
Implicit Causal Representation Learning via Switchable Mechanisms

Shayan Shirahmad Gale Bagi

Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
sshirahm@uwaterloo.ca

Zahra Gharaee

Department of Systems Design Engineering
University of Waterloo
Waterloo, Canada
zahra.gharaee@uwaterloo.ca

Oliver Schulte

School of Computing Science
Simon Fraser University
Burnaby, Canada
oschulte@cs.sfu.ca

Mark Crowley

Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
mark.crowley@uwaterloo.ca

Abstract

Learning causal representations from observational and interventional data in the absence of known ground-truth graph structures necessitates implicit latent causal representation learning. Implicit learning of causal mechanisms typically involves two categories of interventional data: hard and soft interventions. In real-world scenarios, soft interventions are often more realistic than hard interventions, as the latter require fully controlled environments. Unlike hard interventions, which directly force changes in a causal variable, soft interventions exert influence indirectly by affecting the causal mechanism. However, the subtlety of soft interventions impose several challenges for learning causal models. One challenge is that soft intervention's effects are ambiguous, since parental relations remain intact. In this paper, we tackle the challenges of learning causal models using soft interventions while retaining implicit modeling. Our approach models the effects of soft interventions by employing a *causal mechanism switch variable* designed to toggle between different causal mechanisms. In our experiments, we consistently observe improved learning of identifiable, causal representations, compared to baseline approaches.

1 Introduction

One of the long-standing challenges in causal representation learning is how to recover the ground-truth causal graph of a system solely from observations. Termed the *identifiability of causal models* problem, this endeavor is crucial. Without achieving identifiability, we risk erroneously attributing causal relationships to learned representations. Furthermore, statistical models can masquerade as Directed Acyclic Graphs (DAGs) where edges lack causal significance, further complicating our pursuit.

When considering the challenge of identifying causal models, it is known that the Markov condition in graphs is insufficient for this task [26]. Thus, without additional assumptions or data, we find Preprint. Under review.

ourselves limited to learning only a *Markov Equivalence Class* (MEC) of the causal model. Existing works have made different assumptions about availability of ground-truth causal variables labels [34], model parameters [1], availability of paired interventional data [3, 31], and availability of intervention targets [17] to ensure identifiability of causal models.

Interventional data are usually obtained through *soft* or *hard* interventions. Hard interventions usually involve controlled experiments and they sever the connection of an intervened variable with its parents [24]. In terms of Structural Causal Models (SCM), hard interventions set the causal mechanism relating a causal variable to its parents, to a constant. Due to ethical or safety reasons, it may not be possible to perform hard interventions in many real-world applications. On the other hand, the effects of soft interventions are more subtle since parent variables can still affect their children. These effects can be modeled by a change in the set of parents, the causal mechanisms, and the exogenous variables [7]. Consequently, hard interventions can also be seen as a special case of soft interventions where the causal mechanism is set to a constant. Illustrated in Figure 1, a prominent challenge in causal representation learning lies in dealing with the ambiguity surrounding the effects of soft interventions. The observed alterations in object colors fail to distinctly elucidate whether they stem from parental influences or the applied interventions.

Additionally, a lack of comprehension regarding causal graphs can pose significant challenges in causal representation learning. In certain applications, the causal graph can be constructed using domain knowledge, allowing us to subsequently learn the causal variables [2, 18, 20]. However, this is not universally applicable, necessitating the direct learning of the causal graph itself. In a Variational AutoEncoder (VAE) framework, there are generally two approaches for causal representation learning: Explicit Latent Causal Models (ELCMs) [34, 1, 35, 37, 17, 15] and Implicit Latent Causal Models (ILCMs) [3]. In ELCMs, the latents are the causal variables and the adjacency matrix of the causal graph is parameterized and integrated into the prior of the latents such that the prior of latents is factorized according to the Causal Markov Condition [27]. This approach to causal representation learning is highly susceptible to becoming stuck in local minima as it is hard to learn representations without knowing the graph, and it is hard to learn the graph without knowing the representations. ILCMs [3] were introduced to circumvent this “chicken-and-egg” problem by using *solution functions*, which can implicitly model edges in the causal graph rather than explicitly modeling the entire adjacency matrix of the causal model. In ILCMs the *latents are the exogenous variables* and there is no explicit parameterization for the graph.

In implicit causal representation learning, the task involves recovering the exogenous variables \mathcal{E} from observed variables \mathcal{X} and learning solution functions. In [3], interventions are assumed to be hard, but this is often unrealistic and does not align with real-world problems. **In this paper, we propose a novel approach for Implicit Causal Representation Learning via Switchable Mechanisms (ICRL-SM).** We will introduce the *causal mechanism switch variable* as a way of modeling the effect of soft interventions and identifying the causal variables. Our experiments on both synthetic and large real-world datasets, highlight the efficacy of proposed method in identifying causal variables and promising future directions in implicit causal representation learning. Our key contributions can be summarized as follows:

- I.** A novel approach for implicit causal representation learning with soft interventions.
- II.** Employing causal mechanisms switch variable to model the effect of soft interventions.
- III.** Theory for identifiability up to reparameterization from soft interventions.

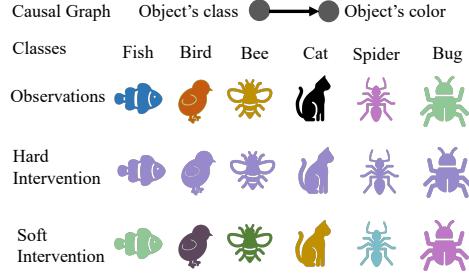


Figure 1: Difference between hard interventions and soft interventions: As seen in the middle row, hard interventions sever connections with parents. Therefore, an object’s class cannot have any effect on the object’s color when we intervene on color. On the other hand, soft interventions, as shown in the bottom row, allow for such effects.

2 Related Work

Causal representation learning has recently garnered significant attention [27, 14]. The primary challenge in this problem lies in achieving identifiability beyond the Markov equivalence class [26]. Solely relying on observational data necessitates additional assumptions regarding causal mechanisms, decoders, latent structure, and the availability of interventional data [22, 28, 36, 25, 15, 1, 40, 13, 34]. Recent works have focused on identifying causal models from collected interventional data instead of making strong assumptions about functions of the causal model. Interventional data facilitates identifiability based on relatively weak assumptions [1, 6, 3, 39, 33]. This type of data can be further categorized based on whether it involves soft or hard interventions, and whether the manipulated variables are observed and specified or latent. Our focus in this paper is on examining soft interventions, encompassing both observed and unobserved variables.

Table 1: Comparison of proposed method with other recent related work on causal learning from interventional data

Methods	Causal Mechanisms	Mixing functions	Interventions	Explicit/Implicit	Identifiability
CausalDiscrepancy [38]	Nonlinear	Full row rank polynomial	Soft	Explicit	Permutation and Affine
CauCA [33]	Nonlinear	Diffeomorphism	Soft	Explicit	Different based on assumptions
Linear-CD [29]	Linear	Linear	Hard	Explicit	Permutation
Scale-I [30]	Nonlinear	Linear	Hard/Soft	Explicit	Scale/Mixed
ILCM [3]	Nonlinear	Diffeomorphism	Hard	Implicit	Permutation and reparameterization
dVAE [21]	Nonlinear	Diffeomorphism	Hard	Implicit	Permutation and reparameterization
ICRL-SM (ours)	Nonlinear	Diffeomorphism	Soft	Implicit	Reparameterization

2.1 Explicit models vs. Implicit models

Table 1 presents a comparison of the assumptions and identifiability results between our proposed theory and other related works on causal representation learning with interventions. In causal representation learning with interventions, one approach assumes a given causal graph and concentrates on identifying causal mechanisms and mixing functions. For instance, Causal Component Analysis (CauCA) [33] explores soft interventions with a known graph. Alternatively, when the graph is not provided, explicit models seek to reconstruct it from interventional data [6, 17], potentially resulting in a chicken-and-egg problem in causal representation learning [3]. Current methods face the challenge of simultaneously learning the causal graph and other network parameters, especially in the absence of information about causal variables or the graph. Addressing these challenges, [3] recently introduced ILCM, which performs *implicit* causal representation learning exclusively using *hard* intervention data. In contrast, our approach introduces a novel method for learning an implicit model from *soft* interventions. [3] describes methods for extracting a causal graph from a learned implicit model, which could be applied to our method as well. In our experiments, we will compare our method with ILCM and dVAE [21], given their implicit nature and similar experimental settings and assumptions. Additionally, to showcase the superiority of our method over explicit models, we will employ explicit causal model discovery methods like ENCO [16] and DDS [5], in conjunction with various variants of β -VAE.

2.2 Hard interventions vs Soft interventions

The identification of explicit causal models from hard interventions has been extensively explored. [29] investigate causal disentanglement in linear causal models with linear mixing functions under hard interventions. Similarly, [4] focus on identifying causal models with linear causal mechanisms and nonlinear mixing functions, also utilizing hard interventions. In a more general setting with non-parametric causal mechanisms and mixing functions, [32] examine the identifiability of causal models, utilizing multi-environment data from unknown interventions. Similarly, [2] explore identifiability of causal models using multi-environment data from unknown interventions. [30] investigate the identifiability of causal models with nonlinear causal mechanisms and linear mixing functions, considering both hard and soft interventions.

Recent work has expanded the concept of explicit hard interventions to include soft interventions. In their study, [38] address the identification of causal models from soft interventions, leveraging the sparsity of the adjacency matrix as an inductive bias. However, when dealing with implicit models, soft interventions introduce new complexities. Identifiability becomes more challenging, as the causal effect of variables on observed variables is less apparent. This ambiguity arises from the dual possibility of effects originating from interventions or influences from parent variables on the causal variables. Moreover, in scenarios where implicit modeling is retained, the absence of knowledge about parent variables further complicates identifiability. While [3] theoretically establishes identifiability

for hard interventions, practical experiments involving complex causal models with over 10 variables reveal increased ambiguity and confounding factors. Consequently, model identification becomes less straightforward.

3 Methodology

3.1 Data Generating Process

A structural causal model (Definition A1.1) is used to understand and describe the relationships between different variables and how they influence each other through causal mechanisms. A **decoder function**, $g(\mathbf{z}) = \mathbf{x}$, maps a vector of causal values \mathbf{z} to observed values \mathbf{x} . The causal variables \mathcal{Z} are unobserved and the goal is to infer them from interventional data. For each causal variable, a **diffeomorphic solution function**, $s_i : \mathcal{E}_i \rightarrow \mathcal{Z}_i$, deterministically maps a value for exogenous variable \mathcal{E}_i to a value for causal variable \mathcal{Z}_i . In implicit modeling, we learn the solution functions s_i directly, rather than defining them through local mechanisms f_i . We write \mathcal{S} for the set of all solution functions $s_i \in \mathcal{S}$, so $\mathcal{S} : \mathcal{E} \rightarrow \mathcal{Z}$.

Identifying causal models from data can be complex and is often studied within classes of models such as those identifiable up to affine transformations. For example, in the context of nonlinear *Independent Component Analysis (ICA)*, the generative process also involves a mixture function g of latent causal variables $\mathcal{Z} \in \mathbb{R}^n$, resulting in observations $\mathcal{X} \in \mathbb{R}^n$ [15, 41]. However, a significant distinction between causal representation learning and nonlinear-ICA is that in the former, the causal variables \mathcal{Z} may have complex dependencies. Our objective in this paper is to recover \mathcal{E} from \mathcal{X} and eventually map \mathcal{E} to \mathcal{Z} using solution functions.

Identifying a causal model from observational data is not trivial and requires assumptions on the parameters of the model [1]. Adding information about interventions in addition to observations, helps to identify causal variables by exhibiting the effect of changing a causal variable on the observed variables. An interventional data point $(\mathbf{x}, \tilde{\mathbf{x}}, i)$ includes the pre-intervention observation \mathbf{x} , the post-intervention observation $\tilde{\mathbf{x}}$, and intervention target $i \in \mathcal{I}$ where \mathcal{I} is the set of intervention targets selected from the causal variables. The post-intervention data $\tilde{\mathbf{x}}$ is generated by a *soft intervention* that targets one of the causal variables in \mathcal{Z} . To achieve identifiability up to reparametrization, we rely on a series of assumptions within the data generation process, outlined as follows:

Assumption 3.1. (*Data generating assumptions*)

1. **Atomic Interventions:** For every sample $(\mathbf{x}, \tilde{\mathbf{x}}, i)$, only one causal variable is targeted by an intervention.
2. **Known Targets:** Targets of soft interventions are known.
3. **Post-intervention Exogenous Variables:** The exogenous variables' values change only for the corresponding intervened causal variable, while the others maintain their pre-intervention values, thus $e_i \neq \tilde{e}_i$ if $i \in \mathcal{I}$, and $e_i = \tilde{e}_i$ otherwise.
4. **Sufficient Variability:** Soft interventions alter causal mechanisms to introduce sufficient variability [15]. These interventions should modify causal mechanisms to ensure non-overlapping conditional distributions of causal variables (refer to Figure A1).
5. **Diffeomorphic decoder and causal mechanisms:** Diffeomorphism guarantees no information loss and avoids abrupt changes in the function's image.

The **known targets** assumption can be relaxed in applications where such data is not available and the same procedure in [3] can be used to infer the intervention targets. In fact, in our real-world experiments, intervention targets are not available and based on the nature of the datasets, we hypothesize our causal variables to be object attributes and actions to be intervention targets.

3.2 Causal Mechanisms Switch Variable

The major difference of soft intervention with hard intervention is that post-intervention causal variable $\tilde{\mathcal{Z}}_i$ is no longer disconnected from its parents and its causal mechanism \tilde{s}_i is affected by the intervention. This is why identifying the causal mechanisms is more difficult for soft interventions. Soft intervention data yield fewer constraints on the causal graph structure than hard intervention data. For more details refer to string diagrams of soft and hard interventions depicted in Figure A5. Figure 2b shows our main generative model. It includes a data augmentation step that adds the intervention displacement $\tilde{\mathbf{x}} - \mathbf{x}$ as an observed feature that directly represents the effect of a soft intervention in observation space.

Augmented implicit causal model To model the effect of soft interventions, we introduce the causal mechanism switch variable \mathcal{V} [26]. By leveraging \mathcal{V} , we can effectively switch to the pre-

intervention causal mechanisms within post-intervention data. This facilitates the model's ability to solely focus on discerning alterations in the intrinsic characteristics of each causal variable. These changes are encapsulated within their respective exogenous variables, aiding the model in learning the causal relationships more accurately. We propose to use a modulated form of \mathcal{V} to model the soft intervention effects on each causal variable as an additive effect with a nonlinear function h_i such that $\forall i, \tilde{\mathcal{Z}}_i = \tilde{s}_i(\tilde{\mathcal{E}}_i; \tilde{\mathcal{E}}_{/i}) = s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(\mathcal{V}))$. As the parental set for each causal variable is not known, we have to use a modulated form of \mathcal{V} in every causal variable's solution function and the inclusion of $h_i(\mathcal{V})$ enables the model to encompass variations in the parental sets of all causal variables in \mathcal{V} . Therefore, there is a switch variable \mathcal{V}_i for each causal variable \mathcal{Z}_i . Adding switch variables to solution functions leads to the concept of an *augmented implicit causal model*.

Definition 3.2. (Augmented Implicit Causal Models) An Augmented Implicit Causal Models (AICMs) is defined as $\mathcal{A} = (\mathcal{S}, \mathcal{Z}, \mathcal{E}, \mathcal{V})$ where $\mathcal{V} \in \mathbb{R}^n$ is the causal mechanism switch variable which models the effect of soft interventions on solution functions \mathcal{S} :

$$\forall i, \tilde{\mathcal{Z}}_i = \tilde{s}_i(\tilde{\mathcal{E}}_i; \tilde{\mathcal{E}}_{/i}) = s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(\mathcal{V})), \quad (1)$$

where \tilde{s}_i is the new solution function resulting from the soft intervention, $\tilde{\mathcal{E}}_{/i}$ is the altered set of all exogenous variables except i , including the ancestral exogenous variables, due to intervention, and $\tilde{\mathcal{E}}_i$ is the post-intervention exogenous variable.

The usage of \mathcal{V} in soft interventions is analogous to augmented networks in [23] which were mainly designed for hard interventions. Pearl [23] even foresaw this possibility by saying: "One advantage of the augmented network representation is that it is applicable to any change in the functional relationship f_i and not merely to the replacement of f_i by a constant."

By using Taylor's expansion, we can expand the solution functions as follows:

$$s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(\mathcal{V})) = s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(v_0)) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(\left. \frac{\partial^n s_i}{\partial h_i^n} \right|_{h_i=h_i(v_0)} (h_i(\mathcal{V}) - h_i(v_0))^n \right) = s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(v_0)) + R_i \quad (2)$$

where we'll use R_i as a short-hand for Equation 2. We define the **separable dependence** property for solution functions as $\exists h_i(v_0) : s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(v_0)) = s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i})$. An example of such a scenario could be in location-scale noise models such as, $s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(v)) = \tilde{e}_i + loc(e_{/i}) + h_i(v) = \tilde{e}_i + loc(e_{/i}) + v^2 + v$ where v_0 would be zero. By assuming the separable dependence property, we can write the solution function in Equation 2 as:

$$s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}, h_i(\mathcal{V})) = s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}) + R_i = s_i(\tilde{\mathcal{E}}_i; \mathcal{E}_{/i}) + \text{soft intervention effect} \quad (3)$$

As a result, we can switch to pre-intervention solution functions. Subsequently, by modeling soft intervention effects using $h_i(\mathcal{V})$, we can recover pre-intervention solution functions. During inference, we simply disregard the $h_i(\mathcal{V})$ term in the solution functions. Nonetheless, it is possible to train the prior $p(\mathcal{V})$ to ensure that the separable dependence property is maintained for pre-intervention data.

Observability of switch variable The intuition behind using \mathcal{V} is to separate the effect of soft intervention on $\tilde{\mathcal{Z}}_i$ into two: (1) The effect on causal mechanisms and parents, and (2) The effect on exogenous variable \mathcal{E}_i . For example, we can say that causal variables in images of objects are the objects' attributes such as shape, color, and size, and performing actions like "Fold" change these attributes. Furthermore, it can be asserted that the camera angle within a given image may influence the shape of the object. If the images were generated from a hard intervention, the camera angle remains fixed between pre and post intervention. However, the camera angle changes along with the performed actions indicating that the interventions are soft. In this case, if we had a knowledge of how the camera angle affects the attributes of objects, then we could separate the effect of soft intervention. In other words, if \mathcal{V} is observed, then we can extract the effect of the intervention that we are interested in (i.e., the effect on the causal variable itself). For more details, refer to Figure A4.

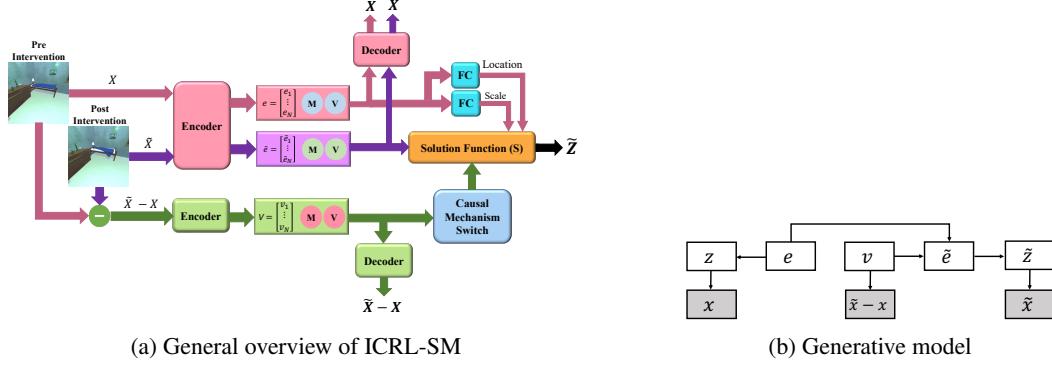
Lacking an understanding of how soft intervention influences the causal model, a more complex model becomes necessary. Consequently, the term R_i in Equation 2 would involve a higher order of $h_i(\mathcal{V})$. Therefore, we assume the observability of \mathcal{V} :

Assumption 3.3. (Observability of \mathcal{V}) Given an intervention sample (x, \tilde{x}, i) and linear decoders, we can approximate the soft intervention effects $h_i(\mathcal{V})$ as follows:

$$\tilde{z} - z = \Delta e_i + R \quad (\text{using Equation 2}), \quad \tilde{x} - x = g(\tilde{z}) - g(z) \approx g(\tilde{z} - z) = g(\Delta e_i + R),$$

where $R = [R_0, R_1, \dots, R_n]$ and n is the number of causal variables. R and Δe_i are the vectors indicating the soft intervention effects and change in effect of the exogenous variable of the intervened causal variable, respectively. Note that elements of R will be all zero except for the intervened causal variable. Consequently, with linear mixing functions and some pre-processing on observed samples (here subtraction), we can observe R_i .

Our synthetic data is generated using a linear decoder, however, the decoder for the real-world datasets is not necessarily linear. Therefore, we do not observe \mathcal{V} from $\tilde{x} - x$ in the real-world dataset. Nevertheless, our findings suggest that incorporating soft interventions through \mathcal{V} leads to superior performance compared to other implicit modeling approaches. Clearly, understanding the impact of soft interventions on the generative system of the dataset would result in improved outcomes.



3.3 Identifiability Theorem for Implicit SCMs with Soft Interventions

In this paper, our focus lies in identifying the causal variables up to reparameterization through soft interventions. We first define identifiability up to reparameterization (Definition 3.4) and subsequently introduce the identifiability theorem 3.5. The proof of theorem is extensive and is available in full in Appendix A1.

We establish identifiability up to reparameterization, allowing for the mapping of causal variables \mathcal{Z} and \mathcal{Z}' between two Latent Causal Models (\mathcal{M} and \mathcal{M}') through component-wise transformations (Definition A1.2). Given our implicit modeling approach, lacking knowledge of the causal graph, we include all exogenous variables in the solution functions, as depicted in Equation 1. Notably, **the causal graph remains unaltered during learning**. To illustrate, we contrast hard interventions, which neglect parent influences, with soft interventions that acknowledge parental effects in a simple example. Consider a basic causal model $Z_1 \rightarrow Z_2$ alongside a location-scale noise model [12] for the solution function, given by $\tilde{z}_2 = \frac{\tilde{e}_2 - \tilde{loc}(e_1)}{\tilde{scale}(e_1)}$. The distribution $p(\tilde{Z}_2)$ mean is $\frac{1}{\tilde{scale}(e_1)} \times \text{mean}(\tilde{e}_2) - \frac{\tilde{loc}(e_1)}{\tilde{scale}(e_1)}$

In the context of hard interventions, we can assume $p(\tilde{Z}_2|Z_1) = p(\tilde{Z}_2) = N(0, 1)$ as there are no parental effects. Consequently, the location and scale networks within the solution function tend to dampen parental effects, given the absence of parental influence in the ground-truth data. Contrarily, soft interventions exhibit parental influence in the ground-truth data, thus $p(\tilde{Z}_2|Z_1) \neq N(0, 1)$. Due to the lack of parental knowledge in implicit modeling, we model $p(\tilde{Z}_2|Z_1) = p(\tilde{Z}_2|\mathcal{E}_2)$, as \mathcal{E}_2 is a known parent of \tilde{Z}_2 . Consequently, parental effects are propagated to \mathcal{E}_i (the corresponding exogenous variable of each causal variable), violating identifiability up to reparameterization. By leveraging \mathcal{V} , we allow parental effects to propagate to \mathcal{V} instead of \mathcal{E}_i .

Definition 3.4. (Equivalence up to component-wise reparameterization) Let $\mathcal{M} = (\mathcal{A}, \mathcal{X}, g, \mathcal{I})$ and $\mathcal{M}' = (\mathcal{A}', \mathcal{X}, g', \mathcal{I})$ be two Latent Causal Models (LCM) based on AICMs $\mathcal{A}, \mathcal{A}'$ with shared observation space \mathcal{X} , shared intervention targets \mathcal{I} , and respective decoders g and g' . We say that \mathcal{M} and \mathcal{M}' are equivalent up to component-wise reparameterization $\mathcal{M} \sim_r \mathcal{M}'$ if there exists a component-wise transformation (Definition A1.2) $\phi_{\mathcal{Z}}$ from the causal variables \mathcal{Z} to the causal variables \mathcal{Z}' and a component-wise transformation $\phi_{\mathcal{E}}$ between \mathcal{E} and \mathcal{E}' such that:

1. Indices are preserved (i.e., $\phi_i(z_i) = z'_i$ and $\phi_i(e_i) = e'_i$). Corresponding edges are preserved (i.e., $\mathcal{Z}_i \rightarrow \mathcal{Z}_j$ holds in \mathcal{G} iff $\mathcal{Z}'_i \rightarrow \mathcal{Z}'_j$ holds in \mathcal{G}' . Edges $\mathcal{E}_i \rightarrow \mathcal{Z}_i$ should be preserved as well.)
2. The exogenous transformation preserves the probability measure on exogenous variables $p_{\mathcal{E}'} = (\phi_{\mathcal{E}})_* p_{\mathcal{E}}$ (Definition A1.4).

3. The causal transformation preserves the probability measure on causal variables $p_{\mathcal{Z}'} = (\phi_{\mathcal{Z}})_* p_{\mathcal{Z}}$ (Definition A1.4).

Theorem 3.5. (Identifiability of latent causal models.) Let $\mathcal{M} = (\mathcal{A}, \mathcal{X}, g, \mathcal{I})$ and $\mathcal{M}' = (\mathcal{A}', \mathcal{X}, g', \mathcal{I})$ be two LCMs with shared observation space \mathcal{X} and shared intervention targets \mathcal{I} . Suppose the following conditions are satisfied:

1. Data generating assumptions explained in Assumption 3.1.
 2. Soft interventions satisfy Assumption 3.3.
 3. The causal and exogenous variables are real-valued.
 4. The causal and exogenous variables follow a multivariate normal distribution.
- Then the following statements are equivalent:
- Two LCMs \mathcal{M} and \mathcal{M}' assign the same likelihood to interventional and observational data i.e., $p_{\mathcal{M}}^{\mathcal{X}, \mathcal{I}}(x, \tilde{x}, i) = p_{\mathcal{M}'}^{\mathcal{X}, \mathcal{I}}(x, \tilde{x}, i)$.
 - \mathcal{M} and \mathcal{M}' are disentangled, that is $\mathcal{M} \sim_r \mathcal{M}'$ according to Definition 3.4.

3.4 Training Objective

Consequently, there will be three latent variables in ICRL-SM:

1. A causal mechanism switch variable \mathcal{V} .
2. The pre-intervention exogenous variables \mathcal{E} .
3. The post-intervention exogenous variables $\tilde{\mathcal{E}}$.

As the data log-likelihood $\log p(x, \tilde{x}, x - \tilde{x}) \equiv \log p(x, \tilde{x})$ is intractable, we utilize an ELBO approximation as training objective:

$$\begin{aligned} \log p(x, \tilde{x}) &\geq E_{q(e, \tilde{e}, v|x, \tilde{x})} [\log p(x, \tilde{x}|e, \tilde{e}, v)] - KLD(q(e, \tilde{e}, v|x, \tilde{x})||p(e, \tilde{e}, v)) \\ &= E_{q(v|\tilde{x}-x) \cdot q(e|x) \cdot q(\tilde{e}|\tilde{x})} [\log(p(x|e)p(\tilde{x}|\tilde{e})p(\tilde{x}-x|v))] - KLD(q(v|\tilde{x}-x) \cdot q(e|x) \cdot q(\tilde{e}|\tilde{x})||p(\tilde{e}|e, v)p(v)p(e)). \end{aligned} \quad (4)$$

The observations are encoded and decoded independently. The KLD term regularizes the encodings to share the latent *intervention model* $p(\tilde{e}|e, v)p(v)p(e)$ that is shared across all data points. The components of this model can be interpreted as follows:

1. $p(e)$ is the prior distribution over exogenous variables e .
2. $p(v)$ is the prior distribution over switch variables v .
3. $p(\tilde{e}|e, v)$ is a transition model that shows how the exogeneous variables change as a function of the intervention.

We factorize the posterior with a mean-field approximation $q(v, e, \tilde{e}|x, \tilde{x}) = q(v|\tilde{x}-x) \cdot q(e|x) \cdot q(\tilde{e}|\tilde{x})$ and, following our data generation model (Figure 2b), the reconstruction probability as $p(x, \tilde{x}|e, \tilde{e}, v) = p(x|e)p(\tilde{x}|\tilde{e})p(\tilde{x}-x|v)$. The prior over latent variables is factorized as $p(\tilde{e}, e, v) = p(\tilde{e}|e, v)p(v)p(e)$ (Figure 2b). Pre-intervention exogenous variables are mutually independent, hence, $p(e) = \prod_i p(e_i)$ and $p(v) = \prod_i p(v_i)$. We assume $p(e_i)$ and $p(v_i)$ to be standard Gaussian. Furthermore, as we assume $e_i = \tilde{e}_i$ for all non-intervened variables, the $p(\tilde{e}|e, v)$ will be as follows:

$$p(\tilde{e}|e, v) = \prod_{i \notin I} \delta(\tilde{e}_i - e_i) \prod_{i \in I} p(\tilde{e}_i|e, v) = \prod_{i \notin I} \delta(\tilde{e}_i - e_i) \prod_{i \in I} p(\tilde{z}_i|e_i) \left| \frac{\partial \tilde{z}_i}{\partial \tilde{e}_i} \right| \quad (5)$$

The last equality is obtained from the Change of Variable Rule in probability theory, applied to the solution function $\tilde{z}_i = s_i(\tilde{e}_i; e_{/i}, h_i(v))$. Furthermore, we write $p(\tilde{z}_i|e, v) = p(\tilde{z}_i|e_i)$ since only e_i is a known parent of \tilde{z}_i in implicit modeling. We assume $p(\tilde{z}_i|e_i)$ to be a Gaussian whose mean is determined by e_i . We implement the solution function using a location-scale noise models [12] as also practiced in [3], which defines an invertible diffeomorphism. For simplicity, in our experiments, we are only going to change the *loc* network in post-intervention. Therefore, $h_i(v)$ will be used as:

$$\tilde{z}_i = \tilde{s}_i(\tilde{e}_i; e_{/i}, h_i(v)) = \frac{\tilde{e}_i - (loc_i(e_{/i}) + h_i(v))}{scale_i(e_{/i})}, \quad (6)$$

where $loc_i : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ and $scale_i : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ are fully connected networks calculating the first and second moments, respectively. The general overview of the model is illustrated in Figure 2a.

4 Experiments and Results

The experiments conducted in this paper address two downstream tasks; (1) Causal Disentanglement to identify the true causal graph from pairs of observations (x, \tilde{x}, i) , and (2) Action Inference to make

supervised inferences about actions generated from the post-intervention samples using information about the values of the manipulated causal variables. Moreover, we conducted additional experiments designed as an ablation study, the results of which are presented in A4. All models are trained using the same setting and data with known intervention targets.

4.1 Datasets

Synthetic Dataset We generate simple synthetic datasets with $\mathcal{X} = \mathcal{Z} = \mathbb{R}^n$. For each value of n , we generate ten random DAGs, a random location-scale SCM, then a random dataset from the parameterized SCM. To generate random DAGs, each edge is sampled in a fixed topological order from a Bernoulli distribution with probability 0.5. The pre-intervention and post-intervention causal variables are obtained as:

$$z_i = \text{scale}(z_{pa_i})e_i + \text{loc}(z_{pa_i}) \xrightarrow{\text{Soft-Intervention}} \tilde{z}_i = \text{scale}(z_{pa_i})\tilde{e}_i + \widetilde{\text{loc}}(z_{pa_i}), \quad (7)$$

where the *loc* and *scale* networks are changed in post intervention. The pre-intervention *loc* and post-intervention $\widetilde{\text{loc}}$ network weights are initialized with samples drawn from $\mathcal{N}(0, 1)$ and $\mathcal{N}(3, 1)$, respectively. The *scale* is constant 1 for both pre-intervention and post-intervention samples. Both e_i and \tilde{e}_i are sampled from a standard Gaussian. The causal variables are mapped to the data space through a randomly sampled $SO(n)$ rotation. For each dataset, we generate 100,000 training samples, 10,000 validation samples, and 10,000 test samples.

Action Datasets Causal-Triplet datasets tailored for *actionable* counterfactuals [19] feature paired images where several global scene properties may vary including camera view and object occlusions. Thus, the images can be viewed as outcomes of soft interventions, wherein actions affect objects alongside subtle alterations. These datasets [19] consist of: images obtained from a photo-realistic simulator of embodied agents, ProcTHOR [9], and the other contains images repurposed from a real-world video dataset of human-object interactions [8]. The former one contains 100 k images in which 7 types of actions manipulate 24 types of objects in 10 k distinct ProcTHOR indoor environments. The latter consists of 2,632 image pairs, collected under a similar setup from the Epic-Kitchens dataset with 97 actions manipulating 277 objects. Based on the nature of actions in this dataset, the causal variables should represent attributes of objects such as shape and color. As the dataset consists of images we train all the methods with ResNet encoder and decoder. For the ProcThor dataset the number of causal variables are 7. For the Epic-Kitchens dataset, we randomly chose 20 actions from the dataset as 97 causal variables will be too complex in a VAE setup.

4.2 Metrics

For the causal disentanglement task, we are going to use the DCI scores [10]. Causal disentanglement score quantifies the degree to which \mathcal{Z}_i factorises or disentangles the \mathcal{Z}^* . Causal disentanglement D_i for \mathcal{Z}_i is calculated as $D_i = (1 - H_K(P_{i.})) = (1 + \sum_{k=0}^{K-1} P_{ik} \log_K P_{ik})$ where $P_{ij} = \frac{R_{ij}}{\sum_{k=0}^{K-1} R_{ik}}$ and R_{ij} denotes the probability of \mathcal{Z}_i being important for predicting \mathcal{Z}_j^* . Total causal disentanglement is the weighted average $\sum_i \rho_i D_i$ where $\rho_i = \frac{\sum_j R_{ij}}{\sum_{ij} R_{ij}}$. Causal Completeness quantifies the degree to which each \mathcal{Z}_i^* is captured by a single \mathcal{Z}_i . Causal completeness is calculated as $C_j = (1 - H_D(\tilde{P}_{.j})) = (1 + \sum_{d=0}^{D-1} \tilde{P}_{dj} \log_D \tilde{P}_{dj})$. D and K here are equal to the dimension of \mathcal{Z}^* and \mathcal{Z} which is n . For the action inference task, we will use classification accuracy as a metric. As we assume intervention targets are known, we train all models using known intervention targets for a fair comparison.

5 Results

5.1 Causal Disentanglement

We generated a dataset for the soft interventions and trained the models of ICRL-SM, ILCM, β -VAE and D-VAE for 10 different seeds, which generated 10 different causal graphs. We selected 4 causal variables to encompass complex causal structures, including forks, chains, and colliders. Table 2 displays the Causal Disentanglement and Causal Completeness scores for all models, computed on the test data.

The results in Table 2 indicate that our method ICRL-SM can identify the true causal graph in most cases. The worst results are seen for graphs $G5$ and $G10$. As mentioned in [27, 25], causal graphs are sparse and in the $G5$ case, where the graph is fully connected, the proposed method cannot identify the causal variables well. Furthermore, in the next experiment we are going to examine the factors affecting causal disentanglement such as the number of edges in the graph and the intensity of soft

Table 2: Comparison of identifiability results

Model	Name	Causal Disentanglement				Causal Completeness			
		β -VAE	d -VAE	ILCM	ICRL-SM	β -VAE	d -VAE	ILCM	ICRL-SM
	G1	0.38	0.54	0.71	0.82	0.51	0.69	0.78	0.87
	G2	0.30	0.72	0.75	0.83	0.49	0.77	0.80	0.87
	G3	0.28	0.51	0.68	0.98	0.49	0.56	0.78	0.98
	G4	0.16	0.50	0.65	0.68	0.38	0.69	0.77	0.78
	G5	0.27	0.44	0.53	0.42	0.45	0.54	0.66	0.50
	G6	0.52	0.62	0.71	0.98	0.66	0.69	0.86	0.98
	G7	0.39	0.49	0.71	0.75	0.70	0.73	0.89	0.89
	G8	0.47	0.54	0.50	0.59	0.6	0.63	0.62	0.68
	G9	0.30	0.68	0.83	0.85	0.40	0.76	0.86	0.87
	G10	0.39	0.39	0.52	0.32	0.53	0.56	0.82	0.70

intervention effect. These findings can explain why ICRL-SM cannot identify causal variables in G_{10} despite its sparsity.

Table 3: Table comparing action and object accuracy across various methods on Causal-Triplet datasets under different settings. Z and z_i show whether all causal variables (Z), or only the intervened causal variable (z_i) are used for the prediction task. R_{64} denote images with resolutions 64×64 .

Method	Epic-Kitchens			ProcTHOR		
	Action Accuracy		Object Accuracy	Action Accuracy		Object Accuracy
	$Z; R_{64}$	$z_i; R_{64}$	$Z; R_{64}$	$z_i; R_{64}$	$Z; R_{64}$	$z_i; R_{64}$
β -VAE [11]	0.27	0.18	0.19	0.06	0.39	0.30
d -VAE [21]	0.19	0.69	0.20	0.17	0.35	0.81
ILCM [3]	0.21	0.59	0.14	0.14	0.30	0.70
ICRL-SM (ours)	0.16	0.86	0.16	0.18	0.28	0.93
					0.40	0.82

5.2 Factors Affecting Causal Disentanglement

In this experiment, we consider the graph G_3 , which has the best identifiability, and change the intensity of soft intervention and number of edges in its data generation process. To change the intensity, the post-intervention \tilde{loc} network weights are initialized with samples drawn from $N(1, 1)$ (almost similar to loc) and $N(10, 1)$ (significantly different from loc). To change the number of edges, we consider a chain and fully-connected graph.

Table 4: Left table depicts the action and object accuracy of three explicit models, with experiments conducted applying an image with resolution of R_{64} as the input to the Resnet50 encoder with the intervened causal variable (z_i). Right table shows the comparison of ICRL-SM performance on different configurations of G_5

Datasets	Methods	Action Accuracy	Object Accuracy	Edges	Post-intervention causal mechanism	Causal Disentanglement	Causal Completeness
Epic-Kitchens	ENCO [16]	0.69	0.13	Chain	Default	0.98	0.98
	DDS [5]	0.44	0.09	Full	Default	0.89	0.89
	Fixed-order	0.79	0.14	Default	Significantly different	0.68	0.73
	ICRL-SM (ours)	0.86	0.18	Default	Almost similar	0.85	0.86
ProcTHOR	ENCO [16]	0.45	0.53				
	DDS [5]	0.64	0.67				
	Fixed-order	0.65	0.54				
	ICRL-SM (ours)	0.93	0.82				

The results in Table 4 further confirms the sparsity of causal graphs as the causal disentanglement is much worse in the fully-connected graph than the default graph of G_3 . The result for significantly different post-intervention causal mechanisms indicate that the switch variable cannot approximate intense effects of soft intervention and more supervision is required to observe \mathcal{V} . Similar post-intervention causal mechanisms also do not have sufficient variability to disentangle the causal variables as mentioned in Theory 3.5.

5.3 Action Inference

In this experiment, we show the performance of ICRL-SM in the real-world Causal-Triplet datasets. In these datasets \mathcal{V} i.e., soft intervention effects, are not directly observable. Nevertheless, our findings suggest that incorporating soft interventions through \mathcal{V} leads to superior performance compared to

other implicit modeling approaches. Clearly, understanding the impact of soft interventions on the generative system of the dataset would result in improved outcomes.

The results in Table 3 indicate that when including all causal variables to predict actions, ICRL-SM performs at par with the baseline methods. However, including all causal variables in the action or object inference may cause spurious correlations. Therefore, we have also experimented with including only the related causal variable in action and object inference. In this setting, ICRL-SM significantly outperforms the baseline methods which means that it can better disentangle the causal variables. We have also compared ICRL-SM with explicit causal representation learning methods. ENCO [16] and DDS [5] have variable topological order of causal variables during training. Furthermore, we have included a specific setting where the topological order is fixed during training. As shown in Table 4, our proposed method has superior performance to explicit models as well.

6 Conclusion

ICRL-SM, our novel model, enhances implicit causal representation learning during soft interventions by introducing a causal mechanism switch variable. Evaluations on synthetic and real-world datasets demonstrate ICRL-SM’s superiority over state-of-the-art methods, highlighting its practical effectiveness. Our findings emphasize ICRL-SM’s ability to discern causal models from soft interventions, marking it as a promising avenue for future research.

References

- [1] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 372–407. PMLR, 2023.
- [2] Shayan Shirahmad Gale Bagi, Zahra Gharaee, Oliver Schulte, and Mark Crowley. Generative causal representation learning for out-of-distribution motion forecasting. In *International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 31596–31612. PMLR, 2023.
- [3] Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *NeurIPS*, 2022.
- [4] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing, 2023.
- [5] Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable DAG sampling. In *The Tenth International Conference on Learning Representations, ICLR*. OpenReview.net, 2022.
- [6] Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data, 2013.
- [7] Juan D. Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020.
- [8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis.*, 130(1):33–55, 2022.
- [9] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.
- [10] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, ICLR*, 2018.
- [11] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR*, 2017.
- [12] Alexander Immer, Christoph Schultheiss, Julia E. Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 14316–14332. PMLR, 2023.
- [13] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020.

- [14] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *CoRR*, abs/2206.15475, 2022.
- [15] Sébastien Lachapelle, Pau Rodríguez, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *1st Conference on Causal Learning and Reasoning, CleaR*, volume 177 of *Proceedings of Machine Learning Research*, pages 428–484. PMLR, 2022.
- [16] Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. In *The Tenth International Conference on Learning Representations, ICLR*. OpenReview.net, 2022.
- [17] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning, ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 13557–13603. PMLR, 2022.
- [18] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6155–6170. Curran Associates, Inc., 2021.
- [19] Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Conference on Causal Learning and Reasoning, CleaR*, volume 213 of *Proceedings of Machine Learning Research*, pages 553–573. PMLR, 2023.
- [20] Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022*, pages 17060–17071. IEEE, 2022.
- [21] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 2020.
- [22] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.
- [23] Judea Pearl. *Causality*, cambridge university press (2000). *Artif. Intell.*, 169(2):174–179, 2005.
- [24] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley and Sons, 2016.
- [25] Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *NeurIPS*, 2022.
- [26] Bernhard Schölkopf. Causality for machine learning. *CoRR*, abs/1911.10500, 2019.
- [27] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- [28] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *J. Mach. Learn. Res.*, 23:241:1–241:55, 2022.
- [29] Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions, 2023.
- [30] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions, 2023.
- [31] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16451–16467. Curran Associates, Inc., 2021.
- [32] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions, 2023.
- [33] Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis, 2023.

- [34] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 9593–9602. Computer Vision Foundation / IEEE, 2021.
- [35] Shuai Yang, Kui Yu, Fuyuan Cao, Lin Liu, Hao Wang, and Jiuyong Li. Learning causal representations for robust domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.
- [36] Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. Causality-based feature selection: Methods and evaluations. *ACM Comput. Surv.*, 53(5), 2020.
- [37] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 2019.
- [38] Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions, 2023.
- [39] Jiaqi Zhang, Chandler Squires, Kristjan H. Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *CoRR*, abs/2307.06250, 2023.
- [40] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with NO TEARS: continuous optimization for structure learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems NeurIPS*, pages 9492–9503, 2018.
- [41] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: sparsity and beyond. In *NeurIPS*, 2022.

Appendix

A1 Proof of Identifiability Theorem

In order to prove our model is identifiable we need a two additional definitions and some previously stated assumptions.

Definition A1.1. Structural Causal Models

A structural causal model (SCM) is a tuple $\mathcal{C} = (\mathcal{F}, \mathcal{Z}, \mathcal{E}, \mathcal{G})$ with the following components:

1. The domain of causal variables $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \dots \times \mathcal{Z}_n$.
2. The domain of exogenous variables $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 \times \dots \times \mathcal{E}_n$.
3. A directed acyclic graph $\mathcal{G}(\mathcal{C})$ over the causal and exogenous variables.
4. A causal mechanism $f_i \in \mathcal{F}$ which maps an assignment of parent values for the parents \mathcal{Z}_{pa_i} plus an exogenous variable value for \mathcal{E}_i to a value of causal variable Z_i .

Definition A1.2. (Component-wise Transformation) Let ϕ be a transformation (1-1 onto mapping) between product spaces $\phi : \prod_{i=1}^n \mathcal{X}_i \rightarrow \prod_{i=1}^n \mathcal{Y}_i$. If there exist local transformations ϕ_i such that $\forall i, j, \forall x, \phi(x_1, x_2, \dots, x_n)_i = \phi_i(x_j)$, then ϕ is a component-wise transformation.

Definition A1.3. (Diffeomorphism) A diffeomorphism between smooth manifolds M and N is a bijective map $f : M \rightarrow N$, which is smooth and has a smooth inverse. Diffeomorphisms preserve information as they are invertible transformations without discontinuous changes in their image.

Definition A1.4. (Pushforward measure) Given a measurable function $f : A \rightarrow B$ between two measurable spaces A and B , and a measure p defined on A , the pushforward measure $f_* p$ on B is defined for measurable sets E in B as:

$$(f_* p)(E) = p(f^{-1}(E))$$

where $*$ denotes the pushforward operation. In other words, the pushforward measure $f_* p$ assigns a measure to a set in B by measuring the pre-image of that set under f in the space A .

Lemma A1.5. The transformation $\phi_{\mathcal{Z}} : \mathcal{Z} \rightarrow \mathcal{Z}'$ between the causal variable of two LCMs \mathcal{M} and \mathcal{M}' defined in Definition 3.4 is a component-wise transformation, if $\forall i, j, i \neq j \quad \tilde{\mathcal{E}}'_i \perp\!\!\!\perp \tilde{\mathcal{E}}'_j$ and the causal variables follow a multivariate normal distribution conditional on the pre-intervention exogenous variables where \tilde{E}'_i denote the post-intervention exogenous variable of causal variable i in \mathcal{M}' .

proof: We consider the case where the exogenous variables are mapped to causal variables by a location-scale noise model such that $\tilde{z}_i = \frac{\tilde{e}_i - \text{loc}(e_{/i})}{\text{scale}(e_{/i})}$.

$$\forall i, j, i \neq j \quad \tilde{\mathcal{E}}'_i \perp\!\!\!\perp \tilde{\mathcal{E}}'_j \rightarrow E[\tilde{\mathcal{E}}'_i \tilde{\mathcal{E}}'_j] = E[\tilde{\mathcal{E}}'_i] E[\tilde{\mathcal{E}}'_j]$$

let's add these three constants $-E[\tilde{\mathcal{E}}'_i]\widetilde{loc'_j}(e'_{/j})$, $-E[\tilde{\mathcal{E}}'_j]\widetilde{loc'_i}(e'_{/i})$, $\widetilde{loc'_i}(e'_{/i})\widetilde{loc'_j}(e'_{/j})$ to the both sides of the equality and then divide both sides by $\widetilde{scale'_i}(e'_{/i})\widetilde{scale'_j}(e'_{/j})$:

$$\begin{aligned}
& E \left[\frac{\tilde{\mathcal{E}}'_i \tilde{\mathcal{E}}'_j - \tilde{\mathcal{E}}'_i \widetilde{loc'_j}(e'_{/j}) - \tilde{\mathcal{E}}'_j \widetilde{loc'_i}(e'_{/i}) + \widetilde{loc'_i}(e'_{/i})\widetilde{loc'_j}(e'_{/j})}{\widetilde{scale'_i}(e'_{/i})\widetilde{scale'_j}(e'_{/j})} \right] = \\
& \frac{E[\tilde{\mathcal{E}}'_i]E[\tilde{\mathcal{E}}'_j] - E[\tilde{\mathcal{E}}'_i]\widetilde{loc'_j}(e'_{/j}) - E[\tilde{\mathcal{E}}'_j]\widetilde{loc'_i}(e'_{/i}) + \widetilde{loc'_i}(e'_{/i})\widetilde{loc'_j}(e'_{/j})}{\widetilde{scale'_i}(e'_{/i})\widetilde{scale'_j}(e'_{/j})} \\
& \rightarrow E \left[\left(\frac{\tilde{\mathcal{E}}'_i - \widetilde{loc'_i}(e'_{/i})}{\widetilde{scale'_i}(e'_{/i})} \right) \left(\frac{\tilde{\mathcal{E}}'_j - \widetilde{loc'_j}(e'_{/j})}{\widetilde{scale'_j}(e'_{/j})} \right) \right] = \left(\frac{E[\tilde{\mathcal{E}}'_i] - \widetilde{loc'_i}(e'_{/i})}{\widetilde{scale'_i}(e'_{/i})} \right) \left(\frac{E[\tilde{\mathcal{E}}'_j] - \widetilde{loc'_j}(e'_{/j})}{\widetilde{scale'_j}(e'_{/j})} \right) \\
& \rightarrow E[\tilde{\mathcal{Z}}'_i \tilde{\mathcal{Z}}'_j | \mathcal{E}'] = E[\tilde{\mathcal{Z}}'_i | \mathcal{E}'] E[\tilde{\mathcal{Z}}'_j | \mathcal{E}'] \\
& \rightarrow E[\tilde{\mathcal{Z}}'_i \tilde{\mathcal{Z}}'_j | \mathcal{E}'] - E[\tilde{\mathcal{Z}}'_i | \mathcal{E}] E[\tilde{\mathcal{Z}}'_j | \mathcal{E}] = 0 \\
& \rightarrow E[\tilde{\mathcal{Z}}'_i \tilde{\mathcal{Z}}'_j | \mathcal{E}'] - E[\tilde{\mathcal{Z}}'_i | \mathcal{E}'] E[\tilde{\mathcal{Z}}'_j | \mathcal{E}'] - E[\tilde{\mathcal{Z}}'_i | \mathcal{E}'] E[\tilde{\mathcal{Z}}'_j | \mathcal{E}'] + E[\tilde{\mathcal{Z}}'_i | \mathcal{E}'] E[\tilde{\mathcal{Z}}'_j | \mathcal{E}'] = 0 \\
& \rightarrow E[\tilde{\mathcal{Z}}'_i \tilde{\mathcal{Z}}'_j | \mathcal{E}'] - E[\tilde{\mathcal{Z}}'_j E[\tilde{\mathcal{Z}}'_i | \mathcal{E}'] | \mathcal{E}'] - E[\tilde{\mathcal{Z}}'_i E[\tilde{\mathcal{Z}}'_j | \mathcal{E}'] | \mathcal{E}'] + E[\tilde{\mathcal{Z}}'_i | \mathcal{E}'] E[\tilde{\mathcal{Z}}'_j | \mathcal{E}'] = 0 \\
& \rightarrow E[(\tilde{\mathcal{Z}}'_i - E[\tilde{\mathcal{Z}}'_i | \mathcal{E}'])(\tilde{\mathcal{Z}}'_j - E[\tilde{\mathcal{Z}}'_j | \mathcal{E}']) | \mathcal{E}'] = 0 \\
& \rightarrow cov(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \mathcal{E}') = 0
\end{aligned}$$

Typically, the aforementioned equalities would be valid for any diffeomorphic solution function $\tilde{s}_i : \tilde{\mathcal{E}}_i \rightarrow \tilde{\mathcal{Z}}_i$. However, in this paper, we specifically focus on solution functions represented by a location-scale noise model.

Assuming that the causal variables follow a **multivariate normal distribution conditional on the pre-intervention exogenous variables**, $cov(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \mathcal{E}') = 0$ would imply that $\tilde{\mathcal{Z}}'_i \perp\!\!\!\perp \tilde{\mathcal{Z}}'_j | \mathcal{E}'$. Let's define $\phi_{\mathcal{E}} = g'^{-1} \circ g : \mathcal{E} \rightarrow \mathcal{E}'$ where g and g' are the decoders in \mathcal{M} and \mathcal{M}' . As stated in Assumption 3.1, the decoders are diffeomorphism, hence, $\phi_{\mathcal{E}}$ is a diffeomorphism. Furthermore, let's denote \tilde{s} as the set of all solution functions in post-intervention which are also diffeomorphism as stated in Assumption 3.1. Consequently:

$$\begin{aligned}
& (\phi_{\mathcal{E}}^{-1} \text{ is diffeomorphic}) \forall i, j, i \neq j \quad \tilde{\mathcal{Z}}'_i \perp\!\!\!\perp \tilde{\mathcal{Z}}'_j | \mathcal{E}' \rightarrow \tilde{\mathcal{Z}}'_i \perp\!\!\!\perp \tilde{\mathcal{Z}}'_j | \phi_{\mathcal{E}}^{-1}(\mathcal{E}') \rightarrow \tilde{\mathcal{Z}}'_i \perp\!\!\!\perp \tilde{\mathcal{Z}}'_j | \mathcal{E} \\
& \rightarrow p(\tilde{\mathcal{Z}}'_i | \mathcal{E}) p(\tilde{\mathcal{Z}}'_j | \mathcal{E}) = p(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \mathcal{E}) \\
& (\text{all functions in } \tilde{s} \text{ are diffeomorphism}) \rightarrow p(\tilde{\mathcal{Z}}'_i | \tilde{s}(\mathcal{E})) p(\tilde{\mathcal{Z}}'_j | \tilde{s}(\mathcal{E})) = p(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \tilde{s}(\mathcal{E})) \\
& \rightarrow p(\tilde{\mathcal{Z}}'_i | \tilde{\mathcal{Z}}) p(\tilde{\mathcal{Z}}'_j | \tilde{\mathcal{Z}}) = p(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \tilde{\mathcal{Z}})
\end{aligned}$$

The association between $\tilde{\mathcal{Z}}'$ and $\tilde{\mathcal{Z}}$ arises from their shared observation space. We know that every causal variable in \mathcal{M}' depends at least on one of the causal variables in \mathcal{M} . If one of the causal variables in \mathcal{M}' depended on more than one causal variable in \mathcal{M} , it would create dependency between two variables in \mathcal{M}' and violate the above equality. Therefore, no variable in \mathcal{M}' depends on more than one causal variable in \mathcal{M} . Consequently, the transformation $\phi_{\mathcal{Z}}$ is a component-wise transformation.

Theorem A1.6. (Identifiability of latent causal models.) Let $\mathcal{M} = (\mathcal{A}, \mathcal{X}, g, \mathcal{I})$ and $\mathcal{M}' = (\mathcal{A}', \mathcal{X}, g', \mathcal{I})$ be two LCMs with shared observation space \mathcal{X} and shared intervention targets \mathcal{I} . Suppose the following conditions are satisfied:

1. Identical correspondence assumptions explained in 3.1.
2. Soft interventions satisfy Assumption 3.3.
3. The causal and exogenous variables are real-valued.
4. The causal and exogenous variables follow a multivariate normal distribution.

Then the following statements are equivalent:

-Two LCMs \mathcal{M} and \mathcal{M}' assign the same likelihood to interventional and observational data i.e.,

$$p_{\mathcal{M}}^{\mathcal{X}}(x, \tilde{x}) = p_{\mathcal{M}'}^{\mathcal{X}'}(x, \tilde{x}).$$

- \mathcal{M} and \mathcal{M}' are disentangled, that is $\mathcal{M} \sim_r \mathcal{M}'$ according to Definition 3.4.

Proof We will proceed to prove the equivalence between statements 1 and 2 by showing the implication is true in each direction.

A1.1 $\mathcal{M} \sim_r \mathcal{M}' \Rightarrow p_{\mathcal{M}}^{\mathcal{X}}(x, \tilde{x}) = p_{\mathcal{M}'}^{\mathcal{X}'}(x, \tilde{x})$

This direction is fairly straightforward. According to Definition 3.4, the fact that $M \sim_r M'$ implies that $\phi_{\mathcal{E}}$ is measure preserving. Therefore, $p_{\mathcal{M}'}^{\mathcal{E}}(e', \tilde{e}') = (\phi_{\mathcal{E}})_* p_{\mathcal{M}}^{\mathcal{E}}(e, \tilde{e})$. Furthermore, considering that ancestry is preserved, $\phi_{\mathcal{Z}}$ is measure preserving, and that causal variables are obtained from their ancestral exogenous variables in implicit models, we have $p_{\mathcal{M}'}^{\mathcal{Z}}(z', \tilde{z}') = (\phi_{\mathcal{Z}})_* p_{\mathcal{M}}^{\mathcal{Z}}(z, \tilde{z})$. Since models are trained to maximize the log likelihood of $p(x, \tilde{x}, \tilde{x} - x)$ and the latent spaces in M and M' have the same distribution, the decoders should yield the same observational distributions $p_{\mathcal{M}}^{\mathcal{X}}(x, \tilde{x}) = p_{\mathcal{M}'}^{\mathcal{X}'}(x, \tilde{x})$.

A1.2 $p_{\mathcal{M}}^{\mathcal{X}}(x, \tilde{x}) = p_{\mathcal{M}'}^{\mathcal{X}'}(x, \tilde{x}) \Rightarrow \mathcal{M} \sim_r \mathcal{M}'$

Let's define $\phi_{\mathcal{E}} = g'^{-1} \circ g : \mathcal{E} \rightarrow \mathcal{E}'$. Since we can express $e = s^{-1}(z)$, we can now define $\phi_{\mathcal{Z}}$ as

$$\phi_{\mathcal{Z}} = s' \circ g'^{-1} \circ g \circ s^{-1} : \mathcal{Z} \rightarrow \mathcal{Z}'. \quad (8)$$

Therefore, $\phi_{\mathcal{E}} = s'^{-1} \circ \phi_{\mathcal{Z}} \circ s$. Because g and g' are **diffeomorphisms**, $\phi_{\mathcal{E}}$ is a diffeomorphism as well. Furthermore, since $p_{\mathcal{M}}^{\mathcal{X}} = p_{\mathcal{M}'}^{\mathcal{X}'}$ and $\phi_{\mathcal{E}}$ is a diffeomorphism, then $p_{\mathcal{M}'}^{\mathcal{E}} = (\phi_{\mathcal{E}})_* p_{\mathcal{M}}^{\mathcal{E}}$. Consequently, $\phi_{\mathcal{E}}$ is measure-preserving. Similarly, $\phi_{\mathcal{Z}}$ is measure-preserving as well since causal mechanisms are **diffeomorphisms**.

Step 1: Identical correspondence of edges and nodes Let's define the set U as $U = \{\mathcal{E} \times \mathcal{E} | \forall I, J \in \mathcal{I} : \text{supp } p_{\mathcal{M}}^{\mathcal{E}, \mathcal{I}}(e, \tilde{e}|I) \cap \text{supp } p_{\mathcal{M}'}^{\mathcal{E}, \mathcal{I}}(e, \tilde{e}|J)\}$. Then, assuming **atomic interventions** and **counterfactual exogenous variables**, $p_{\mathcal{M}}^{\mathcal{E}, \mathcal{I}}(U|I) = p_{\mathcal{M}'}^{\mathcal{E}, \mathcal{I}}(U|J) = 0$. Therefore, we can say that $p_{\mathcal{M}}^{\mathcal{E}}(e, \tilde{e}) = \sum_{I \in \mathcal{I}} p_{\mathcal{M}}^{\mathcal{E}, \mathcal{I}}(e, \tilde{e}|I) p_{\mathcal{M}}^{\mathcal{I}}(I)$ is a discrete mixture of non-overlapping distributions $p_{\mathcal{M}}^{\mathcal{E}, \mathcal{I}}(e, \tilde{e}|I)$. Similarly, we can say that $p_{\mathcal{M}'}^{\mathcal{E}}(e, \tilde{e})$ is a discrete mixture of non-overlapping distributions. It can be concluded that as $\phi_{\mathcal{E}}$ must map between these distributions, there exists a bijection that also induces a permutation $\psi : [n] \rightarrow [n]$. Note: If we had non-atomic interventions or non-counterfactual exogenous variables, then these distributions would have some overlapping. With overlapping distributions, we can no longer claim there is a bijection mapping between these distributions.

In space \mathcal{Z} , the interventions should also be **sufficiently variable** in order to have non-overlapping $p_{\mathcal{M}}^{\mathcal{Z}, \mathcal{I}}(z, \tilde{z}|I)$ distributions. In the case of soft interventions, \tilde{z} is affected by all ancestral exogenous variables which could be ancestors of other causal variables as well. Consequently, if the changes in causal mechanisms are not sufficient, the effect of ancestral exogenous variables on causal variables will share some similarities and create overlapping distributions. Similar to $p_{\mathcal{M}}^{\mathcal{E}}(e, \tilde{e}|I)$, we can say that there is a permutation between $p_{\mathcal{M}}^{\mathcal{Z}}(z, \tilde{z}|I)$ as well. Furthermore, as we assume the target of interventions are known we have:

$$\forall I \in \mathcal{I} : p_{\mathcal{M}}^{\mathcal{Z}}(z, \tilde{z}|I) = p_{\mathcal{M}'}^{\mathcal{Z}}(z, \tilde{z}|I) \quad (9)$$

Consequently, the permutation ψ is an identity transformation. The effect of soft intervention with known targets on these conditional distributions is shown in Figure A1.

Step 2: Component-wise $\phi_{\mathcal{Z}}$

According to Lemma A1.5, in order to prove that $\phi_{\mathcal{Z}}$ is a component-wise transformation, we need to prove that \tilde{z}'_i and \tilde{z}'_j are independent $\forall i, j, i \neq j$. In implicit modeling we do not know the parents of each causal variable, hence, we assume the distribution of \tilde{z}'_i to be conditioned only on \mathcal{E}'_i as in Equation 5 since \mathcal{E}'_i is a known parent of \tilde{z}'_i . The mean of a conditional distribution can be calculated as:

$$E[\tilde{z}'_i | e'_i] = \mu_{\tilde{z}'_i} + \rho \frac{\sigma_{\tilde{z}'_i}}{\sigma_{e'_i}} (e'_i - \mu_{e'_i}) \quad (10)$$

where ρ and σ are the correlation coefficient and variance of the random variables, respectively. On the other hand, we model \tilde{z}'_i using switch mechanisms as:

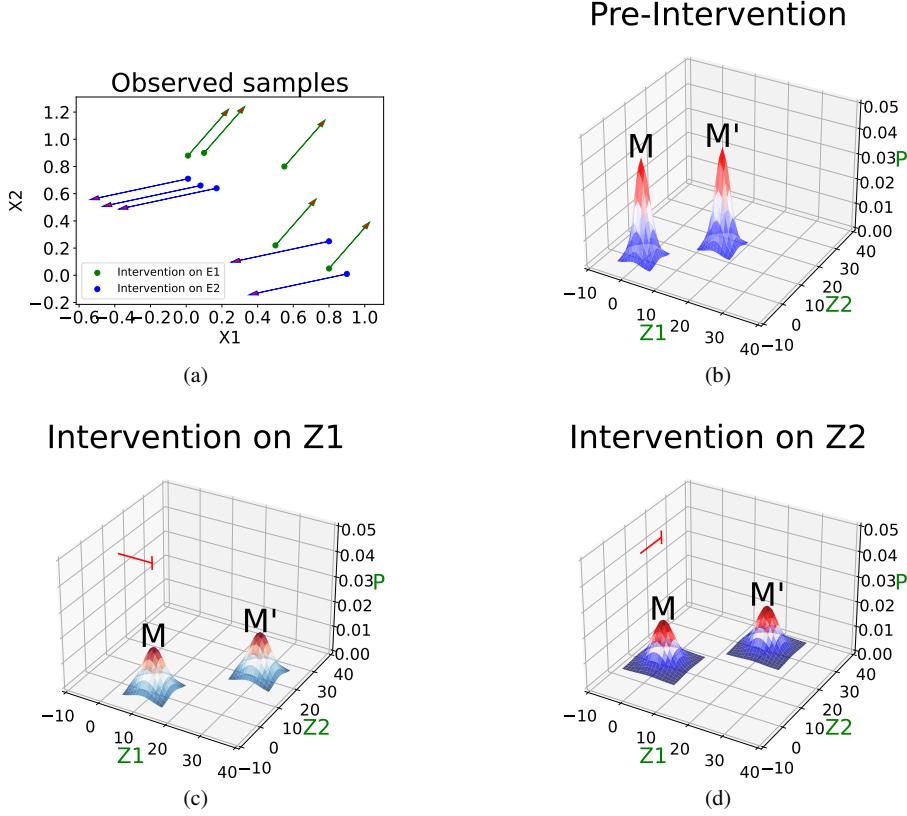


Figure A1: The distribution of observed and causal variables in two causal models \mathcal{M} and \mathcal{M}' , which belong to the equivalence class up to reparameterization. (a) There are 10 observed samples in which Z_1 or Z_2 has been intervened on. (b) The distribution of causal variables when $I = 0$ (no intervention) is identical to each other but the range of value of causal variables are different and can be mapped to each other using ϕ_Z . (c) The intervention on Z_1 ($I = 1$). (d) The intervention on Z_2 ($I = 2$). For $I = 1$ and $I = 2$ the distributions are again identical to each other but are different for different targets of intervention as soft interventions change the conditional distribution (condition on parents) of causal variables. Also, for each value of I , the distributions of \mathcal{M} and \mathcal{M}' should move in one direction as targets are known.

$$\tilde{z}'_i = s_i(\tilde{e}'_i; e'_{/i}, h(v'))$$

By using Taylor's expansion we can write above equation as:

$$\begin{aligned} s_i(\tilde{e}'_i; e'_{/i}, h_i(v')) &= s_i(\tilde{e}'_i; e'_{/i}, h_i(v'_0)) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\partial^n s_i}{\partial h_i^n} \Big|_{h_i=h_i(v'_0)} (h_i(v') - h_i(v'_0))^n \right) \\ &= s_i(\tilde{e}'_i; e'_{/i}, h_i(v'_0)) + R_i \end{aligned}$$

Furthermore, we assume **separable dependence** such that:

$$\exists v'_0 \text{ such that } \forall i \quad s_i(\tilde{e}'_i; e'_{/i}, h_i(v'_0)) = s_i(\tilde{e}'_i; e'_{/i})$$

An example of such a scenario could be in location-scale noise models, where a soft intervention changes the location parameter of the model as:

$$\begin{aligned} s_i(e'_i; e'_{/i}) &= e'_i + \text{loc}(e'_{/i}) \rightarrow \tilde{s}_i(\tilde{e}'_i; e'_{/i}) = s_i(\tilde{e}'_i; e'_{/i}, h_i(v')) \\ &= \tilde{e}'_i + \text{loc}(e'_{/i}) + h_i(v') = \tilde{e}'_i + \text{loc}(e'_{/i}) + v'^2 + v' \end{aligned}$$

In this example, for $v'_0 = 0$, $s_i(\tilde{e}'_i; e'_{/i}, h_i(v'_0)) = s_i(\tilde{e}'_i; e'_{/i})$.

Consequently, we can write the following equality from Equation 10:

$$E[\tilde{\mathcal{Z}}'_i | e'_i] = E[s_i(\tilde{\mathcal{E}}'_i; \mathcal{E}'_{/i}) + R_i | e'_i] = \mu_{\tilde{\mathcal{Z}}'_i} + \rho \frac{\sigma_{\tilde{\mathcal{Z}}'_i}}{\sigma_{\mathcal{E}'_i}} (e'_i - \mu_{\mathcal{E}'_i})$$

By taking the partial derivative of both side with respect to $\tilde{\mathcal{E}}'_j$ we have:

$$\forall j \neq i \quad E\left[\frac{\partial s_i(\tilde{\mathcal{E}}'_i; \mathcal{E}'_{/i})}{\partial \tilde{\mathcal{E}}'_j} \cdot \frac{\partial \tilde{\mathcal{E}}'_i}{\partial \tilde{\mathcal{E}}'_j} + \frac{\partial s_i(\tilde{\mathcal{E}}'_i; \mathcal{E}'_{/i})}{\partial \mathcal{E}'_{/i}} \cdot \frac{\partial \mathcal{E}'_{/i}}{\partial \tilde{\mathcal{E}}'_j} + \frac{\partial R_i}{\partial \tilde{\mathcal{E}}'_j} | e'_i\right] = 0$$

If we did not have the causal mechanism switch variable ($h_i(\mathcal{V}')$), the equation above would only hold if s_i was constant in parents, which is not the case due to the presence of soft interventions, or if $\frac{\partial s_i(\tilde{\mathcal{E}}'_i; \mathcal{E}'_{/i})}{\partial \tilde{\mathcal{E}}'_i} \cdot \frac{\partial \tilde{\mathcal{E}}'_i}{\partial \tilde{\mathcal{E}}'_j} = -\frac{\partial s_i(\tilde{\mathcal{E}}'_i; \mathcal{E}'_{/i})}{\partial \mathcal{E}'_{/i}} \cdot \frac{\partial \mathcal{E}'_{/i}}{\partial \tilde{\mathcal{E}}'_j}$. The latter scenario would imply that $\frac{\partial \tilde{\mathcal{E}}'_i}{\partial \tilde{\mathcal{E}}'_j} \neq 0$, hence, $\tilde{\mathcal{E}}'_i \perp\!\!\!\perp \tilde{\mathcal{E}}'_j$. However, by introducing the causal mechanism switch variable \mathcal{V} and assuming it is observed, we can account for the effects of soft interventions through $h_i(\mathcal{V}')$. In this case, $\frac{\partial \tilde{\mathcal{E}}'_i}{\partial \tilde{\mathcal{E}}'_j} = 0$ as exogenous variables are commonly assumed to be independent in practice. Consequently:

$$\begin{aligned} \forall i, j \quad & \tilde{\mathcal{E}}'_i \perp\!\!\!\perp \tilde{\mathcal{E}}'_j \\ \rightarrow \forall i, j \quad & p(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \tilde{\mathcal{Z}}_i, \tilde{\mathcal{Z}}_j) = p(\tilde{\mathcal{Z}}'_i | \tilde{\mathcal{Z}}_i)p(\tilde{\mathcal{Z}}'_j | \tilde{\mathcal{Z}}_j) \\ \rightarrow \phi_{\mathcal{Z}} \text{ is a component-wise transformation.} \end{aligned}$$

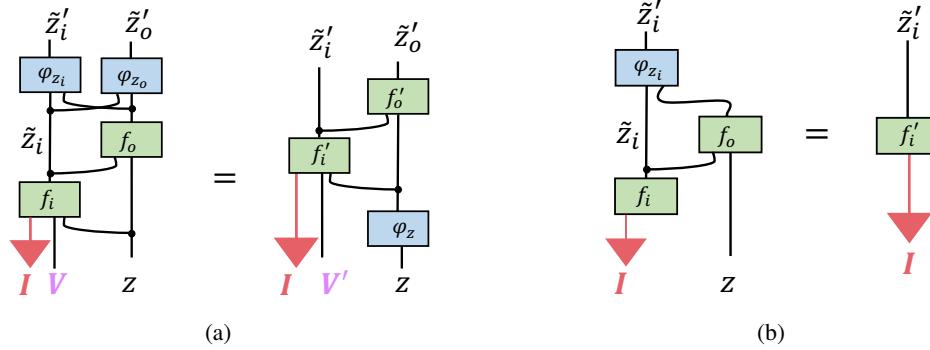


Figure A2: (a) String diagram of the causal variables \mathcal{Z} and \mathcal{Z}' . The triangle indicates sampling I from its distribution. The left-hand side diagram is when $\phi_{\mathcal{Z}}$ is applied last and the right-hand side diagram is when $\phi_{\mathcal{Z}}$ is applied first. I is the intervention which affects intervened causal variable's mechanism variable. V is used to model the effect of intervention on mechanisms and parents. (b) String diagrams after discarding \tilde{z}'_o and the disentangled effect of soft intervention on \tilde{z}_i modeled by V .

Step 3: Component-wise $\phi_{\mathcal{E}}$

Using the result from previous step that $\phi_{\mathcal{Z}}$ is a component-wise transformation, the string diagrams for connections between \mathcal{E} and \mathcal{E}' will be as shown in Figure A3. $\phi_{\mathcal{E}_i}$ will only depend on \mathcal{E}_A , where $A = \text{anc}_i$ is the ancestors of variable i , and e_i . Because $s(e)_{\text{anc}_i}$, $s(e)_i$, and $s'^{-1}(z')_i$ only depend on ancestors and $\phi_{\mathcal{Z}}$ is a component-wise transformation. The first equality in Figure A3 follows from the definition of $\phi_{\mathcal{E}_i}$. The second equality holds when we first apply $\phi_{\mathcal{Z}_A}$ and then apply the causal mechanisms. It can be concluded from the most right-hand side diagram in Figure A3 that the transformation from $\mathcal{E}'_i \times \mathcal{E}_A \rightarrow \mathcal{E}'_i$ is constant in \mathcal{E}_A . Therefore, $\phi_{\mathcal{E}_i}$ is a component-wise transformation.

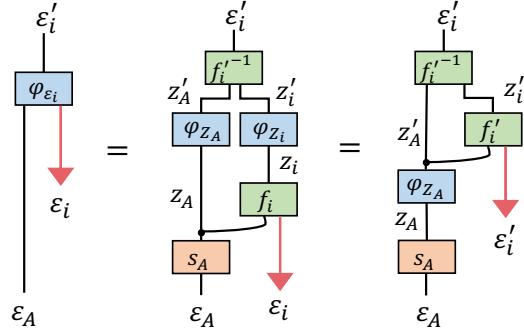


Figure A3: String diagrams for connections between \mathcal{E} and \mathcal{E}' . The triangle indicates sampling variables from their corresponding distributions.

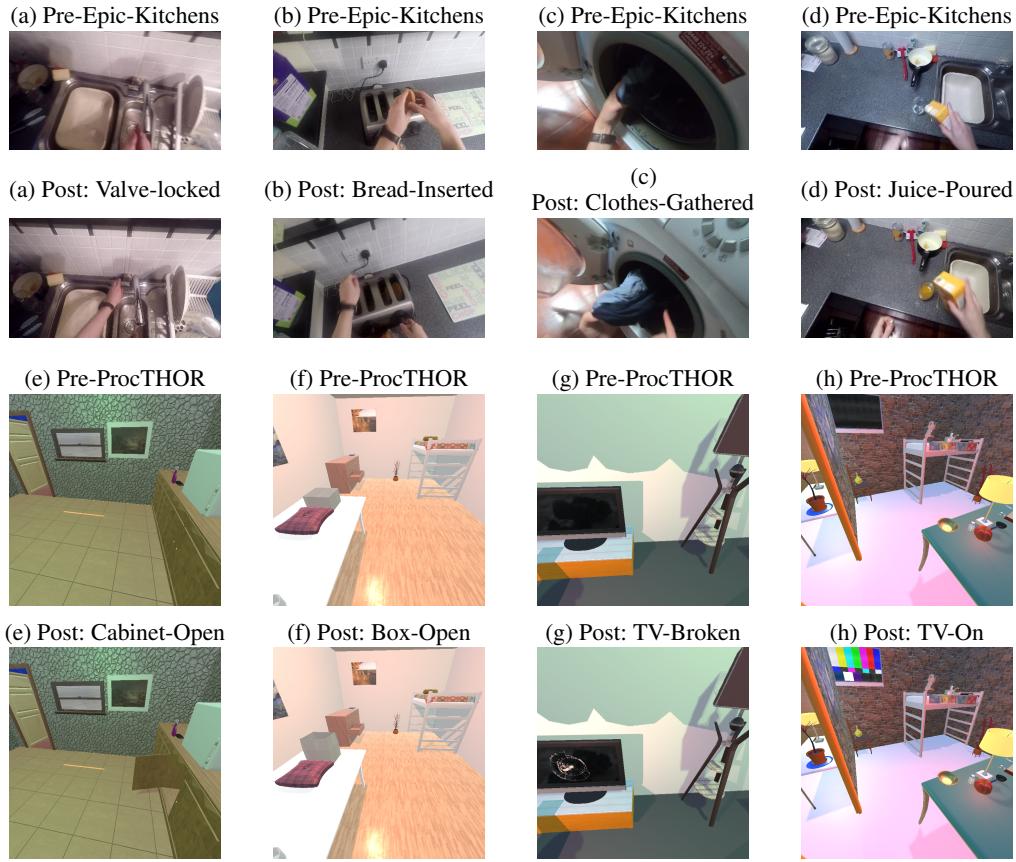


Figure A4: In the Causal-Triplet dataset [19], visual representations capture both pre and post-intervention scenarios. The first two rows showcase data samples from Epic-Kitchens, while the third and fourth rows feature samples from ProcTHOR. Each image in the post-intervention condition is accompanied by labels specifying the corresponding action and intervened object. In the images in the first two rows, the agent is performing an action on an object but the camera angle has also changed. So we can say that for example the distribution of causal variables conditioned on the camera angle has been changed due to soft intervention.

A2 Soft vs. Hard intervention

In a causal model, an intervention refers to a deliberate action taken to manipulate or change one or more variables in order to observe its impact on other variables within the causal model. Interventions help to study how changes in one variable directly cause changes in another, thereby revealing causal relationships.

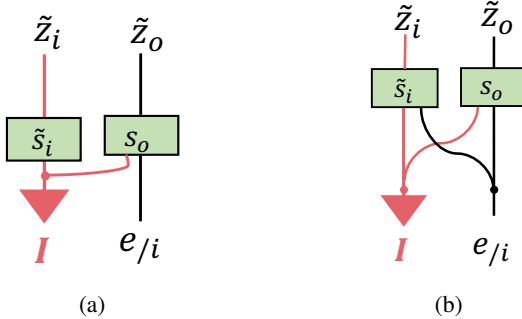


Figure A5: Causal graph models in the presence of Hard (a) and Soft (b) interventions. There are no connections from parents to \tilde{Z}_i in hard interventions (a). Whereas, parents are connected to \tilde{Z}_i in soft interventions (b). Let's consider an implicit model and use $/i$ to denote all variables except variable i . The major difference of soft intervention (b) with hard intervention (a) is that \tilde{Z}_i is no longer disconnected from its parents and its causal mechanism \tilde{s}_i is affected by the intervention. Thus, with a hard intervention, we know the post-intervention parents of a node \tilde{Z}_i (there are none), whereas with soft interventions, the parents themselves may not change.

Based on the levels of control and manipulation in a causal intervention, we can have soft vs. hard interventions. A hard intervention involves directly manipulating the variables of interest in a controlled manner such as Randomized Controlled Trials (RCTs). In other words, a hard intervention sets the value of a causal variable Z to a certain value denoted as $do(Z = z)$ [24].

On the other hand, soft intervention involves more subtle or less controlled manipulation of variables and changes the conditional distribution of the causal variable $p(Z|Z_{pa}) \rightarrow \tilde{p}(Z|Z_{pa})$ which can be modeled as $\tilde{z}_i = \tilde{f}_i(z_{pa_i}, \tilde{e}_i)$ [7].

Looking at interventions from a graphical standpoint, a hard intervention entails that the intervened node is solely impacted by the intervention itself, with no influence coming from its ancestral nodes. Conversely, in the context of a soft intervention, the representation of the intervened node can be influenced not only by the intervention but also by its parent nodes.

As an example, suppose we are trying to understand the causal relationship between different types of diets and weight loss. The *soft intervention* in this scenario could be a switch from a regular diet to a low-carb diet. Switching to a low-carb diet is a voluntary choice made by the individual and there are no external forces or regulations compelling them to make this change (non-coercive).

The intervention involves a modification of the individual's diet rather than a complete disruption since they are adjusting the proportion of macronutrients (fats, proteins, and carbs) they consume, which is less disruptive than a radical change in eating habits (gradual modification). The individual has autonomy to choose and tailor their diet according to their preferences and health goals so they are empowered to make informed decisions about their dietary choices (behavioural empowerment).

Conversely, if the government or an authority were to intervene and enforce a mandatory low-carb diet through legal means, this would constitute a *hard intervention*. In this scenario, regulations would be implemented, prohibiting the consumption of specific carbohydrate-containing foods. Regulatory agencies would be established to oversee and ensure adherence to the low-carb diet mandate, taking actions such as removing prohibited foods from the market, restricting their import and production, and so on. Individuals caught consuming banned foods would be subject to fines, legal repercussions, or other penalties.

A3 Experiments

This section contains additional details about ICRL-SM design architectures, datasets, and experiments settings.

A3.1 Datasets

A3.1.1 Synthetic

We generate simple synthetic datasets with $\mathcal{X} = \mathcal{Z} = \mathbb{R}^n$. For each value of n , we generate ten random DAGs, a random location-scale SCM, then a random dataset from the parameterized SCM. To generate random DAGs, each edge is sampled in a fixed topological order from a Bernoulli

distribution with probability 0.5. The pre-intervention and post-intervention causal variables are obtained as:

$$z_i = \text{scale}(z_{pa_i})e_i + \text{loc}(z_{pa_i}) \xrightarrow{\text{Soft-Intervention}} \tilde{z}_i = \text{scale}(z_{pa_i})\tilde{e}_i + \widetilde{\text{loc}}(z_{pa_i}), \quad (11)$$

where the loc and scale networks are changed in post intervention. The pre-intervention loc and post-intervention $\widetilde{\text{loc}}$ network weights are initialized with samples drawn from $\mathcal{N}(0, 1)$ and $\mathcal{N}(3, 1)$, respectively. For ablation studies, we change the mean of these Normal distributions. The scale is constant 1 for both pre-intervention and post-intervention samples. Both e_i and \tilde{e}_i are sampled from a standard Gaussian. The causal variables are mapped to the data space through a randomly sampled $SO(n)$ rotation. For each dataset, we generate 100,000 training samples, 10,000 validation samples, and 10,000 test samples.

A3.1.2 Causal-Triplet

The Causal-Triplet datasets are consisted of images containing objects in which an action is manipulating the objects shown in Figure A4. Examples of actions and objects in these datasets are given in Table A1 and A2.

Table A1: Actions and objects present in the Causal-Triplet images (ProcTHOR Dataset).

ProcTHOR Dataset								
Object	Television	Bed	Bed	Television	Laptop	Book	Box	
Action	Break	Clean	Dirty	Turn off	Turn on	Open	Close	

Table A2: Actions and objects present in the Causal-Triplet images (Epic-Kitchens Dataset).

Epic-Kitchens Dataset									
Object	Tofu	Rice	Hob	Bag	Cupboard	Garlic	Tap	Wrap	Rice
Action	Insert	Pour	Wash	Fold	Open	Pat	Move	Check	Transition
Object	Wrap	Skin	Button	Lid	Plate	Egg	Sponge	Oil	Cheese
Action	Flip	Gather	Press	Lock	Wrap	Drop	Water	Carry	Stretch
Object	Wrap	Skin	Button	Lid	Plate	Egg	Sponge	Oil	Cheese
Action	Flip	Gather	Press	Lock	Wrap	Drop	Water	Carry	Stretch
Object	Wrap	Skin	Button	Lid	Plate	Egg	Sponge	Oil	Cheese
Action	Flip	Gather	Press	Lock	Wrap	Drop	Water	Carry	Stretch

Based on the actions and objects, we treat our causal variables as attributes of objects which can be changed by actions. Therefore, actions in these datasets are considered as interventions. Assume that z_1 corresponds to the attributes of an object, e.g. a door, the target of opening or closing (action's target) is z_1 .

We use actions' labels in these datasets to detect the targets of interventions to determine which causal variable has been intervened upon. Note that informing the model about the target of intervention is not same as informing about the action itself (See Table 3). We use 5000 images of these datasets to train all models.

A3.2 Architecture Design

Based on the ICRL-SM architecture depicted in Figure 2a, we devised a location-scale solution function (Equation 6) in which the loc_i and scale_i , and h_i networks each comprise of fully connected networks. These networks consist of two layers each, with 64 hidden units per layer and ReLU activation functions. The encoder and decoder parameters for latents \mathcal{E} and $\tilde{\mathcal{E}}$ are shared and we use a separate encoder and decoder with the same architecture for the latent \mathcal{V} . For our synthetic dataset experiments, the encoder and decoder are consisted of fully connected networks with 2 hidden layers and 64 units in each hidden layer. For the Causal-Triplet datasets, we utilized ResNet-based networks. The same encoder and decoder architectures are used for all baseline models in the experiments. ResNet50 encoder, ResNet50 decoder, and classifiers with 1 hidden layer and 64 hidden units are used for predicting actions and objects for experiments in Table 4 and Table 3. ResNet18 encoder, ResNet18 decoder, and classifiers with 2 hidden layer and 2 hidden units are used for predicting actions and objects for experiments in Table A4 and Table A3.

A3.3 Training

To enforce the condition described in Equation 5 for $i \notin \mathcal{I}$, we assign the post-intervention exogenous variables the same value as the pre-intervention exogenous variables. In mathematical terms, this translates to $\forall i \notin \mathcal{I}$, we set $\tilde{e}_i = e_i$.

In our experiments, we do not pretrain the networks, however, for the baseline models we follow the training procedure in [3]. We also use consistency in our experiments to ensure that the encoder and decoder are inverse of each other. Consistency regularizer is used as $\sum_i E_{\hat{x} \sim p(\hat{x}|e), x \sim p(x)}[(x - \hat{x})^2]$ where \hat{x} are the reconstructed samples.

For optimization, Adam optimizer is used with default hyperparameters. In the synthetic experiments, learning rate changes from $3e-4$ to $1e-8$ with a cosine scheduler. In the Causal-Triplet experiments in Table 4 and Table 3 learning rate changes from 0.002 to $1e-8$ with a cosine scheduler. For Table A4 and Table A3 experiments learning rate changes from 0.0001 to $1e-8$ with a cosine scheduler. In all experiments the batch size is set to 64. In the main Causal-Triplet experiments we train the models for 400 epochs, in the appendix Causal-Triplet experiments we train the models for 2000 epochs, and in the synthetic experiments we train the models for 100 epochs. In the appendix experiments, the graph parameters for explicit models are frozen after 1000 epochs.

All models are trained using Nvidia GeForce RTX4090 GPUs. Each of the Causal-Triplet experiments takes 3-8 hours to train the models and each of the synthetic experiments takes 2-3 hours to train the models.

We save the models' weights with best validation loss and evaluate them using those weights with test data.

A4 Ablation study

A4.1 Scalability

While our primary research objective centered on addressing identifiability challenges in implicit causal models under soft interventions, we also conducted an investigation into the scalability of our proposed model. To comprehensively assess its performance, we designed experiments covering a range of causal graphs, featuring 5 to 10 variables, with 10 different seeds for each variable, following a similar experimental setup as our 4-variable causal graph experiments. The outcomes of these experiments, comparing ICRL-SM and ILCM, are presented in Figure A6. By increasing the number of variables in the graph, confounding factors and ambiguities of causal relations increase as well. Consequently, more supervision on \mathcal{V} is required to better separate the effect of causal variables themselves on the observed variables.

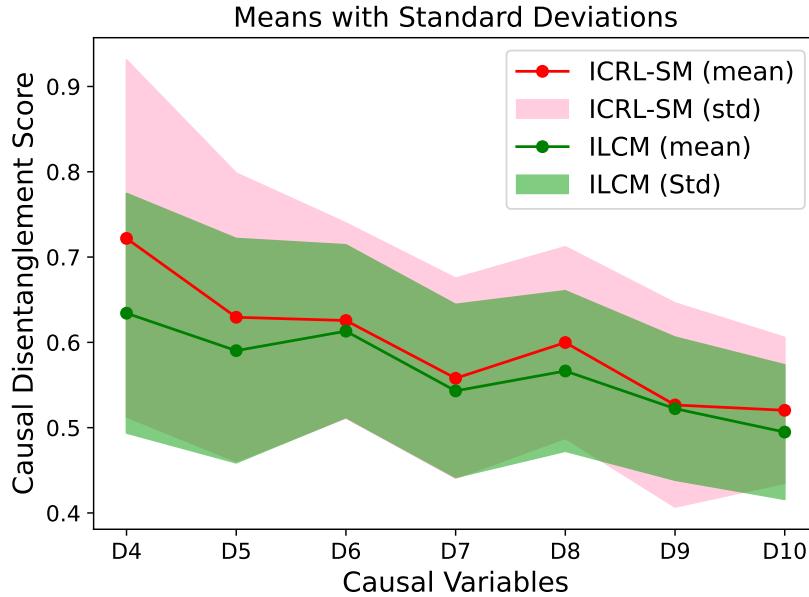


Figure A6: Causal disentanglement for different number of variables

A4.2 Backbone model

We trained the models using a simpler backbone model, ResNet18, to see how it affects performance. The input image resolution is 64×64 and we use the intervened causal variables to predict action

and object classes. The results are shown in Table A4 and A3. It can be seen from the results that the proposed method outperforms other explicit and implicit models even with a simpler model.

Table A3: Table comparing action and object accuracy across various methods on Causal-Triplet datasets using ResNet18 model.

Method	Epic-Kitchens		ProcTHOR	
	Action Accuracy	Object Accuracy	Action Accuracy	Object Accuracy
$\beta - VAE$ [11]	0.15	0.04	0.20	0.36
$d - VAE$ [21]	0.16	0.02	0.15	0.38
ILCM [3]	0.19	0.04	0.15	0.42
ICRL-SM (ours)	0.35	0.04	0.40	0.69

Table A4: Action and object accuracy of three explicit models are compared with ICRL-SM. Experiments are conducted applying image with resolution of R_{64} as the input to the Resnet18 encoder with the intervened causal variable (z_i).

Datasets	Methods	Action Accuracy	Object Accuracy
Epic-Kitchens	ENCO [16]	0.14	0.03
	DDS [5]	0.16	0.05
	Fixed-order	0.14	0.05
	ICRL-SM (ours)	0.35	0.04
ProcTHOR	ENCO [16]	0.16	0.28
	DDS [5]	0.34	0.35
	Fixed-order	0.34	0.38
	ICRL-SM (ours)	0.40	0.69