2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

# Evaluating Gaze Detection for Children with Autism Using the ChildPlay-R Dataset

Nursena Boluk[1] and Hatice Kose[2]

[1] Faculty of Computer and Informatics Engineering, Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey

[2] Faculty of Computer and Informatics Engineering, Department of Artificial Intelligence and Data Engineering, Istanbul Technical University, Istanbul, Turkey

*Abstract*— **This study focuses on detecting two-dimensional (2D) gaze points in interactions between children both with and without autism and adults. Within the scope of this objective, a novel subset called ChildPlay-R has been created based on the open-source ChildPlay dataset. This new dataset includes explicitly the interactions of children with and without autism with adults in similar uncontrolled environments and relevant 2D gaze points. Moreover, the analyses using the Modified Spatiotemporal Gaze Detection (M-STGD) model, compared to the traditional Spatiotemporal Gaze Architecture (STGA), reveal significant improvements in the Area Under the Curve values and out-of-sample precision. These advancements indicate the M-STGD model's potential to provide a deeper understanding of the social engagement patterns of children with autism. This study contributes to the automatic detection of 2D gaze points in social interactions and supports the social development of children with autism.**

## I. INTRODUCTION

From the earliest moments of life, gaze behavior emerges as a powerful and nonverbal form of communication, marking one of the fundamental elements of interactions [3]. From this early period, gaze behavior plays a significant role in social interactions. Individuals with autism, however, experience difficulties in social interactions compared to their typically developing peers. These challenges are displayed in less eye contact, atypical gaze patterns, and a reduced response to their names by individuals with autism [4].

Eye contact is crucial in social interactions, signifying attention, engagement, and interest. Consequently, the evaluation of social communication skills in children at risk for developmental disorders has become critical through gaze studies [8]. Therapeutic activities that seriously impact the development of these social skills in children with autism take a long time [4]. Nevertheless, the manual annotation and evaluation of video recordings of these therapeutic activities are time-consuming and require considerable effort. Despite advancements in automatic gaze estimation from cameras [9], [12], [16] and wearable camera-based predictions [8], making this process more efficient and reducing the cost of video coding remains a challenging problem. The primary cause of this challenge lies in the scarcity of open-source interaction datasets involving children with autism [2], [14], [15]. Among these limited datasets, only one features gaze target labels [2]. However, incorporating 3D-processed data to protect participants' privacy, this dataset adds an extra layer of complexity to gaze studies. While crucial for maintaining confidentiality, this approach significantly complicates the analysis and interpretation of gaze data, presenting additional hurdles for researchers in this field.

In response to this problem, this study introduces a reduced version of the open-source ChildPlay dataset titled ChildPlay-R. This ChildPlay-R dataset contains fifteen videos, five describing interactions between children with autism and adults, and the remaining ten showing interactions involving children without autism and adults. The availability of such an immediately accessible dataset is expected to accelerate interaction studies with children with autism.

Furthermore, this study proposes a modified version of the STGA model, the M-STGD model. The STGA [9] and M-STGD models are compared using the ChildPlay-R dataset regarding interactions between children with and without autism and adults.

## II. RELATED WORK

In autism detection [13], [18] and therapy [4], [6], [10], a notable increase has been observed in the successful utilization of technological approaches that have garnered positive feedback. Among these technologies, wearable cameras, eye-tracking systems, cameras, and interactive robots are prominent.

Wearable cameras provide datasets for behavioral analysis, but their prolonged use can cause discomfort or distraction, particularly in sensitive children with autism. These devices may create a sense of pressure, inhibiting the children from displaying their natural behaviors [5].

Eye-tracking systems [17], [18] are used in critical areas such as eye contact and joint attention, which are pivotal in autism. However, the requirement for children to hold a fixed position can be challenging for children with autism [5], [17].

Video cameras [9], [13], [10], are less restrictive than other observed technologies and are highly effective in recording the natural behaviors of children [16]. These cameras do not require special hardware and do not exert additional pressure on the children. However, they face challenges in usability and generalizability [9], [16].

Robotic technology [1], [6], [10] offers interactive and educational tools to enhance the social skills of children with autism. With their programmable and predictable behaviors,

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

robots facilitate children's engagement in social interactions and are increasingly used in robot-assisted therapy studies [6], [10].

Technological tools are essential for improving our understanding and therapy of social interactions in children with autism. These tools enable the detailed study of crucial aspects such as mutual gaze, joint attention, and the recognition of emotional responses.

Mutual gaze [7], [11] involves two individuals making eye contact. Children with autism often struggle with eye contact, which poses challenges to their social development. Research in this area focuses on improving their social skills by enhancing their ability to engage in mutual gaze [5].

Joint attention [1], [6], [10] is the ability of two or more individuals to focus their attention together on the same point or object. This joint focus is another area that can be particularly challenging for individuals with autism. Studies using these technological tools aim to improve the ability of children with autism to focus their attention on specific objects or activities and thus improve their social interaction potential.

Lastly, the method of attention target detection is employed to ascertain the focus points [9], [12], [16] of individuals, helping researchers and educators better understand their interests and learning styles. This technique allows for a deeper understanding of their engagement patterns by identifying where a participant's visual attention is concentrated. Such insights are invaluable for personalizing educational content and therapeutic activities to align with the child's interests and learning needs, ultimately supporting more effective interventions in autism therapy.

## III. DATASET

This study utilized a reduced version of the ChildPlay dataset [16] named ChildPlay-R to examine social interactions between children (either with or without autism) and adults. The ChildPlay dataset [16] consists of short video clips labeled with rich gaze information that capture children engaging in play and interacting with adults. This dataset contains a selection of 401 video clips from 95 videos collected from open-source videos from various environments, such as therapy centers and kindergartens, via YouTube. The ChildPlay dataset was annotated by seven annotators, incorporating 2D gaze points and gaze labels. Gaze labels were segregated into seven categories: inside-frame, outside-frame, gaze-shift, occluded, eyes-closed, uncertain, and not-annotated. Frames labeled as occluded, eyes-closed, uncertain, and not-annotated were excluded from this study to sharpen the focus on different gaze behaviors [16].

The reduced dataset used in this research contains data that adheres to specific criteria: i) the interaction must involve one child, and ii) the interaction should occur around a table. These criteria were chosen to steer future research towards a more controlled study, enabling a concentrated examination of interactions and the clustering potential of the child's focus area. Under these criteria, 35 video clips extracted from 15 videos were selected for the reduced dataset. Of these videos, five featured interactions with children with autism. Autism-related information was explicitly indicated in the respective videos, sourced from an autism center, or related to autism studies. To the best of our knowledge, ChildPlay-R is the only instantly accessible open-source dataset that includes 2D gaze points label specifically designed for studying interactions between autistic children and adults, making it an important resource for researchers.

A "subject-independent" approach has been adopted by using leaving-one-actor-out method to enhance the general applicability of the model. Considering that more than one video can belong to the same child in the dataset, we picked children who are represented with only one video for the test set. This test set includes one video of a child with autism and two videos of children without autism, allocating 20% of the video for testing and 80% for training. Given the total number of frames, the videos of children with autism contain 5706 frames, whereas those of children without autism include 12532 frames. Consequently, in terms of the total frame count, the test set represents 16% of the overall frames.

Data augmentation techniques were also applied to increase the dataset's diversity and robustness. These techniques, which include color-changing, head position jittering, cropping, and flipping, were implemented randomly and independently of each other 50% of the time [16]. These color-changing operations were adjusted for brightness, contrast, and saturation. This strategy aimed to imitate a more comprehensive visual scenario, expanding the dataset and potentially enhancing the model's predictive accuracy across various environments.

## IV. METHOD

### A. *Spatiotemporal Gaze Architecture*

The Spatiotemporal Gaze Architecture, as cited in [9], contains three main parts: the scene branch, the head conditioning branch, and the recurrent attention prediction module. The general structure of the STGA system is shown in Figure 1.

The Spatiotemporal Gaze Architecture uses a head convolution process within the head conditioning branch, applying it to the image cropped based on the head's position. This process combines with the binary image of the head position, creating a head feature map.

The scene branch enhances the original image by adding a layer containing the binary image. After being subjected to the scene convolution process, this layered image results in the scene feature map. This scene feature map is multiplied by the attention map produced by the head conditioning branch, highlighting areas in the image based on the head's position. The resulting weighted scene feature map combines with the head feature map to enrich the features [9].

These enriched features are then processed through the encoder module. The Spatiotemporal Gaze Architecture then integrates temporal information using a convolutional long short-term memory network within the recurrent attention prediction module. The deconvolution module upscales the

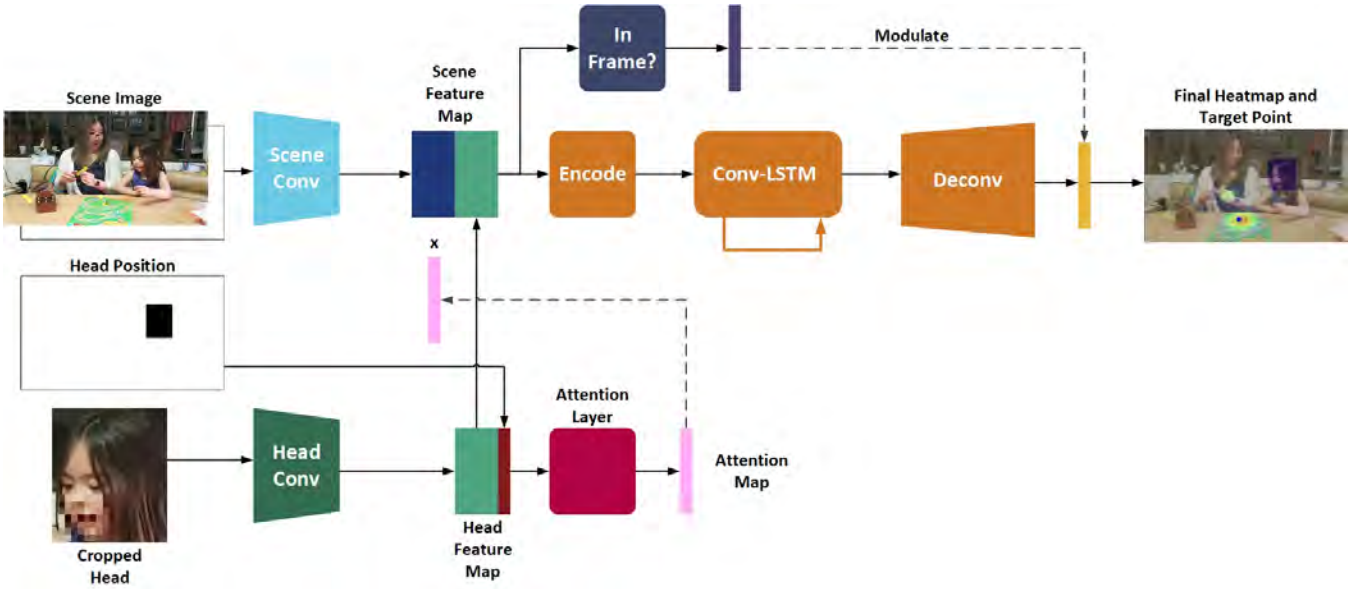


Fig. 1: General structure of the STGA system [9]

extracted features to a full-size feature map. This full-size feature map is then modulated with $\alpha$ to decide whether the possible 2D gaze point is within the image frame. Eventually, a heatmap predicting the attention target point in the image is generated [9].

### B. Modified Spatiotemporal Gaze Detection

This study introduces the Modified Spatiotemporal Gaze Detection (M-STGD) system, an adaptation of the STGA initially designed for adult 2D gaze point detection. The modification aims to extend the application of the model to determine 2D gaze points between children with autism and children without autism.

The number of neurons in the second layer was reduced from 128 to 64 to increase performance and efficiency in the Decoder, the last stage of the M-STGD architecture. Such an adjustment serves multiple purposes: it combats overfitting, reduces the model's complexity, enhances the accuracy of image reconstruction, and increases overall learning efficiency. Moreover, weights not associated with the batch normalizations in the deconvolution layers were frozen to ensure stability and prevent overfitting.

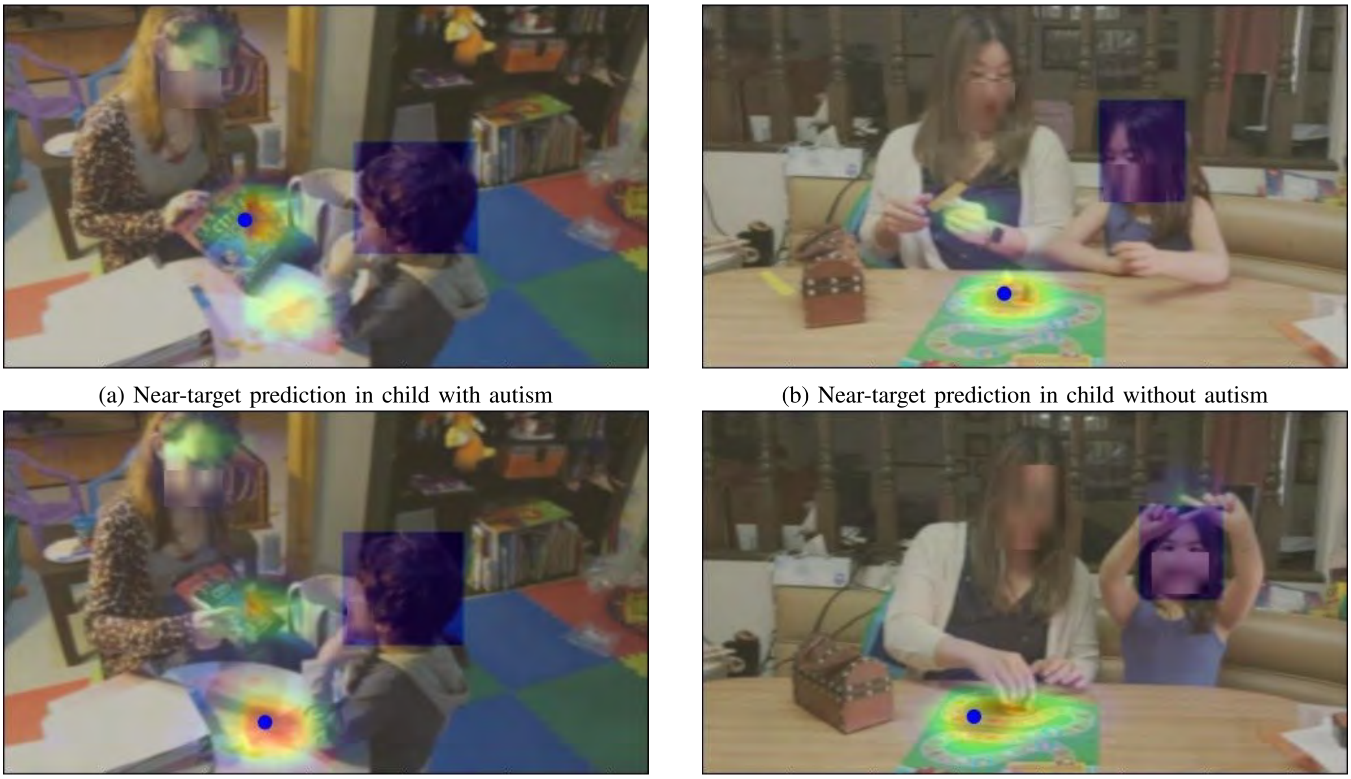
After detailed testing and optimization, the optimal hyperparameters were determined, comprising a learning rate of $5 \times 10^{-6}$, a training duration set to 2 epochs, the Adam optimizer for effective optimization, and a batch size of 16. This set of hyperparameters has been determined to be the best for achieving superior model performance and efficiency.

## V. RESULT

This study analyzes how the children's 2D gaze points during interactions are assessed by applying two models: the STGA and the M-STGD. The effectiveness of these models was determined using a set of performance metrics, namely the Area Under the Curve (AUC), the $L^2$ distance, and the Out-of-Frame Average Precision (AP) [9].

AUC is computed based on the classification of each cell in the feature image as potential 2D gaze points. The ground truth is a Gaussian confidence mask centered around the 2D gaze points annotated by labels. The heatmap illustrates the prediction confidence score at different thresholds on the ROC (Receiver Operating Characteristic). The heatmap loss, $\mathcal{L}_h$ (loss), is calculated using the Mean Squared Error loss (MSE) when the 2D gaze points are within the frame relative to the ground truth. Furthermore, the binary cross-entropy loss method is used to calculate the In-Frame loss, $\mathcal{L}_f$, with the overall training loss being a composite of both heatmap and In-Frame losses [9]: $\mathcal{L} = w_h \cdot \mathcal{L}_h + w_f \cdot \mathcal{L}_f$. The distance performance metric evaluates the $L^2$ distance between the actual 2D gaze point and the estimated maximum pixels of the heatmap [9]. Moreover, the AP for each image is r01ified by comparing it to its actual value using a scalar $\alpha$.

In this study, the STGA model, which is typically trained on adult data, was tested on the ChildPlay-R dataset for children with autism and without autism. For children with autism, the STGA model achieved an AUC of 79% and an L2 distance of 0.33. In contrast, for children without autism, the model recorded an AUC of 87% and an $L^2$ distance of 0.13. Subsequently, the weights from the STGA model were transferred to the M-STGD model, which was then trained on the ChildPlay-R dataset. This adjustment improved accuracy in determining the 2D gaze points for both groups of children. Specifically, for children with autism, the AUC decreased to 81%, and the L2 distance was reduced to 0.27. For children without autism, the AUC increased to 90%, and the $L^2$ distance experienced a slight increase of 0.0041. The combined performance metrics for the 2D gaze points detection across STGA and M-STGD models are summarized in Table I, highlighting the comparative analysis



(a) Near-target prediction in child with autism

(b) Near-target prediction in child without autism

(c) Off-target prediction in child with autism

(d) Off-target prediction in child without autism

Fig. 2: Comparative analysis of 2D gaze points prediction in child with and without autism

for each group.

TABLE I: Combined Performance Metrics for 2D Gaze Points Detection

| Model | Children w/ Autism | | | Children w/o Autism | | |
|---|---|---|---|---|---|---|
| | AUC↑ | $L^2$ D↓ | AP↑ | AUC↑ | $L^2$ D↓ | AP↑ |
| STGA | 0.7931 | 0.334 | 0.9747 | 0.874 | **0.1368** | 0.9739 |
| M-STGD | **0.8143** | **0.2713** | **0.9756** | **0.9099** | 0.1409 | **0.9751** |

Figure 2 shows correct and incorrect predictions of the M-STGD model in some sample scenes involving children with and without autism interacting with adults. The output of this model, which predicts the gaze target point of the child, is a heat map. This heat map provides a representation of where the child is looking in the image. A blue dot and a blue box have been added to the image for clarity in the demonstration. The blue dot represents the model's predicted attention target point and the blue box indicates the head position of the individual whose gaze target point has been predicted. Also, the individuals have been blurred to prevent identification for anonymization.

## VI. CONCLUSION

This study proposes a deep learning-based approach for detecting the 2D gaze points during social interactions between children with and without autism and adults to provide guiding information for developing effective intervention methods for children with autism who have difficulties in social interaction. For this purpose, using the ChildPlay open-source dataset, a subset was created containing videos of the child with and without autism playing and interacting with one or more adults. This dataset was analyzed using the STGA and the newly proposed M-STGD models. The analysis showed that the M-STGD model outperformed the STGA model regarding higher AUC values, improved $L^2$ distance outcomes, and increased out-of-frame average precision values. These findings suggest that the M-STGD model is more adapted to accurately analyzing interactions across both groups of children. The results of this research could contribute to a better understanding and support of interactions with children with autism. Improvements made in the M-STGD model may provide more accurate information for applications such as educational and therapeutic programs. In addition, this study emphasizes the advantage of using technological tools to analyze social interactions by detecting 2D gaze points.

## VII. ACKNOWLEDGMENTS

This study is part of the "TECH-TOYS: Acquire Digital Biomarkers in Infancy with Sensorized Toys for Early Detection and Monitoring of Neurodevelopmental Disorders" Project and is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the Grant Number 221N351. The authors thank TUBITAK for their support.

The authors extend their gratitude to the creators of the

ChildPlay dataset for their efforts in labeling children's interaction videos and making these labels available to the public.

## References

[1] S. M. Anzalone, J. Xavier, S. Boucenna, L. Billeci, A. Narzisi, F. Muratori, D. Cohen, and M. Chetouani. Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment. *Pattern Recognition Letters*, 118:42–50, 2019. Cooperative and Social Robots: Understanding Human Activities and Intentions.

[2] E. Billing, T. Belpaeme, H. Cai, H.-L. Cao, A. Ciocan, C. Costescu, D. David, R. Homewood, D. Hernandez Garcia, P. Gómez Esteban, H. Liu, V. Nair, S. Matu, A. Mazel, M. Selescu, E. Senft, S. Thill, B. Vanderborght, D. Vernon, and T. Ziemke. The dream dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy. *PLOS ONE*, 15(8):1–15, 08 2020.

[3] T. B. Brazelton, E. Tronick, L. Adamson, H. Als, and S. Wise. *Early Mother-Infant Reciprocity*, chapter 9, pages 137–154. John Wiley Sons, Ltd, 1975.

[4] A. Budarapu, N. Kalyani, and S. Maddala. Early screening of autism among children using ensemble classification method. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 162–169, 2021.

[5] J.-J. Cabibihan, H. Javed, M. Aldosari, T. W. Frazier, and H. Elbashir. Sensing technologies for autism spectrum disorder screening and intervention. *Sensors*, 17(1), 2017.

[6] D. Cazzato, P. L. Mazzeo, P. Spagnolo, and C. Distante. Automatic joint attention detection during interaction with a humanoid robot. In A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, editors, *Social Robotics*, pages 124–134, Cham, 2015. Springer International Publishing.

[7] O. Celiktutan, W. Wu, K. Vogeley, and A. L. Georgescu. A computational approach for analysing autistic behaviour during dyadic interactions. In J.-J. Rousseau and B. Kapralos, editors, *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, pages 167–177, Cham, 2023. Springer Nature Switzerland.

[8] E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R. M. Jones, A. Rozga, and J. M. Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3), sep 2017.

[9] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg. Detecting attended visual targets in video. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5395–5405, 2020.

[10] P. R. S. De Silva, K. Tadano, A. Saito, S. G. Lambacher, and M. Higashi. The development of an assistive robot for improving the joint attention of autistic children. In *2009 IEEE International Conference on Rehabilitation Robotics*, pages 694–700, 2009.

[11] Z. Guo, K. Kim, A. Bhat, and R. Barmaki. An automated mutual gaze detection framework for social behavior assessment in therapy for children with autism. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 444–452, New York, NY, USA, 2021. Association for Computing Machinery.

[12] K. Higuchi, S. Matsuda, R. Kamikubo, T. Enomoto, Y. Sugano, J. Yamamoto, and Y. Sato. Visualizing gaze direction to support video coding of social attention for children with autism spectrum disorder. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, page 571–582, New York, NY, USA, 2018. Association for Computing Machinery.

[13] J. Liu, Z. Wang, K. Xu, B. Ji, G. Zhang, Y. Wang, J. Deng, Q. Xu, X. Xu, and H. Liu. Early screening of autism in toddlers via response-to-instructions protocol. *IEEE Transactions on Cybernetics*, 52(5):3914–3924, 2022.

[14] S. S. Rajagopalan, A. Dhall, and R. Goecke. Self-stimulatory behaviours in the wild for autism diagnosis. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 755–761, 2013.

[15] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, C. Kim, H. Rao, J. C. Kim, L. L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye. Decoding children's social behavior. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3414–3421, 2013.

[16] S. Tafasca, A. Gupta, and J. Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20878–20889, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society.

[17] P. Venuprasad, T. Dobhal, A. Paul, T. N. M. Nguyen, A. Gilman, P. Cosman, and L. Chukoskie. Characterizing joint attention behavior during real world interactions using automated object and gaze detection. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, New York, NY, USA, 2019. Association for Computing Machinery.

[18] W. Zhou, M. Yang, J. Tang, J. Wang, and B. Hu. Gaze patterns in children with autism spectrum disorder to emotional faces: Scanpath and similarity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:865–874, 2024.