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

Author names withheld

Editor: Under Review for CLeaR 2023

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 `CausalTriplet`, 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 in `CausalTriplet`. Our code and datasets will be available at https://sites.google.com/view/causaltriplet.

Keywords: Causal Representation Learning, Out-of-Distribution Generalization

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 along this line 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, we aim to bridge this gap by revisiting recent hypotheses and methods in more realistic settings. Beside the 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 all existing methods (Locatello et al., 2020a; von Kügelgen

© 2023 .

et al., 2021; Klindt et al., 2022; Lachapelle et al., 2022; 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 $X \rightarrow Y$, 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 `CausalTriplet`. 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 on `CausalTriplet`. 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), 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-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 mecha-

	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.

nism shifts between paired observations to identify the underlying causal variables (Locatello et al., 2020a; Zimmermann et al., 2021; von Kügelgen et al., 2021; Lippe et al., 2022a,b,c; Ahuja et al., 2022a; Brehmer et al., 2022). 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 `CausalTriplet`, 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 in causal representation learning (Locatello et al., 2020a; von Kügelgen et al., 2021; Lachapelle et al., 2022; Brehmer et al., 2022). Yet, how to represent such interventions remains largely unexplored. In fact, learning a robust representation 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 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 obejct-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 action categories from image pairs, *i.e.*, modeling the conditional probability $\mathbb{P}(\mathbf{a} \mid \mathbf{x}, \tilde{\mathbf{x}})$.

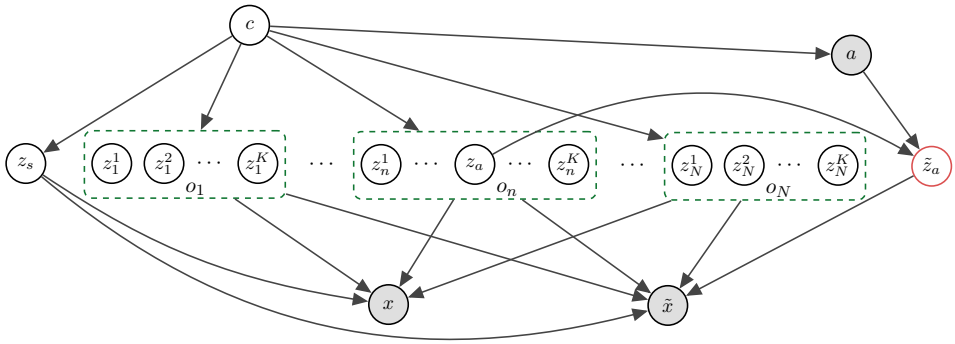


Figure 1: Causal graph for a pair of scene observations $(x, \tilde{x})$ before and after an action a. The data generating process of each raw observation is described by a set of latent factors, including global scene-level factors z_s and local object-level factors z_n^k, which are statistically dependent 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.

Crucially, we assume that the effect of an action $\mathbb{P}(\tilde{\mathbf{z}}_a \mid \mathbf{z}_a, \mathbf{a})$ stay invariant between training and test data, whereas other mechanisms may change. Such distribution shifts $\mathbb{P}_{\text{tr}}(\mathbf{a}, \mathbf{z}) \neq \mathbb{P}_{\text{te}}(\mathbf{a}, \mathbf{z})$ 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. Building a robust intervention model 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 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 manip-

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

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

ulated 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.

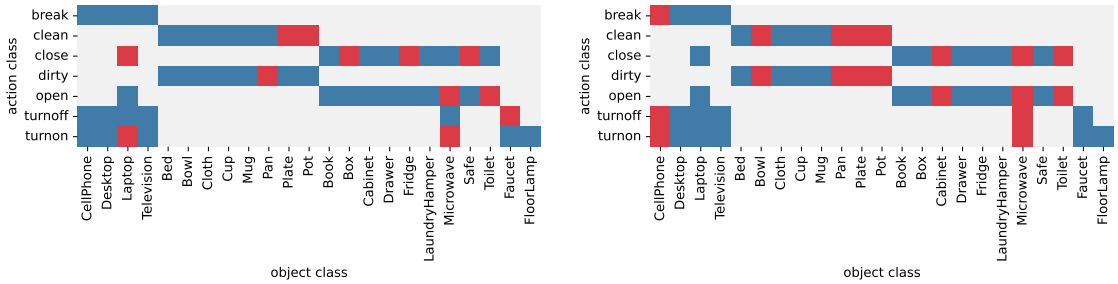
3.2. Benchmark Data

Existing datasets for causal representation learning (von Kügelgen et al., 2021; Ahmed et al., 2021; Lippe et al., 2022a,c), as listed in Table 1, are heavily simplified in many aspects when 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).

Unlike the standard i.i.d. setting, 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 datraset details are summarized in Appendix A



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

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

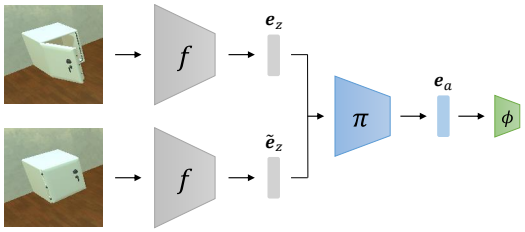


Figure 5: Overview of our benchmark model for reasoning about the action class from a pair of single-object images. We consider different regularizers for learning robust image representation $\mathbf{e}_z$ and action representation $\mathbf{e}_a$.

	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

Figure 6: 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.

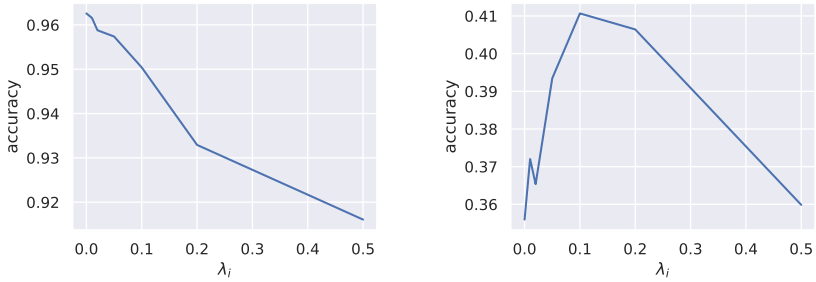
4. Experiments and Results

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

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

To this end, we consider a variety of visual representations, including modern distributed representations (*e.g.*, ResNet, CLIP), oracle structured representations (*e.g.*, ground-truth knowledge of disentanglement / object-centric) and learned structured representations (*e.g.*, GroupViT). 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), and 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.



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

Figure 7: 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.

4.1. Simulated Single-object Images

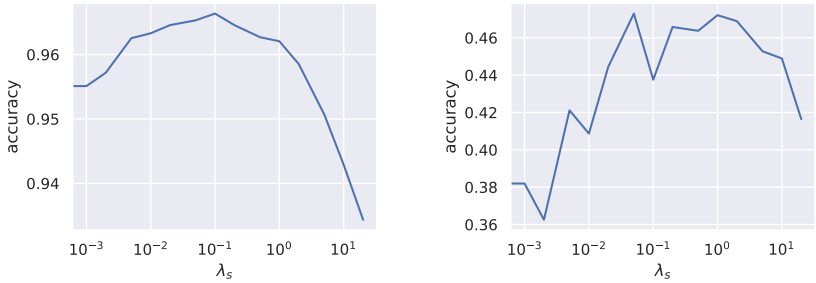
4.1.1. Compositional Distribution Shifts

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 Fig. 6, 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 8: 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.

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. 7 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. 8 and Fig. 13 shows that the effect of the sparsity regularizer on test accuracy and visual representations, respectively. Compared with the vanilla baseline, the sparsity regularizer with a moderate coefficient ($\lambda_s \approx 0.1$) demonstrates a prominent strength on both the ID and OOD test sets, improving the OOD robustness by nearly 30%. Nevertheless, same as Lippe et al. (2022b,c), the sparsity regularizer requires intervention labels at the variable-level, *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, dirty vs. clean, turnon vs. turnoff), collecting such detailed annotations in real-world problems can be difficult.

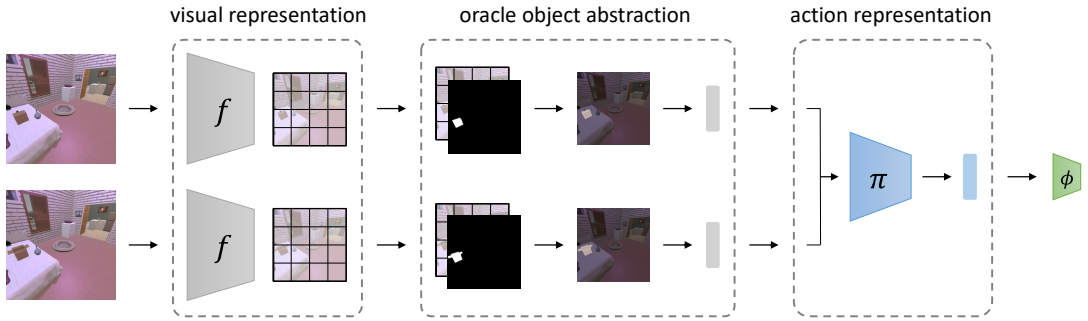
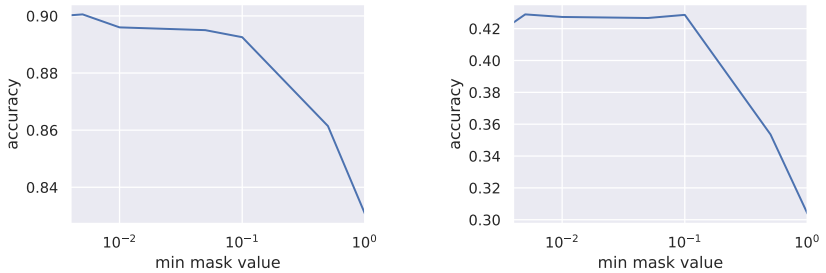


Figure 9: Overview of the intervention model built with the oracle object-centric representation.



(*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.

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. Fig. 6 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 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

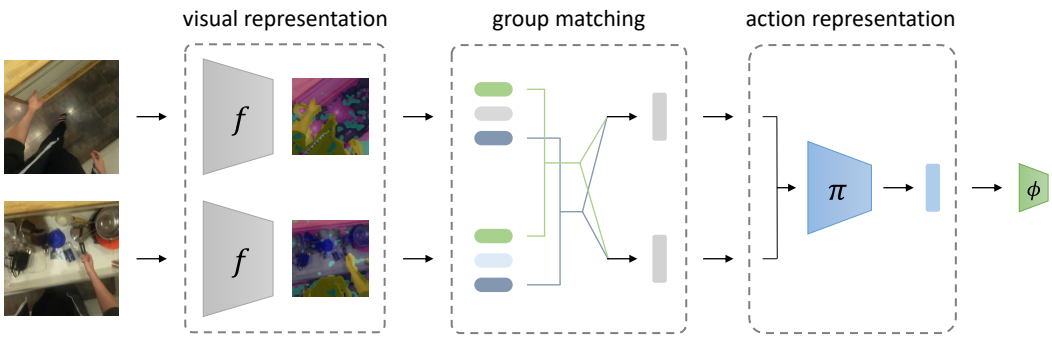


Figure 11: Overview of the intervention model built with the group-structured visual representation.

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. 2 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. 2) is significantly worse than the counterparts in the single-object images (Fig. 6). 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.

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. 2 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.

	ID	OOD
resnet	0.83 ± 0.01	0.30 ± 0.08
oracle-mask	$\mathbf{0.90} \pm 0.01$	$\mathbf{0.42} \pm 0.06$
group-avg	0.77 ± 0.01	0.36 ± 0.07
group-dense	0.79 ± 0.01	0.38 ± 0.05
group-token	0.81 ± 0.01	0.39 ± 0.05

Table 2: Intervention classification accuracy in simulated multi-object scenes. The pre-trained GroupViT 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 3: Intervention classification accuracy in real-world multi-object scenes. All encoders are pre-trained, and kept frozen in experiments.

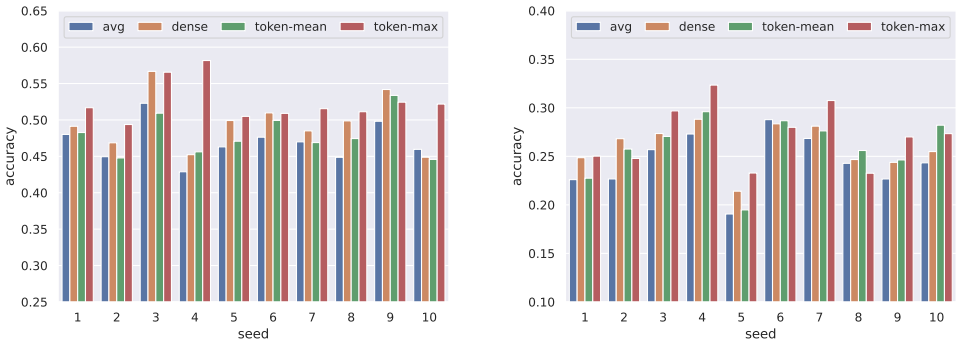
Learned object-centric representation. We further investigate intervention models built with learned object-centric representations. Due to the significant challenge of learning such representations from scratch, we turn our attention to pre-trained models. Specifically, we use GroupViT (Xu et al., 2022), a set-structured visual representation pre-trained on massive amounts of internet data, as our image encoder. As shown in Fig. 11, GroupViT decomposes an input scene into a number of ($N = 8$) semantically-related regions, each with its own feature vector. We consider three different group matching schemes before passing them to the downstream action encoder,

- *group-avg*: average-pooling over the feature vectors within each image, which degenerates to a distributed representation (only one pair of feature vectors);
- *group-dense*: densely pair the feature vectors across the two images, pass all combinations to the action encoder, and aggregate $N \times N$ relation embeddings with average-pooling;
- *group-token*: selectively pair the feature vectors across the two images based on group tokens, only pass the matched pairs to the action encoder, and aggregate N relation embeddings with average-pooling.

Tab. 2 shows the results of each design choice on the ID and OOD test sets. While the GroupViT encoder is frozen, the learned intervention classifier reaches an accuracy comparable to the ResNet counterpart. In fact, the GroupViT-based models even outperform the ResNet baseline on unseen OOD objects, suggesting the potential of large-scale pre-training for OOD generalization. Among the GroupViT-based variants, token-based matching yields the best result. Nevertheless, the improvement from *group-avg* to *group-token* is less than 4%, which is much smaller than that from the ResNet baseline to the oracle object-centric representation, indicating that there is still large room for improvement in object-centric learning on our proposed benchmark.

4.2.2. REAL-WORLD MULTI-OBJECT SCENES

Setup. We finally migrate to intervention modeling in real-world observations. Reasoning about abstract transformations between image pairs in Epic-Kitchens (Damen et al., 2022) is generally more challenging than in the controlled simulation environment in the previous sections, due to significant global changes in camera locations and perspectives, as well as frequent local occlusions by arms and hands. 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, for GroupViT-based models incorporating



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

Figure 12: 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.

token-based group matching, we consider two operations for embedding aggregations: average-pooling (denoted by *token-mean*) and max-pooling (denoted by *token-max*).

Results. Tab. 3 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. Corroborating with Tab. 2, this result confirms 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. 12 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 `CausalTriplet`, 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, `CausalTriplet` 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 `CausalTriplet` 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.

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. 2

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. 2, 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. 2

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. 1, 4

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 18

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, 11

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. 1

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

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. 11

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

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural Expectation Maximization. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 2

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. 2

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

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*, February 2022. 2

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, 2, 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

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, 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. 2

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

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, 4

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. 10

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*, May 2022. 10

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

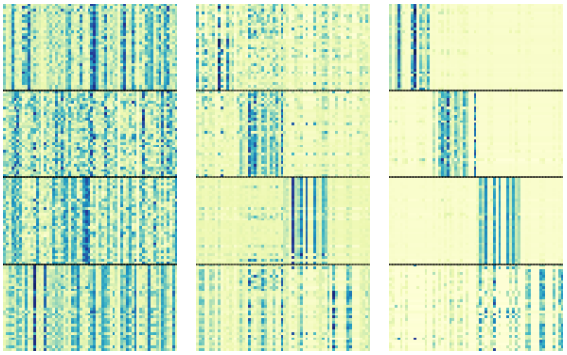
Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *Advances in Neural Information Processing Systems*, volume 34, pages 16451–16467, 2021. 1, 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. 11

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. 2

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. 2

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



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

Figure 13: 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.

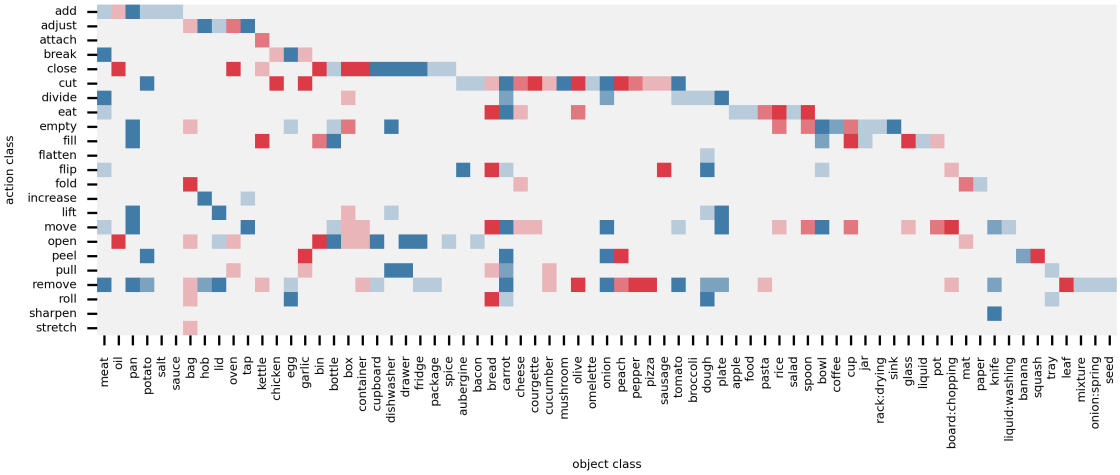


Figure 14: Data split in `CausalTriplet` collected from real-world observations.

Appendix A. Benchmark Details

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 repurposed from the Epic-Kitchens-100 dataset, one of the largest egocentric video datasets known for strong human-object interactions. Due to frequent motion blurs in the video dataset, we revmoe the examples in which the intervened object is not visually clear. More

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

Table 4: Hyper-parameters in style shift experiments.

specifically, we use the confidence score from an off-the-shelf object detector (Damen et al., 2018) to estimate the image quality and filter out the ones with a detection confidence score lower than 0.4. Furthermore, we delete 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.