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. Accurate prediction of physical interaction outcomes is a crucial component of human intelligence and is important for safe and efficient deployments of robots in the real world. While there are existing vision-based intuitive physics models that learn to predict physical interaction outcomes, they mostly focus on generating short sequences of future frames based on physical properties (e.g. mass, friction and velocity) extracted from visual inputs or a latent space. However, there is a lack of intuitive physics models that are tested on long physical interaction sequences with multiple interactions among different objects. We hypothesize that selective temporal attention during approximate mental simulations helps humans in physical interaction outcome prediction. With these motivations, we propose a novel scheme: **P**hysical **I**nteraction **P**rediction via Mental Simulation with Span Selection (PIP). It utilizes a deep generative model to model approximate mental simulations by generating future frames of physical interactions before employing selective temporal attention in the form of span selection for predicting physical interaction outcomes. To the best of our knowledge, attention has not been used with deep learning to tackle intuitive physics. For model evaluation, we further propose the large-scale SPACE+ dataset of synthetic videos with long sequences of three prime physical interactions in a 3D environment. Our experiments show that PIP outperforms human, baseline, and related intuitive physics models that utilize mental simulation. Furthermore, PIP's span selection module effectively identifies the frames indicating key physical interactions among objects, allowing for added interpretability, and does not require labor-intensive frame annotations. PIP is available on https://sites.google.com/view/piphysics

Keywords: Computer vision, scene understanding, physical reasoning

* Equal Contribution

1 Introduction

The ability to predict the outcomes of physical interactions among objects is a vital part of human intelligence [39,29]. Yet, it is very challenging for AI systems to acquire this ability. The key to tackling this challenge lies in understanding commonplace physical events. AI systems need to possess this ability before they can be safely and efficiently deployed in the physical world [13,57,33].

With the rapid advancements in computer vision, deep learning and embodied AI [18,4,12], there is an increase in intuitive physics models that aim to predict physical interaction outcomes. Many of these physical reasoning models [2,50,31,55,22] are inspired by *intuitive physics* in humans, which are found in cognitive science and neuroscience research [36,1,16,46,30,10]. One of the hypothesis from intuitive physics research postulates that humans predict physical interactions via the process of mental simulation. With only a few initial visual inputs of physical interaction, we can mentally reconstruct the scene with some initial approximations of the physical proprieties and dynamics of the objects. We can then predict the outcomes of physical interactions using this estimated information and the generated future visual states of objects during mental simulation. However, existing intuitive physics models are tested on short video sequences from datasets with mostly one continuous physical interaction among objects. Furthermore, it is uncertain whether humans can estimate physical properties accurately from visual inputs, and whether accurate physical property prediction is always useful for predicting physical interaction outcomes. In some cases, despite biases in estimations of physical properties [17,42,38], humans have been found to have adequately precise physical interaction outcome predictions [37]. This suggests that we might have other cognitive abilities on top of the physical property estimation that enable good physical interaction outcome prediction.

Past research has shown that humans make rational probabilistic inferences about physical interaction outcomes in a "noisy Newtonian" framework, assuming Newton's laws plus noisy observations [1]. We use noisy and approximate physical simulations to account for property, perceptual, dynamic and even collision uncertainties [1,21,24,5,35,43]. We posit that one of the beneficial cognitive abilities in humans for effective physical interaction outcome prediction is the ability to perform mental simulation with selective temporal attention to focus on physically relevant moments [15]. This might be because noisy observations and simulations are counterproductive except in moments when crucial physical interactions (e.g. collision events [35,46,21]) are present. We then posit that the selected moments in the mental simulation are used to predict the outcome.

Inspired by our hypothesis that humans use selective temporal attention in noisy mental simulations to reduce the negative effects of noise, we propose PIP, an intuitive physics model with future frame generation and span selection for predicting physical interaction outcomes. The span selection module serves as the temporal attention mechanism to focus on key physical interaction moments in the generated frames. Since state-of-the-art generative models in video generation still have artifacts and prediction errors in their generations [47,53], we

simply use the well-established convolutional LSTM (ConvLSTM) [52,23] for future frame generation to approximate noisy mental simulations as a start.

Our contributions include: (a) PIP, a novel model for effective predictions of physical interaction outcomes among objects in long sequences disjointed interactions, (b) the SPACE+ dataset, the largest synthetic video dataset with long sequences of multiple disjointed object interactions for three fundamental physical interactions (*stability, contact* and *containment*) in a 3D environment, and (c) our experiments shown that PIP outperforms intuitive physics-inspired baselines and human performance while identifying the salient moment in frames of physical interactions, which makes PIP more interpretable.

2 Related Work

Several synthetic video datasets based on fundamental physical interactions among objects in 3D environments have been developed [55,22,3,11,8] with the growing importance of physical reasoning in AI research [10]. As a result, a diverse range of intuitive physics models [6,19,31,32,14,11] were also proposed for performing physical reasoning of object interactions. However, we find Physics 101 [50], Interpretable Intuitive Physics Model [55], and PhyDNet [23] to be the most relevant to our work as their intuitive physics models were also trained on video datasets of physical interactions.

Physics 101 [50] introduced a video dataset containing over 101 real-world physical interactions of objects in four different physical scenarios. It further proposed an unsupervised representation learning model to tackle the Physics 101 dataset. The model learns directly from unlabeled videos to output the estimates of physical properties of objects, and the generative component of the model can then be used for predicting the outcomes of physical interactions.

Interpretable Intuitive Physics Model [55] proposed an encoder-decoder framework for predicting future frames of collision events. The encoder layers will extrapolate the physical properties such as mass and friction from the input frames. The decoder then disentangles latent physics vectors by outputting optical flow. For a collision event, a bilinear grid sampling layer takes the optical flow and the input frames to produce a prediction of its outcome in the form of a future frame. The dataset used for training the model is a synthetic video dataset of collision events with 11 different object combinations of 5 unique basic objects generated using the Unreal Engine 4 (UE4) game engine.

PhyDNet [23] leverages the physical knowledge extracted from partial differential equations (PDE) to improve unsupervised video prediction on videos with physical interactions and dynamics. PhyDNet does so in a two-branch approach. PhyDNet's architecture separates the PDE dynamics from unknown complementary information. PhyCell, a deep recurrent physical model, performs PDE-constrained predictions for PDE dynamics, while a ConvLSTM [52] is used to model the complementary information. PhyDNet outperforms state-of-art methods in unsupervised video prediction of physical interaction outcomes.

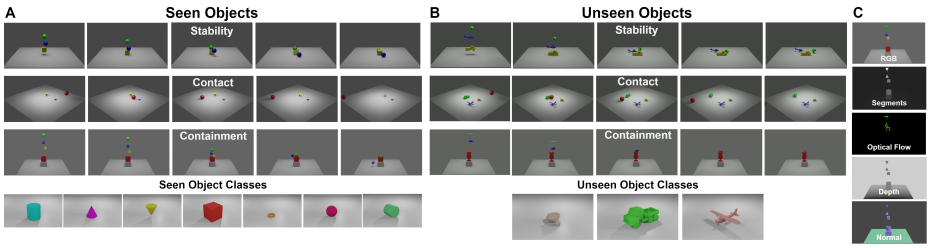

Fig. 1. Examples from SPACE+ dataset: (A) Frames of the three physical interaction tasks from the SPACE+ dataset for the seen object scenario. (B) Frames of the same tasks with new object classes for the unseen object scenario. (C) Visual information for one frame: RGB, object segmentation, optical flow, depth and surface normal vector.

The related works focus primarily on extracting physical properties of objects and dynamics from visual inputs for generating future frames, which are later used for predicting the outcome of physical interactions. In this work, we propose a new direction for mental simulation in predicting physical interaction outcomes by incorporating selective temporal attention. Our method first generates the future frames to model approximate mental simulation, then uses span selection to focus on key moments in the simulation.

3 SPACE+ Dataset

The proposed SPACE+ dataset, an improved extension of the SPACE dataset [11]. The original SPACE dataset comprises three novel video datasets synthesized by the SPACE simulator from 3D scenarios based on three fundamental physical interactions: stability, contact and containment. The SPACE dataset allows for the configuration of several parameters such as object shapes, the number of objects, object spawn locations and container types (only applicable to the containment task) during the generation.

The SPACE dataset has 15,000 unique scenarios with 5,000 scenarios for each of the three tasks. From there, 15,000 videos are generated lasting 3 seconds each at a frame rate of 50 frames per second (FPS), adding up to 2 million frames. However, there is an exception for the stability task, which is inherently unbalanced. During the stability task, for scenarios where two or three objects are spawned and land on top of each other, the objects that spawn above other objects will have higher chances of being unstable.

Our SPACE+ dataset expands and improves upon the existing SPACE dataset. Without altering the adjustable parameters, we further generate 42,057 unique scenarios on top of the original 15,000 scenarios created using the SPACE simulator. These scenarios follows the data distribution of the original SPACE dataset. We collect up to 57,057 videos with over 8 million frames in total. The overall data distribution ratio for SPACE+ is balanced with a ratio of 47:53 (positive:negative) for physical interaction outcomes, as shown in Figure 2. Beyond

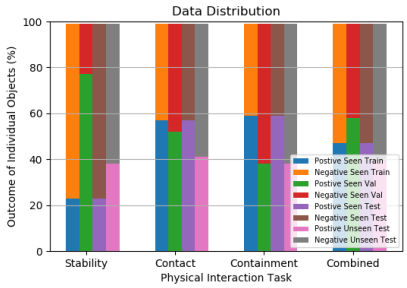

Fig. 2. Data distribution of the SPACE+ dataset used for training and testing.

scaling up the size of the dataset, we also add new object classes for all three fundamental physical interactions in the SPACE+ dataset, as shown in Figure 1B. These new object classes will be used in an unseen object scenario that will help us evaluate the generalizability of our models and human performance, since unexpected physical interactions might arise due to the new and complex shapes of the new objects. The original object classes $O = \{$*cylinder, cone, inverted cone, cube, torus, sphere, flipped cylinder*$\}$ in the SPACE dataset are shown in Figure 1A, and will be used in the seen object scenario, i.e. our models and humans will be able to train on or familiarize themselves with these object classes in the various physical interaction tasks before predicting their physical interaction outcomes in the tasks. For the SPACE+ dataset, besides the RGB frames, we also follow the SPACE paper in providing the object mask, segmentation map, optical flow map, depth map and surface normal vector map, as shown in Figure 1C. Therefore, SPACE+ is the largest dataset of its kind as we have scaled up the SPACE dataset [11] by three folds and, further, added in unseen object classes, which aims to evaluate the generalizability of the trained model for unseen object shapes.

4 PIP

As shown in Figure 3, PIP utilizes a ConvLSTM for future frame prediction to mimic noisy mental simulations and span selection to incorporate selective temporal attention. To the best of our knowledge, attention has not been used with deep learning to tackle intuitive physics tasks [10]. 2D/3D residual networks (ResNets) [26,25,27] and a pretrained BERT [9] are used to encode the necessary visual and task information for span selection [45]. PIP enables interpretability through span selection without the costly and subjective frame-based annotations needed for typical key frame selection approaches [54].

4.1 Mental Simulation

We use a ConvLSTM for future frame prediction to mimic noisy mental simulation as it is well-established and forms the backbone of recent video prediction

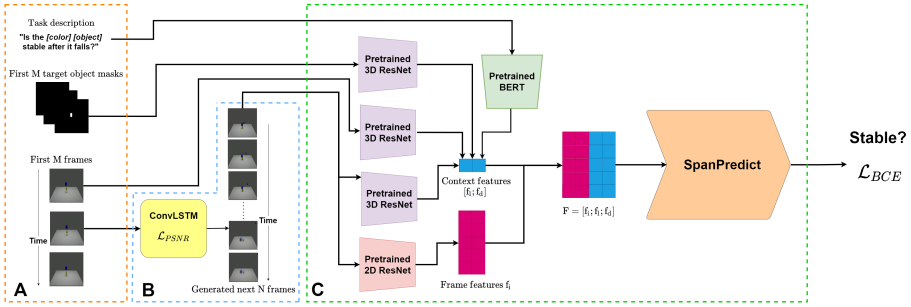

Fig. 3. PIP model architecture. (A) **Data inputs**: the original data inputs for our physical interaction prediction task comprise of the first M frames, the first M target object masks and the task description. (B) **Mental simulation**: the first M frames are fed into the mental simulation module that consists of a ConvLSTM to generate the next N frames. (C) **Span selection**: the original data inputs and the generated N frames are fed into the span selection module, where pretrained models will encode them into features before classification. All models are trained.

approaches [56,7]. The input frames are individually encoded into features with convolutional and transposed convolutional layers modelled after deep convolutional generative adversarial networks (DCGAN) [41,23] and individually fed as inputs into the ConvLSTM in sequence. We make use of teacher forcing [48] where we provide ground-truth frames to the model instead of generated frames to improve model learning. Starting from a specified frame, we train the ConvLSTM for future frame prediction.

A peak signal-to-noise ratio (PSNR) loss is used to train the convolutional layers and the ConvLSTM. The weights from each of them are shared across all three tasks in the combined training scenario.

4.2 Span Selection

PIP includes a span selection module to focus on salient frames while learning to predict physical interaction outcomes. It further allows for added interpretability by identifying the frames that are important for physical interaction prediction. Furthermore, this is done without labor-intensive frame-based annotations.

We use SpanPredict [45], a model used in natural language processing for document classification with only classification labels in the absence of ground-truth spans. Likewise, we focus on physical interaction outcome prediction in videos with only classification labels in the absence of ground-truth spans.

We modify SpanPredict to take in features for each generated frame. We obtain image features for each frame $\mathbf{f}_{i,t} \in \mathbb{R}^i$ by passing them down a pretrained 2D ResNet50. To facilitate multi-task learning in the combined task and standardize inputs for all four tasks, we encode different language features for each of the three fundamental tasks. This helps to prevent model size from increasing with the number of tasks. For the stability, contact and containment tasks, we

create the queries *"Does the [color] [object] get contacted by the red ball?"*, *"Is the [color] [object] contained within the containment holder?"* and *"Is the [color] [object] stable after it falls?"* respectively for each object in a scene. We obtain subword tokens for a query $Q = \{q_1, q_2, ..., q_n\}$ using the WordPiece tokenizer [51] and process the sequence as $[CLS]$ Q $[SEP]$, as per the standard format for single sentence inputs into BERT models. We then feed the processed sequence into a pretrained BERT model (specifically bert-base-uncased). The language features $\mathbf{f}_l \in \mathbb{R}^l$ are derived from the embeddings corresponding to the $[CLS]$ token, and are concatenated to each generated frame's feature.

In addition, for each generated frame's feature, we concatenate features $\mathbf{f}_d \in \mathbb{R}^{d\times 3}$ from the first 3 frames, the first 3 segmentation masks for the target object and all generated frames. We pass each of them through a different pretrained 3D ResNet34 to get their features. Intuitively, these 3 sources of information provide the model with prior knowledge, object tracking and global contextual information respectively. The combined features for each generated frame $\mathbf{f}_t = [\mathbf{f}_{i,t}; \mathbf{f}_l; \mathbf{f}_d]$ are stacked to form a sequence of features $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_T]$.

In the following paragraph, we will briefly explain SpanPredict [45]. We set the number of spans, and for each span we provide a pair of trainable attention weights, $\mathbf{w}_p, \mathbf{w}_q \in \mathbb{R}^{i+l+d\times 3}$. This allows for flexibility for physical interactions that might require two or more disjointed segments in a video to determine their occurrence. Using these attention weights, we get vectors $\tilde{\mathbf{p}} = \text{softmax}(\mathbf{F}^{\mathrm{T}}\mathbf{w}_{\mathrm{p}})$ and $\tilde{\mathbf{q}} = \text{softmax}(\mathbf{F}^{\mathrm{T}}\mathbf{w}_{\mathrm{q}})$, which represent the probabilities of each frame being the start and end of a salient span respectively. We then produce a span representation $\mathbf{r}$ for each span using the cumulative sum function. Firstly, we sum up the set of probabilities for each span cumulatively such that $\mathbf{p} = \text{cumsum}(\tilde{\mathbf{p}})$ and $\mathbf{q} = \text{cumsum}(\tilde{\mathbf{q}}_{::-1})$, where $\tilde{\mathbf{q}}_{::-1}$ is $\tilde{\mathbf{q}}$ with its elements reversed. Intuitively, each element in $\mathbf{p}$ and $\mathbf{q}$ represents the probability that the start of a span has occurred by that element when coming from the left of the sequence and the probability that the end of a span has occurred by that element when coming from the right of the sequence respectively. We then combine both start and end positional information as $\tilde{\mathbf{r}} = \mathbf{p} \odot \mathbf{q}$ to assign larger weights to frames that have high mass under both $\mathbf{p}$ and $\mathbf{q}$, i.e. frames that are between the start and end points. Finally, we normalize $\tilde{\mathbf{r}}$ such that its elements sum to 1: $\mathbf{r} = \frac{\mathbf{p}\odot\mathbf{q}}{\Sigma_t(\mathbf{p}\odot\mathbf{q})_t+\epsilon}$, where ϵ is a small constant. $\mathbf{r}$ gives us the final score of each frame's contribution to the span. We weigh the combined features $\mathbf{F}$ by $\mathbf{r}$, then average its values across its temporal dimension to get $\mathbf{m} = \text{average}(\mathbf{Fr}) \in \mathbb{R}^{i+l+d\times 3}$. To get each span's contribution to the final classification, we use a third attention weight $\mathbf{w}_z$ to get $\mathbf{z} = \mathbf{m}\mathbf{w}_z$, and we repeat this process for every span with the same $\mathbf{w}_z$. Finally, the contribution scores for all spans are summed up and passed through a sigmoid layer to predict $\hat{y} \in \{0, 1\}$. An additional explicit penalty is also included in the form of the generalized Jensen-Shannon divergence [34] to make the spans more concise and distinct (i.e. minimize overlapping frames for multiple span selections) [45].

5 Experiments

5.1 Experimental Setup

The SPACE+ dataset is divided into stability, contact and containment tasks, and we further create a new combined task that contains an equal number of samples from each of the three fundamental tasks.

For each of the three fundamental tasks and the combined task, we use 1,000 scenes from the SPACE+ dataset that is representative of the full dataset and split it into 60% for the training set, and 20% for the validation and test set each. For the combined task, 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, i.e. the inputs are the same across objects in the same scene except for the object masks, and the labels are different across objects. To ensure a fair comparison of the models' performance with human performance, we used the same number of samples in the test set for both and did not use the full SPACE+ dataset for training.

For the unseen object scenario introduced with SPACE+, we also take 200 scenes for each of the three fundamental tasks and the combined task. The combined task contains 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 so as 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 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. For each object in a scene, there are also 150 segmentation masks indicating its location in the 150 frames.

5.2 Evaluation Metric

To evaluate the performance of our selected methods, we use classification accuracy of physical interaction outcome predictions on the test sets of both seen and unseen object scenarios: score $= \mathbb{1}_{\hat{y}=y}$, where $y \in \{0, 1\}$ is the ground-truth label.

5.3 Human Baseline

We conduct a simple human experiment to obtain a benchmark for human performance in predicting physical interaction outcomes. Similar to other related works [22,31,32], we recruited ten participants anonymously from the internet

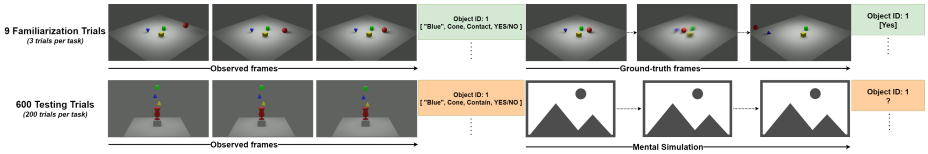

Fig. 4. Human trial setup on physical interaction prediction tasks. Trial structure for familiarization trials (*top*) and test trials (*bottom*) with the observed frames, task queries and ground-truth frames.

for the experiments. The participants first undergo a familiarization trial with nine questions (three questions for each of the three physical interaction tasks) for only the seen objects. In each familiarization trial, the participants are first shown a video with three continuous observed frames containing the initial moments of a physical interaction scenario. The participants are then asked to predict the outcome of the physical interaction by indicating either "YES" or "NO" for the specified objects. After the participants have completed indicating their prediction, they are shown the remaining parts of the video and are thus able to evaluate their predictions, as shown in Figure 4. After completing the familiarization trials, they proceed to the actual test trials beginning with the scenarios with seen objects and then with unseen objects. The test trials are similar to the familiarization trials, but the full videos are not revealed to the participants at the end of each submission. Upon completion, their results are computed, and they are informed of their task-specific accuracy and the standard deviation of their performance for both the seen and unseen object scenarios.

5.4 Baseline

We establish baseline performance by building a model similar to PIP without the mental simulation and span selection modules. This baseline model takes in the first 3 frames, the first 3 segmentation masks and the BERT language features, and encodes the frames and the segmentation masks with separate pretrained 3D ResNet34s. It then uses linear layers for classification of the concatenated features. The results, averaged across 5 runs with different seeds, are shown in Table 1 as "Baseline". We use 3D ResNets as past intuitive physics models used well-established convolutional neural networks with great success.

5.5 PhyDNet

We modify PhyDNet [23] for a performance comparison between our approach and the incorporation of physical dynamics in mental simulations. PhyDNet is a state-of-the-art generative model for predicting physical dynamics and interactions in videos. Like PIP, this modified PhyDNet model takes in the first 3 frames, the first 3 segmentation masks and the BERT language features, and encodes the frames and the segmentation masks with separate pretrained 3D

ResNet34s. However, this model generates 37 subsequent frames using the first 3 frames with the PhyDNet model instead of a ConvLSTM. It then uses another pretrained 3D ResNet34 to encode the 37 frames as features, before all the features are concatenated, and linear layers are used for classification of the combined features. The results, averaged across 5 runs with different seeds, are shown in Table 1 as "PhyDNet".

5.6 PIP

Implementation Details Models are implemented using PyTorch [40] and BERT is implemented using Hugging Face's Transformers [49]. We train using the Adam optimizer [28] for 20 epochs with a fixed learning rate of 1e-3 and a batch size of 2. We set a teacher forcing [48] rate of 0.1 for the ConvLSTM. Our ConvLSTM module consists of 3 ConvLSTM layers, 6 convolutional layers and 6 transposed convolution layers. We fine-tune the pretrained 2D ResNet50, 3D ResNet34s and BERT. We use PSNR loss to train the ConvLSTM and binary cross-entropy loss to train the entire model. To decide when to stop training, we monitor the validation classification accuracy for physical interaction outcome prediction. Our best model is selected based on its classification accuracy on the validation set of the seen object scenario. We train and test PIP over 5 runs with different random seeds and average the results. The number of spans extracted is set to 3. Experiments were run across NVIDIA GPU servers (RTX A6000, V100 and GeForce RTX 2080 Ti) and the total time for each epoch is about 2-3 hours for a total of about 40-60 hours for the entire process of 20 epochs.

Ablation Study We conduct an ablation study to examine the effect of the span selection module. Like PIP, this ablation model takes in the first 3 frames, the first 3 segmentation masks and the BERT language features, and encodes the frames and the segmentation masks with separate pretrained 3D ResNet34s. Using the first 3 frames, this ablation model also generates 37 subsequent frames with a ConvLSTM. However, it uses another pretrained 3D ResNet34 to encode the 37 frames as features, before all the features are concatenated, and linear layers are used for classification of the combined features. We use the same hyperparameters and seed runs to train and test this model as those for PIP. The results, averaged across 5 runs with different see, are shown in Table 1 as "PIP w/o SS".

6 Results and Analysis

6.1 Human Performance

Based on the results obtained from the human experiments, as shown in Table 1 and Figure 5A, we observe that human performance is consistent with an average standard deviation of 4.33 across all tasks in the seen object scenario. Hence,

	Seen Objects (%)				Unseen Objects (%)			
Methods	Stability	Contact	Containment	Combined	Stability	Contact	Containment	Combined
Human	80.24	59.54	76.19	69.54	65.88	56.07	**75.40**	61.95
Baseline	92.35	65.63	78.70	60.78	63.17	54.06	55.47	54.98
PhyDNet	**92.36**	61.23	79.47	62.03	63.78	55.41	59.10	55.61
PIP w/o SS	92.18	61.84	80.39	58.70	**67.07**	52.14	60.27	54.59
PIP (Ours)	92.33	**87.50**	**86.45**	**77.71**	66.41	**56.33**	55.99	**62.23**

Table 1. Accuracy results for seen (*left*) and unseen (*right*) object scenarios for all four physical interaction outcome prediction tasks.

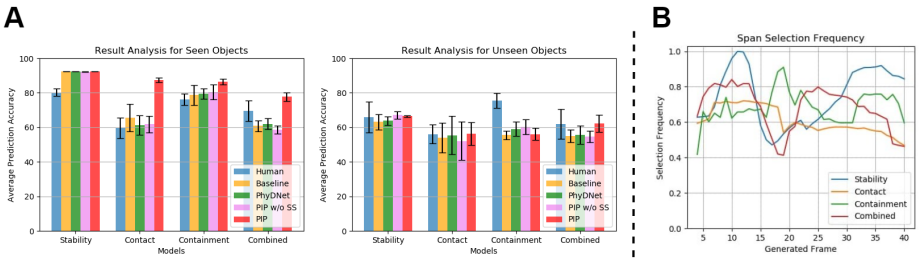

Fig. 5. (A) Average test prediction accuracy and standard deviation for seen (*left*) and unseen object (*right*) scenarios across all models and seeds. (B) PIP's frame selection frequencies on the test set for seen object scenarios across all seed runs.

the performance of ten participants is representative of the general human performance, despite the smaller sample size. Human performance is also affected by the complexity of the object shapes. Human performance has an average decrease of 6.54% from the seen object scenario to unseen object scenario across all tasks and a higher average standard deviation of 6.82, suggesting lower consistency. However, it has the lowest decrease when compared to the other models, suggesting the generalizability of human performance. Furthermore, human performance for the containment task in the unseen object scenario has the lowest decrease of 0.79% and is significantly higher than that of the other model methods by a difference of at least 15.13%. We believe this anomaly is due to the fact that humans can employ heuristics based on the estimation of physical properties (e.g. the width of the unseen object in comparison to the width of the container's entrance determines containment success) to improve predictions in the containment task. This ability is a known complement to the human mental simulation process [46].

6.2 PIP Test Performance

Based on our test results from Table 1, PIP achieve accuracies of 92.33% for *stability*, 87.50% for *contact*, 86.45% for *containment* and 77.71% for *combined* in the seen object scenario. PIP surpasses human performance by an average of 14.62% across all tasks with seen objects, the baseline model by 11.63% and the

ablation model by 12.71%. PIP also surpasses the modified PhyDNet by 16.31% excluding for the stability task, where PIP performs slightly worse. Furthermore, the average standard deviation of PIP is only 1.36, which is the lowest compared to other models and human performance. Lastly, PIP is the only model that outperforms human performance in all four tasks in the seen object scenario and three tasks in the unseen object scenario.

The results suggests that PIP is effective in predicting the outcomes of physical interactions, as it outperforms most of the models significantly in seen object scenarios with a double-digit margin for some tasks. PIP is also highly consistent in its prediction accuracy for all tasks. Moreover, through comparison of PIP with the ablation results, it can be seen that the span selection mechanism improves the prediction performance. The only anomaly is in stability performance. For the stability task, the results between the different models are relatively close. These high accuracy predictions for the stability task could indicate "shortcut learning" [20] rather than the model learning the physical understanding behind object interactions. Furthermore, the performance for the stability task is also significantly higher than the other tasks. We hypothesize that the high performance of stability is partly due to the imbalance of physical states as mentioned in Section 3. However, this imbalance in the distribution of object physical states in the stability tasks reflects an accurate representation of real-world physical dynamics where there is a higher probability of instability in multiple object scenarios. This is further supported by our analysis of the SPACE+ dataset in the supplementary material. We also evaluate the generalizability of the models by testing them on unseen object scenarios. We show in Table 1 that all the methods perform more closely in relation to one another in unseen object scenarios with an maximum difference of 7.64% except for human performance in the containment task.

6.3 Span Selection

Based on our results, we show that PIP outperforms our ablation model by a huge average margin of 12.71%. This supports our premise that span selection as a form of selective temporal attention helps mental simulations to improve predictions of physical interaction outcomes. PIP also outperforms the modified PhyDNet model by 16.31%, suggesting that PIP contributes more to mental simulations than the incorporation of learned physical dynamics. Furthermore, PIP's span selection allows us to understand how important each frame is in its contribution to physical interaction outcome predictions, providing added interpretability. This added interpretability is a significant advantage as it gives us greater insights on how to improve model performance and build trust in applications where safety is a priority.

In our experiments, it is difficult for the model to follow a strict threshold of 0 for $\mathbf{r}$ during salient frame selection. It is also difficult to use a static threshold, even if it is normalized by generated frame sequence length N (i.e. $\frac{1}{N}$), due to the way $\mathbf{r}$ is calculated. Hence, we propose a new way to calculate the threshold

for $\mathbf{r}$ as such:

$$\mathbf{p\_threshold} = \text{cumsum}([\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}]) \in \mathbb{R}^N$$

$$\mathbf{q\_threshold} = \mathbf{p\_threshold}_{::-1}$$

$$\mathbf{r\_threshold} = \frac{\mathbf{p\_threshold} \odot \mathbf{q\_threshold}}{\Sigma_t(\mathbf{p\_threshold} \odot \mathbf{q\_threshold})_t},$$

where N is the generated frame sequence length. Intuitively, this sets a uniform distribution normalized by sequence length as the threshold for $\mathbf{p}$ and $\mathbf{q}$ and calculates the threshold for $\mathbf{r}$ in the same way $\mathbf{r}$ is obtained from $\mathbf{p}$ and $\mathbf{q}$. We show in Figure 6 an example of PIP's effective selection of salient frames using this threshold calculation.

In Figure 5B, we show the frequencies of each generated frame being selected across all four tasks and all five seed runs, from frame 4 to frame 40, for the seen object scenario. Furthermore, upon inspection of the generated frames, we found that peaks in Figure 5B indicate moments of key physical interactions among objects. For example, in Figure 6, we illustrate for the stability task that frames 8-13, which corresponds to the first peak in the stability task's span selection frequencies, capture the first physical interactions among the ground and the falling object(s).

The span selection frequencies also highlight the complexity of each task. For example, the stability task's selected frames are mainly distributed into two distinct windows of frames 8-13 and 30-40 with high frequencies. This suggests that the stability task has two consistent and distinct moments of key physical interactions, which might allow for overfitting from generative models if they focus on these moments. On the other hand, for the contact task with a balanced distribution of selected frames, overfitting is more difficult. This highlights PIP's significance, since it outperforms all other models significantly for the contact task in the seen object scenario.

Finally, for the combined task, the span selection frequencies generally follow the trends of the three fundamental tasks in the first part before frame 15 with a small peak. The frequencies after frame 15 generally follow those of the contact task. More importantly, at frames 18 and 19, the frequencies decrease significantly in stark contrast to the containment task. Furthermore, the frequencies show a decreasing trend after a peak at frame 25, in contrast to the stability and containment tasks. This suggests that the combined task helps PIP to learn novel features that allow for generalizability across the three fundamental tasks. These features improve PIP's robustness, as seen in Table 1 where PIP has lowest accuracy decrease of 15.48% from the seen object scenario to the unseen scenario for the combined task, whereas there is a decrease of at least 25.92% for the other tasks.

7 Future Work

PIP currently performs worse on each of the three fundamental tasks when it is trained on the combined task than when it is both trained and tested on

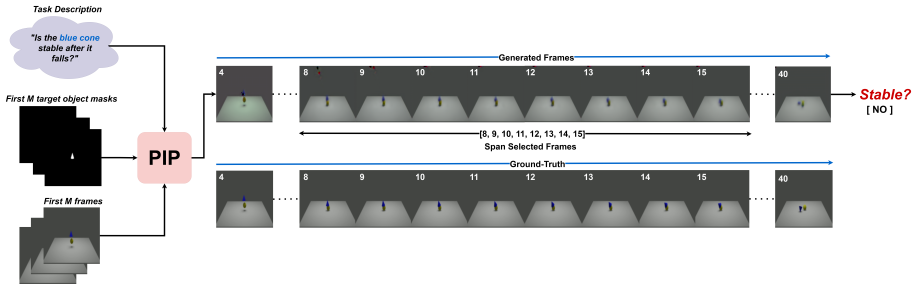

Fig. 6. An example of PIP's generation and span selection corresponding to the first window of peak span selection frequencies in the stability task. For visualizations of key physical interaction moments in the other tasks, refer to our supplementary material.

each of the three tasks. This is a common problem in multi-task learning [44]. Since PIP is only trained on a small number of samples for each of the three tasks in the combined task scenario, future work can be done with more data to improve PIP's robustness in multi-task settings. Furthermore, we assume that the generation artifacts and errors in the ConvLSTM's generations accurately model "noisy Newtonian" dynamics in human mental simulations. Future work can be done to better model "noisy Newtonian" dynamics in our model's mental simulations by investigating how the type of noise (e.g. disappearing objects, wrong trajectories), the variation in different physical properties (e.g. size and shape) and more constrained setups (e.g. attention on objects before generation) affect deep learning performance for the different physical interaction tasks.

8 Conclusion

Our ability to effectively predict the outcomes of physical interactions among objects in the real world is vital to ensure safety and success in performing complex tasks. This intuitive understanding of commonplace physical interactions is critical for complex real-world tasks such as human-robot collaboration and self-driving cars, which require reacting to ever-changing physical dynamics. In this work, we propose a new direction for intuitive physics models by proposing PIP, an intuitive physics model with selective temporal attention via span selection to improve physical interaction outcome prediction in noisy mental simulations. We evaluate PIP on the SPACE+ dataset, and show that PIP outperforms baseline and related intuitive physics models and human performance, while identifying key physical interaction moments and providing added interpretability.

Acknowledgments. This project is supported by funding allocation to C.T. by the Agency for Science, Technology and Research (A*STAR) under its SERC Central Research Fund (CRF) and its Centre for Frontier AI Research (CFAR), A*STAR's funding allocation to S.P. under its RIE 2020 AME programmatic grant RGAST2003 and project T2MOE2008 awarded by Singapore's MoE under its Tier-2 grant scheme.

References

1. Battaglia, P.W., Hamrick, J.B., Tenenbaum, J.B.: Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences **110**(45), 18327–18332 (2013)
2. Battaglia, P.W., Pascanu, R., Lai, M., Rezende, D., Kavukcuoglu, K.: Interaction networks for learning about objects, relations and physics. arXiv preprint arXiv:1612.00222 (2016)
3. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Fei-Fei, L., et al.: Physion: Evaluating physical prediction from vision in humans and machines. arXiv preprint arXiv:2106.08261 (2021)
4. Bengio, Y., Lecun, Y., Hinton, G.: Deep learning for ai. Communications of the ACM **64**(7), 58–65 (2021)
5. Bramley, N.R., Gerstenberg, T., Tenenbaum, J.B., Gureckis, T.M.: Intuitive experimentation in the physical world. Cognitive Psychology **105**, 9–38 (2018)
6. Brubaker, M.A., Sigal, L., Fleet, D.J.: Estimating contact dynamics. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 2389–2396. IEEE (2009)
7. Chai, Z., Yuan, C., Lin, Z., Bai, Y.: Cms-lstm: Context-embedding and multi-scale spatiotemporal-expression lstm for video prediction. arXiv preprint arXiv:2102.03586 (2021)
8. Dasgupta, A., Duan, J., Ang Jr, M.H., Tan, C.: Avoe: A synthetic 3d dataset on understanding violation of expectation for artificial cognition. arXiv preprint arXiv:2110.05836 (2021)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Duan, J., Dasgupta, A., Fischer, J., Tan, C.: A survey on machine learning approaches for modelling intuitive physics. arXiv preprint arXiv:2202.06481 (2022)
11. 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)
12. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A survey of embodied ai: From simulators to research tasks. arXiv preprint arXiv:2103.04918 (2021)
13. Duchaine, V., Gosselin, C.: Safe, stable and intuitive control for physical human-robot interaction. In: 2009 IEEE International Conference on Robotics and Automation. pp. 3383–3388. IEEE (2009)
14. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. Advances in neural information processing systems **29**, 64–72 (2016)
15. Firestone, C., Scholl, B.: Seeing stability: Intuitive physics automatically guides selective attention. Journal of Vision **16**(12), 689–689 (2016)
16. Fischer, J., Mikhael, J.G., Tenenbaum, J.B., Kanwisher, N.: Functional neuroanatomy of intuitive physical inference. Proceedings of the national academy of sciences **113**(34), E5072–E5081 (2016)
17. Fleming, R.W.: Visual perception of materials and their properties. Vision research **94**, 62–75 (2014)
18. Forsyth, D., Ponce, J.: Computer vision: A modern approach. Prentice hall (2011)
19. Fragkiadaki, K., Agrawal, P., Levine, S., Malik, J.: Learning visual predictive models of physics for playing billiards. arXiv preprint arXiv:1511.07404 (2015)

20. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
21. Gerstenberg, T., Tenenbaum, J.B.: Intuitive theories. Oxford handbook of causal reasoning pp. 515–548 (2017)
22. Groth, O., Fuchs, F.B., Posner, I., Vedaldi, A.: Shapestacks: Learning vision-based physical intuition for generalised object stacking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 702–717 (2018)
23. Guen, V.L., Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11474–11484 (2020)
24. Hamrick, J.B., Smith, K.A., Griffiths, T.L., Vul, E.: Think again? the amount of mental simulation tracks uncertainty in the outcome. Cognitive Science (2015)
25. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
27. Kataoka, H., Wakamiya, T., Hara, K., Satoh, Y.: Would mega-scale datasets further enhance spatiotemporal 3d cnns? arXiv preprint arXiv:2004.04968 (2020)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
29. Kubricht, J.R., Holyoak, K.J., Lu, H.: Intuitive physics: Current research and controversies. Trends in cognitive sciences **21**(10), 749–759 (2017)
30. Kubricht, J.R., Holyoak, K.J., Lu, H.: Intuitive physics: Current research and controversies. Trends in cognitive sciences **21**(10), 749–759 (2017)
31. Lerer, A., Gross, S., Fergus, R.: Learning physical intuition of block towers by example. In: International conference on machine learning. pp. 430–438. PMLR (2016)
32. Li, W., Azimi, S., Leonardis, A., Fritz, M.: To fall or not to fall: A visual approach to physical stability prediction. arXiv preprint arXiv:1604.00066 (2016)
33. Li, W., Leonardis, A., Fritz, M.: Visual stability prediction for robotic manipulation. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 2606–2613. IEEE (2017)
34. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991). https://doi.org/10.1109/18.61115
35. Ludwin-Peery, E., Bramley, N.R., Davis, E., Gureckis, T.M.: Limits on simulation approaches in intuitive physics. Cognitive Psychology **127**, 101396 (2021). https://doi.org/https://doi.org/10.1016/j.cogpsych.2021.101396, `https://www.sciencedirect.com/science/article/pii/S0010028521000190`
36. McCloskey, M.: Intuitive physics. Scientific american **248**(4), 122–131 (1983)
37. Mitko, A., Fischer, J.: When it all falls down: the relationship between intuitive physics and spatial cognition. Cognitive research: principles and implications **5**, 1–13 (2020)
38. Mitko, A., Fischer, J.: A striking take on mass inferences from collisions. Journal of Vision **21**(9), 2812–2812 (2021)
39. Moore, D.S., Johnson, S.P.: Mental rotation in human infants: A sex difference. Psychological science **19**(11), 1063–1066 (2008)

40. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
41. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
42. Rossi, F., Montanaro, E., de'Sperati, C.: Speed biases with real-life video clips. Frontiers in integrative neuroscience **12**, 11 (2018)
43. Smith, K.A., Vul, E.: Sources of uncertainty in intuitive physics. Topics in cognitive science **5**(1), 185–199 (2013)
44. Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., Savarese, S.: Which tasks should be learned together in multi-task learning? In: International Conference on Machine Learning. pp. 9120–9132. PMLR (2020)
45. Subramanian, V., Engelhard, M., Berchuck, S., Chen, L., Henao, R., Carin, L.: Spanpredict: Extraction of predictive document spans with neural attention. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5234–5258 (2021)
46. Ullman, T.D., Spelke, E., Battaglia, P., Tenenbaum, J.B.: Mind games: Game engines as an architecture for intuitive physics. Trends in cognitive sciences **21**(9), 649–665 (2017)
47. Weissenborn, D., Täckström, O., Uszkoreit, J.: Scaling autoregressive video models. arXiv preprint arXiv:1906.02634 (2019)
48. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Computation **1**(2), 270–280 (1989). https://doi.org/10.1162/neco.1989.1.2.270
49. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
50. Wu, J., Lim, J.J., Zhang, H., Tenenbaum, J.B., Freeman, W.T.: Physics 101: Learning physical object properties from unlabeled videos. In: BMVC. vol. 2, p. 7 (2016)
51. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
52. 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)
53. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
54. Yan, X., Gilani, S.Z., Feng, M., Zhang, L., Qin, H., Mian, A.: Self-supervised learning to detect key frames in videos. Sensors **20**(23) (2020). https://doi.org/10.3390/s20236941, https://www.mdpi.com/1424-8220/20/23/6941

55. Ye, T., Wang, X., Davidson, J., Gupta, A.: Interpretable intuitive physics model. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 87–102 (2018)
56. Zhang, L., Lu, L., Wang, X., Zhu, R.M., Bagheri, M., Summers, R.M., Yao, J.: Spatio-temporal convolutional lstms for tumor growth prediction by learning 4d longitudinal patient data. IEEE transactions on medical imaging **39**(4), 1114–1126 (2019)
57. Zheng, B., Zhao, Y., Yu, J., Ikeuchi, K., Zhu, S.C.: Scene understanding by reasoning stability and safety. International Journal of Computer Vision **112**(2), 221–238 (2015)