# It's the Child's Body: The Role of Toddler and Parent in Selecting Toddler's Visual Experience

Tian Xu Yu Chen Linda Smith
Department of Psychological and Brain Sciences, Cognitive Science and Computer Science1101 East 10th Street, Bloomington, IN, 47405 USA

***Abstract*— Human visual experience is tightly coupled to action – to the perceiver's eye, head, hand and body movements. Social interactions and joint attention are also tied to action, to the mutually influencing and coupled eye, head, hand and body movements of the participants. This study considers the role of the child's own sensory-motor dynamics and those of the social partner in structuring the visual experiences of the toddler. To capture the first-person visual experience, a mini head-mounted camera was placed on the participants' forehead. Two social contexts were studied: 1) *parent-child play* wherein children and parents jointly played with toys; and 2) *child play alone* wherein parents were asked to read a document while letting the child play by himself. Visual information from the child's first person view and manual actions from both participants were processed and analyzed. The main finding is that the dynamics of the toddler's visual experience did not differ significantly between the two conditions, showing in both conditions highly selective views that largely reduced noise perceived by the child. These views were strongly related to the child's own head and hand actions. Although the dynamics of children's visual experience appear dependent mainly on their own body dynamics, parents also play a complementary role in selecting the targets for the child's momentary attention.**

***Index Terms*— Embodied cognition, Perception and Action, Child-Parent Interaction**

## I. Introduction

Children learn about their world through their actions, as they hold, explore, stack, and play with objects. Children's bodily interactions during toy play – posture shifts, head turns, reaching, holding and turning objects – all determine the momentary visual information. Children also learn through their interactions with the social partners, who show and demonstrate objects, and introduce interesting objects and actions into play. The overall goal of the present study is to document and analyze the structure of children's dynamic visual experiences as they relate to the active participation with objects in the physical world and as they relate to the actions exhibited by the mature social partners.

Developmentalists such as Gibson [1] and Ruff [2] have studied the powerful dynamic visual information that emerges as infants and children move their eyes, heads and bodies, and as they act on objects in the world. In addition, Bertenthal and Campos [3] have shown how movement, such as crawling and walking over, under, and around obstacles, creates dynamic visual information crucial to children's developing knowledge about space. Computational theorists and roboticists Ballard [4], Metta and Fitzpatrick [5] have demonstrated the computational advantages of what they call "active vision", how an observer (human or robot) is able to understand a visual environment more effectively and efficiently by interacting with it. This is because perception and action form a closed loop: attentional acts are preparatory to and made manifest in action while also constraining perception in the next moment.

In our recent work, we proposed and implemented a multimodal sensing system for recording the visual input from the child's point of view by attaching a mini-camera on the forehead of young children close to eyes. This system provides a record of the head-centered available visual information as the child engages with objects. Past work indicates that the child's view – during toy play with the parent – is substantially different from the visual information recorded in the parent's head camera [6, 7]. The child's view is more dynamic with the objects in view changing rapidly from one moment to the next, with the size of the objects varying dramatically as they are brought closer to the head of the child and occlude others, and with there often being one visually dominant object taking up a large proportion of the child's head camera view. The unique dynamics of the child's view are to be found closely related to the child's own manual actions on objects. However, the parents were also actively engaged with objects during interaction, and with the toddler in joint play.

The question for the present study is whether the unique dynamic pattern of the toddler's visual experience during toy play is primarily the product of their own sensory-motor dynamics or primarily the product of their coupled dynamics with the mature partner. Our method starts with this observation: in everyday life, young children and their social partners engage each other in different ways and to varying degrees. On some occasions, the play is highly coordinated and both the child and the caregiver are engaged in the same task; other times, however, parents may be busy with housework or on the phone, and children are interacting on their own with objects in proximity to their parents making

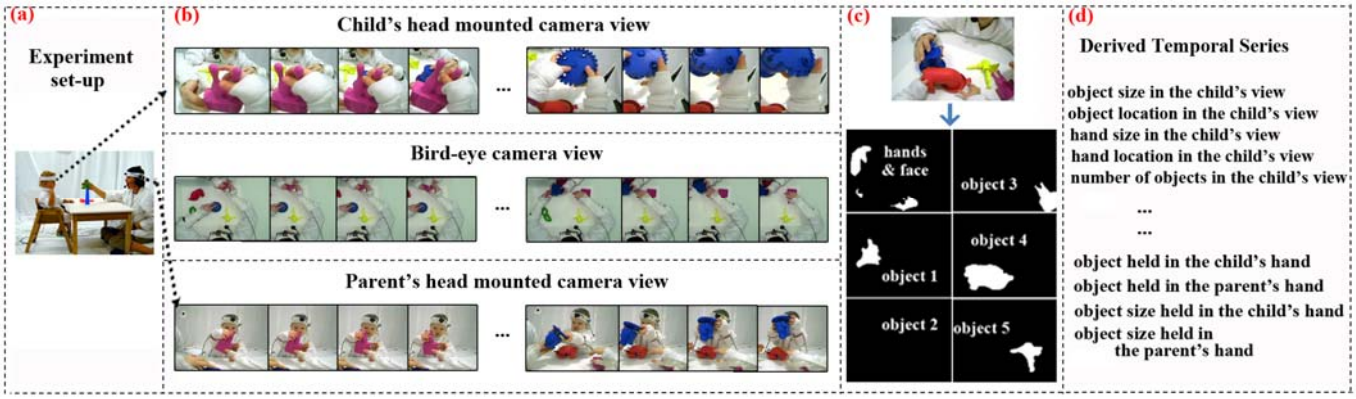


**Figure 1: (a) Multisensory sensing environment: two mini-cameras were placed on both the child's and the parent's foreheads to collect visual information from a first-person view. Another camera was mounted on the top of the table recording the bird-eye view of the whole interaction. (b) Example image sequences from child's head mounted camera, the parent's head mounted camera and the bird-eye camera view. (c) The automatic object detection and image segmentation process: the hand/face portion of the frame, object 1 to 5 were all automatically detected and segmented from the whole image. (d) Various temporal series extracted from data processing, such as object size in the child's view, object held in the child's hand and etc.**

social bids periodically to the parent. Here then, is the empirical question: how different is the dynamic structure of children's visual experiences when the child is playing alone versus when playing with a parent; in self-selected or directed interactions with objects? Is it the child's own sensory-motor dynamics that determine the structure of visual experience? Or this emerges in the social interaction of coordinated play? This is a critical question for understanding embodied attention and also for understanding the coordination of joint attention between parents and toddlers.

In the experiment presented in this paper, parents were instructed to interact with their children in these two different ways. We used a multisensory experimental environment to capture bodily actions of the participants and as well as their momentary visual information. We then compared and analyzed the child's visual experiences in these two social contexts.

## II. Multi-Camera Sensing Environment

As shown in Figure 1 (a, b), mini head-mounted cameras recorded the head-centered view of the two participants. A third camera was placed right above the table to record observation of visual information within the area of table.

**Environment setup**. Parents and children were seated across a 61cm × 91cm × 64cm plain white table, facing each other. A higher chair was provided for children and parents were asked to sit on the floor, in which case, children have the same eye level as the parent. Both participants were asked to wear white shirts (provided by experimenter). White curtains from ceiling to floor surrounded the table. Such room set up largely facilitates the visual object segmentation process since all white pixels from images captured by cameras can be treated as backgrounds with the exception of objects on the table, heads, faces and hands of the participants.

**Head-mounted cameras.** A lightweight mini camera attached to a sports headband was used to capture all visual information from the child's perspective. The headband was then placed on the forehead of participant, close to his or her eyes tightly enough that the camera did not move during experiments, also not too tight to cause any physical discomfort. The angles of both cameras were calibrated so that during each trial, when participants were attending to an object in particular, this object will appear near the center of image frames recorded by a head camera. Both head-mounted cameras have a visual field of 90 degrees, horizontally and vertically. Long and lightweight cables connected cameras to a wall socket, which did not restrict participants in any way during movements. A digital video recorder card in a computer near the experiment room recorded synchronized video streams from two cameras simultaneously.

**Bird's eye view camera**. A high-resolution camera was mounted right above the table and the table edges were aligned with image recorded from this bird-eye view camera. It provided an observation of the entire interaction during each trial from a third-person view, capturing static visual information that was independent from the head or manual actions of both the child and the parent. Also, due to size and weight, image resolutions from two head-mounted cameras were limited while this bird-eye view camera records video in higher quality, providing increased robustness in image processing and object segmentation process.

## III. Experiment Procedure

We used table-top toy play, a common everyday context, but provided task instructions that encourage parents to interact with their children in two different ways: In the *parent-child play* condition, parents were instructed to play naturally with their toddler with the toy, organizing the play as they wished; in the *play alone* condition, parents were asked to read a printed document, responding to the child – with look, nod, and word – when a social bid was made by the child, but not interacting with the toys through her own manual actions.

**Participants**. Twelve children and their parents from the

community of Bloomington, IN were invited to participate in the experiment (Six additional children were recruited, but either did not tolerate the head camera or were excluded because of fussiness before experiment started). For the child participants included, mean age was 21.56, ranging from 16.4 to 24.73 months, 8 females and 4 males.

**Stimuli**. There were two sets of toys that were made of rigid plastics with simple and novel shapes and one single main color. Each set consisting of five toy objects with five different colors (blue, green, red, pink and yellow) was randomly assigned to one of the two experimental conditions across participants.

**Procedure**. Upon entering the experiment room, the child was quickly seated in a high chair and several attractive toys were placed on the table. One experimenter played with the child while the second experimenter placed a sports headband with the mini-camera onto the forehead of the child at a moment that he appeared to be well distracted. After this, the second experimenter placed the second head-mounted camera onto the child's forehead close to eyes and calibrated the horizontal camera position in the forehead and the angle of the camera relative to the head. A third experimenter in the control room checked the centering of the head camera image during calibration to make sure that when the subject was looking at a well-centered object, the object also appeared in the center of the camera view. Then the parents head camera was placed and calibrated while the child played with one experimenter.

Each dyad participated in four interaction trials in total – two in the parent-child play condition and the other two in the play alone condition. In each trial, one of two toy sets was randomly selected to use. Both the order of trial types and the selection of toy sets were count-balanced across participants. In the two trials of *parent-child play* condition, parents were instructed to interact with their child as naturally as possible. At the beginning of each trial, after an experimenter provided a set of five objects on the table, parents started playing with them, and naturally engaging the child to play with these objects. There was no specific instruction on what they had to do or to say. At the end of a trial, after hearing a signal given by one experimenter, parent removed all the objects off the table to start the next trial. Each trial lasted for approximately 80 seconds. In the two trials of *play alone* condition, parents also sat across the table facing their child, but were asked to read a printed article provided by experimenters while letting the child play with toy objects by herself. During the interaction, when the child attempted to communicate by waving and showing a toy object toward their parent, the parent was asked to just briefly respond by looking toward the child and the object with verbal acknowledgement (e.g. "ok", "yeah"), but immediately switching their attention back to the article. Thus, children in this condition primarily played objects by themselves with the presence of their parents in the same environment.

With 12 dyads, 24 trials of parent-child play condition and 24 trials of play alone condition were completed and contributed to data processing and analysis in this study.


**Figure 2: (a) An example image frame from the child's head camera in the parent-child play condition. The size of all objects in this image is about 15% of the whole frame. (b) An example image frame from the child's head camera in the play alone condition. The size of all objects in this image is about 13%.**

## IV. Data Processing

For each pair of participants, three video sequences were collected from three separate views. With a recording rate of 30 frames per second for each camera, about 14,400 (30 × 80 × 3 × 2) image frames were collected for each pair of participants in each condition. The resolution of image frame is 720 × 480 pixels. Figure 1 (c, d) shows the procedure of image segmentation and analysis results. The technical details can be found in [6, 7]. Here we used the same image-processing method to automatically annotate thousands of image frames collected in this study. More specifically, visual information extracted from the collected image frames at each time stamp includes the number of objects, size of each object and location of objects in each camera view. In addition, from a bird's eye view camera, which object was held by either the child or parent can be manually coded by human coders. The present study focused on analyzing and comparing the visual contents from the child's head camera view in the two experimental social contexts and as well as how the child's and the parent's actions may shape the child's visual perception.

## V. Visual Information Analysis and Results

The head camera records the visual information from the *learner's point of view* via a lightweight mini camera mounted on a sports headband and placed low on the forehead. The angle of the camera is adjustable, and has a visual field of approximately $90^{o}$, horizontally and vertically. Yoshida and Smith [8] demonstrated the validity of this method in the tabletop context. They compared head camera view and eye gaze direction and found they were highly correlated. Their study showed that small shifts in eye-gaze direction unaccompanied by a head shift do not yield distinct tabletop views. Indeed, in their study, 90% of head camera video frames corresponded with independently coded eye positions. Therefore, although the human visual field is considerably larger (e.g. $180^{o}$) than the head camera field, and although, head and eye movements can be decoupled (an occasional glimpse without head movement), the restricted geometry and the typical behavior of young children in tabletop play means that the contents of the head camera field is a good

approximation of the contents of the child's visual field.

In the following, we first report the structure and content of visual information from the child's head camera view. In each trial, five objects were placed on the table, with about the same physical size and same distance from the child's head. All of the five objects would appear in the child's head camera view with approximately the same size if the child were to sit back, look straight to the table and take a broad view. However, if the child actively moved his body with head turns, gaze shifts, and posture changes, and if the child used his hands to manually hold, manipulate and play with those toys, then the objects in view would change moment by moment, and the dynamics in visual information captured by the head camera can be seen as a function of child's body movements, directly reflecting perceptual changes and visual selectivity in real-time.

### A. *Size and Number of Objects in View*

The size of an object in the head camera view directly reflects the closeness of the object to the child's head as well as it is non-occlusion by other objects and body parts. In the parent-child play condition, the objects comprised 15.02% of the head camera view on average, and 13.22% in the play alone condition. As shown in Figure 2, these objects took a large proportion of the child's view in both conditions. However, the average size of all objects in view in the parent-child play condition was significantly larger than in the child play alone condition ($t(23)=2.23$, $p<0.05$).

This may be because more objects are in view in the child-parent play condition than in the play alone condition (average number of objects in the child's view: $M_{parent\_child\_play} = 4.04$; $M_{play\_alone}=3.38$; $t(23)=4.93$, $p<0.001$). In the parent-child play condition, the proportion of time when there were five objects in the child's view (M=48.00%) was also significantly higher than the play alone condition (M=28.99%, $t(23)=4.36$, $P<0.001$). The proportion of time when there were only one or two objects in the child's view in the parent-child play condition was less than half of the play alone condition (7.15% vs. 15.86%). These results indicate that the parent have introduced more objects into the child's view. Our previous work suggests that the child's view relative to the adult view often reduces visual clutter, leading to one or a few objects in view at a time. This characteristic is *more marked* when the child is playing alone versus when playing with the parent, a result that suggests that this property may originate in the child's own sensory-motor dynamics. Put another way, the child's visual field was more cluttered when playing with the parent than when playing alone.

### B. *Visual Dominance of Objects*

Our previous work suggests that the child's view, relative to an adult view, is often characterized by a single dominant visual object. The size of an object in the head-camera view indicates its proximity to the viewer's head and its non-occlusion by other objects (since all of the objects have similar sizes). Thus, the size of an object provides a measure

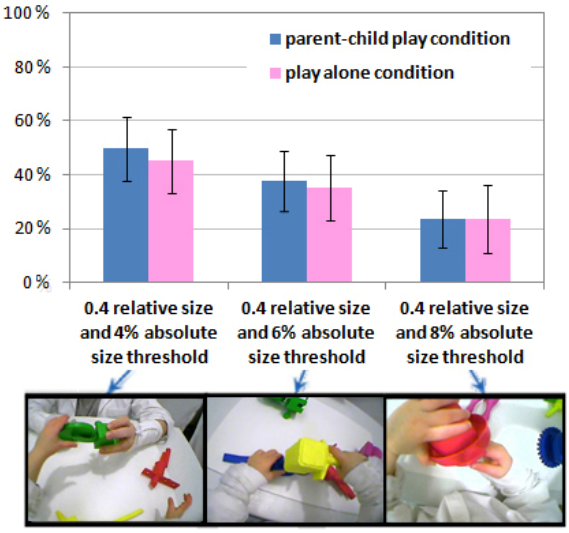


**Figure 3: Proportion of time that there was an object dominating the child's view. A dominating object is defined based on both the absolute size of an object and its relative size compared to other objects appeared in child's view. Three sets of thresholds selected here for defining dominant moments are: 1) 0.4 relative size and 4% absolute size, 2) 0.4 relative size and 6% absolute size, 3) 0.4 relative size and 8% absolute size.**

of selective attention to visual dominance of an object: the largest object in the head camera field - the closest and non-occluded one - is most likely the focus of the child's attention. In order to examine the dominant moments wherein a single object was dominating the child's visual field in the present study, we used two criteria: 1) *absolute size*: the percentage of an image frame that was occupied by an object in this frame; 2) *relative size*: the ratio of the largest object's size to the total size of all objects in view. We defined three types of dominant moments by choosing different parameters in both absolute size and relative size to obtain a more complete picture of visual dominance. First, given that the average absolute size of an object in the child's head camera view was approximately 3.5% in both conditions, we selected three thresholds of absolute size, 4% (just above the average size), 6% and 8% (twice the average size). Then, given that there were five objects in total and the baseline for relative threshold was 0.2 (1/5), 0.4 relative threshold was chosen (the dominant object took up more than 40% of the size of all objects in view). Taken together, three dominant types were defined in an increasingly strict dominance order: 1) 0.4 relative size and 4% absolute size, 2) 0.4 relative size and 6% absolute size, 3) 0.4 relative size and 8% absolute size. Figure 4 shows the proportions of time for three dominant types in two experimental conditions. With 4% absolute dominance, about 50% of time there was an object dominating the child's perceptual field in both conditions. Even with the most conservative measure (0.4/8%), there was still about 25% of time when a single large object dominated the child's head camera view. This result is consistent with our previous findings [6, 7] using the head camera technique: the child's first-person view was highly selective and dynamic.

However, with all three sets of thresholds, there were no significant differences between two conditions (0.4/4%: t(23)=1.40, p=0.17; 0.4/6%: t(23)=0.60, p=0.56; 0.4/8%: t(23)=0.13, p=0.89), which is really surprising. This result suggests that the parent's involvement didn't cause a significant change in the moments when one object dominates the view, a perhaps optimal time for learning about objects and their structure. In brief, with respect to this measure, the child's visual experiences in the two different conditions have approximately the same level of selectivity.

## VI. Hand Actions

### A. Objects Held by the Child and the Parent

The child's view is selective in that it is often dominated by a single visual object and this appears to be a fundamental characteristic of the child's embodied attention: attention is achieved by a single object close to the head and sensors. However, there are multiple ways that this selection might occur: the child could reach out and bring an object close or, for example, the parent could select and show the child an object, honoring the child's embodied approach by bring that object close to the child's head and thus making the object large and potentially dominating in the child's view. The children participated in this study were virtually always holding an object: in the parent-child play condition, the child was holding at least one object in 76.3% of the frames and 74.1% of the frames in the play alone condition. Also when engaged, parents also held objects: in the parent-child play condition, the parent was holding at least one object in 63.3% of the frames (the parents were instructed to read a printed document in the play alone condition without touching any objects). In addition, the child switched the objects in hand 5.5 times per minute in the parent-child play condition and 5.8 times per minutes in the play alone condition; the objects held by the parent's hands switched from one object to another 8.0 times per minute in the parent-child play condition. These facts showed the active manual engagement of both the child and the parent in two conditions.

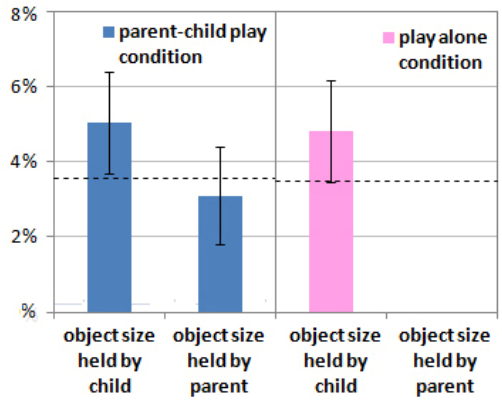


**Figure 4: Average size of objects held by child and parent in two conditions. The size of an object was measured by the percentage of the head camera's view that was occupied by the object. Dash line in each condition indicates the average size of all objects appeared in all frames.**

We calculated the average size of objects in the child's view when the object was held by the child and of objects held by the parent. In both conditions, the average size of objects held by the child was reliably larger than the average size of all objects appeared in the image (see Figure 4): in the parent-child play condition, on average, the objects held by the child took 4.8% of the child's visual field, and the average size of all objects in the image was 3.6%; in the child play alone condition, the average size of objects in the child's hand was 4.9% compared to the average size of all objects which was 3.5% ($F(2,69)=14.85$, $P<0.001$). However, the size of objects held by the parent did not differ much from the baseline size of objects in the parent-child play condition. Large image size in both conditions is associated with holding by the child, not by the parent. Perhaps, parents' role is to influence visual dominance by directing their child's interest to an object, triggering a behavioral cascade on the part of the child that leads to visual dominance.

### B. Manual Actions that Lead up to Object Dominance

There are five possible interactive behavioral patterns that can change the object dominance in child's view: 1) the child selected an object and brought it closer to the head; 2) the parent selected an object and put it closer to the child's head; 3) the parent took the object in the child's hand, then moved it closer to the child's head; 4) the child took the object which was initially in the parent's hand, and brought it closer to the head; 5) the child leaned forward and moved his head closer to the target object, decreasing the distance between the head and the object. To better understand the dynamic processes that led up to the emergence of object dominance in the child's view, we zoomed into the moments right before an object became dominant to investigate the events that might lead to visual selection.

Here we selected 0.4 relative size and 8% absolute size thresholds as measurements for determining whether an object has became dominant – the most conservative measure of visual dominance. In addition, a temporal stability criterion was added to select dominant moments: the dominating object must hold its dominance for at least 500 ms before the child switched to other objects. This is used to exclude the transient dominant moments wherein an object happened to appear and dominate the child's visual field for an extremely short period of time (probably caused by dramatic head turns from the child, etc.). Thus, the results presented in this section were based on stabilized dominant events wherein the dominant object not only took a large proportion of the child's view but also maintained its visual dominance for a sufficient amount of time (this temporal stability criterion was also used in analyzing manual actions from the child and parent). Given the criteria, there were 268 dominant events in total in the parent-child play condition (11.2 events per trial); and 241 dominant events in the play-alone condition (10.04 events per trial).

Next, we measured the manual actions from both the child and the parent from 5 seconds prior to the onset of a dominant

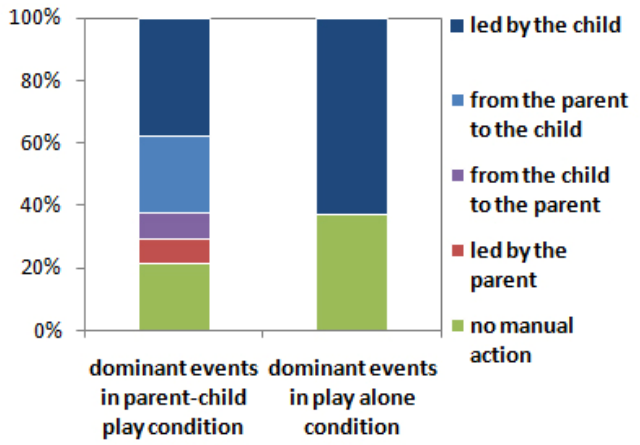


**Figure 5: The proportion of dominant events that belong to different categories. Dominant events are selected with 0.4 relative size and 8% absolute size threshold as measurements.**

event, and investigated the following-leading patterns between their hand actions and object dominance in the child's view. Accordingly, the dominance events can be further defined into five categories: 1) *led by the child*: the object was held by the child when it became dominant, no parent's hand actions were involved right before and during the emergence of this dominant event; 2) *led by the parent*: the parent held the object, put it closer to the child's head and caused this object to become dominant; 3) *from the child to the parent*: the child was holding an object in his or her hand, the parent took it and brought it closer to the child's head and the object became dominant; 4) *from the parent to the child*: the object was initially held by the parent, then the child was attracted to this object, took it from the parent and brought it closer to his or her head; 5) *no manual action*: the object was not in hand when it became dominant.

Figure 5 shows the grouping results. In most cases, the object was in the child's hands when it became dominant (both from *led by the child* category and from *the parent to the child* category). Only in less than 20% cases was the to-be-dominant object in the parent's hands (dominant events from *led by the parent* category and *from the child to the parent* category). However, among all the instances that the child was holding the object right before its dominance, there was a large proportion of events wherein the object was initially held by the parent's hand, then was passed on to the child's hand and within 5 seconds afterward, that object became dominant in the child's view. This is our first indication of a role for parents in potentially selecting the object attended to by the child.

There are two implications from this result. First, in the parent-child play condition, the parent was influencing the objects dominating the child's view through the child's own hand actions: from time to time, the parent held an object, successfully attracted the child's visual attention. then passed the to-be-dominant object to the child and thereafter the child brought this object closer to his head. Second, in the child play alone condition without the parent's manual involvement and without the increase of the child's own manual actions, the child managed to create similar visual dominating experiences which can be due to the child's increased head movements and/or the fact that the parent's hands are in the child's view to block visual objects on the table.

## VII. Conclusion

The sensory-motor dynamics that lead to the signature structure of toddler visual experience – rapidly changing views focused on a single dominant object at a time – appear to the be the child's own dynamics. Parent behaviors may perturb these dynamics only slightly. However, parents are able to embed objects within those dynamics, playing a role in the objects that are selected and attended to by the child. These processes of visual selection are important because the ordinary contexts in the child's real learning environment are often highly cluttered, with multiple visual targets in the physical world and complicated real-time social cues from their caregivers for attention and for leaning. Optimizing learning and joint attention with a toddler may depend not on perturbing the child's sensory-motor dynamics in the service of object play but on perturbing the objects selected - not the manner of attention but the content.

The results raise new questions, requiring continued and deeper analyses, such as what is the difference between the child's internal sensory-motor processes when the child was focusing on the object that was initially chosen by themselves and when the child was holding on to the object selected by their caregivers, and whether and how those different paths will lead to visual dominance that is longer and perhaps more conducive to learning. The present paradigm provides a promising platform to further investigate those questions.

## References

[1] E. J. Gibson, *Principles of perceptual learning and development*. New York: Appleton-Century-Croft, 1969.

[2] H. A. Ruff, "The Development of Perception and Recognition of Objects," *Child Development*, vol. 51, pp. 981-992, 1980.

[3] B. I. Bertenthal and J. J. Campos, "New Directions in the Study of Early Experience," *Child Development*, vol. 58, pp. 560-567, 1987.

[4] D. H. Ballard, "Reference frames for animate vision," presented at the Proceedings of the 11th international joint conference on Artificial intelligence, Detroit, Michigan, 1989.

[5] G. Metta and P. Fitzpatrick, "Better vision through manipulation," *Adaptive Behavior*, vol. 11, pp. 109-128, 2003.

[6] C. Yu, L. B. Smith, H. Shen, A. F. Pereira, and T. G. Smith, "Active Information Selection: Visual Attention Through the Hands," *IEEE Transactions on Autonomous Mental Development*, vol. 2, pp. 141-151, 2009.

[7] L. B. Smith, C. Yu, and A. F. Pereira, "Not your mother's view: the dynamics of toddler visual experience," *Developmental Science*, vol. 14, pp. 9-17, 2010.

[8] H. Yoshida and L. B. Smith, " What's in View for Toddlers? Using a Head Camera to Study Visual Experience," *Infancy*, vol. 13, pp. 229-248, 2008.