Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning

Yuejiang Liu[1, 2, *] YUEJIANG.LIU@EPFL.CH
Alexandre Alahi[2] ALEXANDRE.ALAHI@EPFL.CH
Chris Russell[1] CMRUSS@AMAZON.DE
Max Horn[1] HORNMAX@AMAZON.DE
Dominik Zietlow[1] ZIETLD@AMAZON.DE
Bernhard Schölkopf[1, 3] BS@TUEBINGEN.MPG.DE
Francesco Locatello[1] LOCATELF@AMAZON.DE
1. Amazon, Tübingen, Germany
2. École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
3. Max Planck Institute for Intelligent Systems, Tübingen, Germany

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Recent years have seen a surge of interest in learning high-level causal representations from low-level image pairs under interventions. Yet, existing efforts are largely limited to simple synthetic settings that are far away from real-world problems. In this paper, we present `Causal Triplet`, a causal representation learning benchmark featuring not only visually more complex scenes, but also two crucial desiderata commonly overlooked in previous works: (i) an actionable counterfactual setting, where only certain object-level variables allow for counterfactual observations whereas others do not; (ii) an interventional downstream task with an emphasis on out-of-distribution robustness from the independent causal mechanisms principle. Through extensive experiments, we find that models built with the knowledge of disentangled or object-centric representations significantly outperform their distributed counterparts. However, recent causal representation learning methods still struggle to identify such latent structures, indicating substantial challenges and opportunities for future work. Our code and datasets will be available at https://sites.google.com/view/causaltriplet.

1. Introduction

Causal representation learning, which strives to discover and represent high-level causal variables from low-level sensory observations, is a critical component for combining modern representation learning with classical causal modeling (Schölkopf et al., 2021). Yet, this is a highly challenging and even ill-posed problem in the unsupervised setting, due to the limits of i.i.d. observational data and ambiguities in levels of abstractions. A promising remedy is to exploit independent causal mechanisms (Schölkopf et al., 2012; Parascandolo et al., 2018) and sparse mechanism shifts (Schölkopf et al., 2021) between pairs of observations under interventions. Significant progress has been made in both theoretical underpinnings and practical algorithms (Locatello et al., 2020a; von Kügelgen et al., 2021; Lachapelle et al., 2022; Lippe et al., 2022c; Brehmer et al., 2022). However, existing efforts are still confined to simple synthetic data sets that are far from practical problems. In this work,

[*] Most work done during an internship at Amazon.

we aim to bridge this gap by revisiting recent hypotheses and methods in more realistic settings. Beside much richer visual complexity of high-dimensional data with heavy occlusions, camera motion, ego motion, and visual scenes with multiple objects, we incorporate two key properties.

Desideratum 1: We take the perspective of embodied agents acting in the world (Cohen, 2022; Deitke et al., 2022a), blurring the line between interventions and counterfactuals. In fact, we notice that assuming that only one/few variables change as a result of a sparse shift that maintains the noise realization of all other variables is a counterfactual and not a proper intervention. Arguably, *perfect counterfactuals* are difficult to observe, and not all variables may allow observable counterfactuals, invalidating the identifiability results of existing methods (Locatello et al., 2020a; von Kügelgen et al., 2021; Klindt et al., 2021; Brehmer et al., 2022). Instead, we propose a practical alternative, where only specific variables corresponding to physical and manipulable objects allow for counterfactual observations while others do not (*e.g.*, camera view, self-occlusions, global scene properties that can change over time). We call this "*actionable counterfactuals*" – a setting that can naturally emerge in various situations, such as embodied agents learning to discover causal models through active interventions and human-object interactions observed from a first-person camera view.

Desideratum 2: We argue that the notion of causal representation learning should also move beyond identifying underlying causal factors, and more explicitly account for object affordances that support downstream tasks involving interventions and reasoning. We thus propose a new task modeling high-level actions between image pairs, with a particular emphasis on out-of-distribution robustness from the independent causal mechanisms principle (*i.e.*, if $X \to Y$, then $P(Y|X)$ remains valid under interventions on X). This task necessitates the discovery of not only causal variables that can be independently manipulated but also causal mechanisms behind actions that govern abstract transformations of object states. Solving it can be an important step towards *Intervention-Centric Causal Representation Learning*, where interventions are associated with (and represented alongside) objects, changing an object's default dynamics whenever they take place.

In light of these two desiderata, we introduce a new benchmark for causal representation learning, named `Causal Triplet`. It features (i) pairs of images with high visual complexity and variability, (ii) real-world actions inducing sparse changes in the underlying structures of natural scenes. Through extensive experiments, we show that intervention models built with causally structured representations (*e.g.*, disentangled and object-centric) can substantially outperform distributed representation counterparts. However, recent approaches still have difficulty in correctly discovering the latent structures. In particular, we observe that learning independent causal mechanisms is highly challenging in the presence of spurious correlations between actions and object attributes, introducing shortcuts such as recognizing an object based on the action performed or vice-versa that do not generalize to unseen compositions. Overall, we hope the proposed benchmark will foster advancement of intervention-centric causal representation learning towards real-world contexts.

2. Related Work

Early efforts towards causal representation learning revolve around the notion of *independence*. One line of works attempts to learn disentangled representations (Chen et al., 2016; Higgins et al., 2017; Rolinek et al., 2019; Locatello et al., 2019; Goyal et al., 2021), where all causal variables are assumed *statistically independent*. Unfortunately, causal structures of real-world observations are often non-trivial, introducing strong correlations between latent variables and undermining the foundation of disentanglement (Träuble et al., 2021). Another prominent branch lies in object-

	Causal3DIdent (von Kügelgen et al., 2021)	Causal Pinball (Lippe et al., 2022a)	Interventional Pong (Lippe et al., 2022c)	Causal World (Ahmed et al., 2021)	Causal Triplet (ours)
Scale Variability	✗	✗	✗	✗	✓
Shape Variability	✗	✗	✗	✗	✓
Illumination Variability	✗	✗	✗	✗	✓
Camera Motion	✗	✗	✗	✗	✓
Complex Texture	✗	✗	✗	✗	✓
Object Occlusion	✗	✗	✗	✓	✓
Object Number	one	few	few	few	many
Downstream Task	factor	factor	factor	action	action

Table 1: Benchmark comparison for causal representation learning.

centric representation learning (Greff et al., 2017; Burgess et al., 2019; Greff et al., 2019; Locatello et al., 2020b), which seeks to decompose visual scenes into a set of individual objects that can be *independently manipulated* (Yang et al., 2020), can *move independently* (Atzmon et al., 2020), or *independently re-appear* across samples (Yang et al., 2021). However, it remains unclear when and to what extent objects can be viewed as causal variables.

More recently, there has been a growing interest in learning causal representations from interventions (Ahuja et al., 2022b). In particular, several recent works exploit the *sparsity* of mechanism shifts between paired observations (Locatello et al., 2020a; Zimmermann et al., 2021; von Kügelgen et al., 2021; Ahuja et al., 2022a; Brehmer et al., 2022) or in a temporarily intervened sequence (Klindt et al., 2021; Lippe et al., 2022a,b,c; Lachapelle et al., 2022) to identify the underlying causal variables. Nevertheless, these works are still restricted to simple toy datasets and impractical intervention conditions. In contrast, we propose a new benchmark with high visual complexity (*e.g.*, camera view, self-occlusions, non-static global scene properties) and realistic interventions in the form of actionable counterfactuals from the embodied agent perspective.

3. Benchmark Design

In this section, we present `Causal Triplet`, a new benchmark for causal representation learning in visually complex settings. We will first describe our designed task through the lens of causality, and subsequently, discuss the key properties of the collected datasets.

3.1. Benchmark Task

Modeling paired observations under known or unknown interventions has been a common setting for causal representation learning (Locatello et al., 2020a; von Kügelgen et al., 2021; Lachapelle et al., 2022; Brehmer et al., 2022). Yet, modeling interventions themselves remains largely unexplored. In fact, learning representations of high-level actions is deeply rooted in a variety of practical problems, from perceiving human actions (Fathi et al., 2011) to building interventional world models (Ha and Schmidhuber, 2018; Lei et al., 2022). To bridge this gap, we extend the paired setup to an intervention modeling task, with an emphasis on out-of-distribution robustness.

Problem setting. Consider the data generating process $\mathbf{x} = g(\mathbf{z})$, where $\mathbf{x} \in \mathbb{R}^d$ is a high-dimensional image observation, $\mathbf{z} \in \mathbb{R}^l$ is a set of latent factors of variation, and $g : \mathbb{R}^l \to \mathbb{R}^d$ is the underlying generative mechanism. We assume that the set of latent factors consists of both global scene-level variables z_s and local object-level variables z_n^k, where n is the object index and k is the latent index within the object. Similar to prior work (Lippe et al., 2022b,c), we consider an

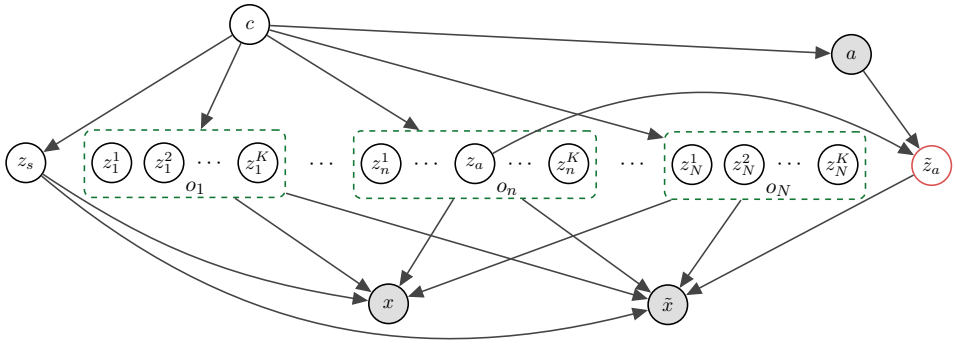


Figure 1: Causal graph for a pair of scene observations $(x, \tilde{x})$ before and after an action a. The data generating process of each raw observation is described by a set of latent factors, including global scene-level factors z_s and local object-level factors z_n^k, which are statistically dependent due to unobserved confounders c. The action is assumed to sparsely influence only one (or a few) object-level factor z_a in the scene. The other latent factors may stay constant in photo-realistic simulations but vary over time in real-world observations.

intervention process $\tilde{\mathbf{z}} = h(\mathbf{z}, \mathbf{a})$, where action $\mathbf{a} \in \mathcal{A}$ is a categorical variable, and $h : \mathbb{R}^l \times \mathcal{A} \to \mathbb{R}^l$ is a deterministic transition function. We assume that each action affects one or a few object-level latent factors, resulting in paired observations $(\mathbf{x}, \tilde{\mathbf{x}})$ before and after intervention. Additionally, we assume that the latent factors and actions have dependences due to unobserved confounding $\mathbf{c}$. For clarity, we refer to the latent factors affected by the action $\mathbf{a}$ as $\mathbf{z}_a$ in the rest of the paper. Fig. 1 shows the causal graph encompassing our assumptions about the data generating process.

Unlike previous works aimed at identifying latent factors, we focus on downstream reasoning about the categories of high-level actions given image pairs, *i.e.*, modeling the conditional probability $\mathbb{P}(\mathbf{a} \mid \mathbf{x}, \tilde{\mathbf{x}})$. Crucially, we assume that the effect of an action $\mathbb{P}(\tilde{\mathbf{z}}_a \mid \mathbf{z}_a, \mathbf{a})$ stays invariant, whereas the joint distribution of latent factors may change between training and test data, *i.e.*, $\mathbb{P}_{\text{tr}}(\mathbf{a}, \mathbf{z}) \neq \mathbb{P}_{\text{te}}(\mathbf{a}, \mathbf{z})$. Such distribution shifts commonly arise in practice, *e.g.*, the training dataset contains only a subset of object classes or object-action combinations that a learning system encounters during deployment. Inferring actions in this setting requires discoveries of both causal variables that can be independently manipulated and causal mechanisms behind actions that govern abstract transformations.

Paired observations as actionable counterfactuals. Paired observations are often assumed to share the same noise realization in prior works (Locatello et al., 2020a; von Kügelgen et al., 2021; Brehmer et al., 2022). We remark that this amounts to a perfect counterfactual, which is largely lacking in nature (Holland, 1986) and only in part plausible for embodied agents that learn to discover causal models through interactions (Cohen, 2022), *e.g.*, a robot intentionally keeps almost everything unchanged while manipulating an object in the environment. Of course, this will allow only a subset of all possible counterfactuals, restricting to those that can be realized by the actions of the embodied agent. We call these "*actionable counterfactuals*". Unfortunately, even this view would be too restrictive. When it comes to real-world observations, many factors (*e.g.*, camera view, object occlusions) may vary from time to time. In our benchmark, we assume that the agent can only perform actionable counterfactuals, where most objects are not manipulated but the global properties

Figure 2: Causal triplet samples collected from photo-realistic simulations.

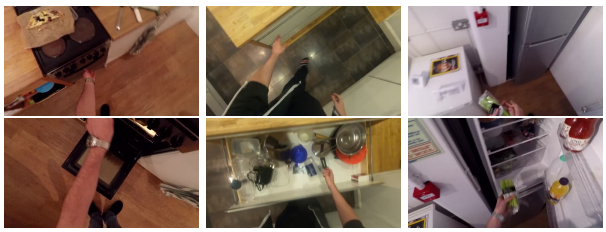


Figure 3: Causal triplet samples collected from real-world observations.

of the visual scene may change. This is a step towards *Intervention-Centric Causal Representation Learning*, where interventions are associated with (and represented alongside) objects.

Sparse mechanism shifts between image representations. Given the causal graph in Fig. 1, the joint distribution of the latent factors can be factorized into a number of causal conditionals,

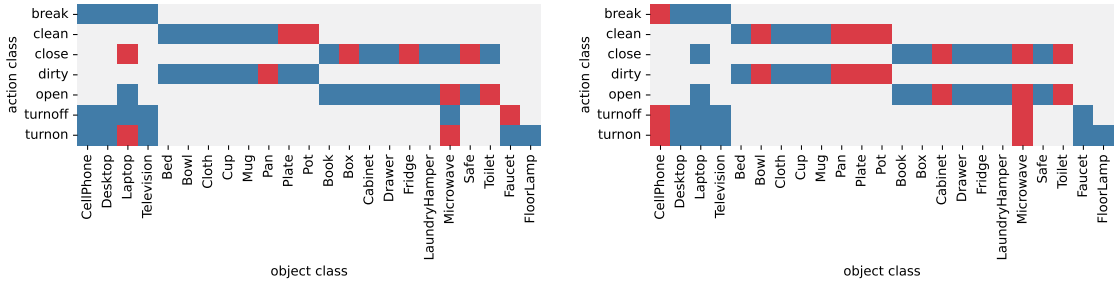
$$\mathbb{P}(\mathbf{z}, \mathbf{c}) = \mathbb{P}(\mathbf{c})\mathbb{P}(\mathbf{z}_s \mid \mathbf{c}) \prod_{n=1}^{N} \mathbb{P}(\mathbf{z}_n^1, \mathbf{z}_n^2, \cdots, \mathbf{z}_n^K \mid \mathbf{c}),$$

where only one/few conditionals are intervened between paired observations (Schölkopf et al., 2021). Intuitively, an action does not manipulate all objects in a scene or all properties of the manipulated object at the same time. Instead, it may sparsely affect a small subset of latent components in a causally structured scene representation (Locatello et al., 2020a; Lachapelle et al., 2022).

Independent causal mechanisms for action representations. Modeling the effect of an action essentially amounts to learning the causal mechanism $p(\tilde{\mathbf{z}}_a \mid \mathbf{z}_a, \mathbf{a})$. By the principle of Independent Causal Mechanism (Schölkopf et al., 2012; Peters et al., 2017), the conditional distribution of each variable given its causes does not inform or influence the other conditional distributions. In other words, $\mathbb{P}(\tilde{\mathbf{z}}_a \mid \mathbf{z}_a, \mathbf{a})$ should remain invariant when other mechanisms, such as $\mathbb{P}(\mathbf{a} \mid \mathbf{c})$ and $\mathbb{P}(\mathbf{z} \mid \mathbf{c})$, change. We hypothesize that this principle is also essential in tackling practical problems like $\mathbb{P}(\mathbf{a} \mid \mathbf{x}, \tilde{\mathbf{x}})$ that is anti-causal. In the presence of distribution shifts resulting from changes of unobserved confounders $\mathbf{c}$, the action representation that takes into account all the latent factors may break due to the non-stability of the statistical dependencies between the action class and unintervened latent factors. In contrast, the action representation only focused on the transition between the intervened latent factors $\mathbb{P}(\mathbf{a} \mid \mathbf{z}_a, \tilde{\mathbf{z}}_a)$ is expected to strongly generalize.

3.2. Benchmark Data

Existing datasets for causal representation learning (von Kügelgen et al., 2021; Ahmed et al., 2021; Lippe et al., 2022a,c), as listed in Tab. 1, are heavily simplified in many aspects compared to real-world problems. To bridge this gap, our benchmark introduces two new datasets, one collected from a photo-realistic simulator of embodied agents and the other repurposed from a real-world video dataset of human-object interactions. The former one contains 7 types of actions manipulating 24 types of objects in $10k$ distinct ProcTHOR indoor environments (Deitke et al., 2022b), resulting in $100k$ pairs of images. Each original image has a resolution of 672×672, and is further cropped and/or resize to a resolution of 224×224 for the experiments in §4. The latter consists of 2632 image pairs, collected under a similar setup from the Epic-Kitchens dataset (Damen et al., 2022).



(*a*) unseen compositions (*b*) unseen objects

Figure 4: Data splits in `Causal Triplet`. We split the training (blue) and test (red) data into two disjoint groups with different (a) object-action combinations or (b) object classes.

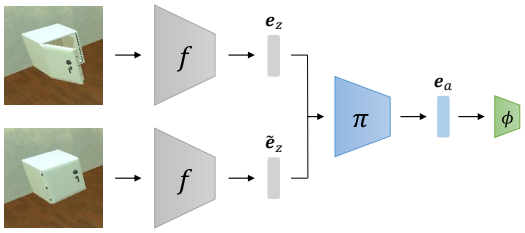


Figure 5: Overview of our benchmark model for reasoning about the action class from a pair of single-object images. We consider different regularizers for learning robust image representation $\mathbf{e}_z$ and action representation $\mathbf{e}_a$.

	compositional shift		systematic shift	
	ID	OOD	ID	OOD
vanilla	0.96 ± 0.01	0.36 ± 0.13	0.94 ± 0.02	0.48 ± 0.08
ICM	0.95 ± 0.01	0.41 ± 0.15	0.94 ± 0.02	0.50 ± 0.09
SMS	0.96 ± 0.01	0.47 ± 0.18	0.94 ± 0.02	0.54 ± 0.07

Table 2: Results of different methods in single-object simulated images. The action representation trained with the independence regularizer (ICM) and the image representation trained with the sparsity regularizer (SMS) result in significantly higher OOD accuracies.

Similar to recent empirical studies towards causal representation learning (van Steenkiste et al., 2019; Dittadi et al., 2020; Montero et al., 2021; Zhang et al., 2021; Liu et al., 2022; Dittadi et al., 2022), we explicitly split the collected data into two groups that differ in the joint distribution of action category $\mathbf{a} \in \mathcal{A}$ and object category $\mathbf{o} \in \mathcal{O}$. More specifically, we consider two types of distribution shifts:

- *compositional shifts*: the sets of object class are identical between training and test, *i.e.*, $\mathcal{O}_{\text{tr}} = \mathcal{O}_{\text{te}}$, but the sets of object-action composition are disjoint, *i.e.*, $(\mathcal{A}_{\text{tr}} \times \mathcal{O}_{\text{tr}}) \cap (\mathcal{A}_{\text{te}} \times \mathcal{O}_{\text{te}}) = \varnothing$
- *systematic shifts*: the training and test sets of object class are disjoint, *i.e.*, $\mathcal{O}_{\text{tr}} \cap \mathcal{O}_{\text{te}} = \varnothing$

We partition each dataset into three parts: 60% for training, 20% for in-distribution (ID) testing, and 20% for out-of-distribution (OOD) testing. More dataset details are summarized in Appendix A.

4. Experiments and Results

The main goal of our experiments is to examine the potential and limitations of recent hypotheses and methods for causal representation learning on `Causal Triplet`. In particular, we seek to understand the following two questions:

- Are causally structured representations (*e.g.*, disentangled, object-centric) helpful for reasoning about interventions between paired observations?
- How effective are recent models at identifying the latent structures in the proposed datasets?

To this end, we consider a variety of visual representations, including

- modern distributed representations, *e.g.*, ResNet (He et al., 2016), CLIP (Radford et al., 2021);
- oracle structured representations, *e.g.*, expert knowledge of disentanglement / object-centric;
- learned structured representations, *e.g.*, Slot Attention (Locatello et al., 2020b), GroupViT (Xu et al., 2022).

We systematically examine their performance in four different settings with growing complexities:

- from compositional shifts (§4.1.1) to systematic shifts (§4.1.2);
- from single-object images (§4.1.2) to multi-object scenes (§4.2.1);
- from photo-realistic simulations (§4.2.1) to real-world observations (§4.2.2).

To model the effect of high-level actions between paired images, we consider neural networks made up of three modules: (i) an image encoder $f(\cdot)$ that extracts an abstract representation of each input image, *i.e.*, $\mathbf{e}_z = f(\mathbf{x})$, (ii) an action encoder $\pi(\cdot)$ that reasons about the relation between paired image embeddings, *i.e.*, $\mathbf{e}_a = \pi(\mathbf{e}_z, \tilde{\mathbf{e}}_z)$, (iii) a classification head $\phi(\cdot)$ that infers the probabilities for each action class from the action embedding.

4.1. Simulated Single-object Images

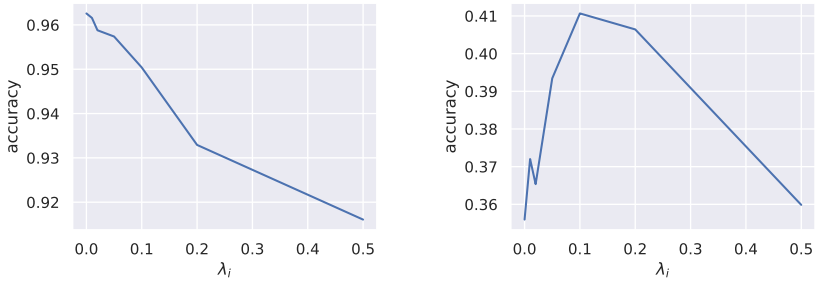
4.1.1. COMPOSITIONAL DISTRIBUTION SHIFTS

Setup. We first consider intervention modeling in single-object images collected from photo-realistic simulations. As shown in Fig. 4(*a*), the training and test data are split into two disjoint groups that differ in object-action combinations. We build the image encoder $f(\cdot)$ with a standard ResNet-18, followed by a two-layer MLP projection head that outputs a 64-dimensional feature vector. To model the action between an image pair, we concatenate the feature vectors obtained from both images, and feed them into a downstream two-layer MLP $\pi(\cdot)$ to extract a 64-dimensional action representation. We finally pass the action representation through a linear classifier $\phi(\cdot)$ that outputs the probabilities for each action class. By default, the model is trained to minimize a standard classification loss $\mathcal{L}_a$, *i.e.*, cross-entropy between the predicted class probabilities $\hat{\mathbf{a}}$ and the action label $\mathbf{a}$.

Vanilla baseline. We start with a vanilla baseline trained in a standard supervised manner. As shown in Tab. 2, the vanilla baseline is highly accurate ($\sim 96\%$) on the ID test set, but generalizes poorly ($\sim 36\%$) to unseen compositions. The large gap is unsurprising though, given that the action label is highly correlated with the object label in the training set, and this correlation varies drastically between the ID and OOD settings. In order to understand the degree to which the vanilla baseline exploits such spurious correlations, we conduct a post-hoc analysis: freeze the trained model and train another classification head $\psi(\cdot)$ to predict the object label. The object classifier built on top of the frozen action embedding turns out to be quite accurate ($\sim 87\%$) on the test set, revealing the high reliance of the learned action embedding on the non-causal object features.

Independence regularizer. To prevent an intervention model from absorbing the spurious correlations demonstrated above, we next consider regularizing the model by minimizing the mutual information between the action embedding and the object class. More specifically, we use the following adversarial training regularizer as a proxy,

$$\min_{f,\pi,\phi} \max_{\psi} \mathcal{L}_a(f, \pi, \phi) - \lambda_i \mathcal{L}_o(f, \pi, \psi) \tag{1}$$



(*a*) ID (*b*) OOD

Figure 6: Effect of the independence regularizer on compositional generalization in single-object images. Action representations containing less information about object classes (larger λ_i) result in smaller performance gaps between the ID and OOD sets that differ in object-action compositions.

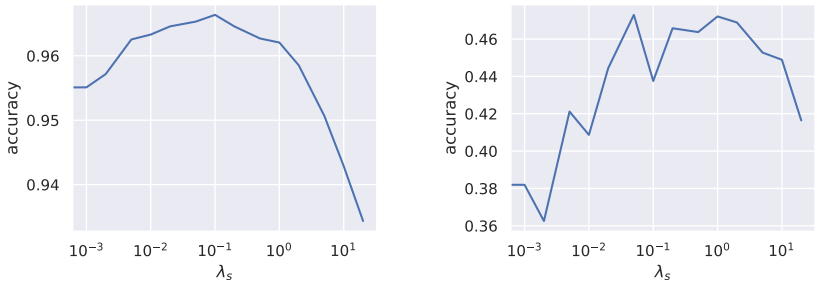
where $\mathcal{L}_o$ is a cross-entropy loss for object classification, and λ_i is the coefficient of the independence regularizer. The adversarial regularizer shares the same spirit with the Hilbert-Schmidt Independence Criterion (HSIC) used in recent works for compositional object recognition (Atzmon et al., 2020; Ruis et al., 2021). We opt for adversarial training as a natural extension of the post-hoc analysis in the previous section. Fig. 6 shows that, while the independence regularizer constrains the performance on the ID test set, it greatly improves the OOD accuracy from 36% to 41%, resulting in $\sim 14\%$ relative performance gain on unseen compositions.

Nevertheless, it is worth noting that the independence regularizer relies on two critical assumptions: (i) the attribute spuriously correlated with the action label is known a priori; (ii) the spurious attribute comes with detailed annotations in the training dataset. Unfortunately, it is typically impractical to examine all possible spurious correlations and gather their corresponding annotations. We next consider another family of regularizations that does not rely on such prior knowledge.

Sparsity regularizer. We further investigate the impact of visual representation structures on the robustness of intervention models. As discussed in §3.1, the action between paired images tends to affect only one or a few generative factors in the causal/disentangled factorization. Motivated by this hypothesis, we introduce a regularizer that imposes sparse changes in a block-disentangled image representation before and after intervention,

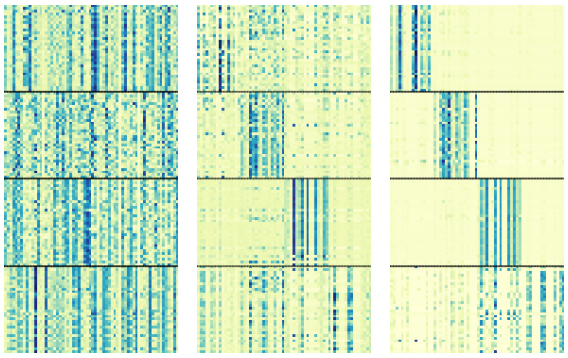
$$\min_{f,\pi,\phi} \mathcal{L}_a(f,\pi,\phi) + \lambda_s \mathcal{L}_s(f,\pi), \quad \mathcal{L}_s = \sum_{k \in \Omega \setminus \Omega_a} ||\mathbf{e}_z^k - \tilde{\mathbf{e}}_z^k||, \tag{2}$$

where k is the block index, Ω is the entire set of latent blocks, and Ω_a is the subset of latent blocks affected by the action, and λ_s is the coefficient of the sparsity regularizer. Fig. 7 and Fig. 8 shows that the effect of the sparsity regularizer on test accuracy and visual representations. Notably, the sparsity regularizer with a moderate coefficient ($\lambda_s \approx 0.1$) demonstrates advantages on both the ID and OOD test sets. In particular, compared with the vanilla baseline in Tab. 2, the sparsity-driven disentanglement lifts the OOD accuracy by nearly 30%. Nevertheless, same as Lippe et al. (2022b,c), the sparsity regularizer requires variable-level intervention labels, *i.e.*, not only is the category of each action labeled, but so is its influence on each latent factor. While this is partially feasible in our simulated data due to the known symmetry between some actions (open vs. close,



(*a*) ID (*b*) OOD

Figure 7: Effect of the sparsity regularizer on compositional generalization in single-object images. Visual representations undergoing sparser changes between paired observations (larger λ_s) result in smaller performance gaps between the ID and OOD sets that differ in object-action compositions.



(*a*) $\lambda_s = 0.0$ (*b*) $\lambda_s = 0.1$ (*c*) $\lambda_s = 1.0$

Figure 8: Effect of the sparsity regularizer on image representation. We randomly sample 32 image pairs from each symmetric group of actions (open vs. close, turnon vs. turnoff, clean vs. dirty, break). We visualize the difference of the encoded feature pairs $||\mathbf{e}_z - \tilde{\mathbf{e}}_z||$ in each row. Larger regularizer coefficients λ_s result in sparser feature changes between paired observations and more distinct block-disentanglement across action classes.

dirty vs. clean, turnon vs. turnoff), collecting such detailed annotations in real-world problems can be difficult.

4.1.2. SYSTEMATIC DISTRIBUTION SHIFTS

Setup. We further consider systematic distribution shifts between training and test in simulated single-object images. As shown in Figure 4(*b*), the object class in the ID and OOD data splits are sampled from two disjoint groups. Same as in the previous section, we compare intervention models trained with different loss functions.

Results. Tab. 2 summarizes the results on the ID and OOD test sets over 10 different runs. Similar to the observations in §4.1.1, both the independence and sparsity regularizers improve robustness on the OOD test set without sacrificing the ID accuracy. Nevertheless, regularizing the action representation to be independent of the object class becomes less effective for robustness under systematic shifts than it is under compositional shifts. We conjecture this is because the major challenge here

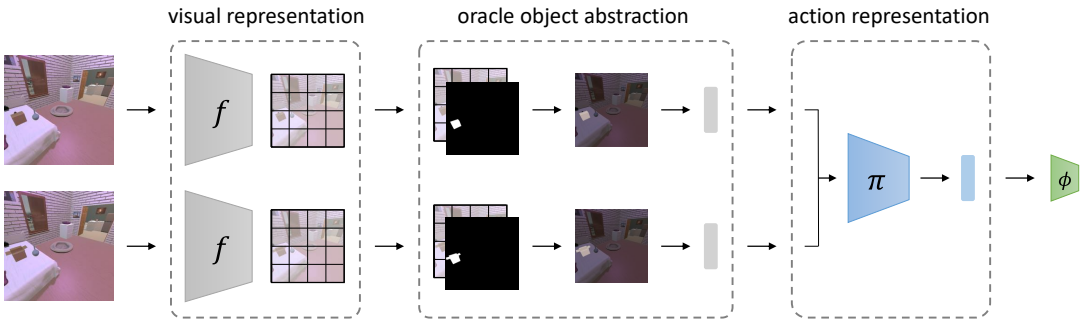
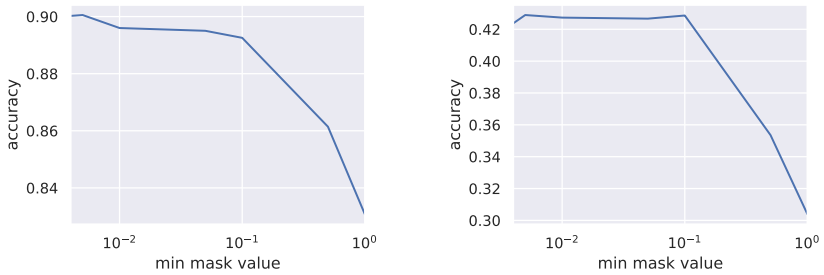


Figure 9: Overview of the intervention model built with the oracle object-centric representation.



(*a*) ID (*b*) OOD

Figure 10: Effect of the oracle object-centric representation in multi-object scenes. Models with clearer object-level abstractions (lower background mask value) perform better on both the ID and OOD sets that differ in object classes.

lies in visual representations of unseen objects, where disentanglement encouraged by the sparsity regularizer remains highly beneficial.

4.2. Multi-object Scenes

4.2.1. SIMULATED MULTI-OBJECT SCENES

Setup. In contrast to the previous sections where each image contains only one object, we next consider more complex scenes composed of multiple objects. Following the data split above, we train and test the model on two disjoint groups of object classes. To represent high-level actions in multi-object scenes, we consider an extra module between the image encoder $f(\cdot)$ and action encoder $\pi(\cdot)$ for object-level abstraction. Tab. 3 summarizes the results of different design choices.

Distributed representation. Consistent with the observations in §4.1.2, the intervention models trained on a subset of object classes generally suffer from significant performance drops on the unseen OOD objects. Notably, while the architecture design of the distributed representation baseline is the same as that in the previous sections, the performance in the multi-object scenes (Tab. 3) is significantly worse than the counterparts in the single-object images (Tab. 2). We postulate that this is largely attributed to the lack of object-level abstractions prior to reasoning about the effect of an intervention. To verify this, we next take a close look at intervention classifiers built with two variants of structured representations that support such high-level abstractions.

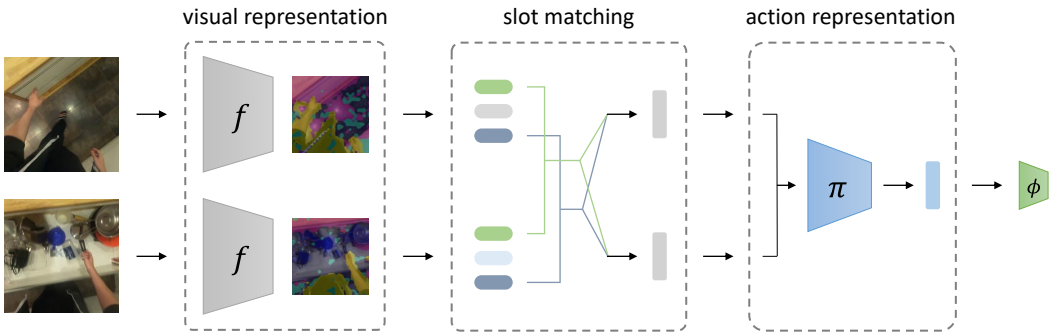


Figure 11: Overview of the intervention model built with object-centric visual representations.

Oracle object-centric representation. A natural way to exploit sparse mechanism shifts at the object-level is to explicitly decompose each scene into a set of constituent objects and selectively attend to the objects that undergo state changes. Nevertheless, learning object-centric representations of complex scenes is often beyond reach in the first place (Singh et al., 2022; Seitzer et al., 2022).

To approximately estimate the potential benefits of this modeling hypothesis, we resort to an *oracle* object-centric representation by making use of the ground-truth instance segmentation obtained from the simulation engine. As shown in Fig. 9, each pixel in the segmentation map is a binary value, with 1 representing the intervened object and 0 representing the rest background. In order to convert it to an object-level attention mask, we first resize the segmentation map to the dimensionality of the feature map, and then perform element-wise multiplication between the resized segmentation map and each channel of the feature map. Moreover, in place of the default background value, we traverse the minimum mask value from 1 to 10^{-3}, which allows us to emulate object-level decomposition and attention in the representation space with varying quality.

Tab. 3 and Fig. 10 show the results of the intervention classifiers built with the oracle object-centric representations of different qualities. Compared with the distributed representation baseline (the minimum mask value equals to 1), incorporating the oracle object-level abstraction boosts performance on the ID and OOD test sets by up to $\sim 8.5\%$ and $\sim 40\%$, respectively. Overall, reducing the background mask value consistently increases classification accuracy, demonstrating a clear advantage of capturing the sparse change at the object level for modeling intervention. Meanwhile, we also notice a mild performance degradation when the min mask value approaches 0. We conjecture this is due to numerical instability in the case where the intervened object is very small in the scene.

Learned object-centric representation. We further investigate intervention models built with learned object-centric representations. Specifically, we first pre-train on our dataset the implicit Slot Attention (Locatello et al., 2020b; Chang et al., 2022), a state-of-the-art unsupervised object-centric learning method, which decomposes an input scene into a number of ($N = 15$) spatially and semantically related regions, each with its own feature vector. To exploit the latent structure for downstream reasoning, we further consider three different slot matching schemes:

- *slot-avg*: average-pooling over slots for each image, which degenerates to a distributed representation (only one pair of feature vectors);
- *slot-dense*: densely pair slots across the two images, pass all combinations to the action encoder, and aggregate $N \times N$ relation embeddings with average-pooling;

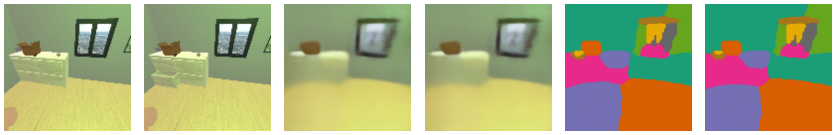


Figure 12: Results of the implicit Slot Attention (Locatello et al., 2020b; Chang et al., 2022) on our simulated multi-object scenes. From left to right: pair of input images, pair of reconstructed images, pair of segmentation masks. More examples are provided in Fig. 15. Overall, the quality of reconstruction and segmentation remains limited.

	ID	OOD
resnet	0.83 ± 0.01	0.30 ± 0.08
oracle-mask	$\mathbf{0.90} \pm 0.01$	$\mathbf{0.42} \pm 0.06$
slot-avg	0.49 ± 0.01	0.15 ± 0.01
slot-dense	0.51 ± 0.01	0.19 ± 0.03
slot-match	0.66 ± 0.01	0.21 ± 0.01

Table 3: Intervention classification accuracy in simulated multi-object scenes. The pre-trained slot-based encoder is frozen in experiments.

	ID	OOD
resnet	0.42 ± 0.03	0.17 ± 0.03
clip	0.45 ± 0.02	0.24 ± 0.02
group-avg	0.47 ± 0.03	0.24 ± 0.03
group-dense	0.50 ± 0.04	0.26 ± 0.03
group-token	$\mathbf{0.52} \pm 0.03$	$\mathbf{0.27} \pm 0.03$

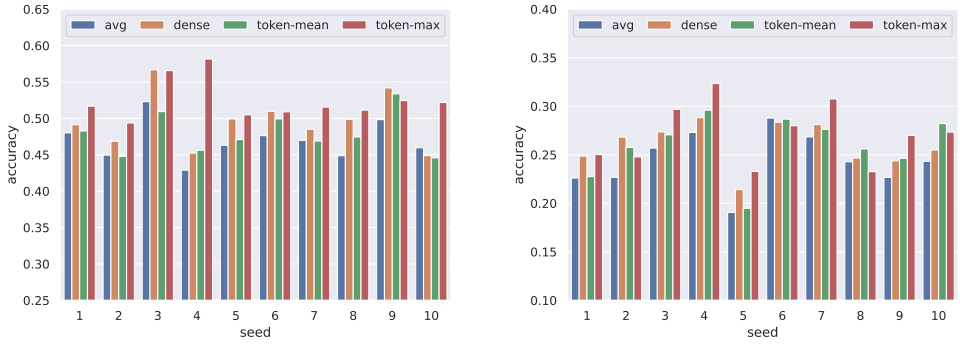
Table 4: Intervention classification accuracy in real-world multi-object scenes. All encoders are pre-trained, and kept frozen in experiments.

- *slot-match*: selectively pair slots across the two images based on the slot similarity, only pass the matched pairs to the action encoder, and aggregate N relation embeddings with max-pooling, as shown in Fig. 11.

Tab. 3 shows the results of each design choice on the ID and OOD test sets. Corroborating with the results from the oracle object-centric approximation, exploiting the latent slot structure leads to significantly higher accuracy for downstream reasoning. In particular, the intervention model built with the slot matching scheme is over 34% more accurate than the vanilla counterpart built with the global average-pooling operation. Nevertheless, the overall performance from the frozen encoder remains limited, as evidenced by the inferior results to the fine-tuned ResNet model as well as the blurry image reconstructure shown in Fig. 12, indicating a large room for improvement in object-centric learning on our benchmark.

4.2.2. REAL-WORLD MULTI-OBJECT SCENES

Setup. We finally migrate to intervention modeling on real-world observations repurposed from the Epic-Kitchens dataset (Damen et al., 2022). Reasoning about abstract transformations between real-world image pairs is generally more challenging than in the controlled simulation environment, due to significant global changes in camera locations and perspectives, frequent local occlusions by arms and hands, as well as limited training data. Given these challenges, we turn our attention to GroupViT (Xu et al., 2022), a set-structured visual representation pre-trained on massive amounts of internet data. Similar to §4.2.1, we assess the impact of visual representations by comparing three pre-trained encoders: (i) ResNet-18 pre-trained on ImageNet (Deng et al., 2009), (ii) CLIP, and (iii) GroupViT pre-trained on internet data. Furthermore, to leverage the latent structure in GroupViT,



(*a*) ID (*b*) OOD

Figure 13: Results of different group matching modules in real-world data on 10 different seeds (random splits). Compared with the *group-avg* baseline, exploiting the learned group structure with *token-max* leads to improved accuracy on both the ID and OOD sets that differ in object classes.

we match the slots between two images based on the group tokens, and aggregate the resulting action embeddings with two choices: average-pooling (denoted by *token-mean*) and max-pooling (denoted by *token-max*).

Results. Tab. 4 shows the results of different models on the ID and OOD test sets. While ResNet, CLIP, and *group-avg* all fall into the class of distributed vector representations, the intervention models built with the latter two generalize much better than the first one to unseen objects, suggesting the strength of large-scale pre-training for OOD generalization. Among all considered models, the token-based group matching appears to be most effective for intervention modeling, outperforming the ResNet baseline on the ID and OOD test sets by $\sim 24\%$ and $\sim 59\%$, respectively. In particular, we observe in Fig. 13 that *token-max* results in better accuracy than *token-mean* on most seeds. Intuitively, when semantic regions are properly matched between two images, attending to the group of the most significant changes using max-pooling can be a good inductive bias to reason about the action class.

5. Conclusion and Discussion

In this paper, we introduced `Causal Triplet`, a causal representation learning benchmark with high visual complexities, actionable counterfactuals, and an interventional downstream task. We revisited the principles of independent causal mechanisms and sparse mechanism shifts in the context of intervention modeling, and empirically estimated their strengths for out-of-distribution robustness by making use of oracle knowledge of the underlying structures. However, we also observed significant challenges for recent methods to discover these structures, indicating fruitful opportunities for future research.

As the first benchmark of its kind, `Causal Triplet` is still subject to two major constraints: the object states in our simulated dataset are binary variables, as opposed to continuous ones in the real world; and the scale of our repurposed real-world dataset is limited, sufficient for transfer learning but not for representation learning from scratch. Nevertheless, we hope our benchmark will call attention to the assumptions made in causal representation learning, serve as a public testbed in challenging yet practical settings, and propel progress towards real-world problems.

Acknowledgments

This work is supported in part by the Swiss National Science Foundation under the Grant 2OOO21-L92326. We thank Maximilian Seitzer, Carl-Johann Simon-Gabriel, Andrii Zadaianchuk, Yuchen Zhu, Harvineet Singh and Milton Montero for helpful discussions. We thank Parth Kothari, Riccardo Cadei, Linyan Yang and Yifan Sun for thoughtful feedback on early drafts, as well as anonymous reviewers for insightful comments.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. In *International Conference on Learning Representations*, 2021. 3, 5

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Weakly Supervised Representation Learning with Sparse Perturbations, June 2022a. 3

Kartik Ahuja, Yixin Wang, Divyat Mahajan, and Yoshua Bengio. Interventional Causal Representation Learning, October 2022b. 3

Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *Advances in Neural Information Processing Systems*, volume 33, pages 1462–1473, 2020. 3, 8

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 4

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised Scene Decomposition and Representation, January 2019. 3

Michael Chang, Thomas L. Griffiths, and Sergey Levine. Object Representations as Fixed Points: Training Iterative Refinement Algorithms with Implicit Differentiation. In *Advances in Neural Information Processing Systems*, October 2022. 11, 12, 20

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180, 2016. 2

Taco Cohen. Towards a Grounded Theory of Causation for Embodied AI, June 2022. 2, 4

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. *International Journal of Computer Vision*, 130:33–55, January 2022. 5, 12

Matt Deitke, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X. Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez D'Arpino, Kiana Ehsani, Ali Farhadi, Li Fei-Fei, Anthony Francis, Chuang Gan, Kristen Grauman, David Hall, Winson Han, Unnat Jain, Aniruddha Kembhavi, Jacob Krantz, Stefan Lee, Chengshu Li, Sagnik Majumder, Oleksandr Maksymets, Roberto Martín-Martín, Roozbeh Mottaghi, Sonia Raychaudhuri, Mike Roberts, Silvio Savarese, Manolis Savva, Mohit Shridhar, Niko Sünderhauf, Andrew Szot, Ben Talbot, Joshua B. Tenenbaum, Jesse Thomason, Alexander Toshev, Joanne Truong, Luca Weihs, and Jiajun Wu. Retrospectives on the Embodied AI Workshop, October 2022a. 2

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *Advances in Neural Information Processing Systems*, June 2022b. 5, 19

J. Deng, W. Dong, R. Socher, L. Li, and and. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 12

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wuthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the Transfer of Disentangled Representations in Realistic Settings. In *International Conference on Learning Representations*, September 2020. 6

Andrea Dittadi, Samuele S. Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and Robustness Implications in Object-Centric Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5221–5285, June 2022. 6

Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding egocentric activities. In *2011 International Conference on Computer Vision*, pages 407–414, November 2011. 3

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent Independent Mechanisms. In *International Conference on Learning Representations*, 2021. 2

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural Expectation Maximization. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 3

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-Object Representation Learning with Iterative Variational Inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2424–2433, May 2019. 3

David Ha and Jürgen Schmidhuber. World Models. *arXiv:1803.10122 [cs, stat]*, March 2018. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 7

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 2017. 2

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81:945–960, 1986. 4

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding. In *International Conference on Learning Representations*, 2021. 2, 3

Sebastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E. Everett, Rémi LE Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 428–484, June 2022. 1, 3, 5

Anson Lei, Bernhard Schölkopf, and Ingmar Posner. Variational Causal Dynamics: Discovering Modular World Models from Interventions, June 2022. 3

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. iCITRIS: Causal Representation Learning for Instantaneous Temporal Effects. In *UAI 2022 Workshop on Causal Representation Learning*, July 2022a. 3, 5

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Intervention Design for Causal Representation Learning. In *UAI 2022 Workshop on Causal Representation Learning*, July 2022b. 3, 8

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In *Proceedings of the 39th International Conference on Machine Learning*, pages 13557–13603, June 2022c. 1, 3, 5, 8

Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards Robust and Adaptive Motion Forecasting: A Causal Representation Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17081–17092, 2022. 6

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124, May 2019. 2

Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-Supervised Disentanglement Without Compromises. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6348–6359, November 2020a. 1, 2, 3, 4, 5

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with

Slot Attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538, 2020b. 3, 7, 11, 12, 20

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of Disentanglement in Generalisation. In *International Conference on Learning Representations*, 2021. 6

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning Independent Causal Mechanisms. In *International Conference on Machine Learning*, pages 4036–4044, July 2018. 1

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA, November 2017. ISBN 978-0-262-03731-0. 5

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, July 2021. 7

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019. 2

Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent Prototype Propagation for Zero-Shot Compositionality. In *Advances in Neural Information Processing Systems*, volume 34, pages 10641–10653, 2021. 8

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 459–466, Madison, WI, USA, June 2012. ISBN 978-1-4503-1285-1. 1, 5

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards Causal Representation Learning, February 2021. 1, 5

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the Gap to Real-World Object-Centric Learning, September 2022. 11

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple Unsupervised Object-Centric Learning for Complex and Naturalistic Videos. In *Advances in Neural Information Processing Systems*, October 2022. 11

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On Disentangled Representations Learned from Correlated Data. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10401–10412, July 2021. 2

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are Disentangled Representations Helpful for Abstract Visual Reasoning? In *Advances in Neural Information Processing Systems*, volume 32, 2019. 6

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 *Advances in Neural Information Processing Systems*, volume 34, pages 16451–16467, 2021. 1, 2, 3, 4, 5

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic Segmentation Emerges From Text Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 7, 12

Yanchao Yang, Yutong Chen, and Stefano Soatto. Learning to Manipulate Individual Objects in an Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2020. 3

Yanchao Yang, Brian Lai, and Stefano Soatto. DyStaB: Unsupervised Object Segmentation via Dynamic-Static Bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2826–2836, 2021. 3

Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. ACRE: Abstract Causal REasoning Beyond Covariation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10643–10653, 2021. 6

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive Learning Inverts the Data Generating Process. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12979–12990, July 2021. 3

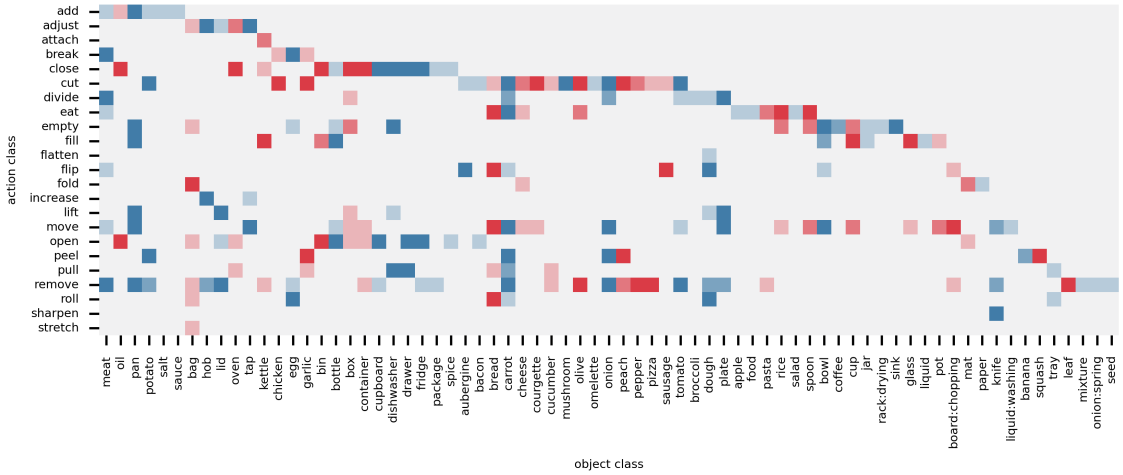


Figure 14: Data split in `Causal Triplet` collected from real-world observations.

Appendix A. Benchmark Details

Dataset Collection. Our simulated data is collected from ProcTHOR, the largest simulation environment for embodied agents. In order to repurpose it to our proposed actionable counterfactual setting, we shortlist the objects that can be independently manipulated, and skip the objects that are physically coupled with others, *e.g.*, tables below cups, bowls, etc. To enlarge the diversity of the collected examples, we randomly sample the position of the embodied agent in the environment and save no more than 20 examples from each distinct room.

Our real-world data is sourced from the Epic-Kitchens-100 dataset, one of the largest ego-centric video datasets known for its strong human-object interactions. In order to ensure the image quality, we removed the examples in which the intervened object is not visually clear. More specifically, we utilized an off-the-shelf object detector to estimate the image quality, filtering out those with a detection confidence score lower than 0.4. Additionally, we remove the action classes that do not result in noticeable visual variations before and after the action, such as "feel" and "spray." Fig. 14 shows the data split used in our experiments in §4.2.2.

Dataset Properties. As summarized in Tab. 1, our gathered datasets possess several crucial properties that the preceding datasets do not. For instance, the variability of local object-level variables, *e.g.*, shape and scale, is highly limited in the Causal3DIdent and CausalWorld datasets. In contrast, the diversity of object shape and scale within each object class is substantial in both our simulated and real-world datasets. Similarly, scene-level variables like the illumination intensity are often fixed in the previous ones, but vary widely in ours, *e.g.*, we randomized the illumination elements (artificial lights and skyboxes) in ProcTHOR (Deitke et al., 2022b), rendering simulated scenes at different times of the day.

Experiment Details. Our implementations are built upon the public libraries of the baseline models. We applied their default settings for image processing and fine-tuning the representations on our datasets. Specifically, we resized images to 128 x 128 for Slot Attention, and 224 x 224 for the other baselines. For pre-trained foundation models (CLIP and GroupViT), we used their official

config	value
batch size with frozen encoder	128
batch size with fine-tuned encoder	32
learning rate	0.001
training epochs in simulated data	50
training epochs in real-world data	200

Table 5: Default hyper-parameters in `Causal Triplet` experiments.

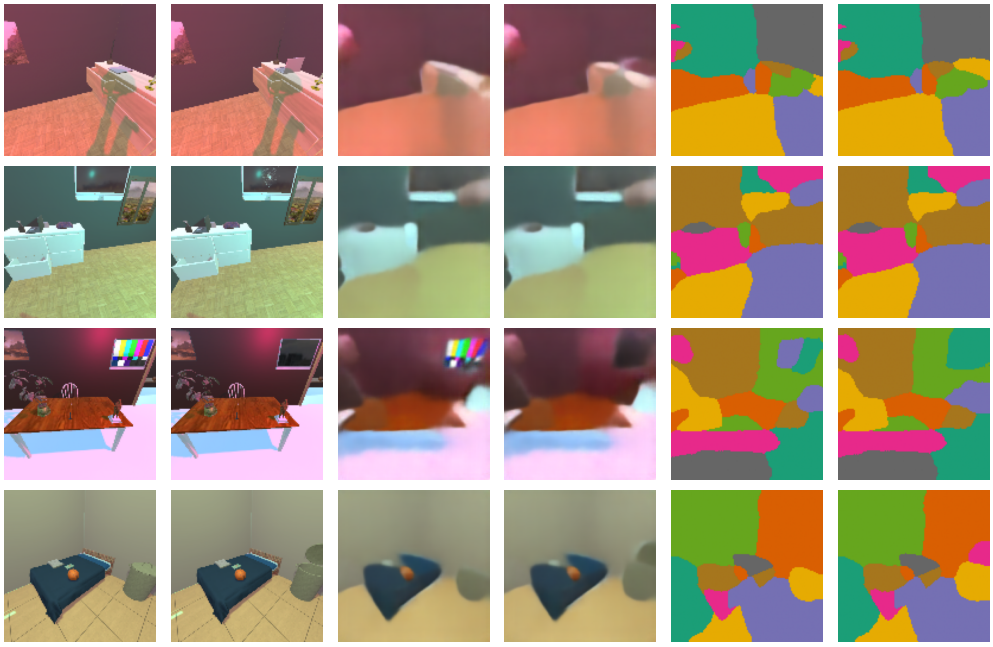


Figure 15: Additional results of the implicit Slot Attention (Locatello et al., 2020b; Chang et al., 2022) on our simulated multi-object scenes. The number of slots is set to $N = 15$ based on the maximum number of individual objects. From left to right: pair of input images, pair of reconstructed images, pair of segmentation masks.

checkpoints. In our experiments, each random seed leads to a unique data split. Common hyperparameters specific to our benchmark are summarized in Tab. 5. In addition to the quantitative results in Tab. 3, visualizations of simulated image pairs, along with their corresponding reconstructions and segmentation masks from implicit Slot Attention, are provided in Fig. 12.

Appendix B. Additional Discussions

Downstream task. The downstream task considered in our benchmark is designed to reason about high-level actions rather than low-level actions for two reasons: (i) from causal standpoint, high-level actions correspond to interventions in the world and induce clear changes on the underlying causal variables, which low-level actions often do not, (ii) from computer vision standpoint, despite the long history of action recognition, the task remains less saturated than object classification. We

add to this challenge with an egocentric camera view, which while corresponds to real-world data capture, hides much of the actor and requires direct reasoning about the effects of actions on objects.

Confounder examples. Unobserved confounders $\mathbf{c}$ between latent variables $\mathbf{z}$ and action classes $\mathbf{a}$ can be ubiquitous in practical problems. One common source is the set of constraints under which an agent can interact with the surrounding objects. For instance, whether or not remote control of electronic devices is feasible jointly influences the size/distance of the intervened object (*e.g.*, TV, light) and the action class (*e.g.*, turnon, turnoff).

Additional limitations. Aside from the major limitations discussed in §5, our benchmark is subject to a few other practical shortcomings, including

- the numbers of object and action classes are limited in our simulated dataset;
- ground truth annotations of causal factors are generally not available for complex scenes;
- actions of the same semantic class sometimes induce varying effects, *e.g.*, the effect of 'open' on a fridge/cupboard may look different from that on a book/laptop.

While our benchmark is significantly closer to the real-world problems than previous ones of its kind, these issues need to be addressed in future research.