Camera-based Assessment of Gendered Toy Preference in Free-Play Parent-Child Interactions

Peitong Li[1], Albert Ali Salah[1], Joyce J. Endendijk[2] and Ronald Poppe[1]

***Abstract*— This paper explores gendered toy preference in parent-child interactions. We focus on free-play, which allows for unique natural and dynamic interactions in which toy preferences might be less constrained by the experimental setting. We operationalize toy preference through the child's visual focus of attention (VFOA). Our analyses of 25 interactions of 12-13 minutes each reveal statistically significant differences between boys and girls in terms of time spent looking at a doll and a jump box. We then investigate whether these effects can also be obtained through automated analyses of the video data. To this end, we leverage an automated VFOA algorithm to predict which toys are attended to. Our automatic algorithm reveals similar patterns as when using manual annotations, albeit with less statistical power. This advancement holds promise for developmental research by providing efficient and objective assessments of children's interactions, potentially guiding early developmental interventions and informing strategies to mitigate gender bias in play environments[1][2].**

I. Introduction

The automated analysis of parent-child interactions (PCIs) has the potential to speed-up research into a wealth of verbal and nonverbal individual and dyadic behaviors and their relation to various developmental constructs [1]. Especially free-play parent-child interactions are rich in terms of affective and social signals [2]. In this paper, we analyze free-play scenarios where toys play an essential role. When children are allowed to freely choose which toys to play with, and how to play with them, we can investigate how a child engages with certain toys. For example, a child might be attracted to specific types of toys, or engage in specific types of play, either alone or with a parent. Such play preferences might aid in forecasting the child's development [3].

Toy play is highly gendered, because it is characterized by large average differences across groups of girls and boys [4]. The large body of research into gendered toy preference builds on either self- or child-reported preferences, or observations in controlled experiments. Furthermore, previous research has predominantly considered individual toy preferences. We argue that a setting in which a child can play with, or is stimulated by, a peer, parent, or caregiver, might invoke a shift in toy preference. We hypothesize that such an effect might occur because of different expectations on the one hand, and different toy affordances in terms of cooperative play on the other [5], [6]. Understanding these early patterns through automated analyses, especially in naturalistic free-play scenarios, can thus provide critical insights into developmental processes and potential early interventions. For instance, early detection of stereotyped behaviors could inform targeted approaches aimed at promoting gender equity and diversified developmental opportunities [7], [8].

In this paper, we depart from controlled experimental setting paradigms to focus on free-play, and explore toy use via camera-based automated approaches. Our contributions:

1) We explore children's gendered toy preference in natural and highly dynamic free-play interactions with a parent through the analysis of visual focus of attention.
2) We develop an automated processing pipeline to estimate visual focus of attention from two views, and to determine joint attention of parent and child towards the various toys.
3) Our analyses provide insights into the potential of examining gendered toy preference in free-play settings and highlights the challenges.

The remainder of this paper is structured as follows. We first discuss related work. Next, we investigate gendered toy preference from manually annotated visual focus of attention labels in Section III. We then turn to automated analysis in Section IV, and compare and discuss our findings in Section V. We conclude in Section VI.

II. Related Work

A. Gendered toy preference

The importance of play preferences in child development is well-documented. Studies have highlighted how different types of play, including toy play, contribute to cognitive, motor, and emotional development. Through play, children develop key skills like problem-solving, social interaction, and emotional resilience [9]. Toy preferences among children often exhibit significant gender[3] differences [10] that reflect broader societal gender norms [4]. Girls' toys are often associated with physical attractiveness, nurturance, and domestic skill, whereas boys' toys are rated as violent, competitive, and somewhat dangerous [11], [12]. Infants begin to exhibit distinct preferences for gender-associated toys by the age of 9 months, though such preferences are

[1]Peitong Li, Albert Ali Salah and Ronald Poppe are with the Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands {p.li1,r.w.poppe,a.a.salah}uu.nl

[2]Joyce J. Endendijk is with the Department of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands j.j.endendijk@uu.nl

[1]https://github.com/Chelseapt/Auto-Parent-Child-Assessment

[2]This is the uncorrected author proof. Please cite as: Li, P., A.A. Salah, J.J. Endendijk, R. Poppe, "Camera-based Assessment of Gendered Toy Preference in Free-Play Parent-Child Interaction," Proc. IEEE Int. Conf. on Development and Learning (ICDL), Prague, 2025.

[3]The term 'gender' is predominantly used to denote biological sex.

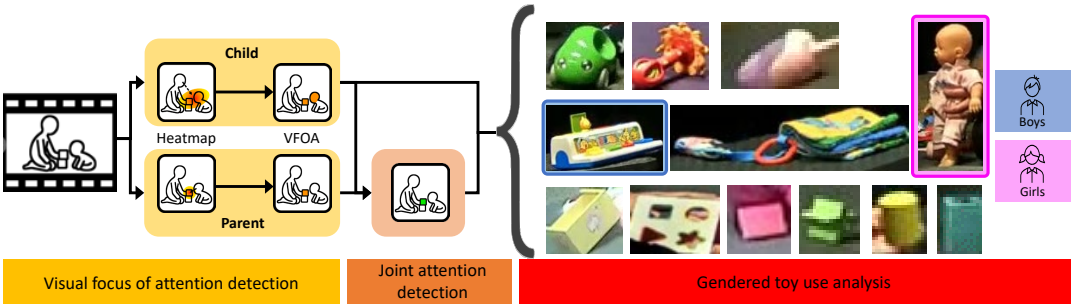

Fig. 1. **Overview of our automated approach**. We use gaze estimation and object detection to determine the visual focus of attention of both child and parent. We combine both to detect joint attention. Then we aggregate detections over an entire interaction to investigate gendered toy preference.

absent at 5 months [13], [14]. Kung et al. link early gender-typed play preferences to later gender-typed occupational interests, suggesting that the toys children prefer in preschool can predict their occupational interests a decade later [15]. Recent studies suggest these early preferences also relate to basic visual attention patterns. Infants tend to focus longer on objects that appear larger and more visually prominent during play [16].

It is important to examine social context in assessing gender-typed play preferences. The presence of a parent or teacher has been shown to magnify the level of gender-typed play in some studies but not others [13], [12], [2]. In addition to overall level differences between boys and girls, toy play has another gendered component, as it is also characterized by beliefs about the appropriateness of toys for each gender [17], [18].

Play environments allow infants more independence in selecting and engaging with toys, than the predominantly used semi-structured toy-play settings. In the latter, parents often guide the infants towards specific toys [19]. When parents differ in their joint attention depending on the gender-appropriateness of the toy, this could be considered as a subtle gender socialization practice that reinforces gender-typical toy play in children [20]. We argue that investigating gendered toy preferences in a free-play interactive setting with a parent could therefore reveal different insights compared to a setting in which only the child appears.

B. Automated analysis of VFOA

Automated analysis of visual focus of attention (VFOA) in parent-child interactions is crucial for understanding social communication skills, especially in children at risk for developmental disorders [21]. Early approaches to infant gaze estimation during parent-infant interaction using cameras focused on whether the infant's gaze was on the parent or not [22]. Free interactions were challenging to observe from a single camera, as a clear view of the faces of the interactants was necessary to observe joint attention. Ceiling-mounted depth cameras [23], head-mounted cameras [24], [25], and eye-tracking glasses [26] were used to overcome such difficulties.

Recent advances in computer vision algorithms enable powerful analysis capabilities. Typically, such algorithms first detect head locations, and then infer gaze targets based on the analyzed orientation of the head and the saliency of the image [27], [28]. By linking gaze to a predefined set of targets, including faces of others, visual focus of attention is estimated. When the VFOA of multiple people is considered, social attention patterns such as joint attention can be leveraged [29]. In the context of PCIs, Fraile et al. used a single viewpoint to investigate gaze patterns, but the interactants were sitting across each other at a table, which helped with constraining the poses [30]. Li et al. investigated a free-play scenario and leveraged [27] to detect joint attention and mutual gaze between the parent and the child [31]. Given the challenging nature of the data, a limited accuracy was achieved. In this paper, we show that using multiple camera viewpoints improves the analysis.

For the automated study of VFOA in PCIs, there are several datasets available. The recently introduced ChildPlay dataset [32] contains 400+ parent-child interactions, and comes with gaze annotations. Despite the tremendous value of such data for the development of automated analysis, we refrain from using ChildPlay because the lack of systematic involvement of a fixed set of toys. Instead, we focus on the YOUth Cohort dataset [33] that consists of 2000+ interactions in the same physical environment. For the purpose of the current study, we produced and make available gaze annotations for a subset of the data.

III. Gendered Toy Preference from Labeled Data

First, we focus on the potential of investigating a child's gendered toy preference in free-play parent-child interactions, by considering annotated visual focus of attention. We discuss the data and annotations used in our experiments, before presenting the results of our analyses. Our experiments with automatically estimated VFOA appear in Section IV.

A. Data and annotation

For our analyses, we use parent-child interaction videos from the YOUth Cohort study [33], which are accessible to researchers following an ethical approval process. These videos are recorded in a confined space (see Figure 5) and feature a parent and a child, approximately 10-12 months old, playing with various-sized toys on the floor. The interactions are relatively unstructured. Parents are instructed

beforehand to introduce the shape sorter and a textile book after some time. But when and how to play with the toys is not mandated, leading to significant variability in play behaviors. Each interaction is captured from four camera angles, but our analyses are limited to two views from opposing perspectives.

TABLE I
NUMBER OF HEAD ANNOTATIONS IN DIFFERENT CAMERA VIEWS.

Input	Parent	Child	2 heads
view1	1,841	1,819	1,791
view2	1,845	1,842	1,818
2 views	1,874	1,874	1,874

Selected recordings. We selected 25 interactions from a pool of videos (i.e., 960×540 pixels). Of these, 15 videos feature a boy, and in the remaining 10, a girl is interacting with the parent. Of the parents, two are fathers and the remaining 23 are mothers. Due to this skewed distribution, we will not investigate the gender of the parent in this paper.

We have temporally cropped the videos to begin when the experimenter exits the scene, leaving the parent and the child to play. The approximate duration is 12-13 minutes per video. The number of interactions considered is modest, which makes our statistical analyses conservative. However, the analysis is dense: the amount of analyzed footage spans 315 minutes (5:15 hours).

TABLE II
NUMBER OF FRAMES THE CHILD SPENT LOOKING AT EACH OF THE TOYS AND THE PERCENTAGE OF TIME SPENT LOOKING AT EACH TOY FOR BOYS AND GIRLS, SEPARATELY.

Toy	Count	Boys	Girls
doll	1746	6.13%	15.80%
car	1683	5.15%	5.72%
jump box	1726	35.50%	18.15%
shape sorter	1101	38.40%	38.97%
flower	1599	8.39%	5.40%
book	755	29.12%	38.57%
baby bottle	1065	2.78%	13.81%
green star	365	17.54%	24.34%
yellow cylinder	479	21.71%	21.65%
blue cube	371	19.32%	29.33%
pink triangle	400	12.39%	21.42%
shape sorter lid	999	28.99%	42.76%

Annotation. For manual annotation, we sampled a frame every 10 seconds. We annotated bounding boxes for all visible toys and heads using the DarkLabel 2.4 tool[4]. We differentiated between the heads of parents and children to aid subsequent analyses. A single coder annotated all heads and objects across two camera views.

In total, 1,874 frames are used in our study, containing at least one visible head of a child and one of a parent across both views. The distribution of the heads within these frames is detailed in Table I. The use of two camera views allows us to process more frames than either single view alone, allowing for a more comprehensive analysis.

Given the inherent ambiguity in determining VFOA due to (partial) occlusions of faces and toys, low resolution of the face in the image, or proximity of objects, VFOA annotations are somewhat subjective. To understand the difficulty of the annotation task, two coders independently annotated the VFOA for children and parents, for all toys. Annotations were based on simultaneous observation of the two camera views. This approach allowed annotators to accurately determine which toys were being focused on by either the parent or the child. Multiple VFOA attentions were allowed per person in each frame. This strategy is more applicable given the numerous challenging scenarios we encounter, where it becomes difficult to ascertain the specific object of attention. For instance, toys are commonly positioned close to each other, an individual's attention can span a larger area, and the line of sight of either parent or child can be obstructed. The modest inter-annotator agreement, measured using Cohen's kappa, yielded results of 63.57% (F1 score 0.671) for the VFOA of the children and 74.15% (F1 score 0.777) for the parent perspective, indicating that our task is challenging.

We consider joint attention to each toy when both child and parent attend to the toy at a specific moment. Joint attention represents a crucial social-cognitive milestone, revealing a child's developing interest in shared activities and social referencing. [34], [35]. We simply calculate joint attention as the binary AND of the VFOA of both interactants. Inter-rater agreement measured using Cohen's kappa for joint attention is 78.24% (F1 score 0.741).

Annotation summary. Table II presents an overview of the attention given to various toys by boys and girls. Green star, yellow cylinder, blue cube, and pink triangle are the shapes for the shape sorter. Each toy's visibility count indicates how often it was marked across 1,874 frames. We observe that there is significant variation in VFOA between toys. For example, the jump box, shape sorter, book, and the lid of the shape sorter are attended to more, especially in comparison to the doll, car, flower, and baby bottle. The occurrences of some toys are correlated. For example, when the lid is on the shape sorter, VFOA is typically annotated for both, or neither. When a shape is about to be placed in the shape sorter, a similar situation occurs. This also happens when the baby bottle is positioned at the doll's mouth.

B. Analysis and results

We first investigate whether we can observe known gendered toy patterns in free-play settings. To this end, we investigate a potential difference in VFOA for different toys by comparing the average VFOA between the PCIs with boys (15 subjects) to PCIs with girls (10 subjects). Per interaction, we calculate the average VFOA to each toy by dividing the number of frames spent looking at a toy by the number of frames the toy is visible. Because a child can attend to multiple (or no) toys at each moment, and toys are not always visible, percentages do not necessarily sum up to 100%. Average VFOA is summarized in Figure 2.

[4]https://github.com/darkpgmr/DarkLabel

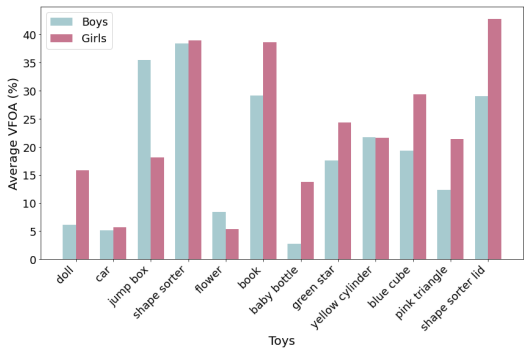

Fig. 2. **Average VFOA (%) for each toy** for boys (blue) and girls (pink), obtained via manual annotation.

Figure 2 reveals differences in VFOA between boys and girls. Overall, girls spent more time looking at the toys. This effect is mainly because boys more often tried to move out of the play area, consequently not focusing on toys.

Difference between toys. The toys in the YOUth Cohort dataset have not been selected to be gender-specific, nor is preference data reported. To this end, our investigations are exploratory and we will align with patterns found in literature. For several individual toys, we notice gender differences. Girls (15.80%) show a notably higher attention ratio for the doll compared to boys (6.13%). Dolls are traditionally more often preferred by girls [17], [36], which aligns with the data. A similar observation can be made for the baby bottle, which is more often attended to by girls (13.81%) than boys (2.78%). In many cases, the baby bottle was used together with the doll. In several occasions, children put the baby bottle in their own mouths, but there would typically not be gaze on the object in these situations.

Boys demonstrated a substantially higher average VFOA (35.50%) for the jump box compared to girls (18.15%). The jump box might be associated with more active play, which can be more aligned with traditional boy play patterns [11]. The attention ratios for the car, generally found to be more preferred by boys, were quite similar between boys and girls. Overall, the VFOA of the car is low. Over all videos, we noticed very little active play with it. None of the children were actively pushing or moving the car, irrespective of the gender of the child.

Apart from its lid, the average VFOA for the shape sorter and its parts was relatively similar for boys and girls. One reason could be that parents are instructed beforehand to introduce the shape sorter at some time and play with it. This might have caused a certain persistence for the parent in engaging with the toy, while the child might not have directly preferred it over other toys. A similar observation can be made for the textile book. Parents were instructed to take out the toy at some point. Many parents then put the child on their laps. Frequently, the child then tried to move away, towards a more preferred toy.

Statistical analysis. We conducted a series of t-tests between the average VFOA for boys and girls for each toy. From all comparisons, we only observed a statistically significant effect for the doll ($t(23) = -2.651, p = 0.022$) and jump box ($t(23) = 3.437, p = 0.002$). The same effect is observed in joint attention, with $t(23) = -2.601, p = 0.0233$ for the doll and $t(23) = 2.650, p = 0.0145$ for the jump box. Given that we consider 12 toys separately, the accepted approach is to apply a Bonferroni correction to the initial $\alpha = 0.05$, reducing our significance level to $\alpha_B = 0.05/12 \approx 0.0042$. After accounting for multiple comparisons, only the difference in the jump box data remains statistically significant.

Overall, the data suggest that some traditional gendered preferences in toy preference continue to be evident, with girls showing a higher interest in dolls and boys showing a stronger preference for more active play toys such as the jump box. However, a pronounced preferrence was not found for the majority of toys. While several co-founding factors have been mentioned previously, it is of note that no gender differences were found for the car. This is at variance with the literature, where there is broad consensus that cars are generally preferred by boys [4]. The softer appearance and color could have reduced the appeal for boys [14].

C. *Qualitative analysis*

The dynamic and varied nature of free-play is reflected in the VFOA. Over an interaction, the attention of both child and parent changes rapidly. We show the annotated VFOA over time for all considered interactions in Figure 3. For viewing considerations, we have merged several toys into a single category.

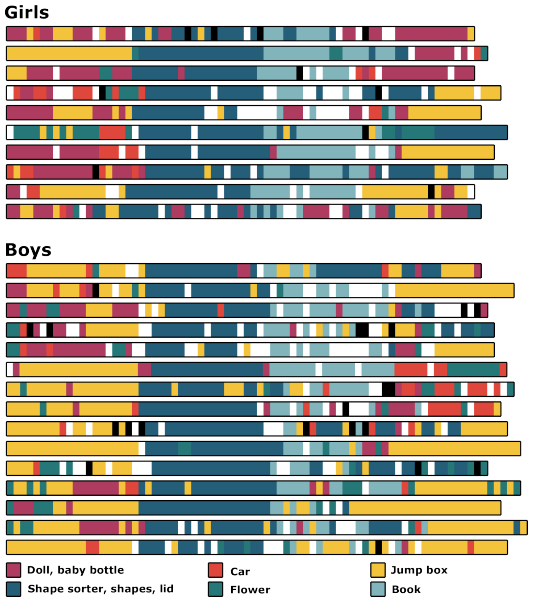

Fig. 3. **Visualization of annotated VFOA** for combined toy categories for all 25 videos, ordered by gender. Black is VFOA to the head of the parent, white is no VFOA to any of the considered targets. When VFOA covered two categories simultaneously, one was chosen randomly.

When comparing the VFOA patterns of these videos, several observations can be made. First, we observe that the shape sorter is introduced at around one third of the time

in all interactions. For most interactions, we observe a prolonged episode of focused interaction. Similarly, we observe that the textile book is introduced around two thirds into the interaction in both videos. While parents were instructed to read the book with their children, several interactions show the child looking at other objects, or outside the scene. From these visualizations, we can quickly verify that parents have followed the instructions, but that the effect of the introduction of the toys varied.

Especially near the start and end of the session, children were free to choose which toy to play with, consequently revealing a stronger toy preference. These observations emphasize the importance of observing play behavior in unconstrained settings. Despite gender patterns, the differences between individual subjects are large. For example, three girls focus predominantly on the doll near the end of the interaction. None of the boys play with the doll in the final phase. However, three boys appear to interact with the car.

IV. Automated Analysis of Gendered Toy Preference

We now turn to the automated analysis of gendered toy preference, to investigate whether our gendered toy preference findings are replicated. We describe the methodology and results of our automated pipeline. We then look at VFOA aggregated over an interaction in Section IV-E.

A. Automated analysis methodology

Figure 1 schematically describes our processing pipeline. We first discuss the estimation of VFOA for a single view, before detailing how we process multiple views. Finally, we assess joint attention by fusing the VFOA data of both parent and child, focusing on simultaneous gaze towards the same toy. Our approach thus not only provides an assessment of individual attention, but also captures moments of joint attention, essential for understanding interactive behaviors.

VFOA estimation. We generate gaze attention heatmaps for the parent and child independently using the method in [27], based on the annotated head locations. This method takes an image frame and the region of the head represented as a bounding box. The algorithm then assesses the probability that a specific 2D location within the frame is the focus of attention. Given the absence of depth information, the emphasis is placed on salient regions, based on the assumption that attention is often directed towards objects that stand out from their environment [34].

The method employs a CNN with two primary components: a head feature extractor and a scene feature extractor. The head feature extractor isolates and analyzes facial regions within video frames to produce feature vectors. The scene feature extractor processes different elements of the scene relevant to gaze direction, generating corresponding feature vectors. These vectors are input into a Multi-Layer Perceptron (MLP) that predicts the focus of gaze in each frame, resulting in a heatmap of gaze likelihood.

Our free-play setting includes several toys that the child (or the parent) can play with (see Figure 1). To determine the VFOA to each toy and the head of the interaction partner (parent or child), we integrate the gaze heatmaps with the annotations of the corresponding toy bounding boxes, following [31]. Specifically, we calculate the average heatmap values within each region, denoted by $VFOA_{\text{avg}}$, and also record the maximum value within each region, denoted by $VFOA_{\text{max}}$. While the former feature is more distinctive when the entire object is attended to, the latter might be more informative when objects cover a larger area in the frame. Given that the toys in our data vary significantly in size, we opt to use both features. This approach improves upon previous methods by using an MLP to dynamically model attention outcomes.

Multi-view estimation. By incorporating multiple camera views, we enhance the accuracy of VFOA assessments and can account for missing head and toy observations and potentially unfavorable viewing angles. For each view and toy, we calculate $VFOA_{\text{avg}}$ and $VFOA_{\text{max}}$ for both parent and child. For an interactant, the attention representation is a vector of $N + 1$ dimensions, where N is the number of objects of interests (i.e., toys) in the scene, and $+1$ is for the head of the other person. Since our setting contains 12 toys, with two interactants and two views, each moment in our data results in 52-dimensional binary feature vectors, indicating attention of both parties to toys and each other.

We set feature values to zero when an object is absent, corresponding to a VFOA probability of zero. Given the challenge of isolating a single target when other objects are nearby, we employ a multi-attention strategy. This approach allows both parents and children to potentially attend to multiple targets simultaneously, enabling multiple dimensions of the label vector to be set to one.

Since we have a classification task with limited input dimensionality and a restricted amount of training data, a simple MLP is used for multi-class classification. The network has a single hidden layer with 20 neurons. This model is capable of learning correlations between objects. When two objects are in close proximity, there may be biases about which object is typically attended to. For example, when a small shape toy is inserted into a shape sorter, the VFOA is typically on the shape. Since the regions of shape and shape sorter overlap, both $VFOA_{\text{avg}}$ and $VFOA_{\text{max}}$ of these two objects will be high. In this case, the MLP can learn to prioritize the smaller shape.The MLP directly outputs the predicted VFOA results.

Joint attention estimation. Joint attention at a toy requires that both parent and child look at it. By combining the VFOA of both parent and child, we obtain a binary indication of joint attention, calculated for each object that we consider. The processing of the parent's and child's VFOA is identical.

B. Automated experiment setup

We developed the automated pipeline detailed in Section IV-A. Our experimentation follows a leave-one-session-out strategy. The child and parent MLPs to determine VFOA for each toy and face are trained on the data of 24 PCIs,

whereas the data of the final recording is used for testing. Results are averaged over all 25 test sessions.

Baseline. To better understand the merits of our automated pipeline quantitatively, we introduce two baselines. Both use heuristics to predict the visual focus of attention on each toy and head from the perspectives of both the child and the parent, involving a series of binary decisions. The first baseline (*50-50*) assumes a naive 50% probability that the visual focus is on any given target. A 50% probability is high, and ignores differences in the prior distribution of VFOA over the various toys. As an alternative, we introduce a second baseline (*prior*) that predicts for each detected toy in the test recording, for both parent and child, the VFOA with a probability that is calculated from the training set. In this baseline, the number of estimated attended targets is typically much lower compared to the 50-50 baseline.

Measures. We employ precision, recall, and F1 score as our evaluation measures. The F1 score is particularly effective in addressing sample imbalance, providing an objective performance measure. For parent and child, we compute these measures for each of the 12 toys and one head. We then average the results across all 13 targets. Only toys and heads that were annotated are considered, as the visibility and number of objects varies per frame.

TABLE III

PERFORMANCE OF AUTOMATED VFOA AND JOINT ATTENTION DETECTION ACROSS VARIOUS COMPUTATIONAL SETTINGS.

Setting	Perspective	Precision	Recall	F1 score
2 views	parent	45.92%	34.76%	38.23%
	child	48.57%	35.01%	40.04%
	joint attention	49.05%	32.35%	38.16%
view1	parent	29.63%	11.99%	15.39%
	child	38.33%	14.96%	20.34%
	joint attention	40.39%	7.35%	10.97%
view2	parent	50.22%	24.29%	28.89%
	child	42.27%	23.76%	28.55%
	joint attention	41.81%	16.82%	21.64%
baseline (50/50)	parent	16.95%	49.48%	23.67%
	child	15.17%	46.41%	21.19%
	joint attention	13.43%	22.94%	15.58%
baseline (prior)	parent	15.71%	17.89%	16.66%
	child	16.73%	19.94%	18.11%
	joint attention	12.27%	5.28%	7.21%

C. Automated results

Multi vs. single-view. From Table III, we observe that the multi-view setup yields higher performance, compared to both single-view scenarios. The multi-view setting improves the F1 score by 20-23% and 10-12% for the first and second view, respectively. Using two views partly mitigates the shortcomings of single views and underscores the value of incorporating multiple perspectives in VFOA detection tasks. Still, both recall and precision values are modest. From the inter-rater reliability analyses, we already concluded that VFOA estimation is not a trivial task.

Compared to the 50-50 baseline, the multi-view's F1 score enhancement is noteworthy. The multi-view exceeds by approximately 14.56% and 18.85% for parent and child perspectives, respectively. The performance of the prior baseline experiments is suboptimal due to the low probabilities of actual visual focus of attention (VFOA).

Parent vs. child. There is a difference in automated VFOA classification performance between parent and child. The child perspective consistently yields a higher precision than the parent perspective across the single- and multi-view settings. This may suggest that the children's interactions with their environment, possibly more concentrated and less obstructed, are easier to track, or that the patterns of children's engagement with objects are more distinctly captured due to their more predictable behavior. An alternative explanation lies in the relative position of the child in the play space. Typically, the child is closer to the toys, with larger viewing angles between toys. In contrast, the parent is more often situated at the side of the play area, with both toys and child in roughly the same viewing direction.

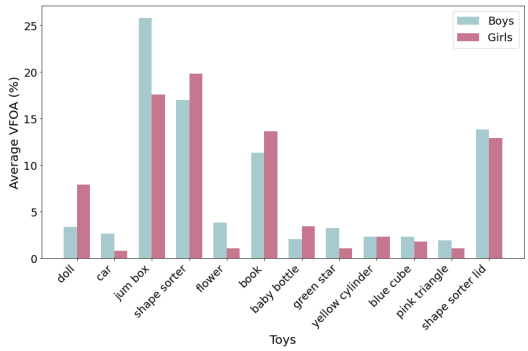

Fig. 4. **Average VFOA (%) for each toy** for boys (blue) and girls (pink), obtained from the automated detection using the multi-view setting.

Toys. We show the automatically assessed VFOA per toy using the multi-view setting in Figure 4. When compared to Figure 2, a couple of observations can be made. First, detected VFOA is, on average, lower than manually annotated VFOA. We discussed before that the lower recall of the automated detection is the main cause. Especially for the smaller objects, such as the shapes of the shape sorter, the car, and the baby bottle, we observe much lower ratios. Small inaccuracies of the gaze heatmap can have a larger effect on these small objects, since the effects cannot be averaged over larger areas. Still, the effects observed with manual annotations, such as girls' preference for the doll and boys' preference for the jump box are apparent.

Statistical analysis. To understand whether the gender differences found when using annotated data remain when applying automated analyses, we perform t-tests between the VFOA estimated for boys and girls, for the doll and jump box only. Differences were statistically significantly different for the jump box ($t(23) = 2.175$, $p = 0.0416$) but not the doll ($t(23) = -1.846$, $p = 0.0926$).

Joint attention. Results for joint attention also appear in Table III. Again, joint attention benefits from the multi-view

setup. Regarding differences between boys and girls in joint attention for the doll and jump box, we conducted two t-tests. Neither of those showed a statistically significant difference for gender, $t(23) = -1.613$, $p = 0.178$ and $t(23) = 1.503$, $p = 0.176$ for the doll and jump box, respectively.

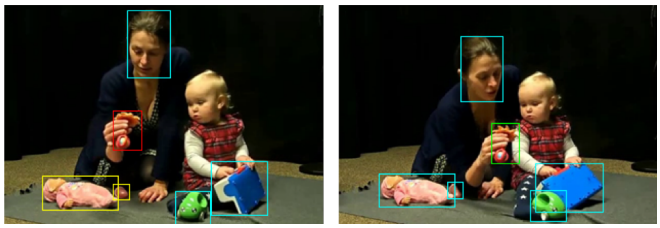

Fig. 5. **Example visualizations of child VFOA predictions** on a video not used in the experiment (printing permitted). Objects and mother's head are potential child VFOA targets. Green: correct prediction, Red: Incorrectly predicted to be attended, Yellow: Missed detection.

D. Visualization of results

To better understand how the automated VFOA estimations relate to manual annotations, we visualize example outputs side-by-side in Figure 5. We observe that the VFOA predictions appear in the line of sight, with incorrect predictions caused by depth ambiguities and the limited resolution to analyze eye gaze. We also observe mismatches between the number of annotated and predicted targets, which significantly affects the performance measures. VFOA annotation is often based on the assumption that manipulation of an object requires and draws attention. However, the automated VFOA predictions have no readily available information about object manipulation.

E. Aggregation over the entire interaction

Our automated analyses were at the frame level but often, we are interested in statistics aggregated over an interaction. Here, we investigate the relation between annotated and automatically predicted VFOA over the entire interaction.

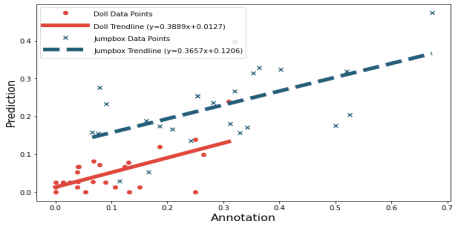

Fig. 6. **Correlation of manually annotated and automatically predicted VFOA** (red for doll and blue dashed for jump box).

VFOA and gendered toy preference. We investigate whether gender differences found for the jump box and doll are replicated with automated measurements. For each toy, we calculated the VFOA ratio based on both annotated and detected results. This approach provided us with 25 samples per toy, allowing us to compute trendlines (see Figure 6).

The Pearson correlation between the predictions and the annotations of doll and jump box are $r(23) = 0.635$, $p < 0.001$ and $r(23) = 0.573$, $p = 0.003$, respectively. Specifically, the trendlines for children's attention to the doll and jump box were found to be $y = 0.389x + 0.013$ and $y = 0.366x + 0.121$, respectively. The slopes reflect the lower recall values for automated VFOA estimations. We also conducted a t-test on the automatically estimated VFOA between boys and girls, for each toy. The results revealed a statistically significant effect for the jump box, with $t(23) = 2.175$, $p = 0.042$, but not the doll ($t(23) = -1.846$, $p = 0.093$). When compensating for multiple tests, none of the differences are statistically significant.

The correlation between our automated results and the ground truth highlights the merits of our automated approach in estimating VFOA. However, the effects found between boys are girls are weaker from a statistical perspective.

Joint attention and gendered toy preference. We performed similar analyses for joint attention on the jump box and doll. Again, the measures are correlated. Two t-tests between the joint attention for boys and girls did not reveal a statistically significant effect for either doll ($t(23) = -1.613$, $p = 0.178$) or jump box ($t(23) = 1.503$, $p = 0.176$).

V. Discussion

Structured vs. free-play. Departing from the predominant controlled research into gendered toy preference [4], we have used a free-play setting. While no ground truth about the children's preferences was available, our findings largely align with the literature. Our findings indicate that children's gaze was predominantly directed towards the shape sorter and book, which aligns with the structured process of the experimental design. This process seems to have effectively channeled the children's attention, suggesting that the order and nature of activities can significantly shape VFOA.

Gendered toy preference. We observed different amounts of VFOA for several toys. VFOA towards the doll and jump box was statistically significant. In the literature, cars are reported to be preferred more by boys [4], which we do not see. The car in our study does not look realistic and it is too large to manipulate with a single hand, which may have been a factor.

Automated vs. manual coding. While manually annotated and automatically estimated VFOA measurements were correlated, the recall of automated VFOA was notably lower. The correlation was approximately 0.6, which indicates that some of the strength of the signal is lost in the automation. Consequently, statistically significant differences between boys and girls in VFOA for different toys were not fully replicated. While the amount of VFOA to the jump box was still statistically significant, this was not true for the doll. Our sample size is low, as manual annotation is very time-consuming, but automated coding will help us to process much more data, which could reveal whether gendered differences are more systematic or not. Moreover, our automated analyses could further be used to discover other factors that influence object manipulation and toy preferences.

Limitations. Due to the complexity of the data, including occlusions and varied zoom levels, we relied on manual

annotations of head and toy positions. Achieving robust automatic detection proved challenging.

While our analyses cover over 5 hours, only 25 children were involved. In our free-play setting, we did not control for potentially co-founding factors such as equal toy visibility. Our analyses show the merits of a free-play setting and the potential for automated processing, but observing more interactions could potentially reveal more patterns.

Finally, we did not explore the role of the parent in shaping the interactions. While the introduction of toys by the parent clearly shaped childrens' VFOA, a similar but more modest effect could have occurred for the other toys.

Future work should consider a larger cohort. Employing sequential analysis on automated analyses frame-by-frame could elucidate interaction patterns. For instance, investigating whether children's gaze follows parents' gaze or vice versa could reveal insights into the child's development.

VI. Conclusion

We have explored gendered toy preference in free-play parent-child interactions. Our analyses reveal distinctions between boys and girls in the time spent looking at various toys, aligning with the literature. To investigate whether similar observations could be made automatically, we have developed an automated processing pipeline to estimate child's and parent's VFOA and joint attention towards various toys. Natural free play settings are more challenging than controlled lab settings. By using two cameras, we addressed the limitations of a single viewpoint, thereby enhancing the reliability of our results. Despite a lower recall in the VFOA estimations of the toys, findings of our analysis of the manual annotations are largely replicated. This indicates the potential to investigate patterns in gendered toy preference or other object manipulations in an automated manner.

References

[1] B. Karaca, A. A. Salah, J. Denissen, R. Poppe, and S. M. de Zwarte, "Survey of automated methods for nonverbal behavior analysis in parent-child interactions," in *Proc. FG*, 2024.

[2] E. Perkovich and H. Yoshida, "Infant sex effect on naturally occurring attention behaviors during interactive object play," in *Proc. ICDL*. IEEE, 2024, pp. 1–6.

[3] L. Sun, D. J. Francis, Y. Nagai, and H. Yoshida, "Early development of saliency-driven attention through object manipulation," *Acta Psychologica*, vol. 243, p. 104124, 2024.

[4] J. T. Davis and M. Hines, "How large are gender differences in toy preferences? A systematic review and meta-analysis of toy preference research," *Archives of Sexual Behavior*, vol. 49, no. 2, pp. 373–394, 2020.

[5] C. L. Martin, O. Kornienko, D. R. Schaefer, L. D. Hanish, R. A. Fabes, and P. Goble, "The role of sex of peers and gender-typed activities in young children's peer affiliative networks: A longitudinal analysis of selection and influence," *Child Development*, vol. 84, no. 3, pp. 921—-937, 2013.

[6] L. A. Serbin, J. M. Connor, C. J. Burchardt, and C. C. Citron, "Effects of peer presence on sex-typing of children's play behavior," *Journal of Experimental Child Psychology*, vol. 27, no. 2, pp. 303–309, 1979.

[7] M. Jover and M. Gratier, "Toward a multimodal and continuous approach of infant-adult interactions," *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, vol. 24, pp. 5–47, 08 2023.

[8] M. Verde Cagiao, C. Nieto, and R. Campos, "Mother-infant co-regulation from 0 to 2 years: The role of copy behaviors. a systematic review," *Infant Behavior and Development*, vol. 68, 08 2022.

[9] D. Whitebread *et al.*, *The role of play in children's development: A review of the evidence*. LEGO Fonden Billund, Denmark, 2017.

[10] B. Francis, "Gender, toys and learning," *Oxford Review of Education*, vol. 36, no. 3, pp. 325–344, 2010.

[11] J. E. O. Blakemore and R. E. Centers, "Characteristics of boys' and girls' toys," *Sex roles*, vol. 53, pp. 619–633, 2005.

[12] B. K. Todd, R. A. Fischer, S. Di Costa, A. Roestorf, K. Harbour, P. Hardiman, and J. A. Barry, "Sex differences in children's toy preferences: A systematic review, meta-regression, and meta-analysis," *Infant and Child Development*, vol. 27, no. 2, p. e2064, 2018.

[13] J. L. Boe and R. J. Woods, "Parents' influence on infants' gender-typed toy preferences," *Sex roles*, vol. 79, pp. 358–373, 2018.

[14] B. K. Todd, J. A. Barry, and S. A. Thommessen, "Preferences for 'gender-typed' toys in boys and girls aged 9 to 32 months," *Infant and Child Development*, vol. 26, p. e1986, 2017.

[15] K. T. Kung, "Preschool gender-typed play behavior predicts adolescent gender-typed occupational interests: A 10-year longitudinal study," *Archives of Sexual Behavior*, vol. 50, no. 3, pp. 843–851, 2021.

[16] L. Sun and H. Yoshida, "Effects of viewed object size and scene saliency on sustained attention in parent-infant object play," in *Proc. ICDL*. IEEE, 2024, pp. 1–6.

[17] J. Endendijk *et al.*, "Boys don't play with dolls: Mothers' and fathers' gender talk during picture book reading," *Parenting*, vol. 14, no. 3-4, pp. 141–161, 2014.

[18] C. Leaper and R. S. Bigler, "Societal causes and consequences of gender typing of children's toys," *Sex roles*, vol. 79, p. 253–259, 2018.

[19] V. Mateus, C. Martins, A. Osório, E. C. Martins, and I. Soares, "Joint attention at 10 months of age in infant–mother dyads: Contrasting free toy-play with semi-structured toy-play," *Infant Behavior and Development*, vol. 36, no. 1, pp. 176–179, 2013.

[20] E. E. de Vries, L. D. van der Pol, M. G. Groeneveld, and J. Mesman, "Fathers' and mothers' sensitivity during free play with gendered toys." *Journal of Family Psychology*, vol. 37, no. 7, pp. 1106–1114, 2023.

[21] K. Higuchi, S. Matsuda, R. Kamikubo, T. Enomoto, Y. Sugano, J. Yamamoto, and Y. Sato, "Visualizing gaze direction to support video coding of social attention for children with autism spectrum disorder," in *Proc. IUI*, 2018, pp. 571–582.

[22] S. Cadavid, M. H. Mahoor, D. S. Messinger, and J. F. Cohn, "Automated classification of gaze direction using spectral regression and support vector machine," in *Proc. ACII Workshops*, 2009.

[23] D. Cazzato, P. L. Mazzeo, P. Spagnolo, and C. Distante, "Automatic joint attention detection during interaction with a humanoid robot," in *Proc. ICSR*. Springer, 2015, pp. 124–134.

[24] E. Chong *et al.*, "Detecting gaze towards eyes in natural social interactions and its use in child assessment," *Proc. IMWUT*, vol. 1, no. 3, pp. 1–20, 2017.

[25] C. Yu and L. B. Smith, "Linking joint attention with hand-eye coordination–a sensorimotor approach to understanding child-parent social interaction," in *The Annual Conference of the Cognitive Science Society*, vol. 2015. NIH Public Access, 2015, p. 2763.

[26] P. Venuprasad *et al.*, "Characterizing joint attention behavior during real world interactions using automated object and gaze detection," in *Proc. ETRA*, 2019.

[27] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg, "Detecting attended visual targets in video," in *Proc. CVPR*, 2020, pp. 5396–5406.

[28] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, "Laeo-net++: Revisiting people looking at each other in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3069–3081, 2022.

[29] A. Gupta, S. Tafasca, A. Farkhondeh, P. Vuillecard, and J.-m. Odobez, "MTGS: A novel framework for multi-person temporal gaze following and social gaze prediction," *Proc. NeurIPS*, 2024.

[30] M. Fraile, C. Fawcett, J. Lindblad, N. Sladoje, and G. Castellano, "End-to-end learning and analysis of infant engagement during guided play: Prediction and explainability," in *Proc. ICMI*, 2022, pp. 444–454.

[31] P. Li, H. Lu, R. W. Poppe, and A. A. Salah, "Automated detection of joint attention and mutual gaze in free play parent-child interactions," in *Proc. ICMI Companion*, 2023, pp. 374–382.

[32] S. Tafasca, A. Gupta, and J.-M. Odobez, "Childplay: A new benchmark for understanding children's gaze behaviour," in *Proc. CVPR*, 2023, pp. 20 935–20 946.

[33] N. C. Onland-Moret *et al.*, "The youth study: Rationale, design, and study procedures," *Developmental Cognitive Neuroscience*, vol. 46, p. A100868, 2020.

[34] Z. Yücel, A. A. Salah, Ç. Meriçli, T. Meriçli, R. Valenti, and T. Gevers, "Joint attention by gaze interpolation and saliency," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 829–842, 2013.
[35] M. Tomasello and M. J. Farrar, "Joint attention and early language," *Child development*, vol. 57, no. 6, pp. 1454–1463, 1986.
[36] M. Kollmayer, M.-T. Schultes, B. Schober, T. Hodosi, and C. Spiel, "Parents' judgments about the desirability of toys for their children: Associations with gender role attitudes, gender-typing of toys, and demographics," *Sex roles*, vol. 79, pp. 329–341, 2018.