A Benchmark for Modeling Violation-of-Expectation in Physical Reasoning Across Event Categories

Arijit Dasgupta
National University of Singapore

Jiafei Duan
A*STAR, Singapore

Marcelo H. Ang Jr
National University of Singapore

Yi Lin
New York University

Su-hua Wang
University of California Santa Cruz

Renée Baillargeon
University of Illinois Urbana-Champaign

Cheston Tan
A*STAR, Singapore

https://arijitnoobstar.github.io/sites/voe.html

Abstract

Recent work in computer vision and cognitive reasoning has given rise to an increasing adoption of the Violation-of-Expectation (VoE) paradigm in synthetic datasets. Inspired by infant psychology, researchers are now evaluating a model's ability to label scenes as either expected or surprising with knowledge of only expected scenes. However, existing VoE-based 3D datasets in physical reasoning provide mainly vision data with little to no heuristics or inductive biases. Cognitive models of physical reasoning reveal infants create high-level abstract representations of objects and interactions. Capitalizing on this knowledge, we established a benchmark to study physical reasoning by curating a novel large-scale synthetic 3D VoE dataset armed with ground-truth heuristic labels of causally relevant features and rules. To validate our dataset in five event categories of physical reasoning, we benchmarked and analyzed human performance. We also proposed the Object File Physical Reasoning Network (OFPR-Net) which exploits the dataset's novel heuristics to outperform our baseline and ablation models. The OFPR-Net is also flexible in learning an alternate physical reality, showcasing its ability to learn universal causal relationships in physical reasoning to create systems with better interpretability.

1 Introduction

Physical-reasoning systems built on the foundations of intuitive physics and psychology are pivotal to creating machines that learn and think like humans [1, 2]. The ability to reason about physical events like humans opens a crucial gateway to multiple real-world applications from robotic assistants, autonomous vehicles, and safe AI tools. These systems are guided by facets of core knowledge [3]. Infant psychologists theorized that human newborns have an in-built physical-reasoning [4, 5] and object representation [6, 7, 8] system. This has inspired researchers to approach the creation of physical-reasoning systems as a start-up software embedded with core knowledge principles [9].

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

Predicting the future of a physical interaction given a scene prior has been the most common task in designing computational physical-reasoning systems. One such task is the tower of falling objects, which has been as a test-bed for evaluating intuitive physics engines [10, 11, 12]. There have also been recent advancements in creating benchmarked datasets and models that combine multiple physical prediction tasks in a 3D environment [13, 14, 15]. Physical reasoning has also been explored in Embodied AI [16] and Visual Question Answering (VQA) [17] with public datasets like **CLEVRER** [18], **CRAFT** [19] and **TIWIQ** [20] providing scenes of general and random interactions between objects and multiple questions concerning the physical outcome.

A parallel track complementary to future prediction is the design of artificial agents that can measure the plausibility of physical scenes. An agent capable of physical reasoning should not only be able to predict the future but also use it to recognize if a scene is *possible* or *impossible*. The Violation-of-Expectation (VoE) paradigm is an empirical diagnostic tool first implemented in infant psychology studies [21, 22] to measure the surprise of infants when shown *possible* or *impossible* scenes. The studies found that infants as young as 2.5 months could express surprise at a constructed scene that violated the principle of object permanence. This was akin to a magic show. VoE has since been used in an array of infant psychology experiments on a range of event categories in physical reasoning [23, 24, 25, 26, 27].

The work in VoE has encouraged recent computational development of models and datasets [28, 29, 30, 31, 32] that challenge artificial agents to independently label *possible* and *impossible* scenes in physical events, goal preferences [33] and more. While these datasets mimic real-world VoE experiments, they provide mainly vision data with little to no heuristics that aid learning. We believe that computational benchmarks in VoE require the embedding of more inductive biases from the body of psychology work that inspired them. A common finding among developmental psychologists on physical reasoning is that infants create abstract representations of objects [6, 7] from which they extract spatial and identity features. These high-level features are coupled with rules of reasoning that infants develop over time via process known as explanation-based learning [34] to form their expectation on how a physical scene should play out [35]. Existing VoE datasets in physical reasoning lack such metadata in their scenarios. These metadata can be embedded into datasets to train models with greater interpretability and effectiveness in physical-reasoning tasks.

Our contributions are three-fold. First, we present a new benchmark for physical reasoning, consisting of the first large-scale synthetic 3D VoE dataset with novel scene-wise ground-truth metadata of abstract features and rules. This dataset is inspired by findings in psychology literature and validated on human trials. Second, we propose a novel heuristic-based and oracle-based model framework to tackle the tasks of our VoE dataset. The model framework outperformed baseline and ablation computer-vision models. Third, we showed that our proposed model framework can learn an alternate reality of physical reasoning by leveraging on the feature and rule heuristics of the VoE dataset. This emphasises that the model is capable of learning universal causal relations in physical reasoning.

2 Related Works

The intersection of computer vision and physical reasoning is heavily grounded in the literature of VoE-based psychology. For example, the **barrier** event is an event category illustrating (or violating) the constraint of solidity. In studies that implemented the barrier event to infants [23, 36], psychologists placed a solid barrier with an object on one side and the *surprising* event occurred when infants were made to believe that the object could pass through the barrier. Like the barrier event, there are other events like **containment** [37, 26, 38], **occlusion** [39, 24], **collision** [27, 40] and **support** [41, 25, 5]. On the computational side of VoE-based physical reasoning, we find that Piloto et al. [28], IntPhys [29] and ADEPT [30] to be the most relevant to our work, as they all employ the VoE paradigm in their datasets and evaluation. They all present 3D datasets of very similar event categories of physical reasoning.

Piloto et al. [28] presents a 3D VoE dataset of 100,000 training videos and 10,000 pair probes of *surprising* and *expected* videos for evaluation. The dataset categorized their videos into 'object persis-

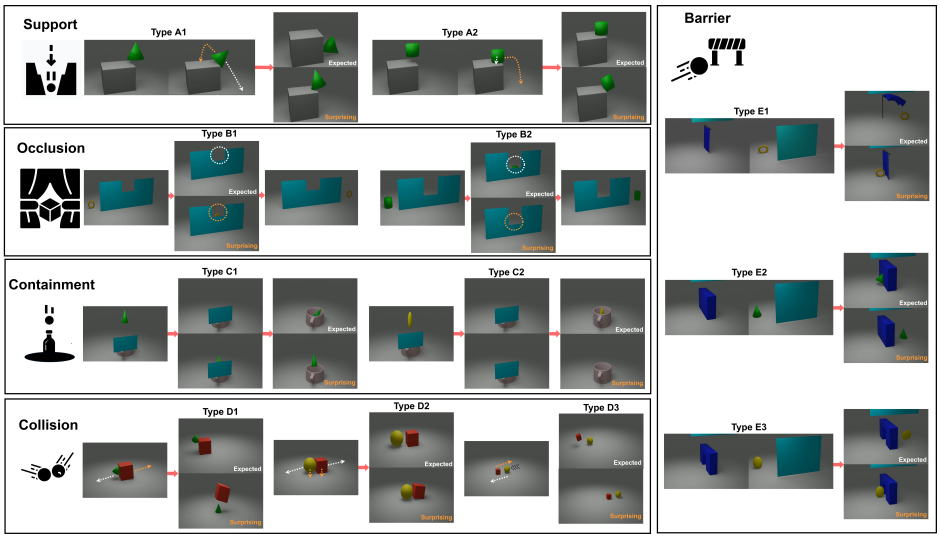

Figure 1: Examples of the different event categories in the VoE dataset with their *expected* and *surprising* outcomes. **Support**: Type A1 (*originally unbalanced*), Type A2 (*originally balanced*). **Occlusion**: Type B1 (*object shorter than occluder*), Type B2 (*object taller than occluder*). **Containment**: Type C1 (*Object fully contained in container*), Type C2 (*object protruding out of container*). **Collision**: Type D1 (*same speed, different size*), Type D2 (*same speed, same size*), Type D3 (*different speed, same size*). **Barrier**: Type E1 (*soft barrier*), Type E2 (*solid barrier*), Type E3 (*barrier with opening*).

tence', 'unchangeableness', 'continuity', 'solidity' and 'containment'. Their Variational Autoencoder model benchmarked on the dataset and showed promise in 'assimilating basic physical concepts'.

IntPhys [29] is a 3D VoE dataset with 15,000 videos of *possible* events and 3,960 videos of *possible* and *impossible* events in the test and dev sets. Only three events on 'object permanence', 'shape constancy' and 'continuity' were present. The study benchmarked the performances of a convolutional autoencoder and generative adversarial network with short and long-term predictions. The models performed poorly but with higher than chance in comparison with their adult human trials.

ADEPT [30] is a model that uses extended probabilistic simulation and particle filtering to predict object expectation. They use ADEPT on a 3D VoE dataset of 1,000 training videos of random objects colliding and 1,512 test videos of *surprising* or *control* stimuli. ADEPT accurately predicts the expected location of objects behind occluders to measure surprise, while replicating adult human judgements on the 'how, when and what' traits of *surprising* scenes [42].

While these datasets provide vision data to replicate experiments in VoE for physical reasoning, they do not explicitly provide any heuristic-based metadata that models can exploit for higher-level interpretable predictions of physical reasoning. Researchers in cognitive AI have been calling for the use rule-based causal reasoning by adopting heuristics [43, 18, 44] in their approaches. We believe that inductive biases and heuristics that guide learning are important for computational physical reasoning. This is especially the case in VoE tasks that only train on *expected* tasks and are made to predict which scenes are *surprising* [28, 29, 30]. This motivated the construction of our VoE dataset.

3 VoE Dataset

3.1 Overview

Figure 1 comprehensively illustrates the composition of the VoE dataset, which comprises synthetic video sub-datasets in five event categories: **support** (A), **occlusion** (B), **containment** (C), **collision**

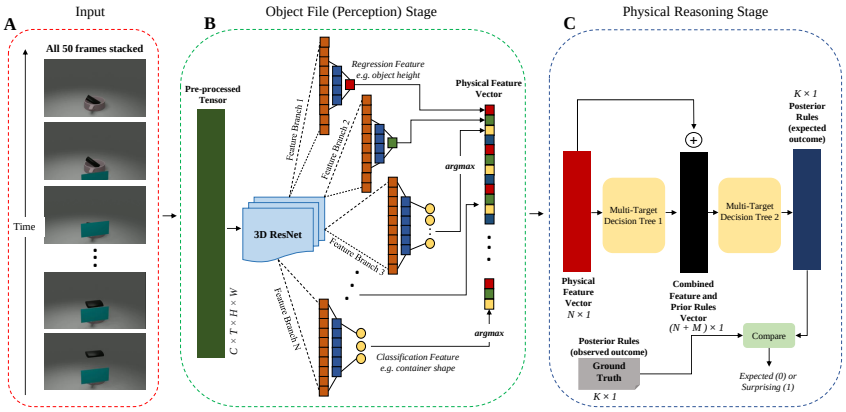

Figure 2: The OFPR-Net architecture. (A) **Input**: the original data inputs for the VoE task comprise 50 stacked frames. (B) **Object File (Perception) Stage**: The pre-processed input is fed into a 3D ResNet which then copies its output to N feature branches. Each branch predicts either a scalar or classification feature. (C) **Physical Reasoning Stage**: The concatenated feature vector is fed into a multi-target decision tree to predict prior rules. The prior rules are concatenated with the feature vector and fed into another multi-target decision tree to predict the posterior rules. The predicted posterior rules are compared with the ground truth to determine if the input scene is *surprising*.

(D) and **barrier** (E). Each one of these event categories are split into further sub-categories that showcase physical variations based on the differing scene stimuli. They are described as follows.

Support (Type A): An object is dropped above the edge of a support. The object's centre of mass is situated either over the edge (**Type A1**) or within the edge (**Type A2**). **Occlusion (Type B)**: An inert object has initial momentum behind an occluder. The object can be shorter than the occluder's middle portion (**Type B1**) or taller (**Type B2**). **Containment (Type C)**: An inert object falls from a short height above a container, with the interaction hidden behind an occluder. The object is short enough to be fully contained inside the container (**Type C1**) or tall enough to protude out of the container top (**Type C2**). **Collision (Type D)**: Two inert objects with initial momentum collide head-on. In the first case (**Type D1**), two objects have the same initial speed with different sizes. In two other cases, both objects have similar size, with either the different (**Type D2**) or same (**Type D3**) initial speeds. **Barrier (Type E)**: An inert object has an initial momentum to pass through a barrier, with their interaction hidden behind an occluder. To explore different barriers, the dataset comprises events with either a soft barrier (**Type E1**), a solid barrier (**Type E2**) or a barrier with opening (**Type E3**).

3.2 Features and Rules

Every event category $\psi \in \{A, B, C, D, E\}$ comes with sets of abstract features f^{ψ}, prior rules r^{ψ}_{prior} & posterior rules r^{ψ}_{post} where $|f^{\psi}| = 20$, $|r^{\psi}_{prior}| = 13$ & $|r^{\psi}_{post}| = 9$. Prior rules are physical conditions about the event which can be answered with the features (e.g. height, width). These prior rules, coupled with the features, suffice to answer posterior rules representing the outcome of the physical interaction. For instance, the features of a containment event could refer to the heights of the object and container. We present prior rules as a question: "is the object taller than the container?" to which the answer 'yes' (based on feature comparison) would aid in answering a posterior rule as a question: "did the object protrude out of the container?". A full list of f^{ψ}, r^{ψ}_{prior} & r^{ψ}_{post} and a complete description of the procedural generation process are in the supplementary.

3.3 Dataset Structure

Each event category ψ has 5,000 different configured trials, amounting to 25,000 trials in the VoE dataset. Every trial showcases an *expected* or *surprising* scene pair of the same stimuli. The training-validation-test dataset split is 75%-15%-10%. This sums to 50,000 videos. At 50 frames per video,

the VoE dataset offers 2,500,000 frames, each with a size of 960×540 pixels. The VoE dataset also provides the depth map and instance segmented frames. Along with the automatically generated ground-truth labels of f^{ψ}, r^{ψ}_{prior} & r^{ψ}_{post} in every video, the frame-wise world position and orientation of all entities are provided. f^{ψ}, r^{ψ}_{prior} & r^{ψ}_{post} are only used for training as they are not relevant to our VoE evaluation (see section 5.4). Nonetheless, they are still provided for the test and validation sets should researchers choose to evaluate performance in predicting f^{ψ}, r^{ψ}_{prior} & r^{ψ}_{post}. All frames were developed in the open-source 3D graphics software Blender [45], using a Python API.

4 OFPR-Net

To understand how abstract features and rules can be used in physical reasoning, we examined a two-system cognitive model developed by infant psychologists [8]. They theorized that early stage physical reasoning is supported by an *object-file* system [6] and a *physical-reasoning* system [4] that serve different functions. The *object-file* system builds temporary spatio-temporal and identity representations of objects. When objects become involved in a causal interaction, the *physical-reasoning* system becomes activated to predict the outcome of the interaction by first categorizing the event and then combining the temporary representations from the *object-file* with its physical knowledge. If the observed outcome does not match the expected outcome, it is signaled as a *surprising* event. To draw parallels between our VoE dataset and the *object-file physical-reasoning* system, the features (f^{ψ}) are analogous to the temporary identity representations of objects recognized by the *object-file* system. The prior and posterior rules (r^{ψ}_{prior} & r^{ψ}_{post}) are analogous to the symbolic structure of the *physical-reasoning* system that provides its physical knowledge. We believe a simplified version of this two-system cognitive model can be computationally represented.

To showcase how a model can exploit the novel heuristics on our dataset, we introduce the Object File Physical Reasoning Network (OFPR-Net): a novel oracle-based model framework for modeling VoE in physical reasoning across event categories. A detailed architecture of the OFPR-Net is shown in Figure 2. The essence of this model is to predict the expected outcome based on the stimuli of a scene, which can then be compared with the oracle of the actual outcome to decide if the scene is *surprising*. As the focus is primarily on showing how the features and rules can be exploited to form expectations of the physical outcome and **not** on classifying the video's actual outcome, we found it suitable to use an oracle in this final step. This would pin the model's performance to the aforementioned focus. A detailed description of the model architecture is in the supplementary.

5 Experiments

5.1 Human Trials

To benchmark human performance on the VoE dataset and validate its trials, we conducted an experiment testing adult humans on their judgement of the surprising level of the videos. 61 participants (50 accepted responses) were recruited to answer an online questionnaire and were compensated with $\$7.50$ each. Every participant was familiarized with 12 trials, where each trial showcased an *expected* and *surprising* version of the same stimuli. All participants were shown the same familiarization trials, and each trial represented a subcategory of every event category (**A1** - **E3**) and was selectively chosen from the training and validation sets. We randomly sampled 10% of the combined VoE test set (250 trials $\leftrightarrow$ 500 scenes), drawn evenly from each event category. Figure 3 (A) illustrates the implementation. Like the VoE human trials in [30, 31], the responses for each participant are standardized in the Z-normal distribution. This accounts for the participants' different usage of the slider, making their responses directly comparable. Full details are in the supplementary.

5.2 Baseline Model

We establish a simple baseline by training a 3D ResNet [46] pretrained on the Kinetics-700 dataset [47]. Four additional fully connected feed-forward layers are augmented to the 3D ResNet. These

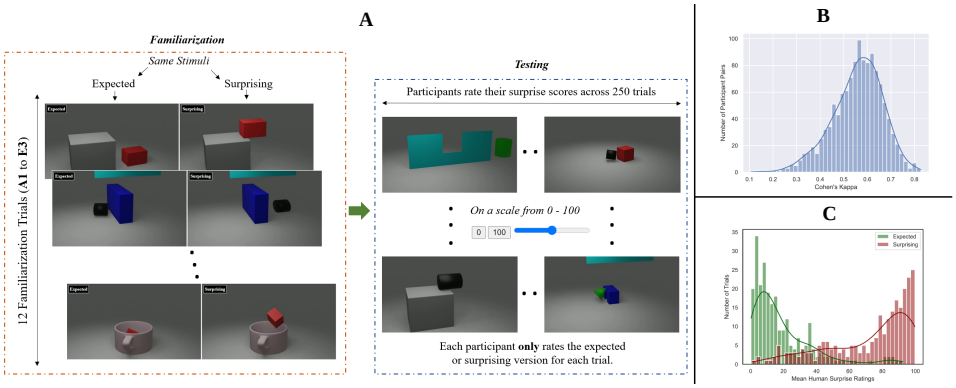

Figure 3: (A) Human trial setup with the familiarization stage and testing stage. (B) Cohen's κ for common scene ratings (Z-scored) between all pairs of human participants. (C) Distributions of the mean human ratings per trial for *expected* and *surprising* scenes.

layers fine tune the model to output a scalar output with a sigmoid activation representing the surprise score from 0 (*expected*) to 1 (*surprising*).

5.3 OFPR-Net

Implementation: The OFPR-Net is implemented using PyTorch [48] and the 3D ResNet (MIT License) implementation by [46] is used with their pre-trained weights (r3d34_K_200ep) on the Kinetics-700 dataset [47]. All regression blocks are trained with mean squared error loss and classifications blocks are trained with categorical cross-entropy loss. The model architecture assumes 24 features (engineered from 20 features to avoid negative scalar features), 13 prior rules and 9 posterior rules, matching what the VoE dataset offers. The multi-target decision trees are fully trained with f^{ψ}, r^{ψ}_{prior} & r^{ψ}_{post} of the training set.

Ablation study : We conducted an ablation study to evaluate the inclusion of the Physical Reasoning stage of the OFPR-Net. Specifically, we removed the feature branches and multi-target decision trees. The Object File stage is modified by replacing the output of the 3D ResNet with 9 classification block branches that attempt to predict the posterior rules (expected outcome). The model pipeline and implementation from the input to preprocessing and the 3D ResNet remain identical to OFPR-Net. Like the OFPR-Net, the predicted posterior rules are compared with the oracle of the ground truth posterior rules to determine if a scene is *surprising* or *expected*. As the Physical Reasoning stage is removed and the Object File stage remains as a modified version, we call this the OF-Net.

5.4 Evaluation Metric

To evaluate model performance on the VoE dataset, we define the Hit Rate, H_r ($= \sum_{i=1}^{J} H_r^i$), similar to [29] where E^i and S^i refer to the surprise level scores of the *expected* and *surprising* versions of trial i from a set of J trials.

$$H_r^i = \begin{cases} 1 & E^i < S^i \\ 0.5 & E^i = S^i \\ 0 & \text{Otherwise} \end{cases} \tag{1}$$

Unlike human responses, all models provide independent surprise ratings. Therefore, we feed both *expected* and *surprising* scenes with the same stimuli on the same model to compute H_r. The formulation of H_r for human responses is adjusted to account for the fact that participants either rate the *expected* or *surprising* version of a trial, to ensure independent ratings. We consider all 625 (25^2) combinations of *expected* scene ratings and *surprising* scene ratings for each trial. The H_r is

computed by taking the average H_r^i (i now referring to the ith combination) of the 625 combinations and then taking the mean of these average scores across all trials.

5.5 Experimental Setup

The aim of the experiment in the present study is to compare the performance of models with humans in their ability to recognize if physical interactions within each event category is *surprising* or expected. To fairly compare model performances with human performance, we evaluated all models with the exact 10% test set used in the human trials (section 5.1). To keep consistent with the testing size, all models were also trained and validated using 10% of the training and validation sets. The experiments are split into 5 tasks: **A, B, C, D, E**. From **A** to **E**, every model is trained on data stipulated purely for each corresponding event category. Following the methodology of existing computational VoE datasets [31, 30, 28, 29, 32], all models are **only trained on *expected* videos**. This sums to 375 training scenes, 150 validation scenes and 100 testing scenes for each event category. Human trial responses for the same 100 test videos are used for comparison. Each model is run on all tasks with 30 epochs for 10 seeded runs each on a single NVIDIA Tesla V100-32GB GPU. See supplementary for preprocessing and hyper-parameter tuning details.

6 Results and Analysis

6.1 Human Performance

To check for inter-rater reliability, we measured the Cohen's κ [49] of our human responses. To classify each response, we assume that a Z score < 0 indicates the *expected* class and a Z score ≥ 0 indicates the *surprising* class. We believe that this is a fair assumption, as participants were shown an equal number of *surprising* and *expected* scenes and the Z-normal standardization accounts for the different usage of the rating slider. Given that the Cohen's κ is measured between a pair of raters, we filtered out all common videos rated by each of the 1225 ($=\binom{50}{2}$) pairs of participants and measured the κ based on them. Figure 3(B) shows the uni-variate distribution of the κ scores for all participant-pairs with a mean of 0.558. The distribution and the mean value reveal that the participants have 'moderate' (close to 'substantial') agreement as defined by [50]. Table 1 shows that the H_r of the human responses are high in all event categories except for support (**Type A**). This is expected, as the support task was especially challenging for humans to tackle. As the view of the support was not perpendicular in the scene view, it is difficult to judge the position of the object mass center over the support edge precisely in the many corner cases where the object's mass center is very near to the support edge. The slight dip in performance for collision (**Type D**) can be explained in [51] and [27], showing that humans often mis-approximate the mass and the violation of object speed in collisions respectively. The mean human rating spread based on the original rating values are shown in Figure 3(C) to visualize the rating spread of *expected* and *surprising* scenes. The plot shows the participants can generally rate *expected* and *surprising* scenes accurately, regardless of the stimuli. This is further substantiated with an AUC-ROC [52] score of 0.938 for the mean human ratings. The plot also illustrates only 6% of *expected* videos were rated ≥ 50 on average, while 22.4% of *surprising* videos were rated < 50 on average. This trend can be explained by considering the interpretation of a 'surprise rating' to a human. Feedback gathered from the human trials revealed some participants would often set a low value for a scene they find surprising to distinguish from other scenes they find more surprising. This reinforces the idea that we cannot take 50 as a universal threshold and further justifies our decision to standardize the ratings.

6.2 Model Performance

Table 1 reveals the average H_r performance of all models for 10 seeded runs for all event categories. The results show that OFPR-Net surpasses the performances of all models across all event categories. In particular, the OFPR-Net outperformed the OF-Net ablation model by an average of 7.01%. This signals that the Physical Reasoning stage of the OFPR-Net boosts the performance in VoE

	Hit Rate (H_r for normal reality)					
Methods	Support (**A**)	Occlusion (**B**)	Containment (**C**)	Collision (**D**)	Barrier (**E**)	Average
Human	0.686	0.844	0.946	0.788	0.883	0.829
Random	0.502	0.502	0.502	0.502	0.502	0.502
Baseline	0.500	0.500	0.500	0.500	0.500	0.500
OF-Net (Ablation)	0.629	0.818	0.811	0.491	0.745	0.699
OFPR-Net (Ours)	**0.676**	**0.907**	**0.855**	**0.532**	**0.768**	**0.748**

Table 1: Hit Rate for Human Trials and all Models across all event categories. Best performing model is **bolded**.

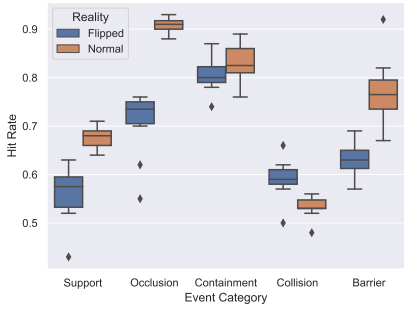

Figure 4: Box plots of Hit Rate for OFPR-Net when trained on only *surprising* videos (flipped reality) or only *expected* videos (normal reality) with the outliers shown.

tasks, showing the importance of learning features and their associations with the outcome via rules. However, the gap in performance is not very significant. This is not surprising, as the OF-Net still uses the posterior rule heuristics of our dataset to develop greater understanding of the expected outcome. For the sake of comparison, a random model that randomly selected a surprise rating in the uniform range $E^i, S^i \in [0, 1]$ is shown in Table 1 to illustrate that the baseline is as good as random, performing significantly worse than models exploiting the heuristics in our dataset. By training on only *expected* scenes, the baseline predicts all scenes as *expected* (i.e. $E^i = S^i = 0$), hence receiving a consistent H_r of 0.5. This baseline illustrates why a purely end-to-end model with no consideration of heuristics or inductive biases will not work on such tasks. Comparing across the event categories, the OFPR-Net performed poorly in the collision (**Type D**) trials. Closer inspection of the N feature branches losses of the Object File stage revealed that the model was poor in predicting the object velocities, which significantly altered the expected outcome in the Physical Reasoning stage of the OFPR-Net. As features like velocities require multiple frames to determine, it is more challenging to predict accurately with the limited data used for training. The human performance was higher than the OFPR-Net by an average of 10.83% and performed better in all tasks except occlusion (**Type B**). Given the high standards of adult human physical reasoning, we find this acceptable. Therefore, surpassing human performance across all event categories remains an open challenge. Interestingly, our model's performance dipped in the same tasks where human performance dipped (**Types A & D**). An explanation for this is that intrinsic features like mass and centre-of-mass are difficult to infer just with vision as the main input modality, like with humans [51].

6.3 Novel Insights

The results confirm that the inherent structure of the OFPR-Net model can tackle the one class classification problem of the VoE dataset. By learning to predict features and their structures and relations with the rules, the OFPR-Net has added interpretability about the physical interaction. Not only can the OFPR-Net predict the expected outcome, but it can store knowledge the features and its basic causal relations to the outcome of the interaction. This added interpretability is crucial to creating safe AI applications. We believe that another advantage of the feature and rule based architecture of the OFPR-Net is that it is flexible to learn any reality presented to it. To test our hypothesis, we ran the OFPR-Net by **only training on the *surprising* versions of all the training**

trials. The model is made to believe that *surprising* scenes depict reality and the *expected* videos violate them. All other task-specific hyper-parameters and implementation were identical to the experimental setup (section 5.5). In this flipped reality scenario, the hit rate is redefined as $1 - H_r$ as the model should label *expected* scenes more surprising than *surprising* scenes. While the OFPR-Net performs better in the normal reality (except collision, **Type D**), the box plots in Figure 4 show that it still performs reasonably and better than chance in the flipped reality. Hence, the OFPR-Net can treat *surprising* videos as the new normal and is more likely to signal *expected* videos as a violation in this new normal. The OFPR-Net signals that it is capable of learning universal causal relationships in physical reasoning. This is mainly possible because of the structural representation of f^{ψ}, r^{ψ}_{prior} & r^{ψ}_{post} in the Physical Reasoning stage. To build systems that are capable of physical reasoning, frameworks that allow universal learning of causal relationships are crucial to "support explanation and understanding, rather than merely solving pattern recognition problems" [1].

7 Limitations and Future Work

When implementing the human trials, each participant only rated either the *expected* or *surprising* version of a trial. While this takes precedence from previous work in computational VoE [31, 30], it meant that the human trial results are not directly comparable to the model output. Future computational VoE work with human trials may consider splitting the data collection into two stages with a set time period between the stages. The first stage can follow the method of our human trials, while the second stage shows the remaining videos not shown in the first stage. The time between the stages allows the participants to forget any stimuli, making the *surprising* and *expected* ratings independent. One constraint of the dataset is that it only considers a few simple event categories and assumes each scene can only be represented by one event. In reality, physical interactions are much more complex, containing multiple event categories and a wider range of features and rules guiding the interactions. This scales up with more active objects in a rich environment. Nevertheless, we believe that providing f^{ψ}, r^{ψ}_{prior} & r^{ψ}_{post} with scenes containing simple interactions is an important step to unveiling the potential of these heuristics. Future versions of the dataset will explore complex interactions with a wider range of rules and features. Frameworks built on our VoE dataset can also explore probabilistic and generative approaches to make use of the heuristics. Finally, researchers may consider developing a curriculum-learning approach that attempts to closely replicate explanation-based learning [34] on VoE tasks.

8 Conclusion

In this work, we showcase a novel approach to modeling VoE across basic event categories of physical reasoning. By leveraging on findings in the psychology literature, we proposed a novel synthetic 3D dataset augmented with ground-truth labels of abstract features and rules in five event categories of physical reasoning. The task of the dataset is to recognize scenes as *expected* or *surprising* with the added challenge of training on only *expected* scenes. Human trials were conducted to benchmark human-level performance. The trials revealed that there was general agreement among participant responses. The participants were also proficient at rating *surprising* videos with high surprise ratings and *expected* videos with low surprise ratings. To showcase how using the abstract rules and features can tackle the challenge of our dataset, we proposed OFPR-Net, a novel oracle-based model framework inspired by a two-system cognitive model [8] of an infant's physical-reasoning system. The OFPR-Net benchmark surpassed the performance of the baseline and ablation models. However, average human-level performance still exceeded that of the OFPR-Net. Therefore, it remains an open challenge to beat the benchmarked human-level performance on our dataset. Finally, we show that the structural nature of the OFPR-Net guided by features and rules is flexible in learning an alternative reality of physical reasoning. This emphasises that a model that exploits heuristics of physical reasoning is capable of learning universal causal relations that are necessary to create systems with better interpretability. This validates the novelty of our dataset and encourages future work in this paradigm to focus on such heuristics and inductive biases for learning in physical reasoning.

Acknowledgments and Disclosure of Funding

This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore under its AME Programmatic Funding Scheme (Award #A18A2b0046). We would also like to thank the National University of Singapore for the computational resources used in this work.

References

[1] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, 2017.

[2] S. Adams, I. Arel, J. Bach, R. Coop, R. Furlan, B. Goertzel, J. S. Hall, A. Samsonovich, M. Scheutz, M. Schlesinger *et al.*, "Mapping the landscape of human-level artificial general intelligence," *AI magazine*, vol. 33, no. 1, pp. 25–42, 2012.

[3] E. S. Spelke and K. D. Kinzler, "Core knowledge," *Developmental science*, vol. 10, no. 1, pp. 89–96, 2007.

[4] R. Baillargeon, J. Li, Y. Gertner, and D. Wu, "How do infants reason about physical events?" pp. 11–48, 2011.

[5] S. J. Hespos and R. Baillargeon, "Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings," *Cognition*, vol. 107, no. 1, pp. 304–316, 2008.

[6] D. Kahneman, A. Treisman, and B. J. Gibbs, "The reviewing of object files: Object-specific integration of information," *Cognitive psychology*, vol. 24, no. 2, pp. 175–219, 1992.

[7] R. D. Gordon and D. E. Irwin, "What's in an object file? evidence from priming studies," *Perception & Psychophysics*, vol. 58, no. 8, pp. 1260–1277, 1996.

[8] M. Stavans, Y. Lin, D. Wu, and R. Baillargeon, "Catastrophic individuation failures in infancy: A new model and predictions." *Psychological Review*, vol. 126, no. 2, p. 196–225, 2019.

[9] T. D. Ullman and J. B. Tenenbaum, "Bayesian models of conceptual development: Learning as building models of the world," *Annual Review of Developmental Psychology*, vol. 2, pp. 533–558, 2020.

[10] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the National Academy of Sciences*, vol. 110, no. 45, pp. 18 327–18 332, 2013.

[11] R. Zhang, J. Wu, C. Zhang, W. T. Freeman, and J. B. Tenenbaum, "A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding," *arXiv preprint arXiv:1605.01138*, 2016.

[12] A. Lerer, S. Gross, and R. Fergus, "Learning physical intuition of block towers by example," in *International conference on machine learning*. PMLR, 2016, pp. 430–438.

[13] J. Duan, S. Y. B. Jian, and C. Tan, "Space: A simulator for physical interactions and causal learning in 3d environments," *arXiv preprint arXiv:2108.06180*, 2021.

[14] D. M. Bear, E. Wang, D. Mrowca, F. J. Binder, H.-Y. F. Tung, R. Pramod, C. Holdaway, S. Tao, K. Smith, L. Fei-Fei *et al.*, "Physion: Evaluating physical prediction from vision in humans and machines," *arXiv preprint arXiv:2106.08261*, 2021.

[15] J. Duan, S. Yu, S. Poria, B. Wen, and C. Tan, "Pip: Physical interaction prediction via mental imagery with span selection," *arXiv preprint arXiv:2109.04683*, 2021.

[16] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *arXiv preprint arXiv:2103.04918*, 2021.

[17] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Computer Vision and Image Understanding*, vol. 163, pp. 21–40, 2017.

[18] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," *arXiv preprint arXiv:1910.01442*, 2019.

[19] T. Ates, M. S. Atesoglu, C. Yigit, I. Kesen, M. Kobas, E. Erdem, A. Erdem, T. Goksun, and D. Yuret, "Craft: A benchmark for causal reasoning about forces and interactions," *arXiv preprint arXiv:2012.04293*, 2020.

[20] M. Wagner, H. Basevi, R. Shetty, W. Li, M. Malinowski, M. Fritz, and A. Leonardis, "Answering visual what-if questions: From actions to predicted scene descriptions," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[21] R. Baillargeon, E. S. Spelke, and S. Wasserman, "Object permanence in five-month-old infants," *Cognition*, vol. 20, no. 3, pp. 191–208, 1985.

[22] R. Baillargeon, "Object permanence in $3\frac{1}{2}$-and $4\frac{1}{2}$-month-old infants." *Developmental psychology*, vol. 23, no. 5, p. 655, 1987.

[23] R. Baillargeon, M. Graber, J. Devos, and J. Black, "Why do young infants fail to search for hidden objects?" *Cognition*, vol. 36, no. 3, pp. 255–284, 1990.

[24] E. S. Spelke, R. Kestenbaum, D. J. Simons, and D. Wein, "Spatiotemporal continuity, smoothness of motion and object identity in infancy," *British journal of developmental psychology*, vol. 13, no. 2, pp. 113–142, 1995.

[25] N. Dan, T. Omori, and Y. Tomiyasu, "Development of infants' intuitions about support relations: Sensitivity to stability," *Developmental Science*, vol. 3, no. 2, pp. 171–180, 2000.

[26] S.-h. Wang, R. Baillargeon, and S. Paterson, "Detecting continuity violations in infancy: A new account and new evidence from covering and tube events," *Cognition*, vol. 95, no. 2, pp. 129–173, 2005.

[27] L. Kotovsky and R. Baillargeon, "Calibration-based reasoning about collision events in 11-month-old infants," *Cognition*, vol. 51, no. 2, pp. 107–129, 1994.

[28] L. Piloto, A. Weinstein, D. TB, A. Ahuja, M. Mirza, G. Wayne, D. Amos, C.-c. Hung, and M. Botvinick, "Probing physics knowledge using tools from developmental psychology," *arXiv preprint arXiv:1804.01128*, 2018.

[29] R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, and E. Dupoux, "Intphys: A framework and benchmark for visual intuitive physics reasoning," *arXiv preprint arXiv:1803.07616*, 2018.

[30] K. Smith, L. Mei, S. Yao, J. Wu, E. Spelke, J. Tenenbaum, and T. Ullman, "Modeling expectation violation in intuitive physics with coarse probabilistic object representations," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8985–8995, 2019.

[31] T. Shu, A. Bhandwaldar, C. Gan, K. Smith, S. Liu, D. Gutfreund, E. Spelke, J. Tenenbaum, and T. Ullman, "Agent: A benchmark for core psychological reasoning," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, 2021, pp. 9614–9625.

[32] K. Gandhi, G. Stojnic, B. M. Lake, and M. R. Dillon, "Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others," *arXiv preprint arXiv:2102.11938*, 2021.

[33] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1–34, 1998.

[34] R. Baillargeon and G. F. DeJong, "Explanation-based learning in infancy," *Psychonomic bulletin & review*, vol. 24, no. 5, pp. 1511–1526, 2017.

[35] Y. Lin, M. Stavans, and R. Baillargeon, "Infants' physical reasoning and the cognitive architecture that supports it," 2020.

[36] E. S. Spelke, K. Breinlinger, J. Macomber, and K. Jacobson, "Origins of knowledge." *Psychological review*, vol. 99, no. 4, p. 605, 1992.

[37] S. J. Hespos and R. Baillargeon, "Infants' knowledge about occlusion and containment events: A surprising discrepancy," *Psychological Science*, vol. 12, no. 2, pp. 141–147, 2001.

[38] Y. Mou and Y. Luo, "Is it a container? young infants' understanding of containment events," *Infancy*, vol. 22, no. 2, pp. 256–270, 2017.

[39] R. Baillargeon and J. DeVos, "Object permanence in young infants: Further evidence," *Child development*, vol. 62, no. 6, pp. 1227–1246, 1991.

[40] L. Kotovsky and R. Baillargeon, "The development of calibration-based reasoning about collision events in young infants," *Cognition*, vol. 67, no. 3, pp. 311–351, 1998.

[41] R. Baillargeon, A. Needham, and J. DeVos, "The development of young infants' intuitions about support," *Early development and parenting*, vol. 1, no. 2, pp. 69–78, 1992.

[42] K. Smith, L. Mei, S. Yao, J. Wu, E. S. Spelke, J. Tenenbaum, and T. D. Ullman, "The fine structure of surprise in intuitive physics: when, why, and how much?" in *CogSci*, 2020.

[43] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi, ""what happens if..." learning to predict the effect of forces in images," in *European conference on computer vision*. Springer, 2016, pp. 269–285.

[44] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum, "A compositional object-based approach to learning physical dynamics," *arXiv preprint arXiv:1612.00341*, 2016.

[45] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[46] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555. [Online]. Available: https://github.com/kenshohara/3D-ResNets-PyTorch

[47] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[49] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[50] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[51] A. Mitko and J. Fischer, "A striking take on mass inferences from collisions," *Journal of Vision*, vol. 21, no. 9, pp. 2812–2812, 2021.

[52] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 7
   (c) Did you discuss any potential negative societal impacts of your work? [No] If the reviewers express specific worries on any specific potential negative societal impact of our work, we shall address them in the camera-ready version of the paper/supplementary
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See Supplementary
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See URL
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Supplementary and Section 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Supplementary for all seeded results and see Figure 4 for error bars
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See section 5.3
   (b) Did you mention the license of the assets? [Yes] See section 5.3
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See URL for new dataset and new code
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not use anyone else's data, we did use their publicly available model with pre-trained weights and cited accordingly
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Supplementary
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Supplementary

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See Supplementary

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Supplementary