THE BABYVIEW DATASET: HIGH-RESOLUTION EGOCENTRIC VIDEOS OF INFANTS' AND YOUNG CHILDREN'S EVERYDAY EXPERIENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Human children far exceed modern machine learning algorithms in their sample efficiency, achieving high performance in key domains with much less data than current models. This "data gap" is a key challenge both for building intelligent artificial systems and for understanding human development. Egocentric video capturing children's experience – their "training data" – is a key ingredient for comparison of humans and models and for the development of algorithmic innovations to bridge this gap. Yet there are few such datasets available, and extant data are low-resolution, have limited metadata, and importantly, represent only a small set of children's experiences. Here, we provide the first release of a large developmental egocentric video dataset – the BabyView dataset – recorded using a high-resolution camera with a large vertical field-of-view and gyroscope/accelerometer data. This 430 hour dataset includes egocentric videos from children spanning 6 months – 5 years of age in longitudinal, at-home contexts. We provide gold-standard annotations for the evaluation of speech transcription, speaker diarization, and human pose estimation, and evaluate models in each of these domains. We train self-supervised language and vision models and evaluate their transfer to out-of-distribution tasks including syntactic structure learning, object recognition, depth estimation, and image segmentation. Although performance in each scales with dataset size, overall performance is relatively lower than when models are trained on curated datasets, especially in the visual domain. Our dataset stands as an open challenge for robust, human-like AI systems: how can such systems achieve human-levels of success on the same scale and distribution of training data as humans?

1 INTRODUCTION

Infants and young children are remarkable learners, becoming capable and engaged social partners within their first two years of life. The pace of this developmental progress far exceeds modern machine learning algorithms in its efficiency and capacity (Frank, 2023). In particular, signature accomplishments of artificial systems such as few-shot learning (Brown et al., 2020) and image classification (Krizhevsky et al., 2012) require hundreds of billions of words of training data and millions of labeled images. In contrast, human learners become proficient in extending labels for newly learned visual concepts (Carey & Bartlett, 1978) and producing language (Frank et al., 2021) from only tens of millions of words and far fewer labeled examples (Zhuang et al., 2021). This "data gap" between human and machine learners is thus a key challenge for the joint goals of understanding human learning and building intelligent artificial systems. Making progress will require not just an understanding of the flexibility of human intelligence, but also an understanding of the efficiency of human learning.

Data availability is a major barrier to progress in our understanding of the gap in learning efficiency between machines and humans. To make effective comparisons between human and machine learners, we need to be able to evaluate models on data comparable to what children see and hear during everyday learning experiences. While models are trained on millions of images and/or videos, these are taken from the adult perspective, providing a very different vantage point on the world that is disconnected from real-world learning environments.

Egocentric video recordings taken from the child's perspective provide a key window into what children both see and hear as they learn about the world around them and from their social partners (Smith et al., 2015; Yoshida & Smith, 2008; Aslin, 2009; Franchak et al., 2011). Developmental psychology studies using these types of video recordings have together revealed that the infant view is dramatically different from that of an adult (Yoshida & Smith, 2008) and varies as children learn to locomote on their own and interact actively with the objects, places, and people around them (Kretch et al., 2014; Long et al., 2022).

Here we present the largest high-resolution developmental egocentric video dataset to date, the BabyView dataset. We collect videos from 28 families predominantly from around the U.S., totalling 430 hours of usable recordings. We capitalize on innovations in the development of head-mounted cameras (Long et al., 2023), obtaining videos with a large vertical field of view and coordinated gyroscope/accelerometer data that can be used to estimate the child's own head movements. We provide pose detection, automated speech transcriptions, and diarization, along with gold-standard annotations for use in evaluating each of these. We then evaluate self-supervised vision and language models on these data relative to existing benchmarks.

2 Related Work

Few developmental egocentric video datasets are available Egocentric video has been an important domain for computer vision (Damen et al., 2022; Grauman et al., 2022) and resulting commercial applications, such as wearable devices. Yet egocentric video datasets are mostly taken from the adult perspective, including the Ego4D dataset, which has become an important standard in this field (Grauman et al., 2022). Head-mounted cameras have also been used in research with children, including both descriptive investigations (Yoshida & Smith, 2008; Aslin, 2009; Franchak et al., 2011; Kretch et al., 2014; Fausey et al., 2016; Bergelson & Aslin, 2017) and computer vision studies (Sheybani et al., 2024; Zhuang et al., 2021). Unfortunately, most prior work did not obtain consent for broad sharing with other research groups and so many major datasets are unavailable for re-analysis.

Those developmental egocentric video datasets that are available have been difficult to use for training models for reasons of both data quantity and quality (Long et al., 2022; Sullivan et al., 2021; Bergelson & Aslin, 2017). For example, the SAYCam dataset – by far the largest available dataset – is relatively low-resolution (480 x 640 pixels), has limited motion-correction (leading to blurry views) and has timestamps imprinted on every frame (Sullivan et al., 2021). The audio quality is quite variable depending on the background noise and context, and the videos have restricted vertical view angle that obscures views of children's hands and what children are interacting with. Further, SAYCam represents video from three children of highly-involved and informed academic parents, all of whom were the first children in their families. These issues have limited the field's ability to make use of automated annotations of the visual or linguistic content of these videos and have restricted the ability to use these data to draw broadly generalizable conclusions. Here we present the largest high-resolution, developmental egocentric video dataset with broad consent from caregivers for reuse within the research community.

Models trained on developmental data show limited performance Self-supervised vision models trained using developmental egocentric video data (Zhuang et al., 2021; Orhan et al., 2020; Zhuang et al., 2022; Orhan & Lake, 2024; Vong et al., 2024) have had some intermediate success. However, these representations trained from egocentric videos significantly underperform those self-supervised models trained on curated datasets, while the latter models approach the accuracy of models trained using fully-supervised methods (Oquab et al., 2023; Caron et al., 2021; He et al., 2021; Chen et al., 2020; He et al., 2020). Thus, it remains unclear whether the current state-of-the-art techniques represent truly general purpose visual learning algorithms. In particular, it is unclear whether gaps in model performance are due to dataset quality and quantity or instead due to the difficulty of learning robust representations from children's more realistic everyday inputs.

Relatedly, in the language domain, recent work has investigated the possibility of training language models (LMs) on small-scale developmental datasets (see e.g., Warstadt et al., 2023; Zhuang et al., 2024; Feng et al., 2024), but most of these have focused on datasets larger than those available from egocentric video data. For example, the text data used in the popular BabyLM competition (Warstadt et al., 2023) are also meant to approximate what a 10-year-old child could receive (including text

Table 1: The BabyView dataset is the only egocentric developmental video dataset with accelerometer/gyroscope data that is available for research.

Dataset	Ego?	Long?	Type	N	Hours	Audio	Transcript	Motion
BV-Home	✓	✓	Infant	28	433	✓	✓	✓
Ego-SingleChild	✓	✓	Infant	1	47	✓	✓	
SAYCam Sullivan et al. (2021)	✓	✓	Infant	3	476	✓	✓	
Ego4D Grauman et al. (2022)	✓		Adult	931	3,670	✓	✓	
Epic Kitchens Damen et al. (2018)	✓		Adult	37	100	✓	✓	

from Wikipedia and other sources), which is very likely more – and different – data than what is required to acquire a language. One exception is Qin et al. (2024), who trained GPT-2 (Radford et al., 2019) on very small amounts of input from a single child and investigated the amount of grammatical knowledge that could be learned.

Here, we evaluate whether data from a new, high-resolution dataset will lead to increases in performance for self-supervised visual and linguistic benchmark models.

3 The BabyView Dataset

We address gaps in data availability by collecting and analyzing a new set of developmental egocentric videos: the BabyView dataset. The current paper describes the first release of the dataset, but data collection is still ongoing and we anticipate future growth in the overall size of the dataset. Recordings were obtained using a high-resolution head-mounted camera for infants and children from 6 months through 5 years of age in both at-home and preschool settings. In the BabyView-Home portion of the dataset, 28 families recorded longitudinal data during everyday activities for a total of 433 hours across all children. All videos are accompanied by accelerometer/gyroscope data that can be used to estimate children's head-motion (Joshi et al., 2010; Karpenko et al., 2011; Joshi et al., 2022). We additionally release the Ego-SingleChild dataset, a related dataset with a different camera (see below). Together, these data comprise the first release of the largest high-resolution egocentric video dataset from the child perspective that will be available to researchers for both descriptive analysis and model building (see Table 1 for comparison to prior datasets).

3.1 Camera and sensor data

The BabyView camera is a GoPro Hero Bones camera attached to a child-safety helmet. This camera was selected because it has gyroscope and accelerometer data, built-in image stabilization features, and relatively high resolution sound and video (Long et al., 2023). The camera is oriented vertically and is neutral with respect to the face plane of the child, enabling the camera to capture both adult faces and objects within a child's hands in the same image, with an effective view angle of 100° vertical by 75° horizontal (see Figure 1a,b)) (Long et al., 2023).

3.2 Dataset Components

BV-Home Twenty-eight families consented to capture home recordings with their infant-toddler (0;5-3;1 years, average age at onboarding = 11 months, SD = .50 years, see Figure 1c). Families were recruited from a convenience sample of researchers in the field of cognitive development (N=9/28 families) and from local advertisements within the State of California. Some English-speaking and English/Spanish bilingual families (N=16/28) completed parent-report measures of children's language development using the long-forms of the MacArthur-Bates Communicative Development Inventories (Marchman et al., 2023; Jackson-Maldonado et al., 2003). See SI for further information on participant consent, detailed demographics, and language questionnaires.

Ego-SingleChild We also release 47 hours of data from a single child of an academic who recorded frequently. They used a Cigno F18 Night Vision 1080P Headband Sport Camera rather than the BabyView camera, which yields shorter and lower-resolution videos.

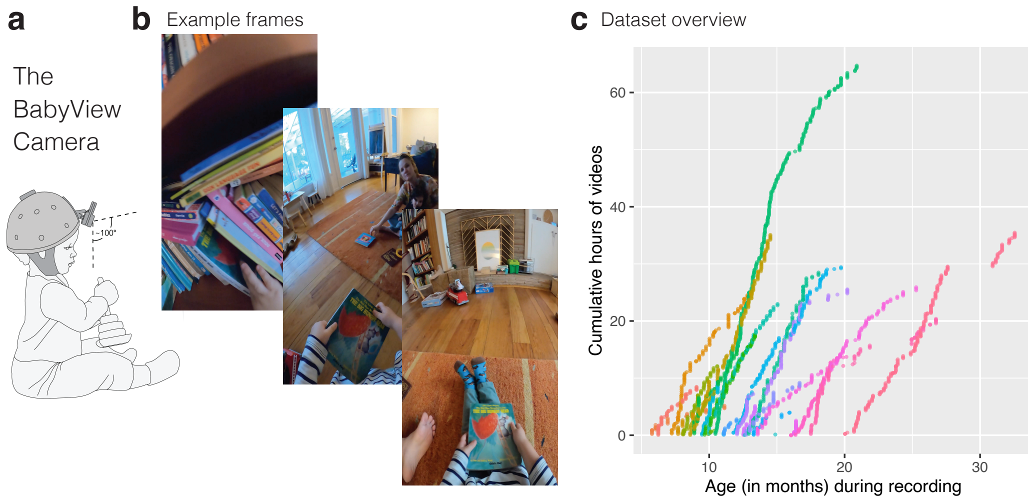

Figure 1: (a) Schematic of a child wearing the BabyView camera illustrating a large vertical field of view. (b) Example frames from a video in the dataset. (c) Cumulative hours of video by each of the participants in the BV-Home subset of the dataset; each color represents an individual child. Data collection is ongoing.

3.3 Data access & ongoing data collection

Egocentric video data from children in their home and school environments necessarily contain more sensitive information than videos in egocentric videos by adults. Families provide full consent for the data that are shared at the time of recording and also have a 6 month period after recording when they can retract any portion of their recording. Thus, all data in this release will be made available in November 2024 once the parental embargo period has lapsed. To ensure BabyView data are accessible to researchers while protecting the privacy of participants, we distribute the data through Databrary (https://nyu.databrary.org/) (Gilmore et al., 2016), similar to previous developmental egocentric datasets (Sullivan et al., 2021; Bergelson & Aslin, 2017). Databrary is an US National Institutes of Health-funded site designed specifically for the distribution of developmental video data. Access to data on Databrary requires investigators be authorized via an institutional agreement that bars reidentification of participants and redistribution of data.

BabyView is an ongoing longitudinal project and our aim is to release further data as the dataset grows. Because of the multi-faceted and growing nature of our dataset, we do not pre-specify train/test splits, recognizing that any split might be appropriate for only a subset of research goals (e.g., examining age-related change, or within- vs. cross-child change).

4 Annotations

4.1 Language annotations

Transcription & diarization pipeline All videos were transcribed using Distil-Whisper-large-v3.[1] As this version only supports English transcription, we discarded utterances for transcription validation that were in languages other than English (BV-Home, N=643 utterances, 24.82%). We also ran a multilingual voice type classifier (Lavechin et al., 2020) on the audio extracted from all BabyView-Home videos, which classified the speech segments as originating from a female adult, male adult, key child (the wearer of the camera), or other child. Each utterance was assigned to one speaker by choosing the model-annotated speaker category that had the greatest overlap with the utterance timestamps. In some cases, an utterance did not overlap with any model-annotated speaker; these were marked as NA (NA rate was 7.18% for BV-Home). For our language model training experiments below, we also ran the same pipeline on the SAYCam audio, though we did not conduct validation on this dataset.

[1]Available at https://huggingface.co/distil-whisper.

Table 2: Language annotation results across the age of the child and the speaker. Child-produced speech and infant-directed speech had the highest error rates.

Dataset	Child age	Speaker	WER	Diarization precision	Diarization recall	N
BV-Home	All Ages	All Speakers	0.38	0.61	0.61	1947
	6-18 m.o.	Adult	0.30	0.79	0.66	1103
		Key-child	1.11	0.48	0.72	190
		Other-child	0.51	0.39	0.64	88
	18-30 m.o.	Adult	0.37	0.77	0.64	271
		Key-child	0.56	0.62	0.76	94
		Other-child	0.21	0.38	0.60	15

Evaluation procedure We hand-annotated a subset of 1947 utterances, stratified across age and participant. Two authors transcribed the speech and labeled the speaker in each segment (N=1.61 hours). For transcription validation, we computed a Word Error Rate (WER), which is is the ratio of the number of word-level errors to the total number of words in the original utterance Gandhi et al. (2023). To evaluate speaker diarization accuracy, we computed precision and recall of the model output by age and speaker.

Child-produced and child-directed speech is challenging for transcription algorithms WER for automated transcriptions was comparable to typical adult performance in the preschool classroom recordings (see Sparks et al. (2024)), but somewhat lower in the naturalistic home environments. Qualitatively, these decrements in performance appear to result from a high prevalence of infant-directed speech that annotation algorithms are less familiar with. Although automated transcriptions perform poorly for the youngest children, we see considerable improvement in WER of child-produced speech of toddler children. The speaker diarization algorithm (Lavechin et al., 2020) was able to identify whether a child vs. adult was speaking 77% of the time, and often could accurately identify the speaker type in the accompanying audio (see Table 2). While combining speaker diarization and automated transcriptions can be very useful, modern transcription algorithms are still considerably less accurate than humans at understanding both child-directed and child-produced speech.

4.2 Human pose annotations

Pose annotations We evaluated how well state-of-the-art pose detectors perform on the BabyView dataset. To do so, we first sampled 353 frames from the dataset (stratified across participants and sessions) and manually annotated the 333 non-blurry frames using LabelStudio (Tkachenko et al., 2020-2022), creating a validation set. To efficiently annotate the frames, we deployed the RTMPose (Jiang et al., 2023) model via MMPose (Contributors, 2020a) as a backend to provide initial pose keypoints and bounding box predictions, which we then manually corrected. The pose annotations followed the format used in the COCO keypoints dataset (Lin et al., 2014; Sun et al., 2019). To evaluate the accuracy of keypoint detections and compare our results with those of other studies, we adopted the Object Keypoint Similarity (OKS) metric, as used by (Sun et al., 2019) (details in SI).

Child egocentric viewpoints are challenging for most pose detection models The BabyView validation set was more challenging for most models than the COCO validation set (Lin et al., 2014), highlighting a new pose benchmark for naturalistic egocentric videos (see Table 3). However, ViTPose-H, the largest model in the group, showed comparable performance between the two validation sets, suggesting that it is more robust to viewpoint variation.

Table 3: Pose Detection performance on COCO2017 Val and BabyView Val. BabyView Validation frames were more challenging the COCO for all models except ViTPose-H.

Architecture	#Params	Input Size	COCO AP	BV AP	COCO AR	BV AR
RTMO-l (Lu et al., 2023)	44.8M	640x640	0.724	0.593	0.762	0.723
YOLOXPose-l (Maji et al., 2022)	87.0M	640x640	0.712	0.588	0.749	0.658
SIMCC-resnet50 (Li et al., 2022)	25.7M	384x288	0.735	0.676	0.790	0.723
RTMPose-l-aic-coco (Jiang et al., 2023)	36.7M	384x288	0.773	0.735	0.819	0.773
HRFormer-pose-base (YUAN et al., 2021)	43.2M	384x288	0.774	0.743	0.823	0.785
ViTPose-H (Xu et al., 2022)	632M	256x192	0.788	0.788	0.840	0.825

5 Benchmarks

5.1 Language representation learning

Next, inspired by the BabyLM challenge, which seeks to learn human-like linguistic representations from small amounts of developmentally-realistic data (Warstadt et al., 2023), we examined the ability to learn linguistic representations from the BV-Home transcripts. For contrast, we compare with high-quality data from the Child Language Data Exchange System (CHILDES), a repository of human-transcribed corpora of children and caregivers' talk (MacWhinney, 2014).

Experiment Setup We pretrained GPT-2 (Radford et al., 2019) with 124M parameters (small) on each dataset for up to 20 epochs (see SI for details). After deduplication, the automatically-transcribed utterances for BV-Home and SAYCam each consisted of $\sim$2M total words. For contrast, the total amount of human-transcribed English-language data available in CHILDES is $\sim$20M words. Hence, we sampled 2M words of conversation from CHILDES (2.4M total words including speaker labels and other metadata) to align the amount of training data across datasets. We then separated each dataset into train and validation splits, using an 85/15 split. We further compared with training on the combination of BV-Home and SAYCam data and $\sim$4M words of conversation (4.8M total words) from CHILDES. We also trained a version on the entirety of the English subset of CHILDES ($\sim$20M words), in line with Feng et al. (2024). For evaluation, we used Zorro (Huebner et al., 2021), a benchmark compatible with child vocabulary that aims to quantify the grammatical knowledge of LMs by assessing their capability to effectively distinguish between minimal pairs of sentences that exhibit various grammatical contrasts.

BV-Home transcriptions provide comparable learning signal for grammatical knowledge All GPT-2 models achieved above-chance performance on the Zorro evaluation, even with only $\sim$2M words of training data (see SI for complete results). With 2M words, there was only a negligible difference between BV-Home (64.13%) and SAYCam data (64.06%) and a minor advantage for CHILDES (66.57%). However, combining BV-Home and SAYCam led to matched performance (69.39%) to CHILDES 4M (69.76%). Training on the full CHILDES English subset of 20M words resulted in significantly higher performance (77.77%), as expected with much more language data. This is also shown in Figure 2; training on more language data results in better performance, in contrast to our vision data scaling experiments shown in Figure 3. Overall, despite the potential data quality issues in BabyView and SAYCam transcripts (introduced by multilingual data and speech recognition errors), we observe that transcriptions of BV-Home and SAYCam are comparable to CHILDES as a learning signal for language models to obtain grammatical knowledge.

5.2 Visual representation learning

We conducted a first set of experiments to investigate the ability of recent self-supervised models to learn useful visual representations from frames taken from these egocentric videos. Enabled by BV-Home, we conduct the largest scale evaluation to date of self-supervised learning methods trained on children's egocentric visual experience.

Experiment Setup We trained a ViT-B/14 DINOv2 (Oquab et al., 2023) from scratch as our reference self-supervised learning algorithm, due to its high performance on a variety of downstream tasks, including object recognition, depth estimation and semantic segmentation. We used the standard

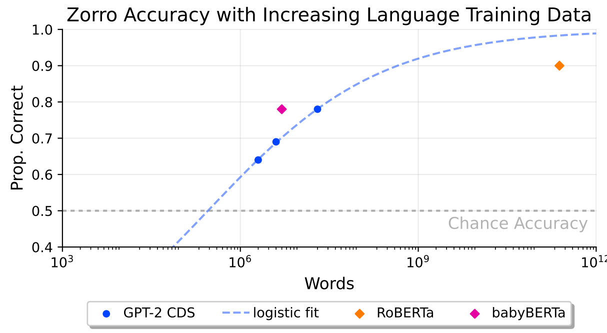

Figure 2: Language data scaling experiments, showing grammatical accuracy on the Zorro benchmark (chance = 0.5) for GPT-2 trained on progressively increasing amounts of child-directed speech (CDS) language data. Within the GPT-2 CDS data points, the first represents 2M words from the BV-Home corpus, the second represents 4M words combined over the BV-Home and SAYCam corpora, and the final point represents 20M words from the CHILDES corpus. Zorro accuracy is also shown for RoBERTa (Liu et al., 2019) [240M words] and BabyBERTa (Huebner et al., 2021) [5M words].

Table 4: Object recognition, depth estimation, and semantic segmentation results on the BabyView & comparison datasets. Downstream generalization accuracy is significantly reduced when learning on frames from egocentric videos relative to curated datasets.

	Object Recognition – Top 1		Depth Estimation	Semantic Segmentation
Dataset	ImageNet kNN	ImageNet linear	NYUv2 RMSE↓	COCOStuff mIoU↑
None (random init.)	10.00	1.43	0.886	0.54
LVD-124M (Oquab et al., 2023)	82.10	84.50	0.307	44.46
ImageNet (Russakovsky et al., 2015)	76.29	77.64	0.456	34.65
Ego4D(Grauman et al., 2022)	43.59	54.39	0.525	23.78
SAYCam(Sullivan et al., 2021)	42.59	52.52	0.518	21.08
BV-Home	40.72	52.19	0.526	22.03
SAYCam + BV-Home	41.76	53.28	0.511	22.53

training configuration from the official code base across all training runs. We sampled Ego4D at 1 FPS, leading to 15M frames, and sampled the BV-Home and SAYCam at 5FPS, leading to about 8M frames per dataset. Despite the inherent redundancy in video data, this ensured a relatively large amount of data, compared with the 1.4M ImageNet training set. We evaluated object recognition accuracy on ImageNet, and after additional training on high-resolution images of the original datasets, we evaluate depth estimation on NYUv2 (Silberman et al., 2012) and semantic segmentation on COCOStuff (Caesar et al., 2018). On top of the frozen ViT, for ImageNet we use kNN and a linear probe, whereas for depth estimation we trained a DPT and for semantic segmention we used a linear probe, following the DINOv2 protocols.

Self-supervised learning from any egocentric data is challenging We anticipated that the more diverse and higher-resolution videos in BV-Home would afford improvements over prior egocentric video datasets (Sullivan et al., 2021). Yet we found that models trained on BV-Home data did not outperform those trained on the SAYCam dataset, despite the difference in data quality (see Table 4), though we found a small improvement in semantic segmentation performance on models trained on BV-Home vs. SAYCam.[2] More broadly, however, we found that the gap in performance is not just specific to data collected from children. Even when training on Ego4D – a roughly 7x larger and more diverse dataset – we see that a significant gap to curated vision datasets remains across all tasks. We further investigated training an additional self-supervised learning method, MoCov3 (Chen et al., 2021) also based on a ViT-B/16 on the full dataset. We obtained 18.7 for kNN and 27.3 for linear on ImageNet, indicating that other self-supervised learning techniques also show a significant gap in performance.

[2]Note results are above random chance: ImageNet – 0.001, NYUv2 – 2, COCOstuff – 0.2.

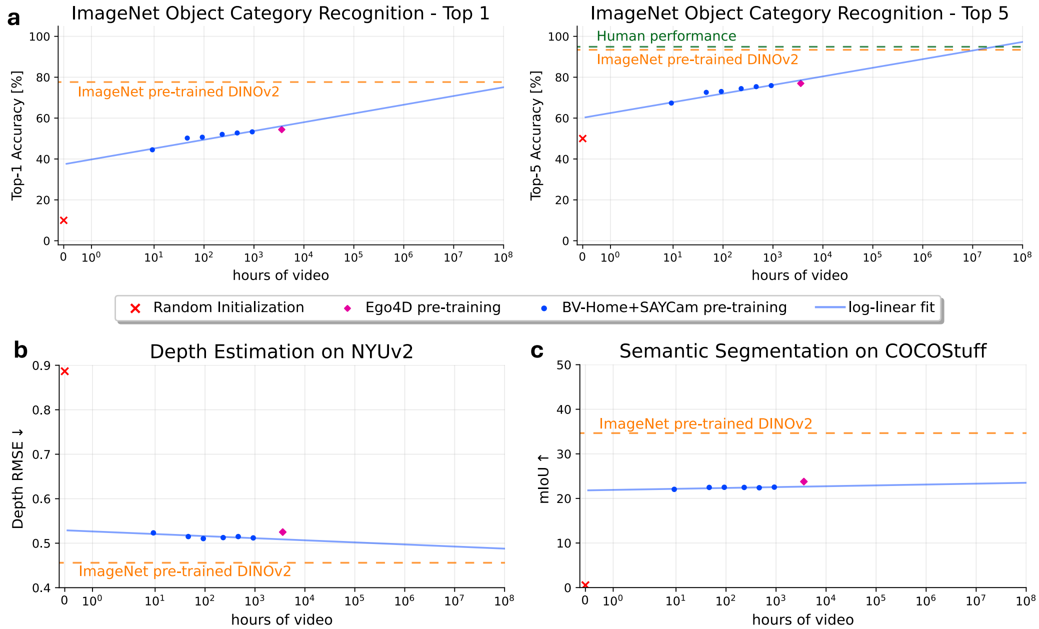

Figure 3: Data scaling experiments for object recognition, depth estimation and semantic segmentation. In **a** we observe a trend that DINOv2 would require upwards of 10^7 hours of video to match human or ImageNet self-supervised ImageNet performance. In **b** and **c** we also observe unfavorable scaling for depth estimation and semantic segmentation.

Insufficient scaling to meet human or self-supervised performance from curated datasets Given a reasonably large amount of training data from egocentric video of children's visual experience, could the current self-supervised state-of-the-art obtain equivalent performance to training on curated vision datasets or human performance? We trained on 1%, 5%, 10% 25%, 50% and 100% of a combined dataset of BV-Home and SAYCam, and extrapolate by fitting log-linear trend lines. For object recognition on ImageNet (see Figure 3a) we observed that more than 10^7 hours would be required to reach human performance (Russakovsky et al., 2015) or ImageNet pre-training performance. In Figures 3b and 3c, we find that a similar trend holds for depth estimation and semantic segmentation, with saturating performance as the scale of data is increased. Note that the first two points on these plots indicate 160K and 800K images, and the last point 16M images. While a similar "data gap" finding has also been reported by Orhan (2021), our new dataset and models yield a somewhat lower estimation of the amount of data needed to achieve human-level performance.

6 General Discussion

We present a new, large-scale high-resolution egocentric video dataset documenting infants' and young children's everyday experiences, accompanied by both dense metadata and gold-standard annotations for several key domains. In contrast to prior work with lower-resolution videos and earlier models (Long et al., 2022), we find that state-of-the-art speech recognition (Gandhi et al., 2023; Radford et al., 2023) and pose detection (Xu et al., 2022; Contributors, 2020a) models perform well on stratified samples of frames and audio recordings from the dataset. Further, language models trained on these data performed comparably to models trained on current gold-standard corpora of hand-transcribed speech. The new BabyView camera thus provides improved data over which supervised algorithms can extract descriptives that will be an important resource for characterizing children's linguistic and social learning environments (Sparks et al., 2024).

Yet our results also suggest that the naturalistic, everyday experiences of children pose a challenging problem for the most advanced of our learning algorithms, especially in the visual domain: current state-of-the-art models fall short relative to existing benchmarks when trained on "human amounts" of visual or linguistic data, requiring unrealistic amounts of additional data to achieve human-level performance (Frank, 2023). In particular, our results suggest that current self-supervised visual learning models are dependent on large, curated datasets with a broad diversity of inputs to construct robust representations.

What might lead to more child-like models of early learning? One idea is that the joint learning of visual and language representations requires more fine-grained and efficient learning algorithms, such as lexicon-level visual grounding (Zhuang et al., 2023; 2024). Further, children's everyday experience contains deep regularities within activity contexts (Clerkin et al., 2017; Clerkin & Smith, 2022; de Barbaro & Fausey, 2022) that are challenging for current models but appear advantageous for human learners. Constructing models that can learn as children do from these skewed input distributions is thus a key challenge for future work. We further speculate that focusing on modeling event-representations in naturalistic video (Zhuang et al., 2020), children's own head-motion via IMU data (Joshi et al., 2022), and attentional guidance from caregivers (Long et al., 2022; Yu et al., 2021) may yield more data-efficient models of early learning.

Our results highlight the need for developmentally appropriate outcome data with which we can be used to evaluate models trained on developmental data. Toddlers cannot classify all ImageNet categories, and a growing literature suggests that object recognition abilities mature throughout middle childhood (Long et al., 2024; Huber et al., 2023). Systematically comparing models' and children's emerging representations may help elucidate the observed gap in model performance.

These data have several limitations. First, these data necessarily incorporate selection bias: parents who opt-in to the study are recording in their homes when they choose to (to avoid privacy issues) and can choose to excise any portion of their data; some naturalistic experiences (e.g., bathtime) are not incorporated into the dataset. Further, with two exceptions, all families are located in the United States, limiting generalizability. Nonetheless, BV-Home incorporates data from a greater diversity of families across race, ethnicity, and family incomes than before (see SI). The potential harms that could arise from this dataset relate to breaches of privacy and trust on the part of the participating families. To guard against these, researchers are required to sign the Databrary data use agreement (Gilmore et al., 2016), which prohibits reidentification or redistribution of videos.

In sum, we present the first release of a new, large-scale, high-resolution developmental egocentric video dataset. Our dataset stands as a challenge to modern AI: how can how can such systems achieve human levels of success on the same scale and distribution of training data as human children?

REFERENCES

Richard N Aslin. How infants view natural scenes gathered from a head-mounted camera. *Optometry and Vision Science*, 86(6):561–565, 2009.

Elika Bergelson and Richard N Aslin. Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, 114(49):12916–12921, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Susan Carey and Elsa Bartlett. Acquiring a single new word. *Linguistics*, 1978.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Elizabeth M Clerkin and Linda B Smith. Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of Sciences*, 119(18): e2123239119, 2022.

Elizabeth M Clerkin, Elizabeth Hart, James M Rehg, Chen Yu, and Linda B Smith. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160055, 2017.

MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020a.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020b.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pp. 1–23, 2022.

Kaya de Barbaro and Caitlin M Fausey. Ten lessons about infants' everyday experiences. *Current Directions in Psychological Science*, 31(1):28–33, 2022.

Caitlin M Fausey, Swapnaa Jayaraman, and Linda B Smith. From faces to hands: Changing visual input in the first two years. *Cognition*, 152:101–107, 2016.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. URL `https://arxiv.org/abs/2408.03617`.

John M Franchak, Kari S Kretch, Kasey C Soska, and Karen E Adolph. Head-mounted eye tracking: A new method to describe infant looking. *Child development*, 82(6):1738–1750, 2011.

Michael C Frank. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 2023.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. *Variability and consistency in early language learning: The Wordbank project*. MIT Press, 2021.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*, 2023.

Rick O Gilmore, Karen E Adolph, and David S Millman. Curating identifiable data for sharing: The databrary project. In *2016 New York Scientific Data Summit (NYSDS)*, pp. 1–6. IEEE, 2016.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Lukas S Huber, Robert Geirhos, and Felix A Wichmann. The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. *Journal of vision*, 23(7):4–4, 2023.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. BabyBERTa: Learning more grammar with small-scale child-directed language. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 624–646, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.49. URL https://aclanthology.org/2021.conll-1.49.

Donna Jackson-Maldonado, Donna J. Thal, Larry Fenson, Virginia A Marchman, Tyler Newton, and Conboy Barbara. *MacArthur-Bates Inventarios del Desarollo de Habilitades Communicativas: User's Guide and Technical Manual*. Brookes Publishing Company, 2003.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

Bharat Joshi, Marios Xanthidis, Sharmin Rahman, and Ioannis Rekleitis. High definition, inexpensive, underwater mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1113–1121, 2022. doi: 10.1109/ICRA46639.2022.9811695.

Neel Joshi, Sing Bing Kang, C Lawrence Zitnick, and Richard Szeliski. Image deblurring using inertial measurement sensors. *ACM Transactions on Graphics (TOG)*, 29(4):1–9, 2010.

Alexandre Karpenko, David Jacobs, Jongmin Baek, and Marc Levoy. Digital video stabilization and rolling shutter correction using gyroscopes. *CSTR*, 1(2):13, 2011.

Kari S Kretch, John M Franchak, and Karen E Adolph. Crawling and walking infants see the world differently. *Child development*, 85(4):1503–1518, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Marvin Lavechin, Ruben Bousbib, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*, 2020.

Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pp. 89–106. Springer, 2022.

Zhenyu Li. Monocular depth estimation toolbox. https://github.com/zhyever/Monocular-Depth-Estimation-Toolbox, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Bria Long, Sarah Goodin, George Kachergis, Virginia A Marchman, Samaher F Radwan, Robert Z Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, et al. The babyview camera: Designing a new head-mounted camera to capture children's early social and visual environments. *Behavior Research Methods*, pp. 1–12, 2023.

Bria Long, Judith E Fan, Holly Huey, Zixian Chai, and Michael C Frank. Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, 15(1):1191, 2024.

Bria L Long, Alessandro Sanchez, Allison M Kraus, Ketan Agrawal, and Michael C Frank. Automated detections reveal the social information in the changing infant view. *Child Development*, 93(1): 101–116, 2022.

Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. *arXiv preprint arXiv:2312.07526*, 2023.

Brian MacWhinney. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press, 2014.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2637–2646, 2022.

Virginia A. Marchman, Philip S. Dale, and Larry Fenson. *The MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual, 3rd Edition*. Brookes Publishing Company, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

A Emin Orhan. How much human-like visual experience do current self-supervised learning algorithms need in order to achieve human-level object recognition? *arXiv preprint arXiv:2109.11523*, 2021.

A Emin Orhan and Brenden M Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, 6(3):271–283, 2024.

Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33, 2020.

Yulu Qin, Wentao Wang, and Brenden M. Lake. A systematic investigation of learnability from single child linguistic input. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Linda B Smith, Chen Yu, Hanako Yoshida, and Caitlin M Fausey. Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, 16(3):407–419, 2015.

Robert Z Sparks, Bria Long, Grace E Keene, Malia J Perez, Alvin WM Tan, Virginia A Marchman, and Michael C Frank. Characterizing contextual variation in children's preschool language environment using naturalistic egocentric videos. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, 2024.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5:20–29, 2021.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. URL `https://github.com/heartexlabs/label-studio`. Open source software available from https://github.com/heartexlabs/label-studio.

Suramya Tomar. Converting video formats with ffmpeg. *Linux journal*, 2006(146):10, 2006.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL `https://aclanthology.org/2020.tacl-1.25`.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

Hanako Yoshida and Linda B Smith. What's in view for toddlers? using a head camera to study visual experience. *Infancy*, 13(3):229–248, 2008.

Chen Yu, Yayun Zhang, Lauren K Slone, and Linda B Smith. The infant's view redefines the problem of referential uncertainty in early word learning. *Proceedings of the National Academy of Sciences*, 118(52):e2107019118, 2021.

YUHUI YUAN, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7281–7293. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/3bbfdde8842a5c44a0323518eec97cbe-Paper.pdf`.

Chengxu Zhuang, Tianwei She, Alex Andonian, Max Sobol Mark, and Daniel Yamins. Unsupervised learning from video with deep neural embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 9563–9572, 2020.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

Chengxu Zhuang, Ziyu Xiang, Yoon Bai, Xiaoxuan Jia, Nicholas Turk-Browne, Kenneth Norman, James J DiCarlo, and Dan Yamins. How well do unsupervised learning algorithms model human real-time and life-long learning? *Advances in Neural Information Processing Systems*, 35:22628–22642, 2022.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. Visual grounding helps learn word meanings in low-data regimes. *arXiv preprint arXiv:2310.13257*, 2023.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. Lexicon-level contrastive visual-grounding improves language modeling. *arXiv preprint arXiv:2403.14551*, 2024.

Author Contributions

BLINDED.

Acknowledgements

We gratefully acknowledge the participating families without whom this work would not be possible. This work was funded by [BLINDED]. We thank many research assistants who have played a key role in the construction of this dataset, including [BLINDED].

A Appendix

A.1 Dataset details

A.1.1 Participant consent

All data collection was approved under [BLINDED] and consent was obtained via one-on-one conversations. Given the sensitive nature of the data, families had multiple opportunities to withdraw their recordings. They could mark videos for deletion during recording and up to six months during the embargo period.

A.1.2 Participant instructions & recording details

All participant instructions were taken from Long et al. (2023) which developed the protocols for using the BabyView Camera, and are publicly available at https://osf.io/kwvxu/.

Families were instructed to record as often as was feasible for their families, with a requested minimum of 45 minutes per week. We use standard, rechargeable 9V battery to provide power to the BabyView camera, which allows for continuous 45-60 minute recordings on a standard charge. Families were then compensated based on the duration (mins) of video recordings they provided on a weekly basis as well as bonuses for questionnaires, totalling 18,370.00 dollars across all families.

A.1.3 BV-Home Additional Participant demographics

Our sample is highly educated, with 21/28 families having at least one parent with a graduate degree, and with all families having at least one parent with a 4-year college degree. 11/28 children are exposed to more than one language at home, including the following languages: English, Chinese, Farsi, French, Gujarati, Japanese, Korean, Malayalam, Portuguese, Spanish, Tagalog, Thai, Vietnamese. Geographically, 20/28 of families live within California, 4/28 live in the Northeastern United States, 1/28 live in the Southern United States, 1/28 live in the Midwestern United States, 1/28 live in Canada, and 1/28 live in South Korea.

Participating children were 64.29% female, 35.71% male, 0.0% African American/Black, 17.86% Asian American/Pacific Islander, 42.89% Caucasian/White, 10.71% Hispanic/Latinx, 39.29% multiracial, 0.0% other.

We only have income information for 25/28 families, as reporting was optional. The average family income of our sample is 221,143 USD (75,000–1,000,000 USD, SD = 201,710 USD). 13/25 families have more than one child in the household, 1/25 families live in a single-parent household, and 2/25 families have more than 2 caregivers living in the household.

A.1.4 BV-Home Language Outcome Questionnaires

Long-form MacArthur Bates CDI language questionnaires (https://mb-cdi.stanford.edu/) were administered every 3 months starting at enrollment. Families were provided compensation for each questionnaire. These parent-report forms assess children's language comprehension and production; aggregate data by age can be viewed at wordbank.stanford.edu. Forms were administered through Web-CDI (https://webcdi.org/). A total of 28 (2 Spanish, 26 English) questionnaires are included in this first release of the dataset.

A.1.5 VIDEO PROCESSING PIPELINE

Videos were manually uploaded by each family to their personalized Google Drive folders. The uploaded videos were automatically downloaded to a secure server where the metadata (accelerometer and gyroscope) were extracted and the videos were compressed then uploaded to a second Google Drive platform. The compression step used the ffmpeg (Tomar, 2006) program to encode video into the libx265 format with a constant rate factor of 23 to enable high quality MP4 videos.

A.2 ANNOTATION DETAILS

A.2.1 POSE KEYPOINT DETAILS AND EVALUATION

The pose keypoints that were evaluated includes 17 keypoints: nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, and right ankle.

The Object Keypoint Similarity (OKS) metric reported is as follows:

$$OKS = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}.$$

In this formula, d_i represents the Euclidean distance between the detected keypoint and the ground truth, v_i indicates the visibility of the ground truth keypoint, s denotes the object scale, and k_i is a constant specific to each keypoint that adjusts the falloff. We report standard metrics for average precision and recall: AP (the average of AP scores at 10 different OKS thresholds: 0.50, 0.55, ..., 0.90, 0.95), and AR (the average of AR scores at OKS = 0.50, 0.55, ..., 0.90, 0.95).

A.2.2 COMPUTE RESOURCES AND INFRASTRUCTURE FOR ANNOTATIONS

Our annotation work was performed on an internal cluster server with an AMD EPYC 9334 32-Core Processor, 756GB memory, 8 NVIDIA A40 GPUs, and Ubuntu 20.04. We used 8 GPUs for speech recognition and 1 GPU for both assisting with annotation and testing pose detection models on the validation set.

A.3 LANGUAGE BENCHMARK DETAILS

A.3.1 LANGUAGE MODEL TRAINING & EVALUATION DETAILS AND DATA PROCESSING

In training our GPT-2 models, we used a learning rate (LR) of 1e-04, linear LR scheduler with no warmup steps, a batch size of 16 per GPU, seed of 42, and Adam optimizer with $\beta = (0.9, 0.999)$ and $\epsilon = 1e - 08$.

The final chosen GPT-2 model for each dataset is the epoch that performed best (had the lowest loss) on the corresponding validation split. The corresponding tokenizer for each model was also trained from scratch on the corresponding dataset.

The training data was set up so that each line corresponded to a single transcribed conversation, which is broken up into chunks of 1024 consecutive tokens by GPT-2 during training. To ensure the data format is consistent for evaluation purposes, we aligned the most important and frequently occurring speaker labels across datasets (mainly based on the existing CHILDES labels): CHI for the target child, MOT for the mother or female adult, and OCHI for other children. All other speaker labels were kept to their default. Around 60% or more of all utterances within each dataset were from CHI or MOT.

See below for an example of part of a single training conversation. Double asterisks surround speaker labels, double newline tokens separate utterances, and an end-of-text token marks the end of the conversation. This format was consistent across all conversations and datasets.

***CHI**: Hi.* `\n\n` ***CHI**: There you go.* `\n\n` ***OCHI**: Do you have a little ball in your cup.* `\n\n` *(...)* `\n\n` ***CHI**: Are those your stars?* `\n\n` ***MOT**: Can you say star?* `\n\n`

CHI: Star. \n\n **CHI**: Look. \n\n **CHI**: Stars. \n\n **MOT**: Stars. See? Look, look at the yellow star, a golden star. <|endoftext|>

We found cases of duplicate conversations and duplicate utterances within conversations among the transcribed data across the three datasets. We removed these to the best of our ability before training.

The Zorro evaluation was inspired by BLiMP (Warstadt et al., 2020) and is a modification for child-directed language (e.g. lower vocabulary). However, it was designed specifically for masked language models such as RoBERTa. To adapt it to GPT-2, we reformatted the Zorro data to match the BLiMP format and used the BLiMP evaluation in the BabyLM evaluation suite [3] since the main difference between the two is the evaluation data. Further, we use the full Zorro test suite and do not filter examples by vocabulary. Hence, our results are not comparable to Qin et al. (2024) which filters Zorro examples by the vocabulary of their training datasets.

To better match the training data format and assess the effects of speaker labels on evaluation, we came up with three variations of Zorro: 1) the original Zorro evaluation sentences, 2) the sentences with the CHI speaker label prepended, and 3) the sentences with the MOT speaker label prepended. To further match the training data, the speaker labels were surrounded by double asterisks, and sentences included double newline tokens (before and after).

As seen in Table 5, all models perform better when the evaluation data is more closely aligned with the training data format (2nd or 3rd variation of Zorro sentences), especially with the MOT speaker label (3rd variation). This is likely because the utterances spoken by the mother or female adults are typically more grammatical than those of the child.

A.3.2 Detailed Language Model Experiment Results

See Table 5 for the Zorro evaluation results of our GPT-2 models, along with the best Zorro evaluation format for each.

Table 5: Quantitative results on the Zorro benchmark

Model	Zorro (Final Avg.)	Best Evaluation Format
BV-Home	64.13%	CHI
SAYCam	64.06%	MOT
CHILDES (2M)	66.57%	MOT
SAYCam + BV-Home	69.39%	CHI
CHILDES (4M)	69.76%	MOT
CHILDES (20M)	77.77%	MOT

A.3.3 Compute Resources and Infrastructure for Language Model Training

Our language model experiments were run on a cloud provider VM instance consisting of four A100s (80GB VRAM each).

A.4 Vision Benchmark Details

A.4.1 Video Preprocessing

BabyView We sample BV-Home at 5 FPS at a resolution of 720x360 for the initial 224 global crop training of DINO, and at 720x1280 for the 518 high resolution final stage of training. This results in a total of 8M frames.

To create datasets of different sizes (1%, 5%, etc.) we randomly select complete clips and append them to a continuously increasing list which we save at different size increments. This ensures that every smaller set of data is a strict subset of the larger set (e.g., the clips in the 1% set are all contained in 5% set etc.). After getting these lists of clips, we extract frames with the same procedure.

[3] https://github.com/babylm/evaluation-pipeline-2023

Because the dataset is at a 9:16 widescreen aspect ratio, significantly different from the mostly 4:3 ImageNet image aspect ratio for which the DINO random cropping strategy was developed, we take random crop with aspect ratio in the 4:3 to 3:4 range with the biggest possible size, before performing the DINO cropping and augmentation. Empirically this results in a 1% improvement in ImageNet classification accuracy.

SAYCam We sample SAYCam at 5 FPS in the native resolution of 480x640. This results in a total of 8.5M frames.

Ego4D We take the complete Ego4D dataset without additional post-processing and sample frames at 1 FPS using ffmpeg at 1/2 of the original resolution. The smallest side of the images we extract ranges from 360 to 960 pixels—sufficient resolution for training (the variance in resolution exists in the original dataset due to the use of different recording devices). We reduce the original resolution to reduce the footprint of the dataset on disk and to lower the computational cost of data loading. This results in a total of 15M frames. We apply the same 3:4 aspect ratio augmentation that we did for BabyView.

A.4.2 TRAINING

DINOv2 To train DINOv2 we use the official code repository.[4] We try to perform minimal modifications of the existing pipeline. We train a ViT-B/14 with a batch size of 1024 with the default ImageNet-1K training config for the default 125K parameter updates. This initial training is done with a global crop of 224x224. All other hyperparameters are kept the same. We experimented with doubling the amount of parameter updates but did not see improvements. Following the DINOv2 paper, we train for an additional 10K parameter updates with a global crop of size 518x518.

MoCov3 To train MoCov3 we use the official code repository.[5] We train a ViT-B/16 with a batch size of 512 with the default ImageNet-1K training configurations for up to 725K parameter updates. Similar to DINOv2, the training is done with an initial global crop of 224x224.

A.4.3 DOWNSTREAM TASKS

ImageNet Category Recognition We use the code from the official DINOv2 repository for kNN classification or for training a linear classifier. Our evaluation procedure, therefore, directly follows the procedure used in DINOv2.

NYUv2 Depth Estimation Following the descriptions in the DINOv2 paper, we use the Monocular Depth Toolbox (Li, 2022). The code interfacing DINOv2 with this package is not released, but the trained depth estimation models and configs are released. After writing the interface code, we verify that the evaluation is correct by training a DPT-based depth estimator using this codebase on top of an off-of-the shelf official DINOv2 checkpoint which matched the performance from the paper.

COCOStuff Semantic Segmentation We interfaced the official DINOv2 code with the mmsegmentation package (Contributors, 2020b). Similarly, the interface code is not released but the models and configs are available. To verify correctness, we trained a linear probe on top of an off-the-shelf official DINOv2 checkpoint and matched the performance from the paper on PASCAL VOC. We used the same config to train a linear probe on COCOStuff as was released for PASCAL VOC. We did not find improvements by training for longer. Future work may investigate training more complex architectures, which was prohibitive for this work due to the time and compute constraints required.

A.4.4 COMPUTE RESOURCES

The DINOv2 vision models in this paper can be trained on a single 8x NVIDIA A40 GPU node. While no multi-node training is required, one full training run of DINOv2 takes about 3 days on 8x A40 GPUs. This translates to about 550 GPU hours per experiment, making it difficult to perform multiple runs to obtain error bars.

[4] https://github.com/facebookresearch/dinov2

[5] https://github.com/facebookresearch/moco-v3

A.5 Data accessibility

No data is available for review due to the parental embargo policy. All data will be hosted on https://nyu.databrary.org/ in November 2024 after the parental embargo period has lapsed. Researchers must be affiliated with a PI at a research-institution, who must request access to the project.

All compressed videos and their associated meta-data will be named according to a standardized format that encodes the subject id and the date at which the recordings were made. A .csv spreadsheet will provide detailed, anonymized information about each individual participant. Separate language outcome data (in standard CDI format) will be provided and linked to the individual subject IDs.

A.6 Licensing

The code and behavioral data published with the benchmark will be licensed under CC BY-NC 4.0. The video dataset is licensed under the terms laid out in the Databrary Access Agreement, see https://databrary.org/about/agreement/agreement.html.

License for Annotation models: YOLOXPose is licensed under the GPL-3.0 license. MMPose, RTMO, SimCC, ViTPose, mmsegmentation, DINOv2, Monocular Depth Toolbox, and LabelStudio are licensed under the Apache-2.0 license. GPT-2 is licensed under the modified MIT License. RTMPose is licensed under the MIT license. All are permissive for this paper release.

We the authors bear all responsibility in case we have violated any rights by the publication of these data and code in these venues.

A.7 Code availability

Anonymized, relevant model training code can be found at https://tinyurl.com/osf-babyview-codebase.