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, containing 106.10 hours of material. The dataset comprises five location classes and 14 activity classes, spanning audio-only, video-only, and multimodal activities. Children wore a vest with an embedded camera that recorded their everyday experiences. 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 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.

ChildLens: An Egocentric Video Dataset for Activity Analysis in Children

Introduction

- Contextualization: Explain the importance of naturalistic data in understanding development and its role in validating theories.
- Challenges in Data Availability: Highlight any gaps or limitations in existing datasets relevant to your research focus (e.g., few longitudinal datasets or those captured from real-world environments).
- Contribution of Your Dataset: Describe the new dataset you are introducing (its scope, duration, type of data captured, and any unique features it offers compared to existing datasets).
- Value of Data Sharing: Emphasize how sharing datasets can enable broader research advances, whether for understanding child development or enabling new computational models in areas like computer vision or audio analysis.
- TODO: research about egocentric video datasets
- Saycam paper als Orientierung
- Relevanz für die psychologische Forschung
- Linda Smith - Indiana (egocentric perspective)
- Roy - Child Talk (Prediction the Birth of a Spoken word)
- Alex Christia
- Bisherige Datensätze sind bisher nur Sprache - Sprache und Video selten zusammen und wenn nur sehr klein (mit anderer Altersgruppe)

The current project attempts to fill this gap by describing a new, openly accessible dataset of more than 415 hours of naturalistic, longitudinal recordings from three children. The SAYCam corpus contains longitudinal videos of approximately two hr per week for three children spanning from approximately six months to two and a half years of age. The data include unstructured interactions in a variety of contexts, both indoor and outdoor, as well as a variety of individuals and animals. The data also include structured annotations of context together with full transcripts for a subsample (described below), and are accompanied by monthly parent reports on vocabulary and developmental status. Together, these data present the densest look into the visual experience of individual children currently available.

Dataset Overview

Activity Classes. The ChildLens dataset contains a total of 14 activity and 5 location classes. The activities are based on the activities of the child in the video and can be divided into *person-only* activities, such as "child talking" or "other person talking", and *person-object* activities, such as "drawing" or "playing with object". You can find a brief description of each class in the appendix. The activities can be further divided into *audio-based*, *visual-based*, and *multimodal* activities, as presented in figure 1. The following list provides 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*, *pretend play*, *reading book*, *making music*

For every activity, the dataset includes information on the start and end times of the activity, the activity class, and whether the child is engaged alone or with somebody else.

For every external person involved in the activity, we capture age and gender. The location classes describe the current location of the child in the video and include *livingroom*, *playroom*, *bathroom*, *hallway*, and *other*.

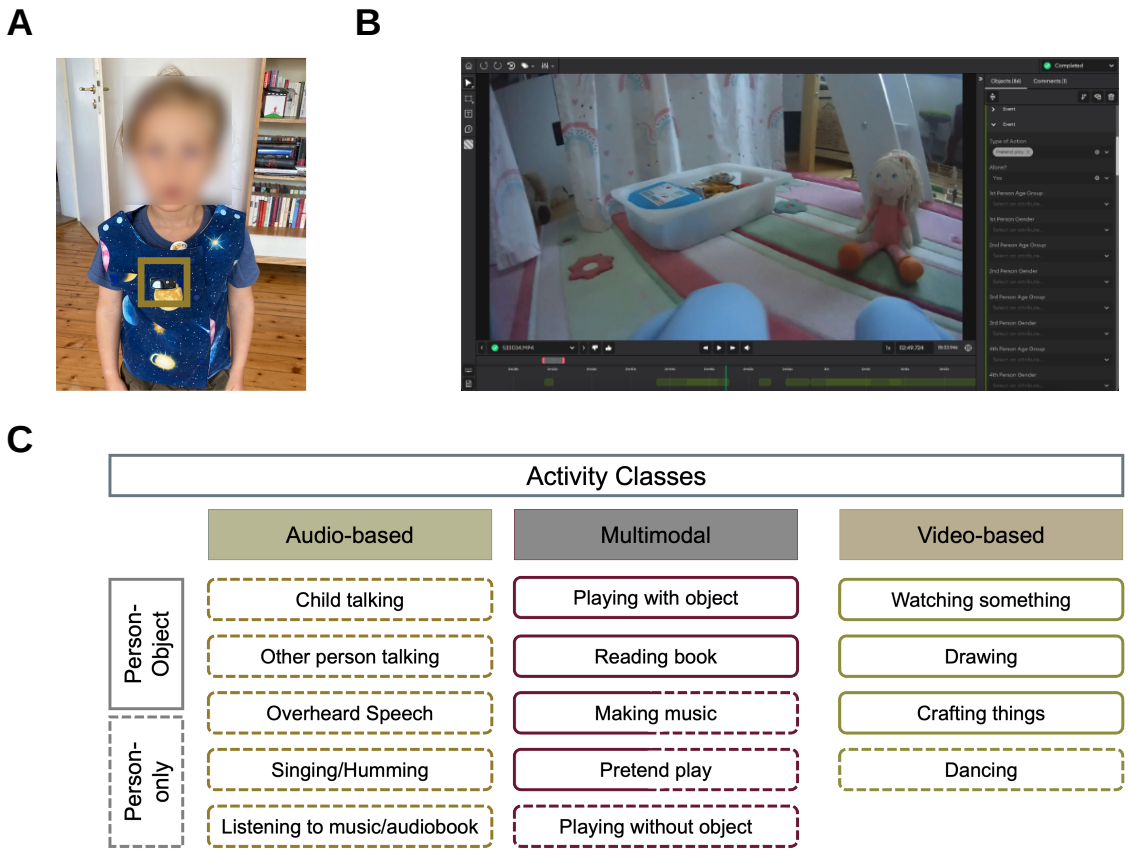

Figure 1. **A** – Vest with the embedded camera worn by the children, **B** – SuperAnnotate platform utilized for video annotation, **C** – Activity classes in the ChildLens dataset.

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). It includes 107 videos from children aged 3, 122 videos from children aged 4, and 114 videos from children aged 5. The duration of recorded video material 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 can be found in figure 2.

This diverse dataset includes a varying number of instances across the 14 activity

classes, ranging from **x** to **x** instances per class. Overlapping instances are counted separately for each activity class; if two activities occur simultaneously in a given video segment, both classes will receive a count for that segment. As a result, parts of the video may be counted multiple times, once for each activity class. The duration of each instance varies by activity. For example, 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.

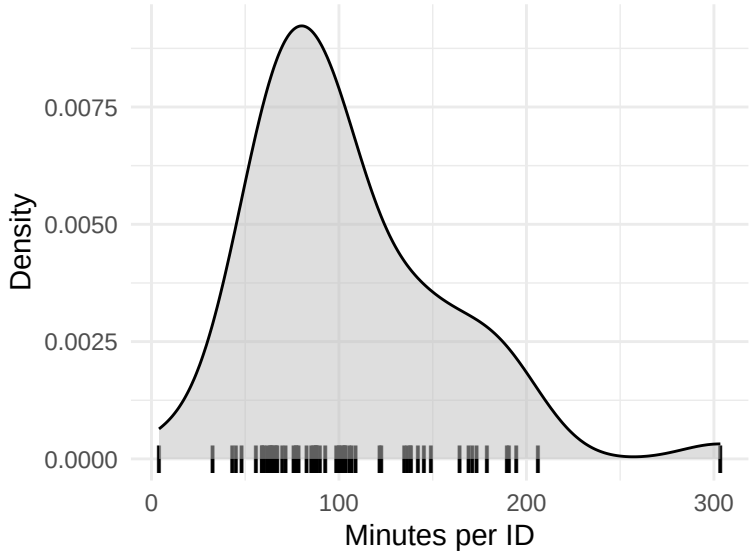

Figure 2. Video recording duration (in minutes) per child ID.

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 and extracted as a separate clip. For example, if a segment shows a child "reading a book" while also "talking," two separate clips 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, which can be seen 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 captured 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 to ensure accuracy and detail. 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, allowing for efficient annotation and review of the videos. Figure 1 provides a screenshot of the SuperAnnotate platform used for video annotation. To ensure the quality of the annotations, the following steps were taken:

1. **Initial round of annotations**: Each set of videos is assigned to specific annotators, who handle the annotations, make changes, and apply corrections as needed. In total, three annotators were actively working on the annotation process.
2. **Quality assurance**: One person is dedicated to quality assurance, ensuring that the annotations are accurate and consistent across all videos.
3. **Review process**: After the initial annotations are completed, the annotations are reviewed by the internal team to ensure that they are accurate and complete. Any discrepancies or errors are corrected before the final submission.

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 video-based on multimodal 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.

Boundary-Matching Network (BMN)

We employ the Boundary-Matching Network (BMN) (Lin, Liu, Li, Ding, & Wen, 2019) for temporal activity localization on untrimmed videos. 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 videos were preprocessed following the MMAction2 guidelines to ensure compatibility with the model architecture. Prior to model training, we analyzed the number of instances per activity class to assess to evaluate the data sufficiency for training and testing purposes. The distribution of activity instances and their total duration across activity classes are presented in Table (**ref?**)(tab:activity-classes-statistics) in the appendix. Our analysis revealed a pronounced class imbalance in the dataset, both in terms of the number of instances and their total duration. Given that the primary aim of this study is to establish initial benchmark results, no data augmentation techniques were employed to address this imbalance. Instead, we focused on the most prevalent activity classes, namely “Playing with Object”, “Drawing”, and “Reading a Book”. The dataset was split into training, validation, and test subsets in an 80-10-10 ratio. The training set was used to optimize the model’s parameters, while the validation set was used for tuning

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	Recall	Recall	Recall	AR@100	AUC
ActivityNet-1.3	0	0	0	72.46	64.47
	Playing with Object	Drawing	Reading a Book		
ChildLens	0	0	0	0	0

hyperparameters and avoiding overfitting. Finally, the test set helped evaluate the model's performance on unseen data, providing a good measure of how well it generalizes.

Implementation Details. We trained the BMN model from scratch on the ChildLens dataset to predict the start and end boundaries of the video-based activity classes. The model was implemented using MMAction2, "an open-source toolbox for video understanding based on PyTorch" (Contributors, 2020). Training was conducted on a Linux server with 48 cores and 256 GB RAM. The model was optimized using the Adam optimizer with a learning rate of 0.001 and a batch size of 16. The training process involved multiple epochs, with early stopping based on validation loss to prevent overfitting.

Evaluation. The performance of the BMN on the ChildLens dataset compared to its original evaluation dataset is summarized in Table 1. A detailed breakdown of recall performance for each activity class is provided in the appendix. Overall, BMN demonstrates satisfactory performance on the ChildLens dataset, effectively generalizing to this new domain.

Voice Type Classifier (VTC)

The Voice Type Classifier (Lavechin et al., 2020) (VTC) is a state-of-the-art model designed to classify audio rawfiles into five distinct voice types: `Key Child (KCHI)`, `Other Child (CHI)`, `Male Speech (MAL)`, `Female Speech (FEM)`, and `Speech (SPEECH)`. Its architecture processes audio by first dividing it into 2-second chunks, which are passed through a SincNet to extract low-level features. These features are then fed into a stack of two bi-directional LSTMs, followed by three feed-forward layers. The output layer uses a sigmoid activation function to produce a score between 0 and 1 for each class. The VTC is trained on 260 hours of audio material obtained from different child-centered audio datasets. Model valuation is performed 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.

The F_1 is a metric that combines precision and recall into a single value, calculated as their harmonic mean. It ranges from 0 to 1, with 1 representing perfect precision and recall, and 0 indicating no correct prediction The interpretation of the F_1 score depends on the specific application of the model. Generally, an F_1 score above 0.8 is considered good, while values above 0.9 are considered excellent. In some cases, a score around 0.5 can still be deemed acceptable, depending on the balance between precision and recall. The F_1 score is computed for each class and averaged to provide an overall measure. No collar is applied to the evaluation, meaning that the prediction have to be exact to be considered

Table 2

Total Duration (in minutes) of all Instances for each VTC Class

	KCHI	CHI	MAL	FEM	SPEECH
Total Duration (min)	100	100	100	100	100

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 applying the VTC to the ChildLens dataset, we mapped our audio-based activity classes to the VTC output classes to enable performance comparison. 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`**
- 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.

Evaluation. Table 3 presents the performance of the Voice Type Classifier (VTC) on the ChildLens dataset compared to the benchmark dataset from the original study. The VTC model achieves an average F_1 score of **xx** on the ChildLens dataset, performing comparably to the benchmark dataset. It performs best on the `CHI` class with an F_1 score

Table 3

Comparison of VTC performance on the ACLEW-Random dataset (used for model evaluation) and the ChildLens dataset, highlighting the F1 measure for each class and the average F1 score

Dataset	KCHI	CHI	MAL	FEM	SPEECH	AVG
ACLEW-Random	68.7	33.2	42.9	63.4	78.4	57.3
ChildLens	59.1	79.2	17.8	33.4	68.3	51.5

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.

Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed ac purus sit amet nisl tincidunt tincidunt. Nullam nec turpis at libero tincidunt tincidunt. Sed nec mi nec nunc tincidunt tincidunt. Nullam nec turpis at libero tincidunt tincidunt. Sed nec mi nec nunc

Dataset Bias

Overall, the dataset demographics are balanced. From the 61 children who participated in the study, 32 children are female and 29 are male.

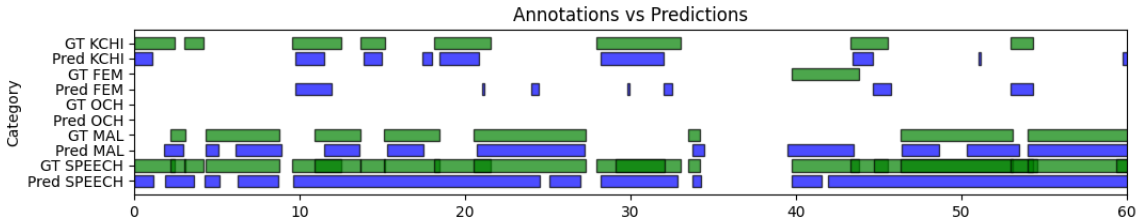

Figure 3. VTC Predictions compared to Ground Truth Annotations

- is there gender bias in the dataset itself (how many female how many male)
- is ther gender bias in some categories (e.g. more female for drawing etc.)

General Discussion

Conclusion

In this paper, we introduced the ChildLens dataset, a novel egocentric video dataset designed for activity analysis in children. The dataset contains a wide range of children's daily live activities, captured in naturalistic environments. We outlined the data collection process and the generation of activity labels, providing detailed information on the labeling strategy employed. Initial results of applying two state-of-the-art models to the dataset were presented, establishing a benchmark for future research. While our current analysis treats audio and video independently, future studies could leverage multimodal approaches to gain deeper insights into children's behavior and activity patterns, advancing the understanding of developmental and interactional contexts.

References

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

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

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	100	100
	Other person talking	100	100
	Overheard Speech	100	100
	Singing/Humming	100	100
	Listening to music/audiobook	100	100
Video	Watching something	2	5.09
	Drawing	62	374.91
	Crafting things	26	109.14
	Dancing	2	0.57
Multimodal	Playing with object	318	1371.08
	Playing without object	25	28.87
	Pretend play	59	158.84
	Reading a book	83	334.19
	Making music	3	2.13

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

Activity Class Statistics