Supplementary for PIP: Physical Interaction Prediction via Mental Simulation with Span Selection

Jiafei Duan[1,2,*], Samson Yu[3,4,6*], Soujanya Poria[3], Bihan Wen[5], and Cheston Tan[1,6]

[1] Institute for Infocomm Research, A*STAR
[2] University of Washington
[3] Singapore University of Technology and Design
[4] National University of Singapore
[5] Nanyang Technological University of Singapore
[6] Centre for Frontier AI Research, A*STAR

Abstract. In this supplementary material, we provide dataset analysis, experimental results and analysis, and software requirements in addition to the main paper.

1 Dataset Analysis

To increase complexity and scale, we introduce the SPACE+ dataset, an improved and expanded version of the SPACE dataset [1] with a larger dataset and additional unseen object classes for the testing model generalizability. Including data from the original SPACE dataset, we collect over 57,057 videos with over 8 million frames in total for seen objects and an additional 11,411 videos of unseen objects with over 1.7 million frames. The detailed breakdown of the SPACE+ dataset is shown in Table 3.

The SPACE dataset is made of three novel video datasets that are synthesized based on three fundamental physical interactions: *stability*, *contact*, and *containment* in a 3D environment. Each interaction scenario is synthesized using the SPACE simulator, developed using Blender, an open-source 3D computer graphics tool with a Python API. The SPACE simulator generates the scenarios for the various physical interactions and their metadata for determining the outcome of these physical interactions. The SPACE dataset comprises 15,000 synthesis videos lasting 3 seconds with a 50 frames per second frame rate. It also comes with other metadata such as the segmentation map, optical flow map, depth map and surface normal vector map of the frames.

For training and testing of PIP, we utilize only 1,000 scenes from the SPACE+ dataset for each task and split it into 60% for the training set and 20% for the

[*] Equal Contribution

	Seen Objects (%)			
	[Stability, Contact, Containment, Combined]			
Seed	Baseline	PhyDNet	PIP w/o SS	PIP
1	[92.35, 57.56, 78.87, 60.83]	[92.36, 68.77, 77.66, 58.85]	[91.70, 53.20, 84.30, 62.20]	[92.16, 89.65, 85.51, 79.13]
2	[92.35, 79.29, 82.87, 59.82]	[92.36, 53.62, 82.89, 65.94]	[92.30, 60.04, 80.08, 59.80]	[92.37, 87.79, 87.79, 76.76]
3	[92.37, 68.12, 78.47, 60.24]	[92.36, 66.25, 76.25, 60.82]	[92.30, 66.80, 84.50, 58.40]	[92.37, 86.96, 84.31, 80.71]
4	[92.35, 57.97, 68.41, 56.49]	[92.36, 57.97, 77.66, 59.00]	[92.30, 64.60, 80.28, 57.80]	[92.37, 87.37, 86.12, 78.74]
5	[92.35, 65.21, 84.90, 66.53]	[92.36, 59.55, 82.89, 65.55]	[92.30, 64.60, 72.80, 55.30]	[92.37, 85.71, 88.53, 73.23]
	Unseen Objects (%)			
	[Stability, Contact, Containment, Combined]			
Seed	Baseline	PhyDNet	PIP w/o SS	PIP
1	[65.38, 61.55, 59.24, 54.09]	[59.23, 42.71, 58.79, 59.00]	[64.60, 41.17, 64.10, 59.30]	[65.65, 63.81, 53.54, 68.98]
2	[69.23, 61.04, 56.00, 59.30]	[64.10, 66.58, 61.24, 49.37]	[68.90, 48.89, 54.12, 53.30]	[65.89, 61.80, 61.80, 57.57]
3	[65.12, 60.55, 56.34, 56.57]	[64.10, 66.08, 51.67, 64.51]	[67.17, 73.86, 65.90, 49.80]	[66.67, 58.04, 58.13, 66.50]
4	[56.39, 41.70, 54.12, 56.32]	[65.10, 41.70, 59.46, 52.60]	[69.70, 48.40, 57.24, 53.59]	[67.44, 52.01, 52.56, 61.76]
5	[59.74, 45.47, 51.67, 48.63]	[66.40, 60.00, 64.36, 52.60]	[65.00, 48.40, 60.00, 57.00]	[66.41, 45.97, 53.90, 56.32]

Table 1. Results for all five seeds used for both seen (*top*) and unseen (*bottom*) object scenarios in our four physical interaction outcome prediction tasks.

Name	URL	License
ConvLSTM [3]	`https://github.com/xibinyue/ConvLSTM-1`	GNU General Publics License v3.0
3D ResNet [2]	`https://github.com/kenshohara/3D-ResNets-PyTorch`	MIT License
SPACE [1]	`https://github.com/jiafei1224/SPACE`	GNU General Publics License v3.0

Table 2. Table of open-source code used.

validation and test set each. For the combined tasks, each of the splits has equal numbers of each of the three fundamental tasks to ensure that the dataset is balanced. The data distributions of the 1,000 scenarios for each task by outcome shown in Figure 1.

Physical Interactions	Scenarios	Frames
Stability	19,551	2,932,650
Contact	19,551	2,932,650
Containment	17,955	2,693,250
Total	57,057	8,558,550
Unseen Stability	3910	586,500
Unseen Contact	3910	586,500
Unseen Containment	3591	538,650
Total	11,411	1,711,650

Table 3. SPACE+ dataset analysis.

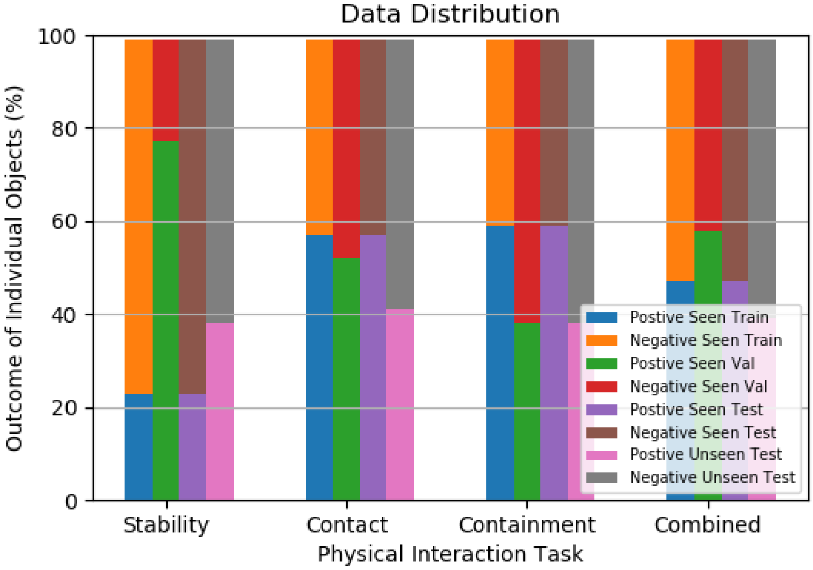

Fig. 1. Data distribution of the SPACE+ dataset used for training and testing by outcome.

2 Experimental Results and Analysis

2.1 Experimental Setup

The SPACE+ dataset is divided into stability, contact and containment tasks, and further create a new combined version that has an equal portion of each of the three fundamental tasks, resulting in four different experiments.

For each of the three fundamental tasks and the combined task, we use 1,000 scenes from the SPACE+ dataset and split it into 60% for the training set, and 20% for the validation and test set each. For the combined tasks, each of the splits has equal numbers of each of the three fundamental tasks to ensure that the dataset is balanced. For each scene, the physical interaction prediction is done for individual objects. To compare the models' performance with human performance using the same number of samples in the test set, we do not use the full SPACE+ dataset for training.

For SPACE+ with the unseen object scenarios, we also take 200 scenes for each of the three fundamental tasks and the combined task. The combined version task equal numbers of each of the three fundamental tasks.

For each scene, there is a 3-second video with a FPS of 50 to make up 150 total frames. To limit the size of the dataset to improve computational runtime, we use a frame interval of 2 where we skip 1 frame every 2 frames, resulting in 75 frames in total. Of these 75 frames, we take the first 3 frames as initial

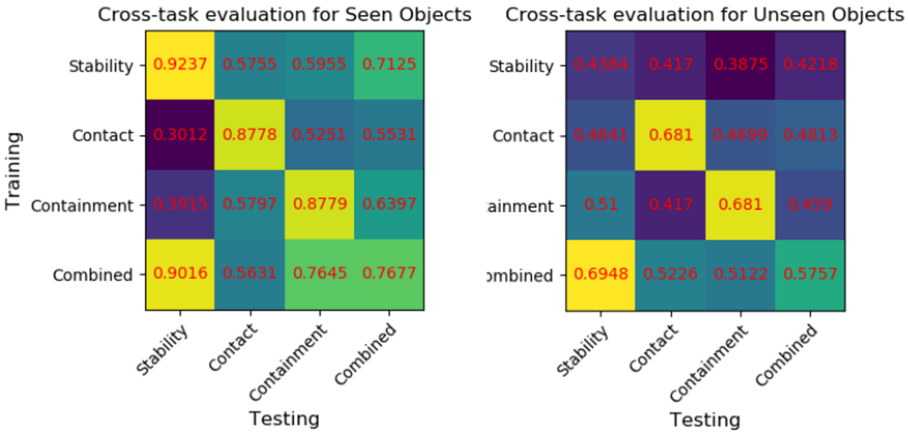

Fig. 2. (A) Average test prediction accuracy for seen objects (*left*) and unseen objects (*right*) across all models. (B) Cross-task results for both seen (*left*) and unseen (*right*) objects for one seed run.

frames to be shown to both human subjects and PIP, since there are no physical interactions among objects in these first 3 frames (i.e. first 6 frames in the original frame sequence with a frame interval of 1) in all scenes. For the ConvLSTM, we provide 37 subsequent frames from the 75 frames in addition to 3 initial frames to train it to learn future frame prediction. This is because 40 frames with a frame interval of 2 (i.e. 80 frames in the original frame sequence with a frame interval of 1) allow all outcomes of physical interactions among objects to be known, while having fewer frames reduces computational runtime. For each object in a scene, there are also 150 segmentation masks indicating its location in the 150 frames.

2.2 Evaluation Metric

To evaluate the physical interaction outcome predictions, we use classification accuracy on the test set of seen objects for both human performance and our PIP model:

$$\text{score} = \begin{cases} 1 & \text{if } \hat{y} = y \\ 0 & \text{otherwise,} \end{cases}$$

where $y \in \{0, 1\}$ is the ground-truth label. We also evaluate the performance of PIP on unseen objects.

2.3 Detailed Test Results

We present the detailed test results in Table 1 for baseline and ablation models and PIP for all the seeds used during training and testing of both seen and unseen object scenarios. A detailed analysis of all the test results are shown in Figure 2.

2.4 Span Selection Threshold

We calculate the threshold for span selection as such:

$$\mathbf{p\_threshold} = \text{cumsum}([\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}]) \in \mathbb{R}^N$$
$$\mathbf{q\_threshold} = \mathbf{p\_threshold}_{::-1}$$
$$\mathbf{r\_threshold} = \frac{\mathbf{p\_threshold} \odot \mathbf{q\_threshold}}{\Sigma_t(\mathbf{p\_threshold} \odot \mathbf{q\_threshold})_{\mathrm{t}}},$$

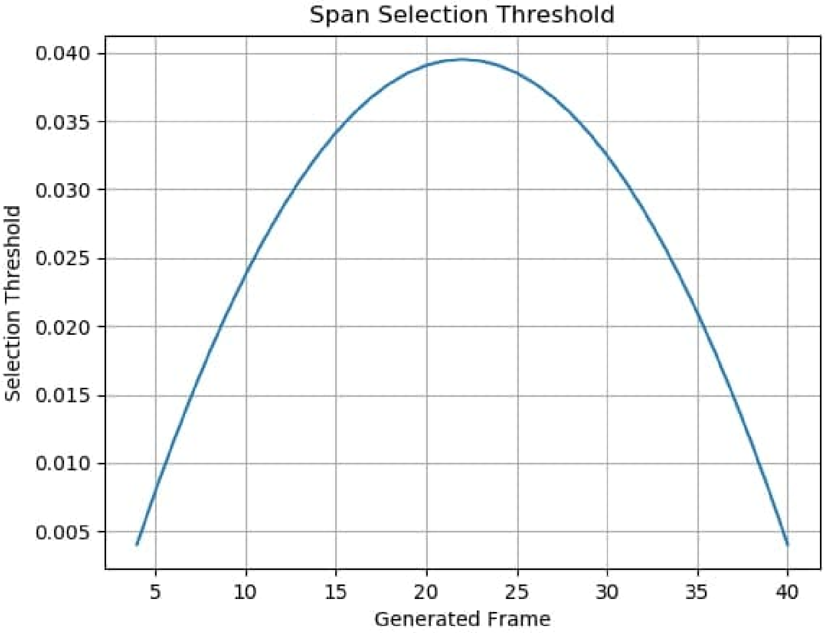

Fig. 3. Span selection threshold for 37 generated frames.

where N is the generated frame sequence length. We show in Figure 3 the span selection values for each frame for our generated frame sequence of length 37. This can be seen in Figure 3.

2.5 PIP Test Visualizations

We evaluate PIP on both seen and unseen object scenarios and extract several generations and span selection examples. The visualizations of the test results are shown in Figure 4 for seen objects and Figure 5 for unseen objects.

3 Software & Code

The open-source code we used in are shown in Table 2 and found at `https://anonymous.4open.science/r/PIP-82D6/config.yml` We use these software libraries and their versions:

- matplotlib: 3.3.4
- natsort: 7.1.1
- numpy: 1.20.2
- opencv-python: 4.5.2.54
- piqa: 1.1.3
- pytorch: 1.8.0
- scikit-image: 0.18.1
- tqdm: 4.59.0
- transformers: 4.9.2
- yaml: 0.2.5

References

1. Duan, J., Yu, S., Tan, C.: Space: A simulator for physical interactions and causal learning in 3d environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2058–2063 (2021)
2. Kataoka, H., Wakamiya, T., Hara, K., Satoh, Y.: Would mega-scale datasets further enhance spatiotemporal 3d cnns? arXiv preprint arXiv:2004.04968 (2020)
3. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. pp. 802–810 (2015)

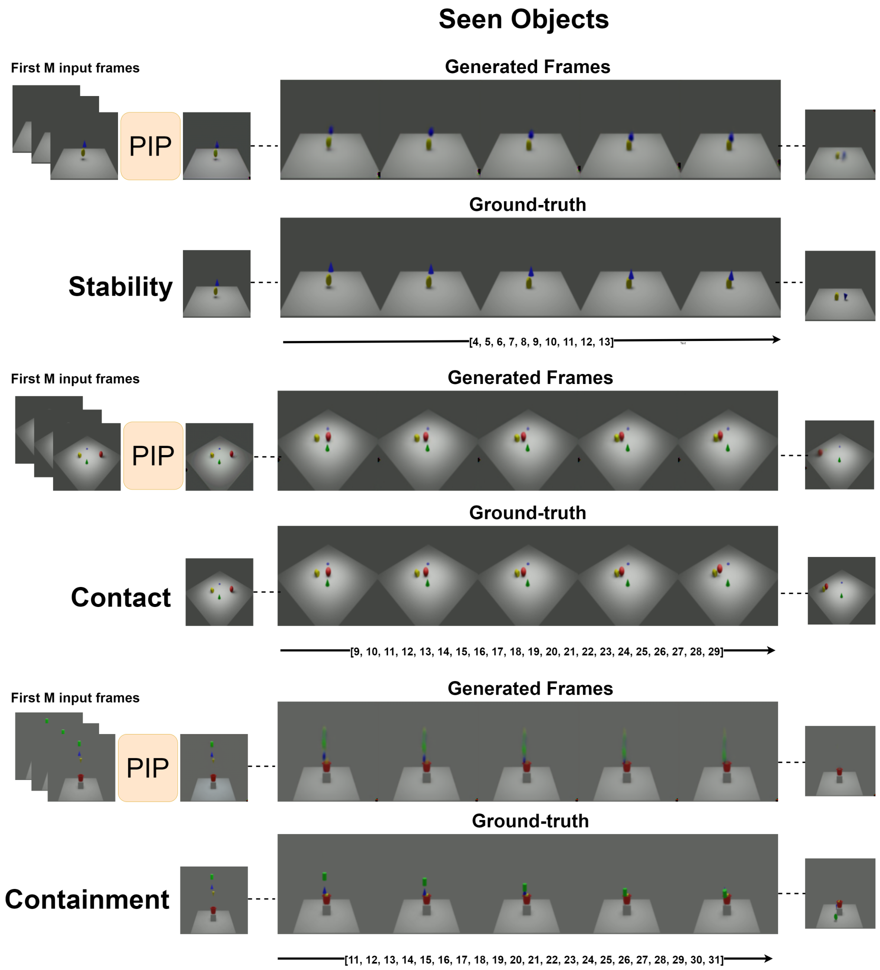

Fig. 4. Examples of generation and span selection with PIP for seen objects.

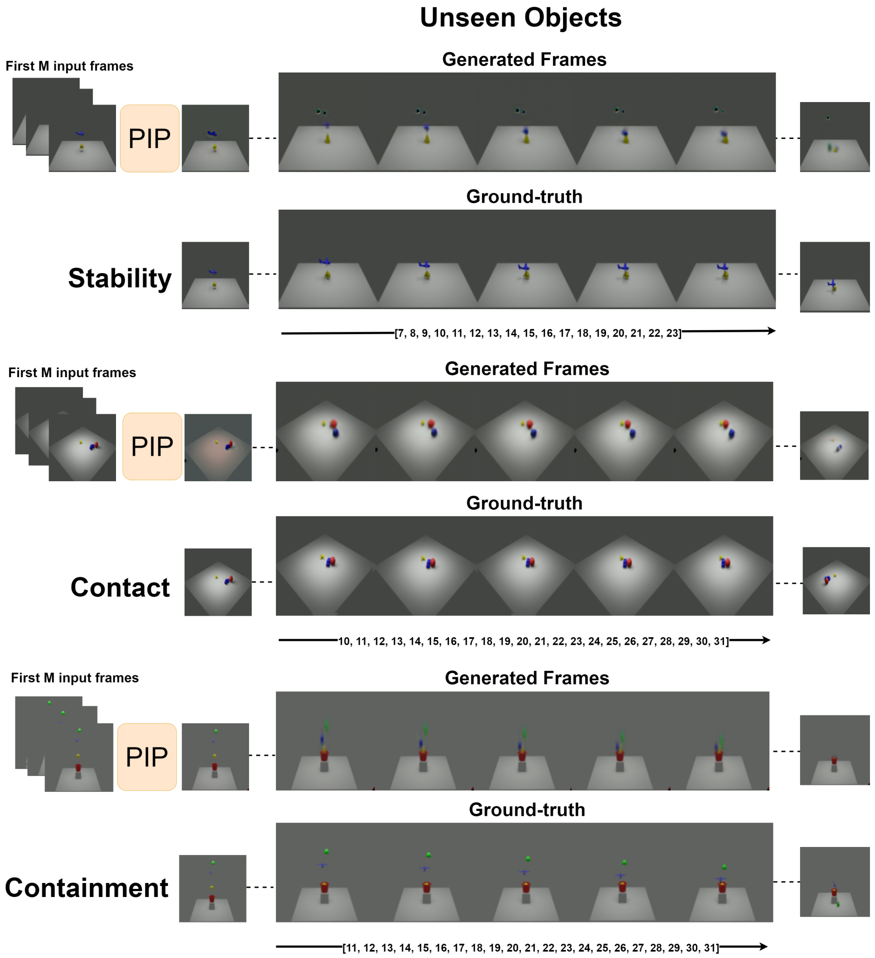

Fig. 5. Examples of generation and span selection with PIP for unseen objects.