Exploring Aspects of Social Interaction using Machine Learning

Nele-Pauline Suffo[1], Pierre-Etienne Martin[2], Anam Zahra[2], Daniel Haun[2], & Manuel Bohn[1, 2]

[1] Institute of Psychology in Education, Leuphana University Lüneburg

[2] Max Planck Institute for Evolutionary Anthropology

Author Note

The authors made the following contributions. Nele-Pauline Suffo: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Manuel Bohn: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Nele-Pauline Suffo, Universitätsallee 1, 21335 Lüneburg. E-mail: nele.suffo@leuphana.de

Abstract

Childrens everyday experiences are known to shape childrens development but only few studies investigate how children actually spent their time at home in naturalistic setting. More particular we were interested in how children's social interactions with others or interactions with objects are observable in their everyday life. To do so, we utilized the Quantex Dataset, an egocentric video and audio dataset of children aged 3-5 years, to investigate the presence of persons, faces, gaze, and objects in children's everyday interactions. We trained a YOLO11 model to detect persons and faces in the videos and analyzed the presence of gaze and objects in the videos. We furthermore applied a pre-trained voice type classifier to detect speech in the audio data. Our results show that children's everyday interactions are characterized by the presence of persons and faces, with the child's gaze directed towards others in 60% of the interactions. Additionally, children interacted with objects in 40% of the videos, with toys being the most common object category. Agr group analysis revealed that children aged 3 years showed more interactions with objects compared to older children. Our findings provide insights into the diversity of children's everyday experiences and highlight the importance of multimodal data for understanding children's social interactions and engagement.

Keywords: Quantex Dataset, egocentric video, audio dataset, children, social interactions, object interactions, gaze, multimodal data, computer vision, audio analysis, developmental psychology

Exploring Aspects of Social Interaction using Machine Learning

Introduction

According to various developmental psychologists, children's everyday experiences play a vital role in their development (Carpendale & Lewis, 2020; Heyes, 2018; Piaget, 1964; Rogoff, Dahl, & Callanan, 2018; Smith, Jayaraman, Clerkin, & Yu, 2018; Tomasello, 2009; Vygotsky, 1978). Everyday interactions, in particular, have been recognized for decades as crucial in the process of actively constructing knowledge (Piaget, 1964) and in transforming sensory experiences into structured understanding (Vygotsky, 1978). Building upon these foundational theories, more recent research has examined the mechanisms of social interaction further. For instance, Tomasello (2009) introduced the concept of shared intentionality, illustrating how collaborative activities enable children to comprehend others' intentions and perspectives, leading to cooperative behaviors and cultural learning .

Whereas theoretical frameworks and controlled laboratory studies have significantly advanced our understanding of children's social development, they often fail to capture the complexities of interactions occurring in naturalistic settings. Observing children in their everyday environments offers a more authentic view of their social behaviors; however, this approach presents challenges due to the extensive data collection and analysis required.

To address these challenges, researchers have increasingly turned to data-driven approaches that utilize sensors and recording devices to gather objective data on social interactions. For instance, Onnela, Waber, Pentland, Schnorf, and Lazer (2014) employed wearable sensors to analyze social interactions in adult work settings, capturing the duration of close proximity between individuals. The study inferred that women were more talkative than men and more likely to be physically close to other women in group settings. Rossano et al. (2022) examined social interactions among 31 two- to four-year-olds using 563 hours of video and audio recordings from a preschool during free play sessions over seven days. Manual interaction labels revealed that four-year-olds engaged in more

cooperative social interactions and experienced fewer conflicts than two-year-olds, with object play and conversations being the most common forms of social engagement in both age groups. Dai et al. (2022) investigated social interactions of 174 preschool children over three years, collecting voice and proximity data using wearable wireless RFID tags to study the co-development of social interactions and language acquisition. They employed manually labeled interaction data to train a temporal segment model that automatically identified periods of free play or class play, concluding that classmates frequently engaged in both contexts. Lemaignan, Edmunds, Senft, and Belpaeme (2018) created a dataset comprising 45 hours of manually labeled social interactions between 45 child-child pairs and 30 child-robot pairs, including video and audio recordings, 3D facial data, skeletal information, and game interactions. By not providing specific instructions to the children, the researchers aimed to capture interactions in naturalistic settings. However, each laboratory session was limited to 40 minutes.

While these studies have advanced our understanding of social interactions, they often focus on controlled environments or are constrained by limited observation periods. Moreover, the manual data collection and analysis involved remain labor-intensive and time-consuming, and the current body of research lacks comprehensive data-driven studies analyzing children's social interactions within their home environments.

To overcome these limitations, our study investigates social interactions in naturalistic home settings over an extended period. We have created the **Quantex** dataset which currently includes 197.20 hours of egocentric video and audio recordings from children aged 3 to 5 years and enables the analysis of specific patterns of social interactions, including:

- Presence of Individuals: Utilizing YOLO11 for person detection to identify when others are present in the child's environment.
- Presence of Faces: Employing YOLO11 face detection to recognize faces the child encounters.

- Object Interactions: Analyzing the objects with which the child interacts using YOLO11 object detection
- Gaze Behaviors: Classifying gaze direction with YOLO11-cls to determine when others are looking at the child.
- Speech Dynamics: Implementing voice type classification to differentiate between the child's speech and that of others, distinguishing between peers and adults.

The primary objectives of this study are to quantify interaction patterns by measuring the frequency and duration of each identified interaction type, both individually and in combination. Additionaly we compare these patterns across different age groups within the 3 to 5-year range to identify developmental variations and milestones. Understanding these interaction patterns can inform developmental psychology about the actual nature of social interactions in children's everyday lives.

Methodology

This chapter outlines the methodology used in this study to collect, annotate, and analyze video and audio recordings of children's everyday interactions. The aim of the study is to investigate key aspects of social interactions and engagement, such as the presence of persons, faces, gaze direction, and objects the child interacts with. The following sections provide a detailed description of the data collection process, the structure and characteristics of the dataset, the annotation strategy, and the preprocessing applied to the data prior to analysis. Additionally, an overview of the automated analysis pipeline is provided, giving details about the models used for person and face detection, gaze classification, object detection, and the application of a pre-trained voice type classifier.

Data Collection

This study collected egocentric video recordings from 76 children, aged 3 to 5 years, over a span of 73 months. Participating families lived in a mid-sized city in Germany. Data collection is ongoing, and the number of children will continue to increase as the study progresses. The data collection process was approved by the local ethics committee, and all participating families provided written informed consent, allowing the researchers to use the data for scientific purposes. In accordance with data privacy regulations, every child was assigned a unique anonymized ID to protect their identity. Moreover, the video recordings are stored on a secure server and are only accessible to the research team, all of whom have signed confidentiality agreements.

To capture the children's everyday experiences, a wearable vest equipped with a *PatrolEyes WiFi HD Infrared Police Body Camera* was used (Figure 2). The camera recorded high-definition video (1920x1080p at 30 fps) with a 140-degree wide-angle lens and also captured audio. The children were free to move around and engage in their usual activities at home without any interference or instructions given to their parents.

As of now, the ongoing data collection process has resulted in a total of 503 video recordings, with a combined duration of 197.20 hours.

Dataset Overview

The Quantex dataset includes video and audio recordings from 76 children aged 3 to 5 years (M=4.53, SD=0.81). The dataset contains 167 videos from three-year-olds, 180 videos from four-year-olds, and 156 videos from five-year-olds. The number of videos per child varies, as parents decide when and how often to record. The recording duration per child ranges from 10.43 to 391.18 minutes (M=155.68, SD=82.62). The total duration of all video recordings in the dataset is 197.20 hours. Figure 1 shows the distribution of video duration per child.

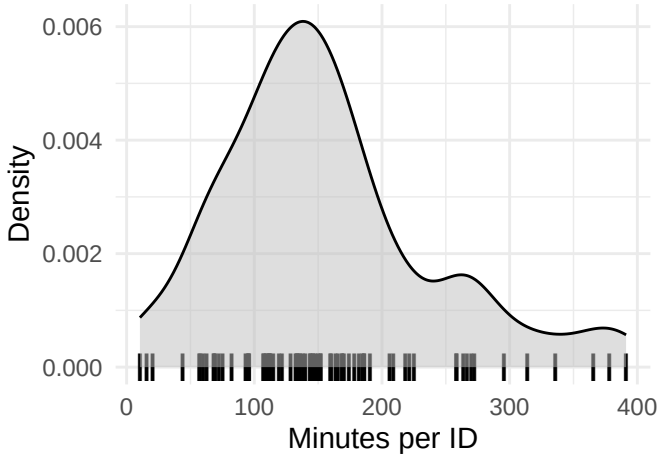

Figure 1. Video recording duration (in minutes) per Child in the Quantex Dataset.

Annotation Strategy

The dataset annotations cover four key elements: persons, faces, gaze direction, and objects the child interacts with. For each detected person (or reflection of a person, such as in a mirror) and face, additional attributes, such as age and gender, are collected. Gaze information indicates whether a detected person's gaze is directed toward the child or not. Faces are annotated even when occluded or blurry to ensure comprehensive coverage of interactions. Partially visible faces are also annotated if key facial features, such as the nose, eyes, or mouth, remain identifiable. Objects are annotated only when the child is actively interacting with them. These objects are categorized into six distinct groups: book, screen, animal, food, toy, and kitchenware, with an additional category for other objects. The annotation strategy is summarized in Figure 2.

The annotations were generated manually by a team of human annotators. Each video was randomly assigned to an initial annotator, and then reviewed by a second annotator to ensure consistency and accuracy. This peer review process helped to identify and resolve discrepancies, ensuring high-quality annotations.

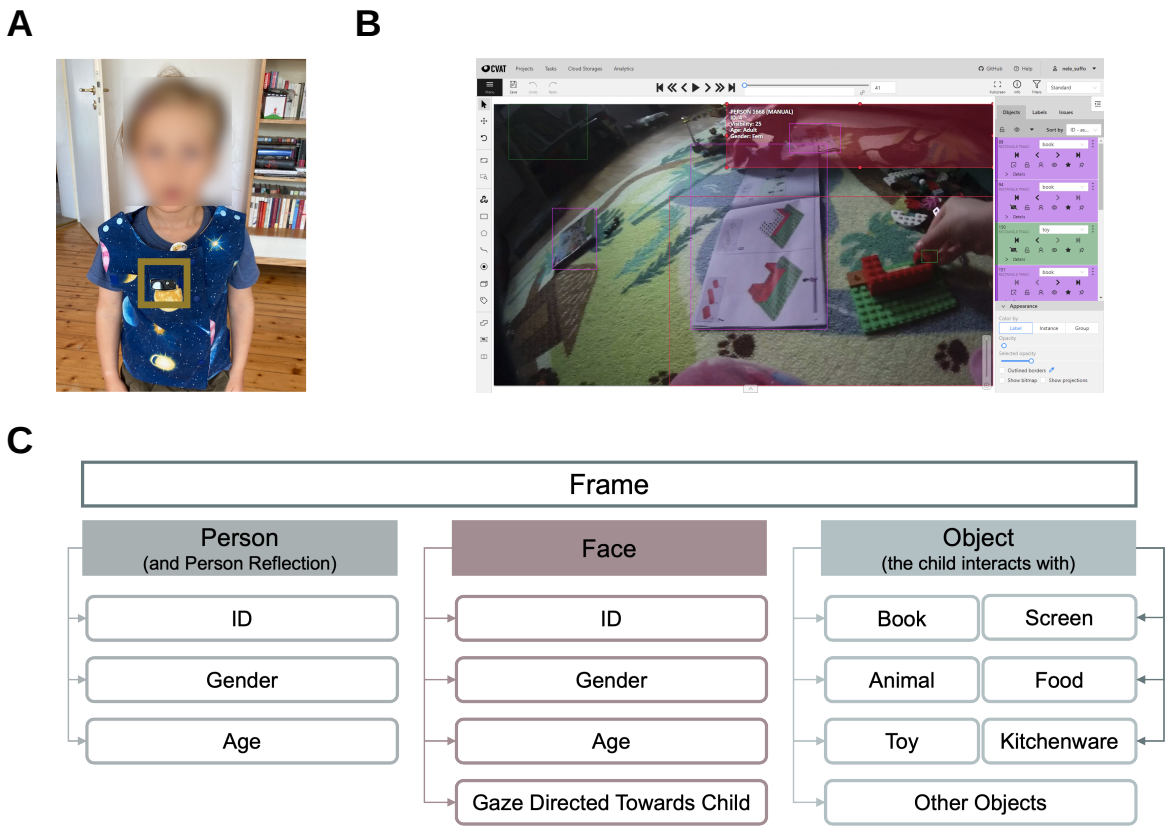

Figure 2. **A** – Vest with the embedded camera worn by the children, **B** – CVAT platform utilized for video annotation, **C** – Annotation Strategy in the Quantex dataset.

Data Preprocessing

For the video data, the annotation strategy required persons, faces, and objects to be labeled even when only partially visible, as long as key features such as facial landmarks (e.g., nose, eyes, or mouth) or parts of a person or object were clearly visible. To prepare the video data for analysis, one frame per second was annotated, corresponding to every 30th frame in the video. This frame sampling was chosen to balance the need for a representative sample of the video while keeping the analysis manageable. Similarly, every 30th raw frame was extracted from the annotated video files for further processing. No preprocessing was applied to the audio data, which was used in its raw form for analysis.

Automated Analysis Pipeline

Our automated analysis pipeline consists of four key modules: person and face detection, gaze classification, object detection, and voice type classification. Each module operates independently, utilizing separate machine learning models. Except for the voice type classifier, all models were trained on the Quantex dataset.

The pipeline follows a sequential process: 1. YOLO11x detection model identifies the presence of individuals (persons and faces) and types of objects the key child interacts with, both in social and independent play contexts, in the video frames. 2. Gaze classification determines whether detected faces are looking at the child. 3. Voice type classification detects the presence of speech and identifies whether the speaker is the key child, another child, or an adult.

By integrating these modules, our pipeline enables a comprehensive analysis of children's everyday experiences, capturing both social interactions and independent play

In the following sections, we describe each module in detail, including training data, model architecture, and evaluation metrics. A full technical analysis of each algorithm is provided in the Supplementary Material.

Person & Face Detection.

Gaze Classification.

Object Detection.

Voice Detection and Classification.

Results

Presence of Aspects of Social Interaction

Presence of a Person.

Presence of a Face.

Presence of Gaze.

Presence of Language.

Co-occurrence of Aspects of Social Interaction

General Discussion

References

Supplementary Material

YOLO11x: Multi-Class Detection of Persons, Faces, and Objects

In our study, we utilized Ultralytics' YOLO11, the "latest iteration in the Ultralytics YOLO series of real-time object detectors" (Jocher & Qiu, 2024), trained on the COCO dataset. Released in October 2024, YOLO11 introduces architectural improvements such as the C2PSA block (Convolutional Block with Parallel Spatial Attention), which enhances spatial attention within feature maps, allowing the model to focus more precisely on critical areas of an image compared to previous YOLO versions. Additionally, YOLO11 incorporates the C3K2 block, designed to be faster and more efficient, enhancing the overall performance of the feature aggregation process (Khanam & Hussain, 2024). These advancements make the YOLO11 detection model, pretrained on COCO, well-suited for training on our egocentric dataset, which captures dynamic movements from a camera perspective on chest height.

Dataset Preprocessing. Our dataset presents unique challenges due to its egocentric viewpoint, as the body parts of the child wearing the camera frequently appear in the footage. To prevent misclassification, we adopt a dedicated annotation scheme where all individuals in the scene are labeled as "person," each assigned a unique ID. The key child, who wears the camera, is consistently assigned ID = 1. During preprocessing, we map the key child (ID = 1) into a separate category, "child body parts," to distinguish their presence from other individuals. Furthermore, we refine the "person" and "face" categories by introducing additional distinctions that standard YOLO models do not inherently make. Specifically, we differentiate between (1) the key child (who wears the camera) and other individuals, (2) adults and children/infants for both full-body detections and faces. These modifications allow for a more precise analysis of social interactions while reducing false positives caused from the key child's own body while capturing detailed information on the presence and classification of people and objects in the child's

environment. To address these challenges, our fine-tuned YOLO11 model is trained to:

- Recognize and differentiate between the key child's body parts and other individuals.
- Distinguish between adults and children/infants for both full-body detections and faces.
- Identify and classify six specific object categories (toy, book, food, kitchenware, screen, other object) relevant to the child's interactions.

While our dataset originally included seven object categories, we merged "animal" and "food" into the "other object" category due to their low occurrences (19 and 1,115 instances, respectively), whereas all other categories had more than 2,000 instances. As a result, our final model was trained on five object categories.

Dataset Splitting. We started data splitting with a dataset comprising a total of 113799 frames from 80 annotated videos. Prior to splitting this dataset into training, validation, and testing datasets, we analyzed how often each class was present in the dataset. The class distribution, displayed in detail in table 1, revealed that the dataset was imbalanced, with the "neither" class (frames without any of the relevant classes) being the most frequent. To address this imbalance, we applied a stratified split to ensure that each dataset preserved the original class distribution. As a result, the final data distribution (see figure **??**) consisted of 91038 frames in the training dataset, 11379 frames in the validation and 11382 frames in the testing dataset, guaranteeing that the model's performance evaluation remained accurate to the real-world data distribution.

Training and Convergence. Model training was conducted on a Linux server equipped with an Intel(R) Xeon(R) Silver 4214Y CPU @ 2.20GHz with 48 cores, a Quadro RTX 8000 GPU and 188 GB of RAM. The model was trained for a total of 86 epochs, taking 200 hours to complete. Training utilized YOLO11's built-in data augmentation, an image size of 640, a batch size of 16, a cosine annealing learning rate scheduler (Loshchilov

Table 1

Dataset splits for the YOLO11 detection model trained on the Quantex dataset. The table shows the distribution of annotated persons, faces, and objects in the training, validation, and testing datasets.

Class	Training	Validation	Testing	Total
Adult	25706	3213	3213	32132
Child/Infant	22403	2801	2800	28004
Adult Face	669	1083	1084	10836
Child/Infant Face	6756	844	845	8445
Book	8370	1046	1046	10462
Toy	13870	1734	1733	17337
Kitchenware	1915	239	240	2394
Screen	3374	422	422	4218
Other Object	15608	1950	1950	19508
Neither	32341	4004	4005	40350

Table 2

Number of images in the training, validation, and testing datasets for the YOLO11 detection model.

	Training	Validation	Testing	Total
Number of images	91039	11380	11380	113799

& Hutter, 2017), and early stopping after 10 epochs without improvement, with a maximum of 200 epochs.

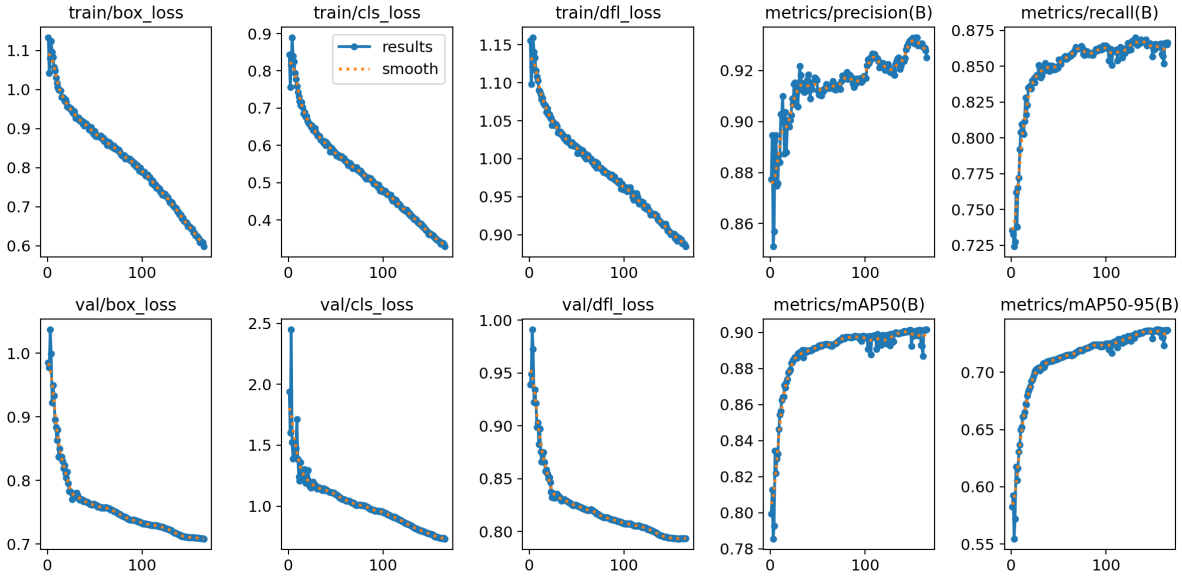

Figure 3. Training and Validation Loss Curves for the YOLO11x detection model.

The loss function of the YOLO11 model comprises three main components: Box Loss, Classification Loss, and Distribution Focal Loss (DFL) (Li et al., 2020; Terven, Cordova-Esparza, Ramirez-Pedraza, Chavez-Urbiola, & Romero-Gonzalez, 2024). *Box Loss* quantifies the difference between predicted bounding boxes and ground truth boxes, ensuring precise localization of detected persons, faces and objects by penalizing inaccuracies in position and size. *Classification Loss* evaluates the model's ability to correctly assign detected instances to their respective classes, reducing false positives and false negatives. *Distribution Focal Loss* enhances the model's ability to detect challenging persons, faces and objects, particularly small or partially occluded ones, by refining the localization of bounding box coordinates and emphasizing hard-to-detect instances. Together, these loss components contribute to a more robust and accurate detection model.

During the training process, we observed that all three loss components decreased over time, indicating effective learning and improved performance, as visible in in figure 3.

Table 3

Evaluation metrics for the YOLO11x detection model trained on the Quantex dataset to detect persons, faces and six object classes. False Positive Rate (FPR) and False Negative Rate (FNR) are given in percentages.

Precision	Recall	F1-Score	FPR	FNR
0.92	0.87	0.90	2.72	10.94

A steady decrease in Box Loss indicates that the model is becoming increasingly accurate in localizing persons, faces and objects within frames. Similarly, the steady convergence of the Classification Loss reveals the model's increasing ability to reliably classifiy the detected instance in one of the relevant classes. The decrease in DFL over time indicates that the model is getting better at focusing on and correctly identifying difficult-to-detect persons, faces or objects, which improves its overall detection capabilities. Conclusively, the loss curves show that the model effectively learned to localize and identify the target classes during the training period.

Model Evaluation Metrics. The performance of the object detection model was evaluated using a confusion matrix and precision-recall (PR) curves. The YOLO11 model achieved a precision of 0.89 and a recall of 0.79 on the testing set, resulting in an F1-score of 0.84. The normalized confusion matrix (see figure 4) reveals strong overall performance with mean Average Precision across all classes mAP=0.86. The averaged precision-recall curve remains close to the top-left corner and signifies that the model maintains high precision and recall across various thresholds, underscoring its effectiveness in detecting the different classes. Notably, the confusion matrix demonstrates minimal confusion between

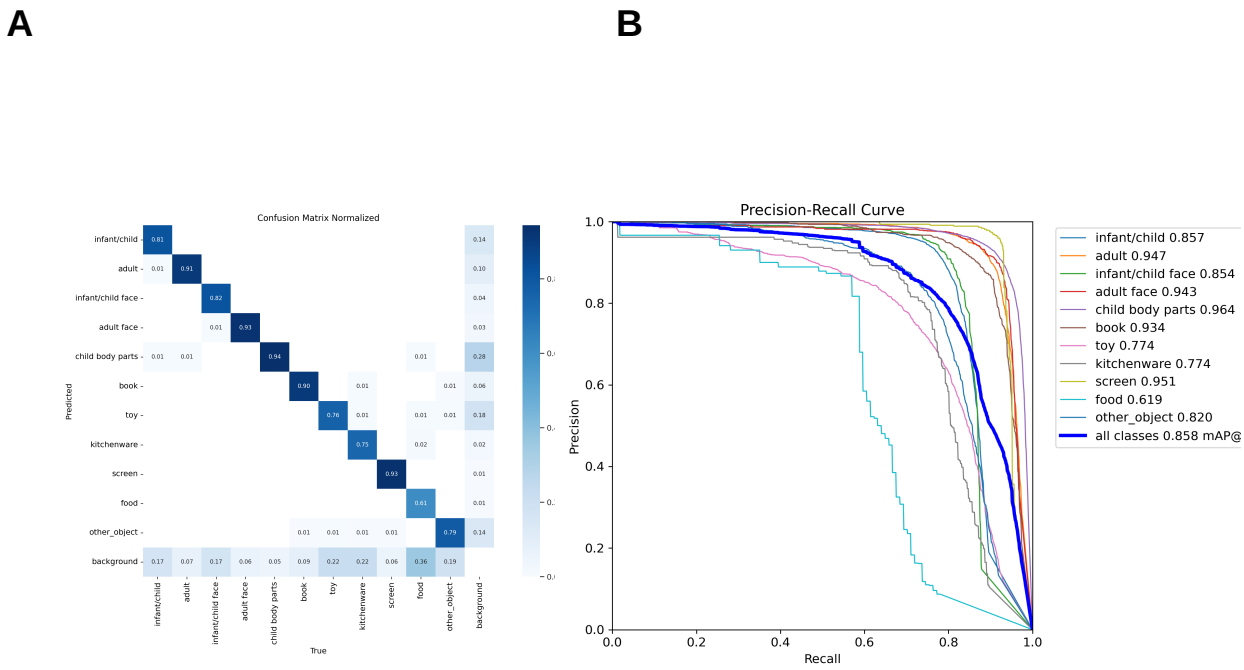

Figure 4. **A** - Confusion Matrix for the YOLO11x detection model trained on the Quantex dataset. **B** - Precision-Recall Curve for the YOLO11x detection model.

most classes, indicating that the model can effectively distinguish between them (see figure 4).

Analysis of the confusion matrix reveals that the classes “infant/child” (0.81), “adult” (0.91), “adult face” (0.91), “child body parts” (0.94), “infant/child face” (0.82), “book” (0.90) and “screen” (0.93) exhibit high Average Precision (AP) scores in the PR curves (see figure 4), signifying excellent precision and recall.

Contrarily, the “toy”, “kitchenware” and “other object” classes show lower performance, with AP scores of 0.77, 0.77, and 0.82, respectively. The confusion matrix supports this observation, indicating that 22% of “kitchenware” and 22% of “toy” items are frequently misclassified as “background”. This misclassification is likely due to the

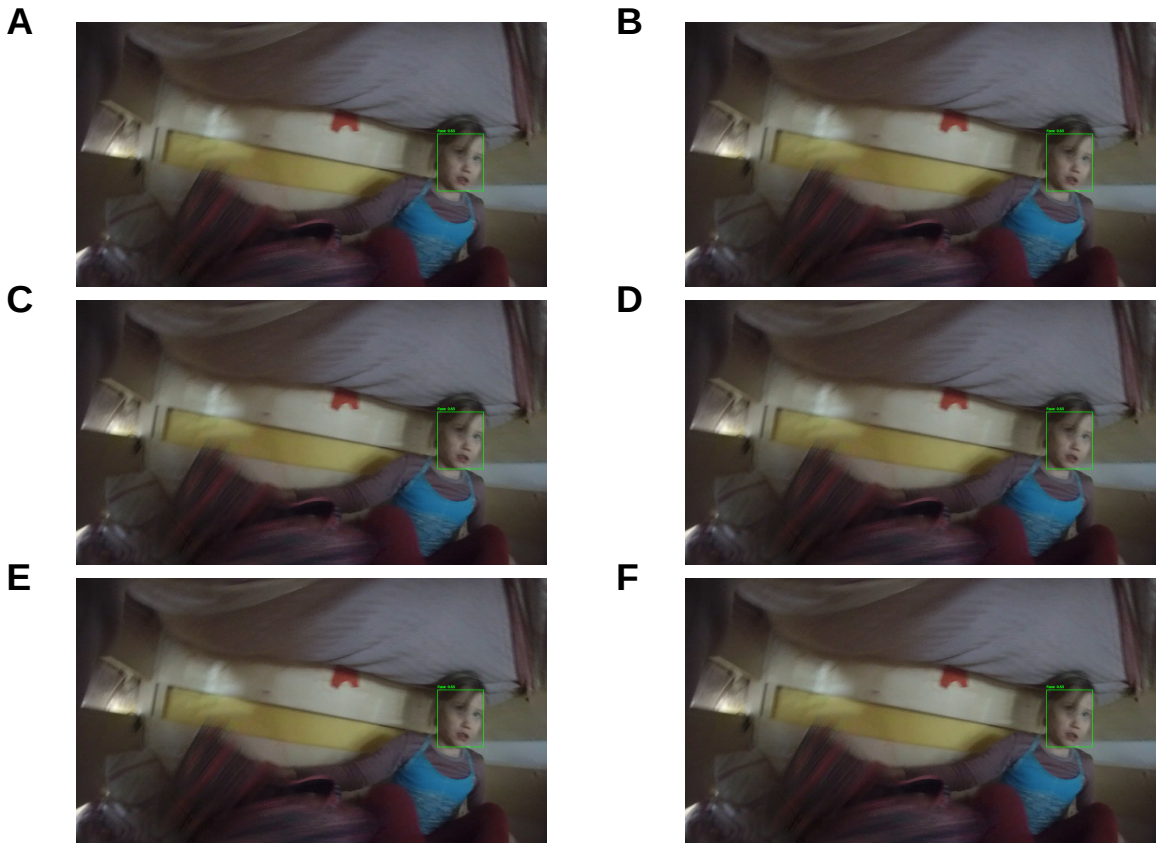

Figure 5. **A**, **B** - Examples of True Positives, **C**, **D** – Examples of False Negatives, **E**, **F** – Examples of False Positives in the YOLO11 person detection model.

annotation strategy, which prioritized labeling only the objects with which the child directly interacts. As a result, many other objects in the scene remain unannotated, despite being visually similar to the target objects, leading to confusion during training. Overall, the object classes appear to be more challenging for the model, particularly "kitchenware", "toy" and "other object". These results suggest that increasing the number of annotated samples, particularly for the objects the child directly interacts with, could improve detection performance. Additionally, future work could also explore model adjustments or data augmentation techniques to enhance the model's ability to distinguish these difficult classes, especially for small or partially visible objects.

Overall, the YOLO11 model demonstrates robust performance in detecting persons,

faces and objects. However, challenges remain in detecting certain object classes, particularly “kitchenware”, “toy” and “other object”, which require further investigation to improve model performance. These findings underscore the complexities inherent in analyzing egocentric video data, where movement and varying perspectives introduce additional challenges.

YOLO11x-cls: Gaze Classification

Selecting an appropriate model for gaze classification in our automated pipeline presented unique challenges due to the egocentric perspective of our dataset. Many gaze estimation methods rely on high-quality eye images, either extracted separately (Zhang, Sugano, Fritz, & Bulling, 2015) or as part of the full face (Zhang, Sugano, Fritz, & Bulling, 2016). However, our dataset often contains blurry or partially occluded faces captured at varying angles, making such approaches not suitable. Additionally, rather than predicting fine-grained gaze direction (e.g., left or right), our focus is on the binary classification of whether a person’s gaze is directed toward the child.

(Cheng, Wang, Bao, & Lu, 2021) provides an overview of the challenges and recent advancements in gaze estimation methods, including approaches that incorporate temporal information, such as Gaze360 (Kellnhofer, Recasens, Stent, Matusik, & Torralba, 2019). While these methods improve tracking across frames, they are not the primary focus of our study, as we analyze social interactions on a frame-by-frame basis.

Given these constraints, we opted for a CNN-based approach trained on ground truth annotations. Recent studies (Shah et al., 2022; Zhang et al., 2020; Zhang, Sugano, Fritz, & Bulling, 2019) have explored different CNN based gaze estimation architectures, among which are VGG, ResNet or YOLO architectures. Based on these findings, we implemented a VGG, ResNet, and YOLO-based model for binary classification (gaze directed toward the child or not). In our preliminary tests, Ultralytics’ YOLO11 architecture (Jocher & Qiu,

2024) demonstrated the best performance, leading us to select the YOLO11x classification model, pretrained on ImageNet, for our gaze classification task.

Given these constraints, we opted for a CNN-based approach trained on ground truth annotations. Recent studies (Shah et al., 2022; Zhang et al., 2020, 2019) have explored different CNN-based gaze estimation architectures, including VGG, ResNet, and YOLO. Based on these findings, we implemented and compared three models; VGG, ResNet, and YOLO; for binary classification (gaze directed toward the child or not). In our preliminary tests, Ultralytics' YOLO11 architecture (Jocher & Qiu, 2024) demonstrated the best performance, leading us to select the YOLO11x classification model, pretrained on ImageNet, for our gaze classification task.

The egocentric nature of our video data, recorded from a chest-mounted camera, posed additional challenges such as motion blur and oblique viewing angles, making gaze classification difficult even for human annotators. To address this, we defined gaze as "directed toward the child" when a person's gaze was oriented toward the child's face, body, or general direction. This included cases where the person was looking directly at the camera (worn by the child) or slightly upward, estimating the likely position of the child's head. Each detected face was annotated as either "gaze" (gaze directed at the child) or "no gaze" (gaze directed elsewhere).

YOLO11's architectural enhancements, such as the C2PSA block, improve the model's ability to focus on critical regions within an image, making it well-suited for our task. We fine-tuned the pretrained YOLO11x-cls model for gaze classification, enabling it to robustly determine whether individuals in the scene were engaged in visual attention toward the child.

Dataset Splitting. For the gaze classification task, we utilized a subset of the Quantex dataset, focusing specifically on frames containing annotated faces. This subset consisted of cut-out faces extracted from the annotated face bounding boxes, resulting in

Table 4

Dataset splits for the YOLO11x gaze classification model trained on the Quantex dataset. The table shows the total number of frames, as well as the number of frames with gaze and no gaze in the training, validation, and testing datasets after data augmentation of the minory class (Gaze). 'Gaze' indicates frames where the person's gaze is directed towards the child, while 'No Gaze' indicates frames where the person's gaze is not directed towards the child. Ratios are given in percentages.

Quantex	Train Ratio	Training	Val & Test Ratio	Validation	Testing	Total
Gaze	50	13160	79	1645	1646	16451
No Gaze	50	13160	21	443	445	14048
Total	100	26320	100	2088	2091	30499

20889 frames from 64 annotated videos. An analysis of the gaze labels showed that 21.25% of the faces were annotated as having gaze directed toward the child, leading to a class imbalance.

To maintain the original distribution of gaze labels across datasets, we first applied stratified dataset splitting, ensuring that the training, validation, and testing sets reflected the natural ratio of gaze and non-gaze frames. To address class imbalance in the training dataset, we applied data augmentation to increase the number of gaze frames. Since our dataset was already a subset of the Quantex dataset, downsampling to achieve balance would have resulted in a loss of valuable data. As a result, the final dataset included a balanced training set with 50% gaze and 50% no-gaze frames, while the validation and testing sets retained their original, unbalanced distribution. The final data distribution is presented in Table 4, with 26320 frames in the training set, 2088frames in the validation set, and 2091 frames in the testing set.

Training and Convergence. We trained the gaze classification model on the same Linux server used for person and object detection model training. Training ran for 37 epochs, completing in 10.40 hours. Similar to the YOLO detection model, we used an input image size of 640, a batch size of 16, a cosine annealing learning rate scheduler (Loshchilov & Hutter, 2017), and early stopping after 10 epochs without improvement, with a maximum of 200 epochs.

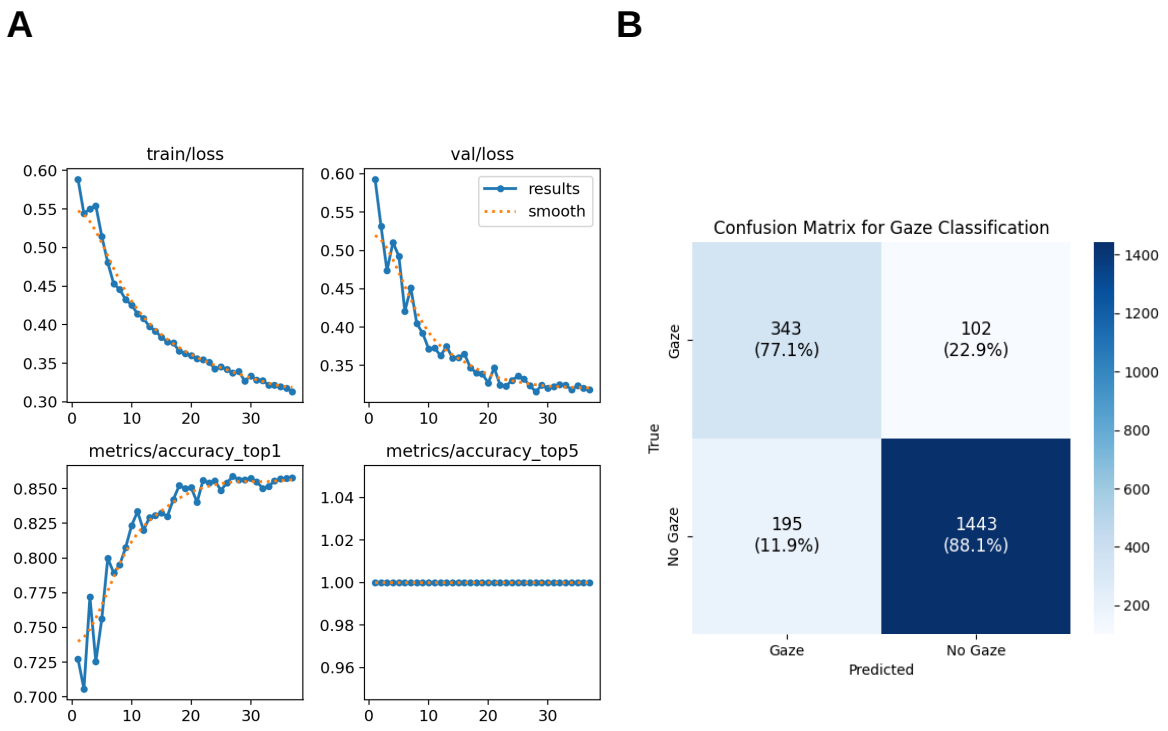

Figure 6. **A** - Training and Validation Loss Curves for the YOLO11x gaze classification model. **B** - Confusion Matrix for the YOLO11x gaze classification model trained on the Quantex dataset.

Figure 6 shows the cross-entropy loss curves for both training and validation. The steady decline in loss over time indicates that the model effectively learned to classify whether a person is looking in the direction of the child or not. The convergence of the

Table 5

Evaluation metrics for the YOLO11x gaze classification model trained on the Quantex dataset to classify whether a person is looking into the direction of the child wearing the camera or not. Precision, recall, and F1-score are given for the testing set.

Precision	Recall	F1-Score
0.63	0.77	0.70

training and validation loss curves suggests that the model generalizes well, as there is no indication of overfitting. Since gaze classification is a binary task, we used top-1 accuracy as the primary evaluation metric. Figure 6 also illustrates the increasing accuracy over time, confirming that the model progressively improves its ability to classify gaze correctly.

Model Evaluation Metrics. The YOLO11x gaze classification model achieved a precision of 0.64, a recall of 0.77, and an F1-score of 0.70 on the testing set. These metrics, summarized in table 5, indicate that the model effectively distinguishes between gaze and no-gaze frames, though there is still room for improvement. The egocentric perspective of the dataset presents challenges, as faces are often partially occluded or blurred, making gaze classification difficult—even for human annotators, who occasionally required a second inspection to determine gaze direction.

One of the primary challenges arises from cut-off faces, where the eyes are not always visible. In such cases, the model must rely on other facial features, such as head orientation

or mouth position, to infer gaze direction. While this approach is often effective, it occasionally leads to misclassifications. Despite these difficulties, the model achieved a satisfactory recall, correctly identifying 77% of all gaze frames. The F1-score of 070 balances precision and recall, providing a comprehensive measure of the model's performance.

More advanced gaze estimation methods—such as incorporating temporal information or leveraging additional facial landmarks—could further improve performance. However, given the constraints of our dataset and the focus on binary gaze classification, the YOLO11x model serves as a strong foundation for analyzing the gaze aspect of social interactions in egocentric video data.

Proximity Heuristic

Proximity between individuals is an important aspect of social interaction. Previous studies have shown that interpersonal distance can provide insights into the nature of relationships and social engagement (Hernández-Heredia, Reyes-Manzano, Flores-Hernández, Ramos-Fernández, & Guzmán-Vargas, 2024; Janssen et al., 2024; Onnela et al., 2014). Since our dataset does not contain explicit proximity labels, we developed a heuristic approach to estimate the distance between the child and other individuals.

Our heuristic relies on the bounding boxes of detected faces, assuming that the size of these bounding boxes provides an implicit cue for proximity. Specifically, we infer that larger bounding boxes correspond to individuals who are closer to the camera, whereas smaller bounding boxes indicate individuals positioned further away. To quantify this relationship, we calculate the area of a detected face bounding box as the product of its width and height. This computed size is then compared to a predefined reference bounding box. For both adults and children/infants, we selected two reference face sizes: one corresponding to a face very close to the camera (within arm's reach) and another

representing a face further away (e.g., in the background or outdoors). By normalizing the detected face sizes relative to these reference sizes, we obtain a proximity score ranging from 0 to 1, where a value of 1 indicates a face extremely close to the camera, and 0 represents a face that is far away.

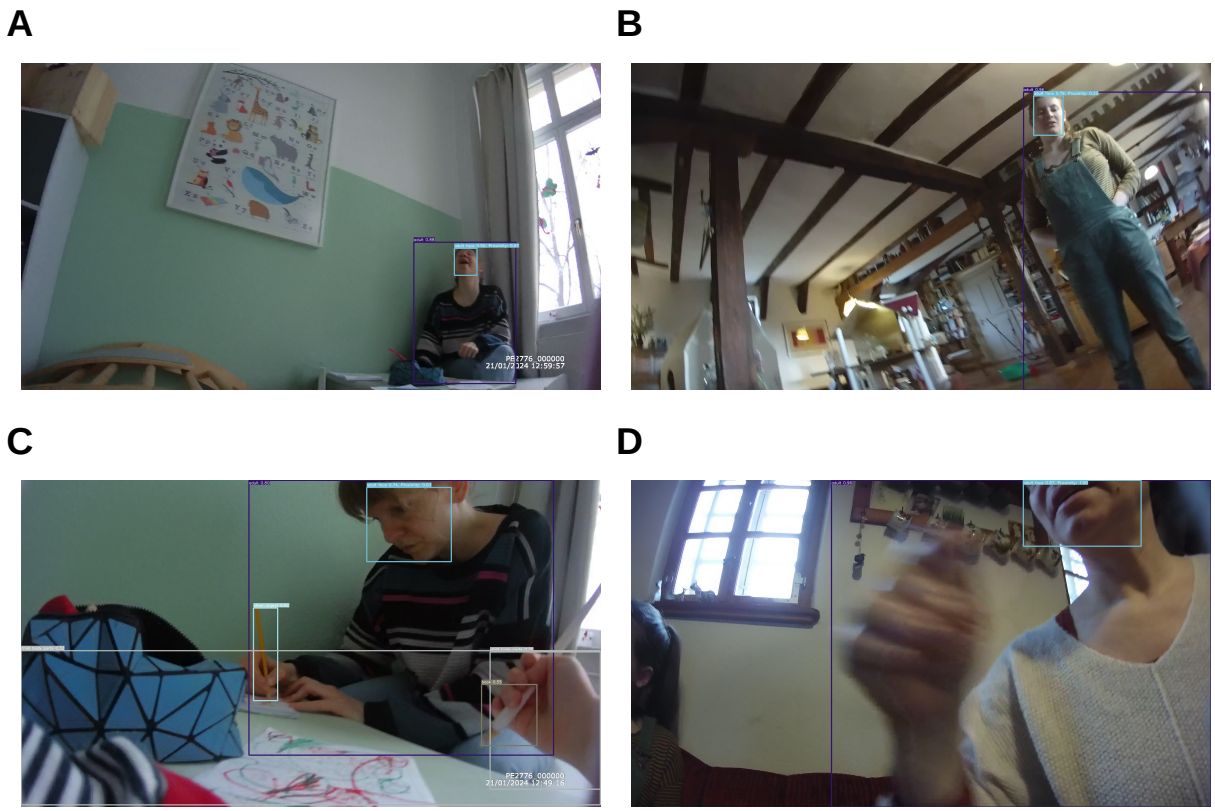

Figure 7. Proximity heuristic examples. Example **A** shows a face far away from the camera (proximity score = 0.07), example **B** depicts a face slighlty closer (proximity score = 0.2). Example **C** shows a face quite close to the camera (proximity score = 0.6), and example **D** illustrates a face extremely close to the camera (proximity score = 1).

To better account for the nonlinear relationship between face size and distance, we applied a logarithmic transformation to the bounding box areas. Since face area decreases quadratically as distance increases, a simple linear mapping would make small differences appear too large for close faces and too small for distant ones. The logarithmic

transformation helps balance this by compressing large values and stretching smaller ones, creating a more natural and meaningful proximity scale. This approach aligns with how humans perceive size, as our sensitivity to changes depends on relative differences rather than absolute ones (Stevens & Marks, 2017). By using this scaling, we ensure that proximity estimates remain consistent across different face sizes and distances.

In addition to absolute size, we incorporate a width-to-height ratio as an additional cue for proximity estimation. This ratio was determined separately for adult and child/infant faces based on reference images depicting full, front-facing faces. If a detected face exhibits a significant deviation from the expected aspect ratio, we infer that the face is partially cropped due to extreme proximity to the camera. In such cases, the proximity score is set to 1. An example illustrating this case is shown in Figure **??**.

This heuristic provides an interpretable and computationally efficient method for estimating proximity in egocentric recordings, allowing us to examine the spatial dynamics of social interactions without the need for explicit depth information or additional sensors.

Voice Type Classification

Regarding the audio component of our interaction analysis, we aimed to use a model that does not only detect the presence of speech but also distinguishes between the key child wearing the camera and other speakers. This distinction is crucial for understanding the dynamics of the interactions and the role of the key child in the social context. To achieve this, we applied the Voice Type Classifier (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020), an open-source model designed to identify five different voice types: key child, other child, female adult, male adult, and speech in general. The model is based on a convolutional neural network (CNN) architecture and was trained on 260 hours of child-centered recordings across 10 different languages.

Although the Quantex dataset does not include explicit audio labels, we are confident

in the model's suitability for our data. Prior testing on a similar labeled dataset which was also collected in our lab, ChildLens, demonstrated that the Voice Type Classifier achieved an F1 score of 58.1 (reference to ChildLens paper), which is comparable to the F1 score of 57.3 reported on the original training dataset. These results indicate that the model generalizes well to our child-centered recordings, making it a reliable choice for our analysis.

We applied the Voice Type Classifier to the extracted audio data from the Quantex dataset using the code provided by the authors of the voice type classifier (Lavechin, 2020). The model was used to detect the presence of speech and classify the speaker into one of the five predefined categories.

References

Carpendale, J., & Lewis, C. (2020). *What Makes Us Human: How Minds Develop through Social Interactions* (1st ed.). Routledge. https://doi.org/10.4324/9781003125105

Cheng, Y., Wang, H., Bao, Y., & Lu, F. (2021). Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. https://doi.org/10.48550/ARXIV.2104.12668

Dai, S., Bouchet, H., Karsai, M., Chevrot, J.-P., Fleury, E., & Nardy, A. (2022). Longitudinal data collection to follow social network and language development dynamics at preschool. *Scientific Data*, *9*(1), 777. https://doi.org/10.1038/s41597-022-01756-x

Hernández-Heredia, T. K., Reyes-Manzano, C. F., Flores-Hernández, D. A., Ramos-Fernández, G., & Guzmán-Vargas, L. (2024). Proximity Sensor for Measuring Social Interaction in a School Environment. *Sensors*, *24*(15), 4822. https://doi.org/10.3390/s24154822

Heyes, C. M. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Cambridge (Mass.): Harvard University press.

Janssen, L. H. C., Verkuil, B., Nedderhoff, A., Van Houtum, L. A. E. M., Wever, M. C. M., & Elzinga, B. M. (2024). Tracking real-time proximity in daily life: A new tool to examine social interactions. *Behavior Research Methods*, *56*(7), 7482–7497. https://doi.org/10.3758/s13428-024-02432-1

Jocher, G., & Qiu, J. (2024). *Ultralytics YOLO11*. Retrieved from https://github.com/ultralytics/ultralytics

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., & Torralba, A. (2019). Gaze360: Physically Unconstrained Gaze Estimation in the Wild. https://doi.org/10.48550/ARXIV.1910.10088

Khanam, R., & Hussain, M. (2024, October 23). YOLOv11: An Overview of the Key Architectural Enhancements. https://doi.org/10.48550/arXiv.2410.17725

Lavechin, M. (2020). *Voice Type Classifier*. Retrieved from https://github.com/MarvinLvn/voice-type-classifier

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. https://doi.org/10.48550/ARXIV.2005.12656

Lemaignan, S., Edmunds, C. E. R., Senft, E., & Belpaeme, T. (2018). The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PLOS ONE*, *13*(10), e0205999. https://doi.org/10.1371/journal.pone.0205999

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., ... Yang, J. (2020, June 8). Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. https://doi.org/10.48550/arXiv.2006.04388

Loshchilov, I., & Hutter, F. (2017, May 3). SGDR: Stochastic Gradient Descent with Warm Restarts. https://doi.org/10.48550/arXiv.1608.03983

Onnela, J.-P., Waber, B. N., Pentland, A., Schnorf, S., & Lazer, D. (2014). Using sociometers to quantify social interaction patterns. *Scientific Reports*, *4*(1), 5604. https://doi.org/10.1038/srep05604

Piaget, J. (1964). Part I: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching*, *2*(3), 176–186. https://doi.org/10.1002/tea.3660020306

Rogoff, B., Dahl, A., & Callanan, M. (2018). The importance of understanding children's lived experience. *Developmental Review*, *50*, 5–15. https://doi.org/10.1016/j.dr.2018.05.006

Rossano, F., Terwilliger, J., Bangerter, A., Genty, E., Heesen, R., & Zuberbühler, K. (2022). How 2- and 4-year-old children coordinate social interactions with peers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1859), 20210100. https://doi.org/10.1098/rstb.2021.0100

Shah, S. M., Sun, Z., Zaman, K., Hussain, A., Shoaib, M., & Pei, L. (2022). A Driver Gaze Estimation Method Based on Deep Learning. *Sensors*, *22*(10), 3959. https://doi.org/10.3390/s22103959

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences*, *22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

Stevens, S. S., & Marks, L. E. (2017). *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects* (1st ed.). Routledge. https://doi.org/10.4324/9781315127675

Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., Chavez-Urbiola, E. A., & Romero-Gonzalez, J. A. (2024, October 12). Loss Functions and Metrics in Deep Learning. https://doi.org/10.48550/arXiv.2307.02694

Tomasello, M. (2009). *Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., & Hilliges, O. (2020). ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. https://doi.org/10.48550/ARXIV.2007.15837

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4511–4520. Boston, MA, USA: IEEE. https://doi.org/10.1109/CVPR.2015.7299081

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2016). *It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation*. https://doi.org/10.48550/ARXIV.1611.08860

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2019). MPIIGaze: Real-World Dataset

and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(1), 162–175.
https://doi.org/10.1109/TPAMI.2017.2778103

Appendix