Joint Attention and 3D Reconstruction: How Computer Vision can Help with Early Childhood Brain Development

Priscilla Zhao
Stanford Graduate School of Education
puzhao@stanford.edu

Yanni Zhao
Stanford Department of Civil Engineering
yanniz@stanford.edu

Raymond Zhang
Stanford Graduate School of Education
zhan1087@stanford.edu

Abstract

The Filming Interactions to Nurture Development (FIND) program uses video coaching to promote positive interactions between caregivers and children. One important interaction in the program is joint attention where both agents are focusing on a single object. In this work, we develop a pipeline that identifies instances of joint attention by using CLIP, a 2D joint attention detector, and finally a monocular depth estimation model to reconstruct the 3D coordinates of the three entities of interest in the scene. Using the 3D coordinates we are able to calculate the area of a 3D triangle that represents the space between the three entities. Our analysis shows that in instances of shared reading the area of the triangle is less than in other playing activities. Our work shows that 1) it is possible to quantify previously only qualitative dimensions of observational studies, and 2) a computational method may be possible to reduce the need for large coding teams when running psychological studies. While there are major limitations to the evaluation methods presented in this study, we believe this work is a meaningful contribution to translational neuroscience research.

1. Introduction

The Filming Interactions to Nurture Development (FIND) is a video coaching program aimed at promoting positive interactions between caregiver and child. The program films caregivers playing with their children, then professional editors create a coaching video of play interactions that promote healthy brain development. This translational neuroscience research initiative has been shown to improve the language skills of families of low socio-economic backgrounds [3]. Our project aims to help scale up this intervention by exploring ways in which computer vision can augment the film editing process. We believe this project will be a good demonstration of the 3D to 2D reconstruction methodologies and contribute to the field of translational neuroscience.

The interactions of FIND are centered around a concept called "serve and return" which reinforces developmentally supportive interactions. The Stanford Center on Early Childhood has distilled these "server and return" interactions down to 5 distinct elements: back and forth, endings and beginnings, name, support and encouragement, and sharing the focus. Among these 5 elements, a common component is "joint attention". Our project will focus on examining joint attention in the context of the FIND research center and how computer vision can automate such quantification without the need for huge manual coding teams.

Joint attention is a behavior in which two people focus on an object or event for the purpose of interacting with each other. Prior work has shown that joint attention has been correlated with increases in the utterance object of shared attention, and could be beneficial to children's early language acquisition [10]. As mentioned above it is one of the key components to serve and return. Working with experts in the field we believe all elements of FIND will contain joint attention. FIND focuses on developing interventions starting with the child's "serve" which will require input from the developmental psychology literature and innovative ways of identifying computer vision methods that would be compatible with infants.

The overall goal of our research group is to create a human-in-the-loop AI video editing system that would speed up the existing manual editing process. Ideally, editors would no longer need to manually look for instances of "serve and return" by examining the raw film but instead use machine learning to identify instances of potential "serve and return" elements and then confirm such instances. The end result is to scale up the capabilities of the center in pro-

ducing high-quality coaching videos to help more families.

Additionally, in order for the center to improve its operations it is important to understand how to design an intervention that allows our machine-learning systems to perform optimally. Does the center need to change its video recording procedures? Would the incorporation of multiple cameras allow for better scene reconstruction that allows for more computer vision methods to be utilized when analyzing different interactions at scale? We hope that this project would allow us to get some insight into what is possible with computer vision techniques with the existing setup and suggest future designs for the intervention.

The scope of this class project is to create a prototype for identifying joint attention activities within the FIND video library. We hope to build a pipeline that will allow for the identify instances of joint attention where there is a specific object of joint attention. We will then use 3D learning methods to reconstruct the object into 3D. This will allow us to quantitatively analyze joint attention interactions within the context of the FIND intervention.

Our approach is to create a computer vision pipeline that incorporates a joint attention detector using foundational models such as contrastive language-image pre-training (CLIP) to detect images of joint attention within a video. Once those instances of joint attention have been identified we will use a joint attention detection model to identify the agents within the image and an object that those agents are focusing on. Finally, we use a monocular depth estimation model to determine where those objects are in relation to the view. The final output will be the 3D coordinates of the location of the object, child, and caretaker. These 3D coordinates will allow us to describe how actors within the view orientate themselves spatially when during joint attention interactions allowing for quantitative analysis of an important psychological phenomenon.

2. Background

The prior work section will be broken into two parts. We draw inspiration from psychology literature and computer vision literature.

2.1. Psychology

We will use manual observational coding schemes to help us determine meaningful directions in our computer vision methods. For example, using head-mounted eye trackers hand-eye coordination predicts joint attention [12]. In a literature review of 27 articles, joint attention was modeled by coordination with the initiation of or responses to pointing, showing, and/or coordinated looks between a person and object [11]. Showing that joint attention requires both recognition and coordination of objects.

2.2. Computer Vision

The existing method allows for predicting attention targets from the third-person point of view [1]. The 2D models mentioned before will give us an idea of where the object of joint attention to reconstructing the 3D image we plan on using NeRF [4]. While joint attention has been done before by using the VideoCoAtt dataset [8], such models do not focus on the context of early childhood interaction. Additionally, some projects look at joint attention in the first person perspective [2], which would not be appropriate for our use case creating videos in which caregivers can examine their own interaction with their child. Our approach is unique because it is done in observational settings (or in the wild), focuses on the infant, and looks to examine the relationship between pose, joint attention, and psychological construct.

2.2.1 Prompt Engineering

With the advancement of foundational models, a variety of tasks can be achieved with ease. CLIP allows for identifying visual concepts from natural language prompts [6]. This model can predict and find an image given a text snippet or prompt without directly optimizing for the task. This sort of zero-shot classification ability requires the careful crafting of text prompts to identify the correct images in a video and phrasing texts the describe the same underline image might result in differences in CLIP similarity score [5]. Part of this project will require prompt engineering to get the best use out of CLIP.

3. Problem Statement

This project aims to build a pipeline or a prototype to detect the 3D coordinates of the caregiver, child, and object that both agents are sharing joint attention of. The inputs of this prototype are the raw videos from the FIND video library and the outputs will be the 3D coordinates of joint attention. The pipeline will be raw video, then detect frames with joint attention interactions, detect the object of joint attention, then estimate the 3D coordinates of the object and actors within the frame. Because there are several steps to this pipeline the problem will be defined into three smaller tasks.

3.1. Automatic Detection of Joint Attention

The first task will be to automatically detect joint attention instances within frames of a video clip. For this project joint attention will be defined explicitly as:

1. Caregiver and child are looking at the same object

2. Caregiver and child looking at each other

3. Do not include when the caregiver is only looking at the child and the child is looking at the object

4. Do not include when the caregiver is acknowledging the child's attention but not looking at the object

In cases where it is uncertain if the caregiver and child are looking at the same object or each other it is rejected as joint attention This project uses a restrictive definition of joint attention, and will also exclude videos where there is more than one person in the scene. This narrow definition of joint attention will allow for easier evaluation of classified joint attention frames.

3.2. Object Detection Through Joint Attention

The second task of the project is to detect both the head location of the caregiver and child as well as the location of the object of joint attention. The coordinate of the head location will be defined as the single point in the middle of the headbox detector used. The object location will also be defined as a single point. This task of the project will take the frames detected from the previous section and run a joint attention model to detect the object and the heads of the child and caretaker.

3.3. Monocular Depth Estimation

The final task of the project is to estimate the 3D of the 2D points given in the previous section. Since we are defining the object and head location of the child and caretaker as single points we will be able to estimate how far the 2D coordinate is in relation to the camera and construct a triangle 3D. This will allow us to perform a descriptive analysis after the entire pipeline.

4. Approach

4.1. Dataset

The dataset will be composed of a library of FIND videos. These FIND videos are spiced into clips that are labeled a specific FIND element. Through prior work with the Stanford Center on Early Childhood[1] and the Stanford Autonomous Agents Lab [2], we were able to recover over 1470 clips of individual FIND elements. These clips vary in length based on the specific context of the study, and the type of element.

[1]https://earlychildhood.stanford.edu/

[2]https://www.autonomousagents.stanford.edu/

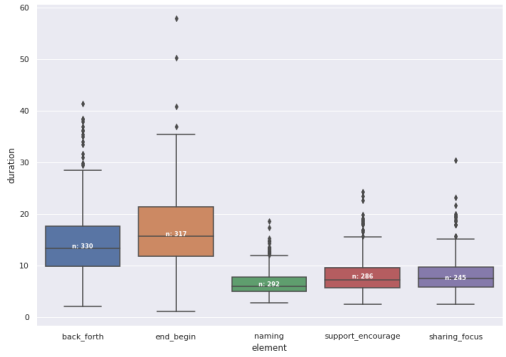

Figure 1. FIND elements grouped by duration

We have all the videos located on a secure google drive. Additionally, we have demographic information related to the families that participated in the FIND intervention, which will be used for analyzing differences between demographics. However, because this data is purely observational or in the wild data it will be a challenge to create a computer vision system that will optimally solve our ask. The following figure shows the entire pipeline for joint attention.

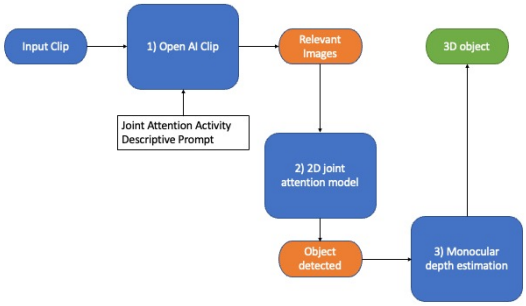

Figure 2. Pipeline for 3D coordinate reconstruction

4.2. Step 1: Joint Attention Frame Detection via CLIP

The first step of the project is to determine if a frame contains joint attention interaction or not. Since the CLIP generates a similarity score based on prompt and image pairs it is important to understand the cut-off threshold for clips with similarity scores below the threshold which would be determined not the be joint attention. The first step is to qualitatively analyze different prompts. After experimenting with a few prompts the following prompts were used **"a photograph of an adult and a child both LOOKING at the SAME object"** and **"a photograph of an adult and children LOOKING at each other"**. In order to determine the threshold, 260 random images were sampled from the video library (randomized the video clip, then ran-

domly select a frame from the clip), and manually labeled as joint attention or not. Then the model was given the image along with two prompts to determine the maximum similarity score among them.

4.2.1 CLIP Similarity Threshold

As we demonstrated earlier, there are 260 image instances in total, out of which 140 are hand-labeled as true joint attention. The similarity scores of these cases are ranging from 0.185 to 0.276. The remaining 120 cases are non-joint attention examples with similarity scores ranging from 0.152 to 0.24. We can calculate the False Positive Rate (FPR), True Positive Rate (TPR), and threshold values for different probability cutoffs using this data. In our case, we hope to achieve a TPR of 0.8, which gives us a threshold value of 0.1941.

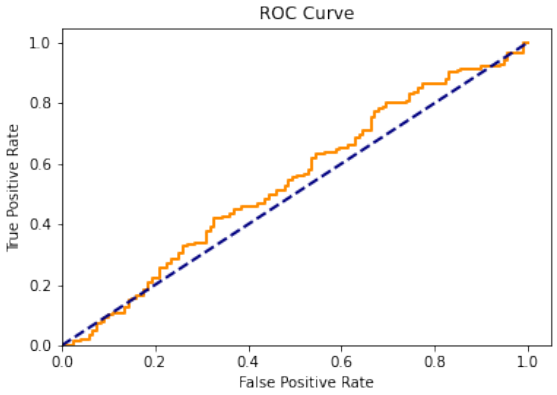

Figure 3. ROC Curve, Threshold for TPR = 0.8: 0.1941

The True Positive Rate (TPR) is defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP is the number of true positives and FN is the number of false negatives.

The False Positive Rate (FPR) is defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

where FP is the number of false positives and TN is the number of true negatives.

An example of a CDF curve for a single video shows that the threshold is around 0.2.

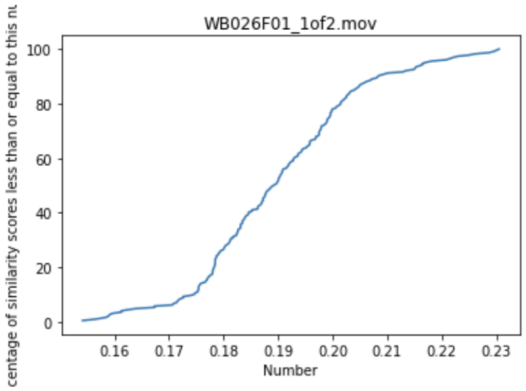

Figure 4. CDF Curve of WB026F01_1of2

4.3. Step 2: 2D Joint Attention Model

Once we have determined which images contained joint attention, a manual check was performed on the images identified through CLIP to feed into the 2D joint attention model. The model used is Attention Flow which uses saliency-augmented attention maps and 2 convolutional attention mechanisms to determine the localization of joint attention in 2D settings [9]. Attention Flow network has three main modules: encoder, generator, and co-attention map generation blocks.

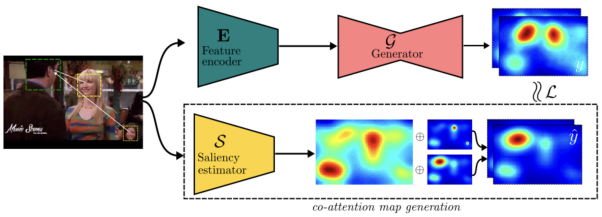

Figure 5. Attention Flow method is composed of three modules, (i) feature encoder, (ii) attention flow generator, and (iii) saliency-based ground truth generation. It estimates a two-channel heatmap, which encodes faces and their co-attention likelihood in the scene.

The output of Attention Flow is an image with the headbox detected and a heat map around the area of joint attention (the single point is defined as the largest number in the heat map).

Figure 6. Example from FIND data set approved for wide release

As shown in the figure above the head box shows one of the agents exhibiting joint attention[3]. The headbox will generate two $(x_1, y_1), (x_2, y_2)$ coordinates in 2-dimensions and the object of joint attention will be a single point in 2D. These coordinates and the image will be passed to the monocular depth estimation model.

4.4. Step 3: Monocular Depth Estimation

The model for monocular depth estimation is the Vision Transformers for Dense Prediction, which proposes a new architecture by using vision transformers [7].

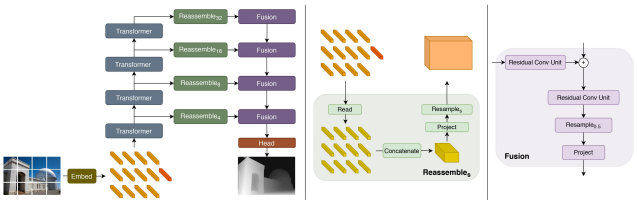

Figure 7. Input image is embedded into tokens by extracting non-overlapping patches followed by a linear project of their flattened representation using DPT-Large. The embedding is then augmented with positional embedding and patch-independent readout tokens. Passing through multiple transformer stages, the tokens are reassembled into image-like representations at multiple resolutions. The fusion module generates the fine-grained prediction.

The model is primarily used for zero-shot monocular depth estimation which fits our task perfectly. This DPT-Large model observes up to 28% improved relative performance when compared to previous methods.

[3]For some reason the model wasn't able to print both headboxes at once for the example

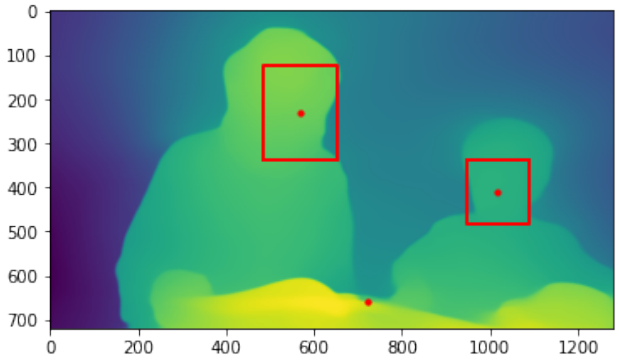

Figure 8. Example of depth estimation with coordinates of joint attention and the center of head boxes

The above example shows the depth estimation generated by DPT-Large and the three points of interest necessary for 3D reconstruction. Using the depth estimation image it is possible to estimate the z coordinate which gives the resulting point in three dimensions. Given the three points of the child, caretaker, and object it is possible to construct a triangle that will represent the relationship between the three entities.

4.4.1 3D Triangle Calculation

To capture the relationship between the three entities in the view, we propose using the area of the triangle created by the three points to represent the relationship. The following equation will allow the calculation of the triangle in 3D.

$$Area = \frac{1}{2} \|\vec{v_1} \times \vec{v_2}\| \quad (1)$$

$$\vec{v_1} = \vec{P_2} - \vec{P_1} \quad (2)$$

$$\vec{v_2} = \vec{P_3} - \vec{P_1} \quad (3)$$

Additionally, we normalize the triangle area by dividing it by the number of pixels in the image. This outcome will be used to quantify previously qualitative measures of joint attention. For example, a smaller triangle area indicates that the three entities are closer together while a larger area shows that the entities are further apart.

5. Analysis

The first analysis is to evaluate the accuracy of steps 1 and 2 of the process. Note that this section is still very noisy because we did not have a proper held-out data set. Additionally, the results are biased since when looking for data to test our models we prioritized instances of joint attention to ensure our modeling was working. This results in over representation of positive cases when performing our our evaluation. Essentially, our accuracy measure is only

based on cases where the model can detect if joint attention assuming there is always joint attention, we did not label enough negative data points to see if there are sufficient true negatives in our evaluation. This is to say our training data in step 2 only has positive instances of joint attention. We caution against using these numbers as a true evaluation of our pipeline, a proper train, dev, and test split should be used when validating the pipeline which includes balancing positive and negative classes. Additionally, step 2 is computationally costly and we were only able to run 9 examples before running out of compute thus not enough to make any concrete claims.

Step	Accuracy	Sample Size
Step 1: CLIP	80%	112
Step 2: Joint Attention	88%	9

To assess the accuracy of the CLIP model, we determined the percentage of cases where the score was above a certain threshold value and a positive label of joint attention. Out of 140 cases, 112 met this criterion, resulting in an accuracy of 0.8.

5.1. Computational Analysis of Joint Attention

We were able to find the normalized triangle area of all 9 videos. We wanted to see if there is a difference between the types joint attention activites that exist. To this end we define two activities: shared reading where caregiver and child are engaged in reading a book together, and playing behavior where caregiver and child are playing with objects that are not books.

Activity	Triangle Area	N
Reading	0.074	4
Playing	0.087	5

Table 1. Traingle Area of Different Interactions

As seen by the table above when there is joint attention and shared reading activity the area of the triangle is smaller compared to the playing behavior, this would indicate that during shared reading caregiver and child are closer together physically.

5.2. Regression Based On Income

By setting the area of the triangle to be the dependent and adding regressors for income and playing behavior we get the following coefficients.

Table 2

	Dependent variable:
	area
income	-0.030
	(0.022)
playing	-0.054
	(0.072)
income:playing	0.027
	(0.026)
Constant	0.135^{*}
	(0.054)
Observations	7
R^2	0.391
Adjusted R^2	-0.217
Residual Std. Error	0.037 (df = 3)
F Statistic	0.643 (df = 3; 3)
Note:	*p<0.1; **p<0.05; ***p<0.01

The table would indicate that as income increases the area of the triangle decreases. The playing variable is a $0, 1$ indicator variable where 0 is playing behavior while 1 is reading behavior. This shows that when reading occurs the triangle area also decreases. The interaction term is when shared reading occurs and increasing income results in the triangle area increase. With only 7 observations these coefficients are not significant, however, may encourage analysis in the future as it does some that the area of the triangle maybe a valid construct to be analyzed.

6. Limitations

There are several limitations to this work. One source of limitation is that the areas of the triangles are not correctly represented. The camera system within each scene is different which adds noise to the estimation. We try to normalize the noise by normalizing the size of the frame but the depth estimation may still not be an accurate representation. Additionally, we only had 9 total observations thus most of our analysis would not have enough statistical significance to make a claim. Finally, we have no ground truth labels for the depth estimation so there are only assumptions about the performance of the DPT-Large model.

We would like to again caution against the over-optimistic accuracy score in the analysis section. During the time of this work, we only had time to run positive cases of joint attention videos, with such unbalanced classes and a small sample size it is unreasonable to expect this pipeline

to perform as well as the metrics. The next phase of this project is to create a train, dev, and test data set that would be appropriate to evaluate the pipeline.

Another key limitation is that we used a deterministic similarity threshold of 0.1941 when running our CLIP model. As shown by 4; examining specific video clips shows that the similarity threshold varies by video clip. Each video has its own unique threshold that would be optimal for detecting frames of joint attention, future work will require understanding how to tune the threshold to each particular video.

The observation of joint attention was only from a single view, future implementations of the program may allow for multiple cameras to capture even more information reducing the noise in 3D reconstruction. This may allow for a richer analysis.

Finally, our pipeline requires human input at each one of the steps. While this may seem like a limitation at first, upon reflection we would suggest that a human-in-a-loop system is necessary for this sort of detection method. There are many nuances and judgment calls required to decide the proper identification of joint attention, while AI can speed up the detection of possible joint attention instances there are too many edge cases for complete automation of the task.

7. Conclusion

This project revealed many insights about using computer vision methods to analyze a psychological phenomenon. While there is still much work to be done as seen in the limitation section, this project provided a meaningful step forward. By looking at the area of a 3D triangle, we were able to suggest that during shared reading the caregiver, child, and object are closer together than during other play activities. This rudimentary result is encouraging and motivating as it shows that a monocular depth estimation method is a meaningful tool for quantifying certain interactions. With the aid of 3D computer vision previously manual tasks that require teams of coders could potentially be sped up and only require individual researchers to confirm the output. Additionally, this project reveals the complexity of evaluating wild data as it is difficult to evaluate this pipeline given all the complexities in the data set.

One major next step is to create a comprehensive evaluation dataset that is a true measure of performance for the pipeline. This would require careful labeling of positive joint attention instances, splitting the train and test sets based on family to prevent leakage, and determining how many frames per video would be sufficient. It is likely that the creation of such a dataset is several projects and will require a large labeling team and input from both experts in psychology and computer vision. One key learning of this project is how complicated applying these computational tools maybe if there are no existing evaluation metrics.

8. CODE:

Link to code is here: https://drive.google.com/drive/folders/1q1SCWCsB_u00_nEIbvRY1THUNgCc_pII?usp=sharing

References

[1] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[2] Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Temporal localization and spatial segmentation of joint attention in multiple first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 2

[3] Andrea Imhof, Sihong Liu, Lisa Schlueter, Tiffany Phu, Sarah Watamura, and Philip Fisher. Improving children's expressive language and auditory comprehension through responsive caregiving: Evidence from a randomized controlled trial of a strength-based video-coaching intervention. *Prevention Science*, 24(1):84–93, 2022. 1

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[5] Joseph Nelson. Prompt engineering: The magic words to using openai's clip, May 2021. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2

[7] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021. 5

[8] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[9] Ömer Sümer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. *CoRR*, abs/2001.03960, 2020. 4

[10] Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child Development*, 57(6):1454, 1986. 1

[11] Pamela J. White, Mark O'Reilly, William Streusand, Ann Levine, Jeff Sigafoos, Giulio Lancioni, Christina Fragale, Nigel Pierce, and Jeannie Aguilar. Best practices for teaching joint attention: A systematic review of the intervention literature. *Research in Autism Spectrum Disorders*, 5(4):1283–1295, 2011. 2

[12] Chen Yu and Linda B. Smith. Hand–eye coordination predicts joint attention. *Child Development*, 88(6):2060–2078, 2017. 2