Supplementary Materials for 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

Abstract

The supplementary materials provides additional information on 5 specific areas. First, we provide more details on features and rules of the VoE dataset. Second, we give a detailed account of the protocols followed and trial process of the human trials. third, we provide a detailed description of the OFPR-Net model architecture. Fourth, we provide more clarity on hyper-parameter tuning and tuned hyper-parameters for each event category. Finally, we show the detailed set of results for our main experiment.

1 VoE Dataset

1.1 Primary Description

Table 1 shows the number of trials for each event category for the full VoE dataset. The table is further stratified for sub-types of each event category. Table 2 shows the equivalent information for the randomly sampled 10% of the dataset. This 10% dataset was used for training all models and testing human judgement. A list of all features in the dataset along with an indication of all causally relevant features per event category is shown in Table 5. Similarly, the list of prior rules and posterior rules can be found in Tables 6 & 7 respectively. Figure 8 shows the distribution of the posterior for each one of the event categories. We filtered to 6 rules with one rule from each event category and 2 rules from **Type D**.

1.2 Procedural Generation

Multiple physical stimuli that affect the outcome of the interaction were randomly sampled to amplify the diversity of the dataset. Common parameters among the 5 sub-datasets included the object's shape $S_{Obj} \in$ {Cube, Cylinder, Torus, Sphere, Cone, Side Cylinder, Inverted Cone} and the object's height and width, $H_{Obj}, W_{Obj} \in [0.4, 1.6]$, where a 1 is equivalent to $2m$ in the 3D environment. The initial contact point of the object on the support in **A** $C_{Obj} \in [0.2, 0.8]$ where $C_{Obj} = 0.2$ indicates that the 20% of the object's width is over the edge. In **B**, the occluder's middle segment height, $H_{Occ}, \in$

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

$[0.1, 0.9]$ with 1 being the height of the occluder. In **C**, the container's shape, $S_{Con} \in \{Mug, Box\}$ was also varied with its height and width, $H_{Con}, W_{Con} \in [0.5, 1.5]$. The height ($\propto$ mass$^{\frac{1}{3}}$) and initial speeds of objects in **D** were sampled as $H_{Obj} \in [0.5, 1.5]$ and $V_{Obj} \in [0.5, 2.5]$. In **E3**, the barrier's opening height and width were sampled in the range, $H_{Bar}, W_{Bar} \in [0.4, 1.4]$.

1.3 Dataset Structure and Composition

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. 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 [1], using a Python API.

2 Human Trials

2.1 Implementation

The task for each participant was to rate how surprising they found scenes assigned to them on an integer slider from 0 (*expected*) to 100 (*surprising*). All responses were filtered via pre-set criteria for accuracy and consistency. 11 participants did not meet our pre-set criteria, hence we excluded their data. Therefore, we establish our analysis in the responses of 50 participants (31 female, 18 male and 1 non-binary) whose ages ranged from 19 to 32. Following the human trial methodology in [2], each participant was either shown the *surprising* scene or the *expected* scene of each trial. This is necessary to ensure each rating is independent and not influenced by comparing to another scene with the same stimuli. The number of *expected* and *surprising* scenes shown to each participant was evenly split. Hence, each participant was shown the *expected* versions of 125 trials and the *surprising* version of the other 125 trials. The *surprising* and *expected* versions of each trial were evenly split among all participants, such that half of all participants rated the *surprising* version of each trial and vice versa. The human trials were conducted online via a custom web application made with Flask [3] to handle the back-end operations managing different scenes for each participant.

2.2 Protocols Followed

We conducted human trials on 61 adult humans to test their performance on the randomly sampled 10% test set. These human trials were not conducted with an Institutional Review Board (IRB) approval. Nevertheless, we made sure that all procedures for these trials met the IRB requirement for exemption. We did not collect personally identifiable information throughout for all 61 participants. These 61 participants were recruited via an online email broadcast. None of the authors or anyone familiar with this work were allowed to take part in the trials. All the briefing, familiarization and test trials were conducted via a custom web application made with the Flask API [3] in Python. The website had HTTPS encryption and was only accessed via a custom password given to each participant and all browser session information that stored the responses were erased upon the completion of the trial. Before erasing any session data, all the participants' responses were anonymized with an ID before they were passed back to the research team via a JSON file. The ID was only used to map to the exact set of videos given to each anonymized participant fur performance evaluation and analysis. The responses contained data of each participants responses, age and gender (none of which are personally identifiable as they cannot be used to track to any participant).

To facilitate payment of $7.50, an anonymized & randomly generated code of 20 alphanumeric characters was generated for each participant which they could use to claim their compensation.

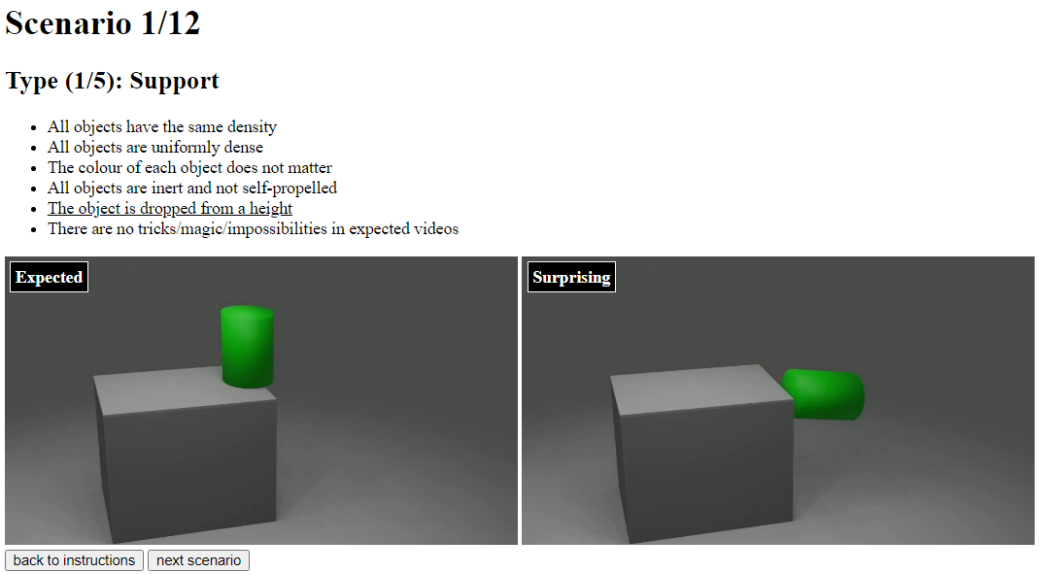

Figure 1: What the participant would see during a familiarization trial

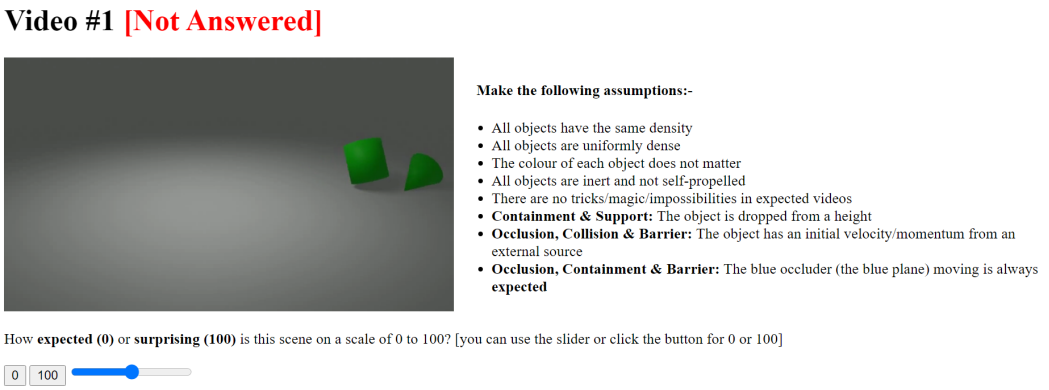

Figure 2: What the participant would see during a testing trial

There is no way to track this code to any participant as there was no identifiable personal data of any form collected throughout the entire duration of our data collection. It must also be noted that the human trial data is **not** used to build the dataset, but to form a benchmark on human performance on the VoE dataset. The dataset itself is procedurally generated using Blender [1] with no input from the human trials. Before the human trials with 61 participants, the human trial workflow was pilot tested on 6 individuals and the feedback was used to improve the trials and ensure that the website is stable with no bugs.

2.3 Trial Process

Figure 1 shows the screen that the participant would see during 1 of 12 familiarization trials. The expected and surprising version of a trial would be shown and a list of assumptions would be listed on the top. Figure 2 show the screen during a testing trial. It shows a single video with a list of assumptions on the right side so that the participants do not forget these assumptions.

Each participant was only shown the surprising version or expected version of each of the 250 trials shown to them. In such trials, it is possible to receive sub-standard responses from participants who do not conduct the trials properly, hence we placed 2 measures to ensure a minimum acceptable quality of responses. The first measure was a consistency check. 25 of the 250 videos shown to each

participant were randomly selected and repeated at another random sequence during the trial. The repeat videos were spaced apart by a few videos so that participants were not aware of the repetition. To check for consistency, we measured if both ratings of the original and repeat videos were on the same side of the slider spectrum with 50 as the threshold (< 50 as one side and ≥ 50 on the other side. As this threshold is not perfect, we accepted responses that were a minimum of 70% consistent. The second measure was to check for basic accuracy by carefully selecting 5 (3 expected and 2 surprising) videos from the validation set that had obvious and extreme expected or surprising outcomes and putting them randomly into the sequence of videos. These 5 videos were the same for all 61 participants. We only accepted responses that rated 4 of the 5 videos on the correct side of the slider spectrum. Although 50 is not a perfect threshold, this step also filtered out participants with an unusual usage of the slider. Therefore, the total number of videos shown for each participant was 280 (including the additional videos for both measures). A total of 11 participants did not meet the standard, hence the total number of accepted responses were 50. Figures 3, 4, 5, 6, 7 show the Cohen's κ [4] among all human participant pairs with common videos for all 5 event categories. It shows 'moderate' to 'substantial' agreement across all 5 event categories.

3 OFPR-Net Architecture

The 50 RGB frames are stacked, pre-processed and propagated through two stages of the OFPR-Net. First, the object file stage extracts all causally relevant features and their values. The data runs through a 3D ResNet and the output is copied into N branches, representing N causally relevant features. Each branch is either a feed-forward regression block that predicts a scalar value of a feature (e.g. object height), or a feed-forward classification block that predicts a categorical feature (e.g. object shape). As different event categories require a distinct set of features, the object file stage must also recognize which features are causally relevant. Therefore, all classification blocks are trained with an additional label for 'irrelevant', signalling that the feature either does not exist (e.g. container height in an occlusion event, **Type B**), or it does not affect the outcome. All regression blocks are also trained to predict -1 in the event of an irrelevant feature. To avoid clashing with the -1 label, all scalar features are engineered to have positive values.

The N concatenated features are propagated through the Physical Reasoning stage. First, the features run through a multi-target decision tree [5] to predict the outcomes of M prior rules. The feature vector is then concatenated with the predicted prior rule vector and is fed into a second multi-target decision tree to predict the K posterior rules (expected outcome). Similar to the features, any irrelevant rules are classified as such. In this model, we assume that we have an oracle of the ground truth posterior rules, pointing to the actual observed outcome. The predicted posterior rules are compared with the ground truth. If they differ, the model signals the scene as *surprising* and vice versa.

4 Experimental Setup

During the experimentation, we feature engineered features 5 to 8 from Table 5 for the Collision (**Type D**) event category to ensure that the values would be positive and not clash with the -1 label for 'irrelevant'. The velocities were split into a scalar quantity for magnitude and a classification for direction (0 for left, 1 for right). This is why the OFPR-Net had 24 features in its architecture instead of 20. As the OFPR-Net, Ablation and Baseline models all have a 3D ResNet backbone, they undergo the same preprocessing steps. Each video frame is scaled to 112×112 and their individual elements scaled from 0 to 1. The input for each training sample maintained a shape of $3 \times 50 \times 112 \times 112$ (Color Channels, Frames, Height, Width) into the 3D ResNet. All pre-trained weights of the 3D ResNet backbone were frozen. For hyper-parameter tuning, we identified 3 major hyper-parameters: learning rate, batch size & optimizer. The domain for learning rate sampling was $\{3 \times 10^{-2}, 1 \times 10^{-2}, 7 \times 10^{-3}, 5 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-3}, 7 \times 10^{-4}, 5 \times 10^{-4}\}$. The domain for the batch size was $\{2, 4, 8, 16, 32, 64\}$ and domain for optimizers were $\{\text{Adam}, \text{SGD}\}$. These hyper-parameters (learning rate, batch size and optimizer choice) were optimized for each of the 6

Table 1: The number of trials for each event category in the full VoE dataset

Set	Support (A)			Occlusion (B)			Containment (C)			Collision (D)				Barrier (E)				Total
	A1	A2	Total	B1	B2	Total	C1	C2	Total	D1	D2	D3	Total	E1	E2	E3	Total	-
Train	1818	1932	3750	1710	2040	3750	1616	2134	3750	1665	404	1681	3750	717	714	2319	3750	18750
Val	377	373	750	318	432	750	446	304	750	317	86	347	750	142	147	461	750	3750
Test	246	254	500	224	276	500	308	192	500	208	74	218	500	97	99	304	500	2500
Total	2441	2559	5000	2252	2748	5000	2370	2630	5000	2190	564	2246	5000	956	960	3084	5000	25000

Table 2: The number of trials for each event category for the 10% random sample of the VoE dataset

Set	Support (A)			Occlusion (B)			Containment (C)			Collision (D)				Barrier (E)				Total
	A1	A2	Total	B1	B2	Total	C1	C2	Total	D1	D2	D3	Total	E1	E2	E3	Total	-
Train	180	195	375	159	216	375	168	207	375	170	34	171	375	73	63	239	375	1875
Val	35	40	75	35	40	75	48	27	75	37	4	34	75	14	12	49	75	375
Test	26	24	50	24	26	50	31	19	50	16	9	25	50	8	15	27	50	250
Total	241	259	500	218	282	500	247	253	500	223	47	230	500	95	90	315	500	2500

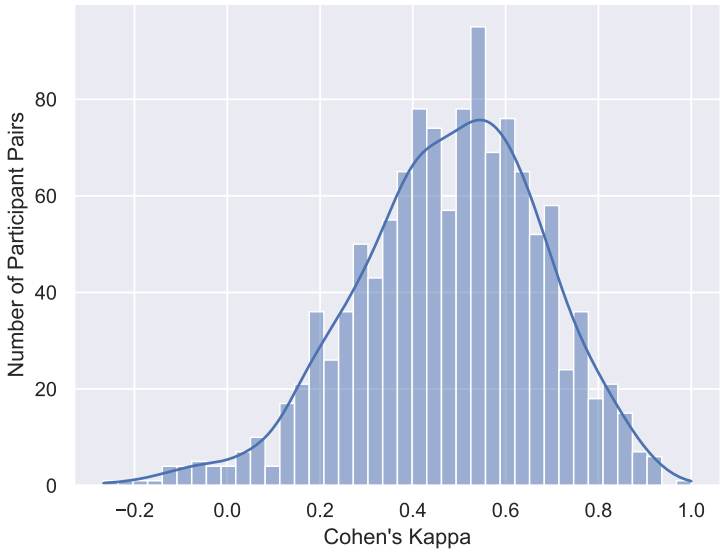

Figure 3: Cohen's κ for all participant pairs for the Support (A) event category

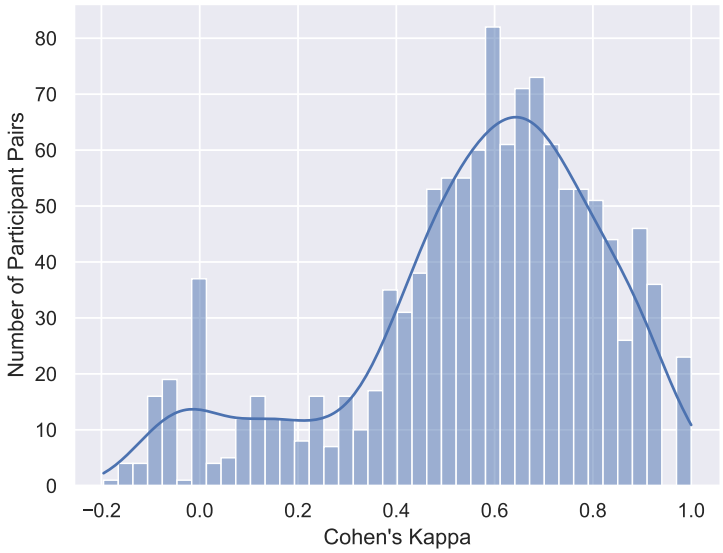

Figure 4: Cohen's κ for all participant pairs for the Occlusion (B) event category

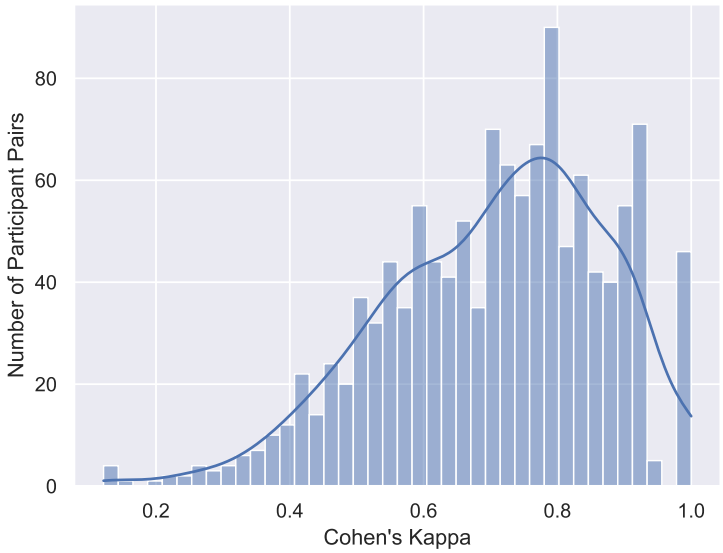

Figure 5: Cohen's κ for all participant pairs for the Containment (C) event category

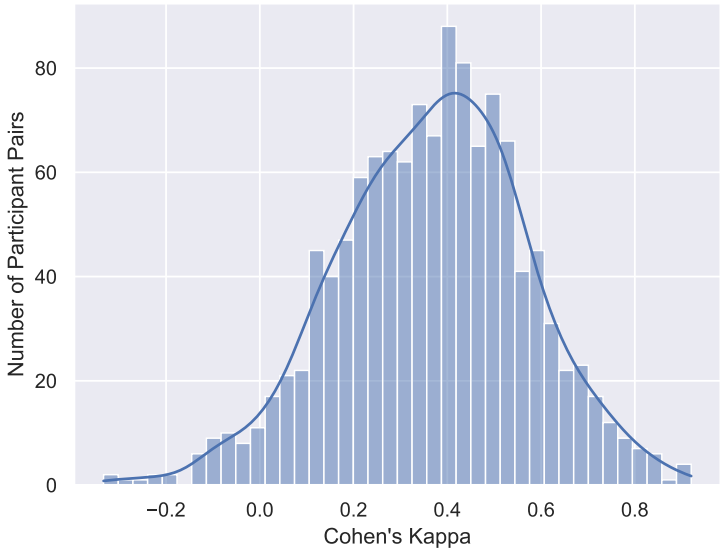

Figure 6: Cohen's κ for all participant pairs for the Collision (D) event category

Table 3: Tuned hyper-parameters for all event categories of the VoE dataset

Hyper-parameter	Support (A)	Occlusion (B)	Containment (C)	Collision (D)	Barrier (E)
Learning Rate	0.007	0.03	0.005	0.01	0.001
Batch Size	16	8	8	16	4
Optimizer	SGD	SGD	Adam	Adam	Adam

Table 4: Hit rate for the OFPR-Net and OF-Net for all event categories across all 10 seeds in the normal reality

OFPR-Net Hit Rate

Seed	Support (A)	Occlusion (B)	Containment (C)	Collision (D)	Barrier (E)
1	0.68	0.91	0.86	0.53	0.8
2	0.69	0.93	0.81	0.53	0.73
3	0.71	0.88	0.76	0.56	0.77
4	0.69	0.92	0.79	0.54	0.67
5	0.68	0.91	0.86	0.48	0.78
6	0.65	0.9	0.87	0.55	0.75
7	0.64	0.91	0.82	0.53	0.92
8	0.66	0.92	0.81	0.52	0.68
9	0.66	0.89	0.83	0.53	0.76
10	0.7	0.9	0.89	0.55	0.82

OF-Net Hit Rate

Seed	Support (A)	Occlusion (B)	Containment (C)	Collision (D)	Barrier (E)
1	0.65	0.82	0.81	0.48	0.75
2	0.63	0.76	0.83	0.49	0.8
3	0.66	0.82	0.81	0.49	0.75
4	0.64	0.86	0.81	0.45	0.77
5	0.58	0.85	0.82	0.52	0.73
6	0.63	0.82	0.74	0.55	0.71
7	0.64	0.83	0.81	0.49	0.76
8	0.64	0.78	0.84	0.44	0.74
9	0.64	0.85	0.78	0.47	0.75
10	0.58	0.79	0.78	0.53	0.69

task datasets via a grid search. A total of 20 grid search samples were extracted for each task and each sample was evaluated on the OFPR-Net for 30 epochs. The tuned hyper-parameters were used across all models for each task.

5 Detailed Results

Table 4 shows the complete set of results for all 10 seeded runs for the OFPR-Net and the OF-Net where each cell refers to a full run on 30 epochs.

References

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

[2] 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.

[3] M. Grinberg, *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.

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

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

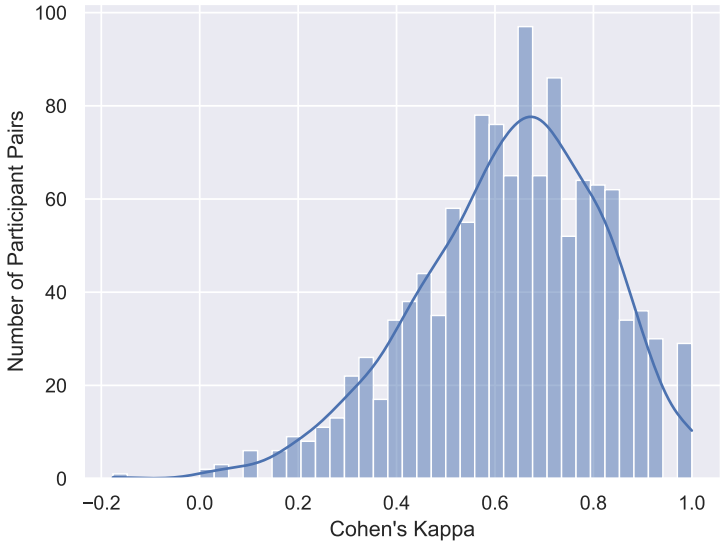

Figure 7: Cohen's κ for all participant pairs for the Barrier (E) event category

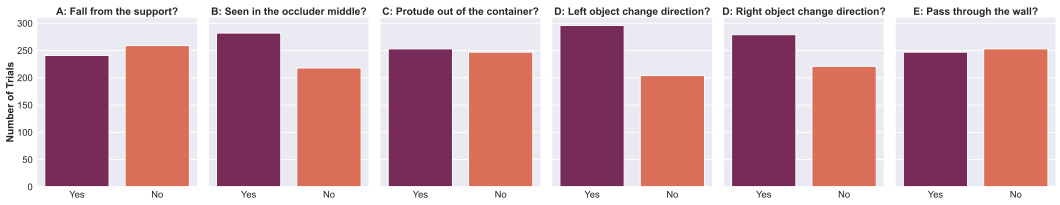

Figure 8: Posterior rule distribution for the VoE dataset for each one of the event categories (Filtered to 6 rules with one rule from each event category and 2 rules from **Type D**). The distribution depicts 'yes' or 'no' answers to a question about each posterior for every trial.

Table 5: List of 20 features in the VoE dataset. The causally relevant features for each event category is marked

	Features	Support (A)	Occlusion (B)	Containment (C)	Collision (D)	Barrier (E)
1	Object Height	✓	✓	✓	-	✓
2	Object Width	✓	✓	✓	-	✓
3	Left Size	-	-	-	✓	-
4	Right Size	-	-	-	✓	-
5	Left Prior Velocity	-	-	-	✓	-
6	Right Prior Velocity	-	-	-	✓	-
7	Left Posterior Velocity	-	-	-	✓	-
8	Right Posterior Velocity	-	-	-	✓	-
9	Container Height	-	-	✓	-	-
10	Container Width	-	-	✓	-	-
11	Wall Opening	-	-	-	-	✓
12	Wall Softness	-	-	-	-	✓
13	Opening Width	-	-	-	-	✓
14	Opening Height	-	-	-	-	✓
15	Contact Height	✓	-	-	-	-
16	Occluder Middle Height	-	✓	-	-	-
17	Object Shape	✓	✓	✓	-	✓
18	Left Shape	-	-	-	✓	-
19	Right Shape	-	-	-	✓	-
20	Container Shape	-	-	✓	-	-

Table 6: List of 13 prior rules in the VoE dataset. The relevant prior rules for each event category is marked

	Prior Rules	Support (A)	Occlusion (B)	Containment (C)	Collision (D)	Barrier (E)
1	does object have majority contact proportion?	✓	-	-	-	-
2	does object have majority volume proportion?	✓	-	-	-	-
3	is object taller than middle?	-	✓	-	-	-
4	is object taller than container?	-	-	✓	-	-
5	is object thinner than container?	-	-	✓	-	-
6	is right object larger?	-	-	-	✓	-
7	are both objects same size?	-	-	-	✓	-
8	is right object faster?	-	-	-	✓	-
9	are both objects same speed?	-	-	-	✓	-
10	is there an opening?	-	-	-	-	✓
11	is the blocker soft?	-	-	-	-	✓
12	is the object thinner than the opening?	-	-	-	-	✓
13	is the object shorter than the opening?	-	-	-	-	✓

Table 7: List of 9 posterior rules in the VoE dataset. The relevant posterior rules for each event category is marked

	Posterior Rules	Support (A)	Occlusion (B)	Containment (C)	Collision (D)	Barrier (E)
1	does support hold object?	✓	-	-	-	-
2	see object in middle?	-	✓	-	-	-
3	did the object fit?	-	-	✓	-	-
4	did the object protrude?	-	-	✓	-	-
5	did right object change direction?	-	-	-	✓	-
6	did left object change direction?	-	-	-	✓	-
7	did right object increase speed magnitude?	-	-	-	✓	-
8	did left object increase speed magnitude?	-	-	-	✓	-
9	did the object pass through the wall?	-	-	-	-	✓