Evaluating Human Cognition of Containing Relations with Physical Simulation

Wei Liang[1,2] **(liangwei@bit.edu.cn)** **Yibiao Zhao**[2] **(ybzhao@ucla.edu)**
Yixin Zhu[2] **(yixin.zhu@ucla.edu)** **Song-Chun Zhu**[2] **(sczhu@stat.ucla.edu)**
[1]School of Computer Science, Beijing Institute of Technology (BIT), Beijing, China
[2]Center for Vision, Cognition, Learning, and Art (VCLA), University of California, Los Angeles (UCLA), 90095, USA

Abstract

Containers are ubiquitous in daily life. By container, we consider any physical object that can contain other objects, such as bowls, bottles, baskets, trash cans, refrigerators, etc. In this paper, we are interested in following questions: What is a container? Will an object contain another object? How many objects will a container hold? We study those problems by evaluating human cognition of containers and containing relations with physical simulation. In the experiments, we analyze human judgments with respect to results of physical simulation under different scenarios. We conclude that the physical simulation is a good approximation to the human cognition of container and containing relations.

Keywords: Container; Simulation; Physical reasoning

Introduction

Containers are ubiquitous objects in daily life, such as bowls, bottles, baskets, trash cans, refrigerators, etc. Containing relation is a general and fundamental relation in the scene. Containers offer containing relations for carrying, hiding, or ensuring the objects remain in a safe place. The contained objects are called contents. The containing relation characterizes the "affordance" that how likely a container can hold its content.

Different from visual object recognition problems, recognition of containers involves the cognitive process of commonsense reasoning, such as analysis of physical properties, geometric shapes, and material properties, etc. Fig.1 shows two examples when a container fails to contain its content: (a) the container with holes can not contain tiny objects or staffs, like beads, sand or water; (b) the container with a low wall fails to contain a big ball.

Containers quantize and organize our perceptual scene space. For example, when people are asked "where the chilled beer is", the answer will usually be that "it is in the refrigerator" without mentioning the exact 3D coordinates. By containers, the perceptual space of 3D scene is discretized and quantized, and objects are often organized in a hierarchy with respect to their containing relations (Zhao & Zhu, 2013). This quantization largely simplifies many tasks, such as planning, detection and tracking.

Inspired by (Battaglia, Hamrick, & Tenenbaum, 2013) and (Zheng, Zhao, Yu, Ikeuchi, & Zhu, 2015), human perceive physical scenes by making approximate and probabilistic inference, and the physical engine helps us to reason about common-sense in complex scenes. When we ask about whether a container will hold another object, human may do similar mental simulations. The definition of containers are related to physical properties of containers and contents. In Fig.1, the container and its contents are not compatible in these two cases. In this paper, we model and infer the containing relations between two objects by imagining what would happen when one puts an object into a container.

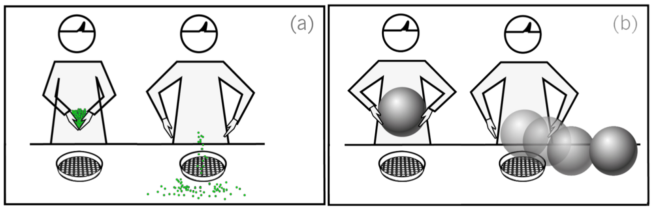

Figure 1: Two typical cases when a container fails to contain its contents: (a) the container with holes can not contain tiny objects; (b) the container with a low wall fails to contain a big ball. The left figures of these two panels illustrate a stimuli of our experiments, and the right figures illustrate simulation results with physical engine or in human mind.

In order to study containers and the factors which affect containing relations, we collected a 3D container dataset and carry out our experiments based on it. In the experiment, we presented some random sampled 3D objects from our dataset to the subjects. The subjects answered questions about container and containing relations according to these pictures. We also built an online physical simulation system with Unity 3D engine on a tablet platform as shown in Fig.2. The system is used for evaluating containing relations between objects and comparing with human judgments.

Related work

Containers in Cognitive Science. Some experiments (Hespos & Baillargeon, 2001; Hespos & Spelke, 2007) showed that even young infants can understand containers. In their first six months of life, infants knew that contents can be occluded by containers. At the end of their first year, infants can develop a more refined concept about container and containment. (Inhelder & Piaget, 1958) studied children's understanding of the conservation and limited capacity of liquid, matters and numbers. For example, six-year children may confuse about what happened when the liquid in a tall skinny container was poured into a short wide container.

Simulation. According to the Simulation Theory (ST), an attributor arrives at a mental attribution by simulating, replicating, or reproducing in his own mind the same state as the

This work is supported by ONR MURI N00014-10-1-0933

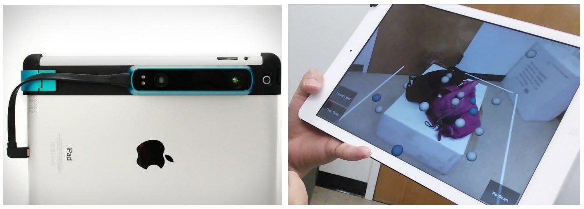

Figure 2: A 3D Structure Sensor attached to a tablet (left) and a physical simulation interface (right) are used in this paper. The interface simulates a few balls falling onto a bag.

target, or by attempting to do so. (Markman, Klein, & Suhr, 2012) reviewed the research of simulation-based models in psychology. Some works (Goldman & Sripada, 2005) examined the simulation approach and the theorizing approach for determining the compatibility between emotions and existing evidence. Some neuroscience research is quite related to simulationist ideas (Chaminade, Meary, Orliaguet, & Decety, 2001; Gallese & Goldman, 1998; Jeannerod, 2001). (Hamrick, Battaglia, & Tenenbaum, 2011; Battaglia et al., 2013) studied the intuitive physics engine as a model to reason stability of a tower built by blocks. They showed the simulation model matched human perceptions. Benefiting from game engines, such as PhysX, Bullet and Unity3D, physical simulation is widely available for game designers as an off-the-shelf component (Kaufmann & Meyer, 2008).

In AI community, container has been studied since 1980s as a wide-accepted example for qualitative reasoning (Williams, Hollan, & Stevens, 1983; Bredeweg & Forbus, 2003; Frank, 1996). In (Collins & Forbus, 1987), container is used to reason liquid. They presented a technique called molecular collection ontology to describe contained stuff. A preliminary knowledge base for qualitative reasoning about containers is developed in (Davis, Marcus, & Chen, 2013), which is expressed in a sorted first-order language of time, geometry, objects, histories, and events. Those studies modeled containers by using logic with a restriction of well-defined task domains, and the observation is not directly obtained from real world signal.

Experiments

3D container dataset

In the experiment, we built a 3D container dataset including 315 real-world 3D objects. The data was collected using a 3D Structure Sensor attached to a tablet platform as shown in Fig.2 (a). The objects in our dataset are full 3D models reconstructed by computer vision algorithms. We then conduct our experiment with these real-world 3D objects. Some results are shown in Fig.3. Comparing with previous cognitive studies, our experiments use daily objects in natural physical scenes.

Participants and stimuli

We conduct human studies with fifty human subjects who are university students around age 25. We are interested in three main questions: i) What is a container? ii) Can an object *A* contain another object *B*? iii) How many objects will a container hold? For each of these three questions, we show a 3D scene as a stimulus to human subjects, and ask them to answer a corresponding question. The objects in the 3D scenes are generated randomly from our 3D container dataset.

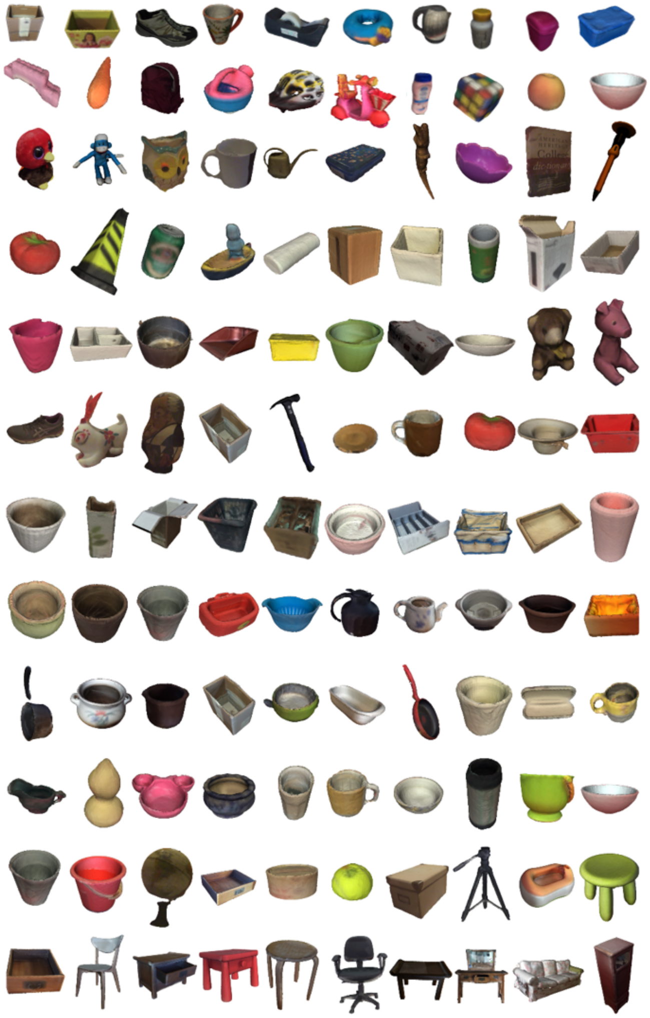

Figure 3: 3D scanned objects in our container dataset

Physical simulation

We set up a physical simulation system with Unity 3D engine to infer the probability for an object to be a container and containing relations between two objects. We place a 3D object as a potential container on a virtual ground, and initialize another object as its potential content over the container with a few random parameters, i.e. relative height, position, pose, and initial speed. Initializing the 3D scene by randomly sampling these parameters, we calculate the frequency of successful cases of containments through physical simulations. In the physics engine, we model the potential container by a ”Mesh Collider” which calculates the collisions for all the triangle faces (around 17000) on the object. And we simplify

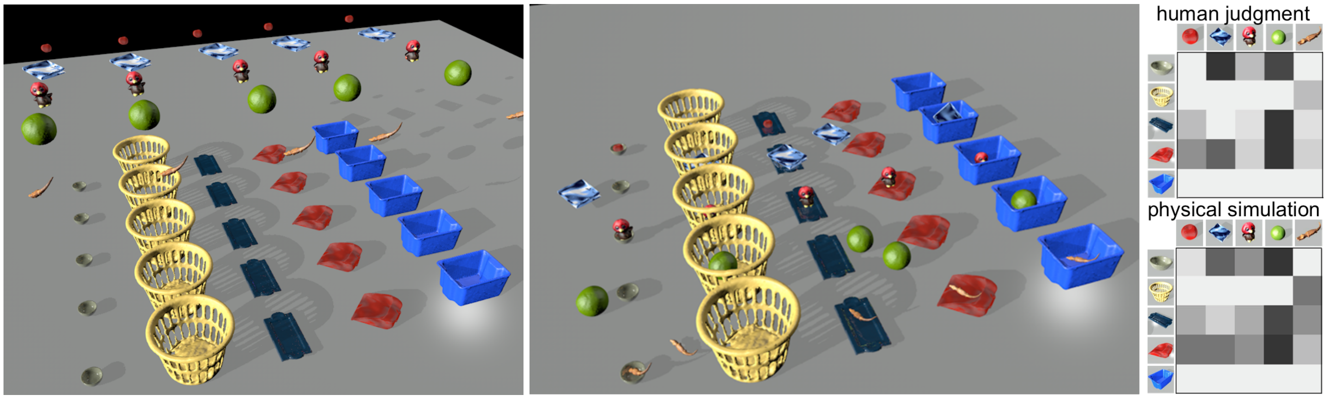

Figure 4: Inferring containing relations by physical simulations. The left figure shows the initial status of 5 containers and 5 contents, and the right figure shows the result of the physical simulation. We simulate the scene 100 times with random relative heights, positions, poses, and initial speeds of objects, and calculate the frequency of successful contained trials. The result confusion matrixes of containing relations are shown on the right, where each cell represents the probability of one object containing another. The human judgement is the average score of ten subjects, and the physical simulation is calculated by the frequency of successful containments with different trials. The lighter the color of a cell, the higher the probability is.

the 3D model of potential content to 255 triangle faces, and approximate its physical dynamics by a "Convex Collider" for the consideration of computation feasibility.

Exp. 1: What is a container?

In this experiment, we let subjects see a 3D object and ask following questions: i) is it a container? ii) is it a convex shape? iii) does it have a hole? iv) does it have a lid? v) is it hollow? vi) is it deformable? vii) what kind of material is it?

The figure on the left of Fig.5 shows the distribution of six attributes associated to these questions. For each attribute, we plot distributions for both container and non-container. For example, most of the containers are concave shapes, and most of the non-containers are convex. The last material attribute takes categorical values of "metal", "paper-based", "fabric", "wood", "glass", and "plastic".

The distribution of object sizes of the dataset are also showed on the right of Fig.5. The size of the object covers from the hand size (a few centimeters) to the body size (a few meters). The size distributions of containers (green dots) and non-containers (red dots) in the dataset are very similar.

Logistic regression analysis for attributes

We analyze the contribution of different attributes to the notion of "container" by logistic regression. We use five binary variables: (convex, has hole, has lid, hollow, deformable), one categorical variable (material), and two continuous variables (height and base area) as predictors. The algorithm aims to analyze the influence of different variables for answering the target question "is it a container or not?".

The results of the regression are shown in Table.1.The attributes convex and hollow with low p-values are statistically significant for discriminating the concept of containers.

Container recognition

We address the containers recognition problem as a computer vision problem. We compare two algorithms: 1) classic computer vision algorithm by pattern-recognition, 2) physical simulation-based method as introduced before.

Table 1: Analysis of logistic regression coefficients.

	Estimate	Std. Err	tStat	pValue
(Intercept)	-3.1168	1.1114	-2.8043	0.005043
convex	-1.8572	0.2692	-6.8999	**5.204e-12**
has hole	0.1248	0.3814	0.3274	0.7434
has lid	1.4893	0.4086	3.6449	0.0002675
hollow	2.2661	0.2736	8.2818	**1.2132e-16**
deformable	-0.7816	0.3067	-2.5485	0.01082
material	0.1712	0.0754	2.2714	0.02312
height	-0.8198	0.5969	-1.3733	0.1697
base area	0.4308	0.2580	1.6702	0.09489

We used a state-of-the-art discriminative classifier based on Hierarchical Kernel Descriptors (Bo, Lai, Ren, & Fox, 2011). In order to apply the classic computer vision method, we project the 3D model to RGB images and depth images from canonical views. And we use the RGB images and RGBD images for training and testing the computer vision algorithm. For comparison, we also test the simulation-based method on the same testing set of 3D objects. The probability is calculated by the expected value for containing another objects in the dataset.

In order to evaluate the generalization ability of these algorithms, we test them on three different scenarios:

i) The single category: both training and testing samples come from the same single category, such as boxes. ii) The mixed category: both training and testing samples come from a collection of multiple categories. iii) The transfer category: the training samples come from one category, such as boxes, while the testing samples come from another category, such as cups. The results are summarized in Table.2. It is worth

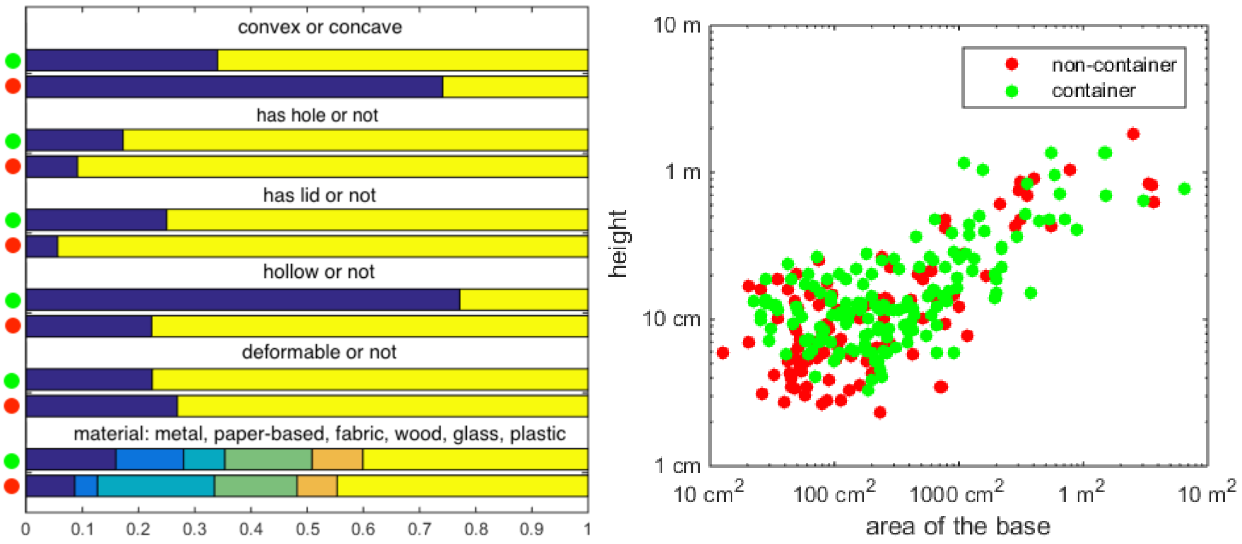

Figure 5: The distribution of different container attributes. In the left bar plot, a pair of horizontal bars represents the distribution of containers and non-containers for each discrete attribute; in the right scatter plot, the green and red dots illustrate the distribution of containers and non-containers with respect to the area of the base and height of these 3D objects.

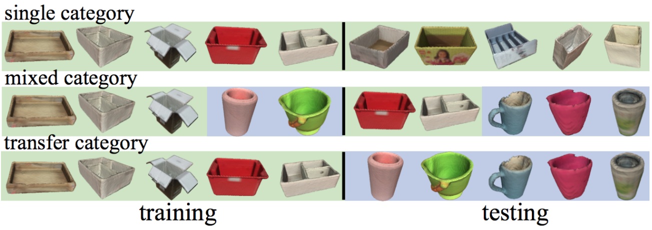

Figure 6: The split of training / testing for container recognition. i) The single category: both training and testing samples come from a same single category. ii) The mixed category: both training and testing samples come from a collection of multiple categories. iii) The transfer category: the training samples come from one category, while the testing samples come from another category. The results show in Table.2.

noting that the (Bo et al., 2011)'s algorithm works well on single category. Because the simulation-based algorithm does not need any training, and the physical laws are generally applicable, physical simulation-based algorithm has advantages for generalizing across categories.

Exp. 2: Will an object contain another?

In the experiment, we evaluate the "affordance" of a container. Human subjects are shown a 3D scene with two 3D objects randomly sampled from the dataset. One is a potential container, another is a potential content. Some of stimuli are shown in Fig.9.

We applied two kinds of approaches to model the containing relations between two objects. i) Regression model. We use features including relative height ratio, base area ratio, and volume ratio, to learn a logistic regression model. ii) Physical simulation model. We compare the results of both models with respect to human judgments in Fig.7 (a,b). And we also show the correlations between two human subjects on the right of Fig.7. We can see that this task is very challenging, as there are diverse judgments even between human subjects. Although the regression method can capture some correlation between the relative size and the containing relation, the results of simulation model show much strong collinearity with the human subject. The area between two blue lines are the variance interval between 25% percentile and 75% percentile, which means a half of the samples will fall into the region between two blue lines. Each point in the graph is a stimulus in the Fig.9. We can not handle the last two challenging cases in current framework. Both containers acquire human intervention to open containers and put in objects, which can not be modeled solely by the rigid-body dynamics.

Table 2: Accuracy of container recognition

	RGB	RGB-Depth	Simulation
single category	0.89	**0.94**	0.93
mixed categories	0.70	0.78	**0.93**
transfer category	0.35	0.59	**0.93**

Exp. 3: How many objects will a container hold?

In this experiment, the stimuli are the same as Exp.2's. The subjects are shown two random 3D objects and ask "how many objects will a container hold?" The qualitative results and quantitative results are shown in Fig.8 and Fig.10. Similarly, the simulation results are more consistent with human judgments than the regression model. Although the results exhibit a large variation, similar variations are also existed among judgments from different subjects.

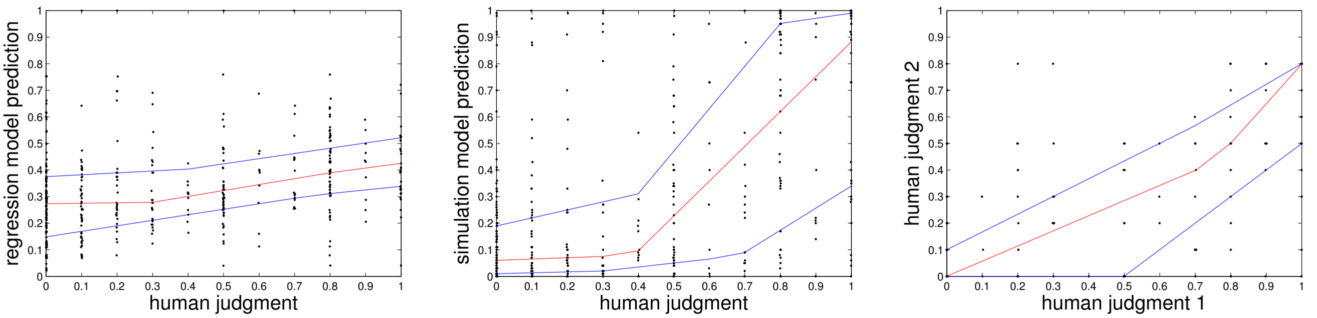

Figure 7: Will an object contain another? The left and middle figures show predictions of the regression model and the simulation model with respect to the human judgments. The right figure shows the human judgments of two different subjects. Each data point represents a stimulus with a pair of objects in Fig.9. The lower blue line, red line, and upper blue line outline the first quartile (25th percentile), second quartile (median), and third quartile (75th percentile) of the distribution respectively.

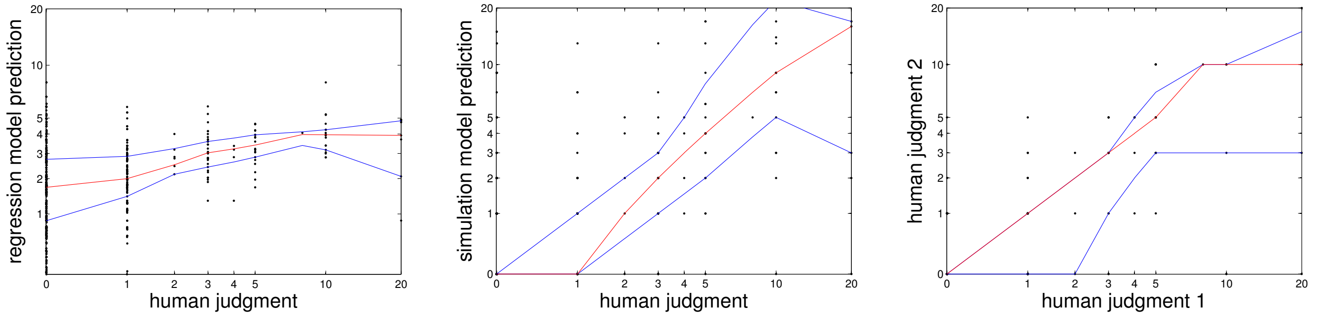

Figure 8: How many objects will a container hold? The left and middle figures show predictions of the regression model and the simulation model with respect to the human judgments. The right figure shows the human judgments of two different subjects. Each data point represents a stimulus with a pair of objects in Fig.10. The lower blue line, red line, and upper blue line outline the first quartile (25th percentile), second quartile (median), and third quartile (75th percentile) respectively.

Conclusions

In this paper, we study a special category of objects "container". We collected a dataset of 315 real-world 3D models including containers and other daily objects. We built a physical simulation system using Unity 3D to infer the "affordance" of containers and containing relations between objects. In the experiment, compared with using regression model of geometric features, the results by physical simulation have stronger correlations with human judgments. We conclude that the physical simulation is a good approximation of human cognition of container and containing relations.

The physical model of the 3D scene quantitatively encodes a large number of static and dynamic variables needed to capture the interactions among objects. These variables include scene configurations, object geometries, masses, material properties, rigidity, fragileness, frictions, collisions, etc. We take advantages of the state-of-the-art 3D scanning technique, which enables us to analyze real-world 3D objects in a physical realistic environment. Although the rigid body dynamics can not exactly follows the real-world motions and parameters, the results are sufficient appealing and promising as a start point for understanding containers.

References

Battaglia, P., Hamrick, J., & Tenenbaum, J. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327-18332.

Bo, L., Lai, K., Ren, X., & Fox, D. (2011). Object recognition with hierarchical kernel descriptors. In *Computer vision and pattern recognition (CVPR)* (pp. 1729–1736).

Bredeweg, B., & Forbus, K. D. (2003). Qualitative modeling in education. *AI magazine*, *24*(4), 35.

Chaminade, T., Meary, D., Orliaguet, J.-P., & Decety, J. (2001). Is perceptual anticipation a motor simulation? a pet study. *NeuroReport*, *12*(17), 3669–3674.

Collins, J. W., & Forbus, K. D. (1987). Reasoning about fluids via molecular collections. In *Aaai* (pp. 590–594).

Davis, E., Marcus, G., & Chen, A. (2013). Reasoning from radically incomplete information: The case of containers. *Advances in Cognitive Systems*, 273–288.

Frank, A. U. (1996). Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science*, *10*(3), 269–290.

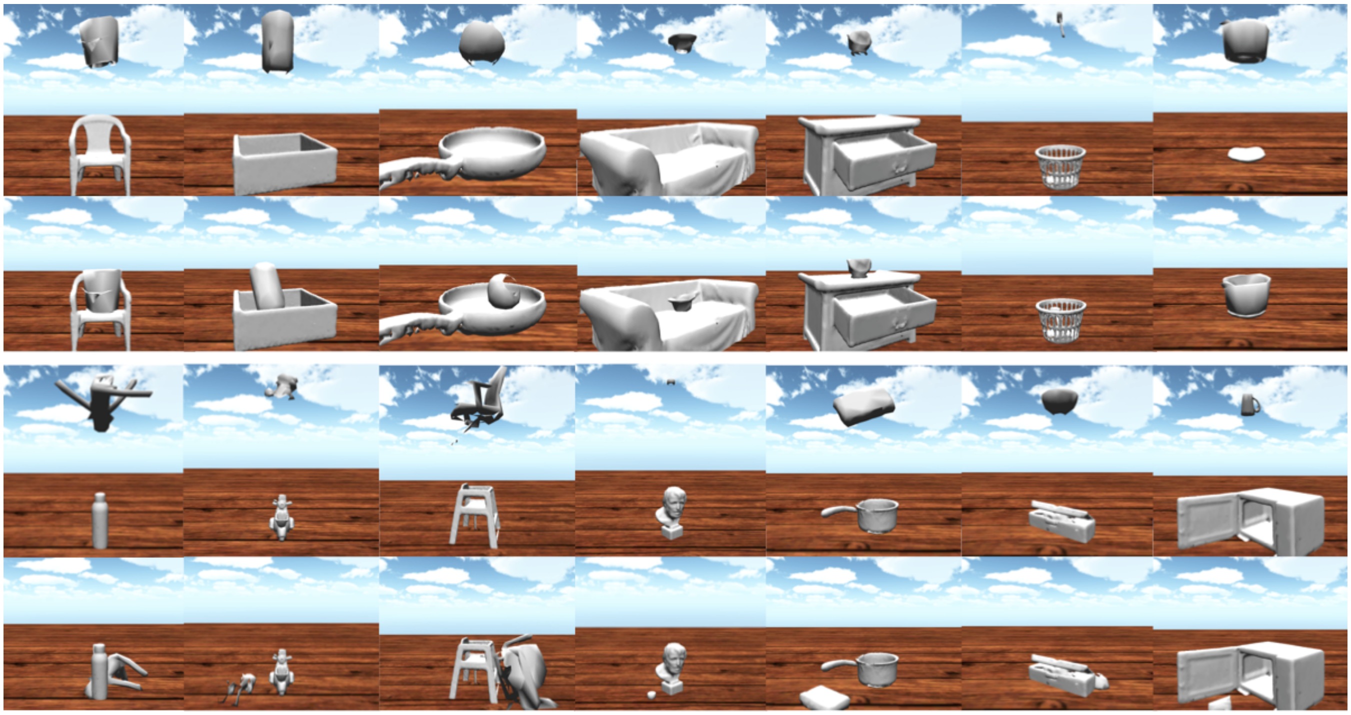

Figure 9: More qualitative results for the Exp.2 ”Will an object contain another?”. The first row and third row are screenshots before simulation as stimuli of our experiment. The second row and fourth row show successful containing cases and failed containing cases after physical simulation respectively.

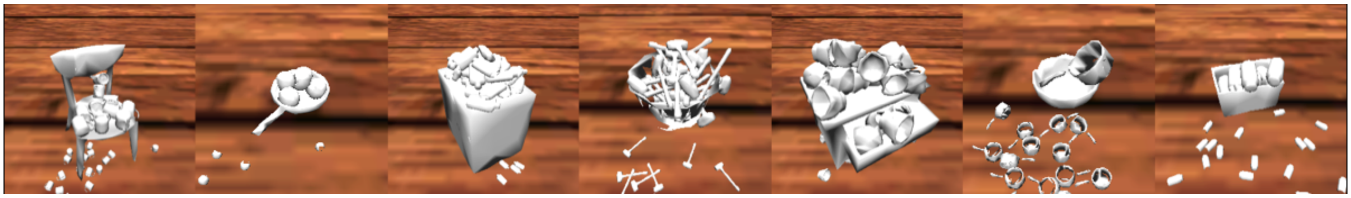

Figure 10: Qualitative results after physical simulation for the Exp.3 ”How many objects will a container hold?”.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493–501.

Goldman, A. I., & Sripada, C. S. (2005). Simulationist models of face-based emotion recognition. *Cognition*, *94*(3), 193–213.

Hamrick, J., Battaglia, P., & Tenenbaum, J. B. (2011). Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1545–1550).

Hespos, S. J., & Baillargeon, R. (2001). Infants' knowledge about occlusion and containment events: A surprising discrepancy. *Psychological Science*, *12*(2), 141–147.

Hespos, S. J., & Spelke, E. (2007). Precursors to spatial language: The case of containment. *The categorization of spatial entities in language and cognition*, 233–245.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (Vol. 22). Psychology Press.

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage*, *14*(1), S103–S109.

Kaufmann, H., & Meyer, B. (2008). *Simulating educational physical experiments in augmented reality*. ACM.

Markman, K., Klein, W., & Suhr, J. (2012). *Handbook of imagination and mental simulation*. Psychology Press.

Williams, M. D., Hollan, J. D., & Stevens, A. L. (1983). Human reasoning about a simple physical system. *Mental models*, 131–154.

Zhao, Y., & Zhu, S.-C. (2013). Scene parsing by integrating function, geometry and appearance models. In *Computer vision and pattern recognition (CVPR).*

Zheng, B., Zhao, Y., Yu, J. C., Ikeuchi, K., & Zhu, S.-C. (2015). Scene understanding by reasoning stability and safety. *IJCV*. doi: 10.1007/s11263-014-0795-4