BISCUIT: Causal Representation Learning from Binary Interactions

Phillip Lippe[1] **Sara Magliacane**[2,3] **Sindy Löwe**[2] **Yuki M. Asano**[1] **Taco Cohen**[4] **Efstratios Gavves**[1]

[1]QUVA Lab, University of Amsterdam
[2]AMLab, University of Amsterdam
[3]MIT-IBM Watson AI Lab
[4] Qualcomm AI Research*, Amsterdam, Netherlands

Abstract

Identifying the causal variables of an environment and how to intervene on them is of core value in applications such as robotics and embodied AI. While an agent can commonly interact with the environment and may implicitly perturb the behavior of some of these causal variables, often the targets it affects remain unknown. In this paper, we show that causal variables can still be identified for many common setups, e.g., additive Gaussian noise models, if the agent's interactions with a causal variable can be described by an unknown binary variable. This happens when each causal variable has two different mechanisms, e.g., an observational and an interventional one. Using this identifiability result, we propose BISCUIT, a method for simultaneously learning causal variables and their corresponding binary interaction variables. On three robotic-inspired datasets, BISCUIT accurately identifies causal variables and can even be scaled to complex, realistic environments for embodied AI.

1 INTRODUCTION

Learning a low-dimensional representation of an environment is a crucial step in many applications, *e.g.*, robotics (Lesort et al., 2018), embodied AI (Kolve et al., 2017) and reinforcement learning (Hafner et al., 2021; Träuble et al., 2022). A promising direction for learning robust and actionable representations is *causal representation learning* (Schölkopf et al., 2021), which aims to identify the underlying causal variables and their relations in a given environment from high-dimensional observations, *e.g.*, images. However, learning causal variables from high-dimensional observations is a considerable challenge and may not always be possible, since multiple underlying causal systems could generate the same data distribution (Hyvärinen et al., 1999). To overcome this, several works make use of additional information, *e.g.*, by using counterfactual observations (Ahuja et al., 2022; Brehmer et al., 2022; Locatello et al., 2020), observed intervention targets (Lippe et al., 2022a, 2023). Alternatively, one can restrict the distributions of causal variables, *e.g.*, by considering environments with non-stationary noise (Khemakhem et al., 2020a; Yao et al., 2022a,b) or sparse causal relations (Lachapelle et al., 2022a,b).

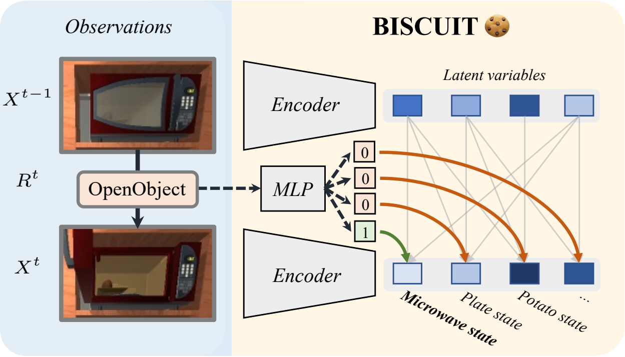


Figure 1: BISCUIT identifies causal variables from images X^{t-1} and X^t, by learning to encode an observable regime variable R^t, *e.g.*, an action, as binary variables. Conditioning each latent on one of these binary variables identifies causal variables in environments like iTHOR (Kolve et al., 2017).

In this paper, instead, we focus on interactive environments, where an agent can perform actions which may have an effect on the underlying causal variables. We will assume that these interactions between the agent and the causal variables can be described by *binary* variables, *i.e.*, that with the agent's actions, we can switch between two mechanisms, or distributions, of a causal variable, similarly to performing soft interventions. Despite being binary, these interactions include a wide range of common scenarios, such as a robot pressing a button, opening/closing a door, or even colliding with a moving object and alternating its course.

[*] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

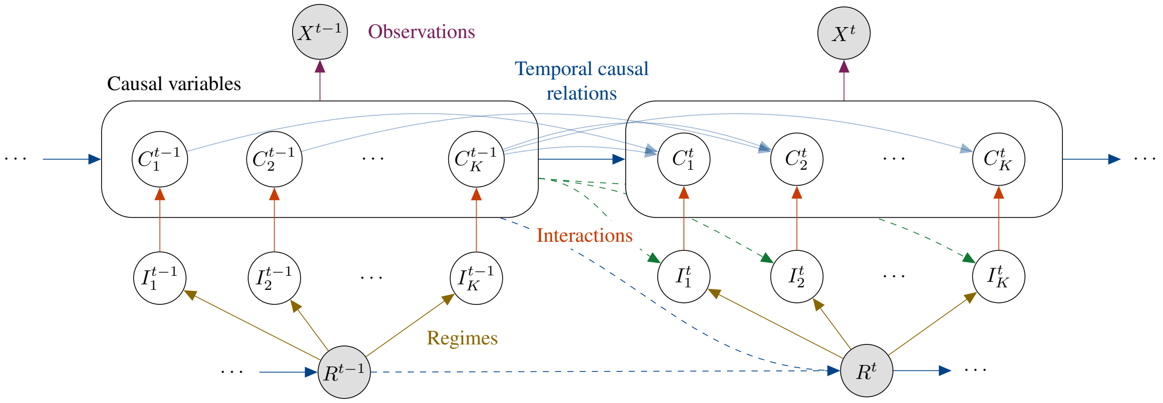


Figure 2: A representation of our assumptions. Observed variables are shown in gray (X^τ and R^τ) and latent variables in white. Optional causal edges are shown as dashed lines. A latent causal variable C_i^t has as parents a subset of the causal factors at the previous time step $C^{t-1} = \{C_1^{t-1}, \ldots, C_K^{t-1}\}$, and its latent binary interaction variable I_i^t. The interaction variables are determined by an observed regime variable R^t and potentially by the variables from the previous time step C^{t-1} (*e.g.*, in a collision). The regime variable can be a dynamical process over time as well, for example, by depending on the previous time step. The observation X^τ is a high-dimensional entangled representation of all causal variables C^τ at time step τ.

In this setup, we prove that causal variables are identifiable if the agent interacts with each causal variable in a distinct pattern, *i.e.*, does not always interact with any two causal variables at the same time. We show that for K variables, we can in many cases fulfill this by having as few as $\lfloor \log_2 K \rfloor + 2$ actions with sufficiently diverse effects, allowing identifiability even for a limited number of actions. The binary nature of the interactions permits the identification of a wider class of causal models than previous work in a similar setup, including the common, challenging additive Gaussian noise model (Hyvärinen et al., 1999).

Based on these theoretical results, we propose BISCUIT (**B**inary **I**nteraction**s** for **C**a**u**sal **I**den**t**ifiability). BISCUIT is a variational autoencoder (Kingma et al., 2014) which learns the causal variables and the agent's binary interactions with them in an unsupervised manner (see Figure 1). In experiments on robotic-inspired datasets, BISCUIT identifies the causal variables and outperforms previous methods. Furthermore, we apply BISCUIT to the realistic 3D embodied AI environment iTHOR (Kolve et al., 2017), and show that BISCUIT is able to generate realistic renderings of unseen causal states in a controlled manner. This highlights the potential of causal representation learning in the challenging task of embodied AI. In summary, our contributions are:

- We show that under mild assumptions, binary interactions with unknown targets identify the causal variables from high-dimensional observations over time.
- We propose BISCUIT, a causal representation learning framework that learns the causal variables and their binary interactions simultaneously.
- We empirically show that BISCUIT identifies both the causal variables and the interaction targets on three robotic-inspired causal representation learning benchmarks, and allows for controllable generations.

2 PRELIMINARIES

In this paper, we consider a causal model $\mathcal{M}$ as visualized in Figure 2. The model $\mathcal{M}$ consists of K latent causal variables $C_1, ..., C_K$ which interact with each other over time, like in a dynamic Bayesian Network (DBN) (Dean et al., 1989; Murphy, 2002). In other words, at each time step t, we instantiate the causal variables as $C^t = \{C_1^t, ..., C_K^t\} \in \mathcal{C}$, where $\mathcal{C} \subseteq \mathbb{R}^K$ is the domain. In terms of the causal graph, each variable C_i^t may be caused by a subset of variables in the previous time step $\{C_1^{t-1}, ..., C_K^{t-1}\}$. For simplicity, we restrict the temporal causal graph to only model dependencies on the previous time step. Yet, as we show in Appendix B.3, our results in this paper can be trivially extended to longer dependencies, *e.g.*, $(C^{t-2}, C^{t-1}) \to C^t$, since C^{t-1} is only used for ensuring conditional independence. As in DBNs, we consider the graph structure to be time-invariant.

Besides the intra-variable dynamics, we assume that the causal system is affected by a regime variable R^t with arbitrary domain $\mathcal{R}$, which can be continuous or discrete of arbitrary dimensionality. This regime variable can model any known external causes on the system, which, for instance, could be a robotic arm interacting with an environment. For the causal graph, we assume that the effect of the regime variable R^t on a causal variable C_i^t can be described by a latent *binary interaction* variable $I_i^t \in \{0, 1\}$. This can be interpreted as each causal variable having two mechanisms/distributions, *e.g.*, an observational and an interventional mechanism, which has similarly been assumed in previous work (Brehmer et al., 2022; Lippe et al., 2022a, 2023). Thereby, the role of the interaction variable I_i^t is to select the mechanism, *i.e.*, observational or interventional, at time step t. For example, a collision between an agent and an object is an interaction that switches the dynamics of the object from its natural course to a perturbed one. In this paper, we consider the interaction variable I_i^t to be an unknown function of the regime variable and the previous causal variables, *i.e.*, $I_i^t = f_i(R^t, C^{t-1})$. The dependency on the previous time step allows us to model interactions that only occur in certain states of the system, *e.g.*, a collision between an agent (modeled by R^t) with an object with position C_i^{t-1} will only happen for certain positions of the agent and the object.

We consider the causal graph of Figure 2 to be causally sufficient, *i.e.*, we assume there are no other unobserved confounders except the ones we have described in the previous paragraphs and represented in the Figure, and that the causal variables within the same time step are independent of each other, conditioned on the previous time step and their interaction variables. We summarize the dynamics as $p(C^t|C^{t-1}, R^t) = \prod_{i=1}^{K} p(C_i^t|C^{t-1}, I_i^t)$. Although C_i^t only depends on a subset of C^{t-1}, w.l.o.g. we model it as depending on all causal variables from the previous time step.

In causal representation learning, the task is to identify causal variables from an entangled, potentially higher-dimensional representation, *e.g.*, an image. We consider an injective observation function g, mapping the causal variables C^t to an observation $X^t = g(C^t)$. Following Klindt et al. (2021); Yao et al. (2022b), we assume g to be defined everywhere for C^t and differentiable almost everywhere. In our setting, once we identify the causal variables, the causal graph can be trivially learned by testing for conditional independence, since the causal graph is limited to edges following the temporal dimension, *i.e.*, from C^{t-1} to C^t. We provide further details on the graph discovery and an example on learned causal variables in Appendix B.4.

3 IDENTIFYING CAUSAL VARIABLES

Our goal in this paper is to identify the causal variables $C_1, ..., C_K$ of a causal system from sequences of observations (X^t, R^t). We first define the identifiability class that we consider. We then provide an intuition on how binary interactions enable identifiability, before presenting our two identifiability results. The practical algorithm based on these results, BISCUIT, is presented in Section 4.

3.1 IDENTIFIABILITY CLASS AND DEFINITIONS

Intuitively, we seek to estimate an observation function $\hat{g}$, which maps a latent space $\hat{\mathcal{C}}$ to observations X, and models each true causal variable C_i in a different dimension of the latent space $\hat{\mathcal{C}}$. This observation function should be equivalent to the true observation function g, up to permuting and transforming the variables individually, *e.g.*, through scaling. Several previous works (Khemakhem et al., 2020a; Lachapelle et al., 2022b; Yao et al., 2022a,b) have considered equivalent identifiability classes, which we define as:

Definition 3.1. *Consider a model $\mathcal{M} = \langle g, f, \omega, \mathcal{C} \rangle$ with an injective function $g(C) = X$ with $C \in \mathcal{C}$ and a latent distribution $p_\omega(C^t|C^{t-1}, R^t)$, parameterized by ω and defined:*

$$p_\omega(C^t|C^{t-1}, R^t) = \prod_{i=1}^{K} p_{\omega,i}\big(C_i^t|C^{t-1}, f_i(R^t, C^{t-1})\big),$$

where $f_i : \mathcal{R} \times \mathcal{C} \to \{0, 1\}$ outputs a binary variable for the variable C_i^t. We call $\mathcal{M}$ identifiable iff for any other model $\widetilde{\mathcal{M}} = \langle \tilde{g}, \tilde{f}, \tilde{\omega}, \tilde{\mathcal{C}} \rangle$ with the same observational distribution $p(X^t|X^{t-1}, R^t)$, g and $\tilde{g}$ are equivalent up to a component-wise invertible transformation T and a permutation π:

$$p_{\mathcal{M}}(X^t|X^{t-1}, R^t) = p_{\widetilde{\mathcal{M}}}(X^t|X^{t-1}, R^t) \Rightarrow g = \tilde{g} \circ T \circ \pi$$

To achieve this identifiability, we rely on the interaction variables I_i^t being binary and having *distinct interaction patterns*, a weaker form of faithfulness on the interaction variables. Intuitively, we do not allow that any two causal variables to have identical interaction variables I_i^t, I_j^t across the whole dataset, *i.e.*, being always interacted with at the same time. Similarly, if all I_i^t are always zero ($\forall t, i\colon I_i^t = 0$), then we fall back into the well-known unidentifiable setting of non-linear ICA (Hyvärinen et al., 1999). Since interaction variables can also be functions of the previous state, we additionally assume that for all possible previous states, the interaction variables cannot be deterministic functions of any other. Thus, we assume that all causal variables have *distinct interaction patterns*, which we formally define as:

Definition 3.2. *A causal variable C_i in $\mathcal{M} = \langle g, f, \omega, \mathcal{C} \rangle$ has a* ***distinct interaction pattern*** *if for all values of C^{t-1}, its interaction variable $I_i^t = f_i(R^t, C^{t-1})$ is not a deterministic function $b\colon \{0, 1\} \to \{0, 1\}$ of any other I_j^t:*

$$\forall C^{t-1}, \forall j \neq i, \nexists b, \forall R^t\colon\ f_i(R^t, C^{t-1}) = b(f_j(R^t, C^{t-1})).$$

This assumption generalizes the intervention setup of Lippe et al. (2022b), which has a similar condition on its binary intervention variables, but assumed them to be independent of the previous time step. This implies that we can create a distinct interaction pattern for each of the K causal variables by having as few as $\lfloor \log_2 K \rfloor + 2$ different values for R^t, if the interaction variables are independent of C^{t-1}. In contrast, other methods in similar setups that also exploit an external, temporally independent, observed variable (Khemakhem et al., 2020a; Yao et al., 2022a,b) require the number of regimes to scale linearly with the number of causal variables. If the interaction variables depend on C^{t-1}, the lower bound of the number of different values for R^t depends on the causal model $\mathcal{M}$, more specifically its interaction functions f_i. Concretely, the lower bound for a causal model $\mathcal{M}$ is the smallest set of values of R^t that ensure different interaction patterns for all C^{t-1} in $\mathcal{M}$. In the worst case, each C^{t-1} may require different values of R^t to fulfill the condition of Theorem 3.2, such that R^t would need to

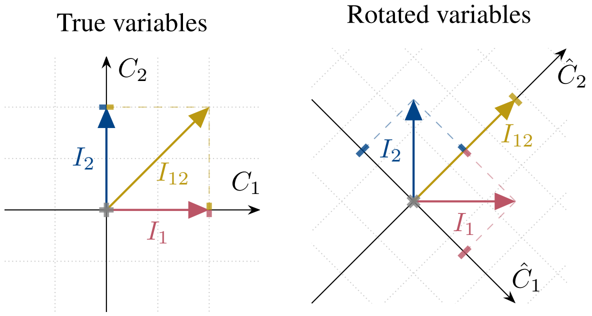


Figure 3: Binary interactions identify the additive Gaussian noise model in Equation (1). The plots show the change of the mean for each variable for a fixed C^{t-1} under interactions affecting only one variable ($I_1 = 1$ or $I_2 = 1$), and under joint interactions I_{12} ($I_1 = I_2 = 1$). Ticks on axes show mean values for each variable, where the mean for ($I_1 = 0, I_2 = 0$) lies at the origin. The colors of the ticks match the interaction color. **Left**: true causal variables C_1 and C_2. Each variable has two possible means after any of the possible interactions. The effect of the interactions can be described by a binary variable per axis. **Right**: a rotated representation. Both $\hat{C}_1$ and $\hat{C}_2$ have three possible means, which cannot be described by a binary variable per axis anymore.

be of the same domain as C^{t-1} (for instance being continuous). At the same time, for models $\mathcal{M}$ in which the condition of Theorem 3.2 can be fulfilled by the same values of R^t for all C^{t-1}, we again recover the lower bound of $\lfloor \log_2 K + 2 \rfloor$ different values of R^t.

3.2 INTUITION: ADDITIVE GAUSSIAN NOISE

We first provide some intuition on how binary interactions, *i.e.*, knowing that each variable has exactly two potential mechanisms, enable identifiability, even when we do not know which variables are interacted with at each time step. We take as an example an additive Gaussian noise model with two variables C_1, C_2, each described by the equation:

$$C_i^t = \mu_i(C^{t-1}, I_i^t) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where ϵ_i is additive noise with variance σ^2, and μ_i a function for the mean with $\mu_i(C^{t-1}, I_i^t = 0) \neq \mu_i(C^{t-1}, I_i^t = 1)$. Due to the rotational invariance of Gaussians, the true causal variables C_1, C_2 and their rotated counterparts $\hat{C}_1, \hat{C}_2$ model the same distribution with the same factorization: $\prod_{i=1}^{2} p_i(C_i^t|C^{t-1}, R^t) = \prod_{i=1}^{2} \hat{p}_i(\hat{C}_i^t|\hat{C}^{t-1}, R^t)$. This property makes the model unidentifiable in many cases (Hyvärinen et al., 2019; Khemakhem et al., 2020a; Lachapelle et al., 2022b; Yao et al., 2022a). However, when the effect of the regime variable on a causal variable C_i can be described by a *binary* variable, *i.e.*, $I_i \in \{0, 1\}$, the two representations become distinguishable. In Figure 3, we visualize the two representations by showing the means of the different variables under interactions, which we detail in Appendix B.6 and provide intuition here. For the original representation C_1, C_2, each variable's mean takes on only two different values for any R^t. For example, for regime variables where $I_1 = 0$, the variable C_1 takes a mean that is in the center of the coordinate system. Similarly, when $I_1 = 1$, the variable C_1 will take a mean that is represented as a pink (for $I_1 = 1, I_2 = 0$) or yellow tick (for $I_1 = 1, I_2 = 1$). In contrast, for the rotated variables, both $\hat{C}_1$ and $\hat{C}_2$ have three different means depending on the interactions, making it impossible to model them with individual binary variables. Intuitively, the only alternative representations to C_1, C_2 which can be described by binary variables are permutations and/or element-wise transformations, effectively identifying the causal variables according to our identifiability class.

3.3 IDENTIFIABILITY RESULT

When extending this intuition to more than two variables, we find that systems may become unidentifiable when the two distributions of each causal variable, *i.e.*, interacted and not interacted, always differ in the same manner. Formally, we denote the log-likelihood difference between the two distributions of a causal variable C_i^t as $\Delta(C_i^t|C^{t-1}) := \log p(C_i^t|C^{t-1}, I_i^t = 1) - \log p(C_i^t|C^{t-1}, I_i^t = 0)$. If this difference or its derivative w.r.t. C_i^t is constant for all values of C_i^t, the effect of the interactions could be potentially modeled in fewer than K dimensions, giving rise to models that do not identify the causal model $\mathcal{M}$.

To prevent this, we consider two possible setups for ensuring sufficient variability of $\Delta(C_i^t|C^{t-1})$: *dynamics variability*, and *time variability*. We present our identifiability result below and provide the proofs in Appendix B.

Theorem 3.3. *An estimated model $\widehat{\mathcal{M}} = \langle \hat{g}, \hat{f}, \hat{\omega}, \hat{\mathcal{C}} \rangle$ identifies the true causal model $\mathcal{M} = \langle g, f, \omega, \mathcal{C} \rangle$ if:*

1. *(**Observations**) $\widehat{\mathcal{M}}$ and $\mathcal{M}$ model the same likelihood:*
$$p_{\widehat{\mathcal{M}}}(X^t|X^{t-1}, R^t) = p_{\mathcal{M}}(X^t|X^{t-1}, R^t);$$
2. *(**Distinct Interaction Patterns**) Each variable C_i in $\mathcal{M}$ has a distinct interaction pattern (Definition 3.2);*

and one of the following two conditions holds for $\mathcal{M}$:

A. *(**Dynamics Variability**) Each variable's log-likelihood difference is twice differentiable and not always zero:*
$$\forall C_i^t, \exists C^{t-1}: \frac{\partial^2 \Delta(C_i^t|C^{t-1})}{\partial (C_i^t)^2} \neq 0;$$

B. *(**Time Variability**) For any $C^t \in \mathcal{C}$, there exist $K + 1$ different values of C^{t-1} denoted with $c^1, ..., c^{K+1} \in \mathcal{C}$, for which the vectors $v_1, ..., v_K \in \mathbb{R}^{K+1}$ with*
$$v_i = \left[\frac{\partial \Delta(C_i^t|C^{t-1}=c^1)}{\partial C_i^t} \quad \cdots \quad \frac{\partial \Delta(C_i^t|C^{t-1}=c^{K+1})}{\partial C_i^t} \right]^T$$
are linearly independent.

Intuitively, Theorem 3.3 states that we can identify a causal model $\mathcal{M}$ by maximum likelihood optimization, if we have distinct interaction patterns (Definition 3.2) and $\Delta(C_i^t|C^{t-1})$ varies sufficiently, either in dynamics or in time. *Dynamics variability* can be achieved by the difference $\Delta(C_i^t|C^{t-1})$ being non-linear for all causal variables. This assumption is common in previous ICA-based works (Hyvärinen et al., 2019; Yao et al., 2022a,b) and, for instance, allows for Gaussian distributions with variable standard deviations. While allowing for a variety of distributions, it excludes additive Gaussian noise models. We can include this challenging setup by considering the *time variability* assumption, which states that the effect of the interaction depends on the previous time step, and must do so differently for each variable. As an example, consider a dynamical system with several moving objects, where an interaction is a collision with a robotic arm. The time variability condition is commonly fulfilled by the fact that the trajectory of each object depends on its own velocity and position.

In comparison to previous work, our identifiability results cover a larger class of causal models by exploiting the binary nature of the interaction variables. We provide a detailed comparison in Appendix B.5. In short, closest to our setup, Khemakhem et al. (2020a) and Yao et al. (2022b) require a stronger form of both our dynamics and time variability assumptions, excluding common models like additive Gaussian noise models. Lachapelle et al. (2022b) requires that no two causal variables share the same parents, limiting the allowed temporal graph structures. Meanwhile, our identifiability results allow for arbitrary temporal causal graphs. Further, the two conditions of Theorem 3.3 complement each other well by covering different underlying distributions for the same general setup. Thus, in the next section, we can develop one joint learning algorithm for identifying the causal variables based on both conditions in Theorem 3.3.

4 BISCUIT

Using the results of Section 3, we propose BISCUIT (**B**inary **I**nteraction**s** for **C**a**u**sal **I**den**t**ifiability), a neural-network based approach to identify causal variables and their interaction variables. In short, BISCUIT is a variational autoencoder (VAE) (Kingma et al., 2014), which aims at modeling each of the causal variables $C_1, ..., C_K$ in a separate latent dimension by enforcing the latent structure of Figure 2. We first give an overview of BISCUIT and then detail the design choices for the model prior.

4.1 OVERVIEW

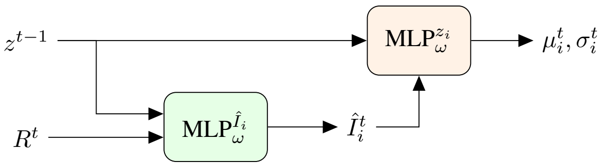


Figure 4: The prior structure of BISCUIT. Based on the previous latents z^{t-1} and the observed regime variable R^t, the $\text{MLP}_\omega^{\hat{I}_i}$ predicts the interaction variable $\hat{I}_i^t$. Then, $\text{MLP}_\omega^{z_i}$ outputs the distribution for the next time step $p(z_i^t|z^{t-1}, \hat{I}_i^t)$, which can be parameterized by, *e.g.*, a mean μ_i^t and std σ_i^t.

BISCUIT consists of three main elements: the encoder q_ϕ, the decoder p_θ, and the prior p_ω. The decoder and encoder implement the observation function g and its inverse g^{-1} (Definition 3.1), respectively, and act as a map between observations x^t and a lower-dimensional latent space $z^t \in \mathbb{R}^M$, in which we learn the causal variables $C_1^t, ..., C_K^t$. The goal of the model is to learn each causal variable C_i^t in a different latent dimension, *e.g.*, z_j^t, effectively separating and hence identifying the causal variables according to Theorem 3.1. Thus, we need the latent space to have at least K dimensions. In practice, since the number of causal variables is not known a priori, we commonly overestimate the latent dimensionality, *i.e.*, $M \gg K$. Still, we expect the model to only use K dimensions actively, with the redundant dimensions not containing any information after training.

On this latent space, the prior p_ω learns a distribution that follows the structure in Definition 3.1, modeling the dynamics in the latent space. As an objective, we maximize the data likelihood of observation triplets $\{X^t, X^{t-1}, R^t\}$ from the true causal model, as stated in Theorem 3.3. The loss function for BISCUIT is:

$$\begin{aligned}\mathcal{L}^t = &-\mathbb{E}_{q_\phi(z^t|x^t)}\left[\log p_\theta(x^t|z^t)\right] + \\ &\mathbb{E}_{q_\phi(z^{t-1}|x^{t-1})}\left[\text{KL}\left(q_\phi(z^t|x^t)||p_\omega(z^t|z^{t-1}, R^t)\right)\right]\end{aligned} \quad (2)$$

with learnable parameter sets ϕ (encoder), θ (decoder), and ω (prior), and KL being the Kullback-Leibler divergence.

For visually complex datasets, the VAE commonly has to perform a trade-off between reconstruction quality and prior modeling, which may cause poorer identification of the causal variables. To circumvent that, we follow Lippe et al. (2022a) by separating the reconstruction and prior modeling stage by training an autoencoder and a normalizing flow (Rezende et al., 2015) in separate stages. In this setup, an autoencoder is first trained to map the observations x^t into a lower-dimensional space. Afterward, we learn a normalizing flow on the autoencoder's representations to transform them into the desired causal representation, using the same prior structure as in the VAE. In experiments, we refer to this approach as BISCUIT-NF, and the previously described VAE-based approach as BISCUIT-VAE.

4.2 MODEL PRIOR

Our prior follows the distribution structure of Definition 3.1, which has two elements per latent variable: a function to model the binary interaction variable, and a conditional distribution. We integrate this into BISCUIT's prior by learning both via multi-layer perceptrons (MLPs):

$$p_\omega(z^t|z^{t-1}, R^t) = \prod_{i=1}^{M} p_{\omega,i}\left(z_i^t|z^{t-1}, \text{MLP}_\omega^{\hat{I}_i}(R^t, z^{t-1})\right). \quad (3)$$

Here, $\text{MLP}_\omega^{\hat{I}_i}$ is an MLP that maps the regime variable R^t and the latents of the previous time step z^{t-1} to a binary output $\hat{I}_i^t$, as shown in Figure 4 . This MLP aims to learn the interaction variable for the latent variable z_i, simply by optimizing Equation (2). The variable $\hat{I}_i^t$ is then used as input for predicting the distribution over z_i^t. For simplicity, we model $p_{\omega,i}$ as a Gaussian distribution, which is parameterized by one MLP per variable, $\text{MLP}_\omega^{z_i}$, predicting the mean and standard deviation. To allow for more complex distributions, $p_{\omega,i}$ can alternatively be modeled by a conditional normalizing flow (Winkler et al., 2019).

In early experiments, we found that enforcing $\hat{I}_i^t$ to be a binary variable and backpropagating through it with the straight-through estimator (Bengio et al., 2013) leads to suboptimal performances. Instead, we model $\hat{I}_i^t$ as a continuous variable during training by using a temperature-scaled $\tanh$ as the output activation function of $\text{MLP}_\omega^{\hat{I}_i}$. By gradually decreasing the temperature, we bring the activation function closer to a discrete step function towards the end of training.

5 RELATED WORK

Causal Discovery from Unknown Targets Learning (equivalence classes of) causal graphs from observational and interventional data, even with unknown intervention targets, is a common setting in causal discovery (Brouillard et al., 2020; Jaber et al., 2020; Mooij et al., 2020; Squires et al., 2020). In recent work, this is even extended to the case in which we have unknown mixtures of interventional data (Faria et al., 2022; Kumar et al., 2021; Mian et al., 2023), which for example can happen if the regime variable is not observed. In our paper, we assume that we observe the regime variable and then reconstruct the latent interaction variables, which resemble the observed context variables by Mooij et al. (2020). Moreover, our work is on a different task, causal representation learning, in which we try to learn the causal variables from high-dimensional data.

Causal Representation Learning A common basis for causal representation learning is Independent Component Analysis (ICA) (Comon, 1994; Hyvärinen et al., 2001), which aims to identify independent latent variables from observations. Due to the non-identifiability for the general case of non-linear observation functions (Hyvärinen et al., 1999), additional auxiliary variables are often considered in this setting (Hyvärinen et al., 2016; Hyvärinen et al., 2019). Ideas from ICA have been integrated into neural networks (Khemakhem et al., 2020a,b; Reizinger et al., 2022) and applied to causality (Gresele et al., 2021; Monti et al., 2019; Shimizu et al., 2006) for identifying causal variables.

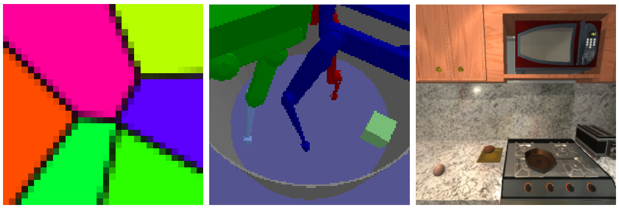
Recently, several works in causal representation learning have exploited distribution shifts or interventions to identify causal variables. Using counterfactual observations, Ahuja et al. (2022); Brehmer et al. (2022); Locatello et al. (2020) learn causal variables from pairs of images, between which only a subset of variables has changed via interventions with unknown targets. For temporal processes, Lachapelle et al. (2022a,b) can model interventions of unknown target via *actions*, which is equivalent to the regime variable in our setting, but require that each causal variable has a strictly unique parent set. On the other hand, Yao et al. (2022a,b) consider observations from m different regimes $u_1, ..., u_m$, where, in our setting, the regime indicator u is a discrete version of R^t. However, they require at least $2K + 1$ different regimes compared to $\lfloor \log_2 K \rfloor + 2$ settings for ours, and have stronger conditions on the distribution changes over regimes (*e.g.*, no additive Gaussian noise models). In temporal settings where the intervention targets are known, CITRIS (Lippe et al., 2022a, 2023) identifies scalar and multidimensional causal variables from high-dimensional images. Nonetheless, observing the intervention targets requires additional supervision, which may not always be available. To the best of our knowledge, we are the first to use unknown binary interactions to identify the causal variables from high-dimensional observations.

6 EXPERIMENTS

To illustrate the effectiveness of BISCUIT, we evaluate it on a synthetic toy benchmark and two environments generated by 3D robotic simulators. We publish our code at `https://github.com/phlippe/BISCUIT`, and detail the data generation and hyperparameters in Appendix C.

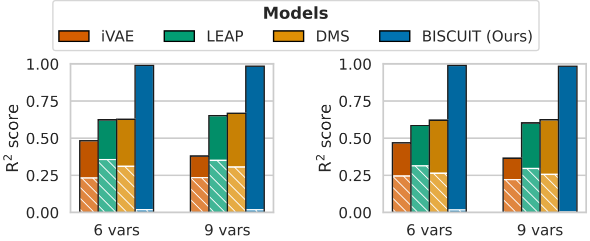
6.1 SYNTHETIC TOY BENCHMARK

To evaluate BISCUIT on various graph structures, we extend the Voronoi benchmark (Lippe et al., 2022a) by replacing observed intervention targets with unobserved binary interactions. In this dataset, each causal variable follows an additive Gaussian noise model, where the mean is modeled by a randomly initialized MLP. To determine the parent set, we randomly sample the causal graph with an edge likelihood of 0.4. Instead of observing the causal variables directly, they are first entangled by applying a two-layer randomly initialized normalizing flow before visualizing the outputs as colors in a Voronoi diagram of size 32×32 (see Figure 5a). We extend the original benchmark by including a robotic arm that moves over the Voronoi diagram and interacts by touching individual color regions/tiles. Each tile corresponds to one causal variable, allowing for both single- and multi-target interactions. The models need to deduce



(a) Voronoi (b) CausalWorld (c) iTHOR

Figure 5: Example figures of our three environments with increasing complexity: Voronoi (Lippe et al., 2023), CausalWorld (Ahmed et al., 2020), and iTHOR (Kolve et al., 2017).



(a) Random Interactions (b) Minimal Interactions

Figure 6: Results on the Voronoi benchmark averaged over 10 seeds. Solid bars show the mean R^2-diag score (higher is better), and striped bars the R^2-sep scores (lower is better, non-visible bars indicate close-to zero values). BISCUIT accurately identifies the causal variables across settings.

these interactions from a regime variable $R^t \in [0,1]^2$ which is the 2D location of the robotic arm on the image. When the robotic arm interacts with a variable, its mean is set to zero, which resembles a stochastic perfect intervention.

Evaluation We generate five Voronoi systems with six causal variables, and five systems with nine variables. We compare BISCUIT to iVAE (Khemakhem et al., 2020a), LEAP (Yao et al., 2022b), and Disentanglement via Mechanism Sparsity (DMS) (Lachapelle et al., 2022b), since all use a regime variable. We do not compare with CITRIS (Lippe et al., 2022a, 2023), because it requires known intervention targets. We follow Lippe et al. (2023) in evaluating the models on a held-out test set where all causal variables are independently sampled. We calculate the coefficient of determination (Wright, 1921), also called the R^2 score, between each causal variable C_i and each learned latent variable z_j, denoted by R^2_{ij}. If a model identifies the causal variables according to Definition 3.1, then for each causal variable C_i, there exists one latent variable z_j for which $R^2_{ij} = 1$, while it is zero for all others. Since the alignment of the learned latent variables to causal variables is not known, we report R^2 scores for the permutation π that maximizes the diagonal of the R^2 matrix, *i.e.*, R^2-diag $= {}^1\!/_K \sum_{i=1}^{K} R^2_{i,\pi(i)}$ (where 1 is optimal). To account for spurious modeled correlation, we also report the maximum correlation besides this alignment: R^2-sep $= {}^1\!/_K \sum_{i=1}^{K} \max_{j\neq\pi(i)} R^2_{ij}$ (optimal 0).

Results The results in Figure 6a show that BISCUIT identifies the causal variables with high accuracy for both graphs with six and nine variables. In comparison, all baselines struggle to identify the causal variables, often falling back to modeling the colors as latent variables instead. While the assumptions of iVAE and LEAP do not hold for additive Gaussian noise models, the assumptions of DMS, including the graph sparsity, mostly hold. Still, BISCUIT is the only method to consistently identify the true variables, illustrating that its stable optimization and robustness.

Minimal Number of Regimes To verify that BISCUIT only requires $\lfloor \log_2 K \rfloor + 2$ different regimes (Theorem 3.3), we repeat the previous experiments with reducing the interaction maps to a minimum. This results in four sets of interactions for six variables, and five for nine variables. Figure 6b shows that BISCUIT still correctly identifies causal variables in this setting, supporting our theoretical results.

Learned Intervention Targets After training, we can use the interaction variables $\hat{I}_1, ..., \hat{I}_M$ learned by BISCUIT to identify the regions in which the robotic arm interacts with a causal variable. Based on our theoretical results, we expect that some of the learned variables are identical to the true interaction variables $I_1, ... I_K$ up to permutations and sign-flips. In all settings, we find that the learned binary variables match the true interaction variables with an average F1 score of 98% for the same permutation of variables as in the R^2 evaluation. This shows that BISCUIT identified the true interaction variables. Thus, in practice, one could use a few samples with labeled interaction variables to identify the learned permutation of the model.

6.2 CAUSALWORLD

CausalWorld (Ahmed et al., 2020) is a robotic manipulation environment with a tri-finger robot, which can interact with objects in an enclosed space by touch (see Figure 5b). The environment also allows for interventions on various environment parameters, including the colors or friction parameters of individual elements. We experiment on this environment by recording the robot's interactions with a cube. Besides the cube position, rotation and velocity, the causal variables are the colors of the three fingertips, as well as the floor, stage and cube friction, which we visualize by the colors of the respective objects. All colors and friction parameters follow an additive Gaussian noise model. When a robot finger touches the cube, we perform a stochastic perfect intervention on its color. Similarly, an interaction with the friction parameters correspond to touching these objects with all three fingers. The regime variable R^t is modeled by the angles of the three motors per robot finger from the current and previous time step, providing velocity information.

This environment provides two new challenges. Firstly, not

Table 1: R^2 scores (diag ↑ / sep ↓) for the identification of the causal variables on CausalWorld and iTHOR.

Models	CausalWorld	iTHOR
iVAE (Khemakhem et al., 2020a)	0.28 / 0.00	0.48 / 0.35
LEAP (Yao et al., 2022b)	0.30 / 0.00	0.63 / 0.45
DMS (Lachapelle et al., 2022b)	0.32 / 0.00	0.61 / 0.40
BISCUIT-NF (Ours)	**0.97** / 0.01	**0.96 / 0.15**

all interactions are necessarily binary. In particular, the collisions between the robot and the cube have different effects depending on the velocity and direction of the fingers of the robot, which are not part of the state of the causal variables at the previous time step. Additionally, the robotic system is present in the observation/image, while our theoretical results assume that R^t is not a direct cause of X^t. We adapt BISCUIT-NF and the baselines to this case by adding R^t as additional information to the decoder, effectively removing the need to model R^t in the latent space.

On this task, BISCUIT identifies the causal variables well, as seen in Table 1. Because the cube position, velocity and rotation share the same interactions, in the evaluation we consider them as a multidimensional variable. Although the true model cannot be fully described by binary interaction variables, BISCUIT still models the binary information of whether a collision happens or not for the cube, since it is the most important part of the dynamics. We verify this in Appendix C.2.3 by measuring the F1 score between the predicted interaction variables and ground truth interactions/collisions. BISCUIT achieves an F1 score of 50% for all cube-arm interactions, which indicates a high similarity between the learned interaction and the ground truth collisions considering that collisions only happen in approximately 5% of the frames. The mismatches are mostly due to the learned interactions being more conservative, *i.e.*, being 1 already a frame too early sometimes. Meanwhile, none of the baselines are able to reconstruct the image sufficiently, missing the robotic arms and the cube as shown in Appendix C.2.3. While this might improve with significant tuning effort, BISCUIT-NF is not sensitive to the difficulty of the reconstruction due to its separate autoencoder training stage.

6.3 ITHOR - EMBODIED AI

To illustrate the potential of causal representation learning in embodied AI, we apply BISCUIT to the iTHOR environment (Kolve et al., 2017). In this environment, an embodied AI agent can perform actions on various objects in an 3D indoor scene such as a kitchen. These agent-object interactions can often be described by a binary variable, *e.g.*, pickup/put down an object, open/close a door, turn on/off an object, etc., which makes it an ideal setup for BISCUIT.

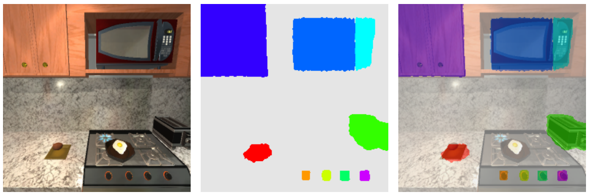


Figure 7: Visualizing the learned interaction variables of BISCUIT for an example input image (left). We show the locations, *i.e.*, values of $R^t \in [0,1]^2$, for which each interaction variable is greater than zero/active as different colors. For readability, only nine interaction variables are shown. The right image is an overlay of both. BISCUIT accurately learns the interactions and adapts them to the input image.

Our goal in this environment is to identify the causal variables, *i.e.*, the objects and their states, from sequences of interactions. We perform this task on the kitchen environment shown in Figure 5c. This environment contains two movable objects, *i.e.*, a plate and an egg, and seven static objects, *e.g.*, a microwave and a stove. Overall, we have 18 causal variables, which include both continuous, *e.g.*, the location of the plate, and binary variables, *e.g.*, whether the microwave is on or off. Causal variables influence each other by state changes, *e.g.*, the egg gets cooked when it is in the pan and the stove is turned on. Further, the set of possible actions that can be performed depends on the previous time step, *e.g.*, only one object can be picked up at a time. For training, we generate a dataset where we randomly pick a valid action at each time step. We model the regime variable R^t as a two-dimensional pixel coordinate, which is the position of a pixel showing the interacted object in the image ($R^t \in [0,1]^2$). This simulates iTHOR's web demo (Kolve et al., 2017), where a user interacts with objects by clicking on them.

We train BISCUIT-NF and our baselines on this dataset, and compare the latent representation to the ground truth causal variables in terms of the R^2 score in Table 1. Although the baselines reconstruct the image mostly well, the causal variables are highly entangled in their representations. In contrast, BISCUIT identifies and separates most of the causal variables optimally, except for the two movable objects (egg/plate). This is likely due to the high inherent correlation of the two objects, since their positions cannot overlap and only one of them can be picked up at a time.

Besides evaluating the causal representation, we also visualize the learned interaction variables of BISCUIT in Figure 7. Here, each color represents the region in which BISCUIT identified an interaction with a different causal variable. Figure 7 shows that BISCUIT has identified the correct interaction region for each object. Moreover, it allows for context-dependent interactions, as the location of the plate influences the region of its corresponding interaction variable.

Finally, we can use the learned causal representation to per-

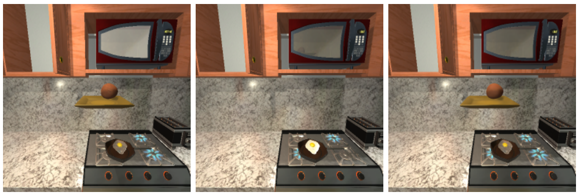


Figure 8: Performing interventions in the latent space. First, the two inputs images are encoded into latent space. Then, we replace the latents of the front-left stove and microwave in the first image by the corresponding latents of the second image. Decoding these new latents creates an unseen scenario where the egg is uncooked, but the stove is turned on. This shows the modularity of BISCUIT's representations.

form interventions and create novel combinations of causal variables. For this, we encode two images into the learned latent space of BISCUIT, and combine the latent representations of the causal variables to have a novel image decoded. For example, in Figure 8, we replace the latents representing the front-left stove and the microwave state in the first image by the corresponding latents of the second image. BISCUIT not only integrates these changes without influencing any of the other causal variables, but generates a completely novel state: even though in the iTHOR environment, the egg is instantaneously cooked when the stove turns on, BISCUIT correctly combines the state of the egg being raw with the stove burning. This shows the capabilities of BISCUIT to model unseen causal interventions.

7 CONCLUSION

We prove that under mild assumptions, causal variables become identifiable from high-dimensional observations, when their interactions with an external system can be described by unknown binary variables. As a practical algorithm, we propose BISCUIT, which learns the causal variables and their interaction variables. In experiments across three robotic-inspired datasets, BISCUIT outperforms previous methods in identifying the causal variables from images.

While in experiments, BISCUIT shows strong identification even for complex interactions, the presented theory is currently limited to binary interaction variables. Although the first step may be to generalize the theory to interaction variables with more than two states, extensions to unknown domains or sparse, continuous interaction variables are other interesting future directions. Instead of assuming distinct interaction patterns, future work can extend these results to partial identifiability, similar to Lachapelle et al. (2022a); Lippe et al. (2022a). Finally, our results open up the opportunity for empirical studies showing the benefits of causal representations for complex real-world tasks like embodied AI.

Author Contributions

P. Lippe conceived the idea, developed the theoretical results, implemented the models and experiments, and wrote the paper. S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, E. Gavves advised during the project and helped in writing the paper.

Acknowledgements

We thank SURF for the support in using the National Supercomputer Snellius. This work is financially supported by Qualcomm Technologies Inc., the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

REFERENCES

Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, Y., Schölkopf, B., Wüthrich, M., and Bauer, S. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Ahuja, K., Hartford, J., and Bengio, Y. Weakly Supervised Representation Learning with Sparse Perturbations. In *Advances in Neural Information Processing Systems 35, NeurIPS*, 2022.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Brehmer, J., de Haan, P., Lippe, P., and Cohen, T. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems 35, NeurIPS*, 2022.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable Causal Discovery from Interventional Data. In *Advances in Neural Information Processing Systems 33, NeurIPS*, 2020.

Comon, P. Independent component analysis, A new concept? *Signal Processing*, 36(3), April 1994.

Dean, T. and Kanazawa, K. A model for reasoning about persistence and causation. *Computational Intelligence*, 5 (2), 1989.

Faria, G. R. A., Martins, A., and Figueiredo, M. A. T. Differentiable Causal Discovery Under Latent Interventions. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*. PMLR, 11–13 Apr 2022.

Gresele, L., von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021.

Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering Atari with Discrete World Models. In *International Conference on Learning Representations*, 2021.

Hyvärinen, A. and Morioka, H. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2016.

Hyvärinen, A. and Pajunen, P. Nonlinear Independent Component Analysis: Existence and Uniqueness Results. *Neural Netw.*, 12(3), apr 1999. ISSN 0893-6080.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis*. John Wiley & Sons, June 2001.

Hyvärinen, A., Sasaki, H., and Turner, R. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*. PMLR, 2019.

Jaber, A., Kocaoglu, M., Shanmugam, K., and Bareinboim, E. Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning. In *Advances in Neural Information Processing Systems 33, NeurIPS*, 2020.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*. PMLR, 2020a.

Khemakhem, I., Monti, R., Kingma, D., and Hyvarinen, A. ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA. In *Advances in Neural Information Processing Systems 33, NeurIPS*, 2020b.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding. In *International Conference on Learning Representations (ICLR)*, 2021.

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., and Farhadi, A. AI2-THOR: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. Web demo `https://ai2thor.allenai.org/demo/`.

Kumar, A. and Sinha, G. Disentangling mixtures of unknown causal interventions. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*. PMLR, 27–30 Jul 2021.

Lachapelle, S. and Lacoste-Julien, S. Partial Disentanglement via Mechanism Sparsity. In *UAI 2022 Workshop on Causal Representation Learning*, 2022a.

Lachapelle, S., Rodriguez, P., Le, R., Sharma, Y., Everett, K. E., Lacoste, A., and Lacoste-Julien, S. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2022b.

Lesort, T., Díaz-Rodríguez, N., Goudou, J.-F., and Filliat, D. State representation learning for control: An overview. *Neural Networks*, 108, 2018.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In *Proceedings of the 39th International Conference on Machine Learning, ICML*, 2022a.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. Intervention Design for Causal Representation Learning. In *UAI 2022 Workshop on Causal Representation Learning*, 2022b.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-Supervised Disentanglement Without Compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.

Mian, O., Kamp, M., and Vreeken, J. Information-theoretic causal discovery and intervention detection over multiple environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI-23, 2023.

Monti, R. P., Zhang, K., and Hyvärinen, A. Causal Discovery with General Non-Linear Relationships using Non-Linear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI*, 2019.

Mooij, J. M., Magliacane, S., and Claassen, T. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 21(99), 2020.

Murphy, K. Dynamic Bayesian Networks: Representation, Inference and Learning. *UC Berkeley, Computer Science Division*, 2002.

Reizinger, P., Gresele, L., Brady, J., von Kügelgen, J., Zietlow, D., Schölkopf, B., Martius, G., Brendel, W., and Besserve, M. Embrace the Gap: VAEs Perform Independent Mechanism Analysis. In *Advances in Neural Information Processing Systems 35, NeurIPS*, 2022.

Rezende, D. J. and Mohamed, S. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, 2015.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 2021.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.*, 7, dec 2006.

Squires, C., Wang, Y., and Uhler, C. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020.

Träuble, F., Dittadi, A., Wuthrich, M., Widmaier, F., Gehler, P. V., Winther, O., Locatello, F., Bachem, O., Schölkopf, B., and Bauer, S. The Role of Pretrained Representations for the OOD Generalization of RL Agents. In *International Conference on Learning Representations*, 2022.

Winkler, C., Worrall, D., Hoogeboom, E., and Welling, M. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

Wright, S. Correlation and causation. *Journal of agricultural research*, 20(7), 1921.

Yao, W., Chen, G., and Zhang, K. Temporally Disentangled Representation Learning. In *Advances in Neural Information Processing Systems 35, NeurIPS*, 2022a.

Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning Temporally Causal Latent Processes from General Temporal Data. In *International Conference on Learning Representations*, 2022b.