ChildLens: An Egocentric Video Dataset for Activity Analysis in Children

Nele-Pauline 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, a novel egocentric video and audio dataset of children aged 3–5 years, featuring detailed activity labels. Spanning 106 hours of recordings, the dataset includes 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 made available for research purposes, providing rich data to advance computer vision and audio analysis techniques while offering new insights into developmental psychology.

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 are crucial for shaping children's development (Carpendale & Lewis, 2020; Heyes, 2018; Piaget, 1964; Rogoff, Dahl, & Callanan, 2018; Smith, Jayaraman, Clerkin, & Yu, 2018; Tomasello, 2009; Vygotsky, 1978). Fundamental theories, such as Piaget's Learning Theory of Cognitive Development (Piaget, 1964), have long recognized the role of everyday interactions in helping children actively construct knowledge, whereas Vygotsky's Sociocultural Theory (Vygotsky, 1978) emphasized how social interactions help transform everyday sensory experiences into structured understanding. Building on these foundational insights, more recent studies have further explored how everyday experiences shape cognitive and social development. For instance, Spangler (Spangler, 1989) showed that toddlers' daily interactions shape their mental and emotional dispositions, predicting later developmental outcomes. Similarly, Tomasello's Cultural Learning Theory (Tomasello, 2009) pointed out how everyday social interactions, particularly those involving shared intentionality, foster uniquely human cognitive abilities by enabling children to understand others' intentions and perspectives. Further expanding on this, Heyes's work on the Cultural Evolution of Thinking (Heyes, 2018) highlighted the importance of experiences like imitation and informal social learning in developing cognitive capacities. Debarbaro et al. (De Barbaro & Fausey, 2022) summarized various studies, emphasizing the need to analyze infants' dynamic, diverse experiences captured through everyday activity sensors, and stressed the significance of long-term analysis to understand developmental patterns and variability. Despite this growing body of work, direct research connecting the diversity of children's daily experiences to broader developmental trajectories remains limited. Whereas many studies focus on specific domains such as language or social cognition, there remains a need for more comprehensive investigations into how diverse daily experiences shape developmental

trajectories.

In the context of children's developmental trajectories, research has focused on areas like language acquisition, theory of mind, and social cognition, utilizing a range of methods and data sources. For instance, Donnelly et al. (Donnelly & Kidd, 2021) used audio-only data to explore the relationship between conversational turn-taking and vocabulary growth in children, whereas Roy et al. (Roy, Frank, DeCamp, Miller, & Roy, 2015) examined how words used in specific contexts are learned more easily, emphasizing the importance of multimodal contexts. In contrast, Rowe (Rowe & Goldin-Meadow, 2009) leveraged video data to investigate how gestures at 14 months predict vocabulary development in children from different socioeconomic backgrounds. Ruffman et al. (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, supporting the minimalist view of theory of mind development. Bergelson (Bergelson et al., 2023), on the other hand, used large-scale audio data to explore the impact of adult speech on children's language production across diverse cultural contexts. These studies demonstrate the value of both audio and video data in understanding children's development, yet they highlight the need for datasets that capture the full diversity of children's everyday experiences.

A significant challenge in this field is the extensive amount of data needed to comprehensively study children's daily lives. Traditional methods, such as manual annotation, are time-consuming and impractical for large-scale datasets. To address this, computational 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, providing insights into gestures and engagement. YOLOv8 (Redmon, Divvala, Girshick, & Farhadi, 2015) offers efficient object detection for analyzing children's interactions with their environment, whereas models like I3D (Carreira & Zisserman, 2017) provide an automataed 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, but their improvement and development depend on the availability of diverse, high-quality datasets. A notable example of such a dataset is ImageNet (Russakovsky et al., 2014), which has been crucial in advancing computer vision models. 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 infants (6–32 months) who wore head-mounted cameras over two years, capturing 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 and captures their naturalistic experiences at home. The dataset consists of 106 hours of video and audio recordings

collected from 61 children wearing camera-equipped vests. It includes detailed activity annotations for five location classes and 14 activity classes, categorizing activities based on whether the child is interacting alone or with others. These annotations, labeled with start and end times, provide a granular view of children's everyday behaviors, crucial for understanding their developmental trajectories. 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.

Dataset Overview

Activity Classes. The ChildLens dataset includes 14 activity classes and 5 location classes. 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"). A brief description of each class can be found in the appendix. 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*

The location classes describe the current location of the child in the video and include *livingroom*, *playroom*, *bathroom*, *hallway*, and *other*.

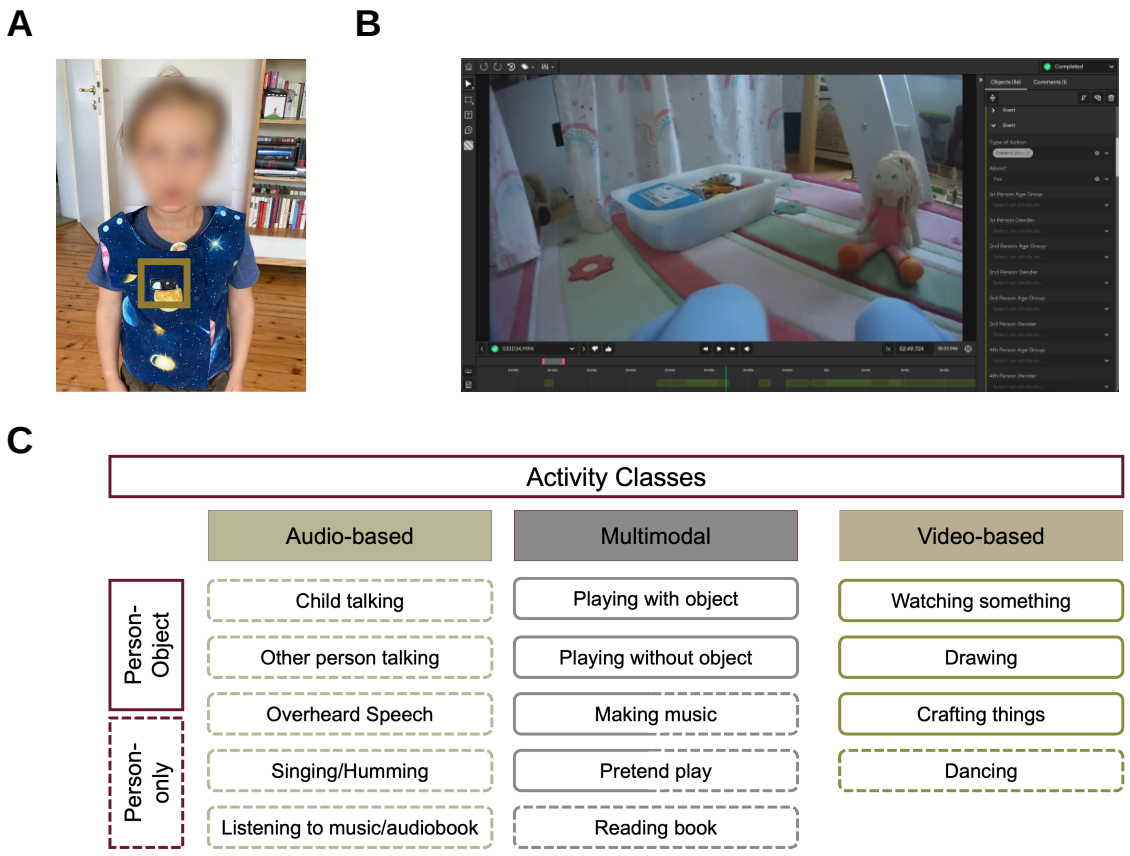

Figure 1. **A** – Vest with the embedded camera worn by the children, **B** – SuperAnnotate platform utilized for video annotation, **C** – Activity classes in the ChildLens dataset.

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, with each class containing 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 total number of instances and summed duration for all

activity classes is available in the appendix.

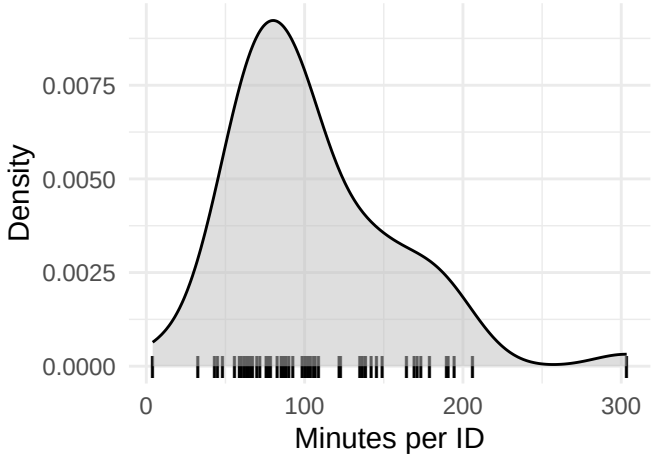

Figure 2. Video recording duration (in minutes) per child ID.

Data Access. The ChildLens dataset will be made available for research purposes, providing a rich resource for studying children's daily activities and interactions. The dataset 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.

As the annotation process is still ongoing, the dataset will be updated regularly. A brief project overview and the latest dataset version can be found here Researchers can submit requests for access through the , which will be carefully reviewed to ensure proper handling and compliance with privacy standards. Please contact person to request access to the dataset.

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 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.

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-degree wide-angle lens and captured everything within the child's field of view with a resolution of 1920x1080p at 30 fps. The camera also recorded audio, allowing us to capture the child's speech and other sounds in the environment. Additionally, the parents were handed a small checklist of activities to record, ensuring that a variety of activities were represented in the videos. 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 based on previous research on child development and the activities that children are known to participate in. We then developed 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. The annotations were conducted using the SuperAnnotate platform, as shown in figure 1, allowing for efficient annotation and review of the videos.

Benchmark Performance

In this chapter, we present the results of applying two model architectures to the ChildLens dataset. While the dataset supports multimodal activity analysis, we focus on two specific tasks: temporal activity localization using video data and voice type classification using audio data. For temporal activity localization, we use the Boundary-Matching Network (BMN) model, a state-of-the-art approach in this domain, and train it from scratch on the unique activity classes in the ChildLens video data. For voice type classification, we apply 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

We employ the Boundary-Matching Network (BMN) (Lin, Liu, Li, Ding, & Wen, 2019) for temporal activity localization on our dataset. 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, BMN achieves an AR@100 of 72.46 and an AUC of 64.47, demonstrating its effectiveness in activity localization.

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 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, the focus was placed 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

Table 1

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

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 1, with AR@100 and AUC reported for both datasets. The results indicate that the BMN model generalizes well to new domains, such as children's everyday activities, despite the ChildLens dataset's focus on social and behavioral interactions in naturalistic settings. 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.

Voice Type Classification

Voice Type Classification can be described as the task of identifying utterances from audio chunks and assigning them to one of the predefined classes (Lavechin et al., 2020). In our case, we are interested in classifying the audio into five distinct voice types: `Key Child (KCHI)`, `Other Child (CHI)`, `Male Speech (MAL)`, `Female Speech (FEM)`, and `Speech (SPEECH)`. The Voice Type Classifier model (Lavechin et al., 2020) is designed to perform this task efficiently, leveraging the capabilities of the open-source pyannote library for speaker diarization (Bredin, 2023; Plaquet & Bredin, 2023).

To investigate the quality and utility of our annotated ChildLens dataset, we implemented and assessed three distinct Voice Type Classifier (VTC) training setups:

1. **Direct Application** of the pretrained VTC model to our data without any modifications as its original training data (260h of child-centered audio recordings closely resembles our dataset).
2. **Fine-tuning** the pretrained VTC model on our ChildLens dataset.
3. **Training the entire architecture from scratch** using only our annotated ChildLens dataset.

This multi-stage evaluation allows us to assess not just model performance, but also the robustness, richness, and consistency of our annotations. Improved results from fine-tuning and competitive performance when training from scratch suggest that our data provides strong, high-quality signal for voice type classification. In this regard, the task serves both as a benchmark and as validation of the training value and generalizability of the ChildLens dataset.

The VTC architecture processes audio by first dividing it into 2-second chunks. Each chunk is passed through a SincNet layer to extract low-level acoustic features, followed by a stack of two bi-directional LSTMs and three fully connected layers. The final output layer

applies a sigmoid activation to produce a score between 0 and 1 for each class. Model performance is evaluated by utilizing the F_1-measure, which combines precision and recall using the following formula:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where precision $= \frac{\text{tp}}{\text{tp+fp}}$ and recall $= \frac{\text{tp}}{\text{tp+fn}}$ with

- tp being the number of true positives,
- fp being the number of false positives, and
- fn being the number of false negatives.

It ranges from 0 to 1, with 1 representing perfect precision and recall. Generally, an F_1 score above 0.8 is considered good, while values above 0.9 are considered excellent. In certain use cases, a score around 0.5 can still be considered acceptable, depending on the balance between precision and recall. The F_1 score is computed per class and averaged to provide an overall measure. No collar is applied to the evaluation, meaning that the predictions have to be exact to be considered correct. The model achieves an F_1 score of 57.3, outperforming the previous state-of-the-art LENA model by 10.6 points.

Data Preparation. Before implementing the different VTC training setups, we mapped our audio-based activity classes to align with the VTC's output classes. The following mapping strategy was applied:

- Child talking → **`Key Child`** & **`Speech`**
- Singing/Humming → **`Key Child`** & **`Speech`**
- Other person talking:
  - If `age = "Child"` → **`Other Child`** & **`Speech`**
  - If `age = "Adult"` & `gender = "Female"` → **`Female Speech`** & **`Speech`**
  - If `age = "Adult"` & `gender = "Male"` → **`Male Speech`** & **`Speech`**

Table 2

Cumulative Duration (in minutes) of Utterances Categorized by Voice Type Class

	KCHI	CHI	MAL	FEM	SPEECH
Total Duration	731.16	40.29	114.18	280.53	1465.55

- Overheard Speech → **`Speech`**

The activity class "Listening to music/audiobook" was not mapped to any VTC class, as it is not covered by the VTC model. The mapping process resulted in new numbers for the total durations for each VTC class, as shown in Table 2.

Implementation Details. To evaluate the practical applicability of our annotated ChildLens dataset, we employed the VTC model in three different training setups, which are defined as follows: Given the similarities between the original training data and the ChildLens dataset, we first implemented the pretrained VTC model without further training. Second, we fine-tuned the pretrained model on the ChildLens dataset while retaining its original weights. We trained the model for 200 epochs on the same Linux server as the BMN model, with a total training time of 12.86 hours. Finally, we also trained the VTC model from scratch using only our annotated data to assess the dataset's standalone value. This setup used the same parameters as fine-tuning and required 200 epochs and 12.86 hours of training.

Evaluation. Table 3 shows the performance of the three different VTC setups on the ChildLens dataset—namely, the original model VTC_{og}, the fine-tuned model on ChildLens VTC_{ft}, and the model trained from scratch on ChildLens VTC_{cl}—compared to the benchmark dataset from the original study. The VTC_{og} model achieves an average F_1 score of `vtc_og_f1` on the ChildLens dataset, performing comparably to the benchmark

Table 3

omparison of Voice Type Classifier (VTC) performance on the ACLEW-Random dataset (used for evaluation) and the ChildLens dataset. Results include: the original pretrained VTC model, VTC fine-tuned on ChildLens (VTC-FT), VTC trained from scratch on ChildLens (VTC-CL), and a variant replacing the SincNet feature extractor with WavLM-Base (SSeRiouSS). The table reports the F1 score per class and the average F1 score (AVG).

Dataset	Architecture	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	60.8	4.6	21.9	38.7	74.6	40.1
ChildLens	VTC-FT	78.5	8.3	45.4	51.8	85.3	53.8
ChildLens	VTC-CL	79.2	7.2	41.0	51.4	81.7	52.1

dataset. It performs best on the `CHI` class with an F_1 score of **xx** and worst on the `MAL` class with an F_1 score of **xx** Compared to the benchmark dataset, the model performs significantly better on the `CHI` class but slightly worse on the `MAL` and `FEM` classes. Analysis of False Positives and False Negatives reveals that the most common confusion occurs between the `MAL` and `FEM` classes. This may be attributed to the deeper pitch of some female voices in the German language. Additionally, the model was trained on a dataset with a different language distribution and younger children, where adults, particularly females, may use a higher pitch when interacting with infants, unlike with older children. Figure 3 provides a visual representation of the VTC predictions compared to the ground truth annotations.

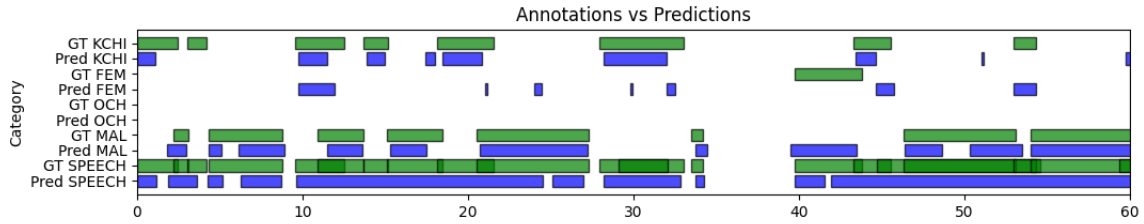

Figure 3. VTC Predictions compared to Ground Truth Annotations

General Discussion

We present the ChildLens dataset, a unique egocentric video-audio dataset that documents children's everyday experiences, spanning a wide age range of 3 to 5 years. 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 focusing on both visual and auditory data, the ChildLens dataset provides comprehensive annotations for a broad spectrum of key activities, offering rich insights into children's social and behavioral interactions in naturalistic settings. This makes it a valuable resource for studying developmental processes in children, with a focus on their cognitive, emotional, and social growth.

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 other datasets focus either 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, missing the multimodal aspect crucial for understanding children's everyday experiences. In contrast, ChildLens includes naturalistic recordings from children's home environments, over an extended period, and features a variety of activity types. The dataset also captures whether children are engaged in activities alone or with others and provides detailed demographic information about all individuals involved. This comprehensive approach enables a deeper and more holistic understanding of children's interactions and developmental trajectories.

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 other datasets. These results indicate that the ChildLens dataset's annotations align well with model predictions, highlighting its quality and potential for multimodal research. Moreover, the successful application of these models demonstrates how the dataset can support and automate the analysis of children's everyday activities.

Expanding the potential for multi-method approaches, activity localization could be further enhanced by incorporating object identification, allowing for better tracking of the objects children interact with during daily routines, as demonstrated in adult-focused studies (Kazakos, Huh, Nagrani, Zisserman, & Damen, 2021). Research by Bambach et al. (Bambach, Lee, Crandall, & Yu, 2015) also emphasizes the importance of hand detection in egocentric video for activity recognition. Their work on using Convolutional Neural Networks for hand segmentation highlights how such techniques can differentiate activities, offering a deeper understanding of children's interactions and behaviors.

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, one limitation of the ChildLens dataset is the 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).

In addition to class imbalance, another potential limitation is the 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 multimodal resource that captures the complexities of children’s lives in naturalistic settings. With its comprehensive annotations and potential to automate the analysis of children’s activities, the dataset enables researchers to gain deeper and more detailed insights into children’s

development. By making such analyses more efficient and accessible, 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

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., … Cristia, A. (2023). 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

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

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

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

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in Early Gesture Explain SES Disparities in Child Vocabulary Size at School Entry. *Science*, *323*(5916), 951–953. https://doi.org/10.1126/science.1167025

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

Spangler, G. (1989). Toddlers' Everyday Experiences as Related to Preceding Mental and Emotional Disposition and Their Relationship to Subsequent Mental and Motivational Development: A Short-Term Longitudinal Study. *International Journal of Behavioral Development*, *12*(3), 285–303. https://doi.org/10.1177/016502548901200301

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

Appendix

List of ChildLens Activity Classes

The dataset contains the following list of activities.

1. **playing with object**: The child is playing with an object, such as a toy or a ball.
2. **playing without object**: The child is playing without an object, such as playing hide and seek or catch.
3. **pretend play**: The child is engaged in imaginative play, such as pretending to be a doctor or a firefighter.
4. **watching something**: The child is watching a movie, TV show, or video on either a screen or a device.
5. **reading book**: The child is reading a book or looking at pictures in a book.
6. **child talking**: The child is talking to themselves or to someone else.
7. **other person talking**: Another person is talking to the child.
8. **overheard speech**: Conversations that the child can hear but is not directly involved in.
9. **drawing**: The child is drawing or coloring a picture.
10. **crafting things**: The child is engaged in a craft activity, such as making a bracelet or decoration.
11. **singing / humming**: The child is singing or humming a song or a melody.
12. **making music**: The child is playing a musical instrument or making music in another way.
13. **dancing**: The child is dancing to music or moving to a rhythm.
14. **listening to music / audiobook**: The child is listening to music or an audiobook.

List of ChildLens Location Classes

1. livingroom

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

2. playroom
3. bathroom
4. hallway
5. other

Activity Class Statistics