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; Piaget, 1964; Rogoff, Dahl, & Callanan, 2018; Smith, Jayaraman, Clerkin, & Yu, 2018; Tomasello, 2009; Vygotsky, 1978; **heyesCognitiveGadg?**). 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.

The present study investigates social interactions in naturalistic home settings over an extended period. The corresponding **Quantex** dataset currently includes 197.20 hours of egocentric video and audio recordings from children aged 3 to 5 years. Here we focus on 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.
- **Person Proximity**: Estimating the distance between the child and others to quantify social engagement.
- **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 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

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.

Participants Recruitment and Data Collection

This study collected egocentric video recordings from 76 children, aged 3 to 5 years, over a span of 73 months [MB: in addition, it would be interesting to know in what time intervall the videos were collected for each child, also a more detailed overview of the number of children per age group would be nice].

Recruitment Process. Participants were recruited from a mid-sized German city through an existing lab database, with approximately equal distribution across age groups (30% 3-year-olds, 35% 4-year-olds, 35% 5-year-olds). 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. Participants were recruited from an internal lab database and contacted via phone. Parents received a detailed explanation of the study's purpose and procedures, consistent with the information in the study brochure. Families who agreed to participate were provided with a vest equipped with an embedded camera, which children were asked to wear for approximately two hours, with flexible extension options. For privacy protection, recordings were limited to the home environment. Recorded videos were securely stored on the MPI-EVA server and subsequently made available to parents. To enhance convenience, parents could choose to have the camera delivered to their home or workplace, or opt for self-pickup.

Distribution of Study Materials and Instructions. At the handover appointment, parents received One or two vests (multiple sizes available), a fully charged camera unit and comprehensive information on the study's purpose and data protection protocols. Parents received hands-on training on camera operation and were encouraged to practice using the equipment. The consent form required listing the participating child and documenting all individuals potentially captured in the recordings. Five copies of data privacy information sheets were provided for distribution to anyone who might be recorded.

A follow-up call was scheduled to ensure recording success, typically planned for the weekend following the handover to ensure enough time for recoding.

Return of Equipment and Final Data Confirmation. On the scheduled date, parents were contacted to confirm recording success and assess any need for additional time. Flexible rescheduling was offered as needed. Once the recording was completed, a pickup or drop-off appointment was arranged to collect the completed consent form, vest, and camera. Parents were also asked to confirm their current email address for sharing the recordings. Children received a "Labyrinth" game as a thank for their participation.

Materials

To capture 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. Children were free to move around and engage in their usual activities at home without any interference or instructions given to their parents.

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.

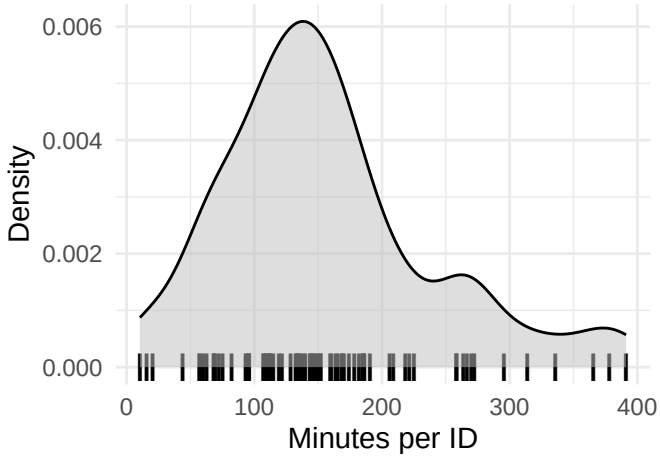

Figure 1. Video recording duration (in minutes) per Child in the Quantex Dataset.

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 are recorded, including a unique identifier, age (infant, child, teen, adult, unknown), and gender (female, male, unknown). 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.

The egocentric nature of our video data, recorded from a chest-mounted camera, posed additional challenges such as motion blur and oblique viewing angles, which made gaze classification difficult, even for human annotators. These factors contributed to occlusion or distortion of gaze information in many frames, making accurate gaze direction annotation particularly challenging.

Objects are annotated only when the child is actively interacting with them, either by holding them in their own hands or when another person in the interaction is holding the

object in their hands. 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 Quantex dataset consists of a total of 634 videos. However, only a subset of 80 videos was annotated, totaling 113799 frames. To balance workload while still obtaining meaningful annotations, we applied a frame sampling strategy, annotating every 30th frame, which corresponds to one frame per second. The goal of these annotations was to create ground truth data for training a model that can later be used to analyze the remaining videos in the dataset. While the video data was subject to structured annotation and validation, the audio data was used in its raw form without preprocessing for analysis.

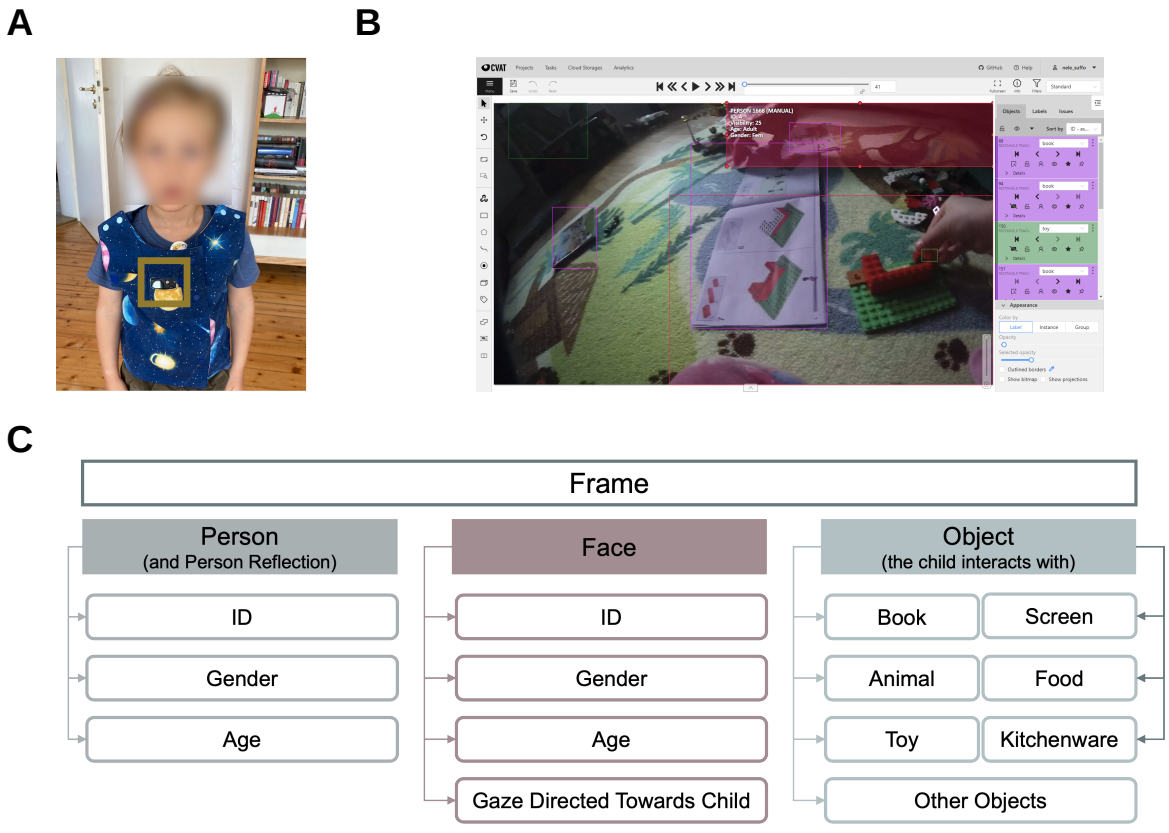

Figure 2. **A** – Vest with the embedded camera worn by the children, **B** – CVAT platform utilized for video annotation, **C** – Annotation Strategy in the Quantex dataset.

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) a
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

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

Outline

This document contains supplementary material for the paper "Exploring Aspects of Social Interaction in Children's Everyday Lives using Machine Learning: A Multimodal Analysis of the Quantex Dataset". First, we provide an overview of the Participant Recruitment and Data Collection, including a detailed description of the Recruitment and Data Collection process. We then describe the Materials used in the study, followed by a structured overview of the Dataset, including the Annotation Strategy. Next, we outline the Automated Analysis Pipeline, including YOLO11x: Multi-Class Detection of Persons, Faces, and Objects, YOLO11x-cls: Gaze Classification), as well as Voice Type Classification. Finally, we report the Results of the automated pipeline, evaluating the performance of the models when applied to all videos in the Quantex dataset.

Methodology

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.

Participants Recruitment and Data Collection

This study collected egocentric video recordings from 76 children, aged 3 to 5 years, over a span of 73 months [MB: in addition, it would be interesting to know in what time intervall the videos were collected for each child, also a more detailed overview of the number of children per age group would be nice].

Recruitment Process. Participants were recruited from a mid-sized German city through an existing lab database, with approximately equal distribution across age groups (30% 3-year-olds, 35% 4-year-olds, 35% 5-year-olds). 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. Participants were recruited from an internal lab database and contacted via phone. Parents received a detailed explanation of the study's purpose and procedures, consistent with the information in the study brochure. Families who agreed to participate were provided with a vest equipped with an embedded camera, which children were asked to wear for approximately two hours, with flexible extension options. For privacy protection, recordings were limited to the home environment. Recorded videos were securely stored on the MPI-EVA server and subsequently made available to parents. To enhance convenience, parents could choose to have the camera delivered to their home or workplace, or opt for self-pickup.

Distribution of Study Materials and Instructions. At the handover appointment, parents received One or two vests (multiple sizes available), a fully charged camera unit and comprehensive information on the study's purpose and data protection protocols. Parents received hands-on training on camera operation and were encouraged to practice using the equipment. The consent form required listing the participating child and documenting all individuals potentially captured in the recordings. Five copies of data privacy information sheets were provided for distribution to anyone who might be recorded. A follow-up call was scheduled to ensure recording success, typically planned for the weekend following the handover to ensure enough time for recoding.

Return of Equipment and Final Data Confirmation. On the scheduled date, parents were contacted to confirm recording success and assess any need for additional time. Flexible rescheduling was offered as needed. Once the recording was completed, a pickup or drop-off appointment was arranged to collect the completed consent form, vest,

and camera. Parents were also asked to confirm their current email address for sharing the recordings. Children received a "Labyrinth" game as a thank for their participation.

Materials

To capture 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. Children were free to move around and engage in their usual activities at home without any interference or instructions given to their parents.

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 3 shows the distribution of video duration per child.

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 are recorded, including a unique identifier, age (infant, child, teen, adult, unknown), and gender (female, male, unknown). 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

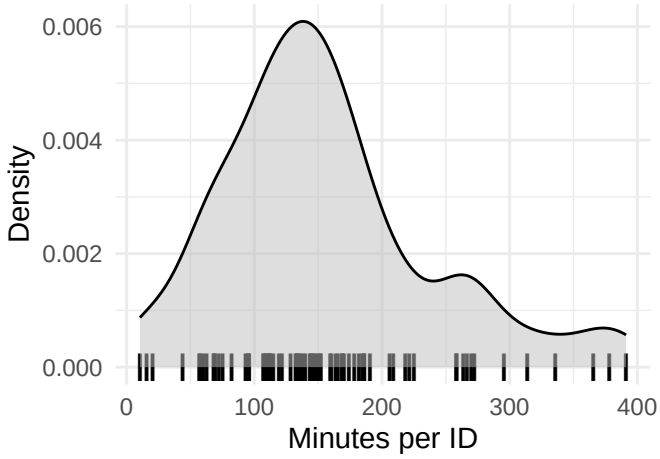

Figure 3. Video recording duration (in minutes) per Child in the Quantex Dataset.

interactions. Partially visible faces are also annotated if key facial features, such as the nose, eyes, or mouth, remain identifiable.

The egocentric nature of our video data, recorded from a chest-mounted camera, posed additional challenges such as motion blur and oblique viewing angles, which made gaze classification difficult, even for human annotators. These factors contributed to occlusion or distortion of gaze information in many frames, making accurate gaze direction annotation particularly challenging.

Objects are annotated only when the child is actively interacting with them, either by holding them in their own hands or when another person in the interaction is holding the object in their hands. 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 4.

The Quantex dataset consists of a total of 634 videos. However, only a subset of 80 videos was annotated, totaling 113799 frames. To balance workload while still obtaining meaningful annotations, we applied a frame sampling strategy, annotating every 30th frame, which corresponds to one frame per second. The goal of these annotations was to create ground truth data for training a model that can later be used to analyze the

remaining videos in the dataset. While the video data was subject to structured annotation and validation, the audio data was used in its raw form without preprocessing for analysis.

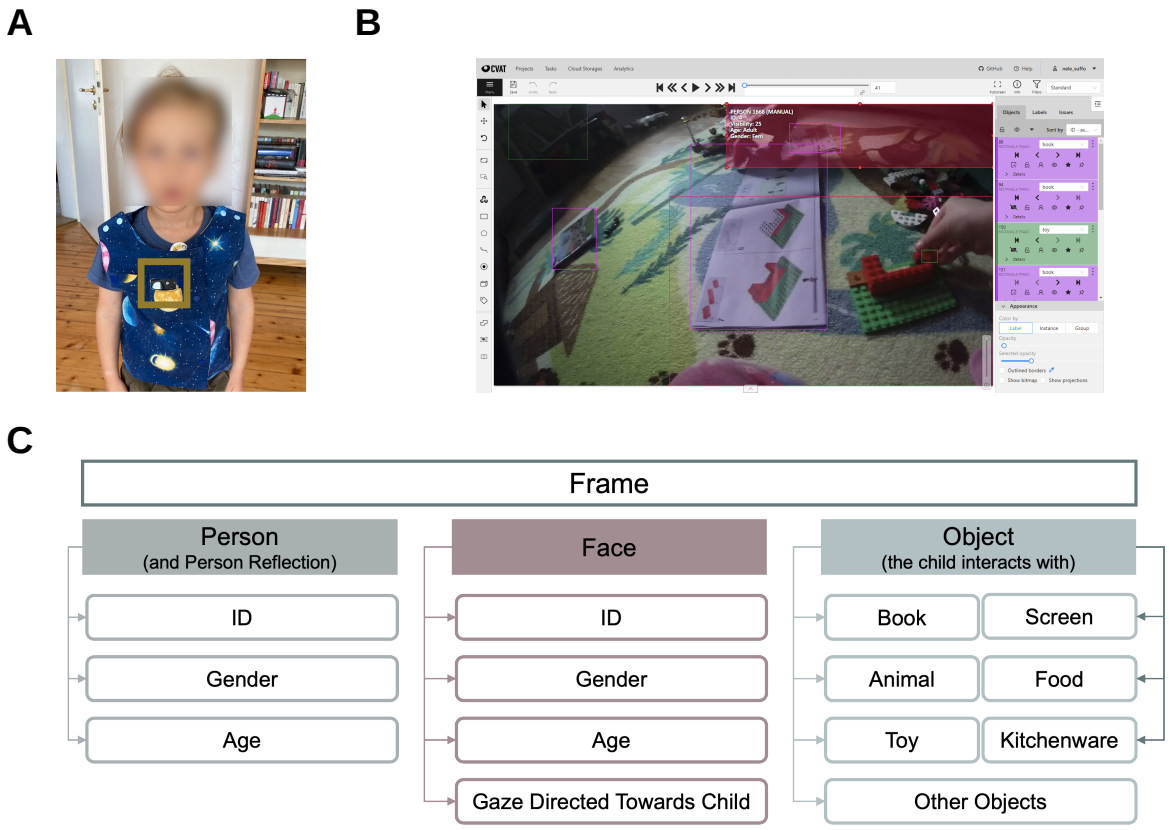

Figure 4. **A** – Vest with the embedded camera worn by the children, **B** – CVAT platform utilized for video annotation, **C** – Annotation Strategy in the Quantex dataset.

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) a

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

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

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 (Lin et al., 2014), a large-scale dataset containing labeled images for 80 object categories commonly found in everyday environments. COCO is widely used for object detection, instance segmentation, and keypoint detection tasks.

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.

For our study, we utilized YOLO11x, the largest model in the YOLO11 model series, which has 56.9M parameters and 194.9 GFLOPs. With the highest accuracy ($\text{mAP}^{val}_{50-95} = 54.7$) among YOLO11 variants, YOLO11x, combined with its architectural advancements, was the optimal choice for our task, as we had the computational resources to support running a larger model.

Dataset Annotation and Preprocessing. Our dataset presents unique challenges due to its egocentric viewpoint, as body parts of the child wearing the camera frequently appear in the footage. To prevent misclassification, we employ a dedicated annotation scheme where all individuals are labeled as "person" with unique IDs, consistently assigning the key child (child wearing the camera) ID = 1. During preprocessing, we map this ID to a separate "child body parts" category to distinguish the child's presence from other individuals.

Additionally, we refine the "person" and "face" categories beyond standard YOLO models by:

- Differentiating the key child from other individuals.
- Distinguishing adults from children/infants in both full-body detections and faces.

Our fine-tuned YOLO11 model also classifies five object categories relevant to the child's interactions: toy, book, kitchenware, screen, and other object. Due to low occurrences, "animal" (19 instances, 0.015%) and "food" (1,115 instances, 0.84%) were included into "other object", leaving five final object categories with at least 2,000 instances each.

Dataset Splitting. The full Quantex dataset consists of 634 videos, all recorded at 30fps, resulting in 19023571 frames. From this dataset, we annotated 80 videos, totaling 113799 frames. Prior to splitting this dataset into training, validation, and testing datasets, we analyzed how often each class was present in the dataset. The distribution, displayed in detail in Table 1, revealed an imbalance, with the "Adult" class being the most frequent. To address this imbalance, we applied a stratified split to ensure that each dataset preserved the original class distribution. Following an 80/10/10 split, the final datasets consisted of 91038 frames for training (80%), 11379 frames for validation (10%), and 11382 frames for testing (10%), ensuring an accurate evaluation of model performance relative to real-world data distribution (see Figure 2).

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	8669	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

Table 2

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

	Training	Validation	Testing	Total
Number of frames	91039	11380	11380	113799

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

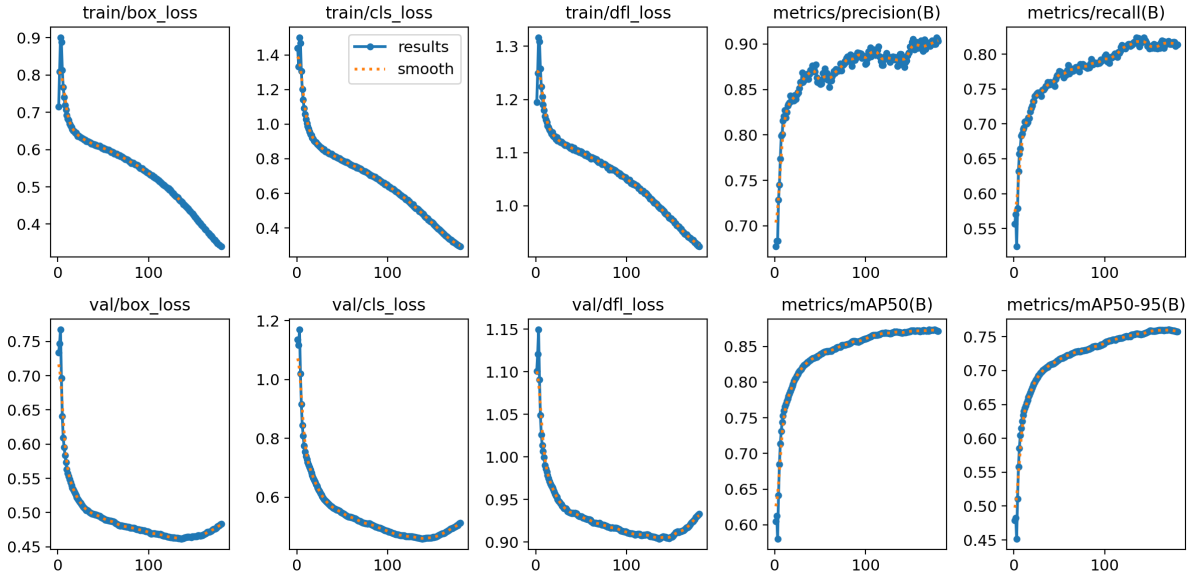

Figure 5. Training and Validation Loss Curves for the YOLO11x detection model.

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* (DFL) 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 5. 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 0.91 and a recall of 0.80 on the testing set, resulting in an F1-score of 0.85. The F1 metric is particularly important as it reflects both the accuracy of identifying each class instance (precision) and the model's ability to detect all occurrences of the class (recall). The normalized confusion matrix (see figure 6) reveals strong overall performance with mean Average Precision across all classes mAP=0.87. 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 most classes, indicating that the model can effectively distinguish between them (see Figure 6). A closer look at class-wise performance reveals that the model excels in detecting people and certain objects. The "infant/child," "adult," and "adult face" classes, as well as "child body parts," "infant/child face," "book," and

Table 3

Mean Average Precision (mAP@0.5) for the YOLO11x detection model trained on the Quantex dataset. The table shows the mAP@0.5 for each class.

class	mAP@0.5
infant/child	0.87
adult	0.95
infant/child face	0.85
adult face	0.95
child body parts	0.97
book	0.94
toy	0.78
kitchenware	0.81
screen	0.96
other object	0.83

Table 4

Detection metrics for the YOLO11x detection model trained on the Quantex dataset. The table reports mean Average Precision at IoU 0.5 (mAP@0.5), along with the averaged precision, recall, and F1-score across all classes.

mAP@0.5	Precision	Recall	F1-Score
0.870	0.91	0.80	0.85

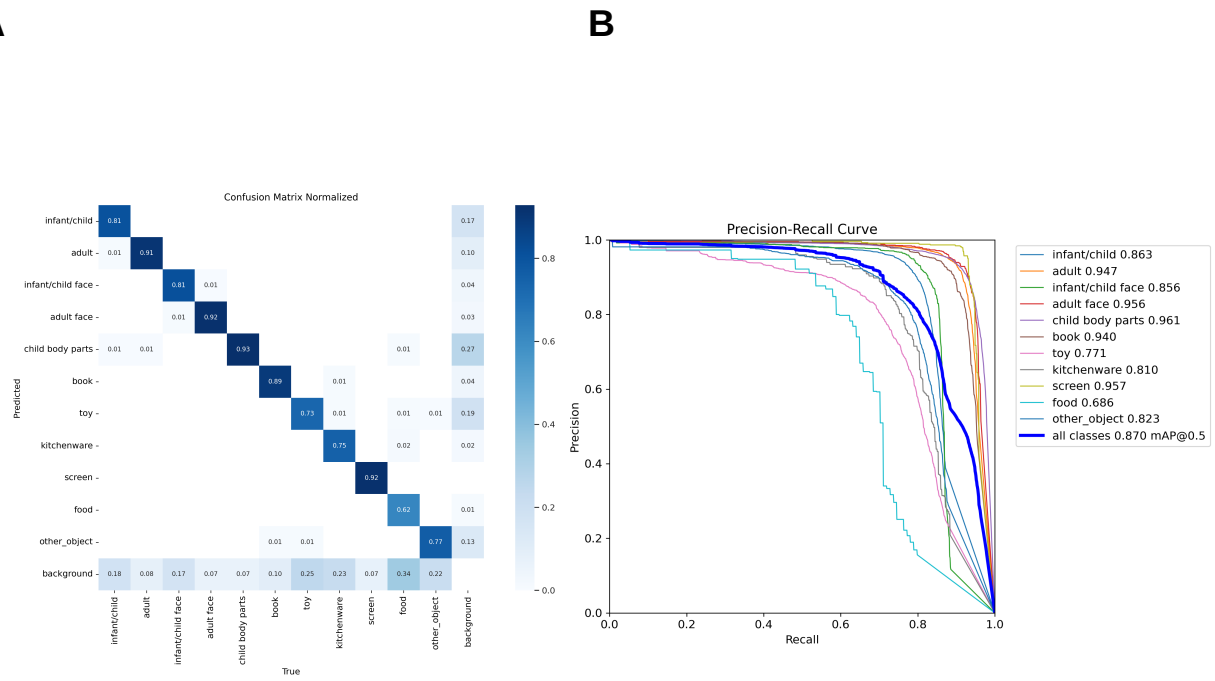

Figure 6. **A** - Confusion Matrix for the YOLO11x detection model trained on the Quantex dataset. **B** - Precision-Recall Curve for the YOLO11x detection model.

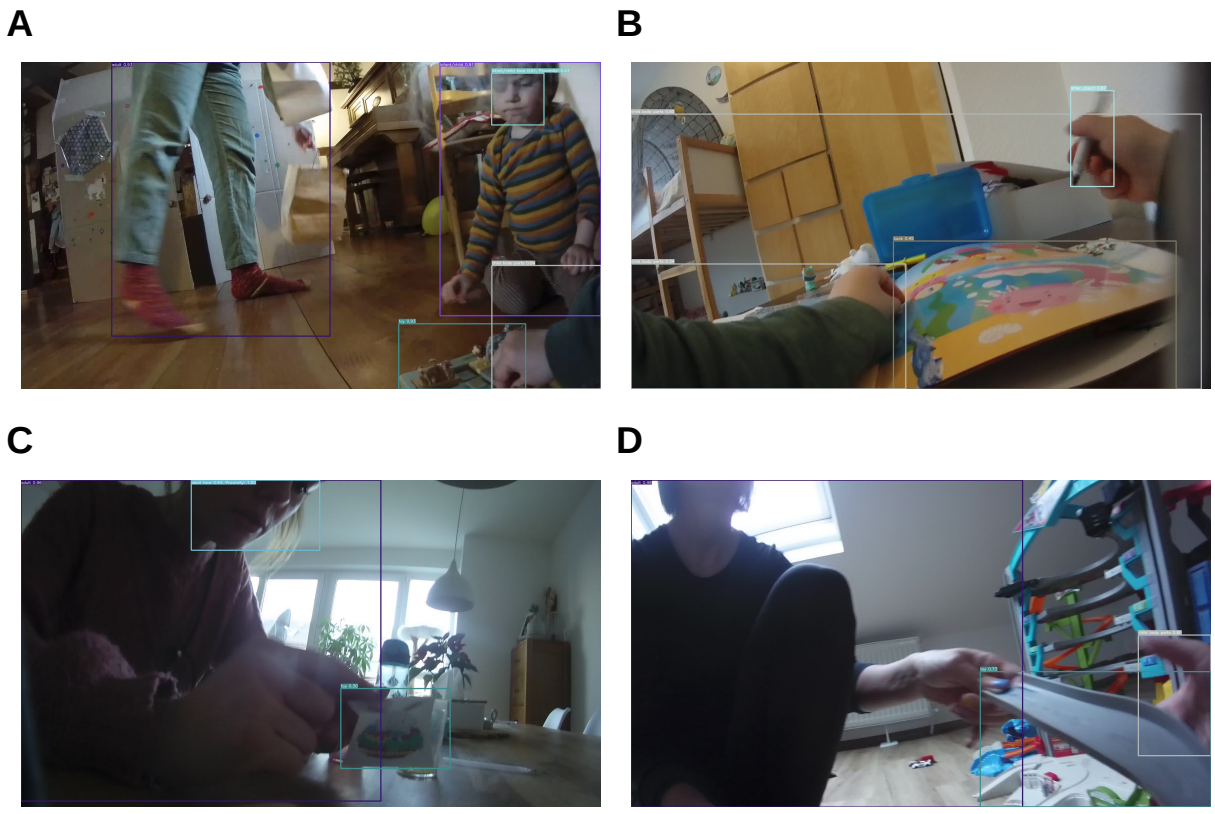

Figure 7. Detection examples for the YOLO11x detection model trained on the Quantex dataset.

"screen," all achieve high Average Precision (AP) scores. This indicates that the model reliably identifies these categories with strong precision and recall (see figure 6).

On the other hand, some object classes remain more challenging. In particular, the "toy," "kitchenware," and "other object" categories exhibit lower AP scores of 0.77, 0.77, and 0.82, respectively. The confusion matrix confirms that 22% of "kitchenware" and 22% of "toy" instances are frequently misclassified as "background." This issue likely stems from the annotation strategy, which prioritized labeling only objects that the child directly interacted with. As a result, other visually similar objects in the background remain unannotated, introducing ambiguity during training. Additionally, "infant/child" and "infant/child face" also show notable false negative rates of 17%, further contributing to

misclassification. These findings suggest that increasing the number of annotated samples—particularly for objects with which the child interacts with—could improve detection performance. Future work could also explore model refinements or data augmentation techniques to enhance the detection of small, occluded, or partially visible objects. In conclusion, the YOLO11 model demonstrates strong performance in detecting people, faces, and several object classes.

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 or not.

Cheng, Wang, Bao, and 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 of our dataset, we opted for a CNN-based approach, which could be 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 ResNet and YOLO-based model for binary classification

(gaze directed toward the child or not). After conducting preliminary tests, Ultralytics' YOLO11 architecture (Jocher & Qiu, 2024) demonstrated the best performance, which led us to select the YOLO11x classification model, pretrained on ImageNet, for our gaze classification task.

We fine-tuned YOLO11x-cls, the largest model in the YOLO11 classification series, which has 28.4M parameters. The model achieved the highest accuracy among all YOLO11 variants (acc_top1 = 79.5), making it the optimal choice for gaze classification in our dataset. Moreover YOLO11's architectural enhancements, such as the C2PSA block, improve the model's ability to focus on critical regions within an image, ensuring robust detection of visual attention toward the child, even under challenging real-world conditions.

Data Annotation and Preprocessing. We defined gaze as being "directed toward the child" when a person's gaze was oriented toward the child's face, body, or general direction of the child. This included instances where the person was looking directly at the camera (worn by the child) or slightly upwards, allowing for an estimation of 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).

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 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

Table 5

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 minority 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

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 5, with 26320 frames in the training set, 2088 frames 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. For details, see Training and Convergence. 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 8 shows the cross-entropy loss curves for both training and validation. The

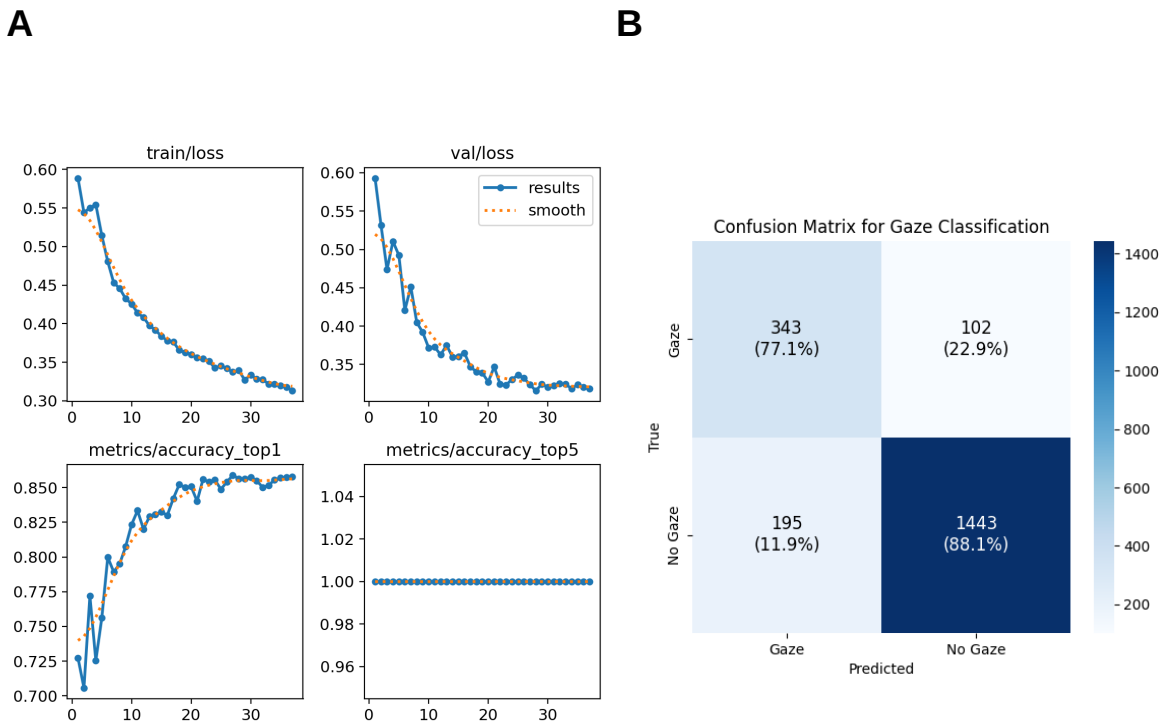

Figure 8. **A** - Training and Validation Loss Curves for the YOLO11x gaze classification model. **B** - Confusion Matrix for the YOLO11x gaze classification model trained on the Quantex dataset.

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 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 8A 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 6, indicate that the model effectively distinguishes between gaze and

Table 6

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
0.64	0.77	0.70

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.

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.

Formula for Proximity Estimation. We infer proximity from the size of 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 closer to the camera, whereas smaller bounding boxes indicate individuals further away. The proximity value is calculated using the face area, defined as the product of the bounding box width and height. This computed size is then compared to predefined 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).

Due to the nonlinear relationship between face size and distance, where 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. To address this, we apply a logarithmic scaling approach, which maps the calculated face area to a proximity score between 0 and 1:

$$\text{Proximity} = \frac{\ln(\text{Face Area}) - \ln(\text{Max Reference Area})}{\ln(\text{Min Reference Area}) - \ln(\text{Max Reference Area})}$$

Where:

- **Face Area** is the area of the detected face, calculated as the width times the height of the bounding box.
- **Max Reference Area** is the reference area corresponding to the furthest detectable face.
- **Min Reference Area** is the reference area corresponding to the closest detectable face.

This logarithmic transformation ensures that proximity values are compressed for large faces (close to the camera) and stretched for small faces (farther away), making them perceptually meaningful. Since human sensitivity to size changes depends on relative differences rather than absolute ones, logarithmic scaling aligns with psychophysical principles of distance perception (Stevens & Marks, 2017).

Incorporating Width-to-Height Ratio. In addition to face size, we consider the width-to-height ratio of the detected face as a cue for proximity. For both adults and children/infants, we defined an expected aspect ratio based on reference images of full, front-facing faces. If a detected face's aspect ratio deviates significantly from the expected ratio, we infer that the face is extremely close to the camera, causing partial cropping of the bounding box.

If the deviation exceeds a defined threshold (ϵ), we set the proximity score to 1:

$$r_{\text{expected}} = \frac{w_{\text{expected}}}{h_{\text{expected}}}$$

$$\text{If} \quad |r_{\text{detected}} - r_{\text{expected}}| > \epsilon, \quad P = 1$$

Where ϵ is a fixed threshold (ϵ=0.5) that determines the degree of deviation considered significant for extreme proximity.

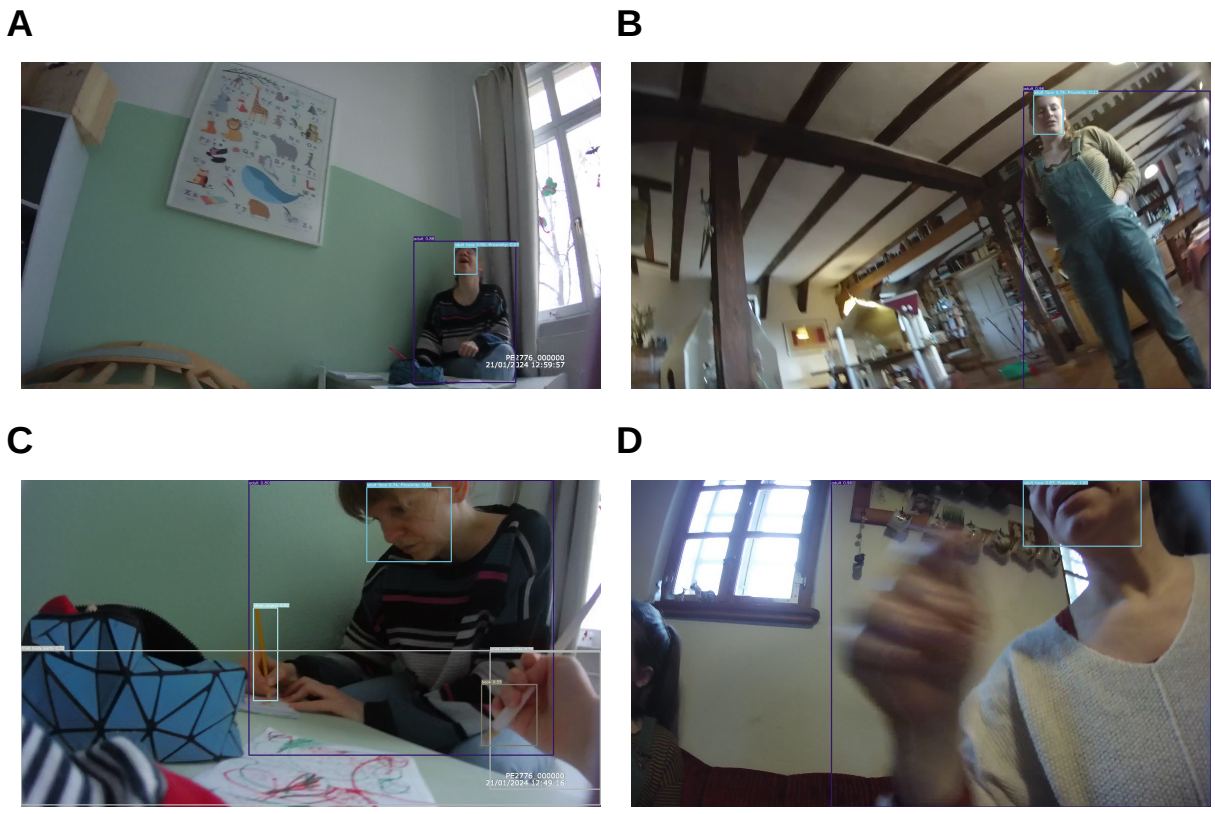

Figure 9. Proximity Heuristic examples. Example **A** shows a face far away from the camera (proximity score = 0.07), example **B** depicts a face slighlty closer (proximity score = 0.2). Example **C** shows a face quite close to the camera (proximity score = 0.6), and example **D** illustrates a face extremely close to the camera (proximity score = 1).

This heuristic approach, which combines face area and aspect ratio, provides an efficient and interpretable method for estimating proximity without requiring additional sensors or depth information. It enables us to examine the spatial dynamics of social interactions in egocentric recordings, offering insights into the relationships between children and others during everyday experiences. Examples of proximity estimation are shown in Figure 9, illustrating the range of proximity scores from 0 (far away) to 1 (extremely close).

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 (KCHI), other child (OCH), female adult (FEM), male adult (MAL), and speech in general (SPEECH). This multi-label classification enables the detection of multiple classes simultaneously, which is particularly useful when faced with overlapping speech segments, which frequently occurs in natural interactions.

The model, as described by Lavechin et al. (2020), is based on a convolutional neural network (CNN) architecture that includes SincNet components. These components are designed to extract meaningful frequency representations, which are then processed by a stack of bi-directional long short-term memory (LSTM) and fully connected layers. The final layer applies a sigmoid activation function to output a score between 0 and 1 for each class, representing the probability of that voice type being present in the audio segment. The network is trained using binary cross-entropy loss, optimizing each class independently.

The Voice Type Classifier was trained on 260 hours of child-centered recordings spanning 10 different languages. Its architecture is specifically designed to handle the unique challenges of child-centered audio, such as overlapping speech and varying acoustic conditions. To address these complexities, the model incorporates a multi-label classification approach, allowing it to detect multiple voice types simultaneously.

Model performance was evaluated using the F1-measure, with the classifier showing a significant performance advantage over the widely used, closed-source LENA system (Ford, Baer, Xu, Yapanel, & Gray, 2008). Specifically, the classifier achieved an improvement of

10.6 in the average F1-measure across the five voice type classes.

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 performs similarly to the original work.

Results

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

Ford, M., Baer, C. T., Xu, D., Yapanel, U., & Gray, S. (2008). *The LENA Language Environment Analysis System: Audio Specifications of the DLP-0121.* Retrieved from https://www.lena.org/wp-content/uploads/2016/07/LTR-03-2_Audio_Specifications.pdf

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

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., 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

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2014). Microsoft COCO: Common Objects in Context. https://doi.org/10.48550/ARXIV.1405.0312

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