ChildLens: An Egocentric Video Dataset for Activity Analysis in Children

Nele-Pauline Suffo[1], Anas Suffo[1], Pierre-Etienne Martin[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

We present ChildLens, an egocentric video and audio dataset capturing naturalistic everyday experiences in in children aged 3–5 years and including detailed activity labels. A total of 106 hours of experiences were recorded from 61 children in their home environment using a 140° wide-lens camera equipped with a microphone embedded in a child-friendly vest. Annotations include five location classes and 14 activity classes, covering audio-only, video-only, and multimodal activities. Captured through a vest equipped with an embedded camera, ChildLens provides a rich resource for analyzing children's daily interactions and behaviors. We provide an overview of the dataset, the collection process, and the labeling strategy. Additionally, we present benchmark performance of two state-of-the-art models on the dataset: the Boundary-Matching Network for temporal activity localization and the Voice-Type Classifier for detecting and classifying speech in audio. Finally, we analyze the dataset specifications and their influence on model performance. The ChildLens dataset will be freely available for research purposes via an institutional repository (listed on the ChildLens website). It provides rich data to advance computer vision and audio analysis techniques and thereby removes a critical obstacle for the study of the context in which children develop.

Keywords: child development, egocentric video, audio dataset, multimodal learning, computer vision, developmental psychology

ChildLens: An Egocentric Video Dataset for Activity Analysis in Children

Introduction

In developmental psychology, everyday experiences play a central role when theorizing about the causes and dynamics of developmental change (Carpendale & Lewis, 2020; Heyes, 2018; Piaget, 1964; Rogoff, Dahl, & Callanan, 2018; Smith, Jayaraman, Clerkin, & Yu, 2018; Tomasello, 2009; Vygotsky, 1978). Famously, the two central processes in Piaget's theory of cognitive development - assimilation and accommodation - describe how children's cognitive abilities change in line with the experiences they make (Vygotsky, 1978). Vygotsky emphasized the role of social interactions between children and (knowledgeable) adults for the acquisition of culturally relevant knowledge (Vygotsky, 1978). Contemporary theorists rest on similar ideas. For example, Tomasello (2009) pointed out how everyday social interactions, particularly those involving shared intentionality, foster uniquely human forms of communication, cooperation and cognition. For Heyes (2018), culturally evolved "cognitive gadgets" are transmitted via language in conversations between adults and children.

These broad theoretical accounts are based on empirical support. Reviewing all such studies is beyond the scope of this paper so we want to point out only a few examples from the domain of language development. Across languages and cultural settings, the amount of language children hear is related to their language development (Bergelson et al., 2023a). There is also a relationship between the amount of conversational turn-taking children are engaged in and their vocabulary growth (Donnelly & Kidd, 2021; Ferjan Ramírez, Lytle, & Kuhl, 2020). Roy, Frank, DeCamp, Miller, and Roy (2015) showed that words that are heard in distinct contexts at distinct times are more likely to be learned. Ruffman et al. (2023) used head-mounted video cameras to study how repeated behaviors in everyday life correlate with the acquisition of mental state vocabulary. Taken together, such studies illustrate how important is to study naturalistic everyday experiences to understand

children's development. However, studies linking everyday experience and development are still vastly underrepresented (De Barbaro & Fausey, 2022; Rogoff et al., 2018) because they come with a set of unique challenges.

Perhaps the most significant obstacle in this field is the extensive amount of data needed to comprehensively study children's everyday experiences. Traditional methods, such as manual annotation, are time-consuming and impractical for large-scale datasets. To address this, Computer Vision or Natural Language Processing models offer scalable solutions for analyzing social interactions and behaviors. For instance, OpenPose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018) allows the tracking of human body, face, and hand poses which provides valuable insights into gestures and engagement. YOLOv8 (Redmon, Divvala, Girshick, & Farhadi, 2015) offers efficient object detection and models like I3D (Carreira & Zisserman, 2017) provide an automated solution for classifying activities in video data. For audio, Wave2Vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020) provides robust speech-to-text and speech representation capabilities, enabling the study of conversational dynamics. Together, these models facilitate the efficient analysis of multimodal data. However, even the best model architecture needs diverse, high-quality datasets to learn from. A notable example of such a dataset is ImageNet (Russakovsky et al., 2014); in fact, the release of this dataset sparked the development of some of the most prominent Computer Vision models available today. Similarly, expanding publicly available datasets in developmental psychology could accelerate progress in studying children's everyday experiences.

Several publicly available datasets have made valuable contributions to our understanding of children's social and communicative behavior. For example, the SAYCam dataset (Sullivan, Mei, Perfors, Wojcik, & Frank, 2021) provides audio-video recordings from 3 infants (6–32 months) who wore head-mounted cameras over two years, to capture naturalistic speech and behaviors. Similarly, the DAMI-P2C dataset (Chen, Alghowinem, Jang, Breazeal, & Park, 2023) includes audio and video recordings of parent-child

interactions during story reading, with annotations for body movements in a controlled environment. The MMDB dataset (Rehg et al., 2013) offers multimodal data (audio, video, physiological) of children (15–30 months) engaged in semi-structured play interactions, recorded in a lab. Another example is the UpStory dataset (Fraile et al., 2024), which features audio and video of primary school children (8-10 years) in dyadic storytelling interactions, also recorded in a lab setting. Additionally, the BabyView dataset (Long et al., 2024) provides high-resolution, egocentric video of children aged 6 months to 5 years, recorded at home and in preschool environments, with annotations for speech transcription and pose estimation. Whereas these datasets vary in age, setting, and target behaviors, they collectively highlight the need for more naturalistic, at-home datasets that can capture the full range of children's daily activities.

To address this gap, we introduce the publicly available ChildLens dataset, which focuses on activity annotations for children aged 3–5 years. The dataset consists of 106 hours of video and audio recordings collected from 61 children in their home environment through a 140° wide-lens camera, recording video and audio, which was embedded in a child-friendly vest. It includes detailed annotations for five location classes and 14 activity classes, which are further categorized based on whether the child is interacting alone or with others. Designed to support research in developmental psychology and computer vision, the ChildLens dataset offers a rich resource for advancing multimodal learning and studying the full spectrum of children's daily activities. As of now, 49% of the dataset are annotated. The remaining videos will be annotated when new funding is available.

Dataset Generation

This section outlines the steps taken to create the ChildLens dataset. We provide detailed information on the video collection process, the labeling strategy employed, and the generation of activity labels.

Step 1: Collection of Egocentric Videos

The ChildLens dataset consists of egocentric videos recorded by children aged 3 to 5 years over a period of 12 months. A total of 61 children from families living in a mid-sized city in Germany, participated in the study. The videos were captured at home using a camera embedded in a vest worn by the children, as shown in figure 1. This setup allowed the children to move freely throughout their homes while recording their activities. The camera, a *PatrolEyes WiFi HD Infrared Police Body Camera*, was equipped with a 140° wide-angle lens and captured the space in front of the child with a resolution of 1920x1080p at 30 fps. The camera also recorded high-quality audio, allowing us to capture the child's speech and other sounds in the environment.

In order to obtain a decent coverage for the different activity classes we planned to annotate, we handed parents a short checklist of activities to record. The focus was on capturing everyday activities that children typically engage in. Parents were therefore asked to include the following elements in the recordings:

- Child spends time in different rooms and performs various activities in each room
- Child is invited to read a book together with an adult
- Child is invited to play with toys alone
- Child is invited to play with toys with someone else (adult or child)
- Child is invited to draw/craft something

Step 2: Creation of Labeling Strategy

To create a comprehensive labeling strategy for the ChildLens dataset, we first defined a list of activities that children typically engage in. This list was inspired by previous research on activities that children are known to participate in (Ginsburg, and the Committee on Communications, & and the Committee on Psychosocial Aspects of Child and Family Health, 2007; Hofferth & Sandberg, 2001). From this, we derived a detailed

catalog of activities that were likely to be captured in the videos and chose to make the activity classes more granular by distinguishing between activities like “making music” and “singing/humming” or “drawing” and “crafting things”.

After an initial review of the videos, we decided to add another class “overheard speech” to capture situations in which the child is not directly involved in a conversation but can hear it. We also added “pretend play” as a separate class to capture situations in which the child is engaged in imaginative play. This approach allowed us to capture the diversity of activities that children engage in and create a comprehensive dataset for activity analysis.

Step 3: Manual Labeling Process

Before the actual annotation process, a setup meeting was held to introduce the annotators to the labeling strategy. To familiarize themselves with the task, the annotators were assigned 25 sample videos to practice and gain hands-on experience. These initial annotations were reviewed by the research team, and feedback was provided to refine the approach. A total of three feedback loops were conducted to ensure that the annotators follow the labeling strategy properly.

The videos were manually annotated by native German speakers who watched each video and labeled the activities present in the footage. Annotators marked the start and end points of each activity. For audio annotations, we implemented a 2-second rule for the categories ‘other person talking’ and ‘child talking’: if the break between two utterances was 2 seconds or less, it was considered a single event; breaks longer than 2 seconds split the activity into separate instances.

Dataset Overview

Activity Classes. The ChildLens dataset includes 14 activity classes and 5 location classes. A brief description of each class can be found in 1. The location classes describe the current location of the child in the video and include *livingroom*, *playroom*, *bathroom*, *hallway*, and *other*. The activity classes are categorized based on the child's interactions within the video and can be divided into *person-only* activities (e.g. "child talking", "other person talking"), and *person-object* activities (e.g. "drawing", "playing with object"). These activities are further categorized into *audio-based*, *visual-based*, and *multimodal* activities, as presented in Figure 1. Below is an overview of the different activity types:

- **Audio-based activities**: *child talking*, *other person talking*, *overheard speech*, *singing / humming*, *listening to music / audiobook*
- **Visual-based activities**: *watching something*, *drawing*, *crafting things*, *dancing*
- **Multimodal activities**: *playing with object*, *playing without object*, *making music*, *pretend play*, *reading book*

Statistics. The ChildLens dataset comprises of 343 video files with a total of 106.10 hours recorded by 61 children aged 3 to 5 years (M=4.52, SD=0.92). This includes 107 videos from children aged 3, 122 videos from children aged 4, and 114 videos from children aged 5. The video duration per child varies between 4.03 and 303.42 minutes (M=104.37, SD=51.65). A detailed distribution of the video duration per child is shown in figure 2.

This diverse dataset includes a varying number of instances across the 14 activity classes. While annotations are still ongoing, the current annotations (49% annotated) include between 2 and 319 instances per class. The duration of each instance varies by activity. For instance, audio-based activities like "child talking" may last only a few seconds, while activities like "reading a book" can span several minutes. The table with the

Table 1

Activity classes in the ChildLens dataset.

Activity Class	Description: The child is ...
Child talking	... talking to themselves or to someone else.
Singing/Humming	... singing or humming a song or a melody.
Listening to music/audiobook	... listening to music or an audiobook.
Watching something	... watching a movie or video on either a screen or a device.
Drawing	... drawing or coloring a picture.
Crafting things	... engaged in a craft activity, such as making a bracelet.
Dancing	... dancing to music or moving to a rhythm.
Playing with object	... playing or interacting with an object, such as a toy or a ball.
Playing without object	... playing without an object, such as playing hide and seek
Pretend play	... engaged in imaginative play, such as pretending to be a doctor.
Reading a book	... reading a book or looking at pictures in a book.
Making music	... playing a musical instrument or making music in another way.
Other person talking	Another person is talking to the child.
Overheard Speech	Conversations the child can hear but is not directly involved in.

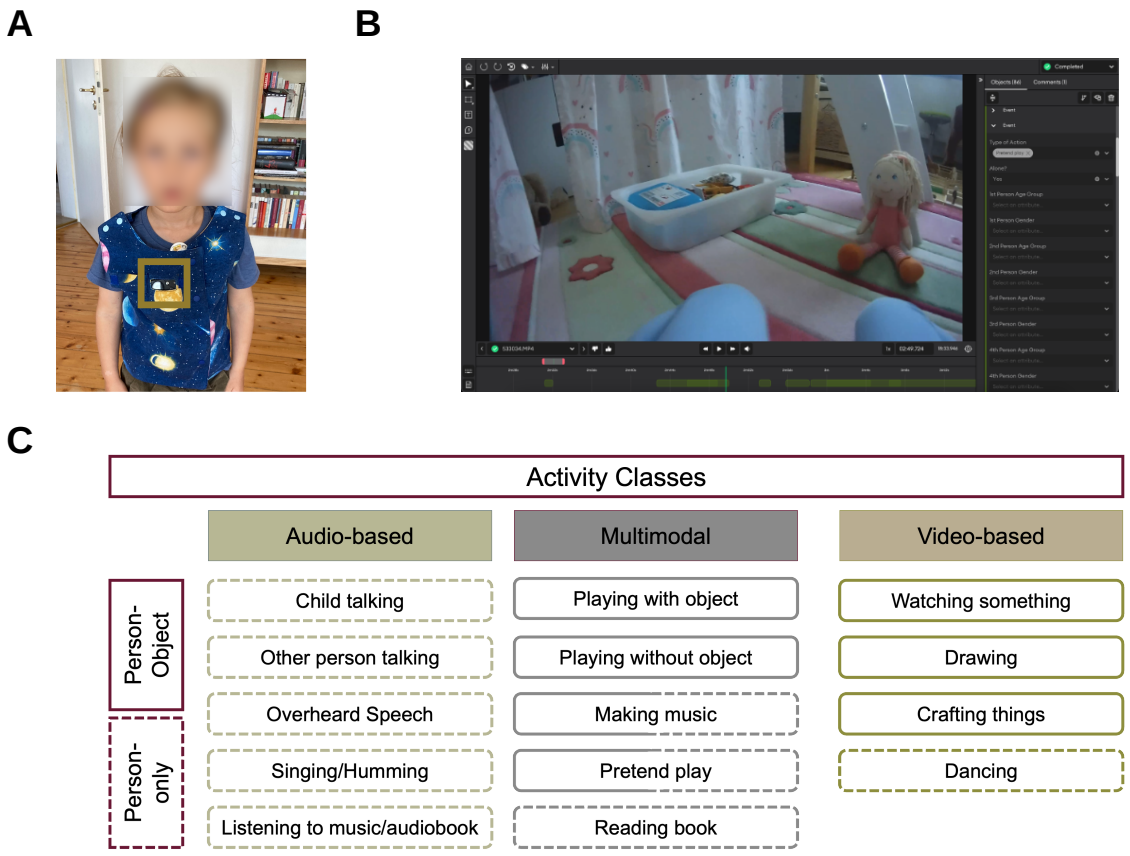

Figure 1. **A** – Vest with the embedded camera worn by the children, **B** – Platform utilized for video annotation, **C** – Activity classes in the ChildLens dataset.

total number of instances and summed duration for all activity classes is available in the appendix.

Data Access. The ChildLens dataset will be made available to scientists for research purposes. It includes video and audio recordings, along with activity labels. Due to the sensitive nature of the data—recordings of children in their homes—access will be restricted.

Researchers can submit requests for access through the , which will be carefully reviewed to ensure proper handling and compliance with privacy standards. Please contact ra@eva.mpg.de to request access to the dataset. As noted above, the annotation process is still ongoing and the dataset will be updated regularly. A brief project overview,

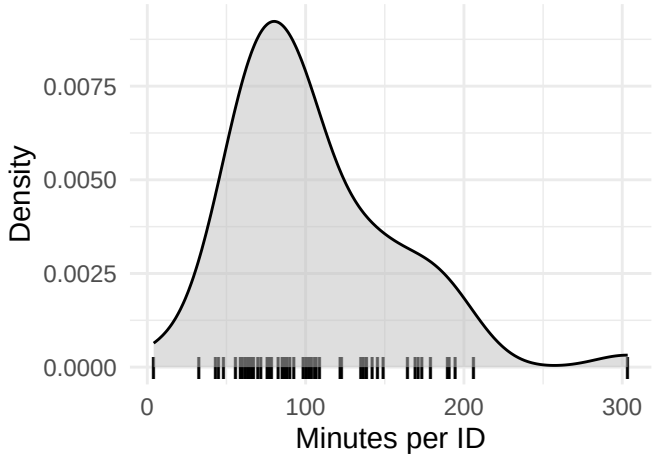

Figure 2. Video recording duration (in minutes) per child ID.

information about how to access the dataset along with the latest dataset version can be found here

Exhaustive multi-label annotations. The dataset provides detailed annotations for each video file. These annotations specify the child's current location within the video, the start and end times of each activity, the activity class, and whether the child is engaged alone or with somebody else. For every person involved in the activity, we capture age class and gender. If multiple activities occur simultaneously in a video, each activity is individually labeled. For example, if a segment shows a child "reading a book" while also "talking," two separate annotations are created: one for "reading a book" and another for "child talking." This exhaustive labeling strategy ensures that each activity is accurately represented in the dataset.

Benchmark Performance

In this chapter, we present the results of applying two model architectures to the ChildLens dataset for two specific tasks: temporal activity localization using video data and voice type classification using audio data. For temporal activity localization, we used the Boundary-Matching Network (BMN) model (Lin, Liu, Li, Ding, & Wen, 2019), a

state-of-the-art approach in this domain, and trained it from scratch on the unique activity classes in the ChildLens video data. For voice type classification, we applied the Voice Type Classifier (VTC) (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020), also state-of-the-art, which was trained on similar data. Both models provide initial results and establish a benchmark for future research.

Temporal Activity Localization

The BMN generates action proposals by predicting activity start and end boundaries and classifying these proposals into activity classes. The architecture consists of two main components: (1) a proposal generation network, which identifies candidate proposals, and (2) a proposal classification network, which classifies these proposals. The model prioritizes proposals with high recall and high temporal overlap with ground truth. BMN performance is evaluated using Average Recall (AR) and Area Under the Curve (AUC) metrics. AR is computed at various Intersection over Union (IoU) thresholds and for different Average Numbers of Proposals (AN) as AR@AN, where AN ranges from 0 to 100. AR@100 reflects recall performance with 100 proposals per video, while AUC quantifies the trade-off between recall and number of generated proposal. On the ActivityNet-1.3 test set (Heilbron, Escorcia, Ghanem, & Niebles, 2015), BMN demonstrates effective activity localization with an AR@100 of 72.46 and an AUC of 64.47.

Data Preparation. The BMN implementation, including video preprocessing and model training, was conducted using the MMAction2 toolbox (Contributors, 2020). Data preparation involved several key steps, such as raw frame extraction and the generation of both RGB and optical flow features for each video. Before training the model, we analyzed the distribution of activity instances across the classes for the annotated videos to assess the data's sufficiency for both training and testing. A detailed summary of the activity instances and their total durations can be found in the appendix.

Our analysis highlighted a significant class imbalance in the dataset, both in terms of

instance count and the total duration of recordings. Given the primary goal of establishing initial benchmark results, no data augmentation methods were employed to mitigate this imbalance. Instead, we focused on the more frequent activity classes, which also had the longest durations: “Playing with Object” (22.85 hours of recording), “Drawing” (6.24 hours of recording), and “Reading a Book” (5.48 hours of recording).

For feature extraction and model training optimization, the videos were divided into clips of 4000 frames each (correspond to approx. 2 minutes and 13 seconds). This resulted in a total of 1130 clips. However, only 995 clips had annotations, so we split these annotated clips into training, validation, and test subsets, using an 80-10-10 split. The training set was used for model optimization, the validation set guided hyperparameter tuning and overfitting prevention, and the test set was reserved for evaluating the model’s generalization ability on unseen data.

Implementation Details. The BMN model was trained from scratch on the ChildLens dataset to predict the start and end boundaries of activity classes in the videos. The model was implemented using MMAction2, an open-source toolbox for video understanding based on PyTorch (Contributors, 2020). Training took place on a Linux server with 48 cores and 187 GB of RAM. The Adam optimizer was used with a learning rate of 0.001 and a batch size of 4. To avoid overfitting, early stopping based on validation loss was applied during training.

Evaluation. The performance of the BMN model on the ChildLens dataset, compared to its evaluation on ActivityNet-1.3, is summarized in Table 2, with AR@100 and AUC reported for both datasets. The results indicate that the BMN model generalizes well to the new domains included in the ChildLens dataset. These benchmark results highlight the potential for integrating the ChildLens dataset with existing models like BMN. Automating the analysis of this dataset can streamline the study of children’s activities and interactions, facilitating more efficient research in developmental psychology and related fields.

Table 2

Comparison of BMN performance on the ActivityNet-1.3 dataset (used for model evaluation) and the ChildLens dataset, highlighting the Average Recall for 100 proposals (AR@100) and the Area Under the Curve (AUC).

Dataset	AR@100	AUC
ActivityNet-1.3	72.46	64.47
ChildLens	77.43	69.21

Voice Type Classification

Voice Type Classification is the task of identifying audio utterances and assigning them to predefined classes. In line with previous work, we focus on the following classes: `Key Child (KCHI)`, `Other Child (CHI)`, `Male Speech (MAL)`, `Female Speech (FEM)`, and `Speech (SPEECH)` (Lavechin et al., 2020). The Voice Type Classifier model (VTC) is designed to perform this task efficiently. Its architecture is composed of a SincNet layer, two bi-directional LSTMs, and three feed-forward layers. The model takes a 2-second audio chunk as input and outputs a score between 0 and 1 for each class. The VTC was originally trained on the BabyTrain dataset which includes 260 hours of child-centered audio in multiple languages from children mostly aged 0-3 years.

The model architecture utilizes the open-source pyannote library for speaker diarization (Bredin, 2023; Plaquet & Bredin, 2023) which provides pre-trained models and pipelines for various audio processing tasks. By adjusting the model architecture and

training process using pyannote, we evaluated the ChildLens dataset's quality through three distinct VTC training setups: We first applied the pretrained VTC model directly to the ChildLens dataset. In the second setup we fine-tuned the VTC model on the ChildLens data. Finally, we trained the VTC model from scratch using only the ChildLens dataset. These setups test the dataset's annotation consistency and standalone value, with performance measured by the F_1-measure, which combines precision and recall.

Data Preparation. We used the following mapping strategy to align our audio-based activity classes with the VTC's output classes. Minutes in parentheses indicates the total duration of annotated audio for each class.

- Child talking & Singing/Humming → **`Key Child`** (859.27 min)
- Other person talking:
  - If age = `"Child"` → **`Other Child`** (43.98 min)
  - If age = `"Adult"` & gender = `"Female"` → **`Female Speech`** (455.90 min)
  - If age = `"Adult"` & gender = `"Male"` → **`Male Speech`** (200.41 min)
- Overheard Speech, Child talking, Singing/Humming, Other person talking → **`Speech`** (2361.87 min)

"Listening to music/audiobook" was excluded, as it's often background audio, lacks speaker age/gender details, and includes music irrelevant to VTC. "Overheard Speech" refers to speech not directed at the key child and was mapped to SPEECH due to missing age/gender annotations. This may underestimate VTC performance, for example, correct FEM predictions have SPEECH as ground truth. We retain this mapping, as re-annotation would be time-consuming. Moreover, the class is conceptually incompatible with the VTC output: it captures the recipient of speech rather than the speaker type. It is intended for more advanced models (for instance NLP architectures) that can infer not just who is talking, but also to whom, based on semantic content.

Table 3

Comparison of Voice Type Classifier (VTC) performance on the ACLEW-Random dataset and the three setups utilizing the ChildLens dataset. The evaluation metrics are reported for the original VTC model (VTC-OG), the VTC fine-tuned on ChildLens (VTC-FT) and the VTC trained from scratch on ChildLens (VTC-CL). The table reports the F1 score per class and the average F1 score (AVG).

Dataset	Model	KCHI	CHI	MAL	FEM	SPEECH	AVG
ACLEW-Random	VTC-OG	68.7	33.2	42.9	63.4	78.4	57.3
ChildLens	VTC-OG	68.5	4.5	19.6	46.7	82.2	44.6
ChildLens	VTC-FT	75.7	12.0	39.2	54.2	85.2	53.2
ChildLens	VTC-CL	74.7	9.6	53.8	52.5	84.7	55.0

Implementation Details. We first applied the pretrained VTC model directly, using the available VTC implementation. We then fine-tuned the model on our ChildLens dataset for 200 epochs (12.86 hours) on the same Linux server as the BMN model. Finally, we trained the VTC model from scratch for 200 epochs (12.86 hours) using the ChildLens dataset to assess its standalone value

Evaluation. Table 3 shows F_1-scores for three VTC setups on the ChildLens dataset: the original model VTC_{og}, fine-tuned model VTC_{ft}, and model trained from scratch VTC_{cl}, compared to the benchmark dataset. VTC_{og} scores 44.60, slightly below the benchmark, with best performance on `KCHI` (68.50) and worst on `CHI` (4.50). Fine-tuning VTC_{ft} training from scratch VTC_{cl} achieve 53.20 and 55, respectively, showing significant improvement. The low `CHI` scores are primarily due to frequent misclassifications as `KCHI`. This likely results from the vocal similarity between the key child and other children in the ChildLens dataset. In many cases, distinguishing between

them may rely more on acoustic cues like proximity to the microphone than on vocal characteristics. In contrast, the original VTC dataset includes audio from infants and toddlers, where the distinction between the key child (often babbling or younger) and other speakers is more pronounced, making classification tasks easier. Importantly, these scores are expected to improve as more annotated data becomes available, since annotation is still ongoing. Figure 3 visualizes how model predictions compare to the ground truth annotations for the VTC_{ft} model.

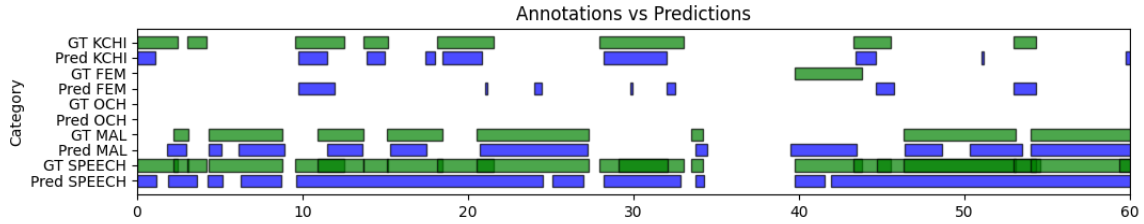

Figure 3. VTC Predictions compared to Ground Truth Annotations

General Discussion

We present the ChildLens dataset, a unique egocentric video-audio dataset that documents naturalistic everyday experiences,in preschool children. This dataset is particularly distinctive due to its diversity in terms of the number of children it includes and the variety of activity labels it covers. By encorporating both visual and auditory data, the ChildLens dataset provides comprehensive annotations for a broad spectrum of key activities, including multi-modal social interactions. These annotations support the training and evaluation of models for the automatic analysis of children's activities and therefore allow for scaling up data collection. A rich and representative dataset like ChildLens is essential for understanding individual differences in children's daily experiences and how they relate to different cognitive and social outcomes.

In comparison to other freely available datasets, the ChildLens dataset stands out due to its broad age span and diverse set of activity labels. Most existing datasets either focus

on toddlers, are limited to dyadic interactions or were recorded in lab settings, with all of them lacking a comprehensive range of activity labels. Furthermore, most of these datasets either capture only audio or video. In contrast, ChildLens includes naturalistic recordings from children's home environments, collected over an extended period, and features a wide variety of activity types. Particularly noteworthy is the activity class `Overheard Speech` which captures speech that children can hear but are not directly involved in. This class is important for studying the impact of overheard speech on children's cognitive development - an area that has largely relied on labor-intensive manual annotations due to the difficulty of automatically distinguishing between child-available and child-directed speech (Bergelson et al., 2023b). Existing models, like the classifier developed by Bang, Kachergis, Weisleder, and Marchman (2023) could be enhanced by incorporating visual features - such as eye contact or the use of gestures - in addition to audio inputs. The ChildLens dataset also captures whether children are engaged in activities alone or with others and provides basic demographic information - age, sex -about all individuals involved.

The usefulness of the ChildLens dataset is demonstrated by its successful application to well-established models. For example, the pretrained Voice-Type Classifier for audio transcription achieves performance comparable to previous datasets, while the Boundary-Matching Network (BMN) produces robust results for activity localization, consistent with its performance on widely used datasets such as ActiviyNet. One way of using the ChildLens dataset to advance methodological development are multi-method approaches. For example, activity localization could be further enhanced by incorporating object identification, allowing for better tracking of the objects children interact with during daily routines. Such an approach has beed used in adult-focused studies (Kazakos, Huh, Nagrani, Zisserman, & Damen, 2021). Research by Bambach, Lee, Crandall, and Yu (2015) also emphasizes the importance of hand detection in egocentric video for activity recognition. Their use of Convolutional Neural Networks (CNNs) for hand segmentation demonstrates how such techniques can help differentiate between activities. To apply a

similar approach to the ChildLens dataset, we would first need to run a pretrained hand detection model on the dataset. In a second step, we would train a CNN to classify frames containing hands into activity classes, following the method described by Bambach et al. (2015). If the performance of the hand detection performance is insufficient, additional hand annotations would be required to improve the model's accuracy.

The integration of visual and auditory data in the ChildLens dataset enables a more detailed and comprehensive understanding of children's daily experiences. Complex activities such as pretend play and reading a book, which require both audio and video for accurate detection, exemplify the strength of this multimodal approach. While previous studies, such as those analyzing disfluencies in children's speech during computer game play (Yildirim & Narayanan, 2009), have demonstrated that combining visual and auditory information can improve performance, few studies have explored this in the context of children's naturalistic activities. With ChildLens, the combination of naturalistic data and multimodal analysis creates new opportunities for in-depth insights into children's cognitive, emotional, and social development, particularly for activities best captured through both modalities.

Despite its strengths, the ChildLens dataset also has its limitations. First, there is class imbalance, especially in underrepresented activity classes, which could affect model training and evaluation. More frequent activities, such as "child talking" (7447 instances, 649 minutes) and "playing with object" (317 instances, 1371 minutes), dominate the dataset, whereas less common activities like "dancing" (2 instances, 0.57 minutes) and "making music" (2 instances, 2.13 minutes) are scarcely represented. Similarly, activities like "pretend play" (59 instances, 158.84 minutes) and "reading a book" (81 instances, 328.70 minutes) appear less frequently. This imbalance may lead to skewed model performance, making it harder to accurately classify rare activities. Possible solutions to this challenge could involve merging rare activity classes into broader categories or excluding them from model training, though these approaches may reduce the dataset's

diversity. Other methods, such as resampling or augmentation, could help balance the dataset and improve model performance (Alani, Cosma, & Taherkhani, 2020; Spelmen & Porkodi, 2018). Second, there is sampling bias. Since the recordings are largely influenced by parental decisions about when and how often activities are captured, certain activities or settings may be overrepresented or underrepresented based on these preferences. Furthermore, the dataset primarily focuses on families from a mid-sized German city, limiting its geographic and cultural diversity. Expanding the dataset to include a broader range of families from different regions and cultures would enhance its generalizability and applicability to various research contexts.

The study of children's everyday experiences is crucial for understanding their cognitive, emotional, and social development. These daily interactions provide important insights into how children learn, grow, and engage with their environment. The ChildLens dataset makes a valuable contribution to this field by offering a rich multi-modal resource that captures children's experiences in naturalistic settings. With its comprehensive annotations and potential to automate the analysis of children's activities, the dataset enables researchers to develop, fine-tune and apply automized processing algorithms that help to scale-up the study of situated development. By virtue of being an openly accessible resource, the ChildLens dataset creates new opportunities for understanding the complexities of early childhood development and provides a foundation for future research in this area.

References

Alani, A. A., Cosma, G., & Taherkhani, A. (2020). Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. Glasgow, United Kingdom: IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207697

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. https://doi.org/10.48550/ARXIV.2006.11477

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1949–1957. Santiago, Chile: IEEE. https://doi.org/10.1109/ICCV.2015.226

Bang, J. Y., Kachergis, G., Weisleder, A., & Marchman, V. (2023). An automated classifier for periods of sleep and target-child-directed speech from LENA recordings. *Language Development Research*, *3.0*(1.0). https://doi.org/10.34842/XMRQ-ER43

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., … Cristia, A. (2023a). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120. https://doi.org/10.1073/pnas.2300671120

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., … Cristia, A. (2023b). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120. https://doi.org/10.1073/pnas.2300671120

Bredin, H. (2023). Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe. *INTERSPEECH 2023*, 1983–1987. ISCA. https://doi.org/10.21437/Interspeech.2023-105

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: Realtime

Multi-Person 2D Pose Estimation using Part Affinity Fields. https://doi.org/10.48550/ARXIV.1812.08008

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

Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. https://doi.org/10.48550/ARXIV.1705.07750

Chen, H., Alghowinem, S., Jang, S. J., Breazeal, C., & Park, H. W. (2023). Dyadic Affect in Parent-Child Multimodal Interaction: Introducing the DAMI-P2C Dataset and its Preliminary Analysis. *IEEE Transactions on Affective Computing*, *14*(4), 3345–3361. https://doi.org/10.1109/TAFFC.2022.3178689

Contributors, M. (2020). *OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark*. Retrieved from urlhttps://github.com/open-mmlab/mmaction2

De Barbaro, K., & Fausey, C. M. (2022). Ten Lessons About Infants' Everyday Experiences. *Current Directions in Psychological Science*, *31*(1), 28–33. https://doi.org/10.1177/09637214211059536

Donnelly, S., & Kidd, E. (2021). The Longitudinal Relationship Between Conversational Turn-Taking and Vocabulary Growth in Early Language Development. *Child Development*, *92*(2), 609–625. https://doi.org/10.1111/cdev.13511

Ferjan Ramírez, N., Lytle, S. R., & Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, *117*(7), 3484–3491. https://doi.org/10.1073/pnas.1921653117

Fraile, M., Calvo-Barajas, N., Apeiron, A. S., Varni, G., Lindblad, J., Sladoje, N., & Castellano, G. (2024). UpStory: The Uppsala Storytelling dataset. https://doi.org/10.48550/ARXIV.2407.04352

Ginsburg, K. R., and the Committee on Communications, & and the Committee on Psychosocial Aspects of Child and Family Health. (2007). The Importance of Play in

Promoting Healthy Child Development and Maintaining Strong Parent-Child Bonds. *Pediatrics*, *119*(1), 182–191. https://doi.org/10.1542/peds.2006-2697

Heilbron, F. C., Escorcia, V., Ghanem, B., & Niebles, J. C. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 961–970. Boston, MA, USA: IEEE. https://doi.org/10.1109/CVPR.2015.7298698

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

Hofferth, S. L., & Sandberg, J. F. (2001). How American Children Spend Their Time. *Journal of Marriage and Family*, *63*(2), 295–308. https://doi.org/10.1111/j.1741-3737.2001.00295.x

Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., & Damen, D. (2021). With a Little Help from my Temporal Context: Multimodal Egocentric Action Recognition. https://doi.org/10.48550/ARXIV.2111.01024

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

Lin, T., Liu, X., Li, X., Ding, E., & Wen, S. (2019). BMN: Boundary-Matching Network for Temporal Action Proposal Generation. https://doi.org/10.48550/ARXIV.1907.09702

Long, B., Xiang, V., Stojanov, S., Sparks, R. Z., Yin, Z., Keene, G. E., … Frank, M. C. (2024, June 14). The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. https://doi.org/10.48550/arXiv.2406.10447

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

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *INTERSPEECH 2023*, 3222–3226. ISCA.

https://doi.org/10.21437/Interspeech.2023-205

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. https://doi.org/10.48550/ARXIV.1506.02640

Rehg, J. M., Abowd, G. D., Rozga, A., Romero, M., Clements, M. A., Sclaroff, S., … Ye, Z. (2013). Decoding Children's Social Behavior. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 3414–3421. Portland, OR, USA: IEEE. https://doi.org/10.1109/CVPR.2013.438

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

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Ruffman, T., Chen, L., Lorimer, B., Vanier, S., Edgar, K., Scarf, D., & Taumoepeau, M. (2023). Exposure to behavioral regularities in everyday life predicts infants' acquisition of mental state vocabulary. *Developmental Science*, *26*(4), e13343. https://doi.org/10.1111/desc.13343

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. https://doi.org/10.48550/ARXIV.1409.0575

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

Spelmen, V. S., & Porkodi, R. (2018). A Review on Handling Imbalanced Data. *2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT)*, 1–11. Coimbatore: IEEE. https://doi.org/10.1109/ICCTCT.2018.8551020

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A Large,

Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective. *Open Mind*, *5*, 20–29. https://doi.org/10.1162/opmi_a_00039

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.

Yildirim, S., & Narayanan, S. (2009). Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio–Visual Information. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(1), 2–12. https://doi.org/10.1109/TASL.2008.2006728

Table 4

Number of video instances and the total duration (in minutes).

Category	Activity Class	Instance Count	Total Duration (min)
Audio	Child talking	7447	649.10
	Other person talking	6113	455.29
	Overheard Speech	1898	299.44
	Singing/Humming	277	82.00
	Listening to music/audiobook	68	222.14
Video	Watching something	2	5.09
	Drawing	62	374.91
	Crafting things	26	109.14
	Dancing	2	0.57
Multimodal	Playing with object	317	1371.06
	Playing without object	25	28.87
	Pretend play	59	158.84
	Reading a book	81	328.70
	Making music	3	2.13

Appendix

Activity Class Statistics