Deconfounding Physical Dynamics with Global Causal Relation and Confounder Transmission for Counterfactual Prediction

Zongzhao Li[1,2], **Xiangyu Zhu**[1,2], **Zhen Lei**[1,2,3*], **Zhaoxiang Zhang**[1,2,3]

[1]NLPR & CBSR, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences
{xiangyu.zhu,zlei}@nlpr.ia.ac.cn {lizongzhao2020,zhaoxiang.zhang}@ia.ac.cn

Abstract

Discovering the underneath causal relations is the fundamental ability for reasoning about the surrounding environment and predicting the future states in the physical world. Counterfactual prediction from visual input, which requires simulating future states based on unrealized situations in the past, is a vital component in causal relation tasks. In this paper, we work on the confounders that have effect on the physical dynamics, including masses, friction coefficients, etc., to bridge relations between the intervened variable and the affected variable whose future state may be altered. We propose a neural network framework combining Global Causal Relation Attention (GCRA) and Confounder Transmission Structure (CTS). The GCRA looks for the latent causal relations between different variables and estimates the confounders by capturing both spatial and temporal information. The CTS integrates and transmits the learnt confounders in a residual way, so that the estimated confounders can be encoded into the network as a constraint for object positions when performing counterfactual prediction. Without any access to ground truth information about confounders, our model outperforms the state-of-the-art method on various benchmarks by fully utilizing the constraints of confounders. Extensive experiments demonstrate that our model can generalize to unseen environments and maintain good performance.

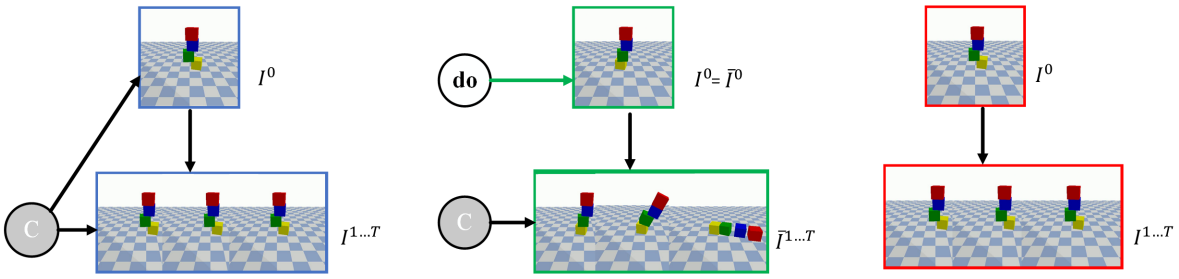
Introduction

One of the main distinctions that differentiate human-like intelligence from others may lie in the understanding, reasoning and predicting ability, especially when the agent is confronted with a novel and complicated environment (Spelke and Kinzler 2007; Martin-Ordas, Call, and Colmenares 2008). Discovering the causality and inferring time-invariant variables from visual input, has served as the core abilities for intelligent agent to build a basic understanding of the world (Glymour, Zhang, and Spirtes 2019). Furthermore, based on the obtained knowledge, agent can forecast the future outcome involving external interventions. In this paper, we utilize counterfactual prediction as our main quantitative indicator to probe model's capability.

In counterfactual prediction tasks, interventions will be employed to the system, producing situations that model has not experienced. Therefore, the model is supposed to discover causal relations between the intervened variables and the alternative outcomes. As shown in Figure 1 (a), the frames display the transformation as well as the reorientation of four objects. We wish the model to predict future trajectories of these objects if we had changed the initial frame by applying $do - operation$ (Pearl 2009) intervention. In our case, $do - operation$ contains object displacement or removal. Note that counterfactual future prediction is different from feedforward future prediction in Figure 1 (b), which learns spatio-temporal regularities from a few past frames to make future predictions. Counterfactual prediction aims to estimate unobservable confounders as references. It benefits from the observation of the original sequence $I^{0...T}$. It takes unobservable confounders into account, which has an important impact on the system. Recently, there are a few methods that model physical dynamics systems from an object-centric view and predict the counterfactual outcome. CoPhyNet (Baradel et al. 2020) models the interactions of objects based on Graph Convolution Networks (Kipf and Welling 2016) and further predicts the future outcomes by GRU (Cho et al. 2014). V-CDN (Li et al. 2020c) proposes a keypoint-based model to infer the hidden confounders and predict the future movements of the keypoints. Although they can discover the relations of objects and make future predictions, there is still room for further improvement.

In this paper, we concentrate on the confounders, which is tied to physics laws and independent of the data, to tackle counterfactual prediction tasks. Our model aims to estimate the unobservable confounders, such as masses, friction coefficients, and gravity, reason interactions across different objects, and discover the relations between the intervened variable with the variable whose alternative future should be predicted. Our model first learns the reflect function to represent each object as a compact vector and uses a Graph Convolution Network (GCN) (Kipf and Welling 2016) to update object embeddings. Then it reasons causal relationships of objects and deconfounds the object embeddings based on a proposed Global Causal Relation Attention (GCRA). GCRA captures the spatial-temporal information across objects and frames, and contributes to the accurate identification of the confounders. Finally, reposed on the inferred confounders and the modified initial state, a new forward module Confounder Transmission Structure (CTS) in-

*Corresponding author.



(a) counterfactual prediction (b) feedforward prediction

Figure 1: The counterfactual prediction task (a) versus the feedforward prediction task (b). I^0 and $I^{1...T}$ refer to the observations of the original sequence. C refer to the confounders and do refer to the $do - operation$. $\bar{I}^0$ refers to the modified initial observation and $\bar{I}^{1...T}$ refer to the counterfactual outcomes. Compared with the feedforward prediciton, the counterfactual one considers confounding variable, which is unobservable whereas governs the behavior of the system. The confounders can be instantiate as the masses, friction coefficients and gravity in physical systems.

tegrates and transmits these information in a residual way to forecast the future trajectory of each object. We evaluate our model in different settings of CoPhy (Baradel et al. 2020) benchmark and show superior performance over the state-of-the-art. Though making long-term prediction is difficult even for human beings, by fully utilizing information from confounders, our model achieves good performance, and generalizes well to novel cases, even without the access to the ground truth confounder labels.

Related Work

Intuitive Physics Understanding and representing intuitive physics is crucial for modelling interactions between objects and predicting the dynamics, which has attracted significant attention in machine learning society (Mottaghi et al. 2016; Kubricht, Holyoak, and Lu 2017; Sun et al. 2018; Ding et al. 2021; Li et al. 2020b, 2019; Liu et al. 2018; Hamrick, Battaglia, and Tenenbaum 2011; Stanić and Schmidhuber 2019). Early approaches (Wu et al. 2015, 2016) make full use of the information from physical properties and object attributes to produce follow-up predictions. Researchers also leverage Convolutional Neural Network (CNN) (Krizhevsky, Sutskever, and Hinton 2012) to investigate methods for reasoning physics from visual input (Wu et al. 2017; Fragkiadaki et al. 2016; Hamrick, Battaglia, and Tenenbaum 2011; Battaglia, Hamrick, and Tenenbaum 2013). In most cases, these methods either acquire the supervision of ground truth information about latent physical parameters or lack the ability to model the relationships across objects appropriately. The Interaction Network (IN) proposed by (Battaglia et al. 2016) takes advantage of Graph Neural Networks (Scarselli et al. 2008) to capture the interactions between entities in the scene. Similar to the Interaction Network, (Janner et al. 2019; Battaglia et al. 2018; Yi* et al. 2020; Baradel et al. 2020; Li et al. 2020a) show encouraging results in physics reasoning. However, because they neglect the causal chain between objects in long-term sequence, they are unable to discover latent interactions effectively.

Causal Inference Causal inference has garnered a lot of attention in recent works (Chalupka, Perona, and Eberhardt 2014; Lopez-Paz and Oquab 2016; Rojas-Carulla, Baroni, and Lopez-Paz 2018) due to the limitations of traditional statistical techniques (Pearl 2009). Methods for measuring the effects of multiple variables are being investigated by researchers. (Lopez-Paz and Oquab 2016; Lopez-Paz et al. 2017) concern discovering causal relations between variables by using neural networks or GANs (Goodfellow et al. 2014). (Kocaoglu et al. 2018) discovers causal effect with true observational and interventional data. (Mao et al. 2019) creates an object-based scene representation and infers causality via video reasoning.

Visual Reasoning People also study the task of visual reasoning in order to better assess model's capability of dynamics modelling (Ehrhardt et al. 2019; Finn, Goodfellow, and Levine 2016; Fraccaro et al. 2017; Hafner et al. 2019; Lei et al. 2018; Ha and Schmidhuber 2018; Finn and Levine 2017). Models need to make future predictions based on the images provided, such as predicting the stability of the physical structures (Lerer, Gross, and Fergus 2016; Groth et al. 2018; Jia et al. 2014; Li, Leonardis, and Fritz 2017; Li et al. 2016), tracking objects in the scenario (Ye et al. 2018), inferring physical properties from raw images (Agrawal et al. 2016), estimating physical plausibility (Riochet et al. 2018; Tompson et al. 2017), or making counterfactual prediction on account of the interventions given by external force (Baradel et al. 2020).

In our work, we first propose GCRA to model interactions between objects in long-term video frames, which is effective for deconfounding process. Then confounder variables are reused in a residual way in CTS, which facilitates the counterfactual prediction. We train our model with no supervision of confounders, so that it can generalize well to novel settings.

Reasoning and Predicting

This section outlines the details of our model, which contains detecting object representations from images, discovering relationships between objects, estimating confounders, and making counterfactual predictions. Our model is trained from visual inputs without supervision of the confounders.

Problem formulation. We study the task of visual reasoning in physical reality. Considering a dataset of various configurations of confounders, such as masses and friction coefficients, and trajectories generated under the effect of physical laws. Each sample in the dataset contains two video sequences of T RGB frames. The first sequence $\mathcal{I} = \{I^0, ..., I^T\}$ is observable for the model, which represents the evolution of the objects' states. The second one $\bar{\mathcal{I}} = \{\bar{I}^0, ..., \bar{I}^T\}$ is called counterfactual sequence, where the initial frame I^0 is transformed to $\bar{I}^0$ after the $do - operation$. The rest of the sequence are counterfactual outcomes accordingly. Given the data, we can formalize the problem as follows. The model takes a raw video $\mathcal{I} = \{I^0, ..., I^T\}$ as input, then it is supposed to model interactions between objects, estimate confounders in the scene. Afterwards, given the modified initial frame $\bar{I}^0$, the model is asked to predict the counterfactual outcomes $\bar{\mathcal{I}} = \{\bar{I}^1, ..., \bar{I}^T\}$.

Overview of the model. Our model takes raw RGB frames as inputs. We firstly extract object-centric features from images and detect corresponding 3D positions using a Convolutional Neural Network (CNN), which uses ResNet18 as backbone.

$$\begin{aligned} &\mathcal{V}^t = f_P(I^t), \bar{\mathcal{V}}^0 = f_P(\bar{I}^0), t = 0, ..., T, \\ &\mathcal{V}^t = \{o_i^t\}, \bar{\mathcal{V}}^0 = \{\bar{o}_i^0\}, i = 1, ..., M, \end{aligned} \tag{1}$$

where f_P is the perception module to detect M objects, the o_i^t corresponds 3D position information for object i at t timestep. Next, we use a Graph Neural Network (GNN) to update object embeddings, denoted as the $\check{\mathcal{V}}^t = \{\check{o}_i^t\}$. The GNN we adapt is the modified version of the Interaction Network (IN) (Battaglia et al. 2016). Then an inference module f_I is applied to take the object embeddings as input and infer latent representations of the confounders, denoted as $\mathcal{C}_i$ for each object in the scene,

$$\mathcal{C} = f_I(\check{o}_i^{0:T}), \quad \mathcal{C} = \{\mathcal{C}_i\}, i = 1, ..., M. \tag{2}$$

Finally, a forward function f_F, aims to predict the counterfactual outcomes, i.e., the 3D coordinates of M objects at $t+1$ timestep based on the estimated confounders $\mathcal{C}$.

$$\bar{\mathcal{V}}^{t+1} = f_F(\bar{\mathcal{V}}^{0:t}, \mathcal{C}), t = 0, ..., T-1. \tag{3}$$

Preliminary. In our pipeline we first utilize the Convolutional Neural Network (CNN) as the perception module to extract object-centric features and convert them into 3D positions through an MLP, denoted as $\mathcal{V}^t = \{o_i^t, i = 1, ..., M\}$. After the o_i^t we have obtained, we use Graph Neural Network (GNN) to model interactions between different objects and update the object embeddings coarsely. We first view the M objects in each frame as a fully-connected object graph $\mathcal{G}_o^t = (\mathcal{V}^t, \mathcal{E}^t)$, the nodes $\mathcal{V}^t = \{o_i^t\}$ are associated to the objects, and the edge $(o_i^t, o_j^t) \in \mathcal{E}^t$ represents the object interactions between object i and j. Specifically,

$$\begin{aligned} &\mathcal{V}^t = \{o_i^t, \}, i = 1, ..., M, t = 0, ...T \\ &e_i^t = \frac{1}{N_j} \sum_{j \neq i} f_R(o_i^t, o_j^t), \check{o}_i^t = f_G(e_i^t, \frac{1}{M} \sum_{i=1}^{M} e_i^t, o_i^t), \\ &\check{\mathcal{V}}^t = \{\check{o}_i^t\}, i = 1, ..., M, t = 0, ...T \end{aligned} \tag{4}$$

where $\check{o}_i^t$ represents the updated object embeddings, f_R is the function to calculate relational reasoning results, the goal of the f_G is to combine the relational embeddings and the original object embeddings.

Global Causal Relation Attention (GCRA)

To approximate the latent representations of confounders by deconfounding the object embeddings, we propose the Global Causal Relation Attention (GCRA) as the inference module f_I. Our goal is to capture the relation information among objects at spatial level as well as temporal level simultaneously. Previous method prefers to run a dedicated RNN for each object and considers the last hidden state of the recurrent network as the confounders (Baradel et al. 2020). It ignores the indirect relations between different objects at different frames. Moreover, the RNN is employed in each object independently and the interactions between objects are not updated during this process. Therefore, sophisticated information across objects and long-term frames cannot be well extracted to estimate confounders.

In our work, we aim to encode the relationships between different objects at different frames in long-term, so that the model can discover and leverage indirect causal chains. To this end, we propose Global Causal Relation Attention (GCRA) to model interactions and infer the confounders over object embeddings in each frame of the video sequence. We extend transformer-based model adapting scaled dot-product attention (Vaswani et al. 2017). It computes the relevance between objects and can be described as follows:

$$\begin{aligned} Attention(Q_i^s, K_j^t, V_j^t) = \textit{softmax}(\frac{Q_i^s (K_j^t)^T}{\sqrt{d_k}}) V_j^t, \\ i, j = 1, ..., M, \\ s, t = 1, ..., T, \end{aligned} \tag{5}$$

where the Q_i^s, K_j^t and V_j^t are queries, keys and values calculated by multiplying the input matrix X and the corresponding W_{qsi}, W_{ktj} and W_{vtj}. X represents the embedding of object, d_k represents the dimension of the key. The subscript i, j denote the index of object, the superscript s, t denote the index of frame. Note that when i, j and s, t are different, two object states from different frames are interacted. Besides, long-term interactions across many frames can be modelled when i and j are far away. The temporal and spatial information are encoded by applying interframe attention as well as intraframe attention. This is of great importance for the confounders estimation. The ablation experiments also demonstrate that GCRA is efficient in various environments.

The outputs of GCRA module are latent representations, which are denoted as the confounders. Specifically, we have:

$$\mathcal{C} = f_I(\check{o}_i^{0:T}), \quad \mathcal{C} = \{\mathcal{C}_i\}, i = 1, ..., M, \tag{6}$$

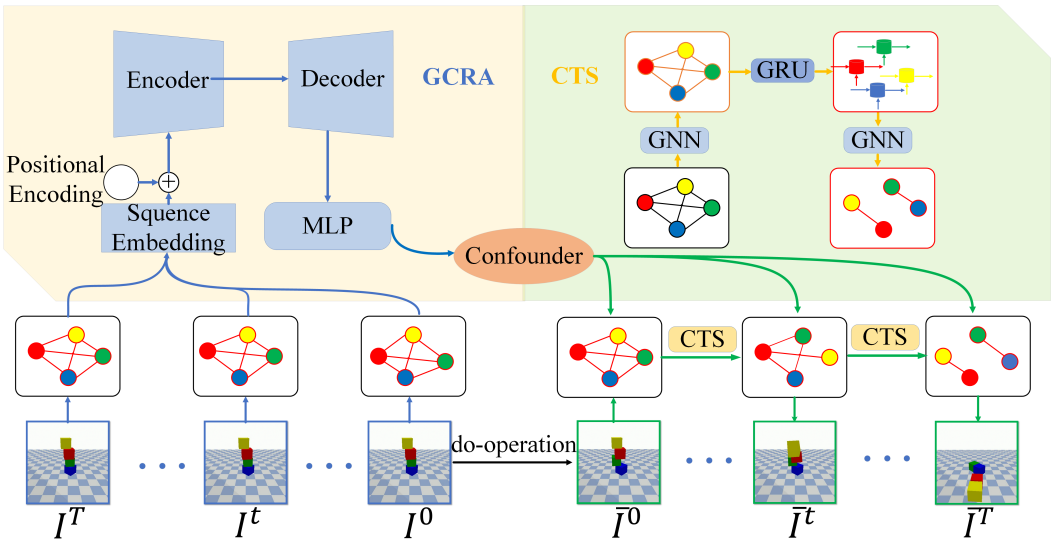


Figure 2: Architecture of our model. The model takes raw images $\mathcal{I} = \{I^0, ..., I^T\}$ as input and first uses a perceptual module to extract abstract representations which constitute a fully-connected graph. Then GCRA is used as an inference module to discover causal relations and infer the confounders. Finally CTS, a forward module that makes counterfactual predictions $\bar{\mathcal{I}} = \{\bar{I}^1, ..., \bar{I}^T\}$ using the estimated confounders as well as the modified initial frame $\bar{I}^0$. We train this model without any access to the ground truth information about confounders.

where f_I represents the GCRA module, depicting deconfounding process. $\mathcal{C}_i$ represents the confounders of i object. This module could generalize to different number of objects as well as different shape of objects in various settings.

Confounder Transmission Structure (CTS)

Conditioned on the inferred confounders, we would like to make counterfactual predictions, i.e., 3D positions of all objects. We propose the Confounder Transmission Structure (CTS) as the forward module f_F, to predict the counterfactual outcomes, by encoding and propagating confounders information. The experiments also validate our effectiveness.

We assume that there is a latent causal graph $\mathcal{G}_c^t$ under the scene. The node represents object attributes like masses and friction coefficients, i.e., confounders. The edge represents contact relations, which can be explained as the interaction force between objects. Firstly, a Graph Neural Network ϕ is used to propagate node and edge information based on the causal graph $\mathcal{G}_c^t$ as well as the modified object graph $\bar{\mathcal{G}}_o^t$. $\bar{\mathcal{G}}_o^t$ is produced by performing $do-operation$ (e.g. change the initial positions of objects or remove one of the objects in the scene). We concatenate $\bar{\mathcal{G}}_o^t$ and $\mathcal{G}_c^t$ since they have the same structure,

$$
\begin{aligned}
&\mathcal{G}_c^t = (\mathcal{C}, \mathcal{E}_c^t), \mathcal{C} = \{\mathcal{C}_i\}, \mathcal{E}_c^t = \{e_{i,j}^t\}, \\
&\bar{\mathcal{V}}^{0:t} = \{\bar{o}_i^{0:t}\}, \bar{e}_{i,j}^{0:t} = [\bar{o}_i^{0:t} : \bar{o}_j^{0:t}], \bar{\mathcal{E}}_o^{0:t} = \{\bar{e}_{i,j}^{0:t}\}, \\
&\bar{\mathcal{G}}_o^t = (\bar{\mathcal{V}}^{0:t}, \bar{\mathcal{E}}_o^{0:t}), [\bar{\mathcal{G}}_o^t : \mathcal{G}_c^t] = ([\bar{\mathcal{V}}^{0:t} : \mathcal{C}], [\bar{\mathcal{E}}_o^{0:t} : \mathcal{E}_c^t]) \quad (7) \\
&(\tilde{o}_i^t, \tilde{e}_{i,j}^t) = \phi(\bar{\mathcal{G}}_o^t, \mathcal{G}_c^t), \\
&t = 0, ..., T, i, j = 1, ..., M,
\end{aligned}
$$

where $\bar{o}_i^{0:t}$ represents the counterfactual outcomes predicted by CTS, $\bar{e}_{i,j}^{0:t}$ is simply stacked by $\bar{o}_i^{0:t}$ and $\bar{o}_j^{0:t}$, edge $e_{i,j}^t$ is randomly initialized and represents the contact information between objects i and object j. It is adaptively learned by the model without the ground truth information. When the model classifies the edge as "no relations", it means that object i and object j of this edge have no direct contact probably, and model tends to decrease the subsequent transmission of information between these two objects.

Then we add residual links that connect $\bar{o}_i^{0:t}$ and the embeddings produced by ϕ. We further concatenate them with the causal graph $\mathcal{G}_c^t$, pass all these information through a recurrent network to aggregate the information at temporal level,

$$
\begin{aligned}
\hat{o}_i &= RNN(\tilde{o}_i^t, \bar{o}_i^{0:t}, \mathcal{G}_c^t), t = 0, ..., T, \\
\hat{e}_{i,j} &= RNN(\tilde{e}_{i,j}^t, \bar{e}_{i,j}^{0:t}, \mathcal{G}_c^t), t = 0, ..., T,
\end{aligned} \quad (8)
$$

where we use GRU as the recurrent network, which enables our model to deal with objects of variable numbers and input sequences of variable lengths.

Taking the outputs of GRU as input, we exert another Graph Neural Network φ to directly make future prediction of object trajectory without forecasting the stability in advance. To accomplish more accurate results, we reuse the confounders and add residual links again. Specifically,

$$
\bar{o}_i^{t+1} = \varphi([\mathcal{G}_c^t : \tilde{o}_i^t : \hat{o}_i^t], [\tilde{e}_{i,j}^t : \hat{e}_{i,j}^t]), t = 0, ..., T-1 \quad (9)
$$

where $\bar{o}_i^{t+1}$ is denoted as the predicted 3D position of objects at next timestep. Finally, we can make long-term counterfactual predictions by applying CTS iteratively. In contrast to the previous work, we have a completely different definition of the causal graph $\mathcal{G}_c^t$ in our paper. Our node denotes the

physical properties and our edge denotes the directly contaction, so that it can better depict the latent causal relations.

Training and Optimizing

The training of perception module is separated from the rest of the whole model. It takes the raw images from video sequences as input and outputs the 3D positions of each object in the image. The inference module and the forward module are jointly trained in an end-to-end fashion without any access to the ground truth information about confounders. We train it with the following loss function:

$$\mathcal{L}_{e2e} = \sum_{t=0}^{T} \sum_{m=1}^{M} \mathcal{L}_{mse}(x_m^t, x_m^{t*}) \tag{10}$$

where $\mathcal{L}_{e2e}$ represents mean squared error measuring the gap between the predicted 3D position x_m^t and the ground truth 3D position x_m^{t*}.

Experiments

Environment. We evaluate our model in two benchmarks (Baradel et al. 2020). One involves stacking blocks in 3D space to construct towers in each sample, another one consists of one moving object and one static object (cylinder or sphere) in each sample. Both of the benchmarks are parameterized by unobservable variables. They are also denoted as the confounders, such as masses, friction coefficients and gravity. We follow the evaluation protocol in the original paper in both two benchmarks. Mean Square Error (MSE) is used to measure the prediction performance (lower score indicates better performance).

$BlocktowerCF$: Each sample shows K ($K = 3, 4$) cubes initialized with a random position and angle. The $do - operation$ includes the removal and displacement for one of the blocks. In our experiments we use all data from the dataset, which contains 51.7k, 14.6k and 7.3k samples for training, validation and testing separately.

$CollisionCF$: Each sample shows a moving object colliding with a static object, and the $do - operation$ is just about displacement for one object. Similar to the previous one, we exploit all samples from the benchmark, which consists of 30k, 7.9k and 6k data for training, validation and testing respectively.

The evaluation of our model's performance can be described in the following two aspects:

- Whether the counterfactual prediction made by the model is accurate?
- How well can the model generalize to the settings that are unseen in the training period, including unseen number of blocks as well as unseen type of objects?

Counterfactual Prediction

We compare our model with the state-of-the-art method by evaluating the mean square error between the counterfactual prediction and the ground truth. The state-of-the-art method CoPhyNet (Baradel et al. 2020) is a counterfactual prediction method as ours. The Interaction Network (IN) (Battaglia et al. 2016) captures the interactions across all objects in the environment, and the Network Physics Engine (NPE) (Chang et al. 2016) models object interactions by considering only neighbouring objects. Both of them are non-counterfactual baselines. They make future predictions directly based on the past frames after $do - operation$ without regard for the confounders. In comparison to the state-of-the-art method, our model performs better in all settings, as demonstrated in Table 1 and Table 2. Furthermore, it also outperforms two non-counterfactual baselines by a large margin. Figure 3 and Figure 4 are the qualitative comparisons between our method and CoPhyNet in two benchmarks, further demonstrating that our model can better infer the future states.

Method	3→3	3→4	4→4	4→3
IN	31.8	52.4	52.1	34.2
NPE	33.1	52.3	52.8	35.0
CoPhyNet	29.4	48.2	45.3	30.1
CoPhyNet∗	23.7	48.0	45.2	26.0
Ours	**22.5**±0.3	**47.6**±0.2	**43.4**±0.1	**24.7**±0.2

Table 1: Comparison between the performance of our method with the state-of-the-art method on the $BlocktowerCF$. NumA→NumB (such as 3→4) means we train the model on the dataset that contains NumA objects, while we test the model on the dataset that contains NumB objects. CoPhyNet is cited from (Baradel et al. 2020). CoPhyNet∗ represents the reproduce results. The value is the MSE (*100) on 3D pose average over time.

Method	all→all	s → c	c → s
IN	70.1	71.5	72.0
NPE	69.7	71.0	69.9
CoPhyNet	17.3	22.0	15.2
CoPhyNet∗	14.8	13.4	13.8
Ours	**13.0**±0.3	**9.6**±0.2	**10.2**±0.4

Table 2: Comparison between the performance of our method with the state-of-the-art method on the $CollisionCF$. The s represents the sphere object, while the c represents the cylinder object. TypeA→TypeB (such as s→c) means we train on the dataset that the moving object is of TypeA, while we test on the dataset that the moving object is of TypeB. CoPhyNet is cited from (Baradel et al. 2020). CoPhyNet∗ represents the reproduced results. The value is the MSE (*100) on 3D pose average over time.

Ablation Study

To have a better understanding of how each submodule contributes to the final results, we conduct experiments about two variants and evaluate them in both $BlocktowerCF$ and $CollisionCF$ dataset, shown in Table 3.

Global Causal Relation Attention (GCRA). We first show the effectiveness of our GCRA module. Compared to the CoPhyNet, we replace the confounders estimation component with the GCRA while the remainder of the model

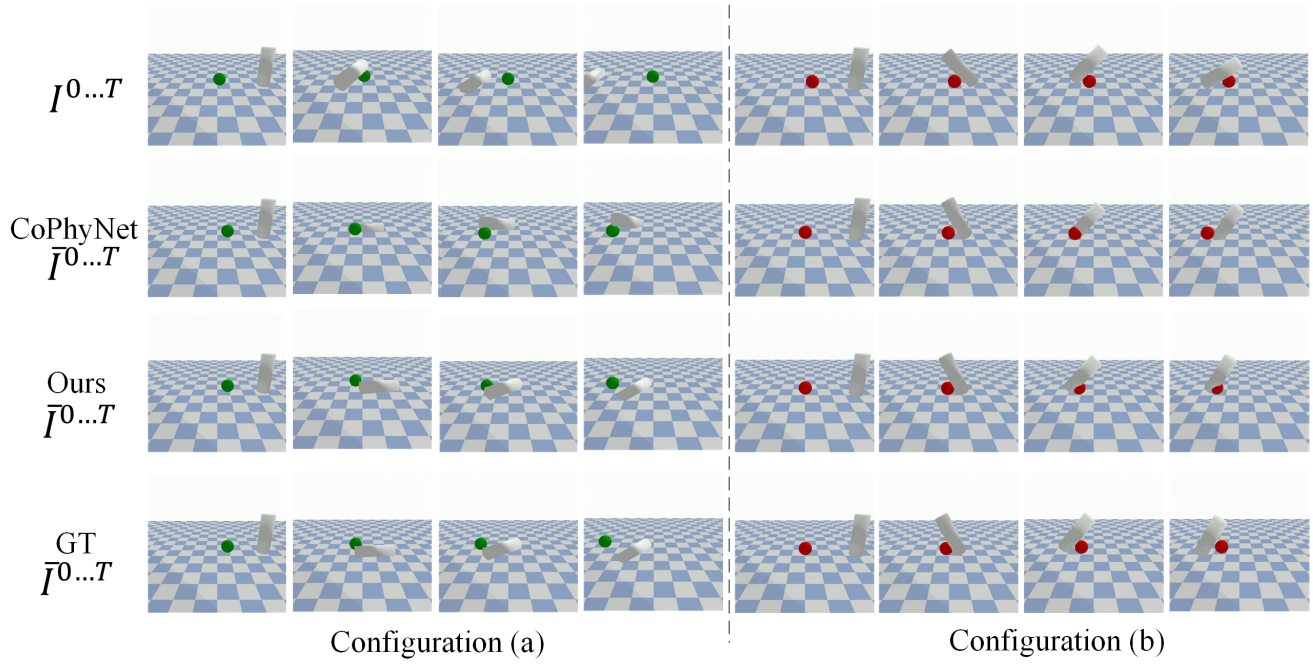


Figure 3: We show the qualitative comparisons of counterfactual predictions between our method and CoPhyNet in the $CollisionCF$ scenario. $I^{0...T}$ represent the original sequence, while $\bar{I}^{0...T}$ represent the counterfactual outcome sequence.

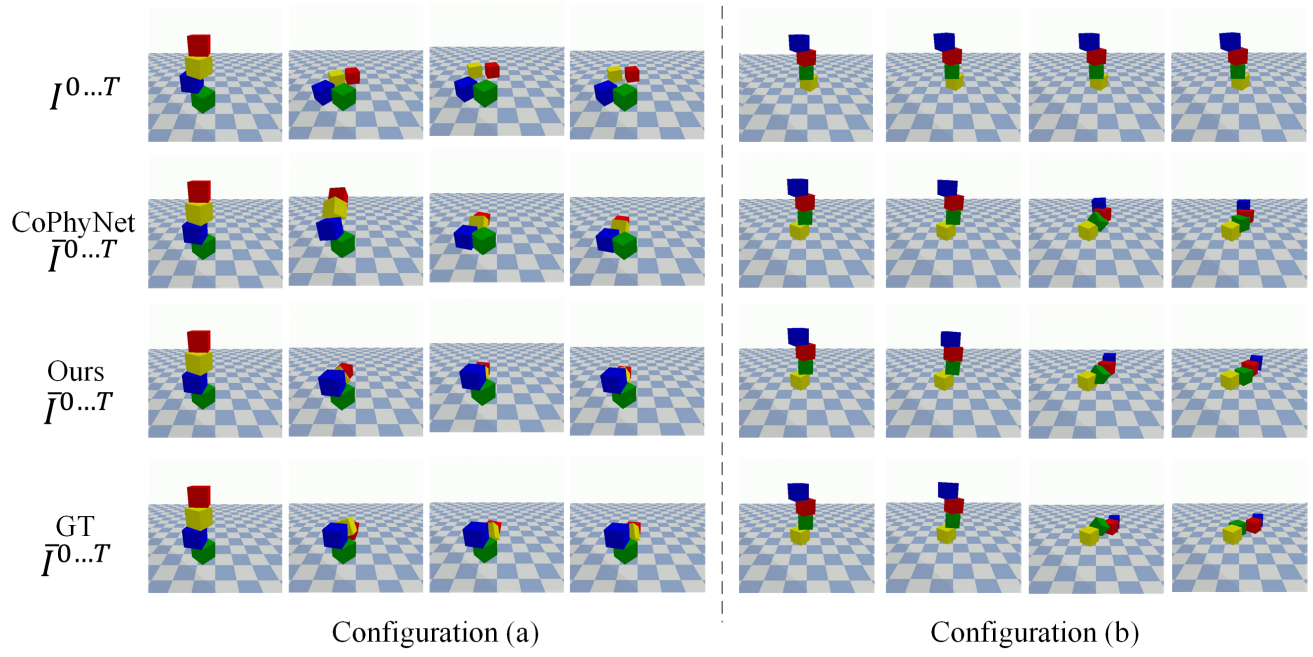


Figure 4: We show the qualitative comparisons of counterfactual predictions between our method and CoPhyNet in the $BlocktowerCF$ scenario. $I^{0...T}$ represent the original sequence, while $\bar{I}^{0...T}$ represent the counterfactual outcome sequence.

Component		BlocktowerCF	CollisionCF
GCRA	CTS		
		0.452	0.148
✓		0.445	0.141
✓	✓	**0.434**	**0.130**

Table 3: Ablative results on 4→4 subset of $BlocktowerCF$ and all→all subset of $CollisionCF$. The value is the MSE on 3D pose average over time.

Method	3→3	3→4	4→4	4→3
CoPhyNet w/o GT	0.294	0.482	0.453	0.301
CoPhyNet GT	0.296	**0.467**	0.481	0.297
Ours w/o GT	0.225	0.476	0.434	0.247
Ours GT	**0.222**	0.493	**0.405**	**0.243**

Table 4: Comparison results of the forward submodule in different methods using ground truth confounder quantities (GT) or using the estimated confounder quantities (w/o GT) as the input on the $BlocktowerCF$. The value is the MSE on 3D pose average over time.

remains the same. GCRA not only bridges the relationships across objects in a single frame, but also applies long-term cross-frame self-attention. As shown in Table 3, GCRA promotes the ability of predicting counterfactual outcomes. It captures more spatial and temporal information about the objects than the vanilla RNN.

Confounder Transmission Structure (CTS). In CTS, we consider an unobservable causal graph additionally to better depict relations across different objects. Besides, we reuse the confounders as well as the updated object embeddings in a residual way. The results are presented in Table 3, demonstrating that our method has better performance than CoPhyNet. Overall, by combining both GCRA and CTS, we achieve the best results.

Generalization

We also evaluate our model's ability to generalize to unseen physical environments. Specifically, in $blocktowerCF$ dataset, the data for training and testing contains different number of objects (3→4 and 4→3). The results shown in Table 1 demonstrate our model can better extrapolate to new physical settings well. Besides, in $CollisionCF$ dataset, we also evaluate the setting that the moving object in training data and testing data has different types (cylinder or sphere). As shown in Table 2, though there is a great discrepancy between the objects, our model displays a strong ability to generalize to novel environments, achieves the best performance among all the other methods.

Forward Module with GT Confounders

To further prove the superiority of our CTS submodule, we assume the ground truth confounder labels are accessible and see if CTS can well utilize the information from confounders. Specifically, we replace the estimated confounders with the ground truth confounder quantities in both CoPhyNet and our model, the results are shown in Table 4. We can see CoPhyNet achieves even worse performance in some settings, showing that it cannot make good use of information from ground truth confounders. In contrast, our prediction module achieves higher performance with the help of ground truth confounders by maintaining vital information during transmission.

Method	4→4	all→all
CoPhyNet	0.463	0.161
Ours	**0.447**	**0.132**

Table 5: The performance using high-dimensional embeddings as input on 4→4 subset in $BlocktowerCF$ and all→all subset in $CollisionCF$. The value is the MSE on 3D pose average over time in validation set.

Modelling using High-dimensional Input

Previous works claim that desired counterfacutal prediction should be performed on the high-dimensional embeddings that encode appearance, context and scenarios. However, most of them use 3D coordinates of objects as the input. In this experiment, we explore whether our method can extend to this challenging problem and handle high-dimensional embeddings besides 3D positions. To this end, we employ Region Proposal Interaction Network (*RPIN*) (Qi et al. 2021), which is an interaction network based on the region proposals proposed recently to extract a feature map as a high-level representation. Based on the features extracted by *RPIN*, we incorporate it with our model by replacing the 3D coordinates $\bar{o}_i^0$ in Equation 1 with the features. Besides, we also replace the input for CoPhyNet. As shown in Table 5, our method still outperforms CoPhyNet, validating that our model can adapt more to this challenging problem with high-level features as the input.

Conclusion

In this paper, we propose a neural network framework combining Global Causal Relation Attention (GCRA) and Confounder Transmission Structure (CTS) to estimate latent confounders and predict counterfactual outcomes. Our model captures the spatial-temporal information through inter-frame and cross-frame attention manners. We also encode the estimated confounders into forward module and propagate them in a residual way. Our model outperforms state-of-the-art counterfactual model on challenging benchmarks. Further experiments also show the robustness and generalization ability of our model when confronted with various input information. We hope that our method could serve as a strong framework for future studies of visual reasoning, especially in intuitive physics learning.

Appendices

Generalization results on BlocktowerCF To further evaluate the generalization ability of our model, we conduct the experiment on unseen confounder combinations in the

$BlocktowerCF$. In comparison to other methods, our model generalizes well in novel environments, as shown in Table 6.

Method	3→3	4→4
IN	0.298	0.480
NPE	0.319	0.476
CoPhyNet	0.289	0.423
CoPhyNet∗	0.282	0.447
Ours	**0.268**±0.002	**0.420**±0.003

Table 6: Comparison between the performance of our method with the state-of-the-art method on the $BlocktowerCF$. NumA→NumB (such as 3→3) means we train the model on the dataset that contains NumA objects, and we test the model on the dataset that contains NumB objects. Test confounder configurations are different from training period (50/50 split). CoPhyNet is cited from (Baradel et al. 2020). CoPhyNet∗ represents the reproduce results. The value is the MSE on 3D pose average over time.

Generalization results on CLEVRER We study the model performance on CLEVRER (Yi* et al. 2020) to see whether our model can generalize to scenes with more objects and more frames. For each video in CLEVRER, several questions in different aspects are posed to measure the model's reasoning and understanding ability. Despite of the descriptive questions ("what happened?") that most visual question answering (VQA) datasets own, CLEVRER contains the explanatory questions ("why did the collision happen?"), predictive questions ("what will happen next?"), and counterfactual questions ("what would happen under an unrealized situation").

To apply our model on CLEVRER, we simplify both the GCRA module and the CTS module to incorporate them with the IODINE (Greff et al. 2019). IODINE uses an amortized iterative variational framework to learn latent object representations. We aim to use our model to enhance the latent representations by modeling object relationships. Therefore, we make a few minor adjustments to our model to better fit this dataset. Inspired by (Tang et al. 2022), we first use the GCRA module to estimate the physical property, and then employ the CTS module to compute the object dynamics. We conduct all experiments on CLEVRER based on ALOE (Ding et al. 2021). The main difference between our method with ALOE is that ALOE takes scene representations from MONet (Burgess et al. 2019), while we leverage the representations produced by CTS. Table 7 shows the performance comparison between our model with other methods. Our model outperforms IODINE in most settings, indicating that our model strengthens the latent representations by estimating the physical property and computing the object dynamics. Further comparison experiments with ALOE show that our model achieves comparable results with state-of-the-art method, which demonstrates the effectiveness of our model when generalized to more complex environments.

Method	Des	Exp	Pre	Cou
MAC (V+) (Yi* et al. 2020)	86.4	22.3	42.9	25.1
NS-DR (Yi* et al. 2020)	88.1	79.6	68.7	42.2
DCL (Chen et al. 2021)	90.7	82.8	82.0	46.5
IODINE (Greff et al. 2019)	92.8	95.6	80.2	71.3
ALOE (Ding et al. 2021)	94.0	**96.0**	87.5	**75.6**
Ours	**94.8**	95.5	**89.2**	75.4

Table 7: Performance (per question accuracy) comparison with the state-of-the-art methods MAC (V+), NS-DR, D-CL, IODINE and ALOE on CLEVRER. Des, Exp, Pre and Cou in the table represent the "Descriptive", "Explanatory", "Predictive" and "Counterfactual" respectively.

Dataset Both the CoPhy and the CLEVRER are synthetic video datasets of physical events. Each video in CoPhy contains 30 frames for $BlocktowerCF$ and 15 frames for $CollisionCF$ at resolution 224 × 224, we use all the frames for training, validating and testing. Each video in CLEVRER contains 128 frames at resolution 480 × 320, we extract images every 4 frames and ensure that at least one collision event is included, as in (Tang et al. 2022).

Acknowledgements

This work was supported by the National Key Research & Development Program (No. 2020YFC2003901), Chinese National Natural Science Foundation Projects #61876178, #61806196, #61976229, #62176256, #62106264, Youth Innovation Promotion Association CAS (#Y2021131), and the InnoHK program.

References

Agrawal, P.; Nair, A.; Abbeel, P.; Malik, J.; and Levine, S. 2016. Learning to poke by poking: Experiential learning of intuitive physics. *arXiv preprint arXiv:1606.07419*.

Baradel, F.; Neverova, N.; Mille, J.; Mori, G.; and Wolf, C. 2020. CoPhy: Counterfactual Learning of Physical Dynamics. In *International Conference on Learning Representations*.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V. F.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; aglar Glehre; Song, H. F.; Ballard, A. J.; Gilmer, J.; Dahl, G. E.; Vaswani, A.; Allen, K. R.; Nash, C.; Langston, V.; Dyer, C.; Heess, N.; Wierstra, D.; Kohli, P.; Botvinick, M.; Vinyals, O.; Li, Y.; and Pascanu, R. 2018. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261.

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*, 110(45): 18327–18332.

Battaglia, P. W.; Pascanu, R.; Lai, M.; Rezende, D. J.; and Kavukcuoglu, K. 2016. Interaction Networks for Learning about Objects, Relations and Physics. In *NIPS*, 4502–4510.

Burgess, C. P.; Matthey, L.; Watters, N.; Kabra, R.; Higgins, I.; Botvinick, M.; and Lerchner, A. 2019. Monet: Un-

supervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*.

Chalupka, K.; Perona, P.; and Eberhardt, F. 2014. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*.

Chang, M. B.; Ullman, T.; Torralba, A.; and Tenenbaum, J. B. 2016. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*.

Chen, Z.; Mao, J.; Wu, J.; Wong, K.-Y. K.; Tenenbaum, J. B.; and Gan, C. 2021. Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning. In *International Conference on Learning Representations*.

Cho, K.; van Merrienboer, B.; aglar Glehre; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*, 1724–1734.

Ding, D.; Hill, F.; Santoro, A.; Reynolds, M.; and Botvinick, M. 2021. Attention over Learned Object Embeddings Enables Complex Visual Reasoning. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Ehrhardt, S.; Monszpart, A.; Mitra, N. J.; and Vedaldi, A. 2019. Taking visual motion prediction to new heightfields. *Computer Vision and Image Understanding*, 181: 14–25.

Finn, C.; Goodfellow, I.; and Levine, S. 2016. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29: 64–72.

Finn, C.; and Levine, S. 2017. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2786–2793. IEEE.

Fraccaro, M.; Kamronn, S.; Paquet, U.; and Winther, O. 2017. A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning. In *NIPS*, 3604–3613.

Fragkiadaki, K.; Agrawal, P.; Levine, S.; and Malik, J. 2016. Learning Visual Predictive Models of Physics for Playing Billiards. In *ICLR (Poster)*.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Greff, K.; Kaufman, R. L.; Kabra, R.; Watters, N.; Burgess, C.; Zoran, D.; Matthey, L.; Botvinick, M.; and Lerchner, A. 2019. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, 2424–2433. PMLR.

Groth, O.; Fuchs, F. B.; Posner, I.; and Vedaldi, A. 2018. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 702–717.

Ha, D.; and Schmidhuber, J. 2018. Recurrent World Models Facilitate Policy Evolution. In *NeurIPS*, 2455–2467.

Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; and Davidson, J. 2019. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2555–2565. PMLR.

Hamrick, J.; Battaglia, P.; and Tenenbaum, J. B. 2011. Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd annual conference of the cognitive science society*, volume 2. Citeseer.

Janner, M.; Levine, S.; Freeman, W. T.; Tenenbaum, J. B.; Finn, C.; and Wu, J. 2019. Reasoning About Physical Interactions with Object-Centric Models. In *International Conference on Learning Representations*.

Jia, Z.; Gallagher, A. C.; Saxena, A.; and Chen, T. 2014. 3d reasoning from blocks to stability. *IEEE transactions on pattern analysis and machine intelligence*, 37(5): 905–918.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *International Conference on Learning Representations*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Kubricht, J. R.; Holyoak, K. J.; and Lu, H. 2017. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10): 749–759.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*, 1369–1379.

Lerer, A.; Gross, S.; and Fergus, R. 2016. Learning physical intuition of block towers by example. In *International conference on machine learning*, 430–438. PMLR.

Li, M.; Yang, M.; Liu, F.; Chen, X.; Chen, Z.; and Wang, J. 2020a. Causal World Models by Unsupervised Deconfounding of Physical Dynamics. *arXiv preprint arXiv:2012.14228*.

Li, W.; Azimi, S.; Leonardis, A.; and Fritz, M. 2016. To fall or not to fall: A visual approach to physical stability prediction. *arXiv preprint arXiv:1604.00066*.

Li, W.; Leonardis, A.; and Fritz, M. 2017. Visual stability prediction for robotic manipulation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2606–2613. IEEE.

Li, Y.; He, H.; Wu, J.; Katabi, D.; and Torralba, A. 2020b. Learning Compositional Koopman Operators for Model-Based Control. In *International Conference on Learning Representations*.

Li, Y.; Torralba, A.; Anandkumar, A.; Fox, D.; and Garg, A. 2020c. Causal discovery in physical systems from videos. *arXiv preprint arXiv:2007.00631*.

Li, Y.; Wu, J.; Tedrake, R.; Tenenbaum, J. B.; and Torralba, A. 2019. Learning Particle Dynamics for Manipulating

Rigid Bodies, Deformable Objects, and Fluids. In *International Conference on Learning Representations*.

Liu, Z.; Freeman, W. T.; Tenenbaum, J. B.; and Wu, J. 2018. Physical primitive decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Scholkopf, B.; and Bottou, L. 2017. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6979–6987.

Lopez-Paz, D.; and Oquab, M. 2016. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*.

Martin-Ordas, G.; Call, J.; and Colmenares, F. 2008. Tubes, tables and traps: great apes solve two functionally equivalent trap tasks but show no evidence of transfer across tasks. *Animal cognition*, 11(3): 423–430.

Mottaghi, R.; Bagherinezhad, H.; Rastegari, M.; and Farhadi, A. 2016. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3521–3529.

Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3: 96–146.

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*.

Riochet, R.; Castro, M. Y.; Bernard, M.; Lerer, A.; Fergus, R.; Izard, V.; and Dupoux, E. 2018. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*.

Rojas-Carulla, M.; Baroni, M.; and Lopez-Paz, D. 2018. Causal Discovery Using Proxy Variables.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Spelke, E. S.; and Kinzler, K. D. 2007. Core knowledge. *Developmental science*, 10(1): 89–96.

Stanić, A.; and Schmidhuber, J. 2019. R-SQAIR: relational sequential attend, infer, repeat. *arXiv preprint arXiv:1910.05231*.

Sun, C.; Shrivastava, A.; Vondrick, C.; Murphy, K.; Sukthankar, R.; and Schmid, C. 2018. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 318–334.

Tang, Q.; Zhu, X.; Lei, Z.; and Zhang, Z. 2022. Object dynamics distillation for scene decomposition and representation. In *International Conference on Learning Representations*.

Tompson, J.; Schlachter, K.; Sprechmann, P.; and Perlin, K. 2017. Accelerating eulerian fluid simulation with convolutional networks. In *International Conference on Machine Learning*, 3424–3433. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

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 *BMVC*, volume 2, 7.

Wu, J.; Lu, E.; Kohli, P.; Freeman, B.; and Tenenbaum, J. 2017. Learning to see physics via visual de-animation. *Advances in Neural Information Processing Systems*, 30: 153–164.

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. *Advances in neural information processing systems*, 28: 127–135.

Ye, T.; Wang, X.; Davidson, J.; and Gupta, A. 2018. Interpretable intuitive physics model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 87–102.

Yi*, K.; Gan*, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; and Tenenbaum, J. B. 2020. CLEVRER: Collision Events for Video Representation and Reasoning. In *International Conference on Learning Representations*.