UvA-DARE (Digital Academic Repository)

CITRIS: Causal Identifiability from Temporal Intervened Sequences

Lippe, P.; Magliacane, S.; Löwe, S.; Asano, Y.M.; Cohen, T.; Gavves, E.

CITRIS: Causal Identifiability from Temporal Intervened Sequences

Phillip Lippe [1 2]   **Sara Magliacane** [2 3]   **Sindy Löwe** [2 4]   **Yuki M. Asano** [1 2]   **Taco Cohen** [5]   **Efstratios Gavves** [1 2]

Abstract

Understanding the latent causal factors of a dynamical system from visual observations is considered a crucial step towards agents reasoning in complex environments. In this paper, we propose CITRIS, a variational autoencoder framework that learns causal representations from temporal sequences of images in which underlying causal factors have possibly been intervened upon. In contrast to the recent literature, CITRIS exploits temporality and observing intervention targets to identify scalar *and* multidimensional causal factors, such as 3D rotation angles. Furthermore, by introducing a normalizing flow, CITRIS can be easily extended to leverage and disentangle representations obtained by already pretrained autoencoders. Extending previous results on scalar causal factors, we prove identifiability in a more general setting, in which only some components of a causal factor are affected by interventions. In experiments on 3D rendered image sequences, CITRIS outperforms previous methods on recovering the underlying causal variables. Moreover, using pretrained autoencoders, CITRIS can even generalize to unseen instantiations of causal factors, opening future research areas in sim-to-real generalization for causal representation learning.

1. Introduction

Causal representation learning (Khemakhem et al., 2020a; Lachapelle et al., 2022; Locatello et al., 2020a; Schölkopf et al., 2021; Yang et al., 2021) focuses on learning repre-

---

[1]QUVA Lab, University of Amsterdam, Amsterdam, The Netherlands [2]Institute of Informatics, University of Amsterdam, Amsterdam, The Netherlands [3]MIT-IBM Watson AI Lab [4]UvA-Bosch Delta Lab, University of Amsterdam, Amsterdam, The Netherlands [5]Qualcomm AI Research, Amsterdam, The Netherlands. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.. Correspondence to: Phillip Lippe <p.lippe@uva.nl>.
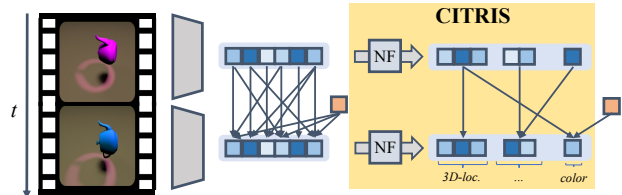
Figure 1. From a sequence of images with interventions (orange), CITRIS learns a normalizing flow, which maps a pretrained autoencoder's entangled latent representation to a causal representation. In this, multiple latent dimensions (blue blocks) can represent a single causal factor, where causal relations exist over time.

sentations of causal factors from high-dimensional observations, such as images. Commonly, in learning causal representations, causal factors are assumed to be scalars. As settings become more complex and high-dimensional, however, so do the causal dynamics, where estimating every scalar causal variable becomes impractical. Consider for instance a set of objects interacting in a three-dimensional space. For each object, we can describe its position by three variables x, y, z and its rotation in multiple angles. However, it is often more natural and sufficient to consider those factors as two *multidimensional* variables, *i.e.* position and rotation, especially when the definition of the default axes is ambiguous. Such multidimensional causal factors are reminiscent of macrovariables explored before in causality (Chalupka et al., 2015; 2016a;b; Hoel et al., 2013; Höltgen, 2021), however, not in causal representation learning yet.

Hence, different from previous causal representation learning approaches, we consider causal factors as potentially multidimensional vectors in this paper. This requires a representation model to learn a latent space of disentangled factors of variation, where latent variables of the same causal factor are grouped together. To identify these multidimensional causal variables, we use sequences of observations where interventions with known targets may have been performed at any time step. This setup resembles a reinforcement learning environment or an interactive real-world setting, with an agent performing actions over time representing interventions on multidimensional causal factors. However, instead of learning a policy to optimize a reward function, we aim at using this setting to identify the causal factors. We assume we can observe the intervention targets in

our data, which occurs for instance when we have access to a set of observation and action trajectories. We refer to this setup as TempoRal Intervened Sequences (TRIS).

Under this setting and common assumptions like an invertible observation function and a stationary causal process, we can identify the *minimal causal variables*, which only model the information of a causal factor that is strictly affected by a provided intervention. With this, we identify the factors which can be directly and independently influenced by, *e.g.*, the action of an agent, and thus of the most practical relevance. Meanwhile, all information that cannot be assigned to a causal factor with certainty and hence not directly influenced, is collected in a separate group of latent variables. As a practical implementation of this, we propose CITRIS for Causal Identifiability from TempoRal Intervened Sequences. CITRIS is a variational autoencoder that learns an assignment of latent variables to causal factors, and promotes disentanglement by conditioning each latent's prior distribution only on its respective intervention target. In experiments on realistically rendered video datasets, CITRIS is able to find and disentangle the causal factors with high accuracy. Moreover, we extend CITRIS to pretrained autoencoders. By using a normalizing flow (Rezende & Mohamed, 2015), CITRIS learns a mapping from the entangled autoencoder representation to a disentangled causal representation, see Figure 1. We empirically show that the normalizing flow can even generalize its disentanglement to unseen instantiations of causal factors, holding promise for future work on generalization of causal representations.

Our contributions are summarized as follows:

- We show that multidimensional causal factors can be identified from temporal sequences with interventions up to their minimal causal variables.

- We propose CITRIS, a VAE architecture for disentangling causal factors in latent space based on this setup.

- Finally, we extend CITRIS to pretrained autoencoders by learning a map from an entangled to a causally disentangled latent space using normalizing flows.

2. Preliminaries and Causal Assumptions

We assume that the underlying latent causal process is a dynamic Bayesian network (DBN) (Dean & Kanazawa, 1989; Murphy, 2002) G over a set of K causal variables. In the corresponding graph $G = (V, E)$, each node $i \in V$ is associated with a causal variable C_i, which can be scalar or vector valued. Each edge $(i, j) \in E$ represents a causal relation from C_i to C_j: $C_i \rightarrow C_j$, where C_i is a *parent* of C_j and $\mathrm{pa}_G(C_i)$ are all parents of C_i in G. Further, we assume that the DBN is first-order Markov, stationary, and without instantaneous effects. This means that in G each causal factor C_i is instantiated at each time step t, denoted by C_i^t, and

its causal parents can only be causal factors at time $t - 1$, denoted as C_j^{t-1}, including its own previous value C_i^{t-1}. In other words, for $t \in [\![1..T]\!]$ and for each causal factor $i \in [\![1..K]\!]$ we can model $C_i^t = f_i(\mathrm{pa}_G(C_i^t), \epsilon_i)$, where $\mathrm{pa}_G(C_i^t) \subseteq \{C_1^{t-1}, \ldots, C_K^{t-1}\}$. We denote the set of all causal variables at time t as $C^t = (C_1^t, \ldots, C_K^t)$, where C^t inherits all edges from its components C_i for $i \in [\![1..K]\!]$ without introducing cycles. In this setting the structure of the graph is time-invariant, *i.e.*, $\mathrm{pa}_G(C_i^t) = \mathrm{pa}_G(C_i^1)$ for any $t \in [\![1..T]\!]$. We also assume all ϵ_i for $i \in [\![1..K]\!]$ are mutually independent noises.

We use a binary intervention vector $I^t \in \{0, 1\}^K$ to indicate that a variable C_i^t in G is intervened upon if and only if $I_i^t = 1$. We consider that the intervention vector components I_i^t might be confounded by another $I_j^t, i \neq j$, and represent these dependencies with an unobserved regime variable R^t (Didelez et al., 2006; Mooij et al., 2020). With this, we construct an augmented DAG $G' = (V', E')$, where $V' = \{\{C_i^t\}_{i=1}^K \cup \{I_i^t\}_{i=1}^K \cup R^t\}_{t=1}^T$ and $E' = \{\{\mathrm{pa}_G(C_i^t) \rightarrow C_i^t\}_{i=1}^K \cup \{I_i^t \rightarrow C_i^t\}_{i=1}^K \cup \{R^t \rightarrow I_i^t\}_{i=1}^K\}_{t=1}^T$. We say that a distribution p is *Markov* w.r.t. the augmented DAG G' if it factors as $p(V') = \prod_{j \in V'} p(V_j \mid \mathrm{pa}_{G'}(V_j))$, where V_j includes the causal factors C_i^t, the intervention vector components I_i^t, and the regime R^t. Moreover, we say that p is *faithful* to a causal graph G', if there are no additional conditional independences to the d-separations one can read from the graph G'. The augmented graph G' can model interventions with an arbitrary number of targets, including observational data. In this paper we will consider *soft* interventions (Eberhardt, 2007), in which the conditional distribution changes, *i.e.*, $p(C_i^t|\mathrm{pa}_G(C_i^t), I_i^t = 1) \neq p(C_i^t|\mathrm{pa}_G(C_i^t), I_i^t = 0)$, which include as a special case *perfect* interventions $\mathrm{do}(C_i = c_i)$ (Pearl, 2009).

3. Identifiability of Minimal Causal Variables

We first describe our setting, TempoRal Intervened Sequences (TRIS). In this setting, we show that identifying the underlying causal factors is not always possible, especially when considering multidimensional causal factors. Therefore, we define the concept of *minimal causal variable*, which represent the manipulable part of each causal factor. Finally, we show under which conditions we can recover the minimal causal variables in TRIS.

3.1. TempoRal Intervened Sequences (TRIS)

In TRIS, we consider data generated by an underlying latent temporal causal process with K *causal factors* $(C_1^t, C_2^t, ..., C_K^t)_{t=1}^T$. At each time step t, we observe a high-dimensional observation X^T representing a noisy, entangled view of all causal factors. The following paragraphs describe the details of this setup, visualized in Figure 2.
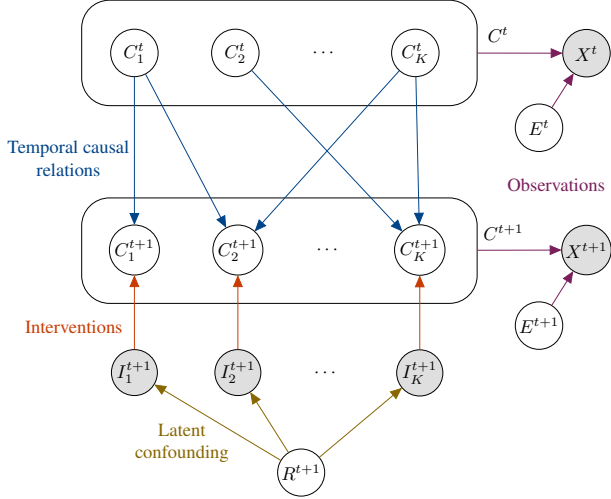
Figure 2. An example causal graph in TRIS, with observed variables shown in gray and latent variables in white. A latent causal factor C_i^{t+1} has as parents a subset of the causal factors at the previous time step $C_1^t, \ldots, C_K^t$, and its intervention target I_i^{t+1}. All causal variables C^{t+1} and noise E^{t+1} cause the observation X^{t+1}. R^{t+1} is a latent confounder between the intervention targets.

Multidimensional Causal Factors: As opposed to most work on causal representation learning, which considers causal factors to be one-dimensional (Khemakhem et al., 2020a; Klindt et al., 2021; Lachapelle et al., 2022), we allow them to be potentially multidimensional, *i.e.*, $C_i \in \mathcal{D}_i^{M_i}$ with $M_i \geq 1$ and in practice we let $\mathcal{D}_i$ be $\mathbb{R}$ for continuous variables (*e.g.*, spatial position), $\mathbb{Z}$ for discrete variables (*e.g.*, the score of a player) or mixed. This allows modeling different levels of causal variables (*e.g.* a 2D-position encoded in a single factor with two dimensions instead of two different causal factors). We define the causal factor space as $\mathcal{C} = \mathcal{D}_1^{M_1} \times \mathcal{D}_2^{M_2} \times ... \times \mathcal{D}_K^{M_K}$.

Observation Function: We define the observation function $h(C_1^t, C_2^t, ..., C_K^t, E_o^t) = X^t$, where E_o^t represents any noise independent of the causal factors that influence the observations, and $h : \mathcal{C} \times \mathcal{E} \rightarrow \mathcal{X}$ is a function from the causal factor space $\mathcal{C}$ and the space of the noise variables $\mathcal{E}$ to the observation space $\mathcal{X}$. We assume that h is bijective, implying that the joint dimensionality of the noise and causal model is limited to the image size. This allows us to identify each causal factor uniquely from observations by learning an approximation of f, while disregarding irrelevant features in the observation space.

Availability of Intervention Targets: Crucially, we assume that in each time-step some causal factors might (or might not) have been intervened upon and that we have access to the corresponding intervention targets, but not the intervention values. We denote these intervention targets by the binary vector $I^t \in \{0, 1\}^K$ where $I_i^t = 1$ refers to an intervention on the causal variable C_i^t.
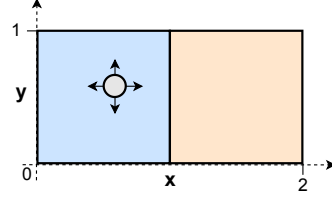

Figure 3. Example causal process of a ball with a two-dimensional position x, y. The ball can only swap between the two boxes (blue and orange) through an intervention, which does not influence the relative position within a box.

3.2. A Necessary Condition for Disentanglement in TRIS

In TRIS, we generally cannot disentangle two causal factors if they are always intervened upon jointly, or, on the contrary, if they are never intervened upon.

Proposition 3.1. *In TRIS, if two causal factors C_i and C_j have only been jointly intervened on or not at all, then there exists a causal graph in which C_i and C_j cannot be uniquely identified from observations X and intervention targets I.*

We provide an example of such a graph in Figure 3, where a ball can move in two dimensions, x and y. If both x and y follow a Gaussian distribution with stationary variances over time, then any two orthogonal axes can describe the distribution equally well (Belouchrani et al., 1997; Hyvarinen & Morioka, 2017), making it impossible to uniquely identify them without interventions. Similarly, if we only observe joint interventions on x, y together, we cannot identify them either due to the same reasoning. We include the proof for this proposition in Appendix B.6.

Additionally, in TRIS where the latent causal factors may correspond to multidimensional vectors, we cannot even completely reconstruct said factors, when by the nature of the system the provided interventions leave some of the causal factor's dimensions unaffected. In the next section, we will instead introduce the concept of minimal causal variables to characterize what we can identify instead.

3.3. Minimal Causal Variables

To visualize this scenario, consider again the example in Figure 3 with a ball in one of two boxes. Over time, the ball can move freely within the box it is currently in, but it can only jump into another box if there is an intervention. The intervention moves the ball to the other box, but keeps the relative position of the ball within the box intact. While one could define this process by a single causal variable x over time, it can also be described by two causal variables: the relative position within the box x' and the current box b. Since only b is affected by the intervention, and we consider causal factors to potentially be multidimensional, we could

not identify which causal factor x' belongs to.

We formalize this intuition as follows. Suppose for each causal factor $C_i \in \mathcal{D}^{M_i}$, there exists an invertible map $s_i : \mathcal{D}_i^{M_i} \to \mathcal{D}_i^{\text{var}} \times \mathcal{D}_i^{\text{inv}}$ that splits the domain $\mathcal{D}^{M_i}$ of C_i into a part that is variant and a part that is invariant under intervention. We denote the two parts of this map as

$$s_i(C_i^t) = (s_i^{\text{var}}(C_i^t), s_i^{\text{inv}}(C_i^t)) \tag{1}$$

The split s must be invertible, so that we can map back and forth between $\mathcal{D}_i^{M_i}$ and $\mathcal{D}_i^{\text{var}} \times \mathcal{D}_i^{\text{var}}$ without losing information. Furthermore, to be called a split in this setup, s must satisfy $s_i^{\text{inv}}(C_i^t) \perp\!\!\!\perp I_i^t \mid pa(C_i^t)$, i.e., $s_i^{\text{inv}}(C_i^t)$ is independent of the intervention variable I_i^t given the parents of C_i^t. Also, both parts of the split must be conditionally independent, i.e. $s_i^{\text{inv}}(C_i^t) \perp\!\!\!\perp s_i^{\text{var}}(C_i^t) \mid pa(C_i^t), I_i^t$. This means that $s_i^{\text{var}}(C_i^t)$ will contain the manipulable, or *variable*, part of C_i^t. In contrast, $s_i^{\text{inv}}(C_i^t)$ is the *invariable* part of C_i^t which is independent of the intervention.

For any causal variable, there may exist multiple possible splits. There is always at least the trivial split where $\mathcal{D}_i^{\text{var}} = \mathcal{D}_i^{M_i}$ is the original domain of C_i, and $\mathcal{D}_i^{\text{inv}} = \{0\}$ is the one-element set (no invariant information). But not all splits are trivial: For the example in Figure 3, we can split the causal factor x in $s^{\text{var}}(x)$ such that the box identifier b is in $s^{\text{var}}(x)$ and the relative position in the box x' is in $s^{\text{inv}}(x)$.

Intuitively, we want to identify the split where s_i^{var} contains *only* the manipulable information:

Definition 3.2. *The* minimal causal split *of a variable C_i^t with respect to its intervention variable I_i^t is the split s_i which maximizes the information content $H(s_i^{\text{inv}}(C_i^t) | pa(C_i^t))$. Under this split, $s_i^{\text{var}}(C_i^t)$ is defined as the* minimal causal variable *and denoted by $s_i^{\text{var}^*}(C_i^t)$.*

Here H denotes the entropy in the discrete case, or the limiting density of discrete points (LDDP) (Jaynes, 1957; 1968) which measures information content of a continuous variable and is invariant under the change of variables induced by the map s. Intuitively, this ensures that only the information which truly depends on the intervention is represented in $s_i^{\text{var}}(C_i)$. The definition of the minimal causal split depends on the characteristics of the provided intervention. In our previous example, when an intervention on x would also change the internal box position, the minimal causal variable would contain the full causal factor x. Our goal becomes to identify these minimal causal variables.

Depending on the interventions, the causal graph among the minimal causal graph may differ from the original graph on $C_1, ..., C_K$. If the interventions are soft, the minimal causal variable, $s_i^{\text{var}}(C_i)$, has as parents the subset of $pa(C_i)$, whose relation to C_i is influenced by the intervention. For instance, consider a three-dimensional causal variable, of which each dimension has a different set of parents. If an

intervention only affects the first dimension, $s_i^{\text{var}}(C_i)$ has the same parents as the first dimension, and $s_i^{\text{inv}}(C_i)$ the same parents as the last two dimensions. For perfect interventions, the intervention-invariant part, $s_i^{\text{inv}}(C_i)$, has no parents since all temporal dependencies are influenced by the intervention. Thus, in this case, the parents of $s_i^{\text{var}}(C_i)$ are the same as of the true causal variable, $pa(C_i)$.

3.4. Learning Minimal Causal Variables

As a practical example of TRIS, we consider a dataset $\mathcal{D}$ of tuples $\{x^t, x^{t+1}, I^{t+1}\}$ where $x^t, x^{t+1} \in \mathbb{R}^N$ represent the observations at time step t and $t+1$ respectively, and I^{t+1} describes the targets of the interventions performed on C^{t+1}. To learn a causal representation, we consider a latent space $\mathcal{Z}$ larger than the latent causal factor space $\mathcal{C}$, i.e. $\mathcal{Z} \subseteq \mathbb{R}^M, M \geq \dim(\mathcal{E}) + \sum_{i=1}^K M_i = \dim(\mathcal{E}) + \dim(\mathcal{C})$. In this latent space, we aim to disentangle the causal factors. Since we may not know the exact dimensions $M_1, ..., M_K$, we overestimate its size in the latent space $\mathcal{Z}$.

Our goal is to approximate the inverse of the observation function h by learning two components. First, we learn an invertible mapping from observations to latent space, $g_\theta : \mathcal{X} \to \mathcal{Z}$. Second, we learn an assignment function $\psi : [\![1..M]\!] \to [\![0..K]\!]$ that maps each dimension of $\mathcal{Z}$ to a causal factor. Learning a flexible assignment function ψ allows us to allocate any dimension size per causal factor without knowing the individual causal factor dimensions $M_1, ..., M_K$ in advance. Further, it also benefits the optimization process, since some variables like circular angles or categorical factors with many categories can have simpler distributions when modelled in more dimensions, which we verify empirically in Section 6.2. In addition to the K causal factors, we use $\psi(j) = 0, j \in [\![1..M]\!]$ to indicate that the latent dimension z_j does not belong to any minimal causal variable. Instead, those dimensions might model $s_i^{\text{inv}}(C_i)$ for some causal factor C_i or the observation noise E_o^t. Finally, we denote the set of latent variables that ψ assigns to the causal factor C_i with $z_{\psi_i} = \{z_j | j \in [\![1..M]\!], \psi(j) = i\}$.

To enforce a disentanglement of causal factors, we model a prior distribution in latent space, $p_\phi(z^{t+1}|z^t, I^{t+1})$, with $z^t, z^{t+1} \in \mathcal{Z}, z^t = g_\theta(x^t), z^{t+1} = g_\theta(x^{t+1})$. This transition prior enforces a disentanglement by conditioning each latent variable on exactly one of the intervention targets:

$$p_\phi\left(z^{t+1}|z^t, I^{t+1}\right) = \prod_{i=0}^K p_\phi\left(z_{\psi_i}^{t+1}|z^t, I_i^{t+1}\right) \tag{2}$$

where $I_0^{t+1} = 0$. Then, the objective of the model is to maximize the likelihood:

$$p_{\phi,\theta}(x^{t+1}|x^t, I^{t+1}) = \left|\frac{\partial g_\theta(x^{t+1})}{\partial x^{t+1}}\right| p_\phi(z^{t+1}|z^t, I^{t+1}) \tag{3}$$

Under the assumptions stated in Section 3.1, we can prove

the following identifiability result for this setup:

Theorem 3.3. *Suppose that ϕ^*, θ^* and ψ^* are the parameters that, under the constraint of maximizing the likelihood $p_{\phi,\theta}(x^{t+1}|x^t, I^{t+1})$, maximize the information content of $p_\phi(z^{t+1}_{\psi_0}|z^t)$. Then, with sufficient latent dimensions, the model ϕ^*, θ^*, ψ^* learns a latent structure where $z^{t+1}_{\psi_i}$ models the minimal causal variable of C_i if $C^{t+1}_i \not\perp\!\!\!\perp I^{t+1}_i|C^t, I^{t+1}_j$ for any $i \neq j$. All remaining information is modeled in z_{ψ_0}.*

We provide the proof for this statement in Appendix B, which relies on I^{t+1}_i not being a deterministic function of any other intervention target. A sufficient condition for this is the interventions being independent of each other, or single-target interventions with observational data. Finding the minimal variables intuitively means that the latent variables z_{ψ_i} model only the information of C_i which strictly depends on the intervention target I^{t+1}_i, thus defining causal variables by their intervention dependency. Going back to the example of the x position of the ball in Figure 3, we would model the box identifier in z_{ψ_1} if $C_1 = x$, while the internal box position is modeled in z_{ψ_0}. We empirically verify the learned split for this example in Appendix D.3. Nonetheless, one can always ensure that for any definition of the causal process, z_{ψ_i} only contains information of causal variable C_i of that process, and no other causal variable.

4. Causal Identifiability from Temporal Intervened Sequences

To identify causal factors from high-dimensional temporal observations with interventions, we propose CITRIS (Causal Identifiability from Temporal Intervened Sequences). Below, we discuss its architecture and variants.

4.1. Variational Autoencoder Setup

Inspired by several previous works (Higgins et al., 2017; Locatello et al., 2020a; Träuble et al., 2021), we implement the framework of Section 3.4 by learning a variational autoencoder (VAE) (Kingma & Welling, 2014), visualized in Figure 4. The encoder q_θ and decoder p_θ approximate the invertible mapping g_θ from observations to latent space, and $p_\phi\left(z^{t+1}|z^t, I^{t+1}\right)$ is the transition prior on the latent variables. In this VAE setup, the objective of the model becomes the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{z^{t+1}}\left[\log p_\theta\left(x^{t+1}|z^{t+1}\right)\right] +$$

$$\mathbb{E}_{z^t,\psi}\left[\sum_{i=0}^{K} D_{\text{KL}}\left(q_\theta(z^{t+1}_{\psi_i}|x^{t+1})||p_\phi(z^{t+1}_{\psi_i}|z^t, I^{t+1}_i)\right)\right] \quad (4)$$

The KL divergence uses the prior definition of Equation (2). This ensures that, conditioned on the previous time step and the interventions, the different blocks of latent variables are independent. Thereby, the assignment function of latent to causal variables, ψ, is learned via a Gumbel-Softmax distribution (Jang et al., 2017) per latent variable. Hence, during training, we sample a latent-to-causal variable assignment from these distributions, while for inference, we can use the argmax to obtain a unique assignment. To encourage information independent of any intervention to be modeled in z_{ψ_0}, we weight the KL divergence of z_{ψ_0} with $1 - \lambda$, where $\lambda > 0$ is a hyperparameter (usually $\lambda = 0.01$).

The prior p_ϕ for each set of latents $z^{t+1}_{\psi_i}$ is implemented by an autoregressive model. For each set of latents $z^{t+1}_{\psi_i}$, the model takes z^t, I^{t+1}_i and z^{t+1} as input, where we sample from ψ and mask the dimensions of z^{t+1} for which $\psi(j) \neq i$. From this input, the model predicts one Gaussian per latent variable. The autoregressive nature of the prior allows complex distributions over the multiple latent dimensions, while still being independent across causal variables.

Target Classifier Since a causal variable C_i is independent of any other target variable when conditioned on I_i, we can use those independence relations to guide the disentanglement in latent space. For this, we propose a target classifier, which is an additional small network trained to predict the intervention targets from the latent variables over time, *i.e.* modeling $p(I^{t+1}|z^t, z^{t+1}_{\psi_i})$ for $i \in [\![0..K]\!]$. To guide the disentanglement, we take the gradients of the same objective with respect to $z^t, z^{t+1}_{\psi_i}$ and ψ, but only for I^{t+1}_i. In other words, we explicitly train $z^{t+1}_{\psi_i}$ having a higher mutual information with the intervention target I^{t+1}_i of its intended causal variable C_i. For all other targets I^{t+1}_j, we change the label to the conditional distribution $p(I_j|I_i)$, since I^{t+1}_j should not be fully identifiable by any other set of latents except $z^{t+1}_{\psi_j}$. This way, the target classifier ensures that information about all causal variables is encoded in the latent space, while also guiding the disentanglement in latent space. We provide further details in Appendix C.3.3, and empirically verify its benefits in Section 6.2.

4.2. Using Pretrained Autoencoders

Despite the improved optimization process by the addition of the target classifiers, VAEs can still struggle to model high-dimensional complex images, especially when small details in the image are relevant. To overcome this issue, we propose an adaptation of CITRIS to pretrained autoencoders. In this setting, the invertible map g_θ is implemented by a deep autoencoder, which is trained to encode and decode the high-dimensional observations to low-dimensional feature vectors, independently of any disentanglement. During training, we add small Gaussian noise to the latents to prevent the latent distribution from collapsing to single delta peaks. However, we do not enforce a certain prior like in a VAE, hence allowing complex marginal distributions.
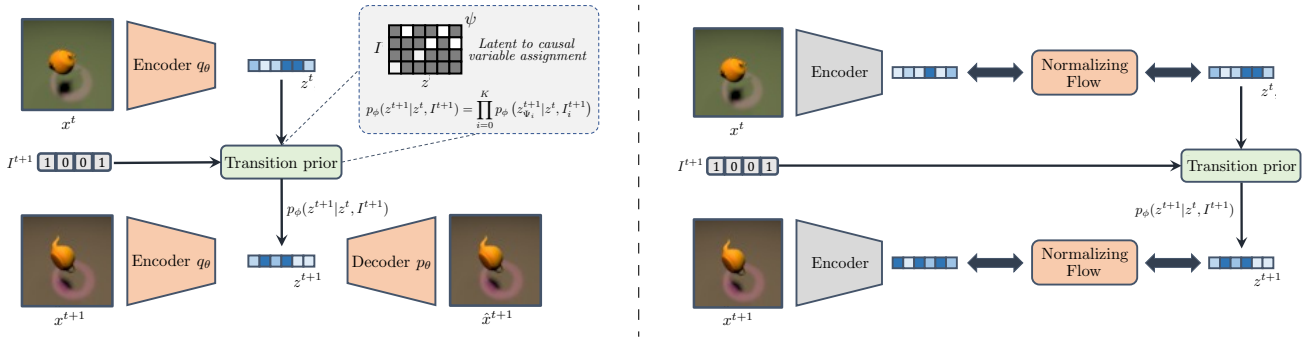
Figure 4. Comparing the VAE and AE+NF setup of CITRIS. **Left**: In the VAE, the encoder and decoder provide an approximate invertible mapping. The transition prior promotes disentanglement in the latent space by conditioning each latent variable on only one intervention target. **Right**: The Normalizing Flow setup uses a pretrained autoencoder, which remains frozen during training. The flow learns to map the autoencoder latents to a new space, that promotes the disentanglement similar to the VAE. The decoder is not needed during training.

In a second step, after the autoencoder converged, we freeze its parameters and learn a normalizing flow (Rezende & Mohamed, 2015) that maps the entangled latent representation to a disentangled version. The invertibility of the normalizing flow ensures that no information is lost when mapping from the entangled to the disentangled latent space, and thus we can use the pretrained decoder to reconstruct the observations without requiring any fine-tuning. Compared to the VAE setup in Section 4.1, we replace the encoder by a successive application of the frozen encoder and a normalizing flow on the encoded latents, shown in Figure 4. Besides that, we deploy the same setup, using both the transition prior structure and target classifier.

Compared to the VAE setup, this approach has three main benefits. First, the autoencoder can be trained on observational data alone, potentially reducing the amount of interventional data required. Second, learning a separate autoencoder provides an opportunity for generalizing causal factors beyond the known dataset. For instance, one could train an autoencoder on two datasets, where only one has interventions, *e.g.* synthetic and real-world data. Then, since the autoencoder uses a joint latent space for both datasets, training the normalizing flow on only the dataset with interventions can lead to a disentanglement function that generalizes to the purely observational dataset. We verify the viability of this approach in a restricted setting in Section 6.2, which opens up great potential to practical applications. Finally, the setup is easier to optimize since the autoencoder can compress the information in an almost unrestricted latent space, while the normalizing flow solely focuses on disentanglement.

5. Related Work

Identifying independent factors of variations from data is a well-studied field in machine learning (Higgins et al., 2017; Klindt et al., 2021; Kumar et al., 2018; Locatello et al., 2019; 2020a;b; Reed et al., 2014). 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 that were transformed by some invertible transformation. Although not generally possible in the non-linear case (Hyvärinen & Pajunen, 1999), ICA was recently extended to this setting by exploiting auxiliary variables under which the latents become conditionally mutually independent (Hyvärinen & Morioka, 2016; Hyvärinen et al., 2019). Several follow-up works extended this work to deep learning architectures like VAEs (Khemakhem et al., 2020a;b; Sorrenson et al., 2020; Zimmermann et al., 2021). 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 and bring it into context to causality. While Lachapelle et al. (2022) can model interventions as external actions, Yao et al. (2022) can model soft interventions through their non-stationary noise. On the other hand, they do not exploit the knowledge of the intervention targets as we do and therefore require additional assumptions in terms of sufficient variation. Moreover, both of these works require scalar causal variables, while CITRIS generalizes to multidimensional causal factors.

A second, related line of work is causal representation learning (Schölkopf et al., 2021), which aims at discovering causal structures and variables from data. For instance, Locatello et al. (2020a) showed that one can identify independent latent causal factors from pairs of observations that only differ in k causal factors. Yang et al. (2021) propose a VAE that integrates a structural causal model in its prior, but requires the true causal variables as labels during training. von Kügelgen et al. (2021) demonstrated that common contrastive learning methods can block-identify the causal variables that remain unchanged under augmentations. CITRIS similarly identifies multidimensional causal factors as a block, and, furthermore, disentangles individual causal factors by grouping the latent space.
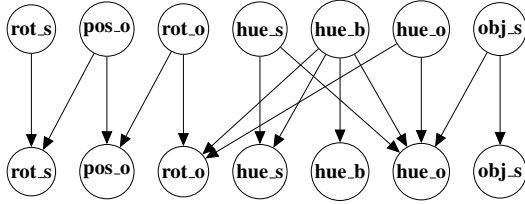
Figure 5. The temporal causal graph among the 7 dimensions of variation. The graph structure covers a wide range of scenarios, such as confounders, chains, and up to 4 parents per variable.

Figure 6. Examples of Interventional Pong. The score represents a causal factor on which we do not have any interventions. The velocity of the ball is encoded in a 4th channel, not visualized here.

6. Experiments

We evaluate CITRIS on two video datasets, and compare it to common disentanglement methods. We include further details for reproducibility in Appendix C, and make our code publicly available.[1]

6.1. Experimental Setup

Temporal Causal3DIdent We evaluate CITRIS on an adapted, temporal version of the Causal3DIdent identifiability benchmark (von Kügelgen et al., 2021; Zimmermann et al., 2021). The dataset consists of 3D renderings of objects, for which we consider seven different causal factors: the object position as multidimensional vector $[x, y, z] \in [-2, 2]^3$; the object rotation with two dimensions $[\alpha, \beta] \in [0, 2\pi)^2$; the hue of the object, background and spotlight in $[0, 2\pi)$; the spotlight's rotation in $[0, 2\pi)$; and the object shape (categorical with seven values). The relations among those variables are shown in Figure 5. Each continuous variable follows a Gaussian distribution over time, where the mean is a (non-linear) function of the parents. Overall, this causes strong dependencies among the variables. We perform perfect interventions with $I_i^t \sim \text{Bernoulli}(0.1)$, ensuring the minimal causal variables to be the true factors. Experiments with confounded interventions are included in Appendix D.1 with similar results. Further details on this dataset are summarized in Appendix C.1.

Interventional Pong As a second benchmark, we adapt the popular Atari game Pong (Bellemare et al., 2013), in which we define 5 causal factors: the x, y position of the ball, the y-positions of the two paddles, and the velocity direction of the ball (angle in $[0, 2\pi)$). Further, we add the score as a sixth causal factor, which is part of the game's dynamics. However, we do not provide any interventions on it. Hence, in CITRIS, this information should be modeled in z_{ψ_0}. The temporal dependencies follow the dynamics of Pong, where the ball can collide with the paddles and walls, and the two paddles move towards the ball to hit it. We provide single-target interventions for the variables. For the paddles, the imperfect interventions override the policy and take a
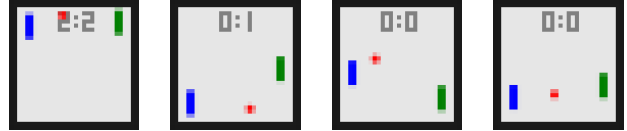
random action (up/down), while keeping a smooth motion. For the causal factors of the ball, we uniformly sample a new value when intervened, causing a possible jump. Example frames of the environment are shown in Figure 6.

Baselines We compare CITRIS to SlowVAE (Klindt et al., 2021), a state-of-the-art disentanglement method for temporal sequences. Notably, SlowVAE assumes that the factors of variation are independent. As we show, this assumption, which often cannot be met in more complex settings like Temporal Causal3DIdent, has detrimental effects on modelling, underlining the importance of modelling the correlations between causal variables. The second baseline we consider is iVAE (Khemakhem et al., 2020a), a VAE model which conditions its components on additional observed variables u. In our setup, u corresponds to the previous time step observation x^t and the intervention targets I^{t+1}. As we aim to find a mapping from image to a causal space, which is independent of those factors, we must adapt iVAE to only condition its prior on u. We refer to this model variant as iVAE*. With this change, the main difference between iVAE* and CITRIS becomes the structure of the prior $p(z_{t+1}|z_t, I_{t+1})$, showcasing the significance of CITRIS's identifiability guarantees in this setup. Hyperparameter details are listed in Appendix C, and the code is provided in the supplementary material.

Correlation Metrics Following common practice, we report the correlation of the learned latent variables to the ground truth causal factors. Since in our setup, multiple latent variables can jointly describe a single causal variable, we first learn a mapping between such, *e.g.*, with an MLP. For CITRIS, we apply one MLP per set of latent variables that are assigned to the same causal factor by ψ. The MLP is then trained to predict *all* causal factors per set of latents, on which we measure the correlation. Thereby, no gradients are propagated through the model. iVAE* and SlowVAE do not learn an assignment of latent to causal factors. As an alternative, we assign each latent dimension to the causal factor it has the highest correlation with. Although this gives the baselines a considerable advantage, it shows whether CITRIS can improve upon the baselines beyond finding a good latent to causal factor assignment. We report both the R^2 coefficient of determination (Wright, 1921) and the Spearman's rank correlation coefficient (Spearman, 1904)

---

[1]https://github.com/phlippe/CITRIS

Table 1. Results on the Temporal-Causal3DIdent dataset. All triplet distances have a distance at random of 1 and optimum of 0. *Oracle* represents the accuracy of a CNN trained supervised to predict the causal factors from images, hence constituting a lower error limit. In the correlation metrics, *diag* refers to the average score of the predicted causal factor to its true value (optimal 1), and *sep* for the average of the maximum correlation per predicted causal variable besides its true factor (optimal 0). Standard deviations over 3 seeds shown in Appendix C.1. CITRIS is able to disentangle the causal factors well, and CITRIS-NF even accurately models all 7 shapes.

	Triplet evaluation distances ↓											Correlation metrics			
	pos_x	pos_y	pos_z	rot_α	rot_β	rot_s	hue_s	hue_b	hue_o	obj_s	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Temporal-Causal3DIdent Teapot															
Oracle	0.02	0.02	0.02	0.02	0.03	0.01	0.02	0.01	0.02	-	0.02	-	-	-	-
SlowVAE	0.13	0.10	0.12	0.50	0.59	0.22	0.64	0.21	0.17	-	0.30	0.65	0.20	0.62	0.27
iVAE[*] - 9dim	0.11	0.09	0.12	0.70	0.76	0.06	0.67	0.02	0.12	-	0.30	0.65	0.11	0.65	0.13
iVAE[*] - 32dim	**0.04**	**0.03**	**0.04**	0.25	0.31	**0.03**	0.58	0.02	0.05	-	0.15	0.78	0.21	0.77	0.17
CITRIS-VAE	0.05	0.04	0.05	0.10	0.20	**0.03**	0.08	0.02	0.05	-	0.07	0.96	0.02	0.95	**0.04**
- No target classifier	0.05	0.04	0.05	0.62	0.66	0.19	**0.04**	0.02	0.17	-	0.20	0.79	0.15	0.76	0.12
CITRIS-NF	**0.04**	**0.03**	**0.04**	**0.06**	**0.10**	**0.03**	**0.04**	**0.01**	**0.04**	-	**0.04**	**0.98**	**0.01**	**0.97**	0.05
Temporal-Causal3DIdent 7-shapes															
Oracle	0.08	0.06	0.08	0.06	0.09	0.04	0.04	0.01	0.04	0.00	0.05	-	-	-	-
SlowVAE	0.44	0.25	0.41	0.69	0.75	0.25	0.57	0.10	0.14	0.37	0.40	0.61	0.23	0.59	0.27
iVAE[*]	0.26	0.23	0.34	0.58	0.65	0.10	0.31	0.02	0.09	0.14	0.27	0.80	0.29	0.77	0.28
CITRIS-VAE	0.15	0.13	0.23	0.54	0.71	0.07	0.05	0.02	0.06	0.18	0.21	0.89	0.10	0.88	0.12
CITRIS-NF	**0.12**	**0.08**	**0.11**	**0.09**	**0.14**	**0.05**	**0.05**	**0.02**	**0.06**	**0.00**	**0.07**	**0.98**	**0.04**	**0.97**	**0.08**

of the predicted values against the ground truth causal factors. However, in our datasets, the causal factors are correlated themselves. This makes it difficult to spot spurious correlations between latents and causal factors. To overcome this issue, we measure the correlations on a test dataset for which we sample the causal factors independently.

Triplet Evaluation To also evaluate the decoding part of the model, we propose another parameter-free evaluation, *triplet evaluation*, to reveal complex dependencies between latent variables. For this, we create triplets of images: the first two are randomly sampled test images, while the third one is created based on a random combination of causal factors of the first two images. For example, in Figure 7, we take the spotlight rotation and object shape from image 1, and all other causal factors from image 2. For evaluation, we then encode the two test images independently, perform the combination of ground-truth causal factors as done for the third image in latent space, and use the decoder to generate a new image, which ideally resembles the ground truth third image. Since the reconstruction error is not descriptive of the errors being made, *e.g.* the rotation in Figure 7, we train an additional CNN in a supervised manner that maps images to the causal factors. With this model, we can extract the causal factors from the generated image, and report the average distance of these to the ground truth causal factors.

6.2. Temporal Causal3DIdent Experiments

Teapot Experiments The Causal3DIdent benchmark constitutes a challenging dataset due to its various interactions of causal factors in the high-dimensional observational space. To slightly simplify the problem, especially for learning a VAE on the high-dimensional observational space, we first show experiments on the teapot shape only, Temporal-Causal3DIdent Teapot. All results are summarized in Ta-
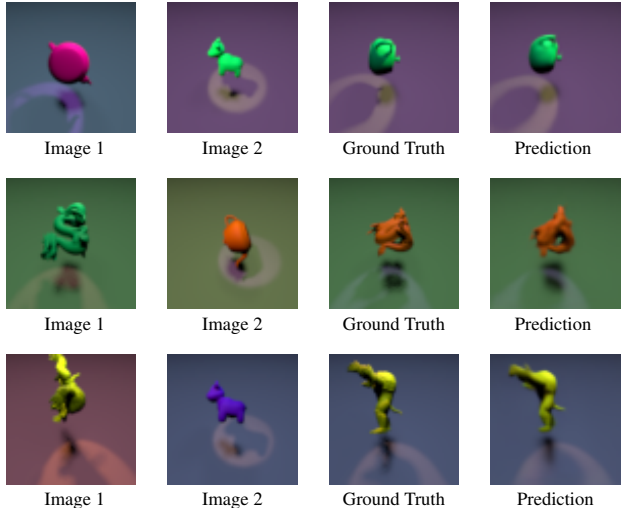

Figure 7. Example triplet evaluations on the Temporal-Causal3DIdent. In the first row, the ground truth combines the spotlight rotation and object shape from Image 1 with all other causal factors from Image 2. The prediction is generated by encoding the two images, and performing the same combination of causal factors in latent space. The shown prediction by CITRIS-NF only slightly differs in the object rotation. For details on the other predictions, see Appendix D.1.

ble 1. Both the VAE and Normalizing Flow version of CITRIS considerably outperformed the two baselines and were able to achieve an average R^2 and Spearman correlation above 0.9, while keeping the correlation between factors low. Moreover, CITRIS-NF achieves close-to optimal scores on the triplet evaluation, and especially outperforms the VAE-based approaches in modeling the rotations. This is because the autoencoder is able to reconstruct images with negligible error, while the VAE optimization involves balancing

Table 2. Results on the Temporal-Causal3DIdent dataset with CITRIS-NF trained on 5 object shapes. The same metrics as in Table 1 are reported, with "Spr" being the Spearman correlation score. The full table is in Appendix D.1. For 2 unseen shapes, CITRIS-NF can yet disentangle their causal factors well.

CITRIS-NF	Triplets ↓	R^2 diag ↑	R^2 sep ↓	Spr diag ↑	Spr sep ↓
5 seen shapes	0.09	0.98	0.05	0.97	0.10
2 unseen shapes	0.23	0.94	0.15	0.93	0.19

Table 3. Results on the Interventional Pong dataset. The same metrics as in Table 1 are reported, with "Spr" being the Spearman correlation score. The full table is in Appendix D.2. The results how that CITRIS can handle imperfect interventions.

	Triplets ↓	R^2 diag ↑	R^2 sep ↓	Spr diag ↑	Spr sep ↓
SlowVAE	0.34	0.61	0.17	0.66	0.23
iVAE*	0.09	0.91	0.04	0.92	0.06
CITRIS-VAE	0.03	0.99	**0.01**	0.99	**0.05**
CITRIS-NF	**0.02**	**1.00**	0.04	**1.00**	0.10

the small difference in output space with its KL divergence regularization in latent space. Further, an ablation study on training CITRIS-VAE without the target classifier verifies that the additional guidance can considerably help in optimization. Finally, in Appendix D.1.4, we show that the original causal graph can be found from CITRIS's latents.

In contrast, the SlowVAE entangles the causal factors due to their strong correlation over time. The iVAE* disentangles the position and the rotation of the spotlight well. However, the hue of the spotlight was highly entangled since its appearance differs with different background and object colors. We further report the performance of an iVAE* with equal number of latent dimensions as causal factors. It is apparent that the smaller latent space is particularly insufficient for modeling the rotation angles, showing the need for higher dimensional latent spaces, not only in CITRIS.

7-shapes Experiments We apply all models on the Temporal Causal3DIdent dataset with all seven shapes (Table 1). The models have to align the rotation axes and central points of all shapes, which caused higher triplet distances for rotations across models. While CITRIS-VAE still outperformed the other baselines, CITRIS-NF significantly improved upon that, still maintaining a disentanglement similar to the single shape experiments. This underlines the optimization benefits of using pretrained autoencoders for disentanglement learning on complex, high-dimensional observations.

Generalization of Causal Representations Since the autoencoder is trained on observational data, we evaluate whether its causal representation can generalize to new, unseen settings. For this, we reuse the same autoencoder as before, but train the Normalizing Flow on an interventional dataset which excludes any observations from two shapes (Head and Cow). Afterwards, we test its zero-shot generalization to the two unseen shapes (see Table 2). Note that optimal performance cannot be achieved here, since the central point and default rotation of an object cannot be generalized to other objects. Nonetheless, the results in Table 2 indicate a strong disentanglement among factors, with slight decreases in position and rotation due to the forementioned limitations. This shows that the learned disentanglement function can indeed generalize to unseen instantiations of causal factors, promising potential for future work on generalizing causal representations to unseen settings with CITRIS.

6.3. Interventional Pong

Finally, we report the disentanglement results of CITRIS on the Interventional Pong dataset in Table 3. The challenge of this dataset is its imperfect and correlated interventions, and the score being an unintervened causal variable. Still, CITRIS is able to disentangle the factors and also assign the score variables to its correct set of latents, *i.e.* z_{ψ_0}, showing that it can indeed handle imperfect interventions. Although CITRIS-NF obtains a higher correlation among learned latents, it again did not affect its triplet generations. SlowVAE entangles the two paddles, since they are strongly correlated by following the same movement policy and having imperfect interventions. iVAE* showed unstable behavior over seeds, and entangled the ball position and velocity.

7. Conclusion

We propose CITRIS, a VAE framework for learning causal representations. CITRIS identifies the minimal causal variables of a dynamical system from temporal, intervened sequences. Furthermore, by using normalizing flows, CITRIS learns to disentangle the representation of pretrained autoencoders. In experiments, CITRIS reliably recovered the causal factors of 3D rendered images. Moreover, we empirically showed that CITRIS can generalize to unseen instantiations of causal factors. This promises great potential for future work on simulation-to-real generalization research for causal representation learning. As future work, CITRIS can be extended to an active learning setup, allowing for more data-efficient causal identifiability methods in practice. Further, future work could consider the setup of robotic simulators (Makoviychuk et al., 2021; Szot et al., 2021), where interventions are available through sequences of actions.

Acknowledgements

References

Ba, L. J., Kiros, J. R., and Hinton, G. E. Layer Normalization. *CoRR*, abs/1607.06450, 2016.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The Arcade Learning Environment: An Evaluation Platform for General Agents. *J. Artif. Int. Res.*, 47(1): 253–279, may 2013. ISSN 1076-9757.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., and Moulines, E. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997. doi: 10.1109/78.554307.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2021. `http://www.blender.org`.

Chalupka, K., Perona, P., and Eberhardt, F. Visual Causal Feature Learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pp. 181–190, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Chalupka, K., Bischoff, T., Perona, P., and Eberhardt, F. Unsupervised Discovery of El Nino Using Causal Feature Learning on Microlevel Climate Data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 72–81, Arlington, Virginia, USA, 2016a. AUAI Press. ISBN 9780996643115.

Chalupka, K., Eberhardt, F., and Perona, P. Multi-Level Cause-Effect Systems. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 361–369, Cadiz, Spain, 2016b. PMLR.

Comon, P. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, April 1994.

Crane, K. Keenan's 3D Model Repository. `https://www.cs.cmu.edu/~kmcrane/Projects/ModelRepository/`, 2021.

Curless, B. and Levoy, M. A Volumetric Method for Building Complex Models from Range Images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pp. 303–312, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237269.

Dean, T. and Kanazawa, K. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(2): 142–150, 1989. doi: https://doi.org/10.1111/j.1467-8640. 1989.tb00324.x.

Didelez, V., Dawid, A. P., and Geneletti, S. Direct and Indirect Effects of Sequential Treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, Arlington, Virginia, USA, 2006. AUAI Press.

Eberhardt, F. Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University*, pp. 93, 2007.

Falcon, W. and The PyTorch Lightning team. PyTorch Lightning, 2019.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pp. 881–889. PMLR, 2015.

Gresele, L., von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. Independent mechanism analysis, a new concept? In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Hendrycks, D. and Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2722–2730. PMLR, 09–15 Jun 2019.

Hoel, E. P., Albantakis, L., and Tononi, G. Quantifying causal emergence shows that macro can beat micro. *Proc. Natl. Acad. Sci. U. S. A.*, 110(49):19790–19795, December 2013.

Höltgen, B. Encoding Causal Macrovariables. *arXiv preprint arXiv:2111.14724*, November 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*, NIPS'16, pp. 3772–3780, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Hyvarinen, A. and Morioka, H. Nonlinear ICA of Temporally Dependent Stationary Sources. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469. PMLR, 20–22 Apr 2017.

Hyvärinen, A. and Pajunen, P. Nonlinear Independent Component Analysis: Existence and Uniqueness Results. *Neural Netw.*, 12(3):429–439, apr 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00140-3.

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 Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019.

Ioffe, S. and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR.

Jang, E., Gu, S., and Poole, B. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.*, 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620.

Jaynes, E. T. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968. doi: 10.1109/TSSC.1968.300117.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020a.

Khemakhem, I., Monti, R., Kingma, D., and Hyvarinen, A. ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12768–12778. Curran Associates, Inc., 2020b.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kingma, D. P. and Dhariwal, P. Glow: Generative Flow with Invertible 1x1 Convolutions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In Bengio, Y. and LeCun, Y. (eds.), *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.

Krishnamurthy, V. and Levoy, M. Fitting Smooth Surfaces to Dense Polygon Meshes. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pp. 313–324, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237270.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

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*, 2022.

Lippe, P., Cohen, T., and Gavves, E. Efficient Neural Causal Discovery without Acyclicity Constraints. In *International Conference on Learning Representations*, 2022.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings*

of Machine Learning Research, pp. 4114–4124. PMLR, 09–15 Jun 2019.

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, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 2020a.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling Factors of Variations Using Few Labels. In *International Conference on Learning Representations*, 2020b.

Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., and State, G. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning, 2021.

Monti, R. P., Zhang, K., and Hyvärinen, A. Causal Discovery with General Non-Linear Relationships using Non-Linear ICA. In Globerson, A. and Silva, R. (eds.), *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*, pp. 186–195. AUAI Press, 2019.

Mooij, J. M., Magliacane, S., and Claassen, T. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

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

Newell, M. E. *The Utilization of Procedure Models in Digital Image Synthesis.* PhD thesis, The University of Utah, 1975. AAI7529894.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Praun, E., Finkelstein, A., and Hoppe, H. Lapped Textures. In *Proceedings of ACM SIGGRAPH 2000*, pp. 465–470, July 2000.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Reed, S., Sohn, K., Zhang, Y., and Lee, H. Learning to Disentangle Factors of Variation with Manifold Interaction. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1431–1439, Bejing, China, 22–24 Jun 2014. PMLR.

Rezende, D. J. and Mohamed, S. Variational Inference with Normalizing Flows. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1530–1538. JMLR.org, 2015.

Rusinkiewicz, S., DeCarlo, D., Finkelstein, A., and Santella, A. Suggestive Contour Gallery. https://gfx.cs.princeton.edu/proj/sugcon/models/, 2021.

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): 612–634, 2021.

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

Spearman, C. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:88–103, 1904.

Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D. S., Maksymets, O., Gokaslan, A., Vondruš, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., and Batra, D. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 251–266. Curran Associates, Inc., 2021.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., and Bauer, S. On Disentangled Representations Learned from Correlated Data. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10401–10412. PMLR, 18–24 Jul 2021.

Turk, G. and Levoy, M. Zippered Polygon Meshes from Range Images. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, pp. 311–318, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916670. doi: 10.1145/192161.192241.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Wright, S. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. CausalVAE: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

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*, 2022.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive Learning Inverts the Data Generating Process. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 2021.

Appendix

TABLE OF CONTENTS

A. Reproducibility Statement

To ensure reproducibility, we publish the code for all models used in this paper at `https://github.com/phlippe/CITRIS`. Further, we include the code for generating the Interventional Pong dataset and the sequences of the causal factors in the Temporal-Causal3DIdent dataset. The complete datasets of this paper are released with corresponding licenses, and links to those datasets are available in the code repository. Moreover, we give a detailed overview of the datasets and more visual examples in Appendix C.1 and Appendix C.2.

Further, for all experiments of Section 6, we have included a detailed overview of the hyperparameters in C.4 and additional implementation details of the evaluation metrics and model architecture components in Appendix C.3. All experiments have been repeated with 3 seeds to obtain stable, reproducible results. We provide an overview of the standard deviations, as well as additional results in Appendix D.

Finally, all experiments in this paper were performed on a single NVIDIA TitanRTX GPU with a 6-core CPU. The overall computation time of all experiments together in this paper correspond to approximately 80 GPU days (excluding hyperparameter search and trials during the research).

B. Proofs

The following section contains the proof for Theorem 3.3. We first give an overview of the notation and additional preliminary discussions in Appendix B.1. Then, we give an outline of the proof. The remaining sections provide the details of the proof.

Additionally, we provide the proof and more details on Proposition 3.1 in Appendix B.6.

B.1. Preliminaries

B.1.1. SUMMARY OF NOTATION

We summarize the notation, which is the same as used for the main paper, as follows:

- We assume K causal factors $C_1, \ldots, C_K$ such that $C_i \in \mathcal{D}_i^{M_i}$ with $M_i \geq 1$;
- We can group all causal factors in a single variable $C = (C_1, \ldots, C_K) \in \mathcal{C}$, where $\mathcal{C}$ is the causal factor space $\mathcal{C} = \mathcal{D}_1^{M_1} \times \mathcal{D}_2^{M_2} \times \ldots \times \mathcal{D}_K^{M_K}$;
- The data is generated by a latent Dynamic Bayesian network with variables $(C_1^t, C_2^t, ..., C_K^t)_{t=1}^T$;
- We assume to know at each time step the binary intervention vector $I^t \in \{0,1\}^{K+1}$ where $I_i^t = 1$ refers to an intervention on the causal factor C_i^{t+1}. As a special case $I_0^t = 0$ for all t;
- For each causal factor C_i, there exists a split $s_i^{\mathrm{var}}(C_i), s_i^{\mathrm{inv}}(C_i)$ such that $s_i^{\mathrm{var}}(C_i)$ represents the variable/manipulable part of C_i, while $s_i^{\mathrm{inv}}(C_i)$ represents the invariable part of C_i;
- The minimal causal split is defined as the one which only contains the intervention-dependent information in $s_i^{\mathrm{var}}(C_i)$, and everything else in $s_i^{\mathrm{inv}}(C_i)$. This split is denoted by $s_i^{\mathrm{var}^*}(C_i)$ and $s_i^{\mathrm{inv}^*}(C_i)$
- At each timestep we can access observations $x^t, x^{t+1} \in \mathcal{X} \subseteq \mathbb{R}^N$;
- $\mathcal{E}$ is the space of the noise variables that affect the observation without changing the encoding of the causal factors. For example, this could be random color shifts in Pong, or brightness shifts in Causal3D, since no causal factor is encoded in color and brightness in these setups respectively;
- Observation function $h : \mathcal{C} \times \mathcal{E} \to \mathcal{X}$, where $\mathcal{E}$ is the space of the noise variables;
- Latent vector $z^t \in \mathcal{Z} \subseteq \mathbb{R}^M$, where $\mathcal{Z}$ is the latent space of dimension $M \geq \dim(\mathcal{E}) + \dim(\mathcal{C})$;
- Inverse of the observation function in the latent space $g_\theta : \mathcal{X} \to \mathcal{Z}$;
- Assignment function from latent dimensions to causal factors $\psi : [\![1..M]\!] \to [\![0..K]\!]$;
- Disentanglement function $\delta^* : \mathcal{X} \to \tilde{\mathcal{C}} \times \tilde{\mathcal{E}}$ with $\tilde{\mathcal{C}} = \mathcal{D}^{\tilde{M}_1} \times \ldots \times \mathcal{D}^{\tilde{M}_K}$ and $\tilde{M}_i$ being the number of latent dimensions assigned to causal factor C_i by ψ^*. We denote the output of δ^* for an observation X as $\delta^*(X) = (\tilde{C}_1, \tilde{C}_2, ..., \tilde{E})$. Then, δ^* is a disentanglement function if there exist a set of deterministic functions $h_0, h_1, ..., h_K$ for which, for any $X = h(C, E)$, $h_i(\tilde{C}_i) = C_i$ for all $i \in [\![1..K]\!]$, and $h_0(\tilde{E}) = E$.
- The representation of δ^* in terms of the learnable function is denoted by g_θ^* and ψ^*;
- Latent variables assigned to each causal factor C_i by ψ are denoted as $z_{\psi_i} = \{z_j | j \in [\![1..M]\!], \psi(j) = i\} = \{g_\theta(x^t)_j | j \in [\![1..M]\!], \psi(j) = i\}$;
- The remaining latent variables that are not assigned to any causal factor are denoted as z_{ψ_0};
- The goal is to learn for each C_i: $p_\phi\left(z_{\psi_i}^{t+1} | z^t, I_i^{t+1}\right) \approx p\left(s_i^{\mathrm{var}}(C_i^{t+1}) | C^t, I_i^{t+1}\right)$;

B.1.2. LIMITING DENSITY OF DISCRETE POINTS

In this section, we give a short overview on the difference between differential entropy and the limiting density of discrete points approach, introduced by Jaynes (1957; 1968). Differential entropy on a continuous random variable X with a distribution $p(X)$ is defined as:

$$H(X) = - \int p(X) \log p(X) dx \tag{5}$$

While for discrete variables, entropy has the intuitive explanation of an uncertainty measure, or the 'information' of a variable, one cannot draw the same relation so easily for continuous variables. This is because differential entropy lacks properties that would be necessary for that. For one, the entropy can become negative. Secondly, and most importantly for the use case in this paper, it is not invariant under invertible transformations, *i.e.* a change of variables. For the example of the random variable X, the entropy of $H(X)$ does not necessarily equal to $H(aX)$ where a is a constant factor, *e.g.* $a = 2$. Thus, it becomes difficult to use differential entropy as a measure of information content of a continuous variable, like in the discrete case.

One approach that was proposed to overcome these issues is the limiting density of discrete points (LDDP) Jaynes (1957; 1968). It adjusts the definition of differential entropy by introducing an *invariant measure* $m(X)$, which can be seen as a reference distribution we measure the entropy of $p(X)$ to. Intuitively, the LDDP adjustment is derived from arguing that the continuous entropy should be derived by taking the limit of increasingly dense discrete distributions. In the limit of infinitely many discrete points, one arrives at the entropy for continuous functions, which becomes:

$$H(X) = - \int p(X) \log \frac{p(X)}{m(X)} dx \tag{6}$$

Note that in some formulations, a constant $\log N$ is added to this equation, where N is the number of discrete points considered which goes against infinity in the limit. Since for this paper, we only require to compare two entropy values with each other and do not require the entropy to take a specific value, we can neglect this constant.

One crucial property of LDDP, which we use in the following proof, is that the entropy stays invariant under a change of variable. This is achieved by transforming the invariant measure $m(X)$ by the exact same invertible transformation as done for $p(X)$. Therefore, when coming back to the example of scaling X by a constant factor, both $p(X)$ and $m(X)$ change in the same way, resulting in $H(X) = H(aX)$.

B.2. Proof Outline

The goal of this section is to proof Theorem 3.3: the global optimum of CITRIS will find the minimal causal variables. We will take the following steps in the proof:

1. (Appendix B.3) Firstly, we show that the function δ^* that disentangles the true latent variables $C_1, ..., C_K$ and assigns them to the corresponding sets $z_{\psi_1}, ..., z_{\psi_K}$ constitutes a global, but not necessarily unique, optimum for maximizing the likelihood of Equation (2).

2. (Appendix B.4) Next, we characterize the class of disentanglement functions Δ^* which all represent a global maximum of the likelihood, *i.e.* get the same score as the true disentanglement. In particular, we show that in all optimal disentanglement functions, each assignment set z_{ψ_i} contains the variable part of the causal factor $s_i^{\text{var}}(C_i)$, but that it might contain also the invariable parts of any other causal factor, thus creating multiple optimal solutions. We do this in two sub-steps:

   (a) First, we assume that all intervention targets are independent, *i.e.* $I_i^{t+1} \perp\!\!\!\perp I_j^{t+1} | C^t$ for any $i \neq j$.
   (b) Secondly, we extend it to a wider group of intervention settings where interventions might be confounded, and show that all of them fall in the same class Δ^*.

3. (Appendix B.5) Finally, we derive Theorem 3.3 by showing that the function $\hat{\delta} \in \Delta^*$, which maximizes the entropy of z_{ψ_0}, identifies the minimal causal mechanisms, which intuitively represent the parts of the causal factors that are affected by the available interventions.

Along the way, we will make use of Figure 8 summarizing the temporal causal graph. For the remainder of the proof, we assume that the prior $p_\phi\left(z^{t+1}|z^t, I^{t+1}\right)$ and the invertible map g_θ are sufficiently complex to approximate any possible function and distribution one might encounter. To simplify the exposition, we also assume that the latent dimension size is unlimited, *i.e.* $M = \infty$, so there are no limitations on how many latent variables z_{ψ_i} can be used to represent a causal factor C_i. In practice, however, this is not a limiting factor as long as we can overestimate the dimensions of the causal factors and noise variables.

Throughout the proof, we will use C^t to refer to the set of all causal factors at time step t, *i.e.* $C^t = \{C_1^t, ..., C_K^t\}$. Similarly, we define $I^{t+1} = \{I_1^{t+1}, ..., I_K^{t+1}\}$.

B.3. Step 1: True Disentanglement δ^* is one of the Global Maxima of the Conditional Likelihood $p(X^{t+1}|X^t, I^{t+1})$

We start by proving the following Lemma:

Lemma B.1. *The true disentanglement function δ^* that correctly disentangles the true causal factors $C_1^{t+1}, ..., C_K^{t+1}$ from observations X^t, X^{t+1} using the true ψ^* assignment function on the true latent variables Z^{t+1} is one of the global maxima of the likelihood of $p(X^{t+1}|X^t, I^{t+1})$.*

We are interested in optimizing $p(X^{t+1}|X^t, I^{t+1})$. We can first try to simplify this equation with the knowledge of the causal graph in Figure 8, *i.e.* using the true underlying generative model, since we aim to show that learning the causal factors and aligning them correspondingly in the prior of Equation (2) represents a global optimum of maximizing $p(X^{t+1}|X^t, I^{t+1})$. Using the conditional independence relations of the graph in Figure 8, we write the joint distribution of all the variables in the true generative model as:

$$p(X^t, X^{t+1}, C^t, C^{t+1}, I^{t+1}) = p(X^{t+1}|C^{t+1}) \cdot \left[\prod_{i=1}^{K} p(C_i^{t+1}|C^t, I_i^{t+1})\right] \cdot p(X^t|C^t) \cdot p(C^t) \cdot p(I^{t+1}) \qquad (7)$$
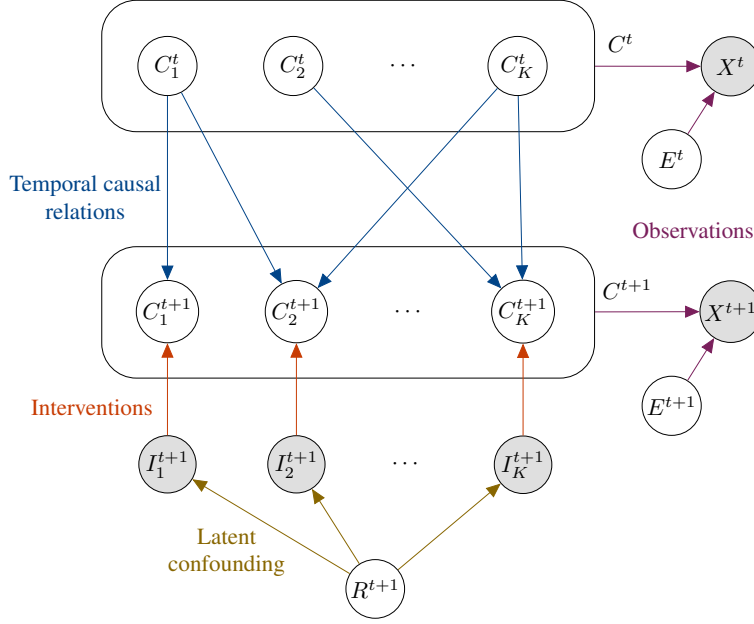
Figure 8. An example temporal causal graph in TRIS, with observed variables shown in gray and latent variables in white. A latent causal factor C_i^{t+1} has as parents a subset of the causal factors at the previous time step $C_1^t, \ldots, C_K^t$, and its intervention target I_i^{t+1}. All causal variables C^{t+1} and the noise E^{t+1} cause the observation X^{t+1}. R^{t+1} is a latent confounder between the intervention targets, but is not strictly necessary (*i.e.* R^{t+1} may also be constant). The goal is to identify the causal factors $C_1, \ldots, C_K$.

We can now condition on X^t and I^{t+1}, marginalize out C^t and C^{t+1} and write the conditional likelihood as:

$$p(X^{t+1}|X^t, I^{t+1}) = \int_{C^{t+1}} \int_{C^t} p(X^{t+1}|C^{t+1}) \cdot \left[\prod_{i=1}^{K} p(C_i^{t+1}|C^t, I_i^{t+1}) \right] \cdot p(C^t|X^t) dC^t dC^{t+1} \tag{8}$$

In our assumptions of Section 3.1, we have defined the observation function h to be bijective, meaning that there exists an inverse f that can identify the causal factors C^t and noise variable E_o^t from X^t. Thus, we can write $p(C^t|X^t) = \delta_{f(X^t)=C^t}$, where δ is a Dirac delta. Since the noise on the observations, E_o^t, is said to be independent of X^{t+1} and C^{t+1}, we can remove it from being in the conditioning set. This leads us to:

$$p(X^{t+1}|X^t, I^{t+1}) = \int_{C^{t+1}} \left[\prod_{i=1}^{K} p(C_i^{t+1}|C^t, I_i^{t+1}) \right] \cdot p(X^{t+1}|C^{t+1}) dC^{t+1} \tag{9}$$

Since we have assumed h to be bijective, we know that for each X^{t+1}, there exist only one combination of C^{t+1} and E_o^{t+1}. Thus, by using the change of variables formula, we can rewrite the equation above by:

$$p(X^{t+1}|X^t, I^{t+1}) = |J_h|^{-1} \cdot \left[\prod_{i=1}^{K} p(C_i^{t+1}|C^t, I_i^{t+1}) \right] \cdot p(E_o^{t+1}) \tag{10}$$

where $J_h = \frac{\partial h(C^{t+1}, E_o^{t+1})}{\partial C^{t+1} \partial E_o^{t+1}}$ denotes the Jacobian of the bijective/invertible observation function h. Equation (10) constitutes a global optimum of the maximum likelihood, since it represents the true underlying dynamics.

We relate this conditional likelihood to the prior setup of CITRIS. We show that assigning C_i^{t+1} to $z_{\psi_i}^{t+1}$, *i.e.*, learning the true assignment function ψ^*, provides us with the same maximum likelihood solution as in Equation (10). We have defined our objective in Section 3 in Equation (2) as:

$$p_\phi \left(z^{t+1}|z^t, I^{t+1} \right) = \prod_{i=0}^{K} p_\phi \left(z_{\psi_i}^{t+1}|z^t, I_i^{t+1} \right) \tag{11}$$

Since we know that g_θ^* is an invertible function between $\mathcal{X}$ and $\mathcal{Z}$, we know that z^t must include all information of X^t. Thus, we can also replace it with $z^t = [C^t, E_o^t]$, giving us:

$$p_\phi\left(z^{t+1}|C^t, E_o^t, I^{t+1}\right) = \prod_{i=0}^{K} p_\phi\left(z_{\psi_i}^{t+1}|C^t, E_o^t, I_i^{t+1}\right) \tag{12}$$

The optimal assignment function ψ^* assigns sufficient dimensions to each causal factor $C_1, ..., C_K$. Since $\mathcal{Z}$ can have a larger space than $\mathcal{E} \times \mathcal{C}$, but $\mathcal{E} \times \mathcal{C}$ is sufficient to describe $\mathcal{X}$, we know that the remaining dimensions of $\mathcal{Z}$ do not contain any information. Thus, the assignment function ψ^* can map them to any causal factor without a change in distribution. Using this assignment function, we now consider $z_{\psi_i^*}^{t+1} = C_i^{t+1}$ for $i = 1, ..., K$. Then, Equation (12) becomes:

$$p_\phi\left(z^{t+1}|C^t, E_o^t, I^{t+1}\right) = \left[\prod_{i=1}^{K} p_\phi\left(z_{\psi_i^*}^{t+1} = C_i^{t+1}|C^t, I_i^{t+1}\right)\right] \cdot p(z_{\psi_0^*}^{t+1}|C^t, E_o^t) \tag{13}$$

where we remove E_o^t from the conditioning set for the causal factors, since know that C^{t+1} and E_o^{t+1} is independent of E_o^t. We further simplify by noting that $z_{\psi_0^*}^{t+1} = E_o^{t+1}$ is independent of any other factor.

$$p_\phi\left(z^{t+1}|C^t, E_o^t, I^{t+1}\right) = \left[\prod_{i=1}^{K} p_\phi\left(z_{\psi_i^*}^{t+1} = C_i^{t+1}|C^t, I_i^{t+1}\right)\right] \cdot p(z_{\psi_0^*}^{t+1} = E_o^{t+1}) \tag{14}$$

Finally, by using g_θ^*, we can replace the distribution on z^{t+1} by a distribution on X^{t+1} by the change of variables formula:

$$p_\phi\left(X^{t+1}|C^t, E_o^t, I^{t+1}\right) = \left|\frac{\partial g_\theta^*(z^{t+1})}{\partial z^{t+1}}\right| \cdot \left[\prod_{i=1}^{K} p_\phi\left(z_{\psi_i^*}^{t+1} = C_i^{t+1}|C^t, I_i^{t+1}\right)\right] \cdot p(z_{\psi_0^*}^{t+1} = E_o^{t+1}) \tag{15}$$

Thereby, it is apparent that g_θ^* is equal to h^{-1}, since both are identical invertible functions between the same spaces ($\mathcal{Z}$ becomes $\mathcal{E} \times \mathcal{C}$ here). Hence, Equation (15) represents the exact same distribution as Equation (10). Therefore, we have shown that the function δ^* that disentangles the true latent variables $C_1, ..., C_K$ and assigns them to the corresponding sets $z_{\psi_1}, ..., z_{\psi_K}$ constitutes a global, but not necessarily unique, optimum for maximizing the likelihood of Equation (2).

Note that while assigning C_i^{t+1} to $z_{\psi_i}^{t+1}$ provides us with the same maximum likelihood solution as in Equation (10), this is not the only possible representation. Additional possible representation will be discussed in Step 2.

B.4. Step 2: Characterizing the Disentanglement Class Δ

Showing that the correct disentanglement constitutes a global optimum is not sufficient for showing that a model trained on solving the maximum likelihood solution converges to it, since there might potentially be multiple global optima. Hence, this section discusses the class of causal representation functions $\delta \in \Delta^*$ which can achieve the same maximum likelihood optimum as the true causal factor disentanglement discussed in Appendix B.3. For this, we first need to distinguish between the *variable* and *invariable* information of a causal variable C_i, which is introduced in Appendix B.4.1. Next, we will discuss the causal representation function class Δ for the setting where interventions are independent, *i.e.* $I_i^{t+1} \perp\!\!\!\perp I_j^{t+1}|C^t$ for any $i \neq j$, and finally extend it to confounded interventions.

B.4.1. INTERVENTION-INDEPENDENT VARIABLES

Interventions allow us to identify a causal variable by seeing the caused change in its conditional distribution. However, especially when talking about multidimensional causal variables, one might have interventions that only affect a subset of the actual causal variable dynamics, while the rest remains independent of the intervention. As we will see later, this can have an influence on the identifiability result, making the found causal factors intervention-dependent.

We start by considering a single causal factor $C_i \in \mathcal{D}_i^{M_i}$ in the setup of Figure 8 under our previously discussed assumptions. Suppose for each causal factor $C_i \in \mathcal{D}^{M_i}$, there exists an invertible map $s_i : \mathcal{D}_i^{M_i} \to \mathcal{D}_i^{\text{var}} \times \mathcal{D}_i^{\text{inv}}$ that splits the domain $\mathcal{D}^{M_i}$ of C_i into a part that is invariant and a part that is variant under intervention. We denote the two parts of this map as

$$s_i(C_i^t) = (s_i^{\text{var}}(C_i^t), s_i^{\text{inv}}(C_i^t)) \tag{16}$$

(a) Original causal graph of C_i
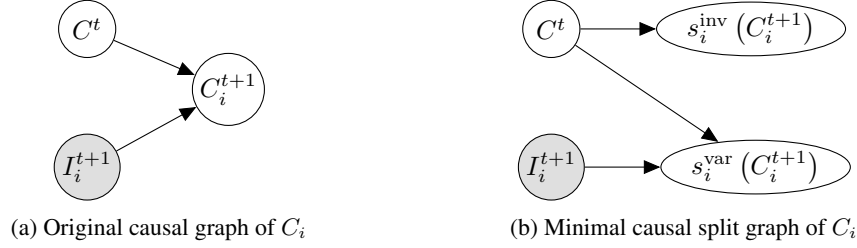
(b) Minimal causal split graph of C_i

Figure 9. Splitting the causal variable C_i in its minimal causal split. (a) In the original causal graph, C_i^{t+1} has C^t (or an eventual subset of it) and I_i^{t+1} as its parents. (b) In the minimal causal split, only the variable part $s_i^{\text{var}}(C_i^{t+1})$ depends on the intervention. The invariable part, $s_i^{\text{inv}}(C_i^{t+1})$, is independent of I_i^{t+1}, hence giving us an additional conditional independence. Note that $s_i^{\text{var}}(C_i^{t+1})$ and $s_i^{\text{inv}}(C_i^{t+1})$ are conditionally independent.
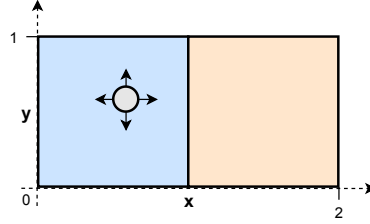

Figure 10. Example for splitting a causal variable into an intervention-dependent and -independent part. The two ground truth causal variables are the x and y positions of the ball. See Appendix B.4.1 for details.

The split s must be invertible, so that we can map back and forth between $\mathcal{D}_i^{M_i}$ and $\mathcal{D}_i^{\text{var}} \times \mathcal{D}_i^{\text{var}}$ without losing information. Furthermore, to be called a split, s must satisfy $s_i^{\text{inv}}(C_i^t) \perp\!\!\!\perp I_i^t \mid \text{pa}(C_i^t)$, *i.e.*, $s_i^{\text{inv}}(C_i^t)$ is independent of the intervention variable I_i^t given the parents of C_i^t. Further, both parts of the split must be conditionally independent, *i.e.* $s_i^{\text{inv}}(C_i^t) \perp\!\!\!\perp s_i^{\text{var}}(C_i^t) \mid \text{pa}(C_i^t), I_i^t$. Hence, we can write their distributions as:

$$p\left(s_i(C_i^{t+1})|C^t, I_i^{t+1}\right) = p\left(s_i^{\text{var}}(C_i^{t+1})|C^t, I_i^{t+1}\right) \cdot p\left(s_i^{\text{inv}}(C_i^{t+1})|C^t\right) \tag{17}$$

This means that $s_i^{\text{var}}(C_i^t)$ will contain the manipulable, or *variable*, part of C_i^t. In contrast, $s_i^{\text{inv}}(C_i^t)$ is the *invariable* part of C_i^t which is independent of the intervention. This relation is visualized in Figure 9.

For any causal variable, there may exist multiple possible splits, but there is always at least the trivial split where $\mathcal{D}_i^{\text{var}} = \mathcal{D}_i^{M_i}$ is the original domain of C_i, and $\mathcal{D}_i^{\text{inv}} = \{0\}$ is the one-element set (no invariant information). However, there might also exist splits in which $s_i^{\text{inv}}(C_i^{t+1}) \neq \emptyset$. For instance, in a multidimensional causal variable $\hat{C} \in \mathbb{R}^3$, if an intervention only affects the first two dimensions while the last one remains unaffected, we obtain the split $s^{\text{var}}([\hat{C}_1, \hat{C}_2]), s^{\text{inv}}(\hat{C}_3)$. Nonetheless, this can even happen for scalar variables, since we do not constrain the possible distributions of C_i. We give an example for such a case below.

Example 1 Consider the scenario in Figure 10 where we have a ball with its two positional dimensions x and y as its causal factors. For now, we only focus on its x position (in the remainder of the section, x^t refers to the position of the ball on the x-axis, not the full observation X^t which we denote by a capital letter). Over time, the ball moves within one of the two boxes, but cannot jump in between boxes. An example of such a conditional could be:

$$p(x^{t+1}|x^t, I_x^{t+1} = 0) = \begin{cases} \min(\max(x^t + \epsilon, 0), 1) & \text{if } x^t < 1 \\ \min(\max(x^t + \epsilon, 1), 2) & \text{otherwise} \end{cases} \tag{18}$$

with $\epsilon \sim \mathcal{N}(0, 0.1)$. Intuitively, the ball therefore moves randomly around its previous position, while being bounded by the box it is in. Due to its modular conditional distribution, we can rewrite the causal variable x and its distribution in terms of two different variables: its position within its current box, $u \in [0, 1]$, and a binary variable indicating in which box the ball is, $b \in 0, 1$ (left/blue vs right/orange in Figure 10). Then, its conditional distribution becomes:

$$p(x^{t+1}|x^t, I_x^t) = p\left(b^{t+1}|x^t, I_x^{t+1}\right) \cdot p\left(u^{t+1}|x^t, I_x^{t+1}\right) \tag{19}$$

Now, suppose that an intervention I_x^{t+1} changes the box the ball is in, while the relative position keeps evolving as it would under no intervention, *i.e.* still depending on its parents. Then, we can yet write its conditional distribution as:

$$p(x^{t+1}|x^t, I_x^t) = p\left(b^{t+1}|x^t, I_x^{t+1}\right) \cdot p\left(u^{t+1}|x^t\right) \tag{20}$$

Using the notation above, we therefore can define the split $s^{\mathrm{var}}(x) = b, s^{\mathrm{inv}}(x) = u$, where b depends on the intervention, while u does not. Note that $s^{\mathrm{var}}(x) = x, s^{\mathrm{inv}}(x) = \emptyset$ is yet another valid split in this case.

Example 2 Consider the same example as before, however, now with a different intervention setup. Suppose that an intervention I_x^{t+1} constitutes a perfect intervention on x, under which $x^{t+1} \perp\!\!\!\perp x^t|I_x^{t+1} = 1$. Then, the previous split $s^{\mathrm{var}}(x) = b, s^{\mathrm{inv}}(x) = u$ is not valid anymore, if the intervention target I_x^{t+1} cannot be deterministically deduced from x^t, since an intervention changes the distribution of the relative position u^{t+1}. Hence, the only valid split is $s^{\mathrm{var}}(x) = x, s^{\mathrm{inv}}(x) = \emptyset$. This shows that the possible space of such splits depends on the available interventions.

Minimal causal variables In the examples above, one can see that for certain situations, a causal variable can have multiple valid splits $s_i^{\mathrm{var}}(C_i), s_i^{\mathrm{inv}}(C_i)$ since intervention-independent information can be modeled in either $s_i^{\mathrm{var}}(C_i)$ or $s_i^{\mathrm{inv}}(C_i)$. The split that will be the most relevant for the identifiability discussion here is the one that assigns only the intervention-dependent information to $s_i^{\mathrm{var}}(C_i)$, and the rest to $s_i^{\mathrm{inv}}(C_i)$. We define this as follows:

Definition B.2. *The* minimal causal split *of a variable C_i^t with respect to its intervention variable I_i^t is the split s_i which maximizes the entropy of $H(s_i^{\mathrm{inv}}(C_i^t)|pa(C_i^t))$. Under this split, $s_i^{\mathrm{var}}(C_i^t)$ is defined as the* minimal causal variable *and denoted by $s_i^{\mathrm{var}^*}(C_i^t)$.*

Additionally, we also define:

Definition B.3. *The <u>minimal causal mechanism</u> of a variable C_i with respect to its intervention I_i is defined as the conditional distribution $p\left(s_i^{\mathrm{var}^*}(C_i^t)|pa(C_i^{t+1}), I_i^{t+1}\right)$.*

We refer to $p\left(s_i^{\mathrm{var}^*}(C_i^t)|C^t, I_i^{t+1}\right)$ as *minimal* causal mechanism, since it is the distribution for which as little as possible information depends on I_i^{t+1}. Hence, the definition of this mechanism depends on the characteristics of the provided intervention. As we will see later, while we cannot guarantee to find the full causal mechanism, we can yet identify the minimal causal mechanism.

The existence of a split where $s^{\mathrm{inv}}(C_i) \neq \emptyset$ for any causal factor C_i creates additional, possible solutions that obtain the same maximum likelihood as the true split. This is because $s^{\mathrm{inv}}(C_i)$ is independent of the intervention target I_i, allowing it to be modeled by any set z_{ψ_j} without losing information. The following subsections further characterize the space of new solutions with the existence of such splits.

B.4.2. INDEPENDENT INTERVENTIONS

In this section, we show by which class of disentanglement functions Δ a maximum likelihood solution of the generative model can be found.

However, to simplify the first steps, we assume that all intervention targets are independent of each other given the causal factors of the previous time step, *i.e.* $I_i^{t+1} \perp\!\!\!\perp I_j^{t+1}|C^t$ for any $i \neq j \in 1, \ldots, K$. By construction, we also assume $I_0^{t+1} = 0$ for all t. We will extend it afterwards in Appendix B.4.3.

Solutions for $s_i^{\mathrm{inv}}(C_i) = \emptyset$ for all $i \in [\![1..K]\!]$ As a first step, we assume that for all causal factors $C_1, ..., C_K$, there does not exist any minimal causal mechanism split besides $s_i^{\mathrm{inv}}(C_i)$ being the empty set. Therefore, all of C_i is dependent on I_i^{t+1}. For this case, consider an arbitrary partition of C_i, $s^0(C_i), s^1(C_i)$, with the same invertibility constraints as $s^{\mathrm{var}}, s^{\mathrm{inv}}$ and conditionally independence between $s^0(C_i), s^1(C_i)$, but with a non-empty invariable part, *i.e.*, $s^1(C_i) \neq \emptyset$. For this partition, the conditional entropy of $s^1(C_i)$ given C^t must be strictly lower when conditioning also on I_i^{t+1}:

$$H\left(s^1(C_i^{t+1})|C^t, I_i^{t+1}\right) < H\left(s^1(C_i^{t+1})|C^t\right) \tag{21}$$

If the conditional entropy was equal, then $s^1(C_i^{t+1})$ would be independent of I_i^{t+1} given C^t, which is only true for $s_i^{\mathrm{inv}}(C_i)$. Since we assume $s_i^{\mathrm{inv}}(C_i)$ is empty, while $s^1(C_i^{t+1})$ is not, this can never happen. Thus, to model a causal factor C_i where $s_i^{\mathrm{inv}}(C_i) = \emptyset$, a maximum likelihood solution can only be found if all of $s_i^{\mathrm{var}}(C_i)$ is conditioned on I^{t+1}. Further, since in this setting $C_j^{t+1} \perp\!\!\!\perp C_i^{t+1}|C^t, I^{t+1}$ for all $i, j \in [\![1..K]\!], i \neq j$, there cannot exist any split across multiple causal factors that violate the entropy inequality above for I_i^{t+1} and I_j^{t+1} while still modeling the true conditional distributions.

Similarly, in this setting under the specified assumptions, the information of I_i^{t+1} cannot be determined by any other target variable $I_j^{t+1}, i \neq j$, since otherwise, the targets would not be independent. Hence, we can write the following entropy inequality for any $i \neq j$:

$$H\left(C_i^{t+1}|C^t, I_i^{t+1}\right) < H\left(C_i^{t+1}|C^t\right) = H\left(C_i^{t+1}|C^t, I_j^{t+1}\right) \tag{22}$$

Therefore, one can only achieve the maximum likelihood solution (*i.e.*, the minimum entropy solution) if all information of C_i^{t+1} is conditioned on I_i^{t+1} for $i = 1, ..., K$ in addition to C^t. This implies that each factor $p_\phi(z_{\psi_i}^{t+1}|z^t, I_i^{t+1})$ described in Equation (2) will have $p(C_i^{t+1}|z^t, I_i^{t+1})$, and therefore $z_{\psi_i}^{t+1} = C_i^{t+1}$ as its maximum likelihood solution.

Nonetheless, this excludes z_{ψ_0}, *i.e.* the factors independent of any intervention, in this case being the noise E_o^{t+1}. Since it is independent of any interventions, any distribution of E_o^{t+1} across the different causal factor sets $z_{\psi_0}, ..., z_{\psi_K}$ will achieve the same likelihood score, as long as the part of E_o^{t+1} across factors is independent. Hence, in conclusion for this scenario, we can guarantee that z_{ψ_i} will model all information of C_i and no other causal factor $C_j, i \neq j$, but can contain additional information from E^{t+1}.

Solutions with invariable parts Next, we consider the scenario where there exists a split with $s_i^{\text{inv}}(C_i) \neq \emptyset$ for some causal variables in $i \in [\![1..K]\!]$. For this case, we can write the maximum likelihood solution of Equation (10) as:

$$p(X^{t+1}|X^t, I^{t+1}) = \left|\frac{\partial g_\theta^*(z^{t+1})}{\partial z^{t+1}}\right| \cdot \left[\prod_{i=1}^K p(C_i^{t+1}|I_i^{t+1}, C^t)\right] \cdot p(E^{t+1}) \tag{23}$$

$$= \left|\frac{\partial g_\theta^*(z^{t+1})}{\partial z^{t+1}}\right| \cdot \left[\prod_{i=1}^K p(s_i^{\text{var}}(C_i^{t+1})|I_i^{t+1}, C^t)\right] \cdot \left[\prod_{i=1}^K p(s_i^{\text{inv}}(C_i^{t+1})|C^t)\right] \cdot p(E^{t+1}) \tag{24}$$

This equation shows that one can assign $s_1^{\text{inv}}(C_1), ..., s_K^{\text{inv}}(C_K)$ to any latent variable set $z_{\psi_0}, ..., z_{\psi_K}$ or split it across, while achieving the same optimal likelihood, since they are independent of any intervention target. The remaining information in $s_1^{\text{inv}}(C_1), ..., s_K^{\text{inv}}(C_K)$ thereby acts the same way as the noise variable E^{t+1}. Thus, there exist multiple maximum likelihood solutions with different splits of information to causal factors.

However, on the other hand, the solution space is yet restricted by the assignment of $s_i^{\text{var}}(C_i)$. In particular, if $s_i^{\text{var}}(C_i)$ cannot be split further into an invariable/intervention-independent part, we can rely on the same results from the previous setting, when considering $s_i^{\text{var}}(C_i)$ as new causal variables. In case there exist another split of C_i which would add more information to $s_i^{\text{inv}}(C_i)$, this part could not be guaranteed to be matched to the causal factor C_i due to its independence. Hence, in conclusion here, we can guarantee that z_{ψ_i} will model all information of $s^{\text{var}}(C_i)$ and no other causal factor $s^{\text{var}}(C_j), i \neq j$, if there does not exist another split of $s^{\text{var}}(C_i)$. The additional information of E^{t+1} as well as $s^{\text{inv}}(C_j)$ can be assigned to any causal variable. In the third step of the proof (Appendix B.5), we discuss how one can yet obtain a unique solution.

B.4.3. CONFOUNDED INTERVENTIONS

In the previous discussion, we have used the assumption that interventions are independent of each other: $I_i^{t+1} \perp\!\!\!\perp I_j^{t+1}|C^t$. This assumption was required for showing that conditioning information of C^{t+1} on any other target will lead to the same entropy as having it without a target, *i.e.* $H\left(C_i^{t+1}|C^t, I_j^{t+1}\right) = H\left(C_i^{t+1}|C^t\right)$. In this section, however, we consider a wider range of interventions. Specifically, we assume that the intervention targets $I_1^{t+1}, ..., I_K^{t+1}$ are confounded by some unobserved variable R^{t+1} besides C^t. This allows the modeling of, for example, single-target interventions or groups of interventions, *e.g.* $I^{t+1} \in \{[0, 0, 0], [1, 1, 0], [0, 1, 1]\}$ for a three-variable case. Under such a setup, the entropy equation from before, *i.e.* $H\left(C_i^{t+1}|C^t, I_j^{t+1}\right) \neq H\left(C_i^{t+1}|C^t\right)$, is not valid anymore since I_j^{t+1} and I_i^{t+1} are not necessarily independent anymore and hence $C_i^{t+1} \not\perp\!\!\!\perp I_j^{t+1}$ can occur for some $i, j \in [\![1..K]\!], i \neq j$.

Despite that, a causal factor C_i^{t+1} is still independent of any other target $I_j^{t+1}, i \neq j$, as long as it is conditioned on its true target and previous time step: $C_i^{t+1} \perp\!\!\!\perp I_j^{t+1}|C^t, I_i^{t+1}$. This is because C^t and I_i^{t+1} are all the parents of C_i^{t+1}, as shown in the causal graph of Figure 8. Further, suppose that there exist information of C_i^{t+1} which is statistically independent of the intervention I_i^{t+1}, *i.e.* $s_i^{\text{inv}}(C_i) \neq \emptyset$. Then, this will also be independent of any other intervention target I_j^{t+1}, since $s_i^{\text{inv}}(C_i) \perp\!\!\!\perp I_j^{t+1}|C^t$, and all paths from C_i to I_j^{t+1} include I_i^{t+1}. Hence, our discussion of the intervention-independent parts follow the same logic as in Appendix B.4.2, and we are left with showing that $s_i^{\text{var}}(C_i)$ is modeled by z_{ψ_i} in any maximum likelihood solution.

For this, we consider a pair of variables C_i, C_j, for which $I_i^{t+1} \not\perp\!\!\!\perp I_j^{t+1}|C^t$, and show under which circumstances we can

guarantee that no information of $s_i^{\mathrm{var}}(C_i)$ will be modeled in z_{ψ_j}. It is sufficient to limit the discussion to pairs of variables, since one latent variable can be only assigned to a single causal variable, hence to the actual one it belongs to, C_i, or any other variable C_j here. A crucial insight to the discussion will be that the influence of I_j^{t+1} to C_i^{t+1} solely relies on I_i^{t+1} being correlated to both variables. Further, one requirement is that no additional conditional independence relations exist, such as $s_i^{\mathrm{var}}(C_i) \perp\!\!\!\perp I_i^{t+1}|I_j^{t+1}, C^t$, which is covered by our faithfulness assumption. Now, under this setup, we consider three cases:

1. for every time step t, the two variables C_i, C_j have always been intervened on together, *i.e.* $I_i^{t+1} = I_j^{t+1}$ for any t;

2. there exist a time step t at which $I_i^{t+1} = 0, I_j^{t+1} = 1$;

3. there exist a time step t at which $I_i^{t+1} = 1, I_j^{t+1} = 0$.

Note that the only excluded case is when for every time step t, $I_i^{t+1} = 0, I_j^{t+1} = 0$. This case refers to not having observed interventions for any of the two variables, and goes back to Appendix B.4.2, where the variable part of C_i is empty, *i.e.* $s_i^{\mathrm{var}}(C_i) = \emptyset$. Hence, in that case, $s_i^{\mathrm{var}}(C_i)$ would have no influence on the modeled solution.

In the first case, since the two factors have always been intervened on together, we know that $I_i^{t+1} = I_j^{t+1}$ for any time step t. Hence, one can assign the information of $s_i^{\mathrm{var}}(C_i^{t+1})$, $s_j^{\mathrm{var}}(C_j^{t+1})$, or the union of both to either intervention target I_i^{t+1} or I_j^{t+1}, without losing any information. Moreover, if $s_j^{\mathrm{var}}(C_i^{t+1})$ has multiple independent dimensions, *i.e.* can be written as a product of multiple, conditionally independent variables, one can even split information of C_i over the two targets. This shows that in the general case, we cannot disentangle between two variables which have always been intervened on together. Similarly, if more than 2 variables have always been intervened on together, we cannot disentangle among all those variables.

For the second case, we can deduce that there must be interventions provided for at least the observational case, *i.e.* $I_i^{t+1} = 0, I_j^{t+1} = 0$, the case where C_j is intervened on but not C_i, *i.e.* $I_i^{t+1} = 0, I_j^{t+1} = 1$, and either the joint intervention on both C_i, C_j or only interventions on C_i, not C_j. The reason why one of the two latter cases needs to exist is that if it would not be the case, $I_i^{t+1} = 0$ would be zero for any t. In that case, the minimal causal mechanism of C_i uses $s_i^{\mathrm{var}}(C_i) = \emptyset$, hence making the modeling of $s_i^{\mathrm{var}}(C_i)$ irrelevant for the maximum likelihood solution.

Thus, from these different intervention settings, it is apparent that there cannot exist a deterministic function f with which we can determine I_i^{t+1} from seeing I_j^{t+1}. If we observe joint interventions on both variables, then for $I_j^{t+1} = 1$, both $I_i^{t+1} = 0$ and $I_i^{t+1} = 1$ can occur. Similarly, if we observe interventions on C_i when C_j is not intervened on, then both $I_i^{t+1} = 0$ and $I_i^{t+1} = 1$ can occur for $I_j^{t+1} = 0$. If both joint interventions and single interventions on C_i have been observed, we cannot determine I_i^{t+1} from I_j^{t+1} at either $I_j^{t+1} = 0$ or $I_j^{t+1} = 1$. Since $s_i^{\mathrm{var}}(C_i^{t+1}) \perp\!\!\!\perp I_j^{t+1}|C^t, I_i^{t+1}$ by definition and $s_i^{\mathrm{var}}(C_i^{t+1}) \not\perp\!\!\!\perp I_i^{t+1}|C^t, I_j^{t+1}$ (the latter because I_i^{t+1} is not a deterministic function of I_j^{t+1}, and therefore the dependence holds), we can write:

$$H\left(s_i^{\mathrm{var}}(C_i^{t+1})|C^t, I_i^{t+1}\right) = H\left(s_i^{\mathrm{var}}(C_i^{t+1})|C^t, I_i^{t+1}, I_j^{t+1}\right) < H\left(s_i^{\mathrm{var}}(C_i^{t+1})|C^t, I_j^{t+1}\right) \tag{25}$$

In conclusion, we cannot find the maximum likelihood solution if any information of C_i^{t+1}, which depends on I_i^{t+1}, is assigned to latent variables z_{ψ_j}. Hence, the maximum likelihood solution will strictly model $s_i^{\mathrm{var}}(C_i)$ in z_{ψ_i}.

Finally, in the third case, we can take a similar argument as for the second case. The only difference is that any of the additional intervention cases (joint or single on C_j), we have that from $I_j^{t+1} = 0$, both $I_i^{t+1} = 0$ and $I_i^{t+1} = 1$ can occur. Hence, the inequality in Equation (25) is still valid, and we cannot replace I_i^{t+1} by I_j^{t+1} for any subset of information of $s_i^{\mathrm{var}}(C_i)$. In summary, the maximum likelihood solution will strictly model $s_i^{\mathrm{var}}(C_i)$ in z_{ψ_i} also for this case.

Therefore, we can summarize the results in the following statement. We can disentangle the intervention-dependent part of any two variables C_i, C_j, if there does not exist a deterministic function f for which $I_i^t = f(I_j^t)$ holds for every time step t.

B.5. Step 3: Deriving the Final Theorem

Now that we have discussed the class of disentanglement functions Δ^* with their corresponding solutions, we can take the final step by adding constraints that ensure a unique solution. In all the settings discussed in Appendix B.4.2 and Appendix B.4.3, the problem is that intervention-independent information can be represented in any of the latent sets $z_{\psi_0}, ..., z_{\psi_K}$ without affecting the optimal likelihood. However, our main goal in getting a causal representation is that we

disentangle information from different causal factors, meaning that we want to guarantee that the latents of z_{ψ_i} will only model information of the causal factor C_i, and no other causal factor $C_j, i \neq j$. Thus, we can do this by collecting all intervention-independent information in z_{ψ_0}. In other words, our ideal solution would be to have the latents of z_{ψ_i} model $s_i^{\text{var}}(C_i)$ $(i = 1, ..., K)$, and z_{ψ_0} to model $\{s_1^{\text{inv}}(C_1), ..., s_K^{\text{inv}}(C_K)\}$, where the split $s_i^{\text{var}}, s_i^{\text{inv}}$ was chosen to maximize the entropy of $p(s_i^{\text{inv}}(C_i^{t+1})|C^t)$. To find both the right splits and collecting all intervention-independent information in z_{ψ_0}, we thus want to find the representation function $\hat{\delta} \in \Delta^*$ which maximizes the entropy of z_{ψ_0} while maintaining the optimal likelihood. If any intervention-independent information would not be modeled in z_{ψ_0}, it implies that there must exist another solution with greater entropy in z_{ψ_0}, since all $s_1^{\text{inv}}(C_1), ..., s_K^{\text{inv}}(C_K)$ as well as $s_1^{\text{var}}(C_1), ..., s_K^{\text{var}}(C_K)$ are conditionally independent of each other (*i.e.* adding parts to z_{ψ_0} cannot reduce the entropy). Further, since we try to maximize the entropy of z_{ψ_0}, we find the information splits $s_i^{\text{var}}, s_i^{\text{inv}}$ that maximize the entropy of its intervention-independent part. This is the same split as we had defined as minimal causal mechanisms in Appendix B.4.1. Thus, we can summarize this result as follows:

Theorem B.4. *Suppose that ϕ^*, θ^* and ψ^* are the parameters that, under the constraint of maximizing the likelihood $p_\phi(g_\theta(x^{t+1})|g_\theta(x^t), I^{t+1})$, maximize the entropy of $p_\phi(z_{\psi_0}^{t+1}|z^t, I^{t+1})$. Then, with sufficient latent dimensions, the model ϕ^*, θ^*, ψ^* learns a latent structure where $z_{\psi_i}^{t+1}$ models the minimal causal variable of C_i if $C_i^{t+1} \not\perp\!\!\!\perp I_i^{t+1}|C^t, I_j^{t+1}$ for any $i \neq j$. All remaining information independent of any interventions is modeled in z_{ψ_0}.*

The conditional independence $C_i^{t+1} \not\perp\!\!\!\perp I_i^{t+1}|C^t, I_j^{t+1}$ ensures that there exists no deterministic function f for which $I_i^{t+1} = f(I_j^{t+1})$. This also includes when I_i^{t+1} is constant, *i.e.*, when C_i is intervened all the time or not at all, since then, C_i becomes independent of I_i^{t+1}.

B.6. Non-Identifiability without Interventions

This section discusses Proposition 3.1 and its accompanying proof. Proposition 3.1 states that in TRIS, if two causal factors C_i and C_j have only been jointly intervened on or not at all, then there exists a causal graph in which C_i and C_j cannot be uniquely identified from observations X and intervention targets I. In other words, without interventions, it cannot be guaranteed that for all possible settings, the causal variables can be uniquely identified. A similar situation occurs in (non-linear) independent component analysis (Hyvärinen et al., 2001; 2019), where the goal is to disentangle and identify N independent sources/signals from observations in which the signals are entangled. In Hyvärinen et al. (2001, Section 7.5), it is shown that if we have two, independent variables s_1, s_2 that follow a standard Gaussian distribution $\mathcal{N}(0, 1)$, then any mixing of the two variables with an orthogonal matrix $\boldsymbol{A}$ results in two variables, x_1, x_2 with $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$, which yet again follow a standard Gaussian distribution. Thus, from the perspective of the probability density function, the two representations cannot be distinguished, and hence no unique solution can be found. This also implies that simply from samples of this distribution, one cannot uniquely identify the true sources.

In TRIS, we do not have two independent variables, but instead variables that have dependencies over time. This corresponds to the setting described in Hyvärinen et al. (2001, Section 18), Belouchrani et al. (1997), or Hyvarinen & Morioka (2017). Suppose that we have two variables s_1^t, s_2^t at a time step t, which follow a Gaussian distribution with constant variance over time:

$$s_1^t \sim \mathcal{N}(\mu_1(s_1^{t-1}, s_2^{t-1}), \sigma^2), x_2^t \sim \mathcal{N}(\mu_2(s_1^{t-1}, s_2^{t-1}), \sigma^2) \tag{26}$$

where μ_1, μ_2 are arbitrary functions on the variables of the previous time step, s_1^{t-1}, s_2^{t-1}. The probability density function of it is:

$$p(s_1^t, s_2^t|\boldsymbol{s}^{t-1}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{s}^t - \boldsymbol{\mu}^t)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{s}^t - \boldsymbol{\mu}^t)\right) \tag{27}$$

where $\boldsymbol{\mu}^t$ represents the mean based on $\boldsymbol{s}^{t-1}$, and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$. Now, consider a mixing of the sources by an orthogonal

matrix $\boldsymbol{A}$, i.e. $\boldsymbol{x}^t = \boldsymbol{A}\boldsymbol{s}^t$. Then, the probability density function of those are:

$$p(x_1^t, x_2^t | \boldsymbol{s}^{t-1}) = \frac{1}{\sqrt{\det(2\pi \boldsymbol{A}^T \boldsymbol{\Sigma} \boldsymbol{A})}} \exp\left(-\frac{1}{2}(\boldsymbol{A}\boldsymbol{s}^t - \boldsymbol{A}\boldsymbol{\mu}^t)^T (\boldsymbol{A}^T \boldsymbol{\Sigma} \boldsymbol{A})^{-1}(\boldsymbol{A}\boldsymbol{s}^t - \boldsymbol{A}\boldsymbol{\mu}^t)\right) \tag{28}$$

$$= \frac{1}{\sqrt{2\pi \det(\boldsymbol{A}^T)\det(\boldsymbol{\Sigma})\det(\boldsymbol{A})}} \exp\left(-\frac{1}{2}(\boldsymbol{s}^t - \boldsymbol{\mu}^t)^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{\Sigma}^{-1} \boldsymbol{A}^T \boldsymbol{A}(\boldsymbol{s}^t - \boldsymbol{\mu}^t)\right) \tag{29}$$

$$= \frac{1}{\sqrt{2\pi \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{s}^t - \boldsymbol{\mu}^t)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{s}^t - \boldsymbol{\mu}^t)\right) \tag{30}$$

$$\tag{31}$$

where $\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{I}$ since $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$ for rotation matrices, and $\det(\boldsymbol{A}) = 1$. This shows that the density functions for the rotated and original signal are identical. Thus, similar to the time-independent case, we cannot uniquely identify the original sources.

Coming back to the setting of TRIS, the two sources s_1^{t-1}, s_2^{t-1} can represent two causal variables, C_1 and C_2, and the orthogonal matrix $\boldsymbol{A}$ represents a valid observation function h. Hence, the previous conclusion also holds here, meaning that there exist a setting in which C_1 and C_2 cannot be uniquely identified from observational data only. This concludes the proof for Proposition 3.1.

(a) Teapot   (b) Cow   (c) Head   (d) Horse   (e) Armadillo   (f) Dragon   (g) Hare

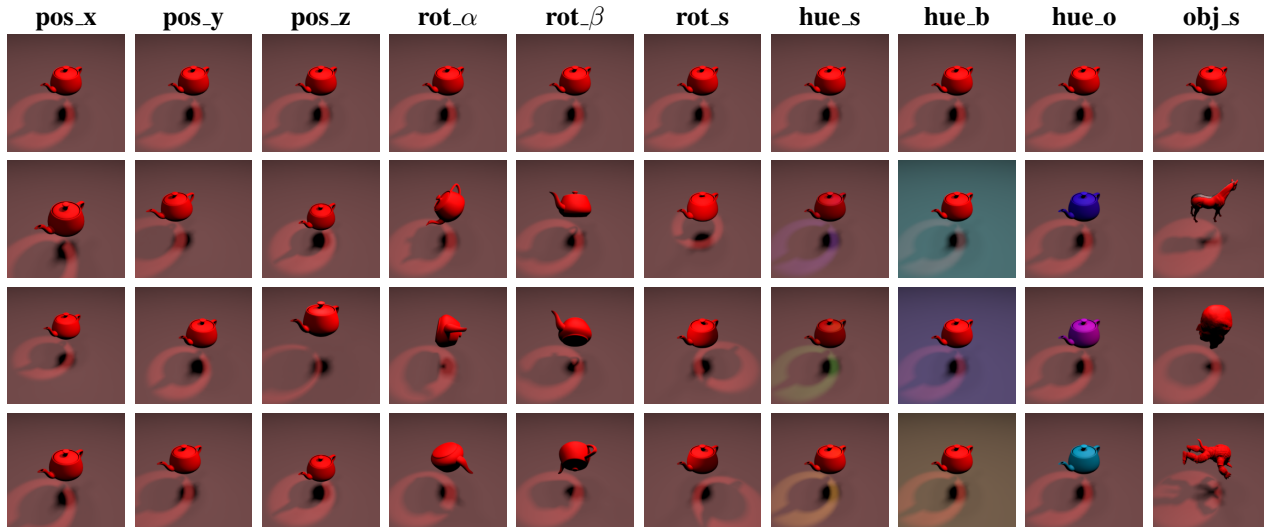*Figure 11.* An example image for each object shape in the Temporal-Causal3DIdent dataset.

Figure 12. Visualizing the different factors of variation in the Temporal Causal3DIdent dataset. Each column represents one dimension of a causal factor, and the different rows show the original image (first row) with only the corresponding causal factor being changed.

C. Experimental Details

In this section, we describe the datasets and hyperparameter used in the experiments of Section 6 in detail. First, we discuss the two datasets, the Temporal Causal3DIdent dataset (Appendix C.1) and the Interventional Pong dataset (Appendix C.2), including giving more examples and insights in the underlying dynamics. The datasets will be publicly released upon publication. In the second part, we discuss the experimental design (Appendix C.3), which includes a closer discussion of the correlation metrics as well as implementation details of CITRIS. Finally, we give an overview of the used hyperparameters in Appendix C.4.

C.1. Temporal Causal3DIdent Dataset

The creation of the Temporal Causal3DIdent dataset closely followed the setup of von Kügelgen et al. (2021). We used the code provided by Zimmermann et al. (2021)[2] to render the images via Blender (Blender Online Community, 2021). We will publish the adapted code for this dataset generation as well as the full datasets used here upon publication with an accompanying license.

C.1.1. CAUSAL FACTOR DESCRIPTION

To begin with, we give a more detailed description of the 7 causal factors here, and provide examples of varying individual factors in Figure 12:

- The **object position** (pos_o) is modeled in 3 dimensions (x - depth dimension, y - horizontal, z - vertical). All values are scaled between -2 as minimum and a maximum of 2, following Zimmermann et al. (2021). For the y and z dimension, this ensures that the object stays within the frame. For the x dimensions, it ensures that the object does not cover the

---
[2]https://github.com/brendel-group/cl-ica

whole camera image, but also does not become too small in resolution for recognizing rotations and shapes.

- The **object rotation** (rot_o) is modeled in 2 dimensions (α - roll angle, β - pitch angle). All dimensions use circular values of $[0, 2\pi)$, meaning that in distributions, we consider values close to 0 and 2π as close as well. The rotation is restricted to two dimensions to guarantee that every object rotation has a unique value assignment of the angles. This can be violated when modeling three angles with a value range of $[0, 2\pi)$.

- The **spotlight rotation** (rot_s) is the positioning of the spotlight as an angle. The value range is $[0, 2\pi)$, where, similarly as the object rotation, we consider it to be circular.

- The **spotlight hue** is the color of the spotlight. The value range is, again, $[0, 2\pi)$, where 0 corresponds to red. Note that the color appearance of the spotlight changes with the object and background color, since we only see the combined reflected color.

- The **background hue** (hue_b) is the color of the background. The value range is $[0, 2\pi)$ with the same color spectrum as the spotlight hue.

- The **object hue** (hue_o) is the color of the object, and follows the same setup as the background hue.

- The **object shape** (obj_s) is a categorical variable describing the object shape. For the 7-shape dataset version, we consider the same object shapes as von Kügelgen et al. (2021): Cow (Crane, 2021), Head (Rusinkiewicz et al., 2021), Dragon (Curless & Levoy, 1996), Hare (Turk & Levoy, 1994), Armadillo (Krishnamurthy & Levoy, 1996), Horse (Praun et al., 2000), Teapot (Newell, 1975). An example image for each of the objects is shown in Figure 11.

C.1.2. DATASET GENERATION

The datasets are generated by starting at a random sample of all causal factors. For each next time step, we generate a sample according to the conditional distributions of each causal factor (see Appendix C.1.3 for details on the distributions). Additionally, we sample intervention targets $I_1^{t+1}, ..., I_7^{t+1}$ for all 7 causal factors. For the datasets in Section 6, we sample the targets from $I_i^{t+1} \sim$ Bernoulli(0.1), and show additionally intervention settings on this dataset in Appendix D.1.3. For each causal factor for which the intervention target is one, we replace its previously sampled value with a new value randomly sampled from a uniform distribution. For angles and hues, the distribution is $U(0, 2\pi)$, while for the positions, we use $U(-2, 2)$. For the object shape, we uniformly sample one out of the seven shapes. After performing the interventions, the sampled vector of causal factors is used to generate an image of a resolution of 64×64 using Blender. Note that for visualization purposes, the depicted images in this section are shown in higher resolution (256×256). This makes it easier to recognize the object shapes and their rotations. However, we use a resolution of 64×64 in experiments to keep the computational cost of the experiments in a reasonable range.

Repeating this generation procedure for several steps results in one long sequence, which we use as a dataset. For the experiments on Temporal-Causal3DIdent Teapot, we generate a sequence of 150,000 images. For the experiments on Temporal-Causal3DIdent 7-shapes, the chosen dataset size was 250,000 images. The large dataset sizes were chosen to prevent any sampling bias and focus the experiments on general identifiability. We noticed that smaller dataset sizes such as 50,000 images still gave good scores on the correlation metrics, but a decrease in the triplet evaluation was noticeable for most causal factors, especially the position and rotation.

C.1.3. TEMPORAL CAUSAL RELATIONS

Below, we define the transition functions used in the causal graph of Temporal Causal3DIdent dataset, as shown in Figure 5 (see Figure 13 for the relations on individual causal dimensions). The chosen function forms are inspired by the ones defined by von Kügelgen et al. (2021) for the Causal3DIdent dataset.
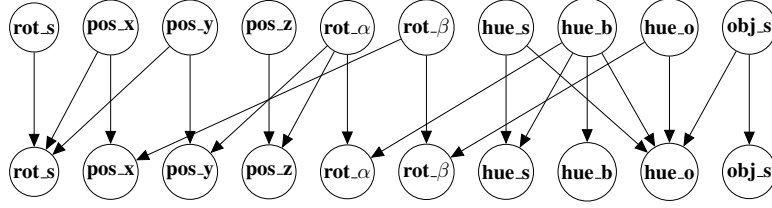
Figure 13. Causal relations between all dimensions of the causal factors of the Temporal Causal3DIdent dataset. The causal graph of Figure 5 summarizes pos_x, pos_y, pos_z into pos_o and rot_α, rot_β into rot_o. See Appendix C.1.3 for details on the conditional distributions.

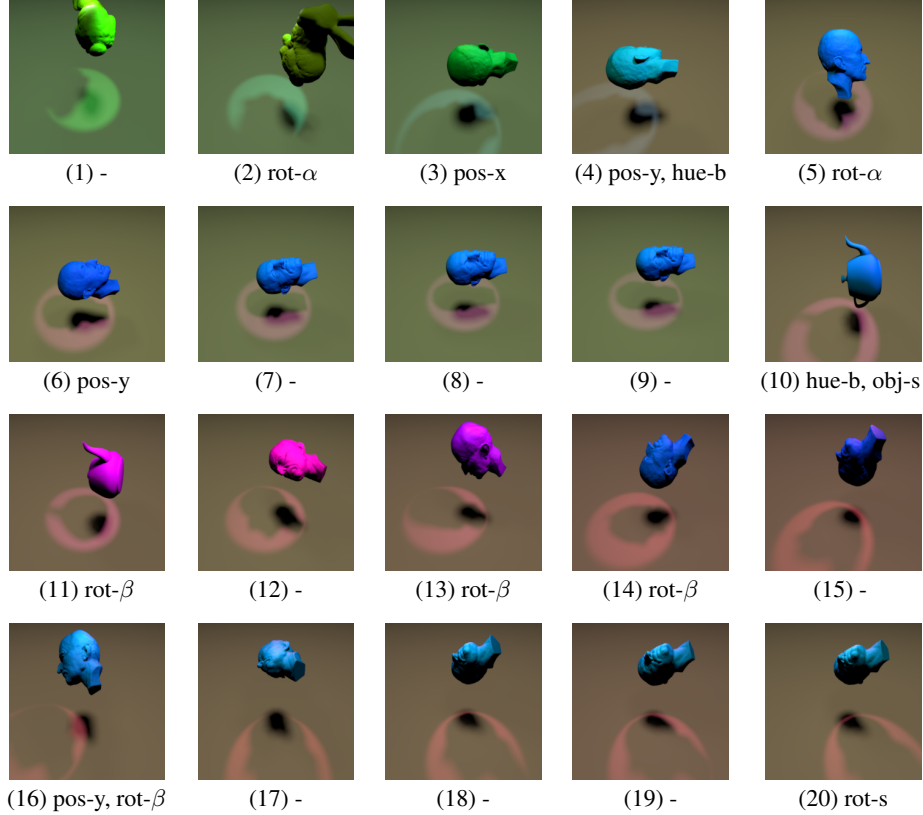|				
(1) -	(2) rot-α	(3) pos-x	(4) pos-y, hue-b	(5) rot-α
(6) pos-y	(7) -	(8) -	(9) -	(10) hue-b, obj-s
(11) rot-β	(12) -	(13) rot-β	(14) rot-β	(15) -
(16) pos-y, rot-β	(17) -	(18) -	(19) -	(20) rot-s

Figure 14. An example sequence with 20 frames in the Temporal Causal3DIdent 7-shapes dataset with higher resolution (from left to right, top to bottom). The causal variables denoted below each image indicate the variables which were intervened on at this time step. For instance, when transitioning from the first to the second image, all variables were sampled according to their temporal dependency except rot-α.

For the position, rotation, and hue values, we sample new values over time with the following functions:

$$\text{pos\_x}^{t+1} = f\left(1.5 \cdot \sin(\text{rot\_}\beta^t), \text{pos\_x}^t, \epsilon_x^t\right) \tag{32}$$

$$\text{pos\_y}^{t+1} = f\left(1.5 \cdot \sin(\text{rot\_}\alpha^t), \text{pos\_y}^t, \epsilon_y^t\right) \tag{33}$$

$$\text{pos\_z}^{t+1} = f\left(1.5 \cdot \cos(\text{rot\_}\alpha^t), \text{pos\_z}^t, \epsilon_z^t\right) \tag{34}$$

$$\text{rot\_}\alpha^{t+1} = f\left(\text{hue\_b}^t, \text{rot\_}\alpha^t, \epsilon_\alpha^t\right) \tag{35}$$

$$\text{rot\_}\beta^{t+1} = f\left(\text{hue\_o}^t, \text{rot\_}\beta^t, \epsilon_\beta^t\right) \tag{36}$$

$$\text{rot\_s}^{t+1} = f\left(\text{atan2}(\text{pos\_x}^t, \text{pos\_y}^t), \text{rot\_s}^t, \epsilon_{rs}^t\right) \tag{37}$$

$$\text{hue\_s}^{t+1} = f\left(2\pi - \text{hue\_b}^t, \text{hue\_s}^t, \epsilon_{hs}^t\right) \tag{38}$$

$$\text{hue\_b}^{t+1} = \text{hue\_b}^t + \epsilon_b^t \tag{39}$$

Object shape	Object hue goal
Teapot	0
Armadillo	$\frac{1}{5} \cdot 2\pi$
Hare	$\text{avg}(\text{hue\_spot}, \text{hue\_back})$
Cow	$\frac{2}{5} \cdot 2\pi$
Dragon	$\pi + \text{avg}(\text{hue\_spot}, \text{hue\_back})$
Head	$\frac{3}{5} \cdot 2\pi$
Horse	$\frac{4}{5} \cdot 2\pi$

Table 4. The causal relation of the object shape, background hue, and spotlight hue to the object hue goal g with which we determine the next step as $\text{hue\_o}^{t+1} = f\left(g, \text{hue\_o}^t, \epsilon_{ho}^t\right)$. The angle mean is defined as $\text{avg}(\alpha, \beta) = \text{atan2}\left(\frac{\sin(\alpha)+\sin(\beta)}{2}, \frac{\cos(\alpha)+\cos(\beta)}{2}\right)$.

(a) $p(\text{pos\_y}, \text{pos\_z})$

(b) $p(\text{pos\_y}, \text{rot\_}\alpha)$
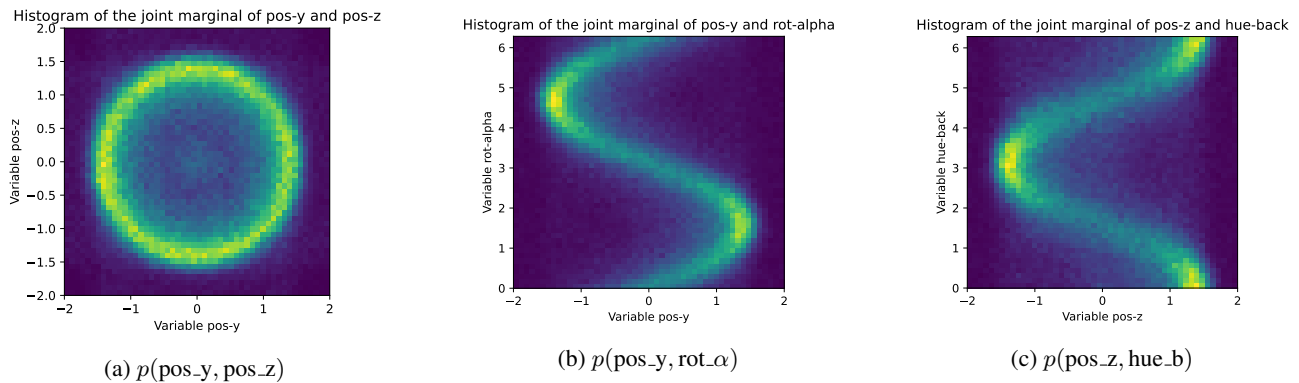
(c) $p(\text{pos\_z}, \text{hue\_b})$

Figure 15. Overview of the joint marginal distributions of selected variables in the Temporal Causal3DIdent dataset, showcasing the correlations among variables. The figures show histograms over the two variables, where yellow indicates a high likelihood/frequency, while dark blue has close to zero samples. (a) The two causal variables pos_y and pos_z share a common confounder, rot_α, which causes them to follow a circle with radius 1.5. (b) rot_α is a parent of pos_y, and one can see that the marginal closely follows its functional form (see Appendix C.1.3). (c) The hue of the background is an ancestor of pos_z, with rot_α in between the two variables. Yet, the marginal clearly follows the cosine signal, showing that the correlation goes beyond parents.

where $f(a, b, c) = \frac{a-b}{2} + c$, and all ϵ-variables being independent samples from a Gaussian distribution with standard deviation 0.1 for positions, and 0.15 for angles and hues. Intuitively, the function f represents that we create a 'goal' position for each variable based on its parents, and move towards the goal by taking the average between goal and current position, with additive noise. This gives us a simulation of a moving system, which, however, also permits large changes without interventions.

The function of the object hue depends on the categorical object shape, which is outlined in Table 4. We use the same setup as von Kügelgen et al. (2021), where the hare is trying to blend into the background and spotlight, while the dragon tries to stand out. The colors of the other objects are spread out across the color ring.

Finally, for the object shape, we use a noisy identity map over time. With a probability of 5%, we change the current object shape with a newly sampled one from a uniform distribution. This introduces additional noise to the object shape besides the interventions.

To showcase the dependency among causal variables, we plot the marginal distribution of tuples of variables in Figure 15. The distributions are plotted based on a histogram of a dataset with 250,000 samples. Despite the occasional interventions, a clear correlation among variables with a confounder and ancestor-descendant relations can be seen. Overall, this shows that the chosen functions in Equation 32 to 39 introduce strong correlations among variables, which makes disentangling the factors a difficult task.
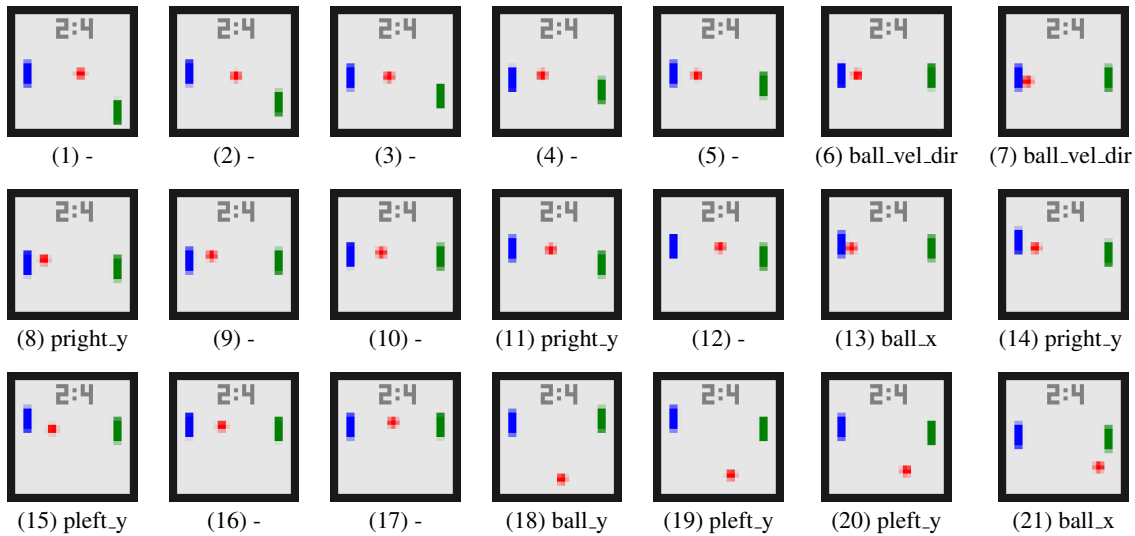
Figure 16. An example sequence with 21 frames in the Interventional Pong dataset. The causal variables denoted below each image indicate the variables which were intervened on at this time step. For instance, when transitioning from the first to the second image, all variables were sampled according to their temporal dependency except. The ball velocity direction is encoded in a 4th channel, which is not visualized here. It encodes the velocity by showing the ball moved by the velocity at the current time step.

C.2. Interventional Pong

The Interventional Pong dataset models the dynamics of the popular Atari game Pong (Bellemare et al., 2013). We show a sequence of 21 frames in Figure 16. Thereby, each frame has a resolution of 32×32.

In this environment, we have six underlying causal factors:

- The **ball x-position** (ball_x) describes the x-position (left to right) of the ball. The value range is limited to $[-1, 1]$ where -1 corresponds to hitting the left border, and 1 the right. Under interventions, we sample the x-position uniformly in the space between the two paddles, which corresponds to a value range of approximately $[-0.7, 0.7]$.

- The **ball y-position** (ball_y) describes the y-position (top to bottom) of the ball. The value range is limited to $[-1, 1]$ where -1 corresponds to hitting the upper border, and 1 the lower border. Under interventions, we sample the y-position uniformly from $U(-1, 1)$.

- The **ball velocity direction** (ball_vel_dir) describes the angle of the velocity of the ball. The value range is $[0, 2\pi)$ where 0 corresponds to the ball heading to the right. Under interventions, we sample the velocity direction uniformly from $U(0, 2\pi)$.

- The **paddle-left y-position** (pleft_y) describes the y-position of the left paddle (blue in the figures). It has the same value range as the causal factor ball_y. Over time, the paddle takes a step towards the ball_y position of the previous time step. The step size is sampled from a Gaussian mean of 0.05 and standard deviation of 0.017. If an intervention is performed on pleft_y, it flips the sign of the step, *i.e.* going up instead of down, with a chance of 50%. This replaces the policy of following the ball with a random one.

- The **paddle-right y-position** (pright_y) describes the y-position of the right paddle (green in the figures), and follows the same dynamics as pleft_y.

- The **score** (score) describes the player's score (gray background numbers). It contains two scores (left player and right player), each with an integer value between 0 and 4. When the ball hits one of the two border, the score for the corresponding player is increased. Further, if a player scores their 5th goal, the score is reset to $0, 0$. No interventions are provided on this causal factor.

Given a previous time step, the game engine performs the following steps in order to determine the next step:
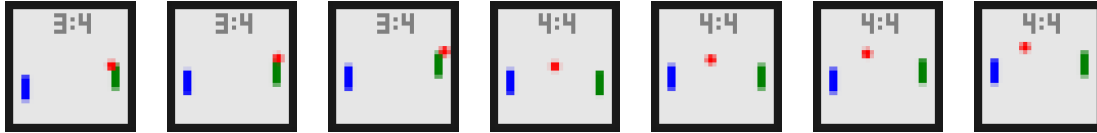
Figure 17. An example sequence of 7 frames (left to right) in the Interventional Pong dataset, where the environment is being reset at the fourth time step. The ball is returned to the center position and gets a new velocity direction, as well as the paddles being moved to new random positions. The score of the respective player is increased as well.
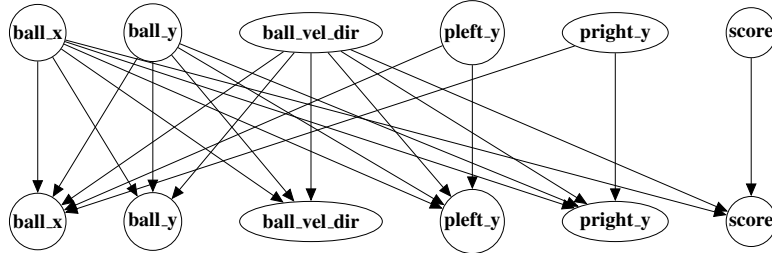

Figure 18. Causal relations between all dimensions of the causal factors of the Interventional Pong dataset. Note that although the graph is very dense, not all relations have an effect on the causal variables at each time step. For instance, the velocity only influences the paddles and the score when the ball x position moved by the velocity ends up in one of the two boundaries.

1. Using the velocity direction at the previous time step, we determine the velocity in x and y direction. We then move the x-position of the ball accordingly.

2. If this new ball-x position hits one of the two borders, we increase the score of the corresponding player and reset the game. Resetting the game includes setting the ball to the center of the field, randomly sampling a ball velocity direction, and replacing the position of the paddles with a new random position between -0.66 and 0.66. The latter is done to prevent the paddles of ending up in one of the two corners when the ball is reset. An example of resetting the environment is shown in Figure 17.

3. If the game was not reset, we do the following steps:

   - We first move the two paddles according to their dynamics, as described before, moving towards the ball.
   - We move the ball y position by the velocity in y direction. If the new ball_y position collides with the upper or lower boundary, we calculate the new y-position it would have after the collision, and mirror the velocity angle on the x-axis.
   - If the new ball x and y position collide with one of the two paddle positions in the previous time step, we reflect the ball accordingly and calculate its new x position, as well as mirroring the velocity angle on the y-axis.
   - Finally, we add Gaussian noise with standard deviation of 0.015 to the ball x and y position, as well as the velocity direction. The small standard deviation was chosen to keep the dynamics similar to Pong and not divert too much into random movements.

4. With a chance of about 35%, we do not perform any intervention. Otherwise, we randomly sample one out of the five causal factors (excluding the score), and perform an intervention as described before.

We provide the full code of the dynamics and dataset generation for this Pong environment in the supplementary. The causal graph implied by these dynamics are shown in Figure 18. Similarly to the Temporal Causal3DIdent dataset, we generate a dataset by sampling one long sequence with a dataset size of 100,000.

C.3. Experimental Design

In this section, we give a more detailed description of elements used in the experiments. We first discuss the evaluation via the correlation metrics and triplet evaluation, and then describe the implementation details of the target classifier.

C.3.1. CORRELATION METRICS

As described in Section 6.1, we evaluate the models by measuring the correlation of the learned latent variables to the ground truth causal factors. Typically, in the literature, the latent variables are one-dimensional. In our setup, however, the latent variables span over multiple dimensions and can describe a causal factor in a non-linear relation. Thus, plotting the correlation per variable would not reveal the true correlation between the learned, multidimensional latent variables and causal factors.

As a first step, we therefore need to learn a mapping between sets of latent variables and true causal factors. We emphasize that during this whole evaluation, the models are frozen and will not be updated. We only make use of their encoders to map observations to latent vectors. For CITRIS, we therefore apply one MLP per set of latent variables that, according to the learned assignment function ψ, belong to the same causal factor. The MLP is then trained to predict *all* causal factors per set of latents, on which we measure the correlation. To this end, we use a two-layer MLP with a latent dimensionality of 128. The MLP size was chosen sufficiently large to uncover any possible correlations between the latent and causal factors. Using larger MLPs did not show any noticeable change in the correlation found, verifying that the MLP is large enough to learn any potential non-linear mapping.

To train the MLP, we split the test dataset into two subsets: one for training the MLP (40% of the dataset), and one for measuring the correlation metrics on (60% of the dataset). The test datasets are generated by randomly sampling sets of causal variables of the respective dataset. We use a dataset size of 25,000 images to prevent any form of overfitting, and ensure a sufficiently large sample size for the correlation metrics. However, in our datasets, the causal factors are correlated themselves. This means that the optimal disentanglement function would yet measure a correlation from, *e.g.*, the latents describing the ball_y position to the paddles' y positions. This makes it difficult to spot spurious correlations between latents and causal factors, since with sufficient latent dimensions, a model can describe the same causal factors in multiple sets of latent dimensions. Since this would not follow the goal of a good disentanglement, we prevent measuring such correlations by perform this evaluation on test datasets without temporal correlations among causal factors. In other words, we generate samples by sampling each causal factor independently, and mapping those into the observation space.

As a loss function for training the MLP, we use a mean-squared-error loss (MSE) for all continuous values excluding the circular values (angles and hues). For the circular values between 0 and 2π, using the MSE loss would be disadvantageous since predicting a value of $2\pi - \epsilon$, when the label is 0, causes a large error although on a unit circle, the two points would be very close. Hence, as a better alternative, we train the MLP by predicting a vector in two dimensions. We then use the cosine distance between the predicted vector and the vector of the ground truth angle projected onto the unit circle. This gives us a loss that respects the circular nature of the causal factor. Since the loss is independent of the length of the predicted vector, we apply a small regularizer to stabilize the vector length. Finally, for the categorical causal factors, we use cross entropy as a training loss.

After training the MLPs, we can use them to create predictions for all causal factors for a set of latents. Then, we determine the correlation metrics between these predictions and the ground truth values for the individual causal factors. Note that for multidimensional causal factors, we calculate the metrics for each dimension individually, and take the average correlation coefficient over dimensions afterwards.

The R^2 correlation coefficient (Spearman, 1904) is a metric which compares the prediction of a variable from a conditional input set to the average prediction. For continuous variables, this is:

$$R^2 = 1 - \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \bar{x})^2} \tag{40}$$

where x_i is the ground truth value of the data point i, $\hat{x}_i$ is the prediction, and $\bar{x} = \sum_i x_i / N$ is the average of the ground truth values. For angles, we adjust the R^2 metric by replacing the difference with the cosine distance, and take the angle mean to calculate $\bar{x}$, since the standard mean of two angles, i.e. $\frac{\alpha+\beta}{2}$, does not take into account the circular nature of the angle values. The mean over more than 2 angles is defined as:

$$\text{avg}(\alpha_1, ..., \alpha_n) = \text{atan2}\left(\frac{1}{n}\sum_i \sin(\alpha_i), \frac{1}{n}\sum_i \cos(\alpha_i)\right) \tag{41}$$

with atan2 being the 2-argument $\arctan$. For the categorical variables, we replace $\bar{x}$ with the most frequent category in the dataset, and distance by a simple equal, *i.e.*, $\delta_{x_i = \hat{x}_i}$. While one could also apply more suitable methods for finding

correlations on categorical variables, we found this simple adjustment sufficient for the two categorical variables we consider in this paper (the object shape in Temporal Causal3DIdent and the score in Pong).

The Spearman's rank correlation (Wright, 1921) ranks both the prediction and ground truth values, and checks the difference in ranking using the Pearson correlation coefficient. We can apply it as is for both categorical and continuous values, excluding angles. For the circular values (angles and hues), we report the average Spearman's rank of their sine and cosine transform, since the original values do not have a unique ranking.

Both metrics result in a correlation matrix, where we have on one dimension the sets of latent variables, and the ground truth causal variables on the other. We show examples of these matrices in Figure 21 and Figure 22. The optimal matrix would contain 1s where the sets of latents was assigned to its respective causal factor by ψ in CITRIS, and zero everywhere else. To summarize such a matrix, we therefore determine the average of these matching values (usually the diagonal). Further, to show the entanglement with other causal factors, we determine for each set of latents the causal factor with the highest correlation besides the matching one, and average these correlations across latents. Since the baselines, SlowVAE and iVAE*, do not learn an assignment function like ψ, we first need to determine one before calculating above's metrics. We do this by first calculating all metrics for all latent dimensions independently, and then assigning a latent variable to the causal factor it has the highest correlation to according to the R^2 correlation coefficient.

C.3.2. TRIPLET EVALUATION

The triplet evaluation provides a parameteric-free evaluation, which can reveal complex dependencies between latent variables in any disentanglement model. To repeat the general approach, we create triplets of images: the first two are randomly sampled test images, while the third one is created based on a random combination of causal factors of the first two images. The test images are taken from a test dataset which contains a sequence of 10,000, generated in the same fashion as the training dataset. Hence, it also contains the common correlations between causal factors. The third image is generated by, *e.g.*, taking the spotlight rotation and object shape from image 1, and all other causal factors from image 2. We overall create a set of 10,000 of such triplets.

For evaluation, we then encode the two test images independently with the model that is supposed to be evaluated. We then perform the combination of ground-truth causal factors as done for the third image in latent space. This means that for each latent dimension z_i, we pick the value of image 1 if the causal factor C_j, where C_j is the factor z_i is assigned to, is taken from image 1 for the ground truth third image. Otherwise, we pick the value of image 2. We repeat this procedure for every latent dimension, and then use the decoder of the model to generate a new image. Note that for the normalizing flow, this also includes inverting the flow before applying the decoder.

Since the reconstruction error is not descriptive of the errors being made, *e.g.* the rotation in Figure 7, we train an additional CNN in a supervised manner that maps images to the causal factors. This CNN has the same architecture as the VAE encoder (see Appendix C.4), and is trained on the same training dataset while using the ground truth causal factors as labels. We use the same training methods for angles as for the MLPs in the correlation metrics. With this model, we can extract the causal factors from the generated image, and report the average distance of these to the ground truth causal factors. However, one common difference between the model's prediction and the ground truth is that the prediction is more blurry, and can cause surprising mispredictions of the CNN when it focused on clear corners in the original dataset. We found it to be beneficial for both the VAE and the AE+NF approach to train the CausalEncoder on reconstructions from the pretrained autoencoder. Since the autoencoder is able to reconstruct the images relatively well and only has minor blurring effects compared to the VAE, the CNN still managed to predict most factors very accurately and, for the Pong dataset and Teapot, with no noticeable difference to the original ground truth data.

C.3.3. TARGET CLASSIFIER IMPLEMENTATION

In this section, we give a detailed description of the target classifier. The target classifier is an additional small network with the goal of predicting the intervention targets from the latent variables over time. While one cannot identify the intervention target 100% of the time, the distribution change between time steps is yet often detectable. This way, the target classifier ensures that information about all causal variables is encoded in the latent space, while guiding the disentanglement in latent space and the assignment function ψ.

In detail, the classifier takes as input the latents z^t and a masked set of z^{t+1} that, according to a sample from ψ, are assigned to the causal variable C_i. We denote this mask with $T_j^{\psi_i} \in \{0,1\}^M$, where $T_j^{\psi_i}$ is sampled from the Gumbel Softmax
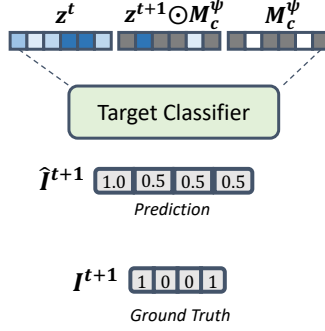
Figure 19. Visualization of the target classifier. From the latents of the first time step, z^t, and masked latents of the second time step, $z^{t+1}_{\psi_i}$ (here shown in the mask setting), the network tries to predict the intervention targets. If the intervention targets are independent and the model has learned a good disentanglement, only I^{t+1}_i should be predicted well.

distribution, modeling the assignment of latent dimension j to the causal factor i. Then, the classifier is trained to recover all intervention targets from those latents by optimizing:

$$\mathcal{L}^{\varphi}_{\text{target}} = -\mathbb{E}_{T,z} \left[\sum_{i=0}^{K} \log p_\varphi \left(I^{t+1} | z^t, z^{t+1} \odot T^\psi_i, T^\psi_i \right) \right] \tag{42}$$

where φ are the parameters of the classifier. Intuitively, we train the parameters φ to predict as many intervention targets from any input as possible, thus revealing any correlations between latents and intervention targets.

When training the classifier for a causal factor C_i, we also take gradients with respect to the latent variables z^t and z^{t+1} as well as the assignment ψ to optimize the target classification for I^{t+1}_i. However, for all other target variables, $z^{t+1} \odot T^\psi_i$ should be uninformative since the intervention target I^{t+1}_i only affects the causal variable C^{t+1}_i while being conditionally independent to all other causal variables. Hence, in this case, the latents and ψ are optimized to make the classifier's prediction follow the target's marginal $p(I^{t+1}_j | I^{t+1}_i)$ (or $p(I^{t+1}_j)$ if $I^{t+1}_i \perp\!\!\!\perp I^{t+1}_j$). This way, the latents and their assignment to causal variables are trained to only contain information from the causal variables that they are assigned to by ψ, and not any others. This training objective can be summarized as:

$$\mathcal{L}^{z,\psi}_{\text{target}} = -\mathbb{E}_{T,z} \left[\sum_{i=0}^{K} \log p_\varphi \left(I^{t+1}_i | Tz_i \right) + \sum_{j \neq i} \sum_{l \in \{0,1\}} p \left(I^{t+1}_j = l | I^{t+1}_i \right) \log p_\varphi \left(I^{t+1}_j = l | I^{t+1}_i, Tz_i \right) \right] \tag{43}$$

where $Tz_i = [z^t, z^{t+1} \odot T^\psi_i, T^\psi_i]$. The second part of the objective is a cross-entropy between the marginal $p(I^{t+1}_j | I^{t+1}_i)$ and the classifier's prediction. Hence, the full loss for CITRIS is $\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \beta_{\text{class}} \cdot \left(\mathcal{L}^{\varphi}_{\text{target}} + \mathcal{L}^{z,\psi}_{\text{target}} \right)$, where $\beta_{\text{class}} > 0$ is a hyperparameter. In common Deep Learning frameworks like PyTorch (Paszke et al., 2019), this two-folded loss can be implemented by stopping the gradients of z and ψ for $\mathcal{L}^{\phi}_{\text{target}}$, and φ for $\mathcal{L}^{z,\psi}_{\text{target}}$. We empirically verify that this simple addition provides considerable benefits in disentanglement and overall optimization of the VAE (see Section 6.2).

C.4. Hyperparameters

C.4.1. NETWORK ARCHITECTURES

All models are implemented in the Deep Learning framework PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon & The PyTorch Lightning team, 2019). The trainings have been performed on a single NVIDIA TitanRTX GPU. Below, we give details about the specific architectures used.

VAEs For all VAE architectures, we make use of the same network architecture to ensure a fair comparison. The encoder is a convolutional architecture, applying several convolutions in sequence with stride 2 for decreasing the resolution. Its output is two parameters per latent variable, which denote the mean and standard deviation for each latent variable. Experiments with increased complexity of the predicted distribution did not show any improvements. For the decoder, we found it beneficial to replace transposed convolutions with bilinear upsampling and using residual blocks. The full architectures are outlined in

Table 5. The architecture for both the encoder and decoder. Every residual block consists of two convolutions, each 64 channels, kernel 3 and stride 1 with BatchNorm+SiLU non-linearity. The upsampling represent bilinear upsampling of the features with a scaling factor of 2. The Tanh activation function on the output scales the output between -1 and 1.

	Layer	Feature Dimension $(H \times W \times C)$	Kernel	Stride	Activation Function
f_{enc}	Conv	$32 \times 32 \times 64$	3	2	BatchNorm+SiLU
	Conv	$32 \times 32 \times 64$	3	1	BatchNorm+SiLU
	Conv	$16 \times 16 \times 64$	3	2	BatchNorm+SiLU
	Conv	$16 \times 16 \times 64$	3	1	BatchNorm+SiLU
	Conv	$8 \times 8 \times 64$	3	2	BatchNorm+SiLU
	Conv	$8 \times 8 \times 64$	3	1	BatchNorm+SiLU
	Conv	$4 \times 4 \times 64$	3	2	BatchNorm+SiLU
	Conv	$4 \times 4 \times 64$	3	1	BatchNorm+SiLU
	Reshape	$1 \times 1 \times 1024$	-	-	-
	Linear	$1 \times 1 \times 256$	-	-	LayerNorm+SiLU
	Linear	$1 \times 1 \times 2 \cdot \text{num\_latents}$	-	-	-
f_{dec}	Linear	$1 \times 1 \times 256$	-	-	LayerNorm+SiLU
	Linear	$1 \times 1 \times 1024$	-	-	-
	Reshape	$4 \times 4 \times 64$	-	-	-
	Upsample	$8 \times 8 \times 64$	-	-	-
	ResidualBlock	$8 \times 8 \times 64$	3	1	-
	Upsample	$16 \times 16 \times 64$	-	-	-
	ResidualBlock	$16 \times 16 \times 64$	3	1	-
	Upsample	$32 \times 32 \times 64$	-	-	-
	ResidualBlock	$32 \times 32 \times 64$	3	1	-
	Upsample	$64 \times 64 \times 64$	-	-	-
	ResidualBlock	$64 \times 64 \times 64$	3	1	-
	Pre-Activation	$64 \times 64 \times 64$	-	-	BatchNorm+SiLU
	Conv	$64 \times 64 \times 64$	1	1	BatchNorm+SiLU
	Conv	$64 \times 64 \times 3$	1	1	Tanh

Table 5. Further, we make use of Batch Normalization (Ioffe & Szegedy, 2015) and the SiLU activation function (Hendrycks & Gimpel, 2016; Ramachandran et al., 2017) for non-linearity. Together, the encoder and decoder have slightly less than 1 million parameters.

Autoencoders For the standard autoencoder training, we found it to be beneficial to increase the complexity of the decoder by applying two residual blocks per resolution instead of one. We also experimented with this increased complexity for the VAE approaches. However, it did not show any noticeable improvement while requiring a smaller batch size and considerably longer training time.

Prior distribution For CITRIS, the prior distribution $p_\phi(z_{\psi_i}^{t+1}|z^t, I_i^{t+1})$ is implemented using an autoregressive model. We use a single autoregressive model that can model the distributions for all latent sets z_{ψ_i} by providing a mask on the intervention vector I^{t+1}. The autoregressive model follow a MADE architecture (Germain et al., 2015) where we assign 16 neurons per layer to each latent variable, and the input to these neurons are the features of all previous latents. Additionally, to the latents, we use the latents of the previous layer as input. The prior is 2 layers deep, and uses the SiLU activation function. Finally, the prior predicts a Gaussian distribution per latent. Similarly to the encoder distribution, one can make the prior more flexible by using normalizing flows. However, in experiments, this showed to not provide any improvement compared to increased computational cost and parameters. For the iVAE*, we use a two-layer MLP with a hidden dimensionality of 128.

Normalizing Flows For the normalizing flow learned on top of the autoencoder, we found a simple affine autoregressive flow to work well. More flexible coupling transformations such as a MixtureCDF layer (Ho et al., 2019) did not show to improve the results. For the autoregressive network, we use the same MADE architecture as for the prior distribution. Further, in between the coupling layer, we use common flow layers such as Activation Normalization and Invertible 1x1

Table 6. An overview of the hyperparameter used for CITRIS-VAE, iVAE* and SlowVAE. This parameter configuration resulted in a training time of about 2 days on one NVIDIA TitanRTX GPU for the Temporal Causal3DIdent, and 12 hours for the Interventional Pong dataset.

Hyperparameter	Value
Batch size	512
Optimizer	Adam (Kingma & Ba, 2015)
Learning rate	1e-3
Learning rate scheduler	Cosine Warmup (100 steps)
KL divergence factor β	1.0
KL divergence factor ψ_0 (λ)	0.01
Number of latents	32 (Causal3DIdent), 16 (Pong)
Number of epochs	600
Target classifier weight	2.0
Gumbel Softmax temperature	1.0

Table 7. An overview of the hyperparameter used for the standard autoencoder. This parameter configuration resulted in a training time of about 2 days on one NVIDIA TitanRTX GPU for the Temporal Causal3DIdent, and 12 hours for the Interventional Pong dataset.

Hyperparameter	Value
Batch size	512
Optimizer	Adam (Kingma & Ba, 2015)
Learning rate	1e-3
Learning rate scheduler	Cosine Warmup (100 steps)
Number of latents	32 (Causal3DIdent), 16 (Pong)
Gaussian noise std	0.05
Number of epochs	1000

Table 8. An overview of the hyperparameter used for CITRIS-NF. This parameter configuration resulted in a training time of 6-10 hours on one NVIDIA TitanRTX GPU.

Hyperparameter	Value
Batch size	1024
Optimizer	Adam (Kingma & Ba, 2015)
Learning rate	1e-3
Learning rate scheduler	Cosine Warmup (100 steps)
KL divergence factor β	1.0
KL divergence factor ψ_0 (λ)	0.01 (Causal3D), 0.1 (Pong)
Number of latents	32 (Causal3DIdent), 16 (Pong)
Number of coupling layers	4 (Causal3DIdent Teapot, Pong), 6 (Causal3DIdent 7-shapes)
Number of epochs	1000
Target classifier weight	2.0
Gumbel Softmax temperature	1.0

Convolutions (Kingma & Dhariwal, 2018).

Target classifier The target classifier uses a single-layer MLP with hidden dimensionality of 128, Layer Normalization (Ba et al., 2016) and a SiLU activation function.

C.4.2. TRAINING HYPERPARAMETERS

We provide an overview of the used hyperparameters in Table 6 (VAE models), Table 7 (Autoencoder training), and Table 8. In general, we use the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 1e-3. We use a Cosine Warmup learning rate scheduler with 100 steps warmup. For the VAEs and autoencoder, we use the MSE loss as reconstruction loss, which corresponds to predicting a Gaussian mean per pixel with fixed standard deviation. We found this to perform considerably better than discrete approaches such as a softmax over the 256 pixel values. To balance the KL divergence and reconstruction loss, we have experimented with scaling the KL divergence like in the β-VAE (Higgins et al., 2017) in the range of $\beta \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$. However, the best results for all three models were achieved when using a value

of $\beta = 1$. Further, we ran a grid search over the γ hyperparameter for SlowVAE for the values $\gamma \in \{1, 2, 4, 8\}$, with the model achieving the best disentanglement results for $\gamma = 1$. Finally, we also experimented with a KL warmup, which slowly increases β to its maximum value. However, no noticeable improvement was found for all three models on the Temporal Causal3DIdent dataset, such that we did not include it for the final experiments.

The autoencoder is trained via the MSE reconstruction loss. As mentioned in Section 4.2, we add small Gaussian noise to the latents during training to prevent the latent space from collapsing into Dirac deltas. We have used a noise standard deviation of 0.05, which did not show to negatively impact the reconstruction loss.

In the normalizing flow, we use 4 coupling layers for the Interventional Pong dataset and the Temporal Causal3DIdent Teapot dataset. However, the more challenging dataset, Temporal Causal3DIdent 7-shapes, showed to benefit from using more coupling layers, hence using 6 here. Note that one considerable benefit of the normalizing flow approach is that it trains much quicker than the VAE once the autoencoder has been trained. Since one autoencoder can be used for multiple different settings, it reduced the overall computational cost and simplified the hyperparameter search.

Table 9. Experimental results for the Temporal-Causal3DIdent Teapot dataset, including standard deviations over 3 seeds. See Table 1 for a detailed discussion on the table and metrics.

	Triplet evaluation distances ↓										Correlation metrics			
	pos_x	pos_y	pos_z	rot_α	rot_β	rot_s	hue_s	hue_b	hue_o	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Oracle	0.02	0.02	0.02	0.02	0.03	0.01	0.02	0.01	0.02	0.02	-	-	-	-
SlowVAE	0.13	0.10	0.12	0.50	0.59	0.22	0.64	0.21	0.17	0.30	0.65	0.20	0.62	0.27
(stds)	±0.002	±0.004	±0.000	±0.009	±0.009	±0.005	±0.004	±0.006	±0.003	±0.003	±0.007	±0.007	±0.005	±0.012
iVAE*	**0.04**	**0.03**	**0.04**	0.25	0.31	**0.03**	0.58	0.02	0.05	0.15	0.78	0.21	0.77	0.17
(stds)	±0.002	±0.001	±0.002	±0.102	±0.243	±0.001	±0.178	±0.002	±0.010	±0.012	±0.036	±0.098	±0.045	±0.043
CITRIS-VAE	0.05	0.04	0.05	0.10	0.20	**0.03**	0.08	0.02	0.05	0.07	0.96	0.02	0.95	**0.04**
(stds)	±0.001	±0.001	±0.000	±0.038	±0.055	±0.000	±0.002	±0.001	±0.001	±0.010	±0.007	±0.002	±0.013	±0.002
CITRIS-NF	**0.04**	**0.03**	**0.04**	**0.06**	**0.10**	**0.03**	**0.04**	**0.01**	**0.04**	**0.04**	**0.98**	**0.01**	**0.97**	0.05
(stds)	±0.001	±0.001	±0.001	±0.002	±0.004	±0.003	±0.002	±0.004	±0.005	±0.002	±0.001	±0.004	±0.001	±0.007

Table 10. Experimental results for the Temporal-Causal3DIdent 7 shapes dataset, including standard deviations over 3 seeds. See Table 1 for a detailed discussion on the table and metrics.

	Triplet evaluation distances ↓											Correlation metrics			
	pos_x	pos_y	pos_z	rot_α	rot_β	rot_s	hue_s	hue_b	hue_o	obj_s	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Oracle	0.08	0.06	0.08	0.06	0.09	0.04	0.04	0.01	0.04	0.00	0.05	-	-	-	-
SlowVAE	0.44	0.25	0.41	0.69	0.75	0.25	0.57	0.10	0.14	0.37	0.40	0.61	0.23	0.59	0.27
(stds)	±0.035	±0.030	±0.021	±0.012	±0.003	±0.003	±0.064	±0.001	±0.004	±0.040	±0.018	±0.053	±0.005	±0.038	±0.006
iVAE*	0.26	0.23	0.34	0.58	0.65	0.10	0.31	0.02	0.09	0.14	0.27	0.80	0.29	0.77	0.28
(stds)	±0.013	±0.038	±0.015	±0.051	±0.043	±0.011	±0.003	±0.000	±0.001	±0.039	±0.002	±0.002	±0.013	±0.001	±0.019
CITRIS-VAE	0.15	0.13	0.23	0.54	0.71	0.07	0.05	0.02	0.06	0.18	0.21	0.89	0.10	0.88	0.12
(stds)	±0.002	±0.000	±0.001	±0.009	±0.009	±0.002	±0.002	±0.000	±0.002	±0.050	±0.004	±0.001	±0.007	±0.002	±0.006
CITRIS-NF	**0.12**	**0.08**	**0.11**	**0.09**	**0.14**	**0.05**	**0.05**	**0.02**	**0.06**	**0.00**	**0.07**	**0.98**	**0.04**	**0.97**	**0.08**
(stds)	±0.001	±0.000	±0.001	±0.000	±0.004	±0.001	±0.001	±0.004	±0.002	±0.000	±0.001	±0.000	±0.003	±0.000	±0.007

D. Additional Experiments and Results

D.1. Temporal Causal3DIdent

D.1.1. DETAILS OF THE MAIN RESULTS

In this section, we provide additional results on the Temporal Causal3DIdent dataset. Firstly, we report the standard deviations over 3 seeds for the results in Table 1, which can be found in Table 9 and Table 10. Generally, CITRIS-VAE and CITRIS-NF showed to be stable across seeds and reach similar performance with different seeds.

For the teapot experiments, the greatest difference is found in the modeling of the rotation angles, since the reconstruction error of a VAE can already be optimized well when modeling the teapot as a sphere. Furthermore, the most common failure mode in the rotations is having a difference of 180 degrees, which increases the triplet distance considerably. The reason is yet again the form of the teapot. Rotating the teapot by 180 degrees in the β angle switches the handle with the spout, which yet again gives a lower reconstruction error than other rotations. CITRIS-NF suffered much less from this problem since the disentanglement is performed independently of the reconstruction error, and the pretrained autoencoder precisely modeled the different rotations.

The iVAE*, on the other hand, showed a more instable behavior. Especially, the hue of the spotlight was very often entangled with different dimensions such as the hue of the object or background. Similarly to CITRIS-VAE, it struggled most with the rotation angles, which also had a high variance in the triplet evaluation. Finally, the SlowVAE showed to be stable across seeds, but achieve lower disentanglement than the other models due to assuming independent causal factors, which was not the case here.

In the experiments on 7 shapes, the most common entanglement was the object shape with the rotation angle. For modeling a change in the object shape independent of the rotation angle, the models need to learn the default rotation angles for each object. Otherwise, one cannot align the different shapes correctly. Furthermore, since having multiple object shapes considerably increased the modeling complexity, the VAE models often had blurry predictions on which the rotation could not be determined. Thus, CITRIS-VAE was able to disentangle most factors well, except the rotation and the object shape. However, CITRIS-NF did not have this difficulty due to relying on an autoencoder, which was trained independently of any latent space prior regularization. This is why CITRIS-NF yet disentangled the different causal factors well, even the rotation.

	Image 1	rot_s, obj_s
Row 1	Image 2	pos_o, rot_o, hue_o, hue_s, hue_b
Row 2	Image 1	rot_o, rot_s, hue_o, hue_s
	Image 2	pos_o, rot_o, hue_b
Row 3	Image 1	pos_o, rot_s, hue_b, obj_s
	Image 2	rot_o, hue_o, hue_s
Row 4	Image 1	rot_s, hue_o, obj_s
	Image 2	pos_o, rot_o, hue_b, hue_s
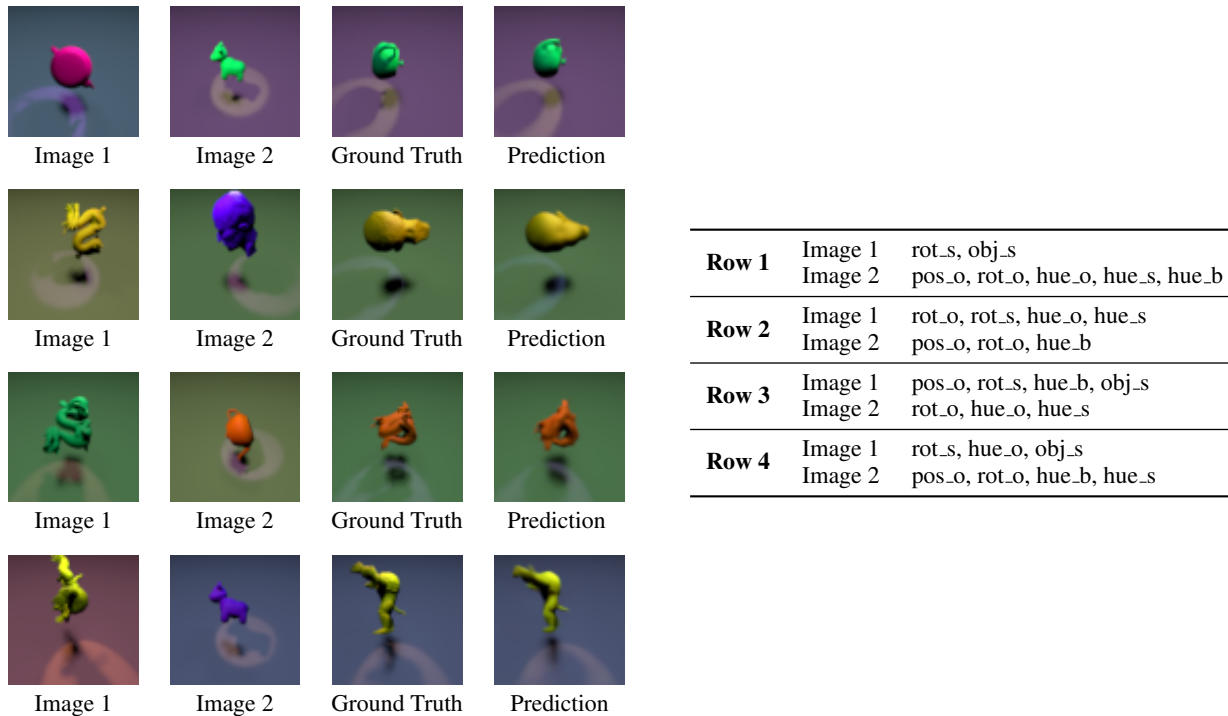
Figure 20. Visualizations of triplet generations. Each row represents one example of the triplet evaluation. The table on the right shows which causal factors were combines from Image 1 and 2 respectively. The predictions were generated by CITRIS-NF trained on all 7 shapes. Despite the blurriness of some predictions, the model still clearly identifies the correct causal factors.

Table 11. Experimental results for the Temporal-Causal3DIdent 5 shapes dataset with additional testing on 2 unknown shapes (Cow and Head), including standard deviations over 3 seeds for CITRIS-NF. See Table 2 for a detailed discussion on the table and metrics. Further, we provide iVAE* as a baseline being trained on the same setting.

	pos_x	pos_y	pos_z	rot_α	rot_β	rot_s	hue_s	hue_b	hue_o	obj_s	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
	Triplet evaluation distances ↓											Correlation metrics			
Oracle	0.08	0.06	0.08	0.06	0.09	0.04	0.04	0.01	0.04	0.00	0.05	-	-	-	-
CITRIS-NF															
5 seen shapes	0.14	0.10	0.14	0.12	0.16	0.06	0.06	0.02	0.07	0.01	0.09	0.98	0.05	0.97	0.10
(stds)	±0.003	±0.004	±0.006	±0.010	±0.012	±0.002	±0.001	±0.002	±0.002	±0.000	±0.004	±0.001	±0.002	±0.000	±0.003
2 unseen shapes	0.32	0.26	0.35	0.36	0.52	0.12	0.10	0.03	0.11	0.10	0.23	0.94	0.15	0.93	0.19
(stds)	±0.026	±0.027	±0.021	±0.037	±0.048	±0.007	±0.009	±0.005	±0.008	±0.033	±0.018	±0.003	±0.012	±0.005	±0.012
iVAE*															
5 seen shapes	0.26	0.18	0.29	0.66	0.68	0.13	0.25	0.03	0.08	0.15	0.27	0.80	0.21	0.81	0.25
2 unseen shapes	0.37	0.28	0.70	0.95	0.94	0.15	0.24	0.03	0.09	0.86	0.46	0.69	0.19	0.68	0.25
CITRIS-NF															
- Cow shape	0.32	0.25	0.36	0.38	0.53	0.10	0.09	0.03	0.10	0.09	0.22	0.93	0.14	0.92	0.22
(stds)	±0.030	±0.024	±0.017	±0.035	±0.049	±0.002	±0.006	±0.004	±0.004	±0.013	±0.015	±0.004	±0.010	±0.004	±0.009
- Head shape	0.32	0.27	0.35	0.33	0.51	0.13	0.11	0.04	0.12	0.12	0.23	0.94	0.15	0.93	0.16
(stds)	±0.023	±0.029	±0.026	±0.038	±0.047	±0.012	±0.011	±0.005	±0.012	±0.052	±0.021	±0.001	±0.014	±0.006	±0.015
iVAE*															
- Cow shape	0.46	0.27	0.98	0.97	0.96	0.14	0.25	0.03	0.09	0.90	0.51	0.69	0.17	0.68	0.26
- Head shape	0.27	0.29	0.42	0.95	0.94	0.16	0.24	0.03	0.10	0.83	0.42	0.69	0.20	0.69	0.25

We show some examples of the triplet generation of CITRIS-NF on this dataset in Figure 20.

In correspondence to the generalization experiments on the Temporal-Causal3DIdent dataset (Table 2), we show a detailed version of the results in Table 11. Overall, the generalization performance shows to be stable across the 2 unseen shapes, Cow (Crane, 2021) and Head (Rusinkiewicz et al., 2021). The most difficult causal factors to generalize are the position and rotation, since both of them are heavily dependent on the object shape and hence also entangled in the autoencoder's representation. Nonetheless, the performance well above random holds promise for future work on exploring these generalization capabilities. Furthermore, we provide a baseline in Table 2 by training an iVAE* on the same setting. To

Table 12. Experimental results for the Temporal-Causal3DIdent Teapot dataset with a limited intervention set, including standard deviations over 3 seeds. CITRIS performs very similar as in the setting of independent interventions (Table 9), showing that it can handle such intervention sets as well.

	Triplet evaluation distances ↓										Correlation metrics			
	pos_x	pos_y	pos_z	rot_α	rot_β	rot_s	hue_s	hue_b	hue_o	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Oracle	0.02	0.02	0.02	0.02	0.03	0.01	0.02	0.01	0.02	0.02	-	-	-	-
SlowVAE	0.10	0.08	0.10	0.39	0.56	0.14	0.61	0.07	0.10	0.24	0.66	0.27	0.65	0.23
(stds)	±0.007	±0.003	±0.001	±0.078	±0.068	±0.000	±0.002	±0.003	±0.003	±0.016	±0.016	±0.014	±0.012	±0.008
iVAE*	0.10	0.07	0.09	0.23	0.37	0.06	0.31	0.02	0.06	0.15	0.81	0.18	0.80	0.18
(stds)	±0.002	±0.001	±0.001	±0.124	±0.213	±0.001	±0.268	±0.001	±0.014	±0.006	±0.045	±0.111	±0.042	±0.068
CITRIS-VAE	**0.05**	**0.03**	0.05	0.09	0.24	**0.03**	0.05	**0.01**	0.05	0.07	0.97	**0.03**	0.96	**0.04**
(stds)	±0.002	±0.001	±0.002	±0.026	±0.091	±0.000	±0.005	±0.000	±0.002	±0.014	±0.001	±0.002	±0.006	±0.007
CITRIS-NF	**0.05**	**0.03**	**0.04**	**0.04**	**0.07**	**0.03**	**0.04**	0.02	**0.04**	**0.04**	**0.98**	0.05	**0.98**	0.08
(stds)	±0.002	±0.000	±0.001	±0.001	±0.003	±0.001	±0.001	±0.000	±0.001	±0.001	±0.002	±0.015	±0.002	±0.023

reduce computational cost and previous experiments showing stable performance across seeds, we run the iVAE* for a single seed. As one would expect, iVAE* heavily struggles with generalizing to unseen shapes since it's en- and decoder have not been trained on these shapes and the new shapes require non-trivial extrapolation. Hence, the model generates arbitrary shapes for the new objects, from which no rotations can be inferred. While the factors independent of the object shape obtain similar performance as for the known shapes, the model's overall disentanglement performance degrades considerably. This underlines the promise of CITRIS-NF to generalize to unseen instances of causal factors.

D.1.2. CORRELATION MATRICES

The results of Table 1 summarized the correlation matrices by reporting the mean on the diagonal and the maximum for any other causal factor. In this section, we additionally show examples of full correlation matrices for all models on the Temporal Causal3DIdent dataset. Figure 21 shows the results for the Temporal Causal3dIdent Teapot dataset. The x-axis shows all dimensions of all causal factors, which includes the x, y and z dimensions of the position causal factor. The optimal model would have 1s between pos_o and pos_x, pos_y, pos_z. The figure shows the difficulty with the rotation for all models. Further, in the iVAE*, one can see the difficulty the model had with the hue of the spotlight.

The results for the 7-shapes datasets are shown in Figure 22. As one could already see from the results in Table 1, only CITRIS-NF is able to disentangle all causal factors well while the other models have especially difficulties with the rotation.

D.1.3. CONFOUNDED INTERVENTIONS

The previous experiments on the Temporal Causal3DIdent dataset were performed on interventions that were independently sampled to show that CITRIS can handle both single-target and joint interventions. In this section, we repeat the experiments on the Teapot dataset, but with a different intervention settings. Instead of sampling each intervention target independently, we consider 6 possible intervention sets:

1. The observational regime, *i.e.*, no causal variable being intervened upon

2. A single target intervention on the object position (pos_o)

3. A single target intervention on the object rotation (rot_o)

4. A joint intervention on the rotation and hue of the spotlight (rot_s, hue_s)

5. A joint intervention on the hue of the object and spotlight (hue_o, hue_s)

6. A joint intervention on the hue of the object and background (hue_o, hue_b)

Note that this set of interventions fulfills the intervention conditions of Theorem 3.3 for all causal variables. At each time step, we perform one of these interventions, and use the same temporal dependencies and dynamics as before.

The results over 3 seeds for these experiments are summarized in Table 12. All models achieved very similar results as for the setting with independent intervention targets. This verifies that CITRIS can also disentangle the different causal factors in a setting with limited diversity in the observed intervention settings.
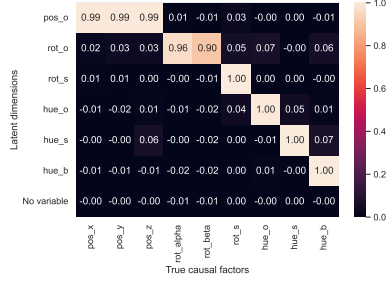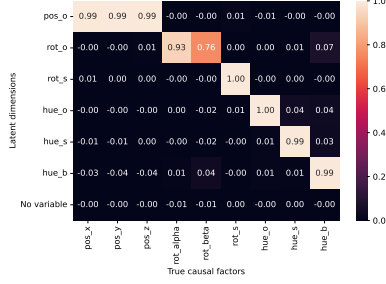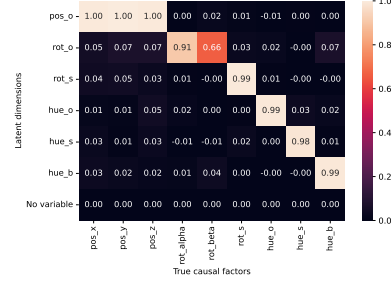
(a) CITRIS-NF, R^2 correlation matrix
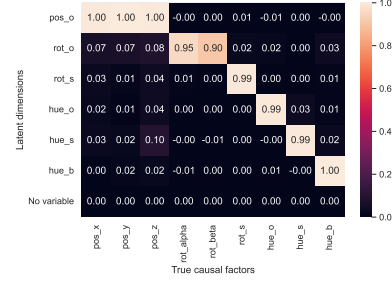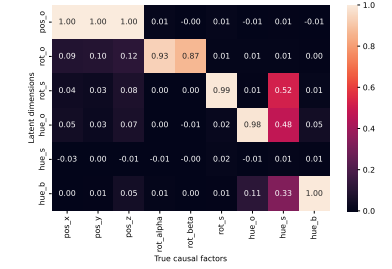
(b) CITRIS-NF, Spearman correlation matrix
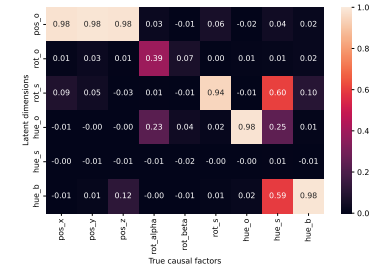
(c) CITRIS-VAE, R^2 correlation matrix
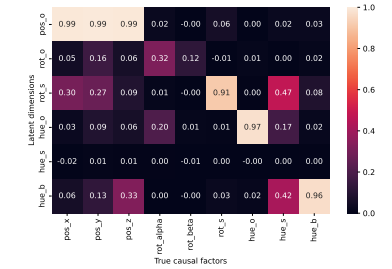
(d) CITRIS-VAE, Spearman correlation matrix

(e) iVAE*, R^2 correlation matrix

(f) iVAE*, Spearman correlation matrix
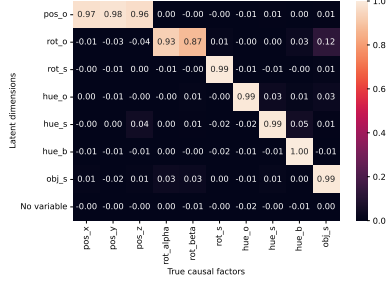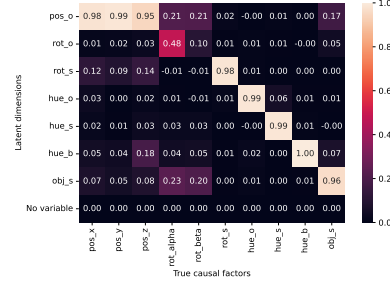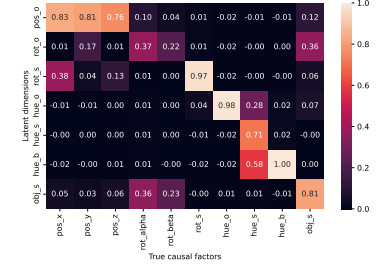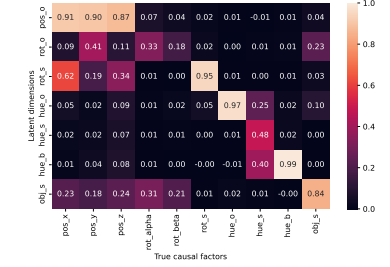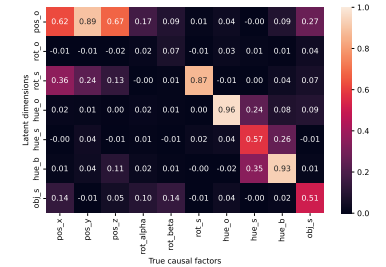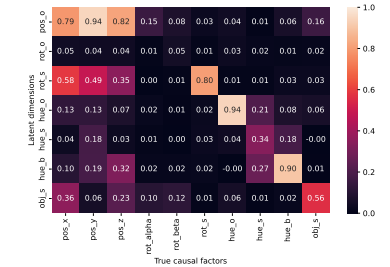
(e) SlowVAE, R^2 correlation matrix

(f) SlowVAE, Spearman correlation matrix

Figure 21. Correlation matrices for the experiments on the Temporal-Causal3DIdent Teapot 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 CITRIS. 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 (R^2 left, Spearman right).

D.1.4. LEARNING THE CAUSAL GRAPH

The transition prior p_ϕ of CITRIS takes as input all latent variables from the previous time-step, z^t, in order to model a distribution over z^{t+1}. Once the model has converged, we can try to learn a sparser prior, which removes unnecessary edges between sets of latent variables that have been assigned to different causal factors. More specifically, we are trying to find the causal graph between $z_{\Psi_0}^t, z_{\Psi_1}^t, ..., z_{\Psi_K}^t$ and $z_{\Psi_0}^{t+1}, z_{\Psi_1}^{t+1}, ..., z_{\Psi_K}^{t+1}$, where the directions of the directions of the edges are

(a) CITRIS-NF, R^2 correlation matrix
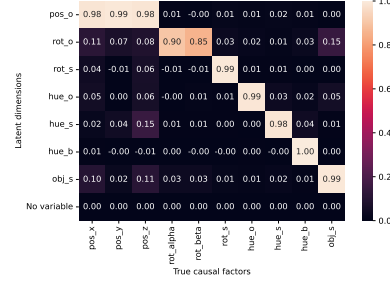
(b) CITRIS-NF, Spearman correlation matrix
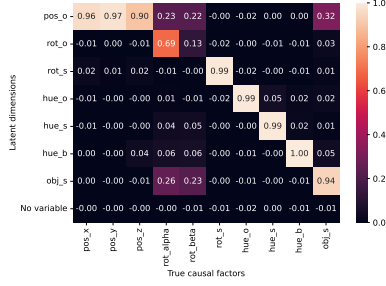
(c) CITRIS-VAE, R^2 correlation matrix

(d) CITRIS-VAE, Spearman correlation matrix

(e) iVAE*, R^2 correlation matrix

(f) iVAE*, Spearman correlation matrix

(e) SlowVAE, R^2 correlation matrix

(f) SlowVAE, Spearman correlation matrix

Figure 22. Correlation matrices for the experiments on the Temporal-Causal3DIdent 7-shapes 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 CITRIS. 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 (R^2 left, Spearman right).

determined by $t \to t+1$, and no edges within a time step exist. We find this graph by using ENCO (Lippe et al., 2022), a continuous-optimization causal discovery method which supports the usage of arbitrary neural networks. The causal graph is parameterized by a weight matrix $\gamma \in \mathbb{R}^{K+1 \times K+1}$, where $\sigma(\gamma_{ij})$ represents the probability of having the edge $z_{\Psi_i}^t \to z_{\Psi_j}^{t+1}$ in the causal graph. To estimate the conditional likelihoods under different causal graphs, we learn a new autoregressive prior $p_\varepsilon(z_{\Psi_i}^{t+1} | z^T \cdot M, I_i^{t+1})$ where M is mask on the latents at time-step t according to the causal graph:
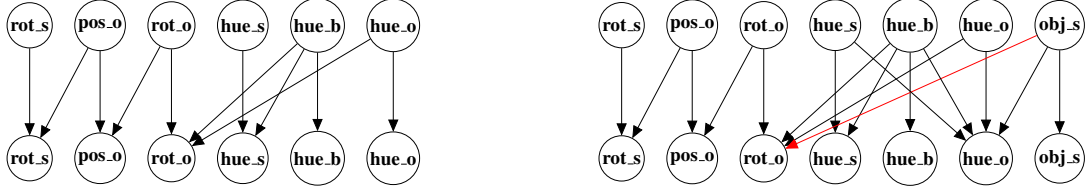
Figure 23. The learned graphs of CITRIS-NF from the Temporal-Causal3DIdent dataset. False positive edges are colored red. **Left**: Temporal-Causal3DIdent Teapot. The learned causal graph is identical to the ground truth graph. Since we have a single object shape, we do not include obj_s as a causal variable in the graph. Additionally, the parent set of hue_o is reduced due to the constant object shape. **Right**: Temporal-Causal3DIdent 7-shapes. The edge obj_s→rot_o is predicted due to disentangling these two variables being the hardest task of the dataset, and small correlations remain.

(a) CITRIS-NF, R^2 correlation matrix

(b) CITRIS-NF, Spearman correlation matrix
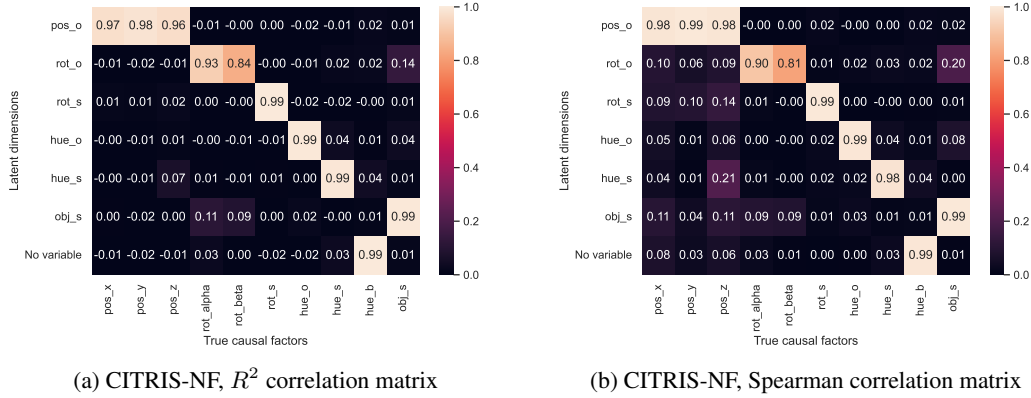
Figure 24. Correlation matrices for the experiments on the Temporal-Causal3DIdent 7-shapes dataset where no intervention targets for the causal variable hue-background are shown (**left**: R^2 correlation, **right**: Spearman correlation). The set z_{ψ_0} is represented by 'no variable' in the plots of CITRIS. As intended, CITRIS learns to map the background hue to z_{ψ_0}, i.e. the latents that represent all information that does not belong to the causal variables for which we are given intervention targets. The remaining causal variables are disentangled as well as when interventions were provided for all variables (compare to Figure 22).

$M_j = \hat{M}_{\psi(j)}, \hat{M}_k \sim \text{Bernoulli}(\gamma_{ki})$. In other words, we sample a causal graph, and then mask a set of latents $z^t_{\Psi_j}$ if the causal variable C_j does not have an edge to C_i: $z^t_{\Psi_j} \not\rightarrow z^{t+1}_{\Psi_i}$. With this prior, we can then determine the gradients for γ on a held-out test set following ENCO's gradient updates. We use a sparsity regularizer of $\lambda = 0.05$ and train for 100 epochs.

We perform the experiments on two learned CITRIS-NF, one for Temporal-Causal3DIdent Teapot, and one for Temporal-Causal3DIdent 7-shapes. The learned graphs are shown in Figure 23. For the teapot experiments, the graph is identical to the ground truth, showing that CITRIS has indeed learned the causal variables. For the 7 shapes experiment, the model has one false positive prediction, the edge obj_s→rot_o. This fits our previous discussion on the results of the 7 shape experiments (Table 10) since the object shape and rotation was the most difficult to disentangle for all models. Hence, there remains some correlation between the object shape and rotation in the CITRIS-NF model.

D.1.5. INTERVENTIONS ON A SUBSET OF VARIABLES

One property of CITRIS is that when interventions are only provided for a subset of variables, CITRIS is yet able to disentangle these variables, while all remaining information is grouped into the latents z_{Ψ_0}. We have used this to disentangle the score in the Interventional Pong experiments (see Section 6.3), and show here that it can also be applied to the more complex Temporal Causal3DIdent dataset. Specifically, we reuse the datasets that have been used for the experiments in Section 6.2, and remove the intervention targets for a subset of variables. This keeps all temporal relations intact, while providing an additional challenge of modeling a multimodal distribution for the excluded variables.

As a first setting, we train CITRIS-NF on the Temporal Causal3DIdent 7-shapes dataset, where the interventions on the hue of the background have been omitted. The correlation matrices in Figure 24 show that CITRIS learned to disentangle all causal variables in a similar accuracy as in the full experiment (see Figure 22), while the background hue is assigned to the latents of z_{ψ_0}. Additionally, we show in Figure 21 results on the Temporal Causal3DIdent Teapot dataset, in which
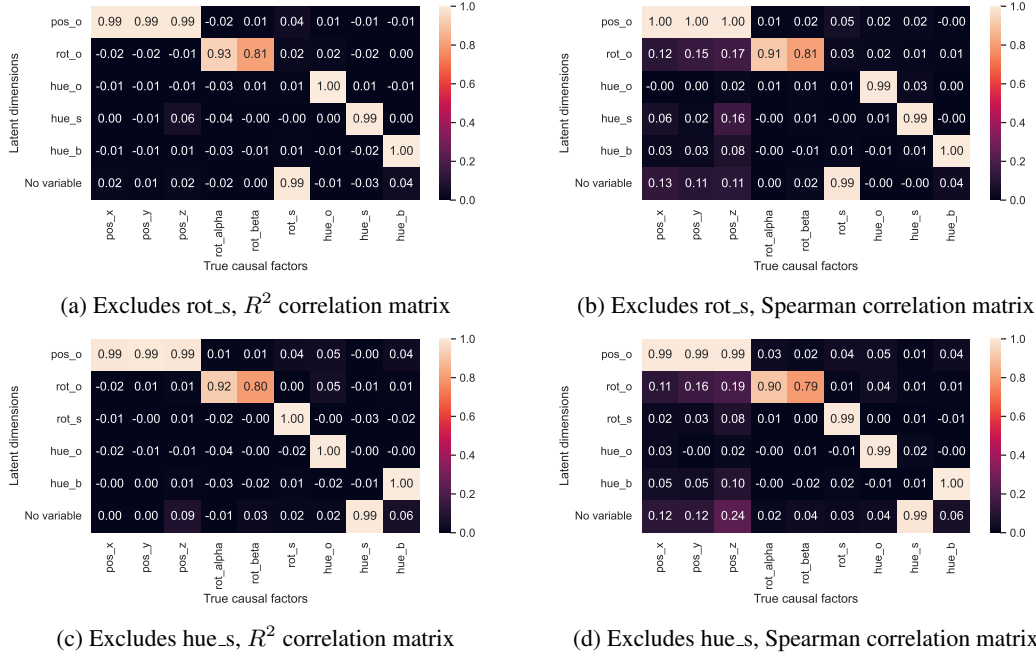
(a) Excludes rot_s, R^2 correlation matrix

(b) Excludes rot_s, Spearman correlation matrix

(c) Excludes hue_s, R^2 correlation matrix

(d) Excludes hue_s, Spearman correlation matrix

Figure 25. Correlation matrices for the experiments on the Temporal-Causal3DIdent Teapot dataset where no intervention targets for the causal variable rotation-spotlight (a,b) and hue-spotlight (c,d) are shown (**left**: R^2 correlation, **right**: Spearman correlation). The set z_{ψ_0} is represented by 'no variable' in the plots of CITRIS. CITRIS assigns the variables without intervention targets to z_{ψ_0}, and the remaining causal variables only show slightly worse disentanglement than for the full intervention set (compare to Figure 21).
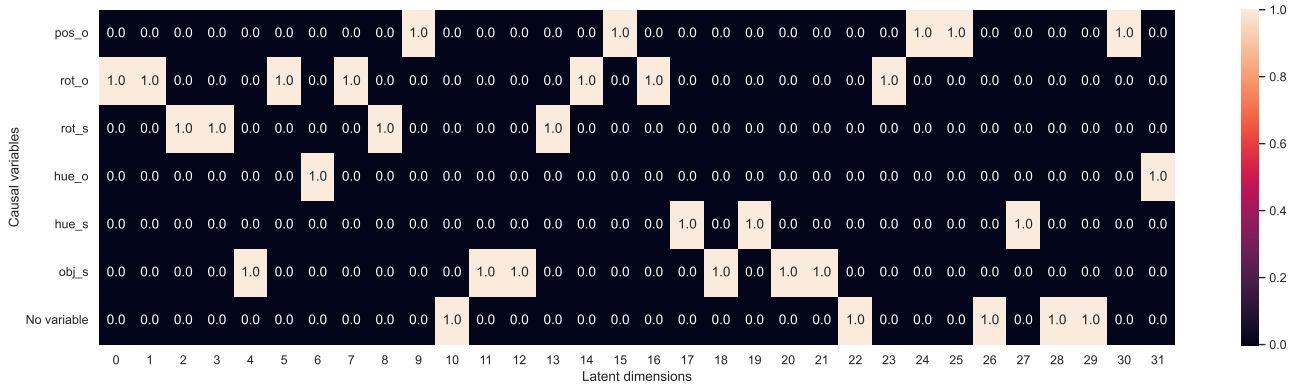
Figure 26. Learned latent-to-causal variable assignment map ψ for an CITRIS-NF trained on the Temporal Causal3DIdent 7-shapes datasets with no interventions on the hue of the background. The map is learned via a Gumbel-Softmax, meaning that we have a softmax distribution over the possible causal variables for each latent variable. Thus, the figure shows the probability that latent dimension i (x-axis) is assigned to a specific causal variable (y-axis), where the distribution usually gets strongly peaked as training progresses. The model has 32 dimensions, of which 5 are assigned to pos_o, 7 to rot_o, 4 to rot_s, 2 to hue_o, 3 to hue_s, 6 to obj_s, and 5 to 'no variable', i.e. z_{ψ_0}. In general, we find that the most complex causal variable, here rot_o and obj_s, are assigned the most dimensions, while circular values have always at least 2 dimensions (the model commonly encodes them as two or more shifted sine-waves).

we exclude intervention targets for the spotlight rotation (a,b) and spotlight hue (c,d) respectively. Both models assign the correct information to the 'no-variable' slot. The disentanglement for all other variables shows to still be very good, but slightly worse than the results on the full intervention set, which comes from additional entanglement in z_{ψ_0}. Still, the experiments show that CITRIS can also handle variables without interventions in complex 3D rendered scenes.

Table 13. Experimental results for the Interventional Pong dataset, extension of Table 3. The six causal factors are the ball x and y position, the velocity direction, the left paddle y position, the right paddle y position, and the player's score. In the correlation metrics, *diag* refers to the average score on its diagonal (optimal 1), and *sep* for the average of the maximum correlation per causal variable besides itself (optimal 0).

|	Triplet evaluation distances ↓							Correlation metrics			
	`ball_x`	`ball_y`	`ball_vel_dir`	`pleft_y`	`pright_y`	`score`	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Oracle	0.01	0.01	0.01	0.01	0.01	0.00	0.01	-	-	-	-
SlowVAE	0.06	0.21	0.99	0.48	0.07	0.23	0.34	0.61	0.17	0.66	0.23
(stds)	±0.006	±0.012	±0.002	±0.121	±0.089	±0.001	±0.001	±0.004	±0.012	±0.007	±0.013
iVAE*	0.03	0.21	0.05	**0.03**	**0.03**	0.20	0.09	0.91	0.04	0.92	0.06
(stds)	±0.008	±0.255	±0.005	±0.011	±0.009	±0.015	±0.039	±0.135	±0.056	±0.131	±0.058
CITRIS-VAE	**0.02**	0.03	0.05	0.04	**0.03**	**0.00**	0.03	0.99	**0.01**	0.99	**0.05**
(stds)	±0.001	±0.004	±0.009	±0.006	±0.002	±0.001	±0.001	±0.001	±0.000	±0.001	±0.010
CITRIS-NF	**0.02**	**0.02**	**0.02**	**0.03**	**0.03**	**0.00**	**0.02**	**1.00**	0.04	**1.00**	0.10
(stds)	±0.002	±0.002	±0.001	±0.002	±0.001	±0.001	±0.001	±0.001	±0.011	±0.001	±0.023

D.1.6. VISUALIZING THE LEARNED ASSIGNMENT FUNCTION ψ

CITRIS learns an assignment between latent and causal variable, ψ. To get an intuition of what a learned function of such looks like, we visualize the parameters of ψ of a trained CITRIS-NF in Figure 26. The model is trained on the Temporal Causal3DIdent 7-shapes dataset, where no interventions have been shown for the background hue (i.e., the same model as in Figure 24). Overall, the probabilities of a latent variable belonging to a certain causal variable are very peaked after training, with the second-highest value per latent variable to be between $1e$-6 and $1e$-9. This shows that the soft relaxation of ψ becomes discrete as training progresses. Further, we find that the most latent variables are assigned to the intuitively most complex causal variables. In the Temporal Causal3DIdent 7-shapes dataset, this corresponds to the rotation of the object and the shape of the object, assigned 7 and 6 latent respectively. The property of learning a flexible latent space size per causal variable is thereby important, since when the causal variables are not known, it would be difficult to manually estimate the number of latent variables per causal variable. Moreover, an interesting property we observe across all models is that circular values like the hues are always encoded in at least two latent dimensions. On closer inspection, the model encodes the angles as two or more shifted sine-waves, verifying our motivation of using multidimensional latent representation even for single-dimensional causal variables.

D.2. Interventional Pong

The full experimental results for the interventional Pong dataset can be found in Table 13, and we plot the correlation matrices for all models in Figure 27. The triplet evaluation distances show that SlowVAE had difficulties with disentangling all the factors. This is since they are highly correlated, while SlowVAE assumes independence in the factors. Both the iVAE* and SlowVAE had troubles with disentangling the y position of the ball (ball_y), which was often entangled together with the ball velocity direction (ball_vel_dir) in the models. Further, the player's score was only consistently recovered by CITRIS. Note that for the triplet evaluation, since there are no interventions given on the player's score, we use the variables z_{ψ_0} of CITRIS of that image of which the score it used for the third image. CITRIS-NF slightly improves upon CITRIS-VAE, mostly due to better modeling of ball_vel_dir. This is because it only affects a single channel and hence has a lower effect in the reconstruction error. Nonetheless, the experiments show that for simpler datasets, CITRIS-VAE can disentangle the causal factors as well as the normalizing flow.
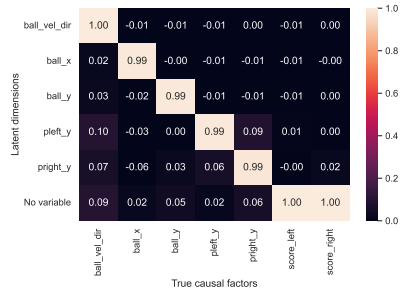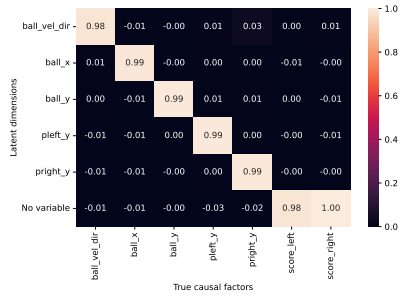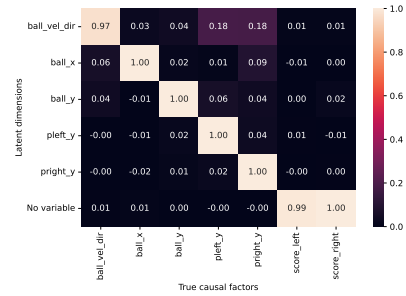
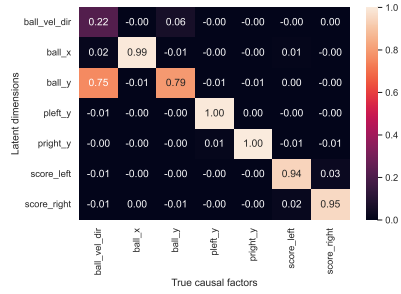(a) CITRIS-NF, R^2 correlation matrix
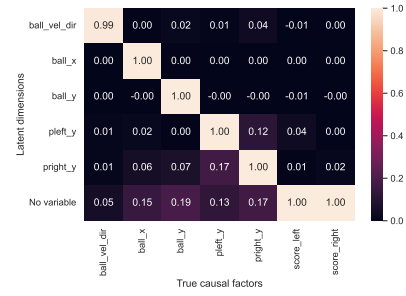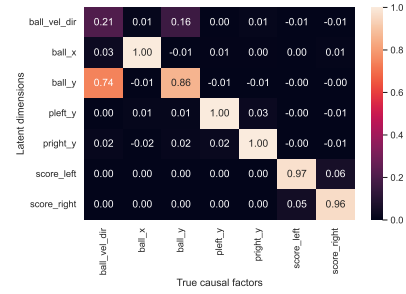
(b) CITRIS-NF, Spearman correlation matrix
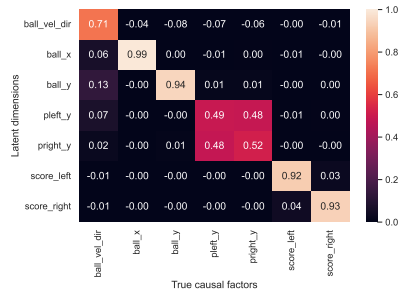
(c) CITRIS-VAE, R^2 correlation matrix

(d) CITRIS-VAE, Spearman correlation matrix
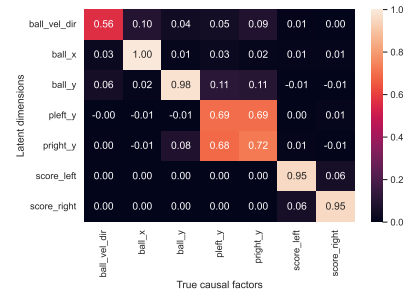
(e) iVAE*, R^2 correlation matrix

(f) iVAE*, Spearman correlation matrix

(e) SlowVAE, R^2 correlation matrix

(f) SlowVAE, Spearman correlation matrix

Figure 27. Correlation matrices for the experiments on the Interventional Pong 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 CITRIS, where no interventions on score_left and score_right have been provided. The x-axis shows the ground truth causal factors. The heatmap is the correlation matrix between those factors (R^2 left, Spearman right).
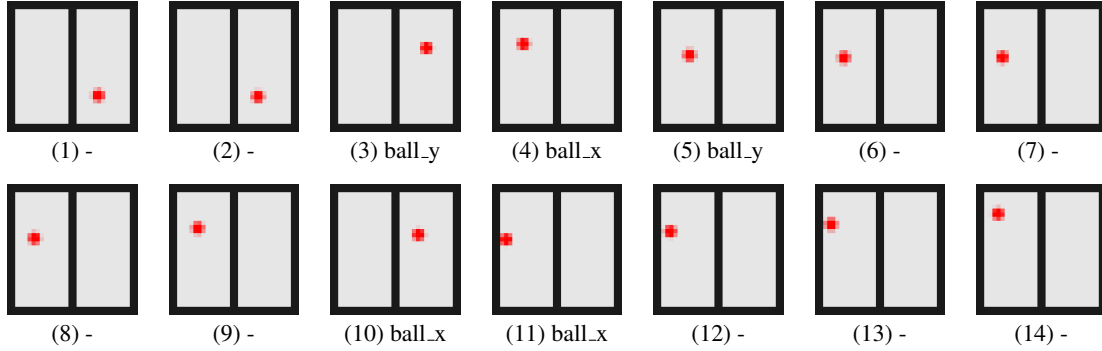
Figure 28. An example sequence with 14 frames in the Ball-in-Boxes dataset. The causal variables denoted below each image indicate the variables which were intervened on at this time step, where an intervention on ball_x changes the box identifier (left or right), and ball_y the y position of the ball.

Table 14. Experimental results for the Ball-in-Boxes dataset. The three evaluation causal factors are the box identifier ball_b, the internal ball x position ball_xin, and the ball y position. In the correlation metrics, *diag* refers to the average score on its diagonal (optimal 1), and *sep* for the average of the maximum correlation per causal variable besides itself (optimal 0). For the correlation metrics, we consider z_{ψ_0} to be the latent variables in which ball_xin should be modeled, *i.e.* have a high correlation with. From the results, one can clearly see that CITRIS has indeed learned to disentangle ball_b and ball_xin, showing that CITRIS learns the minimal causal variables.

	Triplet evaluation distances ↓				Correlation metrics			
	`ball_b`	`ball_y`	`ball_xin`	Mean	R^2 diag ↑	R^2 sep ↓	Spearman diag ↑	Spearman sep ↓
Oracle	0.00	0.01	0.01	0.01	-	-	-	-
CITRIS-VAE	0.00	0.01	0.02	0.01	0.99	0.00	1.00	0.02
(stds)	±0.001	±0.003	±0.005	±0.002	±0.001	±0.001	±0.001	±0.002

D.3. Ball-in-Boxes Example

Throughout the paper, we have referred to the example of a ball in a 2D space, which can be in one of two boxes. It can move freely within a box, but cannot switch between them by itself. Only under interventions on its x position, we switch the box. However, this intervention does not influence the internal position of the ball in the box. Under our theoretical setup of CITRIS, the box identifier represents the minimal causal variable of x, while the internal position should be modeled in z_{ψ_0}.

In this section, we give empirical evidence that this is indeed also the case in practice. For simplicity, we implement it by generating frames of 32×32 resolution similar to our Pong environment (see Figure 28). However, the dynamics are replaced by the ball-in-boxes example: we consider two causal variables, ball_x and ball_y. Both ball_x and ball_y follow a truncated Gaussian distribution, *i.e.* move randomly with Gaussian noise over time in the boundaries of the box the ball is in. Under interventions on ball_y, we uniformly resample the position of the ball. Under interventions on ball_x, we switch the box without influencing the internal ball position. We generate the dataset by using a sequence of $100,000$ frames, and sample the intervention targets from $I_i^t \sim \text{Bernoulli}(0.2)$. On this dataset, we apply CITRIS-VAE with the same experimental setup as for the Pong dataset, including the small regularizer $\lambda = 0.01$, to show that little change in the KL divergence already encourages the disentanglement to z_{ψ_0}.

For evaluation purposes, we perform the correlation and triplet evaluation by predicting both the box identifier ball_b and the box internal position ball_xin instead of just ball_x. Note that this choice does not influence the training of CITRIS in any way, but simply allows us to identify whether the split between ball_b and ball_xin has been learned or not. The results are shown in Table 14. One can clearly see that CITRIS learned to disentangle the internal ball position ball_xin and ball_b, since z_{ψ_0} has a R^2 correlation score of almost one to ball_xin, and close to zero for ball_b. We show a visualization of the R^2 and Spearman correlation matrix in Figure 29. In conclusion, this experiment validates that CITRIS indeed learns the minimal causal variables, *i.e.* the information that strictly depends on the individual interventions.
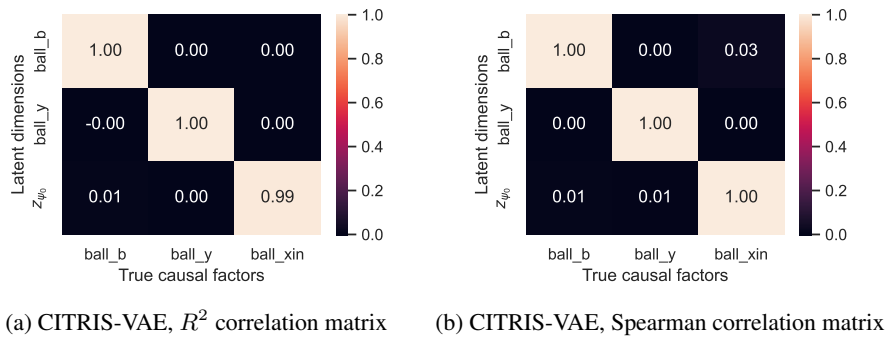
(a) CITRIS-VAE, R^2 correlation matrix   (b) CITRIS-VAE, Spearman correlation matrix

Figure 29. The learned correlation matrices of CITRIS-VAE on the Ball-in-Boxes dataset. The y-axis shows the sets of latent dimensions which are assigned to causal variable 1 (ball_b/ball_x), causal variable 2 (ball_y), and to none of them (z_{ψ_0}). The x-axis shows the true causal factors. Each element of the matrix therefore describes the correlation between a set of latent factors and a ground truth causal factor. CITRIS disentangles the causal factors as intended, and assigning the internal box position ball_xin to z_{ψ_0} since it is independent of the provided interventions.