see Fig. 5): the first one processes the gaze information of a person (typically, a head crop and the person's image location), and its output is processed by the second branch along with the full image to identify salient items related to the person's gaze and generate a heatmap highlighting the candidate visual attention target of the person [17, 36, 42, 48, 57, 90]. More recent works have incorporated additional modalities like temporal information [19], depth [4, 29],

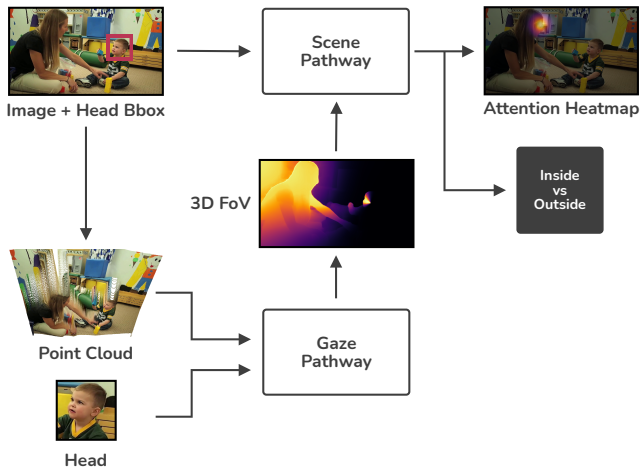

Figure 5: Our proposed gaze following architecture [81]. The Gaze Pathway processes the head crop of a person to predict a 3D gaze vector, which is used with the inferred point cloud to generate the person's 3DFoV map. It is processed along with the image and a head location mask by the Scene Pathway to predict the Attention Heatmap and the In-Out gaze label (*i.e.* whether the person is looking inside the frame or outside).

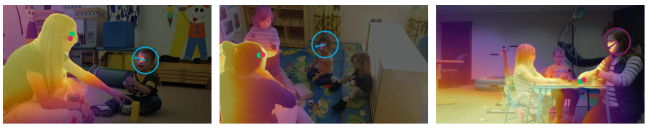

Figure 6: Qualitative results of our method on ChildPlay. The 3DFoV highlights potential gaze targets and excludes other salient items (like people) with non-matching depth. The GT gaze target is given in red and the prediction in green.

or a combination of pose and depth as in our work, which had set the state-of-the-art in gaze following benchmarks [38].

Depth, in particular, is important for ruling out salient items that lie along a person's 2D line of sight but are not visible to the person in 3D space. As standard gaze following datasets do not come with depth information, some works [4, 29] relied on depth (more specifically, disparity) estimated from pre-trained models [70, 88]. However, due to unknown camera parameters, the latter suffer from stretched and distorted reconstructed 3D scenes (point clouds) which are not so suitable for 3D geometric scene analysis.

To address this issue, we recently proposed [81] a more geometrically consistent approach, leveraging a depth estimation algorithm [63] yielding geometry-preserving point clouds, and learning to predict a 3D gaze vector congruent with the estimated point cloud, and allowing to predict the 3D Field of View (3DFoV) of a person. Fig. 5 presents our approach, while Fig. 6 shows qualitative results of our method on ChildPlay samples.

Social Gaze Prediction. Predicting gaze 2D locations is not very informative for autism analysis (and other social tasks), where obtaining attention labels is more relevant. In particular, three social gaze tasks are of general interest:

- *LAEO,* Looking at Each Other (or eye contact): binary label indicating whether a pair of people look at each other;

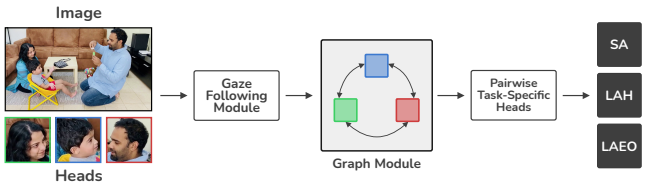

Figure 7: Social gaze inference architecture. Gaze representations from a Gaze Following Module outputs are used by a Graph Module to jointly model all people interactions. Then Task-Specific Network Heads predict the social gaze relation for each person pair.

Model	P_{LAH} Children ↑	P_{LAH} Adults ↑
Gupta [38]	0.435	0.621
Tafasca [81]	0.509	0.681
Gupta* [38]	0.648	0.731
Tafasca* [81]	0.604	0.704

Table 1: LAH performance of gaze following methods on ChildPlay. *indicates models fine-tuned on ChildPlay.

- *SA,* Shared Attention (or joint attention): binary label indicating whether a pair of people look at the same item;
- *LAH,* Looking at Head: binary label indicating whether a person looks at another person's face/head.

LAEO was first studied in computer vision by [55]. Their model used geometric information to predict LAEO. Since then, several deep learning-based works have been proposed to solve this task [12, 24, 53, 54], including the use of a gaze following approach [37]. Regarding the SA task, authors in [28] were the first to study it. Their method combined the predicted 2D gaze cones of people in the scene with a heatmap of object region proposals. A more recent work [79] improved over this by directly inferring shared attention from the raw image. Several gaze following methods [19, 83] also evaluated their method on the SA task and seemed to provide better performance than methods meant specifically for shared attention detection (protocols were not exactly the same).

Despite its importance, the LAH task had never been proposed in the context of analyzing general scenes. Indeed, LAH can be viewed as a generalization of LAEO, indicating one-way eye contact. This is relevant for autism diagnosis as children may look at the clinician and not vice-versa, and detecting such cases could help provide an extended characterization of children's gaze behavior. We thus introduced it in [81], and processed gaze following benchmark datasets to obtain LAH annotations. Sample performance (accuracy of looking or not at a person) is shown in Table 1. We see that models trained on standard gaze following datasets, which consist mostly of adults, provide poor LAH performance for children, and improve significantly after fine-tuning on ChildPlay. This highlights (i) the difficulty of the task, with performance reaching only up to 0.7, and (ii) the need for children-specific datasets like ChildPlay to obtain gaze models more applicable to children.

Finally, note that we recently developed a social gaze prediction methodology, summarized in Fig. 7 for the 3 tasks, which addressed previous limitations by processing all people jointly, leading to state-of-the-art performance.

Figure 8: Toy kit samples (with ADOS and AutoPlay toys).

5.3 Toy-Embedded Sensors

Motivation. As mentioned earlier, although play behavior is a natural demonstration of children's neurodevelopment, only a few works have investigated the sensory-motor aspect of play in the IoT literature. Developing an ASD digital phenotyping method calls for defining manipulation measurement methodologies enabling an autonomous and unbiased play observation, which is privacy-preserving and non-intrusive for the children. Additionally, children's gestures and play activities are intrinsically different from human activities well-studied in literature: the difference between fine-grained and coarse-grained activity drives the need for new methodologies for human activity recognition. Below we describe how we addressed these challenges: measurement of autonomous play activity and recognition of play-related fine-grained activities.

Autonomous Play. Our methodology exploits inertial and indoor-localization sensors to measure the movements of toys manipulated by children, allowing the inference of the driving manipulation activities during play in a privacy-preserving and non-intrusive manner. The measurements rely on the AutoPlay toys-kit that we had developed, which includes commonly used toys (e.g. ball, car, cube, doll, spoon) [9, 30] and extended to incorporate the ADOS toys (see Fig. 8): in both cases, sensors are embedded within the toys in a child-safe way. Each toy embeds one sensor node, comprising both an Inertial Measurement Unit (IMU) and an Ultra-Wide Band sensor (UWB) recording the 3D position and orientation of the embedding object in a room. The toys allow for the analysis of the main sensory-motor classes of play: mouthing, simple manipulation, functional, relational, and functional-relational [66]. The measurement of play behaviors with different settings (ADOS augmented toys and ADOS protocol, AutoPlay toys-kit for autonomous play) allows for the comparison of play behaviors analysis in autonomous situations versus in a semi-standardized protocol, and can provide clinicians with indicators related to children's motor behavior (i.e. giving gesture) assessment, and validates the potential scaling of the IoT based methodology in a non-controlled scenario. The augmented ADOS setting has been chosen because ADOS is the gold standard for autism diagnosis. AutoPlay and ADOS data allow us to model autonomous play behavior by extracting behavioral quantitative features from the observations and investigate which of them and of play behavior patterns are specifically associated with autism, in a way that could help remote screening.

Fine-Grained Human Activity Recognition. Human manipulation-related activities are often composed of sequences of micro-movements whose level of granularity is in the order of the milli-second. To recognize those micro-movements and activities, we use supervised learning methodologies applied to the inertial and 3D indoor localization (UWB) data collected from the toy-embedded sensors, using as labels video-based annotation of play activities related to the children-toys interactions. An initial list of micro-activities for analyzing play behaviors has been identified [30] through an initial study involving typically developing children in the age of 9 to 18 months: functional activities such as drag, stack, and push; exploratory activities like bite and knock; rotational activities like overturn and rollover. In [62], we demonstrated the intrinsic difference between coarse-grained (widely investigated in literature) and fine-grained (characterizing children's motor behavior) human activities, showing the performance drop (around 10% for f1-score) of the well-established methodologies from the Human Activity Recognition literature when applied to children play activities, demonstrating the need for dedicated approaches which we will further investigate during this project.

6 CONCLUSION AND FUTURE WORK

This paper is an effort to review the AI4Autism project and contrast our work with the available literature. In essence, it is an interdisciplinary research collaboration aiming to develop tools and strategies to ease the autism screening process, gain a better understanding of ASD phenotypes, and scale the solutions to less controlled environments where they can make a tangible impact.

Our work so far has focused on several key areas. First, collecting the necessary data to enable any further analyses and modeling, be it datasets annotated with relevant behavioral cues (*e.g.* gaze, gestures), or recordings of screening sessions annotated according to the ADOS-2 protocol. Second, preliminary attempts at the end-to-end one-stage prediction of ASD based on raw inputs, and specifically, pose skeletons. Third, modeling of gaze as an intermediate step for ASD recognition. In particular, we have focused on the gaze-following task and its possible extensions to localize the gaze target and introduce informative semantic components (*e.g.* eye-contact). Lastly, we explored the use of inertial and localization sensors embedded in toys to study play activities in children from a sensory-motor perspective.

There are several remaining steps that we plan to tackle next: (i) finalize data collection and annotations; (ii) on the computational side, incorporate the temporal component in our gaze methods, move to gestures and interactions modeling, and leverage IoT data coupled with visual streams to develop multimodal activity recognition methods; (iii) on the autism side, study ASD predictions from intermediate cues, and once we have a large enough sample of ADOS-2 annotations, study how these behavioral quantities shape patients' profiles and assess whether distinct clusters emerge. Finally, we plan to integrate our methods together into a solution to be tested in a different and more challenging environment.

ACKNOWLEDGMENTS

This research is supported by the AI4Autism project (Digital Phenotyping of Autism Spectrum Disorders in children, grant agreement no. CR- SII5 202235 / 1) of the Sinergia interdisciplinary program of the SNSF.

REFERENCES

[1] DS American Psychiatric Association, American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. Vol. 5. American psychiatric association Washington, DC.

[2] Salvatore Maria Anzalone, Jean Xavier, Sofiane Boucenna, Lucia Billeci, Antonio Narzisi, Filippo Muratori, David Cohen, and Mohamed Chetouani. 2019. Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment. *Pattern Recognition Letters* 118 (2019), 42–50.

[3] Chongyang Bai, Srijan Kumar, Jure Leskovec, Miriam Metzger, Jay Nunamaker, and V. S. Subrahmanian. 2019. Predicting the Visual Focus of Attention in Multi-Person Discussion Videos. In *Proceedings of the International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 4504–4510.

[4] Jun Bao, Buyu Liu, and Jun Yu. 2022. ESCNet: Gaze Target Detection With the Understanding of 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14126–14135.

[5] Emma Baumgartner. 2002. *Il gioco dei bambini*. Carocci.

[6] David Q. Beversdorf and Missouri Autism Summit Consortium*. 2016. Phenotyping, Etiological Factors, and Biomarkers: Toward Precision Medicine in Autism Spectrum Disorders. *Journal of Developmental & Behavioral Pediatrics* 37, 8 (Oct. 2016), 659–673. https://doi.org/10.1097/DBP.0000000000000351

[7] David Q Beversdorf and Missouri Autism Summit Consortium. 2016. Phenotyping, etiological factors, and biomarkers: toward precision medicine in autism spectrum disorders. *Journal of Developmental and Behavioral Pediatrics* 37, 8 (2016), 659.

[8] Erik Billing, Tony Belpaeme, Haibin Cai, Hoang-Long Cao, Anamaria Ciocan, Cristina Costescu, Daniel David, Robert Homewood, Daniel Hernandez Garcia, Pablo Gómez Esteban, et al. 2020. The DREAM Dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy. *PloS one* 15, 8 (2020), e0236939.

[9] Niko Bonomi and Michela Papandrea. 2021. Non-intrusive and Privacy Preserving Activity Recognition System for Infants Exploiting Smart Toys. In *EAI International Conference on IoT Technologies for HealthCare*. Springer, 3–18.

[10] Sarabeth Broder-Fingert, Emily Feinberg, and Michael Silverstein. 2018. Improving Screening for Autism Spectrum Disorder: Is It Time for Something New? *Pediatrics* 141, 6 (June 2018). https://doi.org/10.1542/peds.2018-0965 Publisher: American Academy of Pediatrics Section: Commentary.

[11] Ian Budman, Gal Meiri, Michal Ilan, Michal Faroy, Allison Langer, Doron Reboh, Analya Michaelovski, Hagit Flusser, Idan Menashe, Opher Donchin, et al. 2019. Quantifying the social symptoms of autism using motion capture. *Scientific Reports* 9, 1 (2019), 7712.

[12] Giorgio Cantarini, Federico Figari Tomenotti, Nicoletta Noceti, and Francesca Odone. 2022. HHP-Net: A light Heteroscedastic neural network for Head Pose estimation with uncertainty. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*. 3521–3530.

[13] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv:1812.08008 [cs]* (May 2019). http://arxiv.org/abs/1812.08008 arXiv: 1812.08008.

[14] Kimberly LH Carpenter, Jordan Hahemi, Kathleen Campbell, Steven J Lippmann, Jeffrey P Baker, Helen L Egger, Steven Espinosa, Saritha Vermeer, Guillermo Sapiro, and Geraldine Dawson. 2021. Digital behavioral phenotyping detects atypical pattern of facial expression in toddlers with autism. *Autism Research* 14, 3 (2021), 488–499.

[15] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. 2017. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–20. https://doi.org/10.1145/3131902 arXiv:arXiv:1902.00607v1

[16] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. 2017. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–20.

[17] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. 2018. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*. 383–398.

[18] Eunji Chong, Audrey Southerland, Abhijit Kundu, Rebecca M Jones, Agata Rozga, and James M Rehg. 2017. Visual 3d tracking of child-adult social interactions. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 399–406.

[19] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. 2020. Detecting Attended Visual Targets in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5396–5406.

[20] National Research Council et al. 2011. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. (2011).

[21] Geraldine Dawson and Guillermo Sapiro. 2019. Potential for Digital Behavioral Measurement Tools to Transform the Detection and Diagnosis of Autism Spectrum Disorder. *JAMA pediatrics* 173, 4 (2019), 305–306. https://doi.org/10.1001/jamapediatrics.2018.5269

[22] Cathryn M Delude. 2015. Deep phenotyping: the details of disease. *Nature* 527, 7576 (2015), S14–S15.

[23] Mieke Dereu, Petra Warreyn, Ruth Raymaekers, Mieke Meirsschaut, Griet Pattyn, Inge Schietecatte, and Herbert Roeyers. 2010. Screening for Autism Spectrum Disorders in Flemish Day-Care Centres with the Checklist for Early Signs of Developmental Disorders. *Journal of Autism and Developmental Disorders* 40, 10 (Oct. 2010), 1247–1258. https://doi.org/10.1007/s10803-010-0984-0

[24] Bardia Doosti, Ching-Hui Chen, Raviteja Vemulapalli, Xuhui Jia, Yukun Zhu, and Bradley Green. 2021. Boosting Image-based Mutual Gaze Detection using Pseudo 3D Gaze. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 2 (May 2021), 1273–1281. https://doi.org/10.1609/aaai.v35i2.16215

[25] Susan Dosreis, Courtney L. Weiner, Lakeshia Johnson, and Craig J. Newschaffer. 2006. Autism spectrum disorder screening and management practices among general pediatric providers. *Journal of developmental and behavioral pediatrics: JDBP* 27, 2 Suppl (April 2006), S88–94. https://doi.org/10.1097/00004703-200604002-00006

[26] Amy N. Esler, Vanessa Hus Bal, Whitney Guthrie, Amy Wetherby, Susan Ellis Weismer, and Catherine Lord. 2015. The Autism Diagnostic Observation Schedule, Toddler Module: Standardized Severity Scores. *Journal of Autism and Developmental Disorders* 45, 9 (2015), 2704–2720. https://doi.org/10.1007/s10803-015-2432-7 Publisher: Springer US ISBN: 0162-3257.

[27] Annette Estes, Jeffrey Munson, Sally J Rogers, Jessica Greenson, Jamie Winter, and Geraldine Dawson. 2015. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* 54, 7 (2015), 580–587.

[28] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. 2018. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6460–6468.

[29] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. 2021. Dual Attention Guided Gaze Target Detection in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11390–11399.

[30] Francesca D Faraci, Michela Papandrea, Alessandro Puiatti, Stefania Agustoni, Sara Giulivi, Vincenzo D'Apuzzo, Silvia Giordano, Flavio Righi, Olmo Barberis, Evelyne Thommen, et al. 2018. AutoPlay: a smart toys-kit for an objective analysis of children ludic behavior and development. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 1–6.

[31] Martina Franchini, Daniela Zöller, Edouard Gentaz, Bronwyn Glaser, Hilary Wood de Wilde, Nada Kojovic, Stephan Eliez, and Marie Schaer. 2018. Early adaptive functioning trajectories in preschoolers with autism spectrum disorders. *Journal of pediatric psychology* 43, 7 (2018), 800–813.

[32] Olivier Friard and Marco Gamba. 2016. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in ecology and evolution* 7, 11 (2016), 1325–1330.

[33] K. Funes, L. Nguyen, D. Gatica-Perez, and J.-M. Odobez. 2013. A Semi-Automated System for Accurate Gaze Coding in Natural Dyadic Interactions. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI), Sydney*.

[34] K. Funes and J.-M Odobez. 2016. Gaze Estimation in the 3D Space Using RGB-D sensors: Towards Head-Pose And User Invariance. *Int. Journal of Computer Vision* 118, 2 (june 2016), 194–216.

[35] Rebecca Grzadzinski, Themba Carr, Costanza Colombi, Kelly McGuire, Sarah Dufek, Andrew Pickles, and Catherine Lord. 2016. Measuring Changes in Social Communication Behaviors: Preliminary Development of the Brief Observation of Social Communication Change (BOSCC). *Journal of Autism and Developmental Disorders* 46, 7 (July 2016), 2464–2479. https://doi.org/10.1007/s10803-016-2782-9

[36] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. 2020. Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*. Springer, 502–513.

[37] Hang Guo, Zhengxi Hu, and Jingtai Liu. 2022. MGTR: End-to-End Mutual Gaze Detection with Transformer. In *Proceedings of the Asian Conference on Computer Vision*. 1590–1605.

[38] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. 2022. A Modular Multimodal Architecture for Gaze Target Prediction: Application to Privacy-Sensitive Settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 5041–5050.

[39] Jordan Hashemi, Geraldine Dawson, Kimberly LH Carpenter, Kathleen Campbell, Qiang Qiu, Steven Espinosa, Samuel Marsan, Jeffrey P Baker, Helen L Egger, and Guillermo Sapiro. 2018. Computer vision analysis for quantification of autism risk behaviors. *IEEE Transactions on Affective Computing* 12, 1 (2018), 215–226.

[40] Angie WS Ip, Lonnie Zwaigenbaum, David Nicholas, and Raphael Sharon. 2015. Factors influencing autism spectrum disorder screening by community paediatricians. *Paediatrics & Child Health* 20, 5 (June 2015), e20–e24. https://doi.org/10.1093/pch/20.5.e20 Publisher: Oxford Academic.

[41] Yvette M. Janvier, Jill F. Harris, Caroline N. Coffield, Barbara Louis, Ming Xie, Zuleyha Cidav, and David S. Mandell. 2016. Screening for autism spectrum disorder in underserved communities: Early childcare providers as reporters. *Autism: The International Journal of Research and Practice* 20, 3 (April 2016), 364–373. https://doi.org/10.1177/1362361315585055

[42] Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda Hu. 2021. Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 01–08.

[43] Ami Klin. 2008. In the eye of the beholden: tracking developmental psychopathology. *Journal of the American Academy of Child & Adolescent Psychiatry* 47, 4 (2008), 362–363.

[44] Ami Klin, David J Lin, Phillip Gorrindo, Gordon Ramsay, and Warren Jones. 2009. Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature* 459, 7244 (2009), 257–261.

[45] Nada Kojovic, Shreyasvi Natraj, Sharada Prasanna Mohanty, Thomas Maillart, and Marie Schaer. 2021. Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Scientific Reports* 11, 1 (2021), 15069.

[46] Kenneth Larsen, Astrid Aasland, and Trond H. Diseth. 2018. Identification of Symptoms of Autism Spectrum Disorders in the Second Year of Life at Day-Care Centres by Day-Care Staff: Step One in the Development of a Short Observation List. *Journal of Autism and Developmental Disorders* 48, 7 (July 2018), 2267–2277. https://doi.org/10.1007/s10803-018-3489-x

[47] Kenza Latreche, Nada Kojovic, Martina Franchini, and Marie Schaer. 2021. Attention to face as a predictor of developmental change and treatment outcome in young children with autism spectrum disorder. *Biomedicines* 9, 8 (2021), 942.

[48] Dongze Lian, Zehao Yu, and Shenghua Gao. 2018. Believe It or Not, We Know What You Are Looking At!. In *Asian Conference on Computer Vision*. Springer, 35–50.

[49] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. 2018. A Differential Approach for Gaze Estimation with Calibration. In *British Machine Vision Conference 2018, BMVC 2018*.

[50] Catherine Lord, Pamela C DiLavore, Katherine Gotham, Whitney Guthrie, Rhiannon J Luyster, Susan Risi, Michael Rutter, and Western Psychological Services (Firm). 2012. *Autism diagnostic observation schedule: ADOS-2*. Western Psychological Services, Los Angeles, Calif. OCLC: 851410387.

[51] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter. 2000. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders* 30, 3 (June 2000), 205–223.

[52] Mathew Maenner and et al. 2023. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *Morbidity and mortality weekly report. CDC surveillance summaries / Centers for Disease Control* 72, 2 (2023).

[53] Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. 2019. LAEO-Net: Revisiting People Looking at Each Other in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[54] Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. 2021. LAEO-Net++: revisiting people Looking At Each Other in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). https://doi.org/10.1109/TPAMI.2020.3048482

[55] Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. 2014. Detecting people looking at each other in videos. *International Journal of Computer Vision* 106 (2014), 282–296.

[56] Shelley Mulligan and Barbara Prudhomme White. 2012. Sensory and motor behaviors of infant siblings of children with and without autism. *The American Journal of Occupational Therapy* 66, 5 (2012), 556–566.

[57] Zhixiong Nan, Jingjing Jiang, Xiaofeng Gao, Sanping Zhou, Weiliang Zuo, Ping Wei, and Nanning Zheng. 2021. Predicting Task-Driven Attention via Integrating Bottom-Up Stimulus and Top-Down Guidance. *IEEE Transactions on Image Processing* 30 (2021), 8293–8305.

[58] Basilio Noris, Jacqueline Nadel, Mandy Barker, Nouchine Hadjikhani, and Aude Billard. 2012. Investigating gaze of children with ASD in naturalistic settings. (2012).

[59] Alyssa J. Orinstein, Molly Helt, Eva Troyb, Katherine E. Tyson, Marianne L. Barton, Inge-Marie Eigsti, Letitia Naigles, and Deborah A. Fein. 2014. Intervention for optimal outcome in children and adolescents with a history of autism. *Journal of developmental and behavioral pediatrics: JDBP* 35, 4 (May 2014), 247–256. https://doi.org/10.1097/DBP.0000000000000037

[60] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Kohler. 2018. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 191–199.

[61] Jools Page, Cathy Nutbrown, and Ann Clare. 2013. *Working with babies and children: From birth to three*. Sage.

[62] Shalini Pandurangan, Michela Papandrea, and Mirko Gelsomini. 2022. Fine-Grained Human Activity Recognition-A new paradigm. In *Proceedings of the 7th International Workshop on Sensor-based Activity Recognition and Artificial Intelligence*. 1–8.

[63] Nikolay Patakin, Anna Vorontsova, Mikhail Artemyev, and Anton Konushin. 2022. Single-Stage 3D Geometry-Preserving Depth Estimation Model Training on Dataset Mixtures With Uncalibrated Stereo Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1705–1714.

[64] Adrienne Perry, Julie Koudys, Alice Prichard, and Hilda Ho. 2019. Follow-up study of youth who received EIBI as young children. *Behavior Modification* 43, 2 (2019), 181–201.

[65] Rosalind W Picard. 2009. Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3575–3584.

[66] H Kathleen Pierce. 2009. *Exploratory, functional, and symbolic play behaviors of toddlers with autism spectrum disorders*. The Florida State University.

[67] Karen Pierce, David Conant, Roxana Hazin, Richard Stoner, and Jamie Desmond. 2011. Preference for geometric patterns early in life as a risk factor for autism. *Archives of general psychiatry* 68, 1 (2011), 101–109.

[68] Irène Pittet, Nada Kojovic, Martina Franchini, and Marie Schaer. 2022. Trajectories of imitation skills in preschoolers with autism spectrum disorders. *Journal of Neurodevelopmental Disorders* 14 (2022), 1–13.

[69] Shyam Rajagopalan, Abhinav Dhall, and Roland Goecke. 2013. Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 755–761.

[70] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020).

[71] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. 2017. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*. 1435–1443.

[72] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. 2013. Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3414–3421.

[73] Katherine Rice, Jennifer M Moriuchi, Warren Jones, and Ami Klin. 2012. Parsing heterogeneity in autism spectrum disorders: visual scanning of dynamic social scenes in school-aged children. *Journal of the American Academy of Child & Adolescent Psychiatry* 51, 3 (2012), 238–248.

[74] Omar Rihawi, Djamal Merad, and Jean-Luc Damoiseaux. 2017. 3D-AD: 3D-autism dataset for repetitive behaviours with kinect sensor. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[75] Guillermo Sapiro, Jordan Hashemi, and Geraldine Dawson. 2019. Computer vision and behavioral phenotyping: an autism case study. *Current Opinion in Biomedical Engineering* 9 (2019), 14–20. https://doi.org/10.1016/j.cobme.2018.12.002

[76] 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.

[77] Dean P. Smith, Diane W. Hayward, Catherine M. Gale, Svein Eikeseth, and Lars Klintwall. 2019. Treatment Gains from Early and Intensive Behavioral Intervention (EIBI) are Maintained 10 Years Later. *Behavior Modification* (Oct. 2019), 145445519882895. https://doi.org/10.1177/0145445519882895

[78] Dean P Smith, Diane W Hayward, Catherine M Gale, Svein Eikeseth, and Lars Klintwall. 2021. Treatment gains from early and intensive behavioral intervention (EIBI) are maintained 10 years later. *Behavior modification* 45, 4 (2021), 581–601.

[79] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2020. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3327–3336.

[80] Satoshi Suzuki, Yukie Amemiya, and Maiko Sato. 2019. Enhancement of gross-motor action recognition for children by CNN with OpenPose. In *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, Vol. 1. IEEE, 5382–5387.

[81] 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 (ICCV)*.

[82] Cora M. Taylor, Alison Vehorn, Hylan Noble, Amy S. Weitlauf, and Zachary E. Warren. 2014. Brief Report: Can Metrics of Reporting Bias Enhance Early Autism Screening Measures? *Journal of Autism and Developmental Disorders* 44, 9 (Sept. 2014), 2375–2380. https://doi.org/10.1007/s10803-014-2099-5

[83] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. 2022. End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2192–2200.

[84] Giacomo Vivanti, Margot Prior, Katrina Williams, and Cheryl Dissanayake. 2014. Predictors of outcomes in autism early intervention: why don't we know more? *Frontiers in pediatrics* 2 (2014), 58.

[85] Zachary Warren, Melissa L McPheeters, Nila Sathe, Jennifer H Foss-Feig, Allison Glasser, and Jeremy Veenstra-VanderWeele. 2011. A systematic review of early intensive intervention for autism spectrum disorders. *Pediatrics* 127, 5 (2011), e1303–e1311.

[86] Peter Washington, Catalin Voss, Aaron Kline, Nick Haber, Jena Daniels, Azar Fazel, Titas De, Carl Feinstein, Terry Winograd, and Dennis Wall. 2017. SuperpowerGlass: a wearable aid for the at-home therapy of children with autism. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–22.

[87] Chunhua Weng, Nigam H Shah, and George Hripcsak. 2020. Deep phenotyping: embracing complexity and temporality—towards scalability, portability, and interoperability. *Journal of biomedical informatics* 105 (2020), 103433.

[88] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. 2020. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569* (2020).

[89] Dajie Zhang, Iris Krieber-Tomantschger, Luise Poustka, Herbert Roeyers, Jeff Sigafoos, Sven Bölte, Peter B. Marschik, and Christa Einspieler. 2019. Identifying Atypical Development: A Role of Day-Care Workers? *Journal of Autism and Developmental Disorders* 49, 9 (Sept. 2019), 3685–3694. https://doi.org/10.1007/s10803-019-04056-3

[90] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. 2020. Learning to draw sight lines. *International Journal of Computer Vision* 128, 5 (2020), 1076–1100.

[91] Andrea Zunino, Pietro Morerio, Andrea Cavallo, Caterina Ansuini, and Jessica Podda. 2018. Video Gesture Analysis for Autism Spectrum Disorder Detection. In *ICPR*.

[92] Andrea Zunino, Pietro Morerio, Andrea Cavallo, Caterina Ansuini, Jessica Podda, Francesca Battaglia, Edvige Veneselli, Cristina Becchio, and Vittorio Murino. 2018. Video gesture analysis for autism spectrum disorder detection. In *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 3421–3426.

[93] Lonnie Zwaigenbaum, Margaret L. Bauman, Roula Choueiri, Connie Kasari, Alice Carter, Doreen Granpeesheh, Zoe Mailloux, Susanne Smith Roley, Sheldon Wagner, Deborah Fein, Karen Pierce, Timothy Buie, Patricia A. Davis, Craig Newschaffer, Diana Robins, Amy Wetherby, Wendy L. Stone, Nurit Yirmiya, Annette Estes, Robin L. Hansen, James C. McPartland, and Marvin R. Natowicz. 2015. Early Intervention for Children With Autism Spectrum Disorder Under 3 Years of Age: Recommendations for Practice and Research. *Pediatrics* 136, Supplement 1 (Oct. 2015), S60–S81. https://doi.org/10.1542/peds.2014-3667E

A APPENDIX: ADOS ANNOTATION

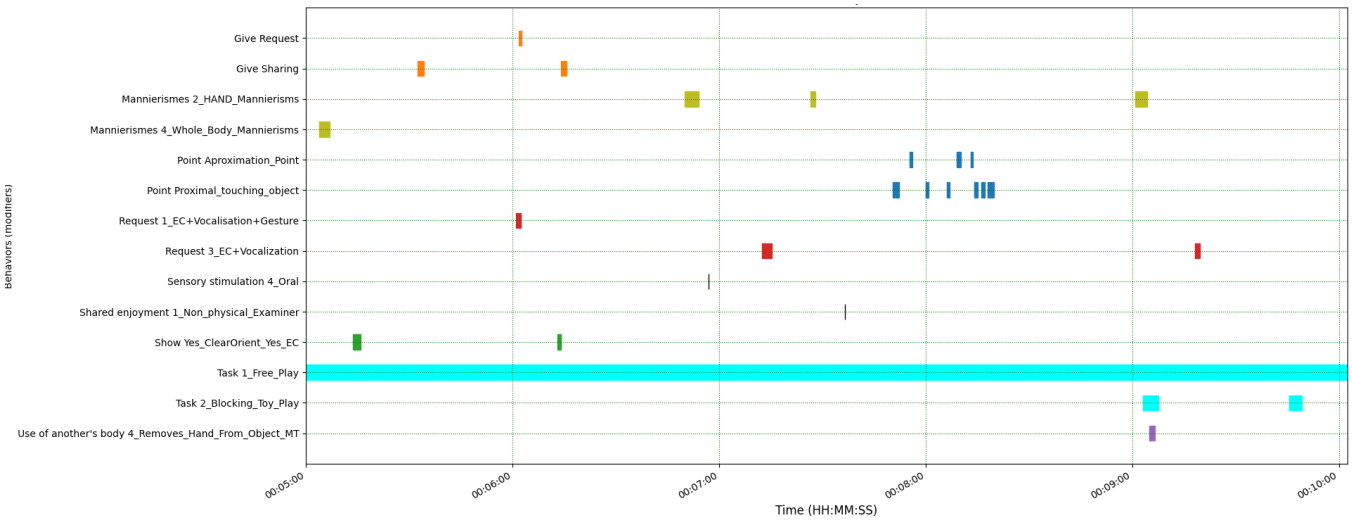

Figure 9: A sample annotation sheet covering 5 minutes of an ADOS session. We can see the duration of different coded behaviors (*e.g.* request using eye contact and vocalization) and ADOS activities (*e.g.* free play).