Intrinsic Physical Concepts Discovery with Object-Centric Predictive Models

Qu Tang [1,2⋆] Xiangyu Zhu [1,2⋆] Zhen Lei [1,2,3] Zhaoxiang Zhang [1,2,3∗]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]MAIS, Institute of Automation, Chinese Academy of Sciences
[3]Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences
{tangqu2020,zhaoxiang.zhang}@ia.ac.cn, {xiangyu.zhu,zlei}@nlpr.ia.ac.cn

1. Training Configuration

1.1. Data Processing

Each data point in ComPhy contains 4 reference videos and 1 target video. Within each set, the objects share the same intrinsic physical properties across all videos and each object in the target video appears at least in one of the reference videos. Reference videos have more interaction among objects, providing information for models to identify objects' physical properties. We train PHYCINE firstly in reference videos, then fine-tune the model in target videos. Each reference video contains 50 frames, and we sample images every 4 frames. Data balance is also important to promote the convergence of the relation model. Following the training strategy in Section 3.6, we balance the data by the object's physical properties in the scene. In detail, we categorize the reference videos into 5 classes as shown in table 1, and the number of videos in each category is similar. During training, sub-dataset 1 is firstly used to learn concepts of object contexts, object dynamics, and collision, then we add sub-dataset 2 to the training data to learn the concept of mass, and finally, we add the reset data to further learn the concept of charge. After all the concepts have been trained, we fine-tune the model in target videos.

Table 1. Categorized training dataset.

sub-dataset	charge	collision	identical-mass
1	56	52	52
2	56	52	56
3	52	56	
4	52	52	52
5	52	52	56

1.2. Hyper-parameters and training

We set the length of each video clip N to be 6, which means the inferred scene representation for the first frame should be able to predict the rest five frames. For the other parameters, we Generally follow the setting of IODINE. We initialize the parameters of the posterior λ by sampling from $U(-0.5, 0.5)$. We use a latent dimensionality of 16 as ALOE which makes dim(λ) = 32 and downscale the image from 320×480 into 64×96. The variance of the likelihood is set to σ = 0.3 in all experiments. We keep the default number of iterative refinements at R = 5 and use K = 8 for both training and testing. We set β = 100.0 for all experiments. We train our models on 8 GeForce RTX 3090 GPUs, which takes approximately 4 days per model. We use ADAM for all experiments, with a learning rate of 0.0003.

2. Necessity of regularization

We discuss the necessity of bottom-up training and variable content reduction with qualitative results. In Figure 1, one collision event occurs between the cylinder and the sphere. From the regeneration results, we can tell that PHYCINE without bottom-up training can not capture the collision but records the trajectories of objects. In addition, for the counterfactual setting "if the sphere is heavier", PHYCINE predicts a likely result: trajectories of both the cylinder and the sphere are affected because of this setting, and the first two frames are not affected because the collision does not happen yet. By contrast, the result of PHYCINE without bottom-up training shows that the model didn't really learn the concept of "mass", because only the sphere is affected.

In order to demonstrate the necessity of reducing variable contents, we conduct an experiment with higher dimension dynamic variables and without interaction modeling (with an FC layer only), Figure. 2 shows that the model can roll out videos with objects interacting even without the interaction model, suggesting that the higher-dimensional dyn learns redundant information (such as change of velocities directions) that should not be entangled and the model

∗Corresponding author. ⋆Equal contribution

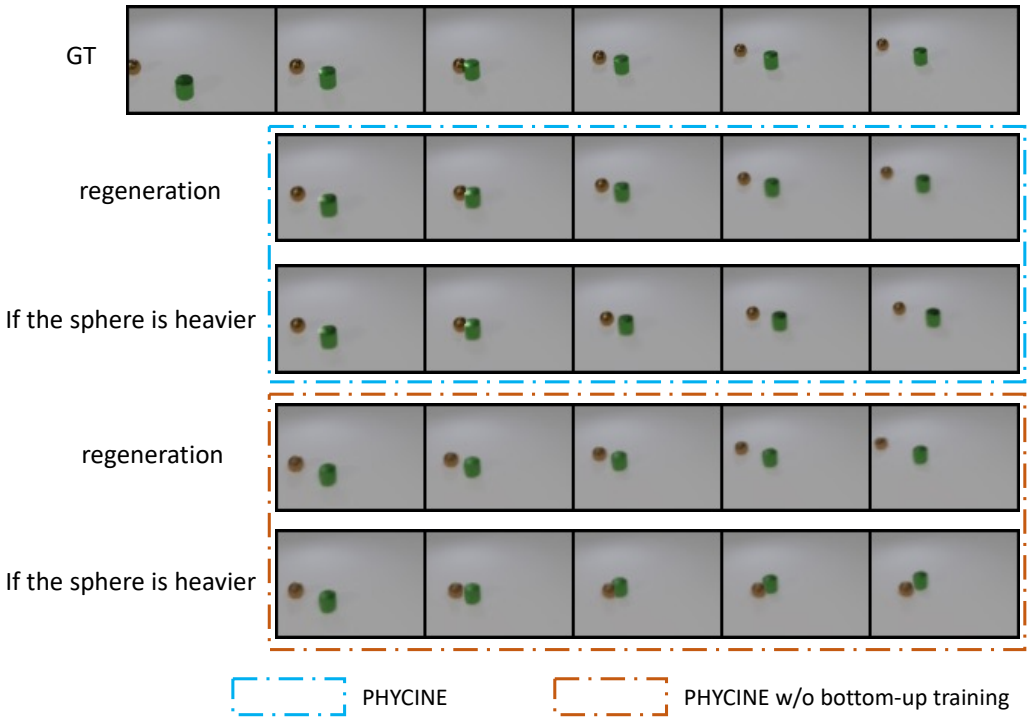

Figure 1. Visualization comparison between PYHCINE and PHYCINE without bottom-up training. Both regeneration and counterfactual results are shown.

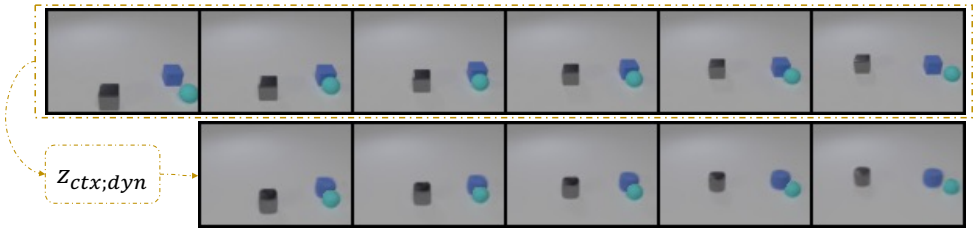

Figure 2. Prediction with higher dimension dynamics.

collapse.

3. Visualizations and numbers of intermediate variables

We visualize the intermediate variables from the force modeling process of the example in Figure. 3 and Table. 2. In frame 3, the two objects were subjected to forces of opposite directions, the force intensities are determined by both dynamics and mass. Their dynamics are changed consequently in frame 5.

F	dyn	dyn	f_d	f_d	f_i	f_i	m	m
1	-0.74 -1.65	2.21 -0.02	0.92 -0.39	0.83 -0.56	0.01	0.01	-4.83	-0.15
3	-0.66 -1.64	2.20 -0.04	0.99 0.16	-0.99 -0.12	3.67	0.87	-4.83	-0.15
5	3.12 -1.03	1.25 -0.17	-0.26 0.96	-0.96 -0.27	0.13	0.03	-4.83	-0.15

Table 2. Variable numbers visualization. Frame (F) 1, 3, and 5 are sampled, with f_d and f_i indicating the force direction and the force intensity applied respectively.

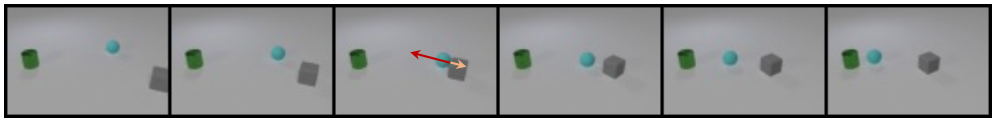

Figure 3. Example for intermediate variable visualization.