Supplementary Materials for On the Learning Mechanisms in Physical Reasoning

Shiqian Li[⋆,1,2,5], **Kewen Wu**[⋆,3,5], **Chi Zhang**[4,5✉], **Yixin Zhu**[1,2✉]

[1] School of Intelligence Science and Technology, Peking University
[2] Institute for Artificial Intelligence, Peking University
[3] Department of Automation, Tsinghua University
[4] Department of Computer Science, University of California, Los Angeles
[5] Beijing Institute for General Artificial Intelligence (BIGAI)

Project Website https://lishiqianhugh.github.io/LfID_Page

A Training Details

We run all our experiments on either NVIDIA A100 80GB or RTX 3090 GPUs. Training details for different settings are specified below.

Learning from Intuition (LfI) We train ViT, Swin Transformer, and BEiT using the same setting. With a balanced number of successful and failed samples, each model takes a batch of 224×224 images with different actions from each task for training. We fine-tune these three pre-trained LfI models for 10 epochs, annealing the learning rate from 1×10^{-4} to 1×10^{-6} using a cosine schedule. The model parameters are optimized using Adam with the binary cross-entropy loss.

Learning from Dynamics (LfD) under Ground-truth Dynamics (GD) We extract ground-truth sequences of lengths 1, 2, 4, and 8 from PHYRE's simulator with a time interval of 1 second. We pad them with the last frame for sequences with a total length shorter than 8. We fine-tune the TimeSformer with the same setting for a fair comparison. Specifically, we tune the TimeSformer pre-trained on Kinetics-600 with a standard input sequence of eight 224×224 images. Similar to LfI, we train the models for 10 epochs, annealing the learning rate from 1×10^{-4} to 1×10^{-6} using a cosine scheduler. The model parameters are also optimized using Adam with the binary cross-entropy loss.

LfD under Approximate Dynamics (AD) For both the serial and parallel optimization schedules, we train the dynamics prediction model PredRNN and the task-solution model TimeSformer using the same number of images. We use the PredRNN based on Memory-Decoupled ST-LSTM as the dynamics predictor. During training, we first reshape the raw images from $224 \times 224 \times 3$ into $28 \times 28 \times 192$. Next, we call the Reverse Scheduled Sampling method to generate input flags to gradually change the training process from using the synthesized frames to using the ground truth. Finally, the initial images and the input flags are fed into the model. PredRNN's output is reshaped back to $224 \times 224 \times 3$ before being sent into TimeSformer as AD for final prediction. The parameters of PredRNN and TimeSformer are optimized using the same training setting as in LfI in both optimization schedules. We set α and β to 1 in the parallel optimization schedule.

⋆ indicates equal contribution.
✉ indicates corresponding authors.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

B Additional Visualizations

We visualize additional dynamics prediction and solution set prediction in Fig. A1 to A4.

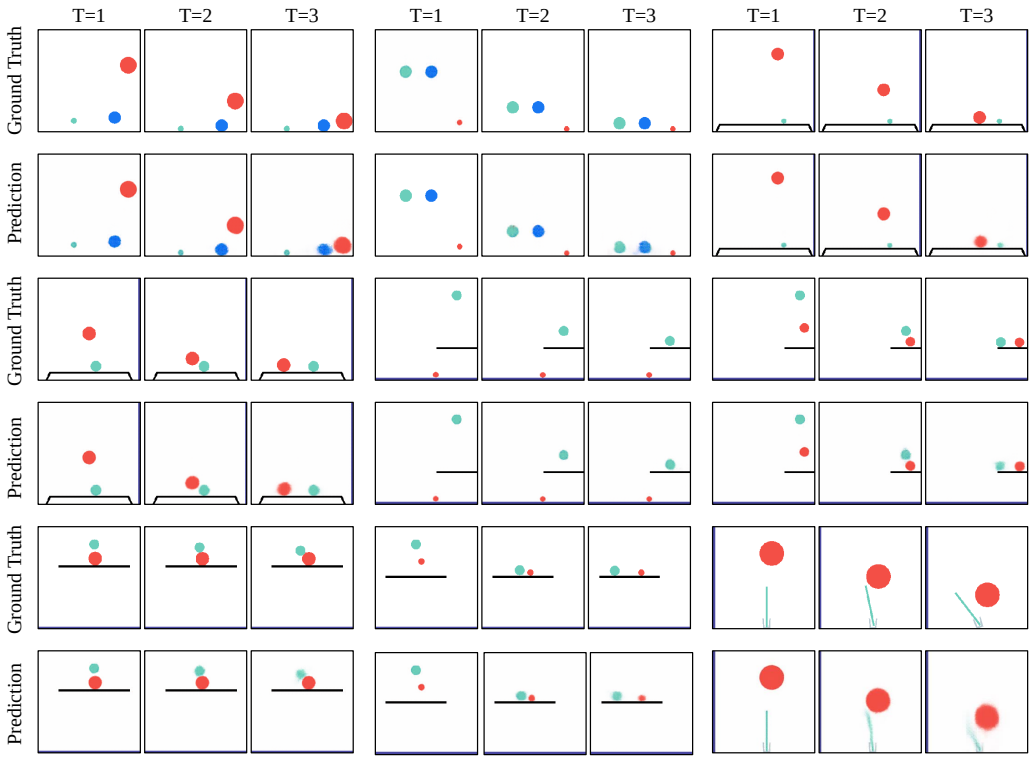

Figure A1: **Predicted dynamics from PredRNN in LfD's serial optimization schedule in easy tasks.**

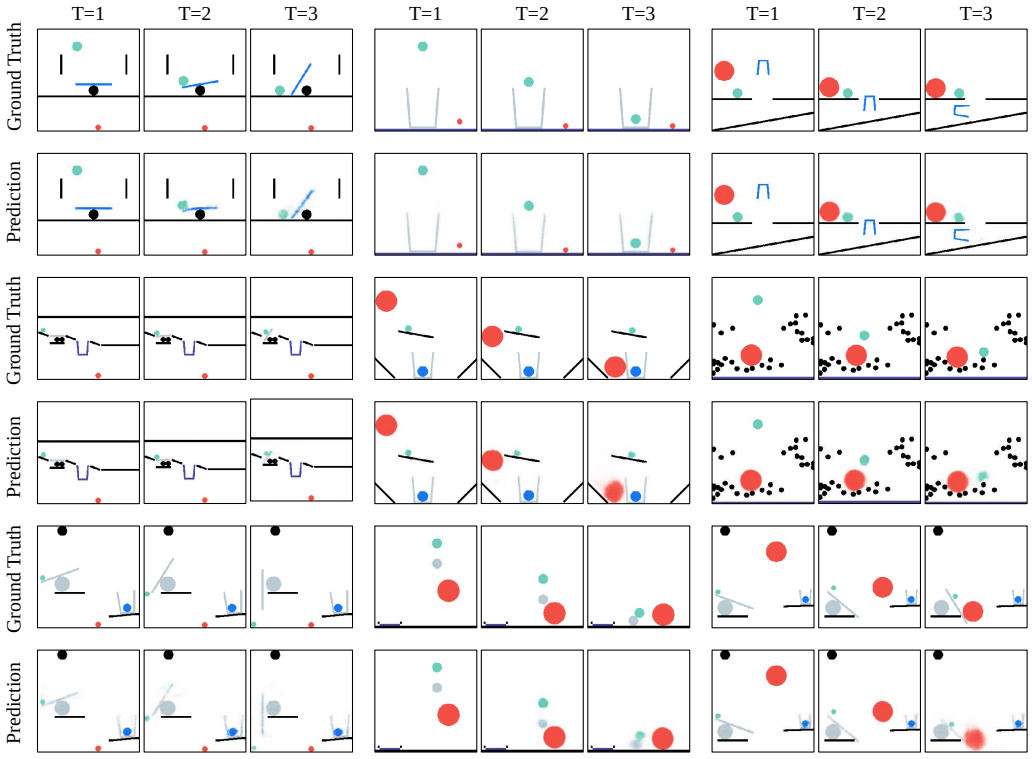

Figure A2: **Predicted dynamics from PredRNN in LfD's serial optimization schedule in hard tasks.**

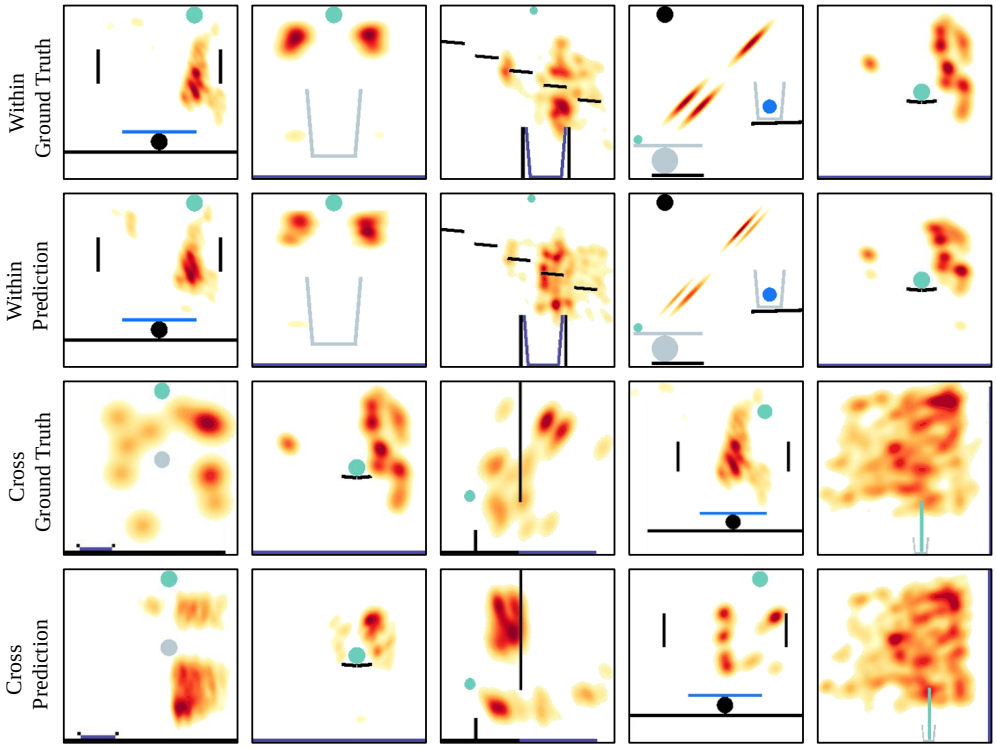

Figure A3: **The ground-truth $\mathbf{P}(\mathbf{y}|\mathbf{X_0})$ distribution heat maps and the ones predicted by Swin Transformer in PHYRE-B.** The heat maps are generated in the same way as in ViT.

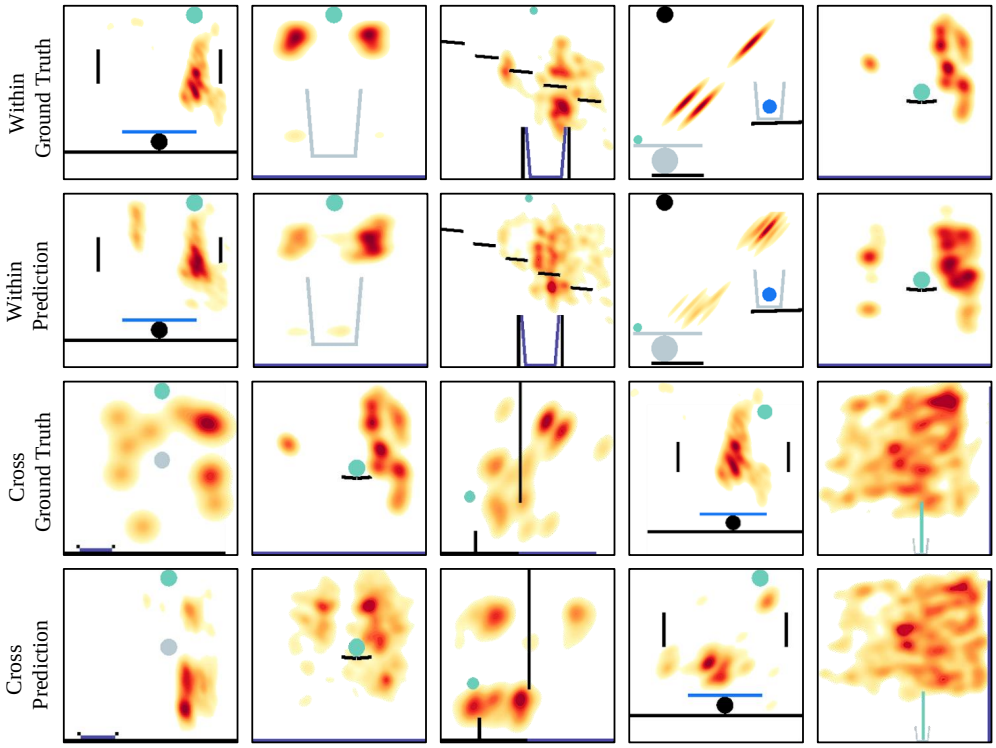

Figure A4: **The ground-truth $\mathbf{P}(\mathbf{y}|\mathbf{X_0})$ distribution heat maps and the ones predicted by BEiT in PHYRE-B.** The heat maps are generated in the same way as in ViT.