INTERVENTION-BASED RECURRENT CAUSAL MODEL FOR NONSTATIONARY VIDEO CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

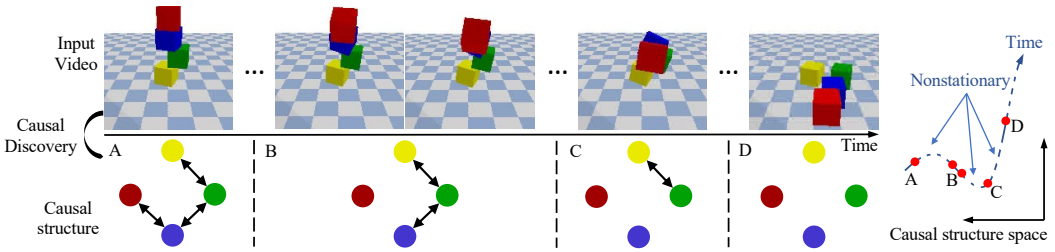

Figure 1: Illustration of nonstationary causal structures in physical systems. Casual structures can be represented as a graph, where edges indicate interaction between objects. Take the example of the supporting force in the block falling sequence, the graph changes over time, posing a challenge to video causal discovery methods.

ABSTRACT

Nonstationary causal structures are prevalent in real-world physical systems. For example, the stacked blocks interact until they fall apart, while the billiard balls move independently until they collide. However, most video causal discovery methods can not discover such nonstationary casual structures due to the lack of modeling for the instantaneous change and the dynamics of the causal structure. In this work, we propose the *Intervention-based Recurrent Casual Model* (IRCM) for nonstationary video casual discovery. First, we extend the existing intervention-based casual discovery framework for videos to formulate the instantaneous change of the causal structure in a principled manner. Then, we use a recurrent model to sequentially predict the causal structure model based on previous observations to capture the nonstationary dynamic of the causal structure. We evaluate our method on two popular physical system simulation datasets with various types of multi-body interactions. Experiments show that the proposed IRCM achieves the state-of-the-art performance on both the counterfactual reasoning and future forecasting tasks.

1 INTRODUCTION

Causal reasoning from visual input is essential for intelligence systems in understanding the complex mechanisms in the physical world. For instance, autonomous vehicles need to infer the unseen causal structures on the road that drives the state evolution of other agents across time to anticipate future events better accordingly. One main obstacle in discovering such causal structures is the dynamic nature of events. In Figure 1, we illustrate the varying casual relationship in a simple multi-body system where the stacked blocks fall to the ground. In nonstationary video sequences, the causal structure can have abrupt changes and/or long-term dependencies, posing challenges for casual graphical models (CGM).

For the first challenge, most CGMs in video causal understanding can not handle abrupt causal relationship changes. Li et al. (2020) (VCDN, Figure 2a) partially address this issue by learning a stationary causal summary graph, where causal structures are learned but fixed throughout the video. Zheng et al. (2018) (DYNOTEARS, Figure 2b) relaxed such fixed structure settings by assuming a

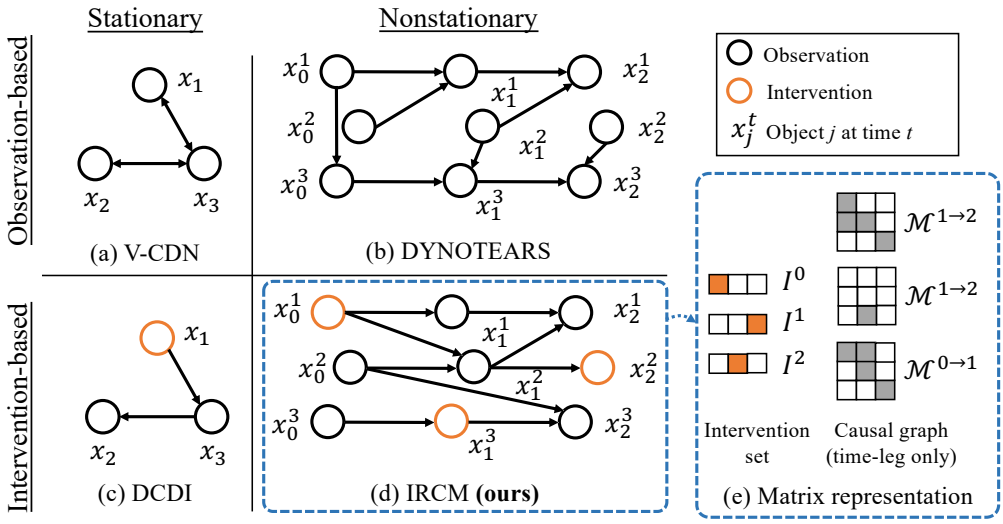

Figure 2: Comparison of existing causal structure representations for causal discovery. For stationary causal models, (a) Li et al. (2020) propose the double-edged causal summary graph and (b) Zheng et al. (2018) model both the instantaneous interactions (*i.e.,* among objects at the same time step) and time-leg interactions (*i.e.,* among objects at different time steps). For nonstationary causal models, (c) Gerhardus & Runge (2020) learns both the causal graph structure and the intervention set (*i.e.,* nodes in orange color). (d) We propose the intervention-based recurrent causal model (IRCM) to use interventions to model nonstationary time-leg interactions. (e) We visualize the matrix representation of the intervention set and causal graph for the given example.

stationary order for the period bigger than 1. On the other hand, Brouillard et al. (2020) (DCDI, Figure 2c) recently proposes a differentiable causal model for a spatial graph to naturally capture the abrupt change of probability distributions during interventions. In this work, we naturally extend the intervention-based causal model to the graph with time-leg edges in videos, *i.e.,* current objects' states are fully determined by previous states (Figure 2d).

For the second challenge, most CGMs in video causal understanding purely depend on the object state observations. That is the causal graph at time t is conditionally independent from the causal graph at time $t-1$ given the object states' observations. Illustrated in Figure 1, CGMs that can be represented as graphs can be modeled as a trajectory in the nonstationary video. In this work, we adopt a recurrent network to sequentially predict CGM to model the trajectories.

Based on the intuitions above, we propose the Intervention-based Recurrent Casual Model (IRCM) to better capture the dynamics in nonstationary videos. As the ground truth CGMs are often not directly measurable, we adopt two popular downstream tasks to benchmark the efficacy of the proposed model: counterfactual reasoning and future state forecasting. Deducing the alternative results countering the reality over the discovered CGM can directly express the impacts of causality. Also, the causal knowledge endows better insights into which factors affect the target variable and how to manipulate the system properly.

We summarize the contribution of this work as follows:

- We introduce the IRCM model to extend the previous intervention-based causal discovery framework to nonstationary video sequences.
- We propose to use recurrent networks to capture the long-term trajectory of Causal Graph Models (CGM) and provide optimization solution to train recurrent networks together with downstream causal models.
- We achieve state-of-the-art performance on two downstream tasks: counterfactual reasoning and future forecasting on two standard benchmark datasets (CoPhy (Baradel et al., 2020), Fabric Manipulation (Brouillard et al., 2020)) by showing an averaged improvement of 11% across 9 metrics.

2 RELATED WORK

Causal Discovery of Stationary Models. Given the input time-series data, the goal is to uncover one fixed directed acyclic graph (DAG), where edges represent the direct causal relationships among variables. There are two main approaches: observation-based and intervention-based. The observation-based approach fully relies on the passive observation of the input system. Constraint-based methods rely on conditional independence tests as constraint-satisfaction to recover Markov-Equivalent Graphs (Spirtes et al., 2000; Entner & Hoyer, 2010; Colombo et al., 2011). Score-based methods assign a score to each DAG, and perform searching in this score space (Chickering, 2002; Zheng et al., 2018). The third class of methods exploits such asymmetries or causal footprints to uniquely identify a DAG (Shimizu, 2014; Zhang & Hyvärinen, 2009).

In practice, domain experts may design interventional experiments and collect additional data of the input system. The intervention-based approach aims to combine such interventional data with the observational data for a better identifiability of the causal structure (Eberhardt, 2012; Eberhardt et al., 2012). However, many of current approaches (Hyttinen et al., 2013; Ghassami et al., 2018b; Kocaoglu et al., 2017; Wang et al., 2017; Shanmugam et al., 2015; Peters et al., 2016; Rothenhäusler et al., 2015; Ke et al., 2019) either assume full knowledge of the intervention, make strong assumptions about the model class, or have scalability limitations. Recently, Brouillard et al. (2020) utilizes the continuous-constrained framework to model the interventions with neural network models. In contrast, our proposed method aims to uncover nonstationary causal structures.

Causal Discovery of Nonstationary Models. To extend to nonstationary data, recent works discover causal models in each sliding window separately, and then compare and merge them. Adams & MacKay (2007) explicitly detect the change points and divide the time series into stationary processes. To implicitly model the change of the causal model, Huang et al. (2015) assume certain smoothness properties and Zhang et al. (2017) use kernel distribution embeddings to describe shifting probabilistic distributions. Later, the problem was reformulated with the online parameter learning framework (Song et al., 2009; Xing et al., 2010). To tackle the varying instantaneous causal relations, both linear (Ghassami et al., 2018a; Huang et al., 2019; Huang & Zhang, 2019; Huang et al., 2020a) and nonlinear (Huang et al., 2020b) causal models are proposed. Our proposed method treats the nonstationary changes of the system as interventions and re-purposes the intervention-based framework to discover time-varying causal graph structures.

Video Causal Discovery. The relevant literature in the computer vision community has accumulated several efforts to tackle down the challenges of video modeling and prediction (Ye et al., 2019; Hsieh et al., 2018; Yi et al., 2020). Nevertheless, one topic that had enjoyed recent success is reasoning objective dynamics in a video sequence. A line of research attempts to solve this task by modeling the correlations in a spatio-temporal context, such as (Yi* et al., 2020; Chen et al., 2021; Bakhtin et al., 2019; Qi et al., 2021; Zhang et al., 2021). However, focusing on modeling the dependencies substantially might not suffice to offer clear interpretations of object dynamics as we humans do. Addressing this issue, the authors of (Baradel et al., 2020) and (Li et al., 2020) try to make efforts to introduce causal knowledge (Schölkopf et al., 2021; Bengio et al., 2020; Runge et al., 2019) to this task. A few works adapt various topics into such a context. Whereas neither of them is able to fully uncover the causal structure underlying the video sequences.: CoPhyNet (Baradel et al., 2020) derives an alternative output based on a known causal graph; VCDN (Li et al., 2020) focus on recovering the stationary causal structures from the video. Instead, our proposed method apply the new intervention-based method to capture nonstationary causal structures.

3 METHODOLOGY

In this section, we present Intervention-based Recurrent Casual Model (IRCM) for non-stationary video causal discovery. We first give an overview of model architecture, as shown in Figure 3, then dive into two components of IRCM , Recurrent Network and Intervention-based Causal Model.

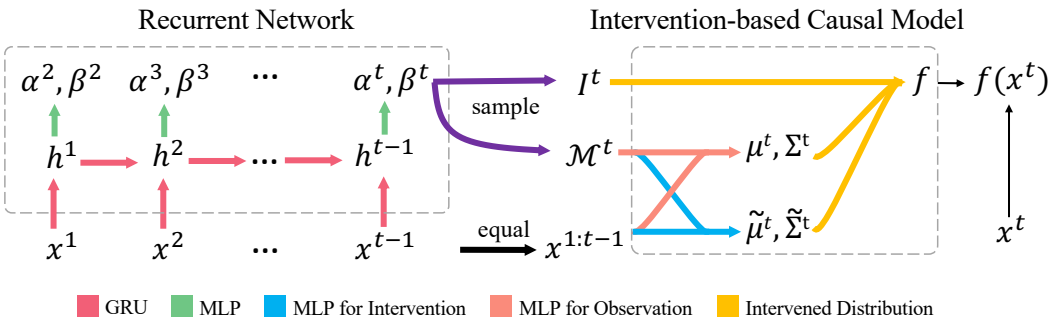

Figure 3: Model designs of the proposed IRCM. The model has two modules: recurrent network (RN) and intervention-based causal model (ICM). Given the input video frame feature $\boldsymbol{x}^{t-1}$, RN updates the hidden states h^t and predicts probability values (α^t, β^t) to sample DAG structure $\mathcal{M}^t$ and intervention set I^t, respectively. With observation $\boldsymbol{x}^{1:t-1}$ and causal structure representation $(\mathcal{M}^t,\ I^t)$, ICM predicts the mean and covariance of multivariate Gaussian distribution for both observation and intervention sets, (μ^t, Σ^t) and $(\tilde{\mu}^t, \tilde{\Sigma}^t)$, resulting the probability density function f.

3.1 Problem Formulation

We factorize the joint probability of a temporal sequence into a sequential form:

$$p(\boldsymbol{x}^{1:T}; \theta) = p(\boldsymbol{x}^1; \theta) \prod_{t=2}^{T} p(\boldsymbol{x}^t | \boldsymbol{x}^{1:t-1}; \theta), \tag{1}$$

where θ is the model parameters to learn. This formulation makes it easy to do future forecasting by conditioning any unknown $\boldsymbol{x}^t$ on observed or previously predicted history $\boldsymbol{x}^{1:t-1}$. For simplicity, we decode multiple frames in an autoregressive way, i.e., at each timestep, we predict $\hat{\boldsymbol{x}}^t$ as the mode of $p(\boldsymbol{x}^t | x^{1:t-1}; \theta)$ and do further prediction conditioning on this prediction.

Furthermore, we decompose the density function into Recurrent Network (RN) and Intervention-based Causal Model (ICM) by:

$$f_\theta(\boldsymbol{x}^t | \boldsymbol{x}^{1:t-1}; \theta) = f_{\text{ICM}}(\boldsymbol{x}^t | \mathcal{M}^t, I^t, \boldsymbol{x}^{1:t-1}; \theta_{\text{ICM}}) \tag{2}$$

$$\mathcal{M}^t, I^t \sim \text{Bern}(\alpha^t, \beta^t) \tag{3}$$

$$\alpha^t, \beta^t = \text{RN}(\boldsymbol{x}^{1:t-1}; \theta_{\text{RN}}) \tag{4}$$

In this way, we extend the framework of Continuous constrained optimization for structure learning to sequential data.

3.2 Model Designs

Intervention-based Casual Model. Formally, given the observed d agents in the scene from time 1 to T, a joint probability distribution $f(\boldsymbol{x})$ depict their state through time. In the context of Causal Graph Model (CGM) (Pearl et al., 2016), a directed acyclic graph (DAG) $\mathcal{G}$ with dT nodes defines $f(\boldsymbol{x})$, where node $\boldsymbol{x}_j^t$ is associates with agent j at time step t. Directed edges represents causal relationships. The distribution of agent states at time t can be factorized as:

$$f(\boldsymbol{x}^t | \boldsymbol{x}^{1:t-1}; \theta) = \prod_{j=1}^{d} f(\boldsymbol{x}_j^t | Pa(\boldsymbol{x}_j^t); \theta), \tag{5}$$

where $Pa(\boldsymbol{x}_j^t)$ pertains to the set of parent nodes of $\boldsymbol{x}_j^t$ in $\mathcal{G}$. Eq. 5 implicitly hypothesizes the causal sufficiency (Peters et al., 2017), i.e., our work does not involve any hidden confounding elements. Also, we neither consider the instantaneous edges nor edges that go back in time in this work. Simply put, $Pa(\boldsymbol{x}_j^t) \subseteq \{\boldsymbol{x}_j^i\}_{i<t}$. This feature makes our causal graph fully identifiable in the context of video sequence as Li et al. (2020).

Eq. 5 allows us to swap $f(\boldsymbol{x}_j^t | Pa(\boldsymbol{x}_j^t))$ with another conditional distribution, which is called interventions. One intervention target set $I \subseteq V$ is a subset of graph nodes where interventions are

exerted. We consider an intervention family $\mathcal{I} = \{I_k\}_{k=1}^{K}$. In particular, $I_1 = \emptyset$ denotes the observed distribution. We furthur use I_k^t to denote intervened nodes at time t in the kth intervention family. Given an interventional family I_{k}, we formalize the intervened distribution at time t by:

$$f^{(k)}(\boldsymbol{x}^t) = \prod_{j \notin I_k^t} f^{(1)}(\boldsymbol{x}_j^t | Pa(\boldsymbol{x}_j^t)) \prod_{j \in I_k^t} f^{(k)}(\boldsymbol{x}_j^t | Pa(\boldsymbol{x}_j^t)). \tag{6}$$

In our case, we use $k = 2$, assuming only one intervention family. Following Brouillard et al. (2020), we use neural networks (NN) to output the parameters of density function $\tilde{f}$, e.g., Gaussian.

$$f^{(1)} = \tilde{f}(.; \mathrm{NN}(., \phi_j^t)), \quad f^{(2)} = \tilde{f}(.; \mathrm{NN}(., \psi_j^t)), \tag{7}$$

where ϕ and ψ are parameters for the observational and interventional density function respectively. Thus, Eq. 6 can be written as:

$$f_{\mathrm{ICM}}(\boldsymbol{x}^t | \mathcal{M}^t, I^t, \boldsymbol{x}^{1:t-1}; \theta_{\mathrm{ICM}}) = \prod_{j \notin I_2^t} \tilde{f}\left(\boldsymbol{x}_j^t; \mathrm{NN}(\mathcal{M}_j^t \odot \boldsymbol{x}; \phi_j^t)\right) \prod_{j \in I_2^t} \tilde{f}\left(\boldsymbol{x}_j^t; \mathrm{NN}(\mathcal{M}_j^t \odot \boldsymbol{x}; \psi_j^t)\right), \tag{8}$$

where $\mathcal{M}_j^t \in \{0,1\}^{dT}$ is a binary vector indicating the parents of $\boldsymbol{x}_j^t$ and $\odot$ is the Hadamard product.

In specific, two separate neural networks with identical architecture are used to predict mean vectors and diagonal covariance matrices to parameterize the multivariate Gaussian distributions for our $\tilde{f}$,

$$\mu^t, \Sigma^t = \mathrm{NN}(\mathcal{M}^t \odot \boldsymbol{x}; \phi^t), \tag{9}$$

$$\tilde{\mu}^t, \tilde{\Sigma}^t = \mathrm{NN}(\mathcal{M}^t \odot \boldsymbol{x}; \psi^t), \tag{10}$$

for observational and interventional distributions respectively. In summary, $\theta_{\mathrm{ICM}} = \{\phi, \psi\}$.

Causal Graph Sampling. Direct prediction of graph structure in its binary from $\mathcal{M}^t$ and I^t is difficult and can lead to mode collapse. Following DCDI (Brouillard et al., 2020), we choose to capture it through multivariate Bernoulli distributions.

In specific, an upstream module Recurrent Network (RN) will predict real matrix α^t and real vector β^t, which are of the same shape as $\mathcal{M}^t$ and I^t. Then we sample binary values in the following way:

$$\mathcal{M}^t \sim \mathrm{Bern}(\alpha^t), \tag{11}$$

$$I^t \sim \mathrm{Bern}(\beta^t). \tag{12}$$

All elements are mutually independent. Optimization difficulty incurred by sampling process is solved by Straight-Through Gumbel estimator (Jang et al., 2016; Maddison et al., 2016).

Recurrent Network. In the Recurrent Network (RN), we are concerned with modeling the distribution of a graph structure given previous observations $\boldsymbol{x}^{1:t-1}$. We consider all time-lagged but not instantaneous causal relations in this model. Thus at time step t, we need to predict graph structure with all its previous $t-1$ frames. We group these graphs into $\mathcal{M}^t \in \{0,1\}^{d^2 \times (t-1)}$. For intervention, it is a vector $I^t \in \{0,1\}^d$.

$$\boldsymbol{h}^t = f_{\mathrm{GRU}}(\boldsymbol{h}^{t-1}, \boldsymbol{x}^t; \theta_{\mathrm{GRU}}) \tag{13}$$

$$\alpha^t, \beta^t = f_{\mathrm{MLP}}(\boldsymbol{h}^t; \theta_{\mathrm{MLP}}) \tag{14}$$

To model the non-stationary nature of real-world physical systems, we use an two-layer Gated Recurrent Unit (GRU) (Chung et al., 2014) to model temporal dependencies and an MLP to predict the likelihood of existing causal relations α^t and successful intervention β^t. In summary, $\theta_{\mathrm{RN}} = \{\theta_{\mathrm{GRU}}, \theta_{\mathrm{MLP}}\}$.

3.3 Learning and Inference

Learning. We do not have access to the ground-truth graph structure. This motivates us to follow DCDI (Brouillard et al., 2020), which serves as the pedestal of our work, to train IRCM in a manner of continuous constrained optimization problem. The core of our objective treats learning by maximizing the regularized log-likelihood in Eq. 8 conditioning on the object states :

$$\begin{aligned} \mathcal{L} = \sum_k \mathbb{E}_{\boldsymbol{x} \sim P\boldsymbol{x}} \log f(\boldsymbol{x}) - \zeta \sum_{(j,t)} ||\mathcal{M}_j^t||_0 - \eta \sum_t ||I^t||_1 \\ \text{s.t. } \mathrm{Tr}(e^{\sigma(\alpha^t)}) - d = 0 \end{aligned} \tag{15}$$

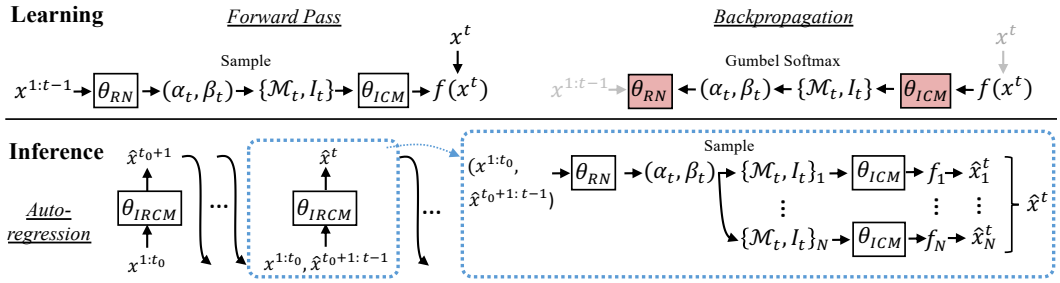

Figure 4: Learning and inference of the proposed IRCM. During learning, we first make the forward pass to predict the Bernoulli distribution parameters (α^t, β^t) and sample causal structure representation $(\mathcal{M}^t, I^t)$ to predict the probability density function f. Then, we backpropagate the loss to update the parameters in neural network models, θ_{RN} and θ_{ICM}. During inference, we recursively feed the predicted $\hat{x^t}$ into the model to estimate the next time step state via the Monte Carlo method.

ζ and η are hyperparameters to control the sparsity of causal graphs and intervention sets respectively. Because we do not consider instantaneous causal relations nor relations go back in time, the learnt graph is guaranteed to be a DAG. Thus, IRCM naturally meets the requirement of the acyclicity constraint $\text{Tr}(e^{\sigma(\alpha^t)}) - d = 0$ (Zheng et al., 2018).

In order to estimate the gradient of α^t and β^t with regard to $\mathcal{L}$, we choose to follow DCDI (Brouillard et al., 2020) utilize the Straight-Through Gumbel estimator (Jang et al., 2016; Maddison et al., 2016). This is equivalent to using discrete Bernoulli samples during forward passing and Gumbel-Softmax samples during backpropagation.

Inference. During inference time, as shown in Figure 4, we use the observed and previously predicted sequence $\{\boldsymbol{x}^{1:t_0}, \hat{\boldsymbol{x}}^{t_0+1:t-1}\}$ to predict the multivariate distribution of $\boldsymbol{x}^t$ (t_0 is the length of observed sequence). We then do a secondary optimization to predict $\hat{\boldsymbol{x}}^t$:

$$\begin{aligned}\hat{\boldsymbol{x}}^t &= \arg\max_{\boldsymbol{x}^t} \left(f(\boldsymbol{x}^t|\boldsymbol{x}^{1:t-1};\theta\right) \\ &= \sum_{(\mathcal{M}^t, I^t)} \arg\max_{\boldsymbol{x}^t} \left(f(\boldsymbol{x}^t|\mathcal{M}^t, I^t; \theta_{\text{ICM}})\right) p(\mathcal{M}^t, I^t|\boldsymbol{x}^{1:t-1}; \theta_{\text{RN}})\end{aligned} \tag{16}$$

$$\hat{\boldsymbol{x}}_j^t = \sum_{(\mathcal{M}^t, I^t)} (\mu_j^t)^{\delta(j \notin I^t)} (\tilde{\mu_j^t})^{\delta(j \in I^t)} p(\mathcal{M}^t, I^t|\boldsymbol{x}^{1:t-1}; \theta_{\text{RN}}), \tag{17}$$

where $\delta(j \in I^t)$ is the delta function indicating if object i is in the intervention set I^t. In practice, we take the Monte Carlo approach to first sample $(\mathcal{M}^t, I^t)$ according to the distribution and average the predicted mean values from either the observation and the intervention set.

4 Experiment

4.1 Experimental setups

Downstream tasks and Datasets. We conduct experiments to understand the efficacy of our proposed IRCM in terms of discovering the causal structure to estiamte the object dynamics across time. More specifically, the counterfactual reasoning and future forecasting in the video sequences are selected to demonstrate this point.

Task 1: Counterfactual Reasoning. This problem is formalized as follows (Baradel et al., 2020): During training, we first infer the causal structure upon a set of visual observations. The objective is to reason the counterfactual outcome given the modified initial object state. The Counterfactual Physics benchmark (CoPhy) (Baradel et al., 2020) dataset contains two types of sequences, observational and counterfactual. The latter sequence is built upon changing the initial object state from the observations with other factors ((such as inertia, gravity or friction)) untouched. CoPhy comprises three physical scenarios in total: BlockTowerCF, BallsCF and CollisionCF. Each scenario provides the 3D positions of all objects in the scene. BlockTowerCF also includes a binary label for stability.

Table 1: Quantitative results on three physical scenarios of the CoPhy dataset and Fabric dataset. We compare with state-of-the-art methods: CoPhyNet (Baradel et al., 2020), and C-CDN (Li et al., 2020). Non-existent experiments are marked by hyphens.

	CoPhy Dataset (Baradel et al., 2020)							Fabric Dataset (Li et al., 2020)	
	BlocktowerCF			BallsCF		CollisionCF			
	MSE ↓	NLL ↓	Acc. ↑	MSE ↓	NLL ↓	MSE ↓	NLL ↓	MSE ↓	NLL ↓
CoPhyNet	0.49	8.52	73.8	1.92	2.76	0.22	6.97	-	-
V-CDN	-	-	-	-	-	-	-	0.0028	0.32
IRCM (Ours)	**0.42**	**7.55**	**77.9**	**1.61**	**2.45**	**0.20**	**6.65**	**0.0024**	**0.30**

Task 2: Future Forecasting. Future forecasting refers to discerning unknown object future given the observed histories. We use the Fabric Manipulation (FM) (Li et al., 2020) dataset for future forecasting task, where 2D coordinates from learned keypoints in the dynamics scene are provided.

Implementation Details. For both tasks, we use the same model architectures and the same settings for learning and inference. On each dataset, we directly use the extracted visual features from video frames in the previous state-of-the-art methods. Below are the details.

Visual Features. For observation x^t, we use the extracted visual features from input videos to improve the model performance. For a fair comparison on CoPhy, we adopt the identical experimental protocols in (Baradel et al., 2020) to examine the generalizability of IRCM. We train and test with 4 objects on BlockTowerCF and BallsCF. The experiments on CollisionCF utilize all types of objects (spheres and cylinders) for both training and test. Moreover, following the settings opted in (Li et al., 2020), we first extract the 2D positions of key points from a pretrained DNN-based mechanism (Kulkarni et al., 2019) to represent the fabrics. Our experiments proceeds by observe first 5 time steps and foresee object states for next 20 time steps for training, and forecast the forthcoming 5 steps upon previous 5 steps for test. We first encode these location information with an MLP as the object states for our model.

Model Architectures. We append two independent three-layer MLPs on a two-layer GRU to predict both α^t and β^t. At the time instance τ, α_τ is then reshaped to a set of $d \times d$ matrices for $\mathcal{M}^t$. Notably, we zero-padded these matrices to ensure there exists $t-1$ individual matrix in total per time instance for backpropagation. For the faster learning convergence, we place an instance normalization layer before each ReLU activation in the MLP model and use the sigmoid activation for the final output to make it a probability value.

Learning and Inference. In our experiments, RMSProp optimizer (Goodfellow et al., 2016) are employed with the learning rate initialized at 8×10^{-5} . Our implementation uses PyTorch. The experiments are executed on four Nvidia GeForce TITAN XPs, with 48 GB of memory in total.

Evaluation Metrics. Since none of the aforementioned datasets provide annotations for the causal graphical model, we gauge model performance by the observed object dynamics which is generated from the unobserved causal structure. Thus the ideal metrics should rely on object states, i.e., coordinates and stability. In particular, we aim to understand how close the outcomes can approximate the ground truth. To this end, we calculate the mean square error (MSE) and the negative log-likelihood (NLL) (Ivanovic & Pavone, 2019) on coordinates of objects between ground-truth and prediction. NLL is the average negative log-likelihood between a ground truth trajectory distribution determined by a kernel density estimate and the predicted trajectory. In addition, the stability classification accuracy are used for our experiment on BlockTowerCF. Lower NLL and MSE and higher accuracy are preferred.

4.2 Benchmark Results

As per comparing methods, we are primarily interested in assessing our IRCM versus two leading studies on estimating agent states in a video sequence in the context of learning CGM. More specifically, CoPhyNet (Baradel et al., 2020), which achieves cutting-edge results on the CoPhy benchmark and the VCDN framework (Li et al., 2020), which performs best on FM, are selected.

Table 2: Ablation studies on the representative BallsCF scenario in CoPhy (Baradel et al., 2020). We justify the design choices of the proposed IRCM in its two modules: Intervention-based Causal Model (ICM) and Recurrent Network (RN).

	IRCM	ICM			RN	
		IRCM w/o $\mathcal{M}, I$	IRCM w/o I	IRCM-local$\mathcal{M}$	IRCM-stationary	IRCM-indep.
MSE ↓	**1.61**	2.95	1.79	1.85	1.93	1.68
NLL ↓	**2.45**	3.82	2.57	2.64	2.70	2.51

CoPhyNet summarizes the problem with a given causal structure to handle the object dynamics over time and approache object interactions with fully-connected graph convolution (Kipf & Welling, 2016; Battaglia et al., 2018). VCDN provides a model that infers a summary graph consists of time-lagged causal relations as shown in Figure 2. To the best of our knowledge. these two methods are the most relevant ones to ours.

We train our algorithm on CoPhy by the exact training objective Eq. 15 on BallsCF, CollisionCF, and FM. For BlockTowerCF, we also include the stability classification term for a fair comparison:

$$\mathcal{L} = \sum_k \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}} \log f^k(\boldsymbol{x}) - \left(\zeta \sum_{(j,t)} ||\mathcal{M}_j^t||_0 + \eta \sum_t ||I^t||_1 + \text{CE}(\hat{\mathcal{S}}^t, \mathcal{S}^t) \right), \tag{18}$$

where the CE term is the cross entropy between predicted and ground-truth stability. We forward the predicted locations and learnt $\mathcal{M}^t$ to a pre-trained GCN for the stability estimation.

It can be seen in Table 1 that out model consistently beat baselines. It demonstrate the necessity of capturing nonstationary causal structures and intervention-based causal discovery.

4.3 Ablation Studies

The proposed IRCM has two main components: Intervention-based Causal Model and Recurrent Network. Below, we justify their design choices with the following ablation studies (Table 2).

Intervention-based Causal Model (ICM). The ICM model relies on the causal DAG structure $\mathcal{M}$ and the intervention set I. Below, we demonstrate their necessities by the ablation studies.

Importance of Causal Graphical Model ($\mathcal{M}, I$). IRCM w/o $\mathcal{M}, I$ treats the counterfactual reasoning task as future forecasting on both sequences by not transferring the learnt causal structure from observatinoal to counterfactual sequences. We can see in Table 2 that this significantly hurt the performance of IRCM . In fact, IRCM w/o $\mathcal{M}, I$ shows the worst scores on both metrics. The comparisons of those values against other methods overwhelmingly demonstrate the necessity and merit to take the causal structure into account for video future forecasting.

Importance of Intervention (I). We justify the advantages of using interventional distribution to discover the causal structure in a video sequence over IRCM w/o I, which directly approximates Eq. 5 from the observations. We can observe the large performance gap between IRCM w/o I and IRCM , demonstrating the impacts of interventions concerning learning the causal structure.

Importance of long-term $\mathcal{M}$. IRCM-markov serves to verify the advantages of IRCM treating $\mathcal{M}$ as a $d^2 \times (t-1)$ matix. The scores of IRCM in Table 2 considerably exceed IRCM-local$\mathcal{M}$. We attribute this to the property of IRCM evidently offering better capability to learn the causal relationships than setting $t = 2$. The advantages of IRCM also convey the message that the impacts of the agent states in several previous time instances can impact on the current agent states. Additionally, the results favor IRCM over CoPhyNet (Baradel et al., 2020) can be attributed to a similar reason.

Recurrent Network (RN). Instead of the sequential modeling of the causal graphical structures with RN, we can predict a single structure or a sequence of structures that are temporally independent.

Importance of Nonstationary Modeling. IRCM-stationary assumes an invariant causal structure over time, thus shares the similar idea with V-CDN (Li et al., 2020), *i.e.,* we assume that the learned

($\mathcal{M}^t$, I^t) and the weight of NN remain static. As shown in Table 2, IRCM significantly outperforms IRCM-stationary, fitting better the time-varying structures in the video sequences. This result emphasizes the importance of considering nonstationary structures in temporal modeling.

Importance of Sequential Modeling. We evaluate the advantages of extrapolating $\mathcal{M}^t$ through our RN against IRCM-indep that learns $\mathcal{M}^t$ independently at each time step. Table 2 suggest that our IRCM significantly outperforms IRCM-indep, demonstrating the advantages of the sequential modeling of causal structures.

5 Conclusion

In this paper, we propose an intervention-based recurrent casual model for video causal discovery. IRCM differs from works the literature in that it introduces the interventions to discover the causal structure for understanding the object dynamics in video sequences. At its core, we introduce a recurrent network to model the interventional distributions. This formulation allows us to grasp the time-varying property that widely exists in video sequences. Experiment results justify that our IRCM delivers better performance in both counterfactual reasoning and future forecasting compared with prior works. One direction is to loose the sufficiency assumption and involve the confounding elements to our framework to enable discovering the causal relationships in real-world applications.

References

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5082–5093, 2019.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual Learning of Physical Dynamics. In *International Conference on Learning Representations*, 2020.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxWIgBFPS.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable Causal Discovery from Interventional Data. In *Advances in Neural Information Processing Systems*, 2020.

Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B. Tenenbaum, and Chuang Gan. Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=bhCDO_cEGCz.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional dags with latent and selection variables. In *UAI*, pp. 850, 2011.

Frederick Eberhardt. Almost optimal intervention sets for causal discovery. *arXiv preprint arXiv:1206.3250*, 2012.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *arXiv preprint arXiv:1207.1389*, 2012.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pp. 121–128, 2010.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. In *Advances in Neural Information Processing Systems*, 2020.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, 31:6266, 2018a.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pp. 1724–1733. PMLR, 2018b.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. *arXiv preprint arXiv:1806.04166*, 2018.

B Huang, K Zhang, M Gong, and C Glymour. Causal discovery from non-identical variable sets. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020a.

Biwei Huang and Kun Zhang. Specific and shared causal relation modeling and mechanism-based clustering. *Advances in neural information processing systems*, 2019.

Biwei Huang, Kun Zhang, and Bernhard Schölkopf. Identification of time-dependent causal model: A gaussian process treatment. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *International Conference on Machine Learning*, pp. 2901–2910. PMLR, 2019.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020b.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2375–2384, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

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

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Nips*, 2017.

Tejas D Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised Learning of Object Keypoints for Perception and Control. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10724–10734, 2019.

Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal Discovery in Physical Systems from Videos. In *Advances in Neural Information Processing Systems*, pp. 9180–9192, 2020.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Haozhi Qi, Xiaolong Wang, Deepak Pathak, Yi Ma, and Jitendra Malik. Learning Long-term Visual Dynamics with Region Proposal Interaction Networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=_X_4Akcd8Re.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems*, 2015.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in Earth system sciences. *Nature communications*, 10(1):1–13, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. *Advances in Neural Information Processing Systems*, 2015.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Le Song, Mladen Kolar, and Eric Xing. Time-varying dynamic bayesian networks. *Advances in neural information processing systems*, 22:1732–1740, 2009.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Yuhao Wang, Liam Solus, Karren Dai Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 2017.

Eric P Xing, Wenjie Fu, and Le Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, pp. 535–566, 2010.

Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10353–10362, 2019.

Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision Events for Video Representation and Reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9736–9746, 2021.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, volume 647. AUAI Press, 2009.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, pp. 1347. NIH Public Access, 2017.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9492–9503, 2018.