Intervention Design for Causal Representation Learning

Abstract

In this paper, we take a first step towards bringing two fields of causality closer together: intervention design and causal representation learning. Intervention design is a well studied task in classic causal discovery, which aims at finding the minimal sets of experiments under which the causal graph can be identified. Causal representation learning aims at recovering causal variables from high-dimensional entangled observations. In recent work in causal representation, interventions are exploited to improve identifiability, similarly to classic causal discovery. Hence, the same task becomes relevant in this setting as well: how many experiments are minimally needed to identify the latent causal variables? Based on the recent causal representation learning method CITRIS, we show that for K causal variables, $\lfloor \log_2(K) \rfloor + 2$ experiments are sufficient to identify causal variables from temporal, intervened sequences, which is only one more experiment than needed for classic causal discovery in the worst case. Further, we show that this bound holds empirically in experiments on a 3D rendered video dataset.

1 INTRODUCTION

Recently, there has been a growing interest in the field of causal representation learning (Brehmer et al., 2022; von Kügelgen et al., 2021; Lippe et al., 2022; Locatello et al., 2020; Schölkopf et al., 2021), which aims at discovering latent, causal factors and their causal relations from high-dimensional observations such as images or videos. A crucial aspect towards reaching this goal is commonly considered to be the access to interventional data. An intervention influences the causal mechanism of one or more variables while leaving other mechanisms and the observation function, *i.e.* the way how we perceive these variables, unchanged. Comparing data under different interventions allows one to find and disentangle the sources of variations, *i.e.* the causal variables, from the high-dimensional observations. In the context of causal representation learning, several works have considered different interventional settings, such as interventions with unknown targets (Brehmer et al., 2022; Locatello et al., 2020; Yao et al., 2022) or observed targets (von Kügelgen et al., 2021; Lachapelle et al., 2022; Lippe et al., 2022), in order to guarantee identifiability of the causal variables. However, interventional data can often be expensive, since it requires a specific experiment in which there is a perturbation of the causal system, e.g. a randomized controlled trial. Hence, a naturally arising question is what is the minimal number of different intervention experiments that suffices for identifying the causal variables.

In this paper, we answer this question by drawing connections from causal representation learning to the area of intervention design (Eberhardt, 2007; Greenewald et al., 2019; He and Geng, 2008; Hyttinen et al., 2012a, 2013; Kocaoglu et al., 2017a,b; Shanmugam et al., 2015; Squires et al., 2020), which aims to find the minimal set of experiments that identifies the causal graph for known causal variables. As a specific setting, we focus on the recent causal representation learning method CITRIS (Lippe et al., 2022) which leverages data from temporal intervened sequences with known intervention targets. Using similar techniques as for intervention design, we show that for CITRIS $\lfloor \log_2(K) \rfloor + 2$ experiments are sufficient to identify the K causal variables, which is exactly one experiment more than needed for causal discovery in the worst case scenario. This opens up several opportunities for future work to leverage intervention design methods from causal discovery also in causal representation learning.

Furthermore, to show that this bound holds also empirically, we conduct experiments on the Temporal Causal3DIdent dataset (von Kügelgen et al., 2021; Lippe et al., 2022). This dataset contains videos of 3D object renderings with 6 causal variables, including the object positions, rotations, colors,

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

and lightning. As expected based on our theoretical results, with a minimal set of $\lfloor \log_2(6) \rfloor + 2 = 4$ experiments, CITRIS is able to disentangle the variables well, obtaining an only slightly higher disentanglement error compared to the same model with access to an unlimited number of intervention experiments.

2 RELATED WORK

2.1 INTERVENTION DESIGN

Intervention design, also known as experiment selection, is a well-studied area in causal discovery (Addanki et al., 2020; Eberhardt, 2007; Eberhardt et al., 2005; Greenewald et al., 2019; Hauser and Bühlmann, 2014; He and Geng, 2008; Hyttinen et al., 2012a, 2013; Kocaoglu et al., 2017a,b; Shanmugam et al., 2015; Squires et al., 2020). Consider an unknown causal graph $G = (V, E)$, where each node $i \in V$ is associated with a causal variable C_i, and each edge $(i, j) \in E$ represents a causal relation from C_i to C_j: $C_i \rightarrow C_j$. Given a set of K causal variables $C_1, ..., C_K$, the goal is to determine the set of experiments $\{E_1, ..., E_N\}$ which identifies the underlying causal structure G most efficiently. In our notation, an experiment E_i is defined as a set of causal variables on which interventions are jointly performed. For example, the observational regime is denoted by $E_i = \emptyset$, *i.e.* all variables are passively observed, and single-target interventions by $E_1 = \{C_1\}, E_2 = \{C_2\}$, and so on.

One can show that under causal sufficiency (*i.e.*, no latent confounders or selection bias), acyclicity (*i.e.*, no cycles in G), and the causal Markov and faithfulness assumptions (*i.e.*, there are no additional independences w.r.t. the ones encoded in the graph G), a set of experiments uniquely identifies the causal graph if for every pair of variables (C_i, C_j), (1) there exists an experiment where C_i has been intervened on, but not C_j, or vice versa, and (2) there exists an experiment under which both C_i and C_j are passively observed, *i.e.* not intervened. The first condition is also referred to as the *unordered pair condition*, and the latter as the *covariance condition* (Eberhardt, 2007; Hyttinen et al., 2013). Using these conditions, Eberhardt (2007) showed that $\lfloor \log_2(K) \rfloor + 1$ experiments are sufficient to guarantee the identifiability of a causal graph with K variables. This bound is for the worst case scenario, and for certain graphs, fewer experiments might be sufficient. For instance, if all variables are independent, the observational regime alone identifies the whole graph.

While in this work we provide a first application of intervention design for causal representation learning, in the intervention design literature, several other settings can be considered, including (1) where the number of variables that can be subject to an intervention simultaneously is limited (Ghassami et al., 2018; Hyttinen et al., 2013; Kocaoglu et al., 2017a; Shanmugam et al., 2015), (2) where latent confounders may exist (Addanki et al., 2020; Hyttinen et al., 2012b; Kocaoglu et al., 2017b), or (3) where additional background knowledge about the underlying causal structure is available (Eberhardt, 2008; Greenewald et al., 2019; Hauser and Bühlmann, 2012; Sen et al., 2017).

2.2 CAUSAL REPRESENTATION LEARNING

Causal representation learning aims at learning representations of causal factors in an underlying system from high-dimensional observations like images (Brehmer et al., 2022; von Kügelgen et al., 2021; Lachapelle et al., 2022; Lippe et al., 2022; Locatello et al., 2020; Schölkopf et al., 2021). One of the first lines of work is Independent Component Analysis (ICA) (Comon, 1994; Hyvärinen et al., 2001). ICA tries to recover independent latent variables transformed by some invertible transformation. ICA was extended to non-linear transformations by exploiting auxiliary variables under which the latents become conditionally mutually independent (Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019), and combined with deep learning architectures like VAEs (Khemakhem et al., 2020a,b; Sorrenson et al., 2020; Zimmermann et al., 2021). Further, recent works draw a connection between causality and ICA (Gresele et al., 2021; Monti et al., 2019). In particular, Lachapelle et al. (2022); Yao et al. (2022) discuss identifiability from temporal sequences. While both can model interventions in their framework, they do not explicitly exploit the knowledge of the intervention targets and require additional assumptions in terms of sufficient variation.

Focusing on causal structures in the data, von Kügelgen et al. (2021) demonstrate that contrastive learning methods can block-identify causal variables by considering augmentations as interventions on the style of the image, while keeping the content unchanged. Locatello et al. (2020) identify independent latent causal factors from pairs of observations that only differ in a subset of causal factors. Brehmer et al. (2022) have recently extended this setup to variables with instantaneous causal effects, but require pairs of observations that share the noise term for all variables, except for a single intervened variable, *i.e.* counterfactual samples. Finally, CITRIS (Lippe et al., 2022) uses temporal sequences with interventions to identify the minimal causal variables, *i.e.* the part of a potentially multidimensional causal variable that is influenced by the provided interventions. Thereby, CITRIS considers the causal variables to have temporal dependencies, but being independent within a time step conditioned on the previous time step. In this paper, we focus on the relation of CITRIS (Lippe et al., 2022) to intervention design, since its setting is the closest to common intervention design methods by explicitly considering intervention targets for n variables. Still, we note that similar derivations may be possible for other causal representation learning methods, like Lachapelle et al. (2022), under additional assumptions.

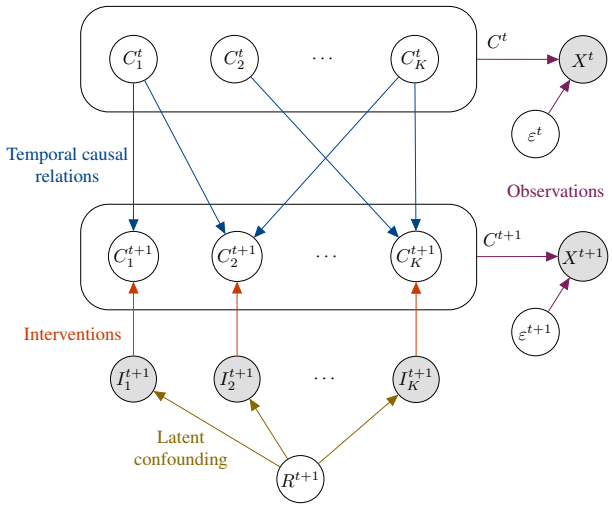


Figure 1: An example causal graph in CITRIS. A latent causal factor C_i^{t+1} has as parents a subset of the causal factors at the previous time step $C_1^t, \dots, C_K^t$, and its intervention target I_i^{t+1}. All causal variables C^{t+1} and the noise ε^{t+1} cause the observation X^{t+1}. R^{t+1} is a latent confounder between the intervention targets.

3 INTERVENTION DESIGN FOR CAUSAL REPRESENTATION LEARNING

We first review the setting and assumptions on which CITRIS (Lippe et al., 2022) aims to identify the causal variables from high-dimensional observations. For simplicity, we limit ourselves to the aspects relevant for designing the intervention experiments, and refer to Lippe et al. (2022) for the full details. We then discuss the identifiability conditions of CITRIS with respect to the interventions, and bring them into context of common intervention design settings. Finally, we derive the minimal number of experiments that are sufficient to fulfill these conditions, and show a simple heuristic with which we can create such minimal sets.

3.1 TEMPORAL INTERVENED SEQUENCES

CITRIS considers a latent, temporal dynamical system with K causal variables, $(C_1^t, C_2^t, ..., C_K^t)_{t=1}^T$, as visualized in Figure 1. The causal graph may contain temporal relations, but no instantaneous effects. This means that we can model each causal factor by $C_i^t = f_i(\text{pa}_G(C_i^t), \epsilon_i)$ for $t = 1, \dots, T$ and $i = 1, \dots, K$, where $\text{pa}(C_i^t) \subseteq \{C_1^{t-1}, \dots, C_K^{t-1}\}$. The graph structure is thereby time invariant, *i.e.*, $\text{pa}_G(C_i^t) = \text{pa}_G(C_i^1)$ for any $t = 1, \dots, T$, and all ϵ_i are mutually independent noises.

Instead of observing the causal variables directly, we measure high-dimensional observations X^t at each time step t. These observations represent an entangled view of all causal factors, C^t, and possible noise variables ε^t.

To represent interventions, we augment the graph with binary intervention target variables $I^t \in \{0, 1\}^K$, where $I_i^t = 1$ if C_i^t has been intervened upon, and $I_i^t = 0$ otherwise. In the causal graph, each intervention variable I_i^t is a parent of its respective causal variable C_i^t. Thereby, CITRIS supports soft interventions (Eberhardt, 2007), which change the conditional distribution, *i.e.*, $p(C_i^t|\text{pa}(C_i^t), I_i^t = 1) \neq p(C_i^t|\text{pa}(C_i^t), I_i^t = 0)$, and includes *perfect* interventions $\text{do}(C_i = c_i)$ (Pearl, 2009) as a special case. Further, CITRIS considers a potential latent confounder R^t between the intervention targets. In alignment with the intervention design setting, R^t could represent the index of the experiment E_i which has been performed at time step t.

Moreover, the causal variables are considered to be potentially multidimensional, *e.g.* representing a 3d position as $C_i \in \mathbb{R}^3$. As shown by Lippe et al. (2022), multidimensional causal variables are not always identifiable from interventions, since interventions may only affect a subset of the variables dimensions. Thus, we are limited to identifying the intervention-dependent parts of variables, *i.e.* the *minimal causal variables*.

3.2 CONDITIONS FOR IDENTIFYING MINIMAL CAUSAL VARIABLES

Under the previously described setting, Lippe et al. (2022) show that the minimal causal variable of a causal factor C_i is identifiable if for any other variable $C_j, i \neq j$, the following condition holds:

$$C_i^{t+1} \not\perp\!\!\!\perp I_i^{t+1} | C^t, I_j^{t+1} \tag{1}$$

Intuitively, we cannot disentangle the causal variables based on interventions if the intervention target variable I_i^{t+1} is constant or can be replaced by any other target variable I_j^{t+1}. To make it more explicit, Equation (1) can be rewritten as the following four conditions:

1. there exists a time step τ such that C_i is intervened, *i.e.*, $I_i^\tau = 1$;
2. there exists another time step τ' such that C_i is not intervened, *i.e.*, $I_i^{\tau'} = 0$;
3. for all other causal factors C_j, there exists a time step $\tilde{\tau}$ when $I_i^{\tilde{\tau}} \neq I_j^{\tilde{\tau}}$;
4. for all other causal factors C_j, there exists a time step $\hat{\tau}$ when $I_i^{\hat{\tau}} = I_j^{\hat{\tau}}$.

The first two conditions ensure that I_i is not constant. The last two conditions guarantee that if the intervention target I_i is not constant, then there cannot exist a deterministic function f for which $I_i^\tau = f(I_j^\tau)$ for all time steps τ.

Instead of considering different time steps, we can also reformulate the conditions in terms of different experiments. Any distribution over intervention targets, $p(I^{t+1})$, can be represented by a distribution over all possible combinations of intervention targets, *i.e.* the experiments $E_1, ..., E_N$. In

this case, the regime variable R^{t+1} can be seen as an experiment indicator, where $R^{t+1} = l$ denotes that the intervention targets at time step $t+1$ follow experiment E_l. Thus, the distribution R^{t+1} follows the distribution over the experiments. For a limited sample size, this distribution may correspond to the number of samples we have per experiment. On the other hand, the conditional distribution of the intervention targets can then be written as a deterministic function of the regime variable:

$$p(I_i^{t+1} = 1 | R^{t+1} = l) = \mathbb{1}\left[C_i \in E_l\right] \tag{2}$$

In other words, $I_i^{t+1} = 1$ if C_i has been intervened on in the experiment E_l (with $R^{t+1} = l$), and otherwise 0. Now, considering a set of the experiments $\{E_1, ..., E_N\}$, we can rewrite the conditions above as follows:

1. there exists an experiment E_l such that $C_i \in E_l$;
2. there exists an experiment E_o such that $C_i \notin E_o$;
3. for all other causal factors $C_j, i \neq j$, there exists an experiment E_{p_j} such that $C_i \in E_{p_j}$ and $C_j \notin E_{p_j}$, or $C_i \notin E_{p_j}$ and $C_j \in E_{p_j}$;
4. for all other causal factors $C_j, i \neq j$, there exists an experiment E_{q_j} such that $C_i, C_j \in E_{q_j}$, or $C_i, C_j \notin E_{q_j}$;

Note the experiments fulfilling these conditions do not need to be mutually exclusive, *i.e.* an experiment can fulfill both condition 1 and 2. With this, our goal becomes finding the minimal number of experiments N' for a given number of causal variables K that can fulfill these four conditions.

3.3 DERIVING THE MINIMAL NUMBER OF EXPERIMENTS

Based on the four conditions, we derive the minimal number of experiments in two steps. Firstly, following condition 3, we need to ensure that every variable C_i has a unique pattern of being intervened or passively observed in different experiments. In other words, the intervention target variables I_i, I_j of two causal variables C_i, C_j cannot be identical for all experiments. For simpler exposition, we express the intervention pattern of a causal variable C_i across N experiments as a binary code $b^i \in \{0,1\}^N$, where $b_l^i = \mathbb{1}\left[C_i \in E_l\right]$. As an example, consider two variables C_1, C_2, for which we have the two experiments $E_1 = \{C_1\}, E_2 = \{C_1, C_2\}$. We represent this pattern with the binary code $b^1 = 11$ for C_1, since C_1 is intervened in both experiments, and $b^2 = 01$ for C_2, since C_2 is only passively observed in E_1. Essentially, the binary code b^i is a concatenation of the binary intervention targets of a variable C_i across experiments $E_1, ..., E_N$.

With this representation, the condition that the intervention target variables of two variables, I_i, I_j, cannot be identical for all experiments, translates to their binary code b^i, b^j being unique, *i.e.* different in at least one digit. To ensure that the binary codes for all variables $b^1, ..., b^K$ are unique, we would need at least $\lceil \log_2(K) \rceil$ experiments, since a binary code of length $L = \lceil \log_2(K) \rceil$ can represent $2^L \geq K$ different values. However, we need to exclude two codes: (1) the code of all zeros $\{0\}^L$, which violates condition 1, and (2) the code of all ones $\{1\}^L$, *i.e.*, interventions at all time, violating condition 2. Removing these two from the available binary codes, we obtain a minimum code length of $\lceil \log_2(K+2) \rceil$ in order to have K unique, valid codes. Consider, for example, two variables C_1, C_2, for which we have four possible binary codes: $\{00, 01, 10, 11\}$. Removing the code for a variable being always passively observed, 00, and the code for a variable always being intervened, 11, we are left with the $\lceil \log_2(2+2) \rceil = 2$ codes $\{01, 10\}$.

In a second step, we need to extend this bound to fulfill condition 4, *i.e.* for all pairs C_i, C_j, there exists an experiment for which $I_i = I_j$ holds. In other words, we need to prevent that $b^i = \neg b^j$ for any i, j, with $\neg b$ representing the one's complement of a binary code b. In the example above, this implies that we cannot use the codes $b^1 = 01$ and $b^2 = 10$ for C_1, C_2, since $b^1 = \neg b^2$. More generally, any subset of more than 2^{L-1} unique codes of length L must contain a pair b^i, b^j for which $b^i = \neg b^j$. Hence, we effectively need to double the previous number of codes to ensure that there exist a subset of K unique codes that are not complementarities of each other. Note, however, that this does not affect the two constant codes, all zeros and all ones, since they are invalid in any code space and are the inverse of each other. Hence, the minimal code length to find K binary codes that are (1) unique, (2) non-constant, and (3) not complementarities of each others, is $\lceil \log_2(2K+2) \rceil = \lceil \log_2(K+1) \rceil + 1 = \lfloor \log_2(K) \rfloor + 2$.[1] This bound simultaneously corresponds to the minimal number of experiments needed to fulfill the conditions 1-4.

One simple way of preventing codes from being complementarities of each other is to add the purely observational regime, *i.e.* $E_0 = \emptyset$, to an experimental set that already fulfills conditions 1-3. This ensures that there exist an experiment E_l for which $C_i, C_j \notin E_l$ and thus $b_l^i = b_l^j$ for all pairs C_i, C_j.

A simple algorithm for creating such sets of experiments is shown in Algorithm 1. Taking again the example of two variables C_1, C_2, we first create all binary codes of length $\lfloor \log_2(2) \rfloor + 1 = 2$: $\mathcal{B} = \{00, 01, 10, 11\}$ (*i.e.* one experiment less than the bound, since we add the observational experiment later). We then remove the code for passively observing a specific causal variable in all experiments: $\mathcal{B} \setminus \{00\} = \{01, 10, 11\}$. Next, we add the experiment in which *all* causal variables are jointly, passively observed: $\mathcal{B} = \{001, 010, 011\}$. From these three codes, we pick two codes for the two causal variables C_1, C_2. Note that any combination of two codes from $\mathcal{B}$ is valid, and we could

[1] These equalities are possible, since we have $K \in \mathbb{N}$, *i.e.* the number of causal variables is a positive integer greater than zero.

Table 1: Results on the Temporal Causal3DIdent dataset with different experimental settings over three seeds. *Full experiments* denotes the setting with full support over all possible intervention experiments (*i.e.*, $I_i^{t+1} \sim \text{Bernoulli}(0.1)$, results taken from Lippe et al. (2022)), and *minimal experiments* follow minimal sets of experiments (ours). R^2 diag and Spearman diag measure the correlation between a causal variable and the latent variables assigned to it by CITRIS. R^2 sep and Spearman sep denote the maximum correlation to any other causal variable. The triplet distance measures the disentanglement by testing the generation of new combinations of causal factors (see Lippe et al. (2022) for detailed descriptions).

Experimental setting	**Triplets** $\downarrow$	R^2 **diag** $\uparrow$	R^2 **sep** $\downarrow$	**Spearman diag** $\uparrow$	**Spearman sep** $\downarrow$
iVAE - Full experiments	0.15 ($\pm$0.01)	0.78 ($\pm$0.04)	0.21 ($\pm$0.10)	0.77 ($\pm$0.05)	0.17 ($\pm$0.04)
CITRIS - Full experiments	0.04 ($\pm$0.00)	0.98 ($\pm$0.00)	0.01 ($\pm$0.00)	0.97 ($\pm$0.00)	0.05 ($\pm$0.01)
CITRIS - Minimal experiments	0.12 ($\pm$0.02)	0.94 ($\pm$0.05)	0.08 ($\pm$0.05)	0.92 ($\pm$0.08)	0.10 ($\pm$0.05)

Algorithm 1 Pseudocode for finding a minimal set of experiments that enable the identification of the minimal causal variables with an observational regime.

Require: Number of variables K

1: Create all possible binary codes of length $L = \lfloor \log_2(K) \rfloor + 1$ as set $\mathcal{B} = \{0,1\}^L$
2: Remove the code of observing a variable passively in all experiments, $\{0\}^L$, from $\mathcal{B}$
3: Extend all codes in $\mathcal{B}$ by appending $\{0\}$, *i.e.* an experiment where all variables are passively observed
4: From the remaining codes in $\mathcal{B}$, (arbitrarily) pick K unique codes $b^1, ..., b^K$, one for each causal variable C_i
5: Create experiments by using the codes as binary intervention targets: $E_l = \{C_i | i \in [\![1..K]\!], b_l^i = 1\}$

Table 2: An example of the selected experiments under the minimal number of experiments setting, generated according to Algorithm 1. E_1 is the observational regime, and E_2, E_3, E_4 create unique intervention patterns for each causal variable.

	E_1	E_2	E_3	E_4
pos_o	-	✓	-	✓
rot_o	-	-	✓	-
rot_s	-	-	✓	✓
hue_o	-	✓	✓	-
hue_b	-	-	-	✓
hue_s	-	✓	-	-

pick them based on some heuristic, for instance minimizing the number of interventions: $b^1 = 001$, $b^2 = 010$. We then create the $\lfloor \log_2(2) \rfloor + 2 = 3$ experiments based on these codes: $E_1 = \emptyset, E_2 = \{C_2\}, E_3 = \{C_1\}$.

In conclusion, we have shown that we can guarantee to find the minimal causal variables of a set of causal variables $C_1, ..., C_K$ with as little as $\lfloor \log_2(K) \rfloor + 2$ experiments, of which one can always be the observational regime. In comparison to the bound derived by Eberhardt (2007) for causal discovery ($\lfloor \log_2(K) \rfloor + 1$) in the worst case scenario, we require just exactly one additional experiment to identify the minimal causal variables. The similarity of the two bounds suggests that we can potentially use similar extensions of the causal discovery domain, with minimal adjustments, for causal representation learning, since we may only have to add one more experiment. Such extensions include, for example, limiting the number of simultaneous interventions (Hyttinen et al., 2013) or selecting the cheapest set of experiments according to some cost function (Ghassami et al., 2018; Kocaoglu et al., 2017a; Lindgren et al., 2018).

4 EXPERIMENTS

To verify that CITRIS can operate in a limited experimental setting, we repeat the experiments of Lippe et al. (2022) on the Temporal Causal3DIdent dataset, but with a smaller set of interventions. The Temporal Causal3DIdent dataset consists of 3D renderings (64×64 pixels) of an object shape under varying positions, rotations, and lights. For simplicity, we fix the shape to a teapot, which leaves six causal variables that causally interact over time. Using our bound derived in Section 3.3, we obtain that $\lfloor \log_2(6) \rfloor + 2 = 4$ intervention experiments are sufficient to identify the variables of the Temporal Causal3DIdent Teapot dataset. Hence, we sample four intervention experiments following Algorithm 1, and show one example of the experiment set in Table 2. The first experiment, E_1, is the observational regime where all variables are passively observed, and the other experiments E_2, E_3, E_4 cover the needed interventions. The data is generated following the same process as in Lippe et al. (2022). For the regime variable, we sample the purely observational experiment E_1 50% of the time, and uniformly between E_2, E_3, E_4 otherwise. Hence, we obtain samples from all experiments, with a bias towards observational data, since this is usually cheaper to obtain.

In Table 1, we show the results of CITRIS on the Temporal Causal3DIdent dataset under different experimental settings. The setting "*full experiments*" is the original setup of Lippe et al. (2022), where the intervention targets are independently sampled from a Bernoulli distribution, *i.e.* $I_i^{t+1} \sim \text{Bernoulli}(0.1)$. Hence, we effectively obtain sam-

ples from all possible experiments. In contrast, the setting "*minimal experiments*" only uses four experiments, as described before. We repeat all experiments with three different seeds and three different minimal sets of experiments (see Appendix A.1 for the specific sets). The results on the minimal experiment set show that CITRIS is still able to disentangle the causal variables decently, with small degradation in performance compared to the full experiments. In general, we find that the model is more likely to entangle variables with a very similar intervention pattern due to possible local minima. For example, for the set of experiments in Table 2, one model seed entangled the background hue (hue_b) and the object position (pos_o), which, in terms of interventions, only differ in experiment E_2. Still, CITRIS under a minimal set of experiments considerably outperforms the best baseline model, an iVAE (Khemakhem et al., 2020a), on a full set of experiments. In conclusion, CITRIS can identify the causal variables even under a minimal set of experiments well, but it is more challenging to optimize due to strong dependencies between intervention targets.

5 CONCLUSION

We show that $\lfloor \log_2(K) \rfloor + 2$ intervention experiments are sufficient to identify K causal variables from high-dimensional observations like images for the causal representation learning method CITRIS. This bound has a strong resemblance to the bound in intervention design, which guarantees the discovery of a causal graph for known variables in just one experiment less than in the bound we present. This suggests that work from the field of intervention design can also be adapted to the causal representation learning domain, and the presented work can provide a first step towards this goal.

References

Raghavendra Addanki, Shiva Kasiviswanathan, Andrew Mcgregor, and Cameron Musco. 2020. Efficient Intervention Design for Causal Discovery with Latents. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. 2022. Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*.

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

Frederick Eberhardt. 2007. Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University*.

Frederick Eberhardt. 2008. Almost Optimal Intervention Sets for Causal Discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, Arlington, Virginia, USA. AUAI Press.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. 2005. On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations among N Variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, Arlington, Virginia, USA. AUAI Press.

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

Kristjan Greenewald, Dmitriy Katz, Karthikeyan Shanmugam, Sara Magliacane, Murat Kocaoglu, Enric Boix Adsera, and Guy Bresler. 2019. Sample efficient active learning of causal trees. *Advances in Neural Information Processing Systems*, 32.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. 2021. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*.

Alain Hauser and Peter Bühlmann. 2012. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *Journal of Machine Learning Research*, 13(1).

Alain Hauser and Peter Bühlmann. 2014. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4).

Yang-Bo He and Zhi Geng. 2008. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov).

Antti Hyttinen, Frederick Eberhardt, and Patrik O. Hoyer. 2012a. Causal Discovery of Linear Cyclic Models from Multiple Experimental Data Sets with Overlapping Variables. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, Arlington, Virginia, USA. AUAI Press.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. 2012b. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1).

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

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. 2001. *Independent Component Analysis*. John Wiley & Sons.

Aapo Hyvärinen and Hiroshi Morioka. 2016. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, Red Hook, NY, USA. Curran Associates Inc.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. 2019. 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.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020a. 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.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. 2020b. ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc.

Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. 2017a. Cost-Optimal Learning of Causal Graphs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. 2017b. Experimental Design for Learning Causal Graphs with Latent Variables. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. 2021. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Sebastien Lachapelle, Pau Rodriguez, Rémi Le, Yash Sharma, Katie E Everett, Alexandre Lacoste, and Simon Lacoste-Julien. 2022. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. In *First Conference on Causal Learning and Reasoning*.

Erik Lindgren, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. 2018. Experimental Design for Cost-Aware Learning of Causal Graphs. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. 2022. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In *Proceedings of the 39th International Conference on Machine Learning, ICML 2022*.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. 2020. Weakly-Supervised Disentanglement Without Compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119. PMLR.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. 2019. 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, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*. AUAI Press.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

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

Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. 2017. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*. PMLR.

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. 2015. Learning Causal Graphs with Small Interventions. *Advances in Neural Information Processing Systems*, 28.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. 2020. Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN). In *International Conference on Learning Representations*.

Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. 2020. Active structure learning of causal DAGs via directed clique trees. *Adv. Neural Inf. Process. Syst.*, 33.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. 2022. Learning Temporally Causal Latent Processes from General Temporal Data. In *International Conference on Learning Representations*.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. 2021. Contrastive Learning Inverts the Data Generating Process. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*. PMLR.

	E_1	E_2	E_3	E_4
pos_o	-	✓	-	✓
rot_o	-	-	✓	-
rot_s	-	-	✓	✓
hue_o	-	✓	✓	-
hue_b	-	-	-	✓
hue_s	-	✓	-	-

(a) Experimental setting 1

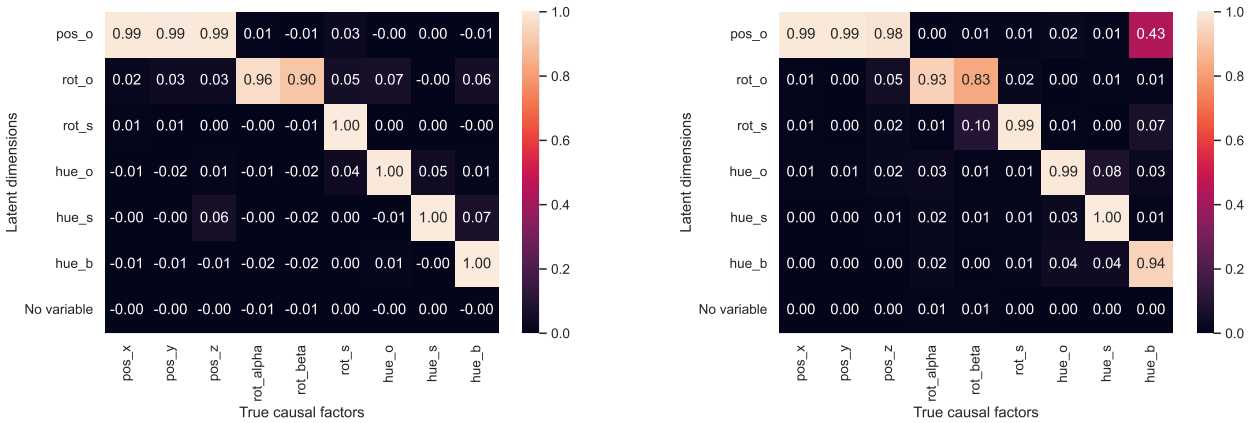
	E_1	E_2	E_3	E_4
pos_o	-	✓	-	-
rot_o	-	-	✓	-
rot_s	-	✓	✓	-
hue_o	-	-	-	✓
hue_b	-	✓	-	✓
hue_s	-	-	✓	✓

(b) Experimental setting 2

	E_1	E_2	E_3	E_4
pos_o	-	✓	-	✓
rot_o	-	-	✓	✓
rot_s	-	-	-	✓
hue_o	-	-	✓	-
hue_b	-	✓	-	-
hue_s	-	✓	✓	-

(c) Experimental setting 3

Figure 2: Minimal sets of experiments used in the Temporal Causal3DIdent experiments. "✓" indicates that a causal variable is intervened in the corresponding experiment. For example, in the first experimental setting (Table 2a), pos_o is intervened in E_2. "-" denotes a passive observation of a variable in this experiment. The three experimental settings were randomly chosen, whereby E_1 is always the observational regime, and each causal variable is at least in one setting only intervened in one experiment, and once intervened in two experiments.



(a) R^2 matrix - full experiments

(b) R^2 matrix - minimal experiments

Figure 3: Correlation matrices of CITRIS-NF for example experiments on the Temporal Causal3DIdent dataset. The y-axis shows the sets of latent dimensions that were assigned to a certain causal factor. The set z_{ψ_0} is represented by 'no variable' in the plots of iCITRIS. The x-axis shows the ground truth causal factors with all dimensions, *i.e.* pos_o represented by pos_x, pos_y, pos_z. The heatmap is the correlation matrix between those factors. **Left**: The results of CITRIS with the full set of experiments (taken from Lippe et al. (2022)). **Right**: An example result of CITRIS on the minimal set of experiments. This experiment was conducted in the experimental setting 1 (Figure 2a).

A EXPERIMENTAL DETAILS

This section provides additional details on the experimental setting. We first show the specific minimal sets of experiments used in the Temporal Causal3DIdent dataset (Appendix A.1). Next, we provide examples of correlation maps for these experiments.

A.1 MINIMAL SETS OF EXPERIMENTS

The chosen sets of experiments are shown in Figure 2. In each set, the experiment E_1 is the observational regime, and E_2, E_3, E_4 cover all conditions for the interventions. The assignment of binary codes to causal variables, as performed in Algorithm 1, is done mostly randomly. To cover all critically different settings, we ensure that for each causal variable, there exists an experimental setting where it is intervened in only one experiment, and that there exists a different experimental setting where it is intervened in two experiments. For instance, pos_o is intervened in only E_2 for the second experimental setting (Figure 2b), and intervened in E_2 and E_4 for the first experimental setting (Figure 2a).

A.2 ADDITIONAL RESULTS

We provide examples of the R^2 correlation matrices between learned latent variables and true causal factors in Figure 3. The left figure shows the result of CITRIS when trained on the full experiment setting, *i.e.* intervention targets are independently sampled. The figure on the right, on

the other hand, shows the correlation matrix for a minimal set of experiments, specifically the setting of Figure 2a. As mentioned in Section 4, the model is still able to disentangle the variables well, but additional correlations between variables with similar intervention patterns can occur. The correlation of 0.43 between pos_o and hue_b is one example for such, since pos_o and hue_b only differ in E_2 in terms of interventions.