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

Abstract

Is dynamics prediction indispensable for physical reasoning? If so, what kind of roles do the dynamics prediction modules play during the physical reasoning process? Most studies focus on designing dynamics prediction networks and treating physical reasoning as a downstream task without investigating the questions above, taking for granted that the designed dynamics prediction would undoubtedly help the reasoning process. In this work, we take a closer look at this assumption, exploring this fundamental hypothesis by comparing two learning mechanisms: Learning from Dynamics (LfD) and Learning from Intuition (LfI). In the **first experiment**, we directly examine and compare these two mechanisms. Results show a surprising finding: Simple LfI is better than or on par with state-of-the-art LfD. This observation leads to the **second experiment** with Ground-truth Dynamics (GD), the ideal case of LfD wherein dynamics are obtained directly from a simulator. Results show that dynamics, if directly given instead of approximated, would achieve much higher performance than LfI alone on physical reasoning; this essentially serves as the performance upper bound. Yet practically, LfD mechanism can only predict Approximate Dynamics (AD) using dynamics learning modules that mimic the physical laws, making the following downstream physical reasoning modules degenerate into the LfI paradigm; see the **third experiment**. We note that this issue is hard to mitigate, as dynamics prediction errors inevitably accumulate in the long horizon. Finally, in the **fourth experiment**, we note that LfI, the extremely simpler strategy when done right, is more effective in learning to solve physical reasoning problems. Taken together, the results on the challenging benchmark of PHYRE [3] show that LfI is, if not better, as good as LfD with bells and whistles for dynamics prediction. However, the potential improvement from LfD, though challenging, remains lucrative.

1 Introduction

Humans possess a distinctive ability of understanding physical concepts and performing complex physical reasoning. The literature on humans learning mechanisms for solving physical reasoning problems can be categorized into two schools of thought [43]: (i) physical intuition at a glance without much thinking, such as judging whether a stacked block tower will collapse [6], and (ii) more

⋆ indicates equal contribution.
✉ indicates corresponding authors.

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

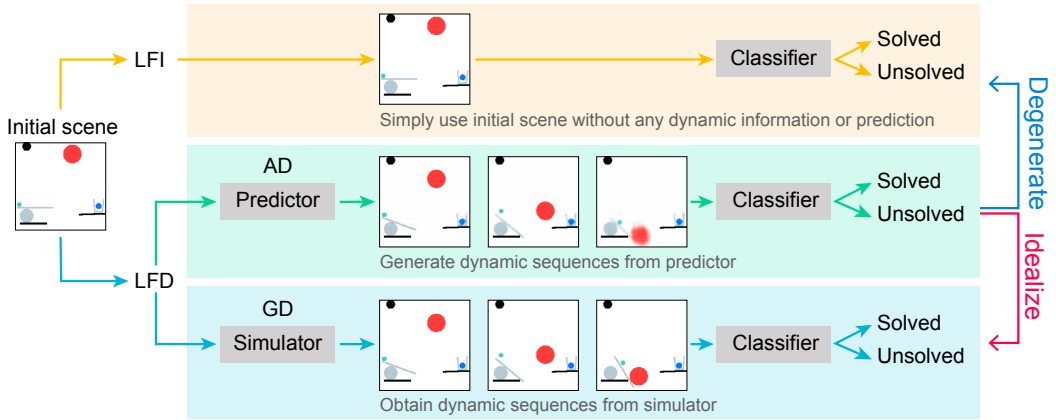

Figure 1: **Comparison of the two learning mechanisms.** Learning from Intuition (LfI) learns from intuition by directly using a classifier (first row in yellow). Based on the source of dynamics, we divide Learning from Dynamics (LfD) into learning with Approximate Dynamics (AD) (second row in green) and Ground-truth Dynamics (GD) (third row in blue). Specifically, learning with GD leverages ground-truth dynamics from the simulator, whereas learning with AD predicts how the objects' positions and poses unfold via a dynamics predictor. In theory, learning with AD should reach the same performance as learning with GD if dynamics prediction were perfect; *i.e.*, learning with GD can be regarded as the ideal case of learning with AD. However, in practice, learning with AD usually degenerates into LfI due to very inaccurate prediction.

extensive unfolding of states under the assumed physical dynamics when facing complex physical tasks [7]. Such a disparity between System-1-and-System-2-like problem-solving strategies [36] motivates us to think over the learning mechanisms for physical reasoning in machines:

> *When performing physical reasoning, is it better for machines to learn from intuition by simply analyzing the static physical structure, or to learn from dynamics by predicting future states?*

Recent physical reasoning benchmarks for machine learning are mostly physics engines focusing on evaluating the task-solving abilities of models. For example, PHYRE [3] and Virtual Tools [1] contain physical scenes with long-term dynamics and complex physics interactions. These environments have various tasks with explicit goals, such as making the green object touch the blue object by *placing a new red ball* into the initial scene. An artificial agent is tasked to predict the final outcome, *e.g.*, whether the placed *red ball* will successfully solve the given task.

Existing problem-solving methods approach such physical reasoning problems by designing various future prediction modules [23, 53]. These modules are devised under the assumption that human brains inherently possess a simplified physics engine (called intuitive physics engine) [18], akin to a computer game simulator, capable of predicting objects' future states and changes.

Although the intuitive theory claims that humans can predict physical outcomes rapidly, it does not directly guide at the computational level more than the hypothesis that we might have a physics-engine-like mechanism in our brain [18, 43]. Critically, although humans can predict dynamics under the intuitive theory, dynamics prediction might not be necessary for all types of physical reasoning tasks. This hypothesis is largely left untouched, especially at the computational level. Just as noted by Lerer et al. [46], directly learning by intuition without dynamics prediction is sufficient for various physical reasoning tasks. Of note, this hypothesis does not contradict the intuitive theory but rather provides a new perspective at the computational level.

In this paper, we conduct a series of experiments to answer the above questions empirically. To the best of our knowledge, ours is the first to **systematically** compare the LfI and LfD paradigms. In the **first experiment**, we verify the simple approach of LfI by training a classifier [15] to predict whether an action would lead to success in problem-solving. Surprisingly, in the preliminary study, such a model already reaches the state-of-the-art (SOTA) performance and even outperforms existing LfD methods in unseen scenarios, indicating better generalization. Inspired by this counter-intuitive result, we conduct more experiments on the two learning mechanisms; see Fig. 1 for an illustration. In the **second experiment**, we first set out to investigate whether LfD could work better than LfI in theory by measuring the performance of a video-based classifier [8] using the ground-truth dynamics,

setting an upper bound for the route of dynamics prediction. In the **third experiment**, we replace the ground-truth dynamics with predicted dynamics from an advanced future prediction model [62] to see how LfD performs in practice. The results from these two experiments reveal that precise dynamics significantly improve model performance, but the predicted approximations fail to work as expected: Approximation brings disturbing biases and causes the performance to degenerate into LfI or even worse in certain tasks. In the **fourth experiment**, through a series of experiments on various LfI models, we conclude that LfI could be a simpler and more practical paradigm in physical reasoning. However, making breakthroughs in physical prediction is still promising though challenging. We hope that our discussions shed light on future studies on physical reasoning.

2 Related Work

Intuitive physics and physical reasoning Since Battaglia et al. [6], the computational aspects of intuitive physics have attracted research attention [43, 75]; intuitive physics and stability has since been further incorporated in complex object [77, 64, 49, 17, 50, 72] and scene [74, 65, 73, 76, 47, 30, 9, 63], and task [26, 27, 32] understanding tasks. The progress enables machines to learn to judge (i) which object is heavier after observing two objects collide [21, 54, 59], (ii) whether a stacked block tower will fall [6, 24, 46], (iii) whether water in two different containers will pour at the same angle if titled [41, 57] or liquid in general [5], and (iv) behaviors of dynamics with various materials [42]. However, this line of work primarily focuses on physical tasks without long-term dynamics, either by knowledge-based approaches [6, 54, 70] or learning-based approaches [24, 46].

More complex physical reasoning problems [1, 3, 69], including those involving question answering [10, 11, 14, 29, 71], have also been studied. In particular, Allen et al. [1] propose to use knowledge-based simulation; Xu et al. [68] adopt a Bayesian symbolic method; Battaglia et al. [7], Girdhar et al. [23], and Qi et al. [53] recruit the graph-based interaction network. However, none of these methods fully justify the necessity of dynamics prediction. In this work, we challenge this fundamental assumption and point out a simpler and efficacious but overlooked solution.

PHYRE and relevant environments Bakhtin et al. [3] introduce the novel physical reasoning task of PHYRE, wherein an agent is tasked with finding action in an initial scene to reach the goal under physical laws. Current methods for solving PHYRE include reinforcement learning (*e.g.*, DQN [3]), forward prediction neural networks with pixel-based representation [23], and object-based representation [23, 53]. Notably, Girdhar et al. [23] adopt different kinds of forward prediction architectures to perform PHYRE tasks but fail to obtain significant performance improvement, whereas Qi et al. [53] design convolutional interaction network to learn long-term dynamics, achieving SOTA performance by leveraging ground-truth information about object states in the physical scenes.

In fact, the physical reasoning task of PHYRE can be regarded as an image classification task on judging whether an initial action would lead to a successful outcome, or a video classification task by considering the dynamics after the initial action is performed. The success of the former sat on deep convolutional neural networks [28, 40] and has now shifted to Transformer-based models [4, 15, 51]. The change from convolutional architectures to attentional models also inspires recent advances in video classification. Models such as TimeSformer [8] and ViViT [2] in this domain also expand into fields such as action recognition [22] and group activity recognition [20].

Unlike PHYRE which directly focuses on physical reasoning in a simplified virtual environment, some benchmarks include physical reasoning in their environments more implicitly and take physics as an aid to finish tasks in real life, such as the ones in autonomous driving [16] and embodied AI [39, 67, 19, 55, 48]. Robotic controller based on physics engines [25, 45, 38, 12, 60], navigation tasks on 3D physical scenes [66], and more broadly task and motion planner [61, 37, 33, 35, 34] may also need physical understanding modules in the system.

Dynamics prediction Predicting dynamics into the future is one of the most extensively studied topics in the vision community. One modern approach is to extract image representation and incorporate an RNN predictor [58, 62] or a cycle GAN-based approach [44]. However, these approaches cannot extract robust representation from the pixels and incur accumulated errors in long-term prediction. To tackle this problem, Janner et al. [31] and Qi et al. [53] focus on object-centric representation; these are task-specific solutions with various inductive biases (*e.g.*, spatial information, the number of objects), and the performance drops when dealing with multiple objects with occlusions [52].

3 The Two Learning Mechanisms

In this section, we define the two learning mechanisms for solving physical reasoning problems. Henceforth, we denote all objects' states at time t as X_t. Given an initial background image I of a physical setup and a random distribution of actions $\mathcal{A}$, the model needs to learn a distribution of the final outcome $P(y|X_0)$, where $X_0 = \{\mathcal{A}, I\}$, and y denotes the possible outcome.

Mechanism 1. Learning from Intuition (LfI) In LfI, the outcome y is learned directly from the initial images and actions using a task-solution model f:

$$P(y|X_0) = f(X_0; \theta), \tag{1}$$

where θ denotes the parameters of the task-solution model $f(\cdot)$. We call this mechanism LfI because $f(\cdot)$ can be viewed as an intuitive map from the initial conditions to the outcome.

Mechanism 2. Learning from Dynamics (LfD) The nature of physics is inherently dynamic. As such, in LfD, y is no longer directly learned from the initial scenes; instead, this approach first learns the underlying dynamics $D = \{X_t | t = 0, 1, \ldots, T\}$ within a time window T using a dynamics prediction module $g(\cdot)$, and then predicts the outcome from the predicted dynamics. Formally, the forward process is described as below:

$$P(y|X_0) = f(D; \theta), \text{ where } D = g(X_0; \phi), \tag{2}$$

where ϕ represents the parameters of the dynamics prediction model $g(\cdot)$. Usually, $g(\cdot)$ is implemented as either an auto-regressive module based on pixel presentation or a graph-based interaction network from object-based representation. In this work, we consider two optimization schedules for $g(\cdot)$: parallel optimization during joint learning of $f(\cdot)$ or serial optimization by learning and fixing $g(\cdot)$ first; please refer to Algs. 1 and 2 for more details.

Algorithm 1: Parallel optimization of LfD

Variables:
I is the initial background image. $\mathcal{A}$ is the action. $f(\cdot)$ and $g(\cdot)$ are the task-solution model and the dynamics prediction model, respectively. D and y denote predicted dynamics and outcome, and D_{gt} and y_{gt} the ground-truth ones. The dynamics loss and cross-entropy loss are denoted as $L_d(D, D_{gt})$ and $L_e(y, y_{gt})$, respectively. α and β are hyperparameters to balance the two losses.

repeat
  Predict the dynamics D from I and $\mathcal{A}$ using $g(\cdot)$;
  Predict the outcome y from D using $f(\cdot)$;
  Compute the total loss $L_{total}(D, y, D_{gt}, y_{gt}) = \alpha L_d(D, D_{gt}) + \beta L_e(y, y_{gt})$;
  Optimize $f(\cdot)$ and $g(\cdot)$ simultaneously using the gradient of the total loss $\nabla L_{total}(D, y, D_{gt}, y_{gt})$.
until max iteration

Algorithm 2: Serial optimization of LfD

Variables:
I is the initial background image. $\mathcal{A}$ is the action. $f(\cdot)$ and $g(\cdot)$ are the task-solution model and the dynamics prediction model, respectively. D and y denote predicted dynamics and outcome, and D_{gt} and y_{gt} the ground-truth ones.

repeat
  Predict the dynamics D from I and $\mathcal{A}$ using $g(\cdot)$;
  Compute the dynamics loss $L_d(D, D_{gt})$;
  Optimize $g(\cdot)$ using the gradient of dynamics loss $\nabla L_d(D, D_{gt})$;
until max iteration
Freeze $g(\cdot)$;
repeat
  Predict the dynamics D from I and $\mathcal{A}$ using the pre-trained $g(\cdot)$;
  Predict the final outcome y from D using $f(\cdot)$;
  Compute the cross-entropy loss $L_e(y, y_{gt})$;
  Optimize $f(\cdot)$ using the gradient of the cross entropy loss $\nabla L_e(y, y_{gt})$;
until max iteration

While the dynamics prediction step and the outcome prediction step can be integrated together, it is worth noting that in LfD, additional architecture changes and supervisory signals are necessary to learn the underlying dynamics, without which the paradigm degenerates into LfI. Ideally, a physics engine or a simulator plays the role of future prediction. However, inversely learning the physical laws [56, 13] is very demanding due to the intrinsic challenges in long-term prediction.

4 Experimental Setup

In this section, we briefly introduce the challenging physical reasoning benchmark of PHYRE-B and the setup of our experiments. We refer the readers to Appendix A for additional training details.

Environment PHYRE-B is a goal-driven benchmark consisting of 25 different task templates of physical puzzles that can be solved by placing a red ball (hence the "B"). Each template has 100 similar tasks, which enables two different evaluation settings. (i) Within-template setting: train on 80% of tasks of each template and test on the rest 20% of each template. (ii) Cross-template setting: train on all tasks of 20 from the 25 templates and test on the rest of five previously unseen templates.

The performance is evaluated by AUCCESS [3], a weighted sum of the accuracy rate related to the number of attempts. To encourage as few attempts as possible to solve a puzzle, the weights are calculated as: $\omega_k = \log(k+1) - \log(k)$, where $k \in \{1, 2, 3...100\}$ represents the range of attempts. Formally, AUCCESS $= \frac{\sum_k \omega_k \cdot s_k}{\sum_k \omega_k}$, where s_k is the success rate with k attempts.

This physical reasoning benchmark is deemed challenging primarily due to the following three reasons: (i) The environment involves a variety of objects of different shapes and mass distribution, such as balls, bars, standing sticks, and jars. (ii) The solution set takes up only a small part of the whole action space, making randomly sampled actions hardly ever work out. (iii) Before reaching the goal, a series of complex physical interactions would happen, involving falling, rotation, collision, and friction. To the best of our knowledge, no success has been claimed on the environment.

Experiment 1: LfD *vs.* LfI In the first experiment, we compare the SOTA LfD model on PHYRE and a transformer-based classifier for LfI. Specifically, we pick the RPIN model [53] to represent LfD and the ViT model [15] for LfI. The RPIN model leverages a convolutional interaction network based on object-centric representation and predicts the object states (bounding boxes and masks) into the future. When solving a PHYRE task, the model first predicts 10-time steps into the future for an action and then recruits an MLP-based task-solution model to predict the outcome. In comparison, ViT is a transformer-based classifier initially designed for image classification. The model patches an image, decomposes it into non-overlapping regions, and performs a series of self-attention. The image-level representation is finally extracted by appending a CLS token.

Experiment 2: LfD under GD The second experiment serves as a diagnostic test for the efficacy of dynamics in physical reasoning tasks. To this end, we directly extract simulator results and feed the image sequence into a video classifier model, setting an upper bound for the LfD paradigm by replacing dynamics prediction with the ground truth. In this experiment, we adopt the TimeSformer model [8] due to its superior performance. The model patches the image sequence and replaces traditional convolution architectures with interleaving spatial attention and temporal attention.

Experiment 3: LfD under AD We dive deep into the ineffectiveness of LfD methods like RPIN by fixing the task-solution model with TimeSformer but replacing the input with trajectories predicted by a learned module. In particular, we train a PredRNN [62] to learn the dynamics. PredRNN is an advanced pixel-based video predictor extending the inner-layer transition function of memory states in LSTMs to zigzag memory flow that propagates in both bottom-up and top-down directions. A curriculum is set up during learning for better long-term prediction [62].

Experiment 4: More on LfI In the final set of experiments, we verify the performance of different LfI methods. In addition to ViT, we consider BEiT [4] and Swin Transformer [51], which have been proven beneficial in image classification tasks.

5 Comparison of Learning from Intuition and Learning from Dynamics

In this section, we start our discussion with a preliminary experiment and consequently explore whether dynamics contribute to better judgments in physical reasoning tasks both in ideal conditions and in reality by comparing the performance of LfI and LfD.

5.1 LfD *vs.* LfI: A surprising finding

To tackle the demanding physical puzzles in PHYRE-B, previous studies use pixel-based or object-based future prediction modules to boost problem-solving performance [23, 53]. The SOTA LfD

method RPIN uses a convolutional interaction network to predict long-term dynamics. Inspired by previous studies on intuitive physics, we set out to see whether a decision can be made directly by a classifier of ViT. Counter-intuitively, while the supervision from the dynamics should ideally help a model in LfD look into the future, the performance of the LfI model reaches better than or on par with SOTA as shown in Tab. 1. The AUCCESS of ViT in the cross-template setting even significantly outperforms RPIN's, demonstrating its high generalization ability when solving unseen physical puzzles. Driven by this surprising and intriguing finding, we conduct additional follow-up experiments to determine if the dynamics prediction component helps physical reasoning.

Table 1: **The performance of RPIN and ViT in solving PHYRE-B puzzles.** We report the AUCCESS in both within and cross settings. RPIN uses objects' information as prior knowledge and extra supervisory signals for dynamics prediction, whereas ViT is trained only supervised by the final outcomes with initial scenes as input.

Model	Mechanism	Input	Supervision	Within	Cross
RPIN	LfD	Initial scenes, bboxs	Bboxs, masks, outcomes	85.49	50.86
ViT	LfI	Initial scenes	Outcomes	84.16±0.30	56.31±1.95

5.2 LfD under GD: Is ground-truth dynamics better than intuition?

The surprising discovery motivates us to ask why the LfD model has similar or even worse performance than LfI. As a first step, we question the merit of dynamics: Do dynamics help solve these physical puzzles? To answer this question, we supply the model with the best-case—ground-truth—dynamics and provide the information to our task-solution model of TimeSformer, due to the video-classification-like nature. We vary the number of time frames supplied to the model: We consider the input of lengths 1, 2, 4, and 8 extracted directly from the simulator with a time interval of 1 second. It is worth noting that using GD, we assume an ideal dynamics prediction model $g(\cdot)$ that accurately predicts the future.

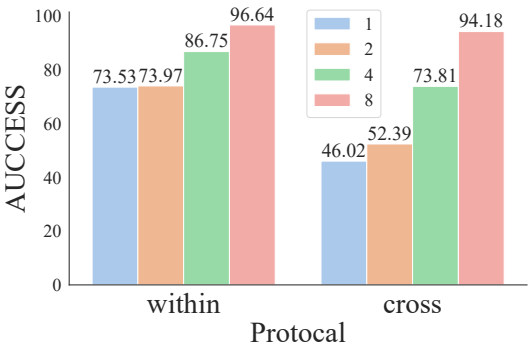

Figure 2: **Performance of TimeSformer on PHYRE with different ground-truth input lengths.** We compare the AUCCESS of within-template and cross-template evaluation settings.

The performance of LfD with GD is shown in Fig. 2. While the model underperforms ViT with fewer than 3 frames, which is out of the scope of our work, we see that AUCCESS is significantly boosted with four or more frames. Specifically, with 8 input frames, AUCCESS can be improved by 23.11 in the within setting and 48.16 in the cross setting, reaching about 95% in both settings. This observation answers the above question: Dynamics *do* help problem-solve in physical reasoning. Such an observation aligns with our intuition that dynamics can tell more about the future states, thus making learning decisions easier for the task-solution model. With sufficient dynamics information, the problem could be "solved" to a large extent.

However, if dynamics *do* help, why do current LfD methods fail to outperform LfI as shown in Sec. 5.1, even considerably lower than the simple baseline in generalization? Is it because of inaccurate dynamics prediction or because of the outdated design in task-solution models? We answer this question in the following set of experiments.

5.3 LfD under AD: How do approximate dynamics perform?

Given the fact that GD does help problem-solving, we explore the reason why AD—the dynamics prediction in practice—does not take effect as expected. Specifically, we fix the task-solution model $f(\cdot)$ with TimeSformer [8] as in Sec. 5.2 and instantiate the dynamics prediction model $g(\cdot)$ with an advanced video prediction model PredRNN [62]. We train the LfD pipeline using both of the two optimization schedules mentioned in Sec. 3 to investigate which schedule can better help in learning useful dynamics information.

The PredRNN first takes one initial frame as the input and outputs three predicted future frames. Next, the TimeSformer takes all the frames as input, including the first initial frame and the three predicted frames, and outputs the final outcome. For parallel optimization, we train an end-to-end framework by integrating the output from PredRNN with the input of TimeSformer. The model parameters ϕ and θ are updated simultaneously by backpropagating both the dynamics-learning loss and the final cross-entropy loss. For serial optimization, the PredRNN is trained first independently using ground-truth future dynamics as supervision and kept fixed, after which the TimeSformer is optimized using the output from the pre-trained PredRNN, supervised by the ground-truth decisions. For both optimization schedules, the PredRNN only takes raw RGB images as the input without extra information, such as bounding boxes or masks of each object, and directly generates the predicted future frames without further rendering.

We report the AUCCESS of both optimization schedules on PHYRE-B in Tab. 2. These results show that independent of the optimization schedule used, AD falls far behind from GD and performs equally or even worse than LfI, indicating that approximate dynamics do little help for the task-solution model in making better judgments. Specifically, by comparing the performance of parallel optimization to TimeSformer with a single frame, we observe a tendency of LfD pipeline's performance degeneration to LfI, ending up with similar AUCCESS. Moreover, Tab. 3 shows that for both within and cross settings, the test dynamics loss in parallel optimization is much higher than the corresponding one in serial optimization, which suggests that the representation learned by LfD pipeline in parallel optimization has difficulty in capturing accurate dynamics and is severely impacted by the cross-entropy loss, as is evident in the visualization results in Fig. 3. Although serial optimization gives the dynamics prediction model ample opportunities to learn better pixel-wise approximations of future frames, it still fails to make better decisions than parallel optimization. We hypothesize that in serial optimization, the task-solution model eventually uses the noisy predicted dynamics, which hinders the performance due to inaccuracy; in parallel optimization, the model can avoid this problem by paying less attention to learning future dynamics and focusing more on directly learning the outcome in a manner similar to LfI. Taken together, these experimental results pinpoint the reasons for the inefficacy of current LfD methods with AD: (i) It is the poor dynamics learned, rather than the task-solution model, that causes the performance degradation. Inaccuracy dynamics prevent accurate reasoning. (ii) Existing LfD methods with AD degenerate into LfI in the best case.

Table 2: **The performance of AD following two optimization schedules in LfD.** We use the same task-solution model TimeSformer for all the experiments. **NF** denotes the number of input frames used by the task-solution model. We also list the results using GD for comparison.

Prediction	NF	Within	Cross
PredRNN (parallel)	4	75.22	46.42
PredRNN (serial)	4	64.90	44.33
/	1	73.53	46.02
Simulator	4	86.75	73.81

By comparing the models above, we conclude that dynamics prediction, in theory, does help solve physical puzzles by providing more temporal information, while struggling to show the full strengths in reality due to the challenges in handling accumulated uncertainty in a long horizon. The results in serial and parallel optimization also reveal the fact that dynamics prediction plays an important role regardless of the task-solution model, but inaccurate dynamics will eventually harm downstream reasoning; in the best case, LfD degenerates into LfI.

Table 3: **The within- and cross-template test losses of AD LfD from the two optimization procedures.**

Opt	Loss	Within	Cross
Parallel	entropy	0.0638	0.5726
	dynamics	0.0039	0.0049
Serial	entropy	0.1285	0.6554
	dynamics	0.0003	0.0021

5.4 More on LfI: How does LfI perform?

Observing the notable success of ViT in the physical reasoning task, we consider testing additional visual classification models to verify the effectiveness of the LfI paradigm. Among a myriad of models, we pick ViT, Swin Transformer, and BEiT for experiments due to their demonstrated superiority in image classification. Of note, TimeSformer with a single input frame can also be considered as a model in the LfI paradigm. All of the models take only the first frame as the input. For equal comparison, we train all of the LfI models with an equal number of actions per task as in RPIN. Nevertheless, the LfI models take no extra prior object information as the input and use only final outcome labels as supervision signals, significantly simplifying the learning process.

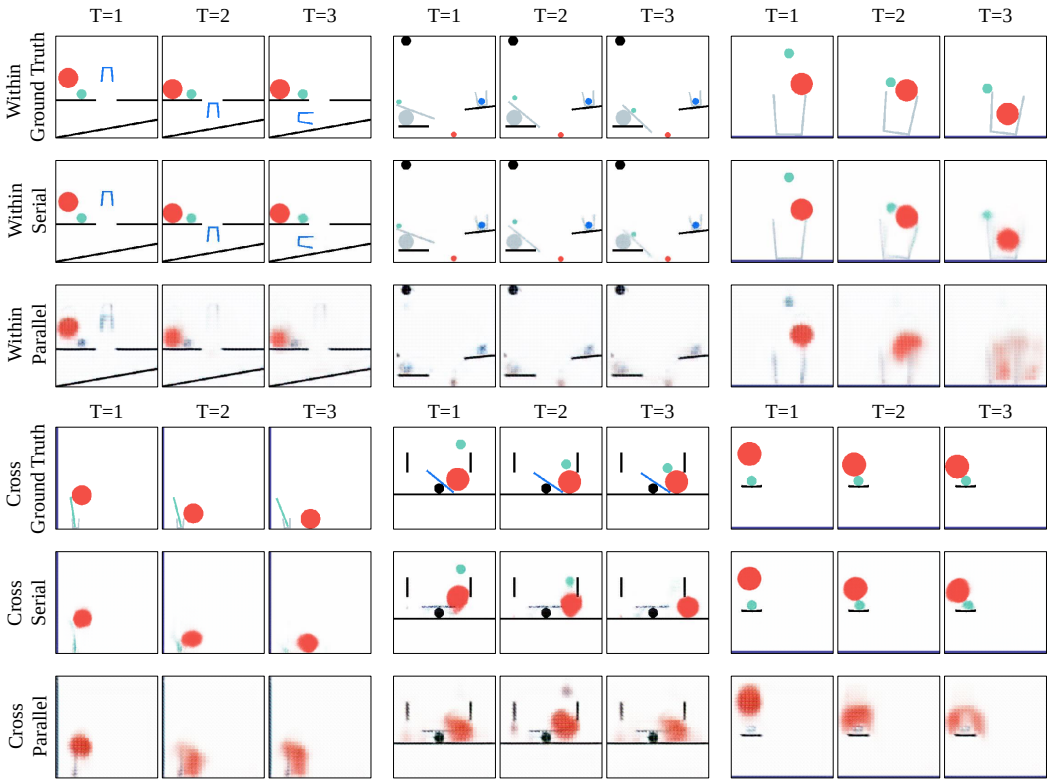

Figure 3: **Predicted dynamics from PredRNN in LfD's two optimization schedules.** In parallel optimization, PredRNN hardly learns any dynamics, whereas the dynamics learned in serial optimization are more accurate. However, serial optimization still fails to perform better than parallel optimization as shown in Tab. 2, indicating that improved but still noisy dynamics don't lead to better problem-solving. We refer the readers to Fig. A1 and A2 for more results.

The results of these LfI models are presented in Tab. 4. The performance of LfI models in the within-template setting is competitive with the SOTA RPIN. In addition, all of the LfI models outperform SOTA in the cross-template setting in AUCCESS; ViT even outperforms by almost 6 points on average. By visualizing the distribution heat maps of the possible solution set learned by ViT and comparing them with the ground truth in within-template and cross-template settings (see Fig. 4), we observe that the ViT model not only achieves high performance but also accurately recovers the underlying solution distributions. We hypothesize that intuitive models, though trained without dynamics information, can still extract high-level spatial knowledge and physical commonsense that generalizes well in unseen scenarios by looking at the data only; for instance, the most suitable red ball position should be close to the moving objects to exert influence. However, we believe that the performance of LfI still has significant room for improvement since the inductive biases and structure designs for physical reasoning are not explicitly considered.

Table 4: **Performance of different LfI models in solving PHYRE-B problems.** For ease of comparison, we also list results from the previous SOTA. We report the AUCCESS in within-template and cross-template settings. Input to ViT, Swin Transformer, and BEiT are only the first frame. For Dec [Joint] and RPIN, we directly use their reported AUCCESS for comparison.

Model	Mechanism	Object Info	Supervision	Within	Cross
ViT	LfI	False	Outcome	84.16±0.30	**56.31±1.95**
Swin	LfI	False	Outcome	84.71±0.33	54.92±2.30
BEiT	LfI	False	Outcome	83.59±0.09	54.07±1.88
Dec [Joint]	LfD under AD	False	Dynamics & Outcome	79.73	52.64
RPIN	LfD under AD	True	Dynamics & Outcome	**85.49**	50.86

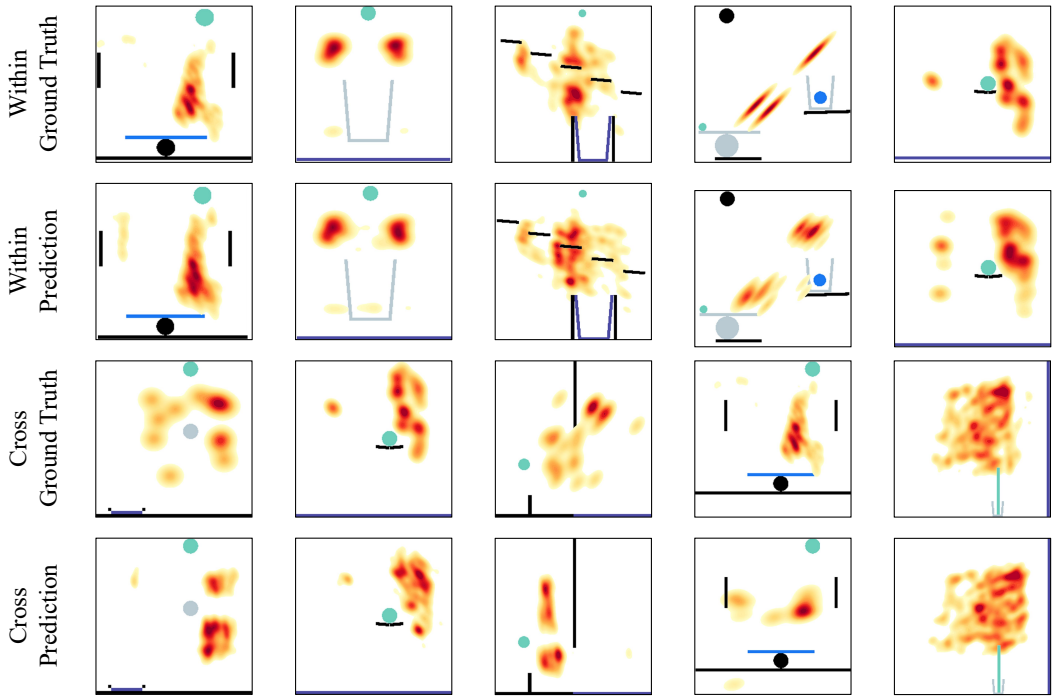

Figure 4: **The ground-truth $P(y|X_0)$ distribution heat maps and the ones predicted by ViT in PHYRE-B.** The maps are generated using the first 10,000 actions from the simulation cache offered by PHYRE-B. We only map the 2D positions of the placed red balls without showing their radii. For prediction heat maps, only actions with a likelihood above 0.8 are used for visualization. A warmer area denotes a higher likelihood of success. ViT learns an accurate solution set as the ground truth. We refer the readers to Fig. A3 and A4 for the maps of Swin and BEiT, respectively.

Besides the promising performance, LfI models also demonstrate the following advantages:

- They are design-efficient without complex hand-crafted tweaks for dynamics prediction modules, which in some cases only introduce harmful distraction or noisy representation; see Sec. 5.3.
- They only take initial scenes as the input and require no extra task-specific prior knowledge about the objects as used in object-centric dynamics prediction.
- They can be easily pre-trained on other computer vision tasks (*e.g.*, image classification on ImageNet), incorporating general domain-agnostic representation to avoid task-specific overfitting.

In summary, we view LfI as a simpler and more effective paradigm for physical reasoning.

6 Conclusion and Discussion

We introduce the concepts of two learning mechanisms in physical reasoning, *i.e.*, Learning from Intuition (LfI) and Learning from Dynamics (LfD). While it is generally believed that learning the dynamics of physics could help downstream reasoning, a beginning trial of RPIN and ViT on PHYRE-B challenges this fundamental assumption: A ViT model effectively learns to perform physical reasoning without any additional supervision from the ground-truth dynamics, object-centric or not. This counter-intuitive and surprising discovery motivates us to ask whether dynamics play an essential role in physical reasoning. We proceed to answer this question by using Ground-truth Dynamics (GD) from PHYRE's simulator and feeding the sequence to a TimeSformer model. Experimental results show that accurate dynamics can boost problem-solving performance. We further explore why Approximate Dynamics (AD) from dynamics predictors perform unfavorably in physical reasoning. We note that the task-solution model is not the one to blame. In addition, despite increasingly accurate dynamics prediction over the years, we notice that noisy dynamics prediction still has

a negative impact on the overall performance in reasoning; during parallel optimization, the LfD paradigm collapses into LfI. We hypothesize that in the long run, uncertainty in dynamics prediction unavoidably accumulates, leading to the inferior final performance. Finally, we dig deeper into the LfI paradigm and check the performance of various classification models in PHYRE. The experimental results show that these models achieve much better cross-template generalization while remaining competitive in within-template generalization. It turns out LfI can still perform well even if accurate dynamics are hard to predict, providing a route for a simpler, more natural, and less task-specific framework for physical reasoning. However, the LfD route, though challenging, remains a lucrative approach to this problem if dynamics could be predicted in a significantly more accurate manner.

Why do dynamics-based models struggle to make accurate predictions? Analyzing the experimental results, we try to provide preliminary explanations on why current dynamics prediction models struggle. We summarize the following three possible reasons:

1. The dynamics prediction itself is challenging, especially in unseen scenarios. For one thing, prediction into the far future is intrinsically difficult as unforeseeable events might steer the dynamics in another direction. For another, the errors will accumulate from earlier frames, leading to exploding uncertainty and noise in the future. Unfortunately, current dynamics prediction methods cannot produce a robust model to predict accurate dynamics in physical scenes.
2. Pixel-based dynamic representation has more information than object-based representation, while object-based representation is more concise. Arguably, pixel-based representation potentially incorporates all necessary information, such as shapes of objects, potential points of collision, and angular velocity. However, such a representation is extremely noisy, hence extracting useful information embedded is difficult. In comparison, object-based representation is by design concise and follows the general principles in the laws of physics. Nevertheless, object-centric methods lose essential clues in the scenes especially when it comes to collision and its aftereffects. The fact that there is not yet a feature representation method that summarizes all necessary information for physical modeling further complicates physical reasoning.
3. Task-solution model design might play a role, though not significant. The self-attention mechanism has constantly proven rewarding in a variety of tasks. The strong performance of LfI could also benefit from this architectural design.

Limitations and future work Building intelligence with physical reasoning ability is a never-ending journey. However, there are still many aspects left to be studied in the future.

- In contrast to the topic's significance is the lack of rich environments for experiments. Therefore, we only conduct experiments on PHYRE to support our insights. In the future, we hope to develop a suite of physical reasoning tasks and further investigate the insights presented in this work. In addition, we would also like to incorporate the latest advancements in the suite to see if LfI is a more feasible approach to physical reasoning.
- In the experiments, we use original LfI models in our experiments without extra design for physical reasoning tasks. We hypothesize that better perceptual modules and more useful spatial analyzing modules related to physical reasoning could further improve the performance of LfI. In particular, the most challenging physical dynamics in PHYRE involve collision which is a non-smooth and angular movement that does not lie in the traditional linear Euclidean space. We hope to incorporate insights for physical modeling into LfI models for further refinement.
- Analyzing the experiments in Sec. 5, an additional question to ask is how accurate the dynamics prediction needs to be for it to be beneficial. We did not figure out a measure for this problem in this work. Carefully masking or introducing noise to specific image areas and testing physics predictors at different levels may lead to an answer.
- While difficult, the dynamics prediction route remains appealing as indicated by its upper bound and the status quo. We are curious how dynamics will benefit other reasoning processes, such as counterfactual and hypothetical reasoning. We argue the dynamics prediction route is still worth additional efforts and look forward to advances made in this field.

Societal impacts No foreseeable negative societal impacts in this work.

Acknowledgment We would like to thank Miss Chen Zhen at BIGAI for making the nice figures and the anonymous reviewers for their constructive comments.

References

[1] Allen, K. R., Smith, K. A., and Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences (PNAS)*, 117(47):29302–29310. 2, 3

[2] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *International Conference on Computer Vision (ICCV)*. 3

[3] Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., and Girshick, R. (2019). Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems (NIPS)*. 1, 2, 3, 5

[4] Bao, H., Dong, L., Piao, S., and Wei, F. (2022). Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*. 3, 5

[5] Bates, C., Battaglia, P. W., Yildirim, I., and Tenenbaum, J. B. (2015). Humans predict liquid dynamics using probabilistic simulation. In *Annual Meeting of the Cognitive Science Society (CogSci)*. 3

[6] Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences (PNAS)*, 110(45):18327–18332. 1, 3

[7] Battaglia, W. P., Pascanu, R., Lai, M., Rezende, J. D., and Kavukcuoglu, K. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems (NIPS)*. 2, 3

[8] Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*. 2, 3, 5, 6

[9] Chen, Y., Huang, S., Yuan, T., Zhu, Y., Qi, S., and Zhu, S.-C. (2019). Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*. 3

[10] Chen, Z., Mao, J., Wu, J., Wong, K. K.-Y., Tenenbaum, B. J., and Gan, C. (2021a). Grounding physical concepts of objects and events through dynamic visual reasoning. In *International Conference on Learning Representations (ICLR)*. 3

[11] Chen, Z., Yi, K., Li, Y., Ding, M., Torralba, A., Tenenbaum, J. B., and Gan, C. (2021b). Comphy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations (ICLR)*. 3

[12] Coumans, E. (2015). Bullet physics simulation. In *ACM SIGGRAPH 2015 Courses*. 3

[13] Dai, B. and Seljak, U. (2021). Learning effective physical laws for generating cosmological hydrodynamics with lagrangian deep learning. *Proceedings of the National Academy of Sciences (PNAS)*, 118(16):e2020324118. 4

[14] Ding, M., Chen, Z., Du, T., Luo, P., Tenenbaum, J., and Gan, C. (2021). Dynamic visual reasoning by learning differentiable physics models from video and language. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. 2, 3, 5

[16] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). Carla: An open urban driving simulator. In *Conference on robot learning*. 3

[17] Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y. N., Lu, H., and Zhu, S.-C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, 4(37). 3

[18] Fischer, J., Mikhael, J. G., Tenenbaum, J. B., and Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. In *Proceedings of the National Academy of Sciences (PNAS)*. 2

[19] Gao, X., Gong, R., Shu, T., Xie, X., Wang, S., and Zhu, S.-C. (2019). Vrkitchen: an interactive 3d virtual environment for task-oriented learning. *arXiv preprint arXiv:1903.05757*. 3

[20] Gavrilyuk, K., Sanford, R., Javan, M., and Snoek, C. G. (2020). Actor-transformers for group activity recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

[21] Gilden, D. L. and Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. *Perception & Psychophysics*, 56(6):708–720. 3

[22] Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A. (2019). Video action transformer network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

[23] Girdhar, R., Gustafson, L., Adcock, A., and van der Maaten, L. (2020). Forward prediction for physical reasoning. *arXiv preprint arXiv:2006.10734*. 2, 3, 5

[24] Groth, O., Fuchs, F., Posner, I., and Vedaldi, A. (2018). Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *European Conference on Computer Vision (ECCV)*. 3

[25] Grzeszczuk, R., Terzopoulos, D., and Hinton, G. (1998). Neuroanimator: Fast neural network emulation and control of physics-based models. *ACM Transactions on Graphics (TOG)*, pages 9–20. 3

[26] Han, M., Zhang, Z., Jiao, Z., Xie, X., Zhu, Y., Zhu, S.-C., and Liu, H. (2021). Reconstructing interactive 3d scene by panoptic mapping and cad model alignments. In *International Conference on Robotics and Automation (ICRA)*. 3

[27] Han, M., Zhang, Z., Jiao, Z., Xie, X., Zhu, Y., Zhu, S.-C., and Liu, H. (2022). Scene reconstruction with functional objects for robot autonomy. *International Journal of Computer Vision (IJCV)*. 3

[28] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

[29] Hong, Y., Yi, L., Tenenbaum, J., Torralba, A., and Gan, C. (2021). Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[30] Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., and Zhu, S.-C. (2018). Holistic 3d scene parsing and reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. 3

[31] Janner, M., Levine, S., Freeman, W. T., Tenenbaum, J. B., Finn, C., and Wu, J. (2019). Reasoning about physical interactions with object-oriented prediction and planning. In *International Conference on Learning Representations (ICLR)*. 3

[32] Jia, B., Lei, T., Zhu, S.-C., and Huang, S. (2022). Egotaskqa: Understanding human tasks in egocentric videos. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[33] Jiao, Z., Niu, Y., Zhang, Z., Zhu, S.-C., Zhu, Y., and Liu, H. (2022). Planning sequential tasks on contact graph. In *International Conference on Intelligent Robots and Systems (IROS)*. 3

[34] Jiao, Z., Zeyu, Z., Jiang, X., Han, D., Zhu, S.-C., Zhu, Y., and Liu, H. (2021a). Consolidating kinematic models to promote coordinated mobile manipulations. In *International Conference on Intelligent Robots and Systems (IROS)*. 3

[35] Jiao, Z., Zeyu, Z., Wang, W., Han, D., Zhu, S.-C., Zhu, Y., and Liu, H. (2021b). Efficient task planning for mobile manipulation: a virtual kinematic chain perspective. In *International Conference on Intelligent Robots and Systems (IROS)*. 3

[36] Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan. 2

[37] Karpas, E. and Magazzeni, D. (2020). Automated planning for robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:417–439. 3

[38] Koenig, N. and Howard, A. (2004). Design and use paradigms for gazebo, an open-source multi-robot simulator. In *International Conference on Intelligent Robots and Systems (IROS)*. 3

[39] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., and Farhadi, A. (2017). Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*. 3

[40] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[41] Kubricht, J., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., and Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In *Annual Meeting of the Cognitive Science Society (CogSci)*. 3

[42] Kubricht, J., Zhu, Y., Jiang, C., Terzopoulos, D., Zhu, S.-C., and Lu, H. (2017a). Consistent probabilistic simulation underlying human judgment in substance dynamics. In *Annual Meeting of the Cognitive Science Society (CogSci)*. 3

[43] Kubricht, J. R., Holyoak, K. J., and Lu, H. (2017b). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10):749–759. 1, 2, 3

[44] Kwon, Y.-H. and Park, M.-G. (2019). Predicting future frames using retrospective cycle gan. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

[45] Lee, S.-H. and Terzopoulos, D. (2006). Heads up! biomechanical modeling and neuromuscular control of the neck. *ACM Transactions on Graphics (TOG)*, pages 1188–1198. 3

[46] Lerer, A., Gross, S., and Fergus, R. (2016). Learning physical intuition of block towers by example. In *International Conference on Machine Learning (ICML)*. 2, 3

[47] Li, C., Liang, W., Quigley, C., Zhao, Y., and Yu, L.-F. (2017). Earthquake safety training through virtual drills. *IEEE Transactions on Visualization and Computer Graph (TVCG)*, 23(4):1275–1284. 3

[48] Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K., Gokmen, C., Dharan, G., Jain, T., et al. (2021). igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*. 3

[49] Liu, H., Zhang, C., Zhu, Y., Jiang, C., and Zhu, S.-C. (2019). Mirroring without overimitation: Learning functionally equivalent manipulation actions. In *AAAI Conference on Artificial Intelligence (AAAI)*. 3

[50] Liu, T., Liu, Z., Jiao, Z., Zhu, Y., and Zhu, S.-C. (2021a). Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters (RA-L)*, 7(1):470–477. 3

[51] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*. 3, 5

[52] Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Argyros, A. (2020). A review on deep learning techniques for video prediction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 3

[53] Qi, H., Wang, X., Pathak, D., Ma, Y., and Malik, J. (2021). Learning long-term visual dynamics with region proposal interaction networks. In *International Conference on Learning Representations (ICLR)*. 2, 3, 5

[54] Sanborn, A. N., Mansinghka, V. K., and Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, 120(2):411. 3

[55] Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al. (2019). Habitat: A platform for embodied ai research. In *International Conference on Computer Vision (ICCV)*. 3

[56] Schmidt, M. and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85. 4

[57] Schwartz, D. L. and Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(1):116. 3

[58] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[59] Todd, J. T. and Warren Jr, W. H. (1982). Visual perception of relative mass in dynamic events. *Perception*, 11(3):325–335. 3

[60] Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems (IROS)*. 3

[61] Toussaint, M., Allen, K., Smith, K. A., and Tenenbaum, J. B. (2018). Differentiable physics and stable modes for tool-use and manipulation planning. In *Robotics: Science and Systems (RSS)*. 3

[62] Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P., and Long, M. (2022a). Predrnn: A recurrent neural network for spatiotemporal predictive learning. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 3, 5, 6

[63] Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., and Huang, S. (2022b). Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[64] Wu, J., Lim, J. J., Zhang, H., Tenenbaum, J. B., and Freeman, W. T. (2016). Physics 101: Learning physical object properties from unlabeled videos. In *British Machine Vision Conference (BMVC)*. 3

[65] Wu, J., Yildirim, I., Lim, J. J., Freeman, B., and Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[66] Wu, Y., Wu, Y., Gkioxari, G., and Tian, Y. (2018). Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*. 3

[67] Xie, X., Liu, H., Zhang, Z., Qiu, Y., Gao, F., Qi, S., Zhu, Y., and Zhu, S.-C. (2019). Vrgym: A virtual testbed for physical and interactive ai. In *Proceedings of the ACM Turing Celebration Conference-China*, pages 1–6. 3

[68] Xu, K., Srivastava, A., Gutfreund, D., Sosa, F., Ullman, T., Tenenbaum, J., and Sutton, C. (2021). A bayesian-symbolic approach to reasoning and learning in intuitive physics. In *Advances in Neural Information Processing Systems (NIPS)*. 3

[69] Xue, C., Pinto, V., Gamage, C., Nikonova, E., Zhang, P., and Renz, J. (2021). Phy-q: A benchmark for physical reasoning. *arXiv preprint arXiv:2108.13696*. 3

[70] Ye, T., Qi, S., Kubricht, J., Zhu, Y., Lu, H., and Zhu, S.-C. (2017). The martian: Examining human physical judgments across virtual gravity fields. *IEEE Transactions on Visualization and Computer Graph (TVCG)*, 23(4):1399–1408. 3

[71] Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, B. J. (2020). Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*. 3

[72] Zhang, Z., Jiao, Z., Wang, W., Zhu, Y., Zhu, S.-C., and Liu, H. (2022). Understanding physical effects for effective tool-use. *RA-L*, 7(4):9469–9476. 3

[73] Zheng, B., Zhao, Y., Yu, J., Ikeuchi, K., and Zhu, S.-C. (2015). Scene understanding by reasoning stability and safety. *International Journal of Computer Vision (IJCV)*, 112(2):221–238. 3

[74] Zheng, B., Zhao, Y., Yu, J. C., Ikeuchi, K., and Zhu, S.-C. (2013). Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

[75] Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., et al. (2020). Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345. 3

[76] Zhu, Y., Jiang, C., Zhao, Y., Terzopoulos, D., and Zhu, S.-C. (2016). Inferring forces and learning human utilities from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

[77] Zhu, Y., Zhao, Y., and Zhu, S.-C. (2015). Understanding tools: Task-oriented object modeling, learning and recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See abstract and introduction in Sec. 1.
   (b) Did you describe the limitations of your work? [Yes] See Sec. 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have read the ethics review guidelines carefully and ensured that out paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will include them in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We will include them in the supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Tab. 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We will include them in the supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See our experiments.
   (b) Did you mention the license of the assets? [Yes] They are open-source for research.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Training Details

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

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.

LfD under 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 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 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.

B Additional Visualizations

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

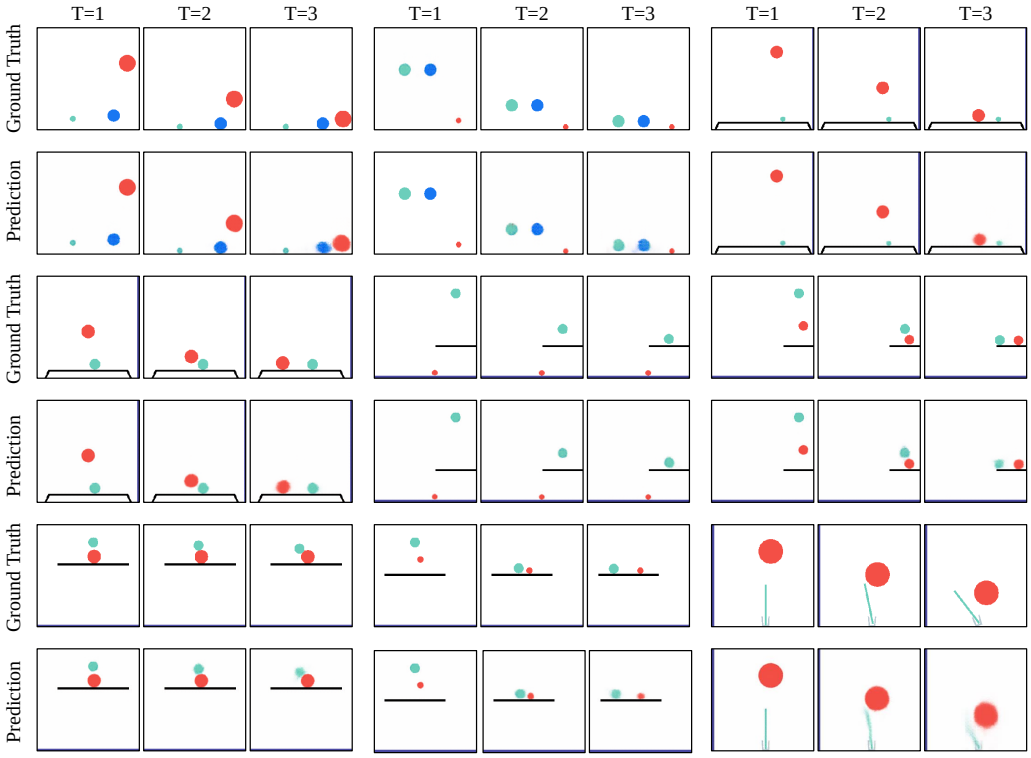

Figure A1: **Predicted dynamics from PredRNN in LfD's serial optimization schedule in easy tasks.**

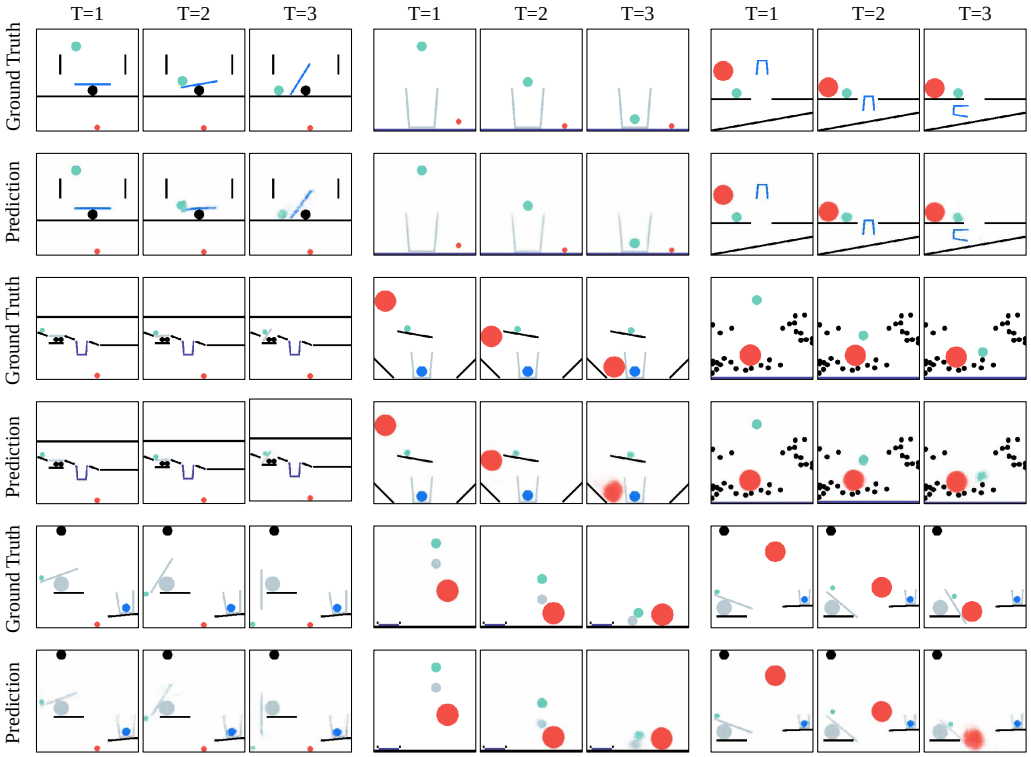

Figure A2: **Predicted dynamics from PredRNN in LfD's serial optimization schedule in hard tasks.**

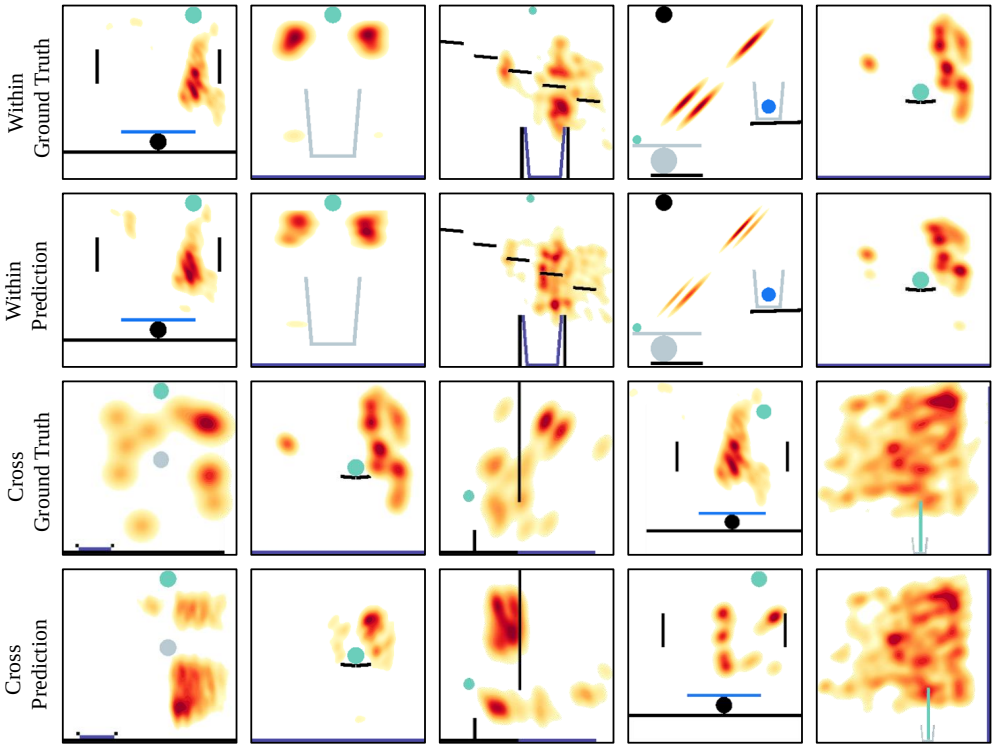

Figure A3: **The ground-truth $\mathbf{P}(\mathbf{y}|\mathbf{X_0})$ distribution heat maps and the ones predicted by Swin Transformer in PHYRE-B.** The heat maps are generated in the same way as in ViT.

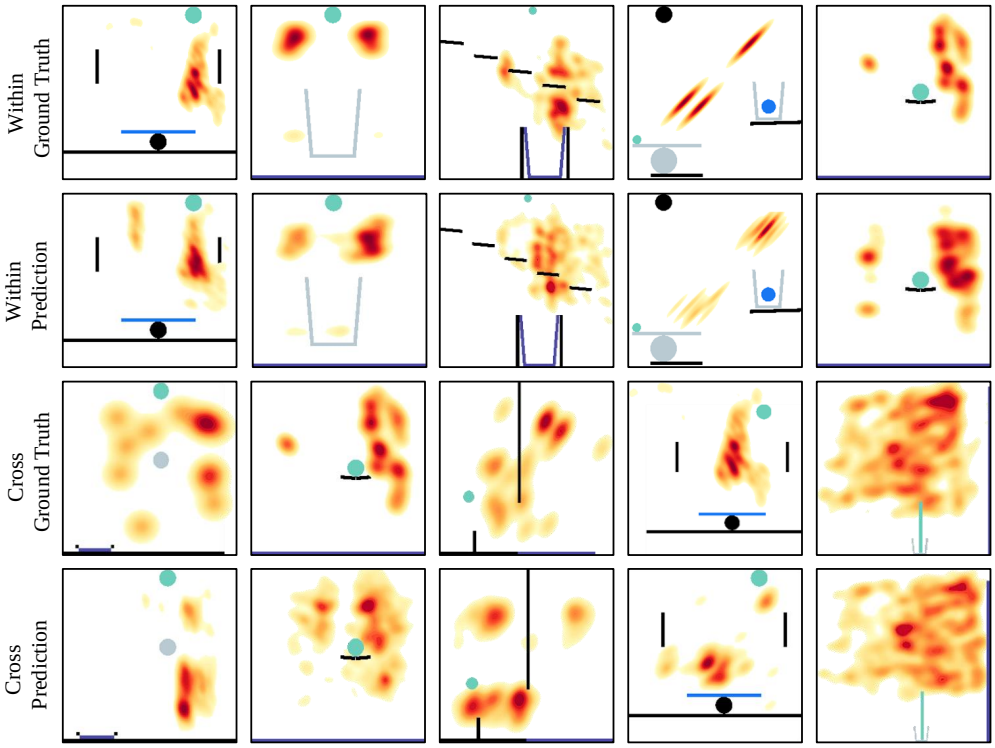

Figure A4: **The ground-truth $\mathbf{P}(\mathbf{y}|\mathbf{X_0})$ distribution heat maps and the ones predicted by BEiT in PHYRE-B.** The heat maps are generated in the same way as in ViT.