

---

# Implicit Causal Representation Learning via Switchable Mechanisms

---

Anonymous Author(s)

Affiliation  
Address  
email

## Abstract

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

## 17 1 Introduction

18 One of the long-standing challenges in causal  
19 representation learning is how to recover the  
20 ground-truth causal graph of a system solely  
21 from observations. Termed the *identifiability*  
22 *of causal models* problem, this endeavor is crucial.  
23 Without achieving identifiability, we risk  
24 erroneously attributing causal relationships to  
25 learned representations. Furthermore, statisti-  
26 cal models can masquerade as Directed Acyclic  
27 Graphs (DAGs) where edges lack causal signifi-  
28 cance, further complicating our pursuit.

29 When considering the challenge of identifying  
30 causal models, it is known that the Markov con-  
31 dition in graphs is insufficient for this task [26].  
32 Thus, without additional assumptions or data,  
33 we find ourselves limited to learning only a  
34 *Markov Equivalence Class* (MEC) of the causal  
35 model. Existing works have made different  
36 assumptions about availability of ground-truth  
37 causal variables labels [34], model parameters  
38 [1], availability of paired interventional data [3, 31], and availability of intervention targets [17] to  
39 ensure identifiability of causal models.

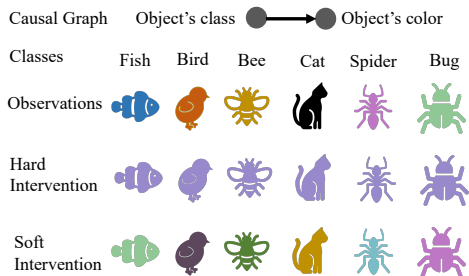


Figure 1: Difference between hard interventions and soft interventions: As seen in the middle row, hard interventions sever connections with parents. Therefore, an object’s class cannot have any effect on the object’s color when we intervene on color. On the other hand, soft interventions, as shown in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

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

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

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

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

79

## 80 2 Related Work

81 Causal representation learning has recently garnered significant attention [27, 14]. The primary  
82 challenge in this problem lies in achieving identifiability beyond the Markov equivalence class [26].  
83 Solely relying on observational data necessitates additional assumptions regarding causal mechanisms,  
84 decoders, latent structure, and the availability of interventional data [22, 28, 36, 25, 15, 1, 40, 13,  
85 34]. Recent works have focused on identifying causal models from collected interventional data  
86 instead of making strong assumptions about functions of the causal model. Interventional data  
87 facilitates identifiability based on relatively weak assumptions [1, 6, 3, 39, 33]. This type of data  
88 can be further categorized based on whether it involves soft or hard interventions, and whether the  
89 manipulated variables are observed and specified or latent. Our focus in this paper is on examining  
90 soft interventions, encompassing both observed and unobserved variables.

### 91 2.1 Explicit models vs. Implicit models

92 Table 1 presents a comparison of the assumptions and identifiability results between our proposed  
93 theory and other related works on causal representation learning with interventions. In causal repre-  
94 sentation learning with interventions, one approach assumes a given causal graph and concentrates  
95 on identifying causal mechanisms and mixing functions. For instance, Causal Component Analysis  
96 (CauCA) [33] explores soft interventions with a known graph. Alternatively, when the graph is

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

97 not provided, explicit models seek to reconstruct it from interventional data [6, 17], potentially  
 98 resulting in a chicken-and-egg problem in causal representation learning [3]. Current methods face  
 99 the challenge of simultaneously learning the causal graph and other network parameters, especially  
 100 in the absence of information about causal variables or the graph. Addressing these challenges, [3]  
 101 recently introduced ILCM, which performs *implicit* causal representation learning exclusively using  
 102 *hard* intervention data. In contrast, our approach introduces a novel method for learning an implicit  
 103 model from *soft* interventions. [3] describes methods for extracting a causal graph from a learned  
 104 implicit model, which could be applied to our method as well. In our experiments, we will compare  
 105 our method with ILCM and dVAE [21], given their implicit nature and similar experimental settings  
 106 and assumptions. Additionally, to showcase the superiority of our method over explicit models, we  
 107 will employ explicit causal model discovery methods like ENCO [16] and DDS [5], in conjunction  
 108 with various variants of  $\beta$ -VAE.

## 109 2.2 Hard interventions vs Soft interventions

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

119 Recent work has expanded the concept of explicit hard interventions to include soft interventions. In  
 120 their study, [38] address the identification of causal models from soft interventions, leveraging the  
 121 sparsity of the adjacency matrix as an inductive bias. However, when dealing with implicit models,  
 122 soft interventions introduce new complexities. Identifiability becomes more challenging, as the  
 123 causal effect of variables on observed variables is less apparent. This ambiguity arises from the dual  
 124 possibility of effects originating from interventions or influences from parent variables on the causal  
 125 variables. Moreover, in scenarios where implicit modeling is retained, the absence of knowledge about  
 126 parent variables further complicates identifiability. While [3] theoretically establishes identifiability  
 127 for hard interventions, practical experiments involving complex causal models with over 10 variables  
 128 reveal increased ambiguity and confounding factors. Consequently, model identification becomes  
 129 less straightforward.

## 130 3 Methodology

### 131 3.1 Data Generating Process

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

140 Identifying causal models from data can be complex and is often studied within classes of models  
 141 such as those identifiable up to affine transformations. For example, in the context of nonlinear  
 142 *Independent Component Analysis (ICA)*, the generative process also involves a mixture function  $g$  of  
 143 latent causal variables  $\mathcal{Z} \in \mathbb{R}^n$ , resulting in observations  $\mathcal{X} \in \mathbb{R}^n$  [15, 41]. However, a significant  
 144 distinction between causal representation learning and nonlinear-ICA is that in the former, the causal

145 variables  $\mathcal{Z}$  may have complex dependencies. Our objective in this paper is to recover  $\mathcal{E}$  from  $\mathcal{X}$  and  
 146 eventually map  $\mathcal{E}$  to  $\mathcal{Z}$  using solution functions.

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

155 **Assumption 3.1.** (*Data generating assumptions*)

156 **1. Atomic Interventions:** For every sample  $(x, \tilde{x}, i)$ , only one causal variable is targeted by an  
 157 intervention.

158 **2. Known Targets:** Targets of soft interventions are known.

159 **3. Post-intervention Exogenous Variables:** The exogenous variables' values change only for the  
 160 corresponding intervened causal variable, while the others maintain their pre-intervention values,  
 161 thus  $e_i \neq \tilde{e}_i$  if  $i \in \mathcal{I}$ , and  $e_i = \tilde{e}_i$  otherwise.

162 **4. Sufficient Variability:** Soft interventions alter causal mechanisms to introduce sufficient variability  
 163 [15]. These interventions should modify causal mechanisms to ensure non-overlapping conditional  
 164 distributions of causal variables (refer to Figure A1).

165 **5. Diffeomorphic decoder and causal mechanisms:** Diffeomorphism guarantees no information loss  
 166 and avoids abrupt changes in the function's image.

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

### 171 3.2 Causal Mechanisms Switch Variable

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

180 **Augmented implicit causal model** To model the effect of soft interventions, we introduce the  
 181 causal mechanism switch variable  $\mathcal{V}$  [26]. By leveraging  $\mathcal{V}$ , we can effectively switch to the pre-  
 182 intervention causal mechanisms within post-intervention data. This facilitates the model's ability to  
 183 solely focus on discerning alterations in the intrinsic characteristics of each causal variable. These  
 184 changes are encapsulated within their respective exogenous variables, aiding the model in learning  
 185 the causal relationships more accurately. We propose to use a modulated form of  $\mathcal{V}$  to model the  
 186 soft intervention effects on each causal variable as an additive effect with a nonlinear function  $h_i$   
 187 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  
 188 not known, we have to use a modulated form of  $\mathcal{V}$  in every causal variable's solution function and  
 189 the inclusion of  $h_i(\mathcal{V})$  enables the model to encompass variations in the parental sets of all causal  
 190 variables in  $\mathcal{V}$ . Therefore, there is a switch variable  $\mathcal{V}_i$  for each causal variable  $\mathcal{Z}_i$ . Adding switch  
 191 variables to solution functions leads to the concept of an *augmented implicit causal model*.

192 **Definition 3.2.** (*Augmented Implicit Causal Models*) An *Augmented Implicit Causal Models (AICMs)*  
 193 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  
 194 the effect of soft interventions on solution functions  $\mathcal{S}$ :

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

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

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

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

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

203 where we’ll use  $R_i$  as a short-hand for Equation 2. We define the **separable dependence** property  
 204 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  
 205 could be in location-scale noise models such as,  $s_i(\tilde{e}_i; e_{/i}, h_i(v)) = \tilde{e}_i + \text{loc}(e_{/i}) + h_i(v) =$   
 206  $\tilde{e}_i + \text{loc}(e_{/i}) + v^2 + v$  where  $v_0$  would be zero . By assuming the separable dependence property,  
 207 we can write the solution function in Equation 2 as:

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

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

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

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

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

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

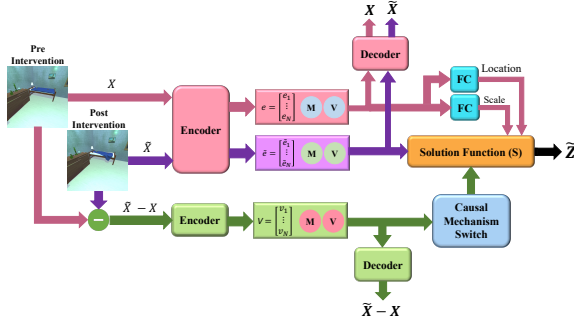
228 where  $R = [R_0, R_1, \dots, R_n]$  and  $n$  is the number of causal variables.  $R$  and  $\Delta e_i$  are the vectors  
 229 indicating the soft intervention effects and change in effect of the exogenous variable of the intervened  
 230 causal variable, respectively. Note that elements of 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

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

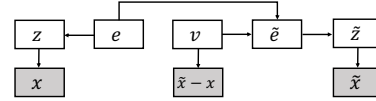
### 238 3.3 Identifiability Theorem for Implicit SCMs with Soft Interventions

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

243 We establish identifiability up to reparameterization, allowing for the mapping of causal variables  $\mathcal{Z}$   
 244 and  $\mathcal{Z}'$  between two Latent Causal Models ( $\mathcal{M}$  and  $\mathcal{M}'$ ) through component-wise transformations



(a) General overview of ICRL-SM



(b) Generative model

245 (Definition A1.2). Given our implicit modeling approach, lacking knowledge of the causal graph, we  
 246 include all exogenous variables in the solution functions, as depicted in Equation 1. Notably, **the**  
 247 **causal graph remains unaltered during learning**. To illustrate, we contrast hard interventions,  
 248 which neglect parental influences, with soft interventions that acknowledge parental effects in a simple  
 249 example. Consider a basic causal model  $Z_1 \rightarrow Z_2$  alongside a location-scale noise model [12] for the  
 250 solution function, given by  $\tilde{z}_2 = \frac{\tilde{e}_2 - \text{loc}(e_1)}{\text{scale}(e_1)}$ . The distribution  $p(\tilde{Z}_2)$  mean is  $\frac{1}{\text{scale}(e_1)} \times \text{mean}(\tilde{\mathcal{E}}_2) -$   
 251  $\frac{\tilde{\text{loc}}(e_1)}{\text{scale}(e_1)}$ . In the context of hard interventions, we can assume  $p(\tilde{Z}_2|Z_1) = p(\tilde{Z}_2) = N(0, 1)$  as there  
 252 are no parental effects. Consequently, the location and scale networks within the solution function tend  
 253 to dampen parental effects, given the absence of parental influence in the ground-truth data. Contrarily,  
 254 soft interventions exhibit parental influence in the ground-truth data, thus  $p(\tilde{Z}_2|Z_1) \neq N(0, 1)$ . Due  
 255 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  $\mathcal{E}_2$   
 256 is a known parent of  $\tilde{Z}_2$ . Consequently, parental effects are propagated to  $\mathcal{E}_i$  (the corresponding  
 257 exogenous variable of each causal variable), violating identifiability up to reparameterization. By  
 258 leveraging  $\mathcal{V}$ , we allow parental effects to propagate to  $\mathcal{V}$  instead of  $\mathcal{E}_i$ .

259 **Definition 3.4.** (Equivalence up to component-wise reparameterization) Let  $\mathcal{M} = (\mathcal{A}, \mathcal{X}, g, \mathcal{I})$   
 260 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  
 261 observation space  $\mathcal{X}$ , shared intervention targets  $\mathcal{I}$ , and respective decoders  $g$  and  $g'$ . We say that  
 262  $\mathcal{M}$  and  $\mathcal{M}'$  are equivalent up to component-wise reparameterization  $\mathcal{M} \sim_r \mathcal{M}'$  if there exists a  
 263 component-wise transformation (Definition A1.2)  $\phi_Z$  from the causal variables  $\mathcal{Z}$  to the causal  
 264 variables  $\mathcal{Z}'$  and a component-wise transformation  $\phi_{\mathcal{E}}$  between  $\mathcal{E}$  and  $\mathcal{E}'$  such that:

- 265 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.,  
 266  $Z_i \rightarrow Z_j$  holds in  $\mathcal{G}$  iff  $Z'_i \rightarrow Z'_j$  holds in  $\mathcal{G}'$ . Edges  $\mathcal{E}_i \rightarrow Z_i$  should be preserved as well.)
- 267 2. The exogenous transformation preserves the probability measure on exogenous variables  
 268  $p_{\mathcal{E}'} = (\phi_{\mathcal{E}})_* p_{\mathcal{E}}$  (Definition A1.4).
- 269 3. The causal transformation preserves the probability measure on causal variables  $p_{\mathcal{Z}'} = (\phi_Z)_* p_{\mathcal{Z}}$   
 270 (Definition A1.4).

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

- 275 1. Data generating assumptions explained in Assumption 3.1.
- 276 2. Soft interventions satisfy Assumption 3.3.
- 277 3. The causal and exogenous variables are real-valued.
- 278 4. The causal and exogenous variables follow a multivariate normal distribution.

279 Then the following statements are equivalent:

- 280 - Two LCMs  $\mathcal{M}$  and  $\mathcal{M}'$  assign the same likelihood to interventional and observational data i.e.,  
 281  $p_{\mathcal{M}}^{\mathcal{X}, \mathcal{I}}(x, \tilde{x}, i) = p_{\mathcal{M}'}^{\mathcal{X}, \mathcal{I}}(x, \tilde{x}, i)$ .
- 282 -  $\mathcal{M}$  and  $\mathcal{M}'$  are disentangled, that is  $\mathcal{M} \sim_r \mathcal{M}'$  according to Definition 3.4.

### 283 3.4 Training Objective

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

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

287 **3.** The post-intervention exogenous variables  $\tilde{\mathcal{E}}$ .

288 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  
 289 approximation as training objective:

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

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

293 **1.**  $p(e)$  is the prior distribution over exogenous variables  $e$ .

294 **2.**  $p(v)$  is the prior distribution over switch variables  $v$ .

295 **3.**  $p(\tilde{e}|e, v)$  is a transition model that shows how the exogeneous variables change as a function of the  
 296 intervention.

297 We factorize the posterior with a mean-field approximation  $q(v, e, \tilde{e}|x, \tilde{x}) = q(v|\tilde{x}-x) \cdot q(e|x) \cdot$   
 298  $q(\tilde{e}|\tilde{x})$  and, following our data generation model (Figure 2b), the reconstruction probability  
 299 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  
 300  $p(\tilde{e}, e, v) = p(\tilde{e}|e, v)p(v)p(e)$  (Figure 2b). Pre-intervention exogenous variables are mutually inde-  
 301 pendent, 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  
 302 Gaussian. Furthermore, as we assume  $e_i = \tilde{e}_i$  for all non-intervened variables, the  $p(\tilde{e}|e, v)$  will be  
 303 as follows:

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

304 The last equality is obtained from the Change of Variable Rule in probability theory, applied to the  
 305 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  $e_i$   
 306 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  
 307 determined by  $e_i$ . We implement the solution function using a location-scale noise models [12] as  
 308 also practiced in [3], which defines an invertible diffeomorphism. For simplicity, in our experiments,  
 309 we are only going to change the *loc* network in post-intervention. Therefore,  $h_i(v)$  will be used as:

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

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

## 312 **4 Experiments and Results**

313 The experiments conducted in this paper address two downstream tasks; (1) Causal Disentanglement  
 314 to identify the true causal graph from pairs of observations  $(x, \tilde{x}, i)$ , and (2) Action Inference to make  
 315 supervised inferences about actions generated from the post-intervention samples using information  
 316 about the values of the manipulated causal variables. Moreover, we conducted additional experiments  
 317 designed as an ablation study, the results of which are presented in A4. All models are trained using  
 318 the same setting and data with known intervention targets.

### 319 **4.1 Datasets**

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

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

325 where the *loc* and *scale* networks are changed in post intervention. The pre-intervention *loc* and  
 326 post-intervention  $\widetilde{\text{loc}}$  network weights are initialized with samples drawn from  $\mathcal{N}(0, 1)$  and  $\mathcal{N}(3, 1)$ ,  
 327 respectively. The *scale* is constant 1 for both pre-intervention and post-intervention samples. Both  
 328  $e_i$  and  $\tilde{e}_i$  are sampled from a standard Gaussian. The causal variables are mapped to the data space  
 329 through a randomly sampled  $SO(n)$  rotation. For each dataset, we generate 100,000 training samples,  
 330 10,000 validation samples, and 10,000 test samples.

331 **Action Datasets** Causal-Triplet datasets tailored for *actionable* counterfactuals [19] feature paired  
 332 images where several global scene properties may vary including camera view and object occlusions.  
 333 Thus, the images can be viewed as outcomes of soft interventions, wherein actions affect objects  
 334 alongside subtle alterations. These datasets [19] consist of: images obtained from a photo-realistic  
 335 simulator of embodied agents, ProcTHOR [9], and the other contains images repurposed from a real-  
 336 world video dataset of human-object interactions [8]. The former one contains 100k images in which  
 337 7 types of actions manipulate 24 types of objects in 10k distinct ProcTHOR indoor environments.  
 338 The latter consists of 2,632 image pairs, collected under a similar setup from the Epic-Kitchens  
 339 dataset with 97 actions manipulating 277 objects. Based on the nature of actions in this dataset, the  
 340 causal variables should represent attributes of objects such as shape and color. As the dataset consists  
 341 of images we train all the methods with ResNet encoder and decoder. For the ProcThor dataset the  
 342 number of causal variables are 7. For the Epic-Kitchens dataset, we randomly chose 20 actions from  
 343 the dataset as 97 causal variables will be too complex in a VAE setup.

## 344 4.2 Metrics











345 For the causal disentanglement task, we are going to use the DCI scores [10]. Causal disentanglement  
 346 score quantifies the degree to which  $\mathcal{Z}_i$  factorises or disentangles the  $\mathcal{Z}^*$ . Causal disentanglement  $D_i$   
 347 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  $P_{ij} = \frac{R_{ij}}{\sum_{k=0}^{K-1} R_{ik}}$   
 348 and  $R_{ij}$  denotes the probability of  $\mathcal{Z}_i$  being important for predicting  $\mathcal{Z}_j^*$ . Total causal disentanglement  
 349 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  
 350 to which each  $\mathcal{Z}_i^*$  is captured by a single  $\mathcal{Z}_i$ . Causal completeness is calculated as  $C_j = (1 -$   
 351  $H_D(\tilde{P}_{.j})) = (1 + \sum_{d=0}^{D-1} \tilde{P}_{dj} \log_D \tilde{P}_{dj})$ .  $D$  and  $K$  here are equal to the dimension of  $\mathcal{Z}^*$  and  $\mathcal{Z}$   
 352 which is  $n$ . For the action inference task, we will use classification accuracy as a metric. As we  
 353 assume intervention targets are known, we train all models using known intervention targets for a fair  
 354 comparison.

## 355 5 Results

### 356 5.1 Causal Disentanglement

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

Table 2: Comparison of identifiability results

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

362 The results in Table 2 indicate that our method ICRL-SM can identify the true causal graph in most  
 363 cases. The worst results are seen for graphs  $G_5$  and  $G_{10}$ . As mentioned in [27, 25], causal graphs are  
 364 sparse and in the  $G_5$  case, where the graph is fully connected, the proposed method cannot identify  
 365 the causal variables well. Furthermore, in the next experiment we are going to examine the factors  
 366 affecting causal disentanglement such as the number of edges in the graph and the intensity of soft  
 367 intervention effect. These findings can explain why ICRL-SM cannot identify causal variables in  
 368  $G_{10}$  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  $64 \times 64$ .

Method	Epic-Kitchens				ProcTHOR			
	Action Accuracy		Object Accuracy		Action Accuracy		Object Accuracy	
	$Z;R_{64}$	$z_i;R_{64}$	$Z;R_{64}$	$z_i;R_{64}$	$Z;R_{64}$	$z_i;R_{64}$	$Z;R_{64}$	$z_i;R_{64}$
$\beta - VAE$ [11]	<b>0.27</b>	0.18	0.19	0.06	<b>0.39</b>	0.30	<b>0.44</b>	0.37
$d - VAE$ [21]	0.19	0.69	<b>0.20</b>	0.17	0.35	0.81	0.40	0.78
ILCM [3]	0.21	0.59	0.14	0.14	0.30	0.70	0.41	0.76
<b>ICRL-SM (ours)</b>	0.16	<b>0.86</b>	0.16	<b>0.18</b>	0.28	<b>0.93</b>	0.40	<b>0.82</b>

## 369 5.2 Factors Affecting Causal Disentanglement

370 In this experiment, we consider the graph  $G3$ , which has the best identifiability, and change the  
371 intensity of soft intervention and number of edges in its data generation process. To change the  
372 intensity, the post-intervention  $\widetilde{loc}$  network weights are initialized with samples drawn from  $N(1, 1)$   
373 (almost similar to  $loc$ ) and  $N(10, 1)$  (significantly different from  $loc$ ). To change the number of  
374 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  $G5$

Datasets	Methods	Action Accuracy	Object Accuracy
Epic-Kitchens	ENCO [16]	0.69	0.13
	DDS [5]	0.44	0.09
	Fixed-order	0.79	0.14
	<b>ICRL-SM (ours)</b>	<b>0.86</b>	<b>0.18</b>
ProcTHOR	ENCO [16]	0.45	0.53
	DDS [5]	0.64	0.67
	Fixed-order	0.65	0.54
	<b>ICRL-SM (ours)</b>	<b>0.93</b>	<b>0.82</b>

Edges	Post-intervention causal mechanism	Causal Disentanglement	Causal Completeness
Chain	Default	0.98	0.98
Full	Default	0.89	0.89
Default	Significantly different	0.68	0.73
Default	Almost similar	0.85	0.86

375 The results in Table 4 further confirms the sparsity of causal graphs as the causal disentanglement is  
376 much worse in the fully-connected graph than the default graph of  $G3$ . The result for significantly  
377 different post-intervention causal mechanisms indicate that the switch variable cannot approximate  
378 intense effects of soft intervention and more supervision is required to observe  $\mathcal{V}$ . Similar post-  
379 intervention causal mechanisms also do not have sufficient variability to disentangle the causal  
380 variables as mentioned in Theory 3.5.

## 381 5.3 Action Inference

382 In this experiment, we show the performance of ICRL-SM in the real-world Causal-Triplet datasets.  
383 In these datasets  $\mathcal{V}$  i.e., soft intervention effects, are not directly observable. Nevertheless, our findings  
384 suggest that incorporating soft interventions through  $\mathcal{V}$  leads to superior performance compared to  
385 other implicit modeling approaches. Clearly, understanding the impact of soft interventions on the  
386 generative system of the dataset would result in improved outcomes.

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

## 396 6 Conclusion

397 ICRL-SM, our novel model, enhances implicit causal representation learning during soft interventions  
398 by introducing a causal mechanism switch variable. Evaluations on synthetic and real-world datasets  
399 demonstrate ICRL-SM’s superiority over state-of-the-art methods, highlighting its practical effective-  
400 ness. Our findings emphasize ICRL-SM’s ability to discern causal models from soft interventions,  
401 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. Proctor: 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.

# 518 Appendix

## 519 A1 Proof of Identifiability Theorem

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

### 522 Definition A1.1. Structural Causal Models

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

524 **1.** The domain of causal variables  $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \dots \times \mathcal{Z}_n$ .

525 **2.** The domain of exogenous variables  $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 \times \dots \times \mathcal{E}_n$ .

526 **3.** A directed acyclic graph  $\mathcal{G}(\mathcal{C})$  over the causal and exogenous variables.

527 **4.** A causal mechanism  $f_i \in \mathcal{F}$  which maps an assignment of parent values for the parents  $\mathcal{Z}_{pa_i}$  plus  
528 an exogenous variable value for  $\mathcal{E}_i$  to a value of causal variable  $Z_i$ .

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

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

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

$$538 (f_*p)(E) = p(f^{-1}(E))$$

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

541 **Lemma A1.5.** The transformation  $\phi_{\mathcal{Z}} : \mathcal{Z} \rightarrow \mathcal{Z}'$  between the causal variable of two LCMs  $\mathcal{M}$   
542 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  
543 the causal variables follow a multivariate normal distribution conditional on the pre-intervention  
544 exogenous variables where  $\tilde{E}'_i$  denote the post-intervention exogenous variable of causal variable  $i$   
545 in  $\mathcal{M}'$ .

546 *proof:* We consider the case where the exogenous variables are mapped to causal variables by a

547 location-scale noise model such that  $\tilde{z}_i = \frac{\tilde{e}_i - \widetilde{\text{loc}}(e_{/i})}{\widetilde{\text{scale}}(e_{/i})}$ .

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

548 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  
549 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 \left[ (\tilde{\mathcal{Z}}'_i - E[\tilde{\mathcal{Z}}'_i | \mathcal{E}']) (\tilde{\mathcal{Z}}'_j - E[\tilde{\mathcal{Z}}'_j | \mathcal{E}']) | \mathcal{E}' \right] = 0 \\
& \rightarrow \text{cov}(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \mathcal{E}') = 0
\end{aligned}$$

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

553 Assuming that the causal variables follow a **multivariate normal distribution conditional on the**  
554 **pre-intervention exogenous variables**,  $\text{cov}(\tilde{\mathcal{Z}}'_i, \tilde{\mathcal{Z}}'_j | \mathcal{E}') = 0$  would imply that  $\tilde{\mathcal{Z}}'_i \perp\!\!\!\perp \tilde{\mathcal{Z}}'_j | \mathcal{E}'$ . Let's  
555 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  
556 Assumption 3.1, the decoders are diffeomorphism, hence,  $\phi_{\mathcal{E}}$  is a diffeomorphism. Furthermore, let's  
557 denote  $\tilde{s}$  as the set of all solution functions in post-intervention which are also diffeomorphism as  
558 stated in Assumption 3.1. Consequently:

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

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

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

568 **1.** Identical correspondence assumptions explained in 3.1.

569 **2.** Soft interventions satisfy Assumption 3.3.

570 **3.** The causal and exogenous variables are real-valued.

571 **4.** The causal and exogenous variables follow a multivariate normal distribution.

572 Then the following statements are equivalent:

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

574  $p_{\mathcal{M}}^{\mathcal{X}}(x, \tilde{x}) = p_{\mathcal{M}'}^{\mathcal{X}}(x, \tilde{x})$ .  
 575 -  $\mathcal{M}$  and  $\mathcal{M}'$  are disentangled, that is  $\mathcal{M} \sim_r \mathcal{M}'$  according to Definition 3.4.

576 **Proof** We will proceed to prove the equivalence between statements 1 and 2 by showing the implica-  
 577 tion is true in each direction.

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

579 This direction is fairly straightforward. According to Definition 3.4, the fact that  $\mathcal{M} \sim_r \mathcal{M}'$  implies  
 580 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  
 581 that ancestry is preserved,  $\phi_{\mathcal{Z}}$  is measure preserving, and that causal variables are obtained from their  
 582 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  
 583 models are trained to maximize the log likelihood of  $p(x, \tilde{x}, \tilde{x} - x)$  and the latent spaces in  $\mathcal{M}$   
 584 and  $\mathcal{M}'$  have the same distribution, the decoders should yield the same observational distributions  
 585  $p_{\mathcal{M}}^{\mathcal{X}}(x, \tilde{x}) = p_{\mathcal{M}'}^{\mathcal{X}}(x, \tilde{x})$ .

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

587 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

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

588 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.  
 589 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,  
 590  $\phi_{\mathcal{E}}$  is measure-preserving. Similarly,  $\phi_{\mathcal{Z}}$  is measure-preserving as well since causal mechanisms are  
 591 **diffeomorphisms**.

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

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

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

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

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

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

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

616 where  $\rho$  and  $\sigma$  are the correlation coefficient and variance of the random variables, respectively. On  
 617 the other hand, we model  $\tilde{\mathcal{Z}}'_i$  using switch mechanisms as:

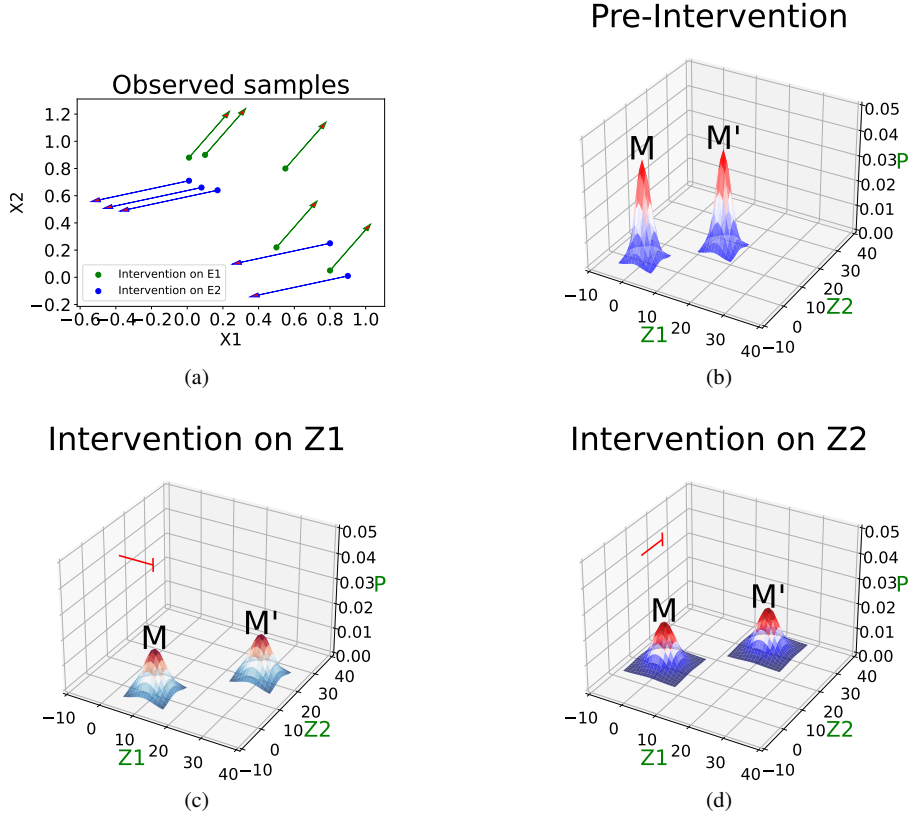


Figure A1: The distribution of observed and causal variables in two causal models  $\mathcal{M}$  and  $\mathcal{M}'$ , which belong to the equivalence class up to reparameterization. (a) There are 10 observed samples in which  $Z_1$  or  $Z_2$  has been intervened on. (b) The distribution of causal 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  $\phi_{\mathcal{Z}}$ . (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.

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

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

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

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

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

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

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

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

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

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

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

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

625 If we did not have the causal mechanism switch variable ( $h_i(\mathcal{V}')$ ), the equation above would only  
626 hold if  $s_i$  was constant in parents, which is not the case due to the presence of soft interventions, or if

627  $\frac{\partial s_i(\tilde{e}'_i; \mathcal{E}'_{/i})}{\partial \tilde{\mathcal{E}}'_i} \cdot \frac{\partial \tilde{\mathcal{E}}'_i}{\partial \tilde{\mathcal{E}}'_j} = -\frac{\partial s_i(\tilde{e}'_i; \mathcal{E}'_{/i})}{\partial \mathcal{E}'_{/i}} \cdot \frac{\partial \mathcal{E}'_{/i}}{\partial \tilde{\mathcal{E}}'_j}$ . The latter scenario would imply that  $\frac{\partial \tilde{\mathcal{E}}'_i}{\partial \tilde{\mathcal{E}}'_j} \neq 0$ , hence,  $\tilde{\mathcal{E}}'_i \not\perp \tilde{\mathcal{E}}'_j$ .

628 However, by introducing the causal mechanism switch variable  $\mathcal{V}$  and assuming it is observed, we  
629 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  
630 variables are commonly assumed to be independent in practice. Consequently:

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

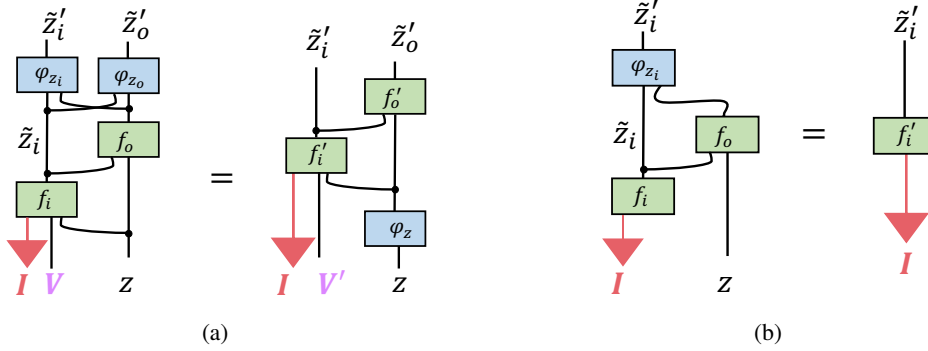


Figure A2: (a) String diagram of the causal variables  $\mathcal{Z}$  and  $\mathcal{Z}'$ . The triangle indicates sampling  $I$  from its distribution. The left-hand side diagram is when  $\phi_{\mathcal{Z}}$  is applied last and the right-hand side diagram is when  $\phi_{\mathcal{Z}}$  is applied 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  $\tilde{Z}'_0$  and the disentangled effect of soft intervention on  $\tilde{Z}_i$  modeled by  $V$ .

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

632 Using the result from previous step that  $\phi_{\mathcal{Z}}$  is a component-wise transformation, the string diagrams  
633 for connections between  $\mathcal{E}$  and  $\mathcal{E}'$  will be as shown in Figure A3.  $\phi_{\mathcal{E}_i}$  will only depend on  $\mathcal{E}_A$ ,  
634 where  $A = anc_i$  is the ancestors of variable  $i$ , and  $e_i$ . Because  $s(e)_{anc_i}$ ,  $s(e)_i$ , and  $s'^{-1}(z')_i$  only  
635 depend on ancestors and  $\phi_{\mathcal{Z}}$  is a component-wise transformation. The first equality in Figure A3  
636 follows from the definition of  $\phi_{\mathcal{E}_i}$ . The second equality holds when we first apply  $\phi_{\mathcal{Z}_A}$  and then apply  
637 the causal mechanisms. It can be concluded from the most right-hand side diagram in Figure A3  
638 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  
639 transformation.



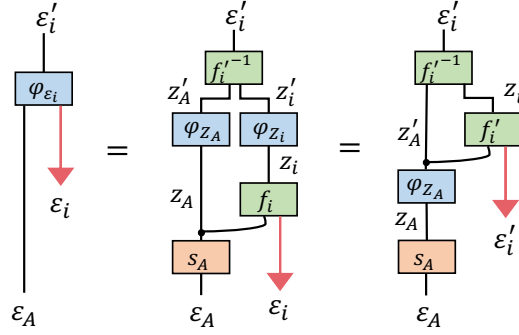


Figure A3: String diagrams for connections between  $\mathcal{E}$  and  $\mathcal{E}'$ . The triangle indicates sampling variables from their corresponding distributions.

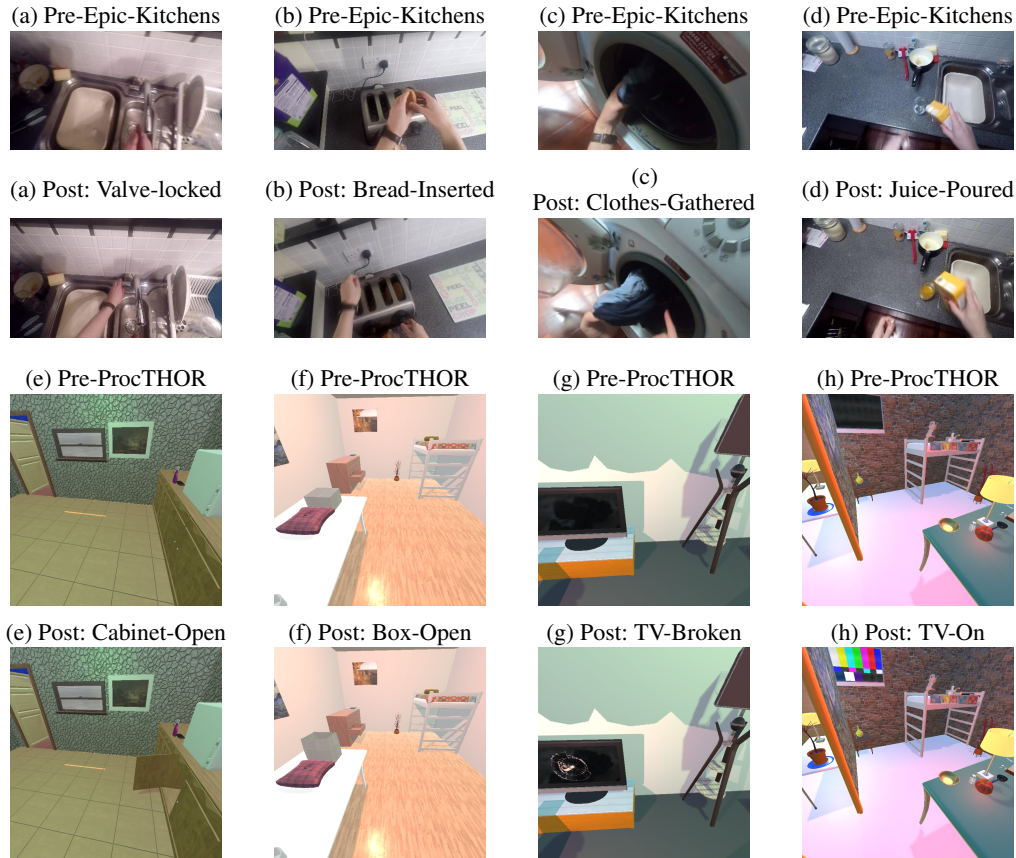


Figure A4: In the Causal-Triplet dataset [19], visual representations capture both pre and post-intervention scenarios. The first two rows showcase data samples from Epic-Kitchens, while the third and fourth rows feature samples from ProcTHOR. Each image in the post-intervention condition is accompanied 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.

## 640 A2 Soft vs. Hard intervention

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

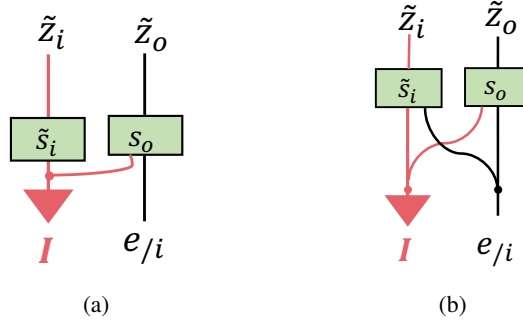


Figure A5: Causal graph models in the presence of Hard (a) and Soft (b) interventions. There are no connections from parents to  $\tilde{Z}_i$  in hard interventions (a). Whereas, parents are connected to  $\tilde{Z}_i$  in soft interventions (b). Let's consider an implicit model and use  $/i$  to denote 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.

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

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

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

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

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

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

## 672 A3 Experiments

673 This section contains additional details about ICRL-SM design architectures, datasets, and experi-  
 674 ments settings.

### 675 A3.1 Datasets

#### 676 A3.1.1 Synthetic

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

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

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

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

### 689 A3.1.2 Causal-Triplet

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

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

ProcTHOR Dataset							
<b>Object</b>	Television	Bed	Bed	Television	Laptop	Book	Box
<b>Action</b>	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										
<b>Object</b>	Tofu	Rice	Hob	Bag	Cupboard	Garlic	Tap	Wrap	Rice	Cheese
<b>Action</b>	Insert	Pour	Wash	Fold	Open	Pat	Move	Check	Transition	Stretch
<b>Object</b>	Wrap	Skin	Button	Lid	Plate	Egg	Sponge	Oil	Water	Dough
<b>Action</b>	Flip	Gather	Press	Lock	Wrap	Drop	Water	Carry	Smell	Mark

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

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

### 701 A3.2 Architecture Design

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

### 714 A3.3 Training

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

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

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

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

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

## 735 A4 Ablation study

### 736 A4.1 Scalability

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

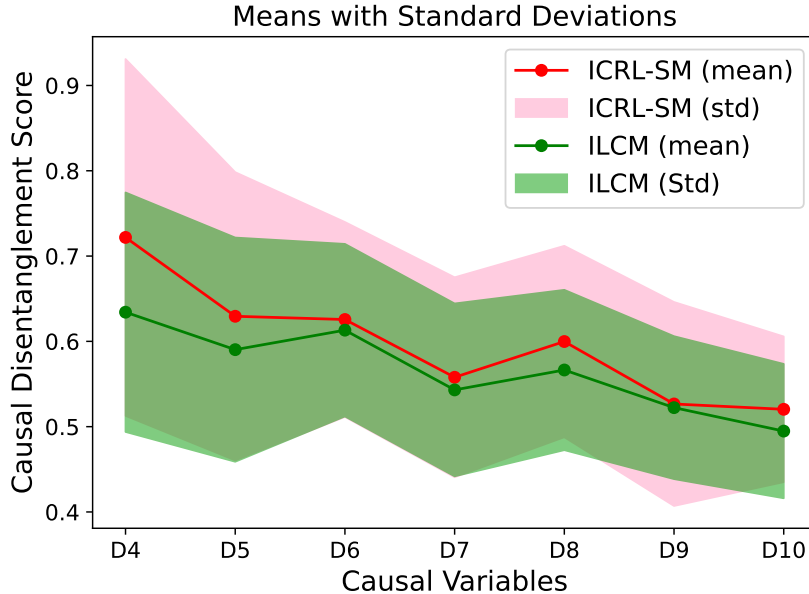


Figure A6: Causal disentanglement for different number of variables

### 746 A4.2 Backbone model

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

749 and object classes. The results are shown in Table A4 and A3. It can be seen from the results that the  
 750 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
<b>ICRL-SM (ours)</b>	<b>0.35</b>	<b>0.04</b>	<b>0.40</b>	<b>0.69</b>

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 casual 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	<b>0.05</b>
	<b>ICRL-SM (ours)</b>	<b>0.35</b>	0.04
ProcTHOR	ENCO [16]	0.16	0.28
	DDS [5]	0.34	0.35
	Fixed-order	0.34	0.38
	<b>ICRL-SM (ours)</b>	<b>0.40</b>	<b>0.69</b>

## 751 **NeurIPS Paper Checklist**

### 752 **1. Claims**

753 Question: Do the main claims made in the abstract and introduction accurately reflect the  
754 paper's contributions and scope?

755 Answer: [\[Yes\]](#)

756 Justification: Our contributions include identifiability of causal models with soft inter-  
757 ventions. In the proposed methods section we give the theory and assumptions for the  
758 identifiability result and in our experiments we evaluate our method using datasets generated  
759 by soft interventions.

760 Guidelines:

- 761 • The answer NA means that the abstract and introduction do not include the claims  
762 made in the paper.
- 763 • The abstract and/or introduction should clearly state the claims made, including the  
764 contributions made in the paper and important assumptions and limitations. A No or  
765 NA answer to this question will not be perceived well by the reviewers.
- 766 • The claims made should match theoretical and experimental results, and reflect how  
767 much the results can be expected to generalize to other settings.
- 768 • It is fine to include aspirational goals as motivation as long as it is clear that these goals  
769 are not attained by the paper.

### 770 **2. Limitations**

771 Question: Does the paper discuss the limitations of the work performed by the authors?

772 Answer: [\[Yes\]](#)

773 Justification: We have some strict assumptions on data generation process and model  
774 which are given in Assumptions 3.3 and 3.1 which may not be plausible to satisfy in some  
775 applications.

776 Guidelines:

- 777 • The answer NA means that the paper has no limitation while the answer No means that  
778 the paper has limitations, but those are not discussed in the paper.
- 779 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 780 • The paper should point out any strong assumptions and how robust the results are to  
781 violations of these assumptions (e.g., independence assumptions, noiseless settings,  
782 model well-specification, asymptotic approximations only holding locally). The authors  
783 should reflect on how these assumptions might be violated in practice and what the  
784 implications would be.
- 785 • The authors should reflect on the scope of the claims made, e.g., if the approach was  
786 only tested on a few datasets or with a few runs. In general, empirical results often  
787 depend on implicit assumptions, which should be articulated.
- 788 • The authors should reflect on the factors that influence the performance of the approach.  
789 For example, a facial recognition algorithm may perform poorly when image resolution  
790 is low or images are taken in low lighting. Or a speech-to-text system might not be  
791 used reliably to provide closed captions for online lectures because it fails to handle  
792 technical jargon.
- 793 • The authors should discuss the computational efficiency of the proposed algorithms  
794 and how they scale with dataset size.
- 795 • If applicable, the authors should discuss possible limitations of their approach to  
796 address problems of privacy and fairness.
- 797 • While the authors might fear that complete honesty about limitations might be used by  
798 reviewers as grounds for rejection, a worse outcome might be that reviewers discover  
799 limitations that aren't acknowledged in the paper. The authors should use their best  
800 judgment and recognize that individual actions in favor of transparency play an impor-  
801 tant role in developing norms that preserve the integrity of the community. Reviewers  
802 will be specifically instructed to not penalize honesty concerning limitations.

### 803 **3. Theory Assumptions and Proofs**

804 Question: For each theoretical result, does the paper provide the full set of assumptions and  
805 a complete (and correct) proof?

806 Answer: [Yes]

807 Justification: We give the full set of our assumptions in the proposed method section and the  
808 detailed proof in Appendix A1.

809 Guidelines:

- 810 • The answer NA means that the paper does not include theoretical results.
- 811 • All the theorems, formulas, and proofs in the paper should be numbered and cross-  
812 referenced.
- 813 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 814 • The proofs can either appear in the main paper or the supplemental material, but if  
815 they appear in the supplemental material, the authors are encouraged to provide a short  
816 proof sketch to provide intuition.
- 817 • Inversely, any informal proof provided in the core of the paper should be complemented  
818 by formal proofs provided in appendix or supplemental material.
- 819 • Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 820 4. Experimental Result Reproducibility

821 Question: Does the paper fully disclose all the information needed to reproduce the main ex-  
822 perimental results of the paper to the extent that it affects the main claims and/or conclusions  
823 of the paper (regardless of whether the code and data are provided or not)?

824 Answer:[Yes]

825 Justification: We provide the full details of our model architecture and training settings in  
826 Appendix A3 and in Section 5.

827 Guidelines:

- 828 • The answer NA means that the paper does not include experiments.
- 829 • If the paper includes experiments, a No answer to this question will not be perceived  
830 well by the reviewers: Making the paper reproducible is important, regardless of  
831 whether the code and data are provided or not.
- 832 • If the contribution is a dataset and/or model, the authors should describe the steps taken  
833 to make their results reproducible or verifiable.
- 834 • Depending on the contribution, reproducibility can be accomplished in various ways.  
835 For example, if the contribution is a novel architecture, describing the architecture fully  
836 might suffice, or if the contribution is a specific model and empirical evaluation, it may  
837 be necessary to either make it possible for others to replicate the model with the same  
838 dataset, or provide access to the model. In general, releasing code and data is often  
839 one good way to accomplish this, but reproducibility can also be provided via detailed  
840 instructions for how to replicate the results, access to a hosted model (e.g., in the case  
841 of a large language model), releasing of a model checkpoint, or other means that are  
842 appropriate to the research performed.
- 843 • While NeurIPS does not require releasing code, the conference does require all submis-  
844 sions to provide some reasonable avenue for reproducibility, which may depend on the  
845 nature of the contribution. For example
  - 846 (a) If the contribution is primarily a new algorithm, the paper should make it clear how  
847 to reproduce that algorithm.
  - 848 (b) If the contribution is primarily a new model architecture, the paper should describe  
849 the architecture clearly and fully.
  - 850 (c) If the contribution is a new model (e.g., a large language model), then there should  
851 either be a way to access this model for reproducing the results or a way to reproduce  
852 the model (e.g., with an open-source dataset or instructions for how to construct  
853 the dataset).
  - 854 (d) We recognize that reproducibility may be tricky in some cases, in which case  
855 authors are welcome to describe the particular way they provide for reproducibility.  
856 In the case of closed-source models, it may be that access to the model is limited in  
857 some way (e.g., to registered users), but it should be possible for other researchers  
858 to have some path to reproducing or verifying the results.



859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our anonymized codes which contains the necessary scripts and instructions to run the experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the full details of our model architecture and training settings in Appendix A3 and in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In our synthetic experiments we initialized the causal graph in the dataests with different seeds. The results of these different seeds are provided in Table 2 and Figure A6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.



- 911 • The factors of variability that the error bars are capturing should be clearly stated (for  
912 example, train/test split, initialization, random drawing of some parameter, or overall  
913 run with given experimental conditions).
- 914 • The method for calculating the error bars should be explained (closed form formula,  
915 call to a library function, bootstrap, etc.)
- 916 • The assumptions made should be given (e.g., Normally distributed errors).
- 917 • It should be clear whether the error bar is the standard deviation or the standard error  
918 of the mean.
- 919 • It is OK to report 1-sigma error bars, but one should state it. The authors should  
920 preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis  
921 of Normality of errors is not verified.
- 922 • For asymmetric distributions, the authors should be careful not to show in tables or  
923 figures symmetric error bars that would yield results that are out of range (e.g. negative  
924 error rates).
- 925 • If error bars are reported in tables or plots, The authors should explain in the text how  
926 they were calculated and reference the corresponding figures or tables in the text.

## 927 8. Experiments Compute Resources

928 Question: For each experiment, does the paper provide sufficient information on the com-  
929 puter resources (type of compute workers, memory, time of execution) needed to reproduce  
930 the experiments?

931 Answer: [Yes]

932 Justification: The details are given in Appendix A3.

933 Guidelines:

- 934 • The answer NA means that the paper does not include experiments.
- 935 • The paper should indicate the type of compute workers CPU or GPU, internal cluster,  
936 or cloud provider, including relevant memory and storage.
- 937 • The paper should provide the amount of compute required for each of the individual  
938 experimental runs as well as estimate the total compute.
- 939 • The paper should disclose whether the full research project required more compute  
940 than the experiments reported in the paper (e.g., preliminary or failed experiments that  
941 didn't make it into the paper).

## 942 9. Code Of Ethics

943 Question: Does the research conducted in the paper conform, in every respect, with the  
944 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

945 Answer: [Yes]

946 Justification:

947 Guidelines:

- 948 • The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- 949 • If the authors answer No, they should explain the special circumstances that require a  
950 deviation from the Code of Ethics.
- 951 • The authors should make sure to preserve anonymity (e.g., if there is a special consid-  
952 eration due to laws or regulations in their jurisdiction).

## 953 10. Broader Impacts

954 Question: Does the paper discuss both potential positive societal impacts and negative  
955 societal impacts of the work performed?

956 Answer: [NA]

957 Justification:

958 Guidelines:

- 959 • The answer NA means that there is no societal impact of the work performed.
- 960 • If the authors answer NA or No, they should explain why their work has no societal  
961 impact or why the paper does not address societal impact.

- 962 • Examples of negative societal impacts include potential malicious or unintended uses  
963 (e.g., disinformation, generating fake profiles, surveillance), fairness considerations  
964 (e.g., deployment of technologies that could make decisions that unfairly impact specific  
965 groups), privacy considerations, and security considerations.
- 966 • The conference expects that many papers will be foundational research and not tied  
967 to particular applications, let alone deployments. However, if there is a direct path to  
968 any negative applications, the authors should point it out. For example, it is legitimate  
969 to point out that an improvement in the quality of generative models could be used to  
970 generate deepfakes for disinformation. On the other hand, it is not needed to point out  
971 that a generic algorithm for optimizing neural networks could enable people to train  
972 models that generate Deepfakes faster.
- 973 • The authors should consider possible harms that could arise when the technology is  
974 being used as intended and functioning correctly, harms that could arise when the  
975 technology is being used as intended but gives incorrect results, and harms following  
976 from (intentional or unintentional) misuse of the technology.
- 977 • If there are negative societal impacts, the authors could also discuss possible mitigation  
978 strategies (e.g., gated release of models, providing defenses in addition to attacks,  
979 mechanisms for monitoring misuse, mechanisms to monitor how a system learns from  
980 feedback over time, improving the efficiency and accessibility of ML).

## 981 11. Safeguards

982 Question: Does the paper describe safeguards that have been put in place for responsible  
983 release of data or models that have a high risk for misuse (e.g., pretrained language models,  
984 image generators, or scraped datasets)?

985 Answer: [NA]

986 Justification:

987 Guidelines:

- 988 • The answer NA means that the paper poses no such risks.
- 989 • Released models that have a high risk for misuse or dual-use should be released with  
990 necessary safeguards to allow for controlled use of the model, for example by requiring  
991 that users adhere to usage guidelines or restrictions to access the model or implementing  
992 safety filters.
- 993 • Datasets that have been scraped from the Internet could pose safety risks. The authors  
994 should describe how they avoided releasing unsafe images.
- 995 • We recognize that providing effective safeguards is challenging, and many papers do  
996 not require this, but we encourage authors to take this into account and make a best  
997 faith effort.

## 998 12. Licenses for existing assets

999 Question: Are the creators or original owners of assets (e.g., code, data, models), used in  
1000 the paper, properly credited and are the license and terms of use explicitly mentioned and  
1001 properly respected?

1002 Answer: [Yes]

1003 Justification:

1004 Guidelines:

- 1005 • The answer NA means that the paper does not use existing assets.
- 1006 • The authors should cite the original paper that produced the code package or dataset.
- 1007 • The authors should state which version of the asset is used and, if possible, include a  
1008 URL.
- 1009 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 1010 • For scraped data from a particular source (e.g., website), the copyright and terms of  
1011 service of that source should be provided.
- 1012 • If assets are released, the license, copyright information, and terms of use in the  
1013 package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets)  
1014 has curated licenses for some datasets. Their licensing guide can help determine the  
1015 license of a dataset.

- 1016
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
  - If this information is not available online, the authors are encouraged to reach out to the asset’s creators.
- 1017
- 1018
- 1019

1020 **13. New Assets**

1021 Question: Are new assets introduced in the paper well documented and is the documentation  
1022 provided alongside the assets?

1023 Answer: [Yes]

1024 Justification: We only have a code repository for replicating experiments and we have  
1025 submitted the anonymized zip file with our submission.

1026 Guidelines:

- The answer NA means that the paper does not release new assets.
  - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
  - The paper should discuss whether and how consent was obtained from people whose asset is used.
  - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.
- 1027
- 1028
- 1029
- 1030
- 1031
- 1032
- 1033
- 1034

1035 **14. Crowdsourcing and Research with Human Subjects**

1036 Question: For crowdsourcing experiments and research with human subjects, does the paper  
1037 include the full text of instructions given to participants and screenshots, if applicable, as  
1038 well as details about compensation (if any)?

1039 Answer: [NA]

1040 Justification:

1041 Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
  - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
  - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.
- 1042
- 1043
- 1044
- 1045
- 1046
- 1047
- 1048
- 1049

1050 **15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human  
1051 Subjects**

1052 Question: Does the paper describe potential risks incurred by study participants, whether  
1053 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)  
1054 approvals (or an equivalent approval/review based on the requirements of your country or  
1055 institution) were obtained?

1056 Answer: [NA]

1057 Justification:

1058 Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
  - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
  - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
  - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.
- 1059
- 1060
- 1061
- 1062
- 1063
- 1064
- 1065
- 1066
- 1067
- 1068