

---

# Smoke and Mirrors in Causal Downstream Tasks

---

Riccardo Cadei<sup>1</sup>, Lukas Lindorfer<sup>1</sup>, Sylvia Cremer<sup>1</sup>, Cordelia Schmid<sup>2</sup>, and Francesco Locatello<sup>1</sup>

<sup>1</sup>Institute of Science and Technology Austria

<sup>2</sup>Inria, Ecole normale supérieure, CNRS, PSL Research University

## Abstract

Machine Learning and AI have the potential to transform data-driven scientific discovery, enabling accurate predictions for several scientific phenomena. As many scientific questions are inherently causal, this paper looks at the causal inference task of *treatment effect estimation*, where the outcome of interest is recorded in high-dimensional observations in a Randomized Controlled Trial (RCT). Despite being the simplest possible causal setting and a perfect fit for deep learning, we theoretically find that many common choices in the literature may lead to biased estimates. To test the practical impact of these considerations, we recorded ISTAnt, the first real-world benchmark for causal inference downstream tasks on high-dimensional observations as an RCT studying how garden ants (*Lasius neglectus*) respond to microparticles applied onto their colony members by hygienic grooming. Comparing 6 480 models fine-tuned from state-of-the-art visual backbones, we find that the sampling and modeling choices significantly affect the accuracy of the causal estimate, and that classification accuracy is not a proxy thereof. We further validated the analysis, repeating it on a synthetically generated visual data set controlling the causal model. Our results suggest that future benchmarks should carefully consider real downstream scientific questions, especially causal ones. Further, we highlight guidelines for representation learning methods to help answer causal questions in the sciences.

Code: <https://github.com/CausalLearningAI/ISTAnt>

Data: <https://doi.org/10.6084/m9.figshare.26484934.v2>

## 1 Introduction

Uncovering the answer to many scientific questions requires analyzing massive amounts of data that humans simply cannot process on their own. For this reason, leveraging machine learning and AI to help answer scientific questions is one of the most promising frontiers for AI research. As a result, AI is now predicting how proteins fold [Jumper et al., 2021], new materials [Merchant et al., 2023], precipitation forecasts [Espeholt et al., 2022], and animal behaviors [Sun et al., 2023]. Even predicting counterfactual outcomes for treatment effect estimation appears to be possible [Feuerriegel et al., 2024]. In scientific applications, these predictions are often incorporated into broader analyses to draw new physical insights. In this paper, we focus on the problem of estimating the strength of the causal effect of some variable on another, which is a common type of question across disciplines [Robins et al., 2000, Samet et al., 2000, Van Nes et al., 2015, Runge, 2023].

While our discussion and conclusions are general, we follow a simple real-world example throughout the paper: behavioral ecologists want to study the social hygienic behavior in ants and, thereby, the ability of the insects to remove small particles from the body surface of exposed colony members. Such grooming behavior performed by nestmates plays an important role in restoring a clean body surface of the contaminated individual, which, in case of infectious particles being groomed off, assures the health of the individual and prevents disease spread through the colony’s [Rosengaus et al., 1998, Hughes et al., 2002, Konrad et al., 2012]. To study whether different microparticles differ systematically in their induction of grooming behavior, the biologists thus perform an experiment under

controlled conditions, where a focal worker ant is treated randomly with either of two microparticle types, and the behavior of two untreated colony members towards the treated ant is filmed in multiple replicates. This is followed by detailed behavioral observation to quantify ant activity, as well as statistical data analysis to determine if treatment has a significant effect. This step could obviously be entirely replaced with deep learning, dramatically accelerating the workflow. In fact, many data sets and benchmarks have been proposed with the specific reason of supporting downstream science in behavioral ecology and biology [Sun et al., 2023, Beery et al., 2018, Kay et al., 2022, Chen et al., 2023] and other scientific disciplines [Beery et al., 2022, Lin et al., 2023, Moen et al., 2019].

Our paper questions the simplicity of this narrative in both theory and practice. While we take experimental behavioral ecology as an example for our motivation and experiments, our theoretical results and experimental conclusions are general, and we expect them to be applicable across disciplines. Our key contributions can be summarized as follows:

- We theoretically show how many design choices can affect the answer to a causal question, from the data used for training, the architecture choices, and even seemingly innocuous standard practices like thresholding the predictions into hard labels, or using held out accuracy for model selection (a common practice in many AI for science benchmarks, e.g., [Sun et al., 2023]). To facilitate future research on representation learning for causal downstream tasks, we *formulate the representation desiderata to obtain accurate estimates for downstream causal queries* together with best practices.
- To showcase the practical impact of these design choices, we design and collect a new dataset, *ISTAnt*, from a real randomized controlled trial, reflecting a real-world pipeline in experimental behavioral ecology, which we will release to accelerate research on representation learning for causal downstream tasks. To the best of our knowledge, this is the *first real-world data set specifically designed for causal inference downstream tasks from high-dimensional observations*.
- On our dataset, we *fine-tune 6 480 state-of-the-art methods* [Dosovitskiy et al., 2020, Zhai et al., 2023, Radford et al., 2021, He et al., 2022, Oquab et al., 2023] in the few- and many-shot settings. Empirically, we confirm that the seemingly innocuous design choices like which samples to annotate, which model to use, whether or not to threshold the labels, and how to do model selection have a major impact on the accuracy of the causal estimate. Since our ground-truth estimate of the causal effect depends on the trial’s design, we propose a *new synthetic benchmark* based on MNIST [LeCun, 1998] controlling for the causal model, and we replicated the analysis.

## 2 Setting

We consider the RCT setting, where binary treatments  $T$  are randomly assigned within an experiment with controlled settings  $\mathbf{W}$ . In many applications, the outcome of interest  $Y$  is not measured directly. Instead, it is collected in high-dimensional observations  $\mathbf{X}$ — e.g., frames from a video of the experiment. Our goal is to estimate the causal effect of  $T$  on  $Y$ , which is quantified by the estimation of the Average Treatment Effect (ATE), i.e.:

$$ATE := \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)]. \tag{1}$$

Assuming an RCT (i.e., Ignorability Assumption [Rubin, 1978]) is the ideal setting for causal inference because the ATE directly identifies in the Associational Difference (AD), i.e.,

$$AD := \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]. \tag{2}$$

However, annotating  $Y$  from the high-dimensional recordings  $\mathbf{X}$  requires costly manual annotations from domain experts. Leveraging state-of-the-art deep learning models, we can hope to alleviate this effort. Instead of labeling all the data, we only partially annotate it. We introduce a binary variable  $S$ , indexing whether a frame is annotated by a human observer or not. We denote the annotated samples with  $\mathcal{D}_s = \{(\mathbf{W}_i, T_i, \mathbf{X}_i, Y_i) : S_i = 1\}_{i=1}^{n_s}$  and the not annotated ones with  $\mathcal{D}_u = \{(\mathbf{W}_i, T_i, \mathbf{X}_i) : S_i = 0\}_{i=1}^{n_u}$ . We use  $\mathcal{D}_s$  to train or fine-tune a deep learning model to estimate the labels on  $\mathcal{D}_u$ . Next, we leverage the Ignorability Assumption on the full RCT to identify the ATE in the AD and consistently estimate it. Ideally, it would be most useful if  $\mathcal{D}_s = \emptyset$ , but for the purpose of this paper, we assume that at least some samples are annotated, for example, during quality controls.

Besides the clear statistical power considerations, recovering the full RCT enables the identification of the causal estimands. Estimating the ATE only on  $\mathcal{D}_s$  may not be feasible even if one aims to

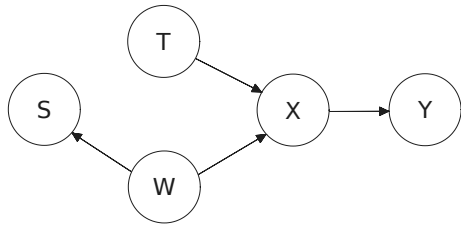
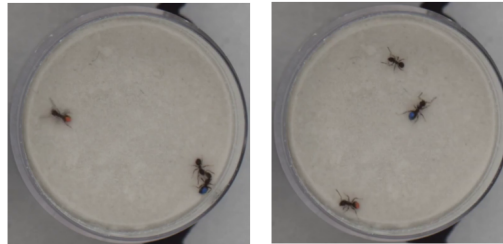


Figure 1: Causal Model for generic partially annotated scientific experiment:  $T$  treatment,  $W$  experimental settings,  $X$  high-dimensional observation,  $Y$  outcome,  $S$  annotation flag.



(a) Grooming (blue to focal)

(b) No Action

Figure 2: Examples of high-dimensional observations  $X$  with corresponding annotated social behaviour  $Y$  from ISTAnt (ours).

adjust for  $W$  due to possible violations of the Positivity Assumption (i.e.,  $0 < P(T = 1|W = w) < 1 \quad \forall w : \mathbb{P}(W = w) > 0$ ). In principle,  $S$  should be assigned randomly (independent from any other variable), but for practical reasons, it is often a function of the experiment settings  $W$ . For example, when annotating grooming in behavioral experiments, experts annotate experiment by experiment, marking the beginning and end of each behavior event, allowing for some selection bias. The experimental setup can be described with the causal model in Figure 1, where we omit the corresponding exogenous random noises for simplicity. For simplicity of exposition, we will now assume a binary outcome, but all the following results naturally generalize to the continuous case.

**Motivating Application and the ISTAnt Dataset.** Ants show strong hygiene behaviors and remove any particles that attach to their body surface, including dust, dirt, and infectious particles. In a process termed “grooming”, they use their mouthparts to pluck off adhering particles, collect and compact them in a pouch in their mouth, and later expulse them as pellets. As a social behavior, ants groom one another to keep all colony members clean and healthy. To understand how social insects like ants may react to changes in their ecosystem, it is of great interest to research in collective hygiene how different particles differentially affect the intensity of grooming by colony members. For this purpose, we recorded groups of three *Lasius neglectus* worker ants interacting in a controlled environment, where we treated one *focal ant* by applying either of two microparticle types to its body surface and observed the grooming activity of the other two towards it. Our exemplary task is to estimate the causal effect of the microparticle type on ant behavior. Sample frames of these recordings used to build our new benchmark are reported in Figure 2.

**Key research question.** Predicting animal behavior is a standard machine learning and computer vision task [Sun et al., 2023, Chen et al., 2023]. At the same time, we hope to use these predictions within the context of a causal downstream task. In this paper, we question whether the naive application of deep learning methods leads to consistent estimates that can be used to draw scientific insights, even if the data we collect is ideal, i.e., a randomized controlled trial. Likewise, in causal inference, the factual effects are always assumed to be given, and the statistical consideration of using machine learning to estimate them is missing.

### 3 Biases in downstream ATE estimation from ML pipelines

In this section, we formalize a model’s bias for a downstream Treatment Effect Estimation and its relationship with (vanilla) prediction accuracy measures. We then highlight possible sources of biases from both the data and the model.

**Definition 3.1** (Treatment Effect Bias). *Let  $f : \mathcal{X} \rightarrow \mathcal{Y}$  a model for  $\mathbb{E}_Y[Y|\mathbf{X} = \mathbf{x}]$ . We define the **treatment effect bias** of  $f$  w.r.t. a treatment  $T$  on an outcome  $Y$  and a signal  $\mathbf{X}$  as:*

$$TEB := \left( \underbrace{\mathbb{E}_{\mathbf{X}|do(T=1)}[f(\mathbf{X})] - \mathbb{E}_{Y|do(T=1)}[Y]}_{\text{Interventional Bias under Treatment}} \right) - \left( \underbrace{\mathbb{E}_{\mathbf{X}|do(T=0)}[f(\mathbf{X})] - \mathbb{E}_{Y|do(T=0)}[Y]}_{\text{Interventional Bias under Control}} \right) \quad (3)$$

$f$  is treatment effect unbiased if  $TEB = 0$ , i.e., the difference among the systematic errors per intervention (over/under estimating) compensates, or directly, the ATE on the predicted outcomes equals the true ATE (despite possible misclassification).

**Lemma 3.1** (Informal). *Assuming the setting described in Section 2. A predictive model  $f$  for the factual outcomes with accuracy  $1-\epsilon$  can lead to  $|TEB(f)| = \frac{\epsilon}{\min_t \bar{P}(T=t)} \geq 2\epsilon$ , which invalidates any causal conclusion when the ATE is comparable with  $\epsilon$  and/or the dataset is unbalanced in  $T$ .*

A formal statement and proof for Lemma 3.1 is reported in Appendix A.1. Lemma 3.1 explicits that misclassification can lead to biased causal conclusion, but not necessarily. Clearly, if the prediction accuracy is perfect (i.e.,  $\epsilon = 0$ ), also the objective of treatment effect estimation is perfect. However, for each error rate  $\epsilon > 0$ , several predictions with different treatment effect biases are possible, from 0 to the worst-case scenario  $\frac{\epsilon}{\min_t \bar{P}(T=t)}$ , which drastically invalidates any causal conclusion for  $\epsilon \gg 0$  or strongly unbalanced dataset with respect to the treatment assignment. Accuracy and similar metrics do not provide a full picture of the goodness of a model for such a downstream task.

Due to the Fundamental Problem in Casual Inference [Holland, 1986], we cannot estimate the treatment effect bias directly. By design (i.e., Ignorability Assumption), the interventional expectations are identified in the conditional ones on the whole population, but not on  $\mathcal{D}_s$  individually due to the effect modifications activated by conditioning on  $S$ . Still, in practice, a validation set, ideally Out-of-Distribution from the training sample in  $\mathcal{D}_s$ , can be considered to approximate the TEB.

**Links to Fairness** This idea of enforcing similar performances (or at least similar systematic errors) among the treated and controlled groups can be revisited in terms of fairness requirements [Verma and Rubin, 2018]. In particular, it strictly relates to Treatment Equality [Berk et al., 2021], where the ratio of false negatives and false positives for both treated and control groups is enforced to be the same, while in TEB we measure the difference, but in a similar spirit. In our setting, the difference is actually a more stable measure since the ratio can be ill-defined when the number of false positive predictions approaches 0. This discussion leaves open where the bias originates and, in the fairness literature, this is reflected in the distinction between bias preserving and bias transforming metrics [Wachter et al., 2021]. For our purposes, the data as a whole is assumed unbiased in principle since we assume an RCT, but the sampling scheme  $S$  could introduce bias in the training data. Orthogonally, the model choices can amplify existing data biases differently or even introduce new ones.

**Data bias from sampling choice** From the assumed causal model illustrated in Figure 1, we have that  $\mathcal{P}^{(X,Y)|S=0}$  generally differs from  $\mathcal{P}^{(X,Y)|S=1}$ . Indeed, conditioning on  $S$  acts as an effect modification on  $X$  and  $Y$ . It follows that the risk in predicting  $Y$  over the annotated population can differ from the expected risk over the whole population, i.e.:

$$\mathbb{E}_{(\mathbf{X},Y)|S=1}[\mathcal{L}(f(\mathbf{X}), Y)] \neq \mathbb{E}_{(\mathbf{X},Y)}[\mathcal{L}(f(\mathbf{X}), Y)]. \quad (4)$$

Due to this distribution shift, we should expect some generalization errors at test time through empirical risk minimization even if  $n_s \rightarrow \infty$ . It follows that the Conditional Average Treatment Effect (CATE) estimate for the experimental settings poorly represented in  $\mathcal{D}_s$  can introduce bias in  $\mathcal{D}_u$ .

*Mitigation: Randomly assigning  $S$  is crucial to suppress any backdoor path and avoid generalization errors. Model selection should also take into account the TEB. Although we cannot estimate it directly, a validation set, ideally Out-of-Distribution in  $\mathbf{W}$ , should be considered to bound the TEB, replacing the interventional distributions with the corresponding conditionals.*

**Model bias from the encoder choice** Since  $\mathbf{X}$  is high-dimensional, we decompose the model  $f$  in  $h \circ e$ , where  $e$  is an encoder potentially pre-trained on a much larger corpus through a representation learning algorithm and  $h$  is a simple decoder (e.g., multi-layer perceptron) for classification. A good representation should be both sufficient and minimal [Achille and Soatto, 2018]. If a representation is only sufficient, redundant information from  $\mathbf{W}$  or  $S$  could be preserved, potentially leading to systematic errors on  $\mathcal{D}_u$  due to spurious correlations and the abovementioned covariates shift. Frozen state-of-the-art models are most likely not minimal for our task, making the sampling choices even more relevant. If the representation is not sufficient, then it is biased by definition.

*Mitigation: Before deploying a new backbone, one should attempt to quantify its biases. If needed, new methodologies to mitigate this bias during adaptation should be investigated. Overall, models with lower bias may be preferable even if they have lower accuracy.*

**Discretization Bias** We can encounter a final source of bias in post-processing the predictions. Indeed, despite the majority of the classification methods directly modeling the conditional expectation  $\mathbb{E}[Y|X = x]$ , we could naively be tempted to binarize this estimate to the most probable prediction or setting a fixed threshold. See indeed how the default choices for the `predict` module, even in established libraries, e.g., Logistic Regression implementation in `sklearn.linear_model` [Pedregosa et al., 2011], is to output the most probable prediction directly. Similarly, even `econML`, the most popular library for causal machine learning [Battocchi et al., 2019], allows for binary outcome prediction methods. Despite being apparently innocent and common practice in classification, discretizing the conditional expectation is biased for downstream treatment effect estimation.

**Theorem 3.1.** [Informal] *Let a binary classification model converge to the true probability of the outcome given its (high-dimensional) signal. Then, its discretization (i.e., rounding the prediction to  $\{0,1\}$  with a fixed threshold) also converges, but to a different quantity with a different expectation. It follows that, for causal downstream tasks from ML pipelines, discretizing the predictions biases the ATE estimation.*

A formal statement and proof of Theorem 3.1 is reported in Appendix A.2. It shows that even if we rely on a consistent estimator of the factual outcome for each subgroup, its discretization would still converge but on a different quantity, i.e., it is biased. There is then no reason to discretize a model for  $\mathbb{E}_X[Y|X]$  if we can model it directly, e.g., using sigmoid or softmax activation [Senn and Julious, 2009, Fedorov et al., 2009]. Likewise, if there is uncertainty over human annotations (e.g., because multiple raters disagree), the soft label should be used and not the majority one.

**Example 1.** *To intuitively visualize this result, consider a generative process following the causal model introduced in Figure 1. Let  $\hat{f}$  a model for  $\mathbb{E}[Y|X = x]$  trained by logistic regression over  $n$  samples and  $\hat{f}^*$  its discretization. Let the Empirical Associational Difference (EAD) of  $\hat{f}$  converge to its AD, then the EAD of  $\hat{f}^*$  still converges but to its own AD, which significantly differs (depending on the randomness in  $\mathbb{P}(Y|X)$  mechanism). In Figure 3, we report the results of a Monte Carlo simulation for an instance of this generative process. A full description of the Structural Causal Model and theoretical derivation of the limits is reported in Appendix B.1.*

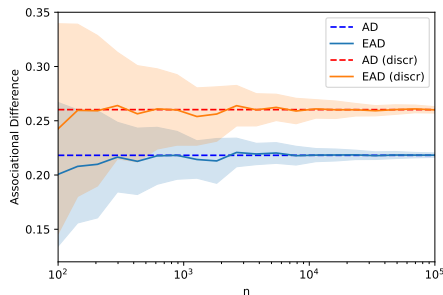


Figure 3: Monte-Carlo simulation of the discretization bias' convergence result.

*Mitigation: Never arbitrarily discretize predictions for downstream treatment effect estimation.*

## 4 Related Works

**Representation learning for scientific applications** The setting we consider is very common, and there are several benchmarks that have studied representation learning as a means to help domain experts in sciences, for example, [Sun et al., 2023, Beery et al., 2018, Kay et al., 2022, Chen et al., 2023] in ecology alone. However, these works focus on downstream prediction accuracy following standard machine learning evaluation practices, which do not necessarily indicate good downstream causal predictions. One notable positive example is Beery et al. [2018], as computing the prediction accuracy separately for different locations allows us to estimate the bias of the model. Overall, we argue that when the ultimate purpose of training a machine learning model is to support scientists in answering some research question that is causal in nature, the specific question should be part of the design and evaluation of the benchmark. For this reason, our paper uniquely starts from the causal downstream task. Only then can we formalize the properties that methods should have in order to do well and be

useful in answering the overarching scientific question. To the best of our knowledge, ours is the first real-world computer vision data set with an associated well-defined and real causal downstream task.

**Causal representation learning** In our analysis (both theoretically and experimentally), we focused on traditional representation learning algorithms, but there is a whole community interested in identifying causal variables from high-dimensional observations [Schölkopf et al., 2021]. Superficially, identifying  $Y$  may be useful to estimate the ATE. However, all existing methods seem to cover two main classes of assumptions that are unfortunately inapplicable to our setting. Interventional methods [Ahuja et al., 2023, Buchholz et al., 2023, Squires et al., 2023, Varici et al., 2023, Zhang et al., 2024] require intervening on the behavior  $Y$ , which is practically impossible and, even if we could, then we would not need to identify it. Multi-view approaches [Ahuja et al., 2022, Brehmer et al., 2022, Locatello et al., 2020, von Kügelgen et al., 2021, Yao et al., 2024] would require access to positive and negative pairs of samples with respect to  $Y$ . However, it is not clear how to construct such pairs in our setting without knowing  $Y$  already. Further, all these approaches only cover continuous variables. A notable exception is Kivva et al. [2021], which covers discrete variables but has a non-degeneracy assumption (Assumption 2.4) that is severely violated in our case (i.e., most pixels are not affected by the behavior variable because ants are small). For these reasons, despite having a very clear causal downstream task, we had to resort to classical representation learning algorithms that are not identifiable. We hope that our data set can serve as a new real-world benchmark for developing algorithms with realistic assumptions that can be applied in practice.

**Other Related Works** In causal inference, only Chakraborty et al. [2022] shows how to use semi-supervised learning to perform imputation on missing effect annotations. Unfortunately, their setting is comparatively very low-dimensional (observations are 200 binary variables). Instead, we consider high-dimensional real-world images in a representation learning setting, which introduces additional new challenges as described in Section 3. Remarkably, they do not discuss discretization bias. Curth et al. [2024] already mentioned that the Positivity and Ignorability/Unconfoundness Assumptions are critical for using machine learning in the context of ATE estimation. However, their work does not explain precisely how confounding effects can arise in the representation learning setting, which we thoroughly addressed. Close in spirit to our discussion are [Angelopoulos et al., 2023, Zrníc and Candès, 2024], considering the role of predictions in statistical estimates. Our setting is related but additionally motivated by the hope of leveraging *causal* identification properties on the prediction-powered dataset.

## 5 Experimental setup

We validate the theoretical results from Section 3 on our new real-world dataset. We assume  $\mathcal{D}_s \cup \mathcal{D}_u$  being a full RCT, and we compare the treatment effect biases among several design choices in annotating and modeling. Overall, we fine-tuned 6 480 different models and tested all the mitigations proposed. We then replicate the experiments on CausalMNIST, a new synthetic benchmark we propose that allows controlling for the causal effect.

### 5.1 New real-world dataset: ISTAnt

We applied microparticles to the body surface of a (focal) *Lasius neglectus* worker ant and recorded the behavioral reaction this treatment elicits in two other worker ants from the same colony. To distinguish between the treated individual and the untreated two nestmates, the latter had been color-coded by a dot of blue or orange paint, respectively, before the experiment. We used two different microparticle treatments to compare grooming responses by the nestmates between treatment types, assigning them at random (i.e., RCT). For five batches, we simultaneously filmed nine ant groups of three ants each under a single camera setup in a custom-made box with controlled lighting and ventilation. In total, we collected 44 videos<sup>1</sup> of 10 minutes at 30fps each for a total of 792 000 frames annotated following a blind procedure, and we run the analysis at 2fps for a total of 52 800 frames. More details about the experiment design are reported in Appendix C. We remark that this is the first real-world data set for treatment effect estimation from high-dimensional observations, which we will release to accelerate future research. Since it encompasses a real-world scientific question, we can, at

<sup>1</sup>One video was discarded for analysis since a leg of one of the two nestmates got stuck in the dot of the color code, impairing its behavior.

best, enforce the Ignorability Assumption by design in the trial. We do not have actual control over the underlying causal model and the causal effect. We take the treatment effect estimation computed with the expert annotations as ground truth.

**Annotation Sampling** Annotating frames individually is significantly more expensive in terms of time and not adopted in practice. The practical gold standard through current software for human annotation is per-video random annotation, where only a few videos taken at random are fully annotated. We compared this criterion with per-video batch ( $W_1$ ) and per-video position ( $W_2$ ) annotation criteria, where only the videos in certain batches or positions were considered in  $\mathcal{D}_s$ . For each of the three criteria, we further considered a many-shots ( $\mathcal{D}_s \gg \mathcal{D}_u$ ) and a few-shots setting ( $\mathcal{D}_s \ll \mathcal{D}_u$ ). Details about the dataset splitting per annotation criteria are in Appendix D.1.

**Modeling** We modeled  $f$  as a composition of a frozen pre-trained encoder  $e$  and a multi-layers perceptron  $h$  fine-tuned on  $\mathcal{D}_s$ . For the encoder, we compared six different established Vision Transformers (ViT), mainly varying the training procedure: ViT-B [Dosovitskiy et al., 2020], ViT-L [Zhai et al., 2023], CLIP-ViT-B,-L [Radford et al., 2021], MAE [He et al., 2022], DINOv2 [Oquab et al., 2023]. For each encoder, we considered the representation extracted (i) by the class encoder (class), (ii) by the average of all the other tokens (mean), or (iii) both concatenated (all). For each representation extracted we trained different heads, varying the number of hidden layers (1 or 2 layers with 256 nodes each with ReLU activation), learning rates (0.05, 0.005, 0.0005) for Adam optimizer [Kingma and Ba, 2014] (10 epochs) and target (independent double prediction of 'blue to focal' and 'orange to focal' grooming, or unique prediction of grooming either 'blue to focal' or 'orange to focal') via (binary) cross-entropy loss. We either discretized or not the output of the model, already in  $[0, 1]$  due to the sigmoid final activation. For each configuration, we repeated the training with five different random seeds. A summary of the architectures and training description is in Appendix D.2.

**Evaluating** For each trained model, we computed the binary cross-entropy loss, accuracy, balanced accuracy, and TEB on validation; and accuracy, balanced accuracy, TEB, and TEB using discretization on the full dataset  $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_u$  (where the average potential outcomes in the TEB are estimated by the sample mean). Since the ATE does not have a reference scale, for interpretation purposes, we replaced the TEB with Treatment Effect Relative Bias (TERB = TEB/ATE) in the visualizations.

## 5.2 CausalMNIST

CausalMNIST is a new synthetically generated visual dataset we designed for downstream treatment effect estimation. It is a colored manipulation of the MNIST dataset [LeCun, 1998], following an underlying generative process in agreement with the causal model assumed in our framework (see Figure 1). We explicitly controlled the ATE and generated 400 different samples from such a population (each one as large as the MNIST dataset, i.e., 60k images), allowing for Monte-Carlo simulations to accurately provide confidence intervals of our estimations. We omitted a comparison among pre-trained encoders since the visual task is relatively simple and can be solved directly by a simple convolutional neural network in a supervised fashion. A full description of the dataset is in Appendix E, together with its experiments, which align with our conclusions from ISTAnt.

## 6 Results

**Annotating criteria matter** Theory suggests that biased annotating criteria (i.e., depending on the experimental settings) can lead to biased treatment effect estimation, wrongly retrieving the conditional treatment effect on unseen experimental settings. Figure 4 validates this observation, particularly in the few-shots regime. Despite the average estimation of the TEB is (almost) always biased, as illustrated in Table 1, the distribution for (per-video) random annotation is more centered towards 0. The benefits of random sampling are less obvious in the many-shots regime since  $\mathcal{D}_u$  becomes less and less Out-of-Distribution. Still, this setting is rarely the case in practice since scientist hope to label  $|\mathcal{D}_s| \ll |\mathcal{D}_u|$  frames to have a concrete advantage in their workflow.

**Encoder Bias** Vanilla classification evaluation (e.g., accuracy, F1-score, etc.) well describes the goodness of a representation for a predictive downstream task. However, it is still unclear how to measure the goodness of a representation for a causal downstream task since we do not directly

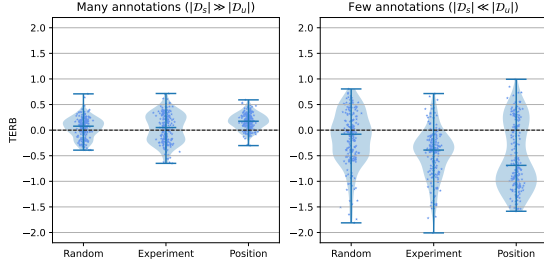


Figure 4: Violin plots comparing the Treatment Effect Relative Bias (TERB) per annotation criteria in few and many annotations regime. Biased annotations lead to biased ATE estimation (i.e.,  $TERB \neq 0$ ) and random annotation should be preferred.

Annot.	Criteria	$t$	$p$ -value
Many	Random	3.581	$4 \cdot 10^{-4}$
	Experiment	1.918	0.0564
	Position	14.982	$\approx 0$
Few	Random	-4.46	$1.3 \cdot 10^{-5}$
	Experiment	-13.417	$\approx 0$
	Position	-11.250	$\approx 0$

Table 1: Two-sided  $t$ -test for  $\mathcal{H}_0 : \mathbb{E}[TEB(f)] = 0$  over the 200 best models in overall Balanced Accuracy per splitting criteria. We found statistical evidence to reject the hypothesis that even the best models alone are unbiased for (almost) each annotation criterion.

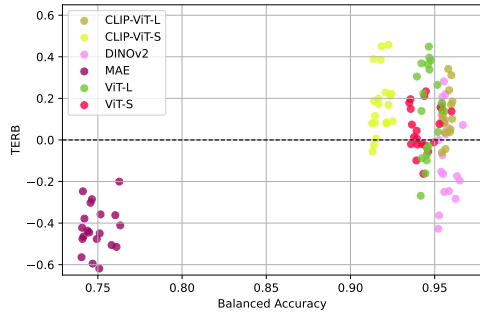


Figure 5: Scatter plot comparing the TERB and balanced accuracy in prediction among the 20 best models per 6 established encoders. Despite different downstream prediction performances, all the encoders (with exceptions of MAE) lead to similar TERB (up to  $\pm 50\%$ ).

Encoder	Fréchet Distance (FD)		
	Random	Experiment	Position
CLIP-ViT-L	$422.6 \pm 87.9$	$461.2 \pm 151.7$	$605.2 \pm 130.0$
CLIP-ViT-S	$329.6 \pm 113.7$	$341.7 \pm 120.2$	$486.8 \pm 97.6$
DINOv2	$360.0 \pm 183.3$	$413.0 \pm 222.5$	$514.4 \pm 244.9$
MAE	$275.0 \pm 20.9$	$211.8 \pm 16.5$	<b><math>760.9 \pm 122.4</math></b>
ViT-L	$499.7 \pm 32.0$	$503.4 \pm 108.6$	$681.7 \pm 159.0$
ViT-S	$308.7 \pm 69.8$	$307.4 \pm 67.7$	$423.9 \pm 103.9$

Table 2: FD distance among  $\mathcal{D}_s$  and  $\mathcal{D}_u$ , representing the average distribution distance ( $\pm$  standard deviation) after normalization per encoder varying splitting criteria (e.g., few and many shots regime) and tokens considered. Representations with higher FD distance on position splitting (where the background changes the most) compared to the other splitting rely on more spurious correlations for our task (i.e., not minimal).

observe the ground truth (fundamental problem of Causal Inference). Even in our simple setting where we can easily identify the treatment effect over the whole population, it is not possible to condition just on a biased subsample (e.g., the validation set). Figure 5 shows clearly how the TERB doesn’t correlate with balanced accuracy on the whole sample once it is sufficiently good (i.e.,  $> 0.9$ ). Even among models with balanced accuracy  $> 0.95$  we estimated TERBs up to  $\pm 50\%$ , which can drastically lead to wrong causal conclusions. Among the different encoders, MAE is significantly underperforming all the others. We postulate the reason for this gap is that the masked reconstruction training leads to overly focus on background conditions instead of the comparatively small ants. Evidence for this hypothesis is reported in Table 2 where we observe that for ‘position’ splitting criteria, the Fréchet Distance between the extracted embeddings by MAE in  $\mathcal{D}_s$  and  $\mathcal{D}_u$  is maxima and significantly higher than for the other splittings, probably due to spurious correlation with the background which is indeed non changing as much for “random” and “experiment” splitting. Despite some (e.g., DINOv2 and CLIP-ViT-L) having better downstream predictive performances, the other encoders all have similar TERB ranges. New criteria to better estimate and bound the treatment effect bias already on validation and methodologies to unbiased these models during training are required.

**Discretization Bias** We considered the absolute value of the TEB over all the 6 480 fine-tuned models, evaluating independently the models predicting both ‘Blue to Focal’ and ‘Orange to Focal’ grooming for a total of 9 720 evaluations. We tested ( $t$ -test):

$$\mathcal{H}_0 : \mathbb{E}[|TEB(f)|] = \mathbb{E}[|TEB(\mathbb{1}_{[0.5,1]}(f))|] \quad vs \quad \mathcal{H}_1 : \mathbb{E}[|TEB(f)|] < \mathbb{E}[|TEB(\mathbb{1}_{[0.5,1]}(f))|] \quad (5)$$

We found strong statistical evidence to confirm that discretizing the model outcome worsens treatment effect estimation ( $t$  statistic= $-10.42$ ,  $p$ -value= $1.07 \cdot 10^{-25}$ ), confirming Theorem 3.1 .



**Prediction is not Causal Estimation** Distinct statistical and causal objectives cannot be used as a proxy for one another. We already formalized this in Lemma 3.1 and partially observed it in Figure 5. In Figure 6, we systematically show it by comparing the rank-correlation among 1 620 models. We further observed that simply computing the TEB on a small validation is a better predictor of the TEB over the full dataset than the metrics focused on prediction accuracy (even on the full dataset). For the few-shot and experiment sampling (the most realistic), if we select the single best model on validation based on the TEB versus the accuracy, we underestimate the effect by 11% and 18%, respectively. While this is not perfect, is a significant improvement. We encourage future research to investigate theoretical generalization guarantees and new techniques to approximate the TEB on validation data.

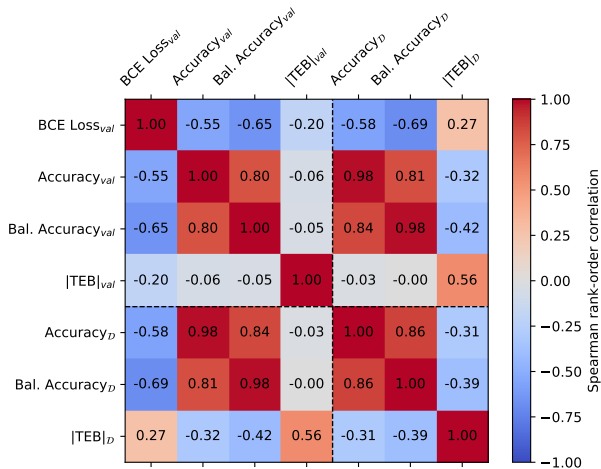


Figure 6: Spearman rank-order correlation matrix comparing different metrics for model selection on validation (subscript *val*) and over the full dataset (subscript  $\mathcal{D}$ ). We considered all the 1 620 fine-tuned models to predict ‘Blue to Focal’ or ‘Orange to Focal’ grooming in few-annotations regime (i.e.,  $|\mathcal{D}_s| \ll |\mathcal{D}_t|$ ). Standard prediction metrics on validation correlate, but they are almost independent of the  $|TEB|_{val}$ . Similarly, they correlate with the prediction metrics on the full dataset but poorly predict the  $|TEB|_{\mathcal{D}}$ . On the other hand,  $|TEB|_{val}$  is the most correlated metric with  $|TEB|_{\mathcal{D}}$ , unlike even the prediction metrics on the full dataset.

**Discussion** Overall, our results clearly show that it is possible to leverage pre-trained deep learning models to accelerate the annotation of experimental data and obtain downstream causal estimates that are consistent with those from domain experts. At the same time, we find that experimental practices need to incorporate the specific needs of these causal downstream tasks. While our theoretical statements are “worst case scenarios” and only indicate that bias *can* arise (but does not always have to), we find empirical validation that it unfortunately easily manifests in practice. Remarkably, the fact that we performed and collected data within a randomized controlled trial, which is the best-case scenario of causal inference, did not alleviate the issue. Therefore, we can expect that the opportunities for bias can be even greater in observational settings, and even greater care is needed in model selection with the TEB and adaptation-time debiasing techniques.

## 7 Conclusions and Limitations

As AI models are increasingly used to answer scientific questions and support human decision-making, it is important to understand how design choices in machine learning pipelines affect the final results. In this paper, we took a closer look at the impact of pre-trained deep learning models in answering downstream causal treatment effect questions. We presented a real-world example in experimental behavioral ecology, creating the first-ever data set for treatment effect estimation from high-dimensional observations. Both theoretically and empirically, we observed that common choices, most notably discretizing the predictions and using in-distribution accuracy for model selection, can significantly affect the downstream conclusions. Two clear limitations of this work are that we did not do anything to the training to mitigate the bias, we kept the backbones frozen, and we did not incorporate the unlabelled data for semi-supervised training. Here, it would be very interesting to study how tools developed in the fairness literature can be extended to causal questions. For future benchmarks targeting scientific applications, we remark that it is vital to include the actual downstream question in the design of the data set. Otherwise, there is a risk that any model produced on that data may be unusable in practice, as it can bias the answer on an otherwise perfectly designed experiment. Finally, we would recommend that future work in causal representation learning starts from a clear downstream task like the one presented in this paper and works backward to reasonable assumptions. To facilitate this process, we will release our data set including all the experimental

variables, so that relevant future work on e.g., discovering confounding or semi-supervised effect discovery, can take place on a real problem.

## Acknowledgments

We thank Piersilvio De Bartolomeis, and the full Causal Learning and Artificial Intelligence (CLAI) group at ISTA for the extremely helpful discussions. Riccardo Cadei was supported by a Google Research Scholar Award and a Google Initiated Gift to Francesco Locatello. We thank the Social Immunity team at ISTA, particularly Michaela Hönigsberger and Wilfrid Jean Louis, for supporting the ecological experiment and Farnaz Beikzadeh Abbasi, Luisa Fiebig and Martin Estermann for annotating ant behavior in ISTAnt.

## References

- Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.
- Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022.
- Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023.
- Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnica. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- Hugh Bailey. The obs project contributors. *Open Broadcasting Software*, 2017.
- Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. <https://github.com/py-why/EconML>, 2019. Version 0.x.
- Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.
- Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The auto arborist dataset: a large-scale benchmark for multiview urban forest monitoring under domain shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21294–21307, 2022.
- Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.
- Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.
- Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.
- Abhishek Chakraborty, Guorong Dai, and Eric Tchetgen Tchetgen. A general framework for treatment effect estimation in semi-supervised and high dimensional settings. *arXiv preprint arXiv:2201.00468*, 2022.
- Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13061, 2023.

- Alicia Curth, Richard W Peck, Eoin McKinney, James Weatherall, and Mihaela van Der Schaar. Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clinical Pharmacology & Therapeutics*, 2024.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gizen, Rob Carver, Marcin Andrychowicz, Jason Hickey, et al. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1):1–10, 2022.
- Valerii Fedorov, Frank Mannino, and Rongmei Zhang. Consequences of dichotomization. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 8(1):50–61, 2009.
- Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.
- William OH Hughes, Jørgen Eilenberg, and Jacobus J Boomsma. Trade-offs in group living: transmission and disease resistance in leaf-cutting ants. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1502):1811–1819, 2002.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: a benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision*, pages 290–311. Springer, 2022.
- S Kimball and P Mattis. The gimp development team. *GNU image manipulation program version*, 2: 34, 2023.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34:18087–18101, 2021.
- Matthias Konrad, Meghan L Vyleta, Fabian J Theis, Miriam Stock, Simon Tragust, Martina Klatt, Verena Drescher, Carsten Marr, Line V Ugelvig, and Sylvia Cremer. Social transfer of pathogenic fungus promotes active immunisation in ant colonies. *PLoS Biology*, 10(4):e1001300, 2012.
- Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- Jacob Lin, Miguel Farinha, Edward Gryspeerdt, and Ronald Clark. Volumetric cloud field reconstruction. *arXiv preprint arXiv:2311.17657*, 2023.
- Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 13–18 Jul 2020.

- Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.
- Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. Deep learning for cellular image analysis. *Nature methods*, 16(12):1233–1246, 2019.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.
- Rebeca B Rosengaus, Amy B Maxmen, Laran E Coates, and James FA Traniello. Disease resistance: a benefit of sociality in the dampwood termite *zootermopsis angusticollis* (isoptera: Termopsidae). *Behavioral ecology and sociobiology*, 44:125–134, 1998.
- Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.
- Jakob Runge. Modern causal inference approaches to investigate biodiversity-ecosystem functioning relationships. *nature communications*, 14(1):1917, 2023.
- Jonathan M Samet, Francesca Dominici, Frank C Curriero, Ivan Coursac, and Scott L Zeger. Fine particulate air pollution and mortality in 20 us cities, 1987–1994. *New England journal of medicine*, 343(24):1742–1749, 2000.
- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- Stephen Senn and Steven Julious. Measurement in clinical trials: a neglected issue for statisticians? *Statistics in medicine*, 28(26):3189–3209, 2009.
- Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, volume 202, pages 32540–32560. PMLR, 2023.
- Jennifer J Sun, Markus Marks, Andrew Wesley Ulmer, Dipam Chakraborty, Brian Geuther, Edward Hayes, Heng Jia, Vivek Kumar, Sebastian Oleszko, Zachary Partridge, et al. Mabe22: a multi-species multi-task benchmark for learned representations of behavior. In *International Conference on Machine Learning*, pages 32936–32990. PMLR, 2023.
- Suramya Tomar. Converting video formats with ffmpeg. *Linux journal*, 2006(146):10, 2006.
- Egbert H Van Nes, Marten Scheffer, Victor Brovkin, Timothy M Lenton, Hao Ye, Ethan Deyle, and George Sugihara. Causal feedbacks in climate change. *Nature Climate Change*, 5(5):445–448, 2015.
- Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.
- Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

- 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. *Advances in neural information processing systems*, 34:16451–16467, 2021.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *West Virginia Law Review*, 123(3), 2021.
- Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *International Conference on Learning Representations*, 2024.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.
- Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024.
- Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

## A Proofs

### A.1 Proof of Lemma 3.1

**Lemma.** Let  $T \sim Be(p_T)$ ,  $Y \sim Be(p_Y)$ ,  $\mathbf{X} \sim \mathcal{P}^{\mathbf{X}}$  and let  $f : \mathcal{X} \rightarrow [0, 1]$  a model for  $\mathbb{E}_Y[Y|\mathbf{X} = \mathbf{x}]$ . Assuming (i) Ignorability (i.e.,  $T \perp Y|do(T = 1), Y|do(T = 0)$ ), (ii)  $\mathbb{E}_{\mathbf{X}}[\mathbb{1}_{[k,1]}(f(\mathbf{X})) = f(\mathbf{X})] = 0$  where  $k \in [0, 1]$ , and (iii)  $f$  with accuracy  $1 - \epsilon$ , i.e., :

$$\mathbb{P}(\mathbb{1}_{[k,1]}(f(\mathbf{X})) = Y) = 1 - \epsilon \quad \text{with } \epsilon \in [0, 1], \quad (\text{Classification Accuracy})$$

then  $|TEB(f)| \leq \frac{\epsilon}{\min_t P(T=t)}$  and the worst-case  $|TEB(f)| = \frac{\epsilon}{\min_t P(T=t)} \geq 2\epsilon$  is reached when all the misclassification over (or under) estimates the factual outcome in the smaller in size treatment group.

*Proof.* Starting from the definition of Treatment Effect Bias and using assumption (i):

$$\begin{aligned} |TEB(f)| &= |(\underbrace{\mathbb{E}_{\mathbf{X}|do(T=1)}[f(\mathbf{X})] - \mathbb{E}_{Y|do(T=1)}[Y]}_{\text{Interventional Bias under Treatment}}) - (\underbrace{\mathbb{E}_{\mathbf{X}|do(T=0)}[f(\mathbf{X})] - \mathbb{E}_{Y|do(T=0)}[Y]}_{\text{Interventional Bias under Control}})| = \\ &= |(\underbrace{\mathbb{E}_{\mathbf{X}|T=1}[f(\mathbf{X})] - \mathbb{E}_{Y|T=1}[Y]}_{\epsilon_1}) - (\underbrace{\mathbb{E}_{\mathbf{X}|T=0}[f(\mathbf{X})] - \mathbb{E}_{Y|T=0}[Y]}_{\epsilon_0})| = \\ &= |\epsilon_1 - \epsilon_0| \leq |\epsilon_1| + |\epsilon_0| \end{aligned} \quad (6)$$

where

$$\epsilon_t := \mathbb{E}_{\mathbf{X}|T=t}[f(\mathbf{X})] - \mathbb{E}_{Y|T=t}[Y] \quad \forall t \in \{0, 1\} \quad (7)$$

represent the overestimation of each conditional outcome expectation (i.e., conditional bias under treatment/control).

By assumption (iii) and the law of total probability:

$$\begin{aligned} \epsilon &= \mathbb{P}(\mathbb{1}_{[k,1]}(f(\mathbf{X})) \neq Y) = \\ &= \mathbb{P}(\mathbb{1}_{[k,1]}(f(\mathbf{X})) \neq Y|T = 0) \cdot \mathbb{P}(T = 0) + \mathbb{P}(\mathbb{1}_{[k,1]}(f(\mathbf{X})) \neq Y|T = 1) \cdot \mathbb{P}(T = 1). \end{aligned} \quad (8)$$

By Jensen's inequality and linearity of the expected value:

$$\begin{aligned} |\epsilon_t| &= |\mathbb{E}_{(\mathbf{X}, Y)|T=t}[f(\mathbf{X}) - Y]| \leq \\ &\leq \mathbb{E}_{(\mathbf{X}, Y)|T=t}[|f(\mathbf{X}) - Y|] = \\ &= \mathbb{E}_{(\mathbf{X}, Y)|T=t}[|f(\mathbf{X}) - \mathbb{1}_{[k,1]}(f(\mathbf{X})) + \mathbb{1}_{[k,1]}(f(\mathbf{X})) - Y|] \leq \\ &\leq \mathbb{E}_{(\mathbf{X}, Y)|T=t}[|f(\mathbf{X}) - \mathbb{1}_{[k,1]}(f(\mathbf{X}))|] + \mathbb{E}_{(\mathbf{X}, Y)|T=t}[|\mathbb{1}_{[k,1]}(f(\mathbf{X})) - Y|] = \\ &= \mathbb{E}_{(\mathbf{X}, Y)|T=t}[|f(\mathbf{X}) - \mathbb{1}_{[k,1]}(f(\mathbf{X}))|] + \mathbb{P}(\mathbb{1}_{[k,1]}(f(\mathbf{X})) \neq Y|T = t). \end{aligned} \quad (9)$$

Combining Equation 8 and 9 we using the assumption (ii), for all  $t \in \{0, 1\}$  we have:

$$\begin{aligned} \epsilon &\geq |\epsilon_0| \cdot \mathbb{P}(T = 0) + |\epsilon_1| \cdot \mathbb{P}(T = 1) - \mathbb{E}_{\mathbf{X}}[\mathbb{1}_{[k,1]}(f(\mathbf{X})) = f(\mathbf{X})] = \\ &= |\epsilon_0| \cdot \mathbb{P}(T = 0) + |\epsilon_1| \cdot \mathbb{P}(T = 1) \end{aligned} \quad (10)$$

And finally, combining this with Equation 6, we get:

$$|TEB(f)| \leq \frac{\epsilon}{\min_t P(T = t)}. \quad (11)$$

The bound we found corresponds to the worst-case scenario where we misclassify, only overestimating or only underestimating, always in the least probable treated group. Since  $T$  is binary, then  $(\min_t P(T = t)) > 0.5$ , and so the thesis.

*Comment:* Assumption (ii) is only used to find the worst-case scenario explicitly. Similar results can be stated bounding this discretization error.

□

## A.2 Proof of Theorem 3.1

**Theorem.** Let  $T \sim Be(p_T)$ ,  $Y \sim Be(p_Y)$  and  $\mathbf{X} \sim \mathcal{P}^{\mathbf{X}}$ . For all  $t \in \{0, 1\}$ , let  $\hat{\tau}_n(\mathbf{X}, t)$  a succession converging in mean  $L^1$  to  $\tau(\mathbf{X}, t) := \mathbb{E}_Y[Y|\mathbf{X}, T = t]$ , i.e.,

$$\mathbb{E}_{\mathbf{X}} [|\hat{\tau}_n(\mathbf{X}, t) - \tau(\mathbf{X}, t)|] \xrightarrow{n} 0 \quad (12)$$

Let  $\hat{\tau}_n^*(\mathbf{X}, t) = \mathbb{1}_{[k, 1]}(\hat{\tau}_n(\mathbf{X}, t))$  for all  $n$  and  $\tau^*(\mathbf{X}, t) = \mathbb{1}_{[k, 1]}(\tau(\mathbf{X}, t))$ , where  $\mathbb{1}_{[k, 1]} : \mathbb{R} \rightarrow \{0, 1\}$  is the indicator function with threshold  $k \in [0, 1]$ . Assuming  $\tau(\mathbf{X}, t)$  having continuous CDF (i.e.,  $\mathcal{F}_{\tau(\mathbf{X}, t)} \in \mathcal{C}^0$ ), then:

$$\mathbb{E}_{\mathbf{X}} [|\hat{\tau}_n^*(\mathbf{X}, t) - \tau^*(\mathbf{X}, t)|] \xrightarrow{n} 0 \quad (13)$$

but

$$\mathbb{E}_{\mathbf{X}}[\tau^*(\mathbf{X}, t)] \neq \mathbb{E}_{\mathbf{X}}[\tau(\mathbf{X}, t)] \quad \forall k \in [0, 1]/\bar{k}, \quad (14)$$

i.e., they are generally different unless for a value  $\bar{k} \in [0, 1]$  depending on the distribution of  $\tau(\mathbf{X}, t)$  (not observed in practice).

*Proof.* Convergence in mean  $L^1$  of that binarized estimator (Equation 13) follows directly from the fact the  $L^1$  convergence implies convergence in distribution and Portmanteau Theorem (using the continuity assumption of  $\tau(\mathbf{X}, t)$  CDF).

It only remains to show that the expectations of their limits generally differ. By developing the expected value of the  $\tau^*(\mathbf{X}, t)$  we have:

$$\mathbb{E}_{\mathbf{X}} [\tau^*(\mathbf{X}, t)] = \int \mathbb{1}_{[k, 1]} \tau(\mathbf{X}, t) dP_{\mathbf{X}} = \quad (15)$$

$$= \mathbb{P}(\tau(\mathbf{X}, t) \geq k) \neq \mathbb{E}_{\mathbf{X}}[\tau(\mathbf{X}, t)] \quad \forall t \in \{0, 1\}, k \in [0, 1]/\bar{k}. \quad (16)$$

where, by definition,  $\bar{k}$  is the  $\alpha$ -quantile for  $\tau(\mathbf{X}, t)$ , with  $\alpha = 1 - \mathbb{E}_{\mathbf{X}}[\tau(\mathbf{X}, t)]$  (uniqueness due to the continuity of  $\tau(\mathbf{X}, t)$  CDF). □

## B Additional Examples

### B.1 Full Description Example 1

Let's consider the following structural causal model in alignment with the generative process in Figure 1. Noises:

$$n_T \sim Be(p_T) \quad (17)$$

$$n_W, n_X \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad (18)$$

$$n_Y \sim \mathcal{N}(0, \sigma_Y^2) \quad (19)$$

where  $p_T \in (0, 1)$  and  $\sigma_Y^2 > 0$ . and structural equations:

$$T := n_T \quad (20)$$

$$W := n_W \quad (21)$$

$$X := T + W + n_X \quad (22)$$

$$Y := \mathbb{1}_{[0, +\infty)}(X + n_Y) \quad (23)$$

By the Law of Total Probability and additivity of Gaussian distributions, it follows:

$$X \sim \mathcal{N}(p_T, 2 + p_T \cdot (1 - p_T)) \quad (24)$$

$$X|T = 1 \sim \mathcal{N}(1, 2) \quad (25)$$

$$X|T = 0 \sim \mathcal{N}(0, 2) \quad (26)$$

$$Y|T = 1 \sim Be\left(\phi\left(\frac{1}{\sqrt{2 + \sigma_Y^2}}\right)\right) \quad (27)$$

$$Y|T = 0 \sim Be(0.5) \quad (28)$$

$$Y^* := \begin{cases} 1 & \text{if } \mathbb{E}_Y[Y|X] > 0.5 \\ 0 & \text{if } \textit{otherwise} \end{cases} \quad (29)$$

Then:

$$Y^*|T = 1 \sim Be(\phi(1/\sqrt{2})) \approx Be(0.76) \quad (30)$$

$$Y^*|T = 0 \sim Be(0.5) \quad (31)$$

And:

$$AD_{Y,T} = \phi\left(\frac{1}{\sqrt{2} + \sigma_Y^2}\right) - 0.5 \quad (32)$$

$$AD_{Y^*,T} = \phi(1/\sqrt{2}) - 0.5 \neq AD_{Y,T} \quad (33)$$

Let  $\hat{f}(x)$  a logistic regression estimator for  $\mathbb{E}[Y|X = x]$  and:

$$\hat{f}^*(x) := \begin{cases} 1 & \text{if } \hat{f}(x) > 0.5 \\ 0 & \text{if } \textit{otherwise} \end{cases} \quad \forall x \in \mathbb{R}. \quad (34)$$

Setting  $p_T = 0.5$  and  $\sigma_Y^2 = 1$ , we run a Monte-Carlo simulation with 50 different random seeds per sample size  $n$ , estimating the associational difference by the empirical associational difference (EAD), i.e., using the sample mean. The results are reported in Figure 3. We observe that  $\hat{f}$  leads to a consistent estimate of the true associational difference, which corresponds to the ATE due to the Ignorability Assumption encoded in the causal model:

$$EAD_{\hat{f}(X),T} \xrightarrow{n} AD_{Y,T} = ATE_{Y,T} \quad (35)$$

and so:

$$EAD_{\hat{f}^*(X),T} \xrightarrow{n} AD_{Y^*,T} = ATE_{Y^*,T} \quad (36)$$

But, according to Theorem 3.1, its discretization is biased:

$$AD_{Y^*,T} - AD_{Y,T} = \left(\phi(1/\sqrt{2}) - \phi(1/\sqrt{3})\right) \approx 0.042 > 0 \quad (37)$$

and more generally, the stronger is the variance in the effect random noise  $n_Y$ , the bigger is the bias.

## C IStAnt

In our study, we analyzed grooming behavior in the ant *Lasius neglectus* in groups of three worker ants. The workers for the experiment were obtained from their laboratory stock colony, which had been collected from the field in 2022 in the Botanical Garden Jena, Germany. Ant collection and all experimental work were performed in compliance with international, national and institutional regulations and ethical guidelines. For the experiment, the body surface of one of the three ants was treated with a suspension of either of two microparticle types (diameter 5  $\mu\text{m}$ ) to induce grooming by the two nestmates, which were individually color-coded by application of a dot of blue or orange paint, respectively. The three ants were housed in small plastic containers (diameter 28mm, height 30mm) with moistened, plastered ground and the interior walls covered with PTFE (polytetrafluoroethane) to hamper climbing by the ants. Filming occurred in a temperature- and humidity-controlled room at 23°C within a custom-made filming box with controlled lighting and ventilation conditions. We set up nine ant groups at a time (always containing both treatments) and placed them randomly on positions 1-9 marked on the floor in a 3x3 grid with a distance of about 3mm from each other. Figure 7 illustrates the filming box and the displaying of the containers in each batch. The experiment was performed on two consecutive days. Videos were acquired using a USB camera (FLIR blackfly S BFS-U3-120S4C, Teledyne FLIR) with a high-performance lens (HP Series 25mm Focal Length, Edmund optics 86-572) in OBS studio 29.0.0 [Bailey, 2017] at a framerate of 30 FPS and a resolution of 2500x2500 pixels. From each original video (105x105 mm), we generated 9 individual videos



.mkv (each 32x32 mm, 770x770 pixels) by determining exact coordinates per container from one frame in GIMP 2.10.36 [Kimball and Mattis, 2023] and cropping of the videos with FFmpeg 6.1.1 [Tomar, 2006]. Annotation was performed over two consecutive days by three observers who had not been involved in the experimental setup or recording and were unaware of the treatment assignments to ensure bias-free behavioral annotation. They annotated the behavior of the ants during video observations, using custom-made software that saves the start and end frames of behaviors marked in a .csv file. In one of the videos, one of the nestmates’ legs got inadvertently stuck to its body surface during the color-coding, interfering with its behavior, so the video was discarded. This left 44 videos from 5 independent setups (n=24 of treatment 1 and n=20 of treatment 2) of 10 minutes each for a total of 792 000 annotated frames. For each video, we provide the following information: the number of the set to which it belongs (1-5); the number of the position within the set reflecting the position of the ant group under the camera (1-9), for which we also provide ‘coordinates’ in the 3x3 grid (taking values -1/0/1 for both X and Y axis); treatment (1 or 2); the hour of the day when the recording was started (in 24h CEST); experimental day (A or B); the top left coordinate of the cropping square from the original video (CropX/CropY); the person annotating the video (given as A, B, C); the date of annotation (1: first day, 2: second day) and in which order the videos were annotated by each person (both reflecting a possible training effect of the person).

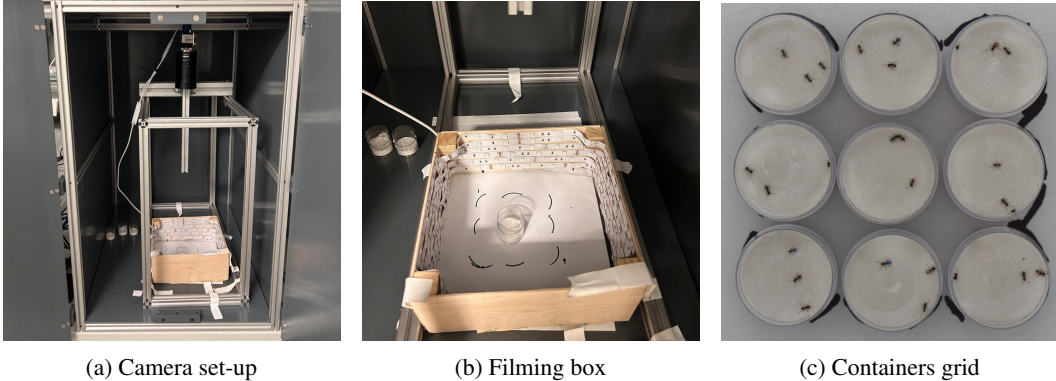


Figure 7: Visualizations of the filming box set-up inside a temperature- and humidity-controlled room.

## D Detailed Experimental Settings

In this section, we provide additional information on the experimental settings for the main experiments (on IStAnT dataset). In particular, we describe the annotation splitting criteria selected, then the modeling choices and the training details. We run all the analyses using 48GB of RAM, 20 CPU cores, and a single node GPU (NVIDIA GeForce RTX2080Ti). The main bottleneck in the analysis is the feature extraction from the pre-trained Vision Transformers. We estimate 96 GPU hours to fully reproduce all the experiments described in the main paper.

### D.1 Splitting Criteria

Let  $W_1 \in \{1, \dots, 5\}$  representing the number of batch experiment and  $W_2 \in \{1, \dots, 9\}$  the relative position of a video inside its batch. We defined the annotation splitting criteria based on the value of the experiment settings  $W_1$  and  $W_2$ , in agreement with Table 3.

where  $\Omega = \{(1, 2), (1, 3), (2, 4), (2, 5), (3, 1), (3, 2), (4, 3), (4, 4), (5, 9)\}$ . For validation (used to generate the Figure 6) we consider 1 000 random frames from  $\mathcal{D}_u$ .

### D.2 Additional Models details

We extracted once the embedding of each frame in the dataset using a pre-trained encoder, and we fine-tuned multi-layer perceptron (MLP) heads for classification according to the training details reported in Table 4. We considered the following encoders for feature extraction, also report the corresponding Hugging Face code ID for reference:

Annotations	Criteria	$\mathcal{D}_s$	$n_s$	$\mathcal{D}_u$	$n_u$
Many	Random	$(W_1, W_2) \notin \Omega$	42 000	$(W_1, W_2) \in \Omega$	10 800
	Experiment	$W_1 \neq 5$	42 000	$W_1 = 5$	10 800
	Position	$W_2 \neq 8$	46 800	$W_2 = 8$	6 000
Few	Random	$(W_1, W_2) \in \Omega$	10 800	$(W_1, W_2) \notin \Omega$	42 000
	Experiment	$W_1 = 1$	10 800	$W_1 \neq 1$	42 000
	Position	$W_2 = 1$	6 000	$W_2 \neq 1$	46 800

Table 3: Annotation splitting criteria details for the extensive experiments on ISTAnt described in Section 5 and 6.

- ViT-B [Dosovitskiy et al., 2020]: `google/vit-base-patch16-224`
- ViT-L [Zhai et al., 2023]: `google/siglip-base-patch16-512`
- CLIP-ViT-B [Radford et al., 2021]: `openai/clip-vit-base-patch32`
- CLIP-ViT-L [Radford et al., 2021]: `openai/clip-vit-large-patch14-336`
- MAE [He et al., 2022],: `facebook/vit-mae-large`
- DINOv2 [Oquab et al., 2023]: `facebook/dinov2-base`

Model/Hyper-parameters	Value(s)
Encoders	[CLIP-ViT-L, CLIP-ViT-S, DINOv2, MAE, ViT-L, ViT-S]
Encoder (token)	[class, mean, all]
MLP (head): hidden layers	[1,2]
MLP (head): hidden nodes	256
MLP (head): activation function	ReLU + Sigmoid output
Tasks	[all, or]
Dropout	No
Regularization	No
Loss	BCELoss (with positive weighting)
Loss: Positive Weight	$\frac{\sum_{i=1}^{n_s} 1 - Y_i}{\sum_{i=1}^{n_s} Y_i}$
Learning Rates	[0.05, 0.005, 0.0005]
Optimizer	Adam ( $\beta_1 = 0.9, \beta_2 = 0.9, \epsilon = 10^{-8}$ )
Batch Size	256
Epochs	10
Seeds	[0,1,2,3,4]

Table 4: Model and training details for the extensive experiments on ISTAnt described in Section 5 and 6.

Encoder (token) refers to which embedded tokens were considered for representation from each ViT. ‘class’ stands for the class taken, ‘mean’ for the mean of all the other tokens and ‘all’ for their concatenation. Task refers to which outcome we aimed to model directly: either the two independent grooming events (‘Blue to Focal’ and ‘Orange to Focal’) or the single grooming event (‘Blue or Orange to Focal’). Overall, we finetuned:

$$\begin{aligned}
 n &= n_{\text{splitting criteria}} \cdot n_{\text{encoders}} \cdot n_{\text{tokens}} \cdot n_{\text{tasks}} \cdot n_{\text{hidden layers}} \cdot n_{\text{learning rates}} \cdot n_{\text{seeds}} \\
 &= 6 \cdot 6 \cdot 3 \cdot 2 \cdot 2 \cdot 3 \cdot 5 = 6480
 \end{aligned}
 \tag{38}$$

heads.

## E CausalMNIST

### E.1 Data generating process

To replicate the results on IStAnt controlling for the causal model, we proposed CausalMNIST: a colored manipulated version of MNIST [LeCun, 1998], defining a simple causal downstream task (treatment effect estimation). Starting from MNIST dataset, we manipulated the background color  $B$  of each image (1: green, 0: red), and the pen color  $P$  (1: white, 0: black) to enforce the following Conditional Average Treatment Effect:

$$\mathbb{E}[Y|do(B = 1), P = 1] - \mathbb{E}[Y|do(B = 0), P = 1] = 0.4 \quad (39)$$

$$\mathbb{E}[Y|do(B = 1), P = 0] - \mathbb{E}[Y|do(B = 0), P = 0] = 0.2 \quad (40)$$

and Average Treatment Effect:

$$\mathbb{E}[Y|do(B = 1)] - \mathbb{E}[Y|do(B = 0)] = 0.3 \quad (41)$$

where  $Y$  is a binary variable equal to 1 if the represented digit is strictly greater than  $d \in \mathbb{R}$ , 0 otherwise. Arjovsky et al. [2019] already proposed a colored variant of MNIST as a benchmark for robustness in a multi-environment setting, but without controlling for any causal model and presenting a causal downstream task. A simple interpretation of this new task is estimating the effect of the background on the chances of writing a big digit (i.e., greater than  $d$ ).

To obtain a sample from such a population manipulating MNIST dataset, we converted each gray image into a RGB, coloring the background  $B$  and the pen  $P$  according to Bayes' rule:

$$P(B = b, P = p|Y = y) = \frac{P(Y = y|B = b, P = p) \cdot P(B = b, P = p)}{P(Y = y)} \quad \forall b, p, y \in \{0, 1\} \quad (42)$$

Since the digits in MNIST dataset are uniformly distributed:

$$Y \sim Be(p_Y) \quad (43)$$

where  $p_Y = (9 - d)/10$ .

We then set:

$$B, P \stackrel{\text{i.i.d.}}{\sim} Be(0.5) \quad (44)$$

and:

$$P(Y = 1|B = 1, P = 1) = p_Y + 0.2 \quad (45)$$

$$P(Y = 1|B = 0, P = 1) = p_Y - 0.2 \quad (46)$$

$$P(Y = 1|B = 1, P = 0) = p_Y + 0.1 \quad (47)$$

$$P(Y = 1|B = 0, P = 0) = p_Y - 0.1 \quad (48)$$

in agreement with the Law of Total Probability and assuming  $d \in \{1, 2, \dots, 7\}$ .

Overall, the final structural causal model can be summarized as follows:

- Noises (independent):

$$n_B \sim Be(0.5) \quad (49)$$

$$n_P \sim Be(0.5) \quad (50)$$

$$n_X \sim P^{n_X} \quad (51)$$

$$n_Y \sim P^{n_Y} \quad (52)$$

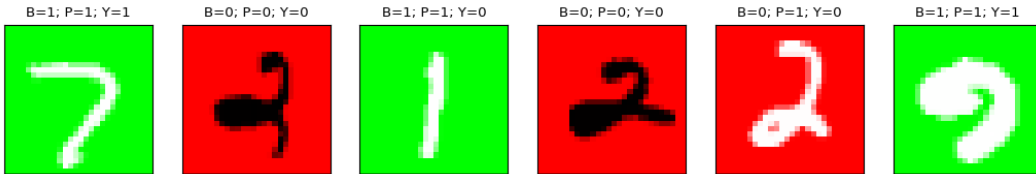


Figure 8: Random samples from CausalMNIST dataset.

- Structural equations:

$$B := n_B \tag{53}$$

$$P := n_P \tag{54}$$

$$\mathbf{X} := f_1(B, P, n_X) \tag{55}$$

$$Y := f_2(\mathbf{X}, n_Y) \tag{56}$$

where  $P^{n_X}, P^{n_Y}, f_1$  and  $f_2$  are unknown and characteristic of MNIST dataset. The corresponding causal model matches the setting described in Section 2 where  $B$  represents the treatment  $T$  and  $P$  the experiment settings  $\mathbf{W}$ . In this analysis, we set  $d = 3$ . Six examples of colored handwritten digits from CausalMNIST are reported in Figure 8.

## E.2 Experimental Setting

**Annotation Sampling** Similarly to ISTAnt experiments, we compare the random annotation, where  $S$  is assigned independently from  $P$ , and a biased annotation, where only the images with a black pen ( $P = 0$ ) are annotated in both few and many annotation setting. The biased annotation criteria don't provide any information in  $\mathcal{D}_s$  about the white pen ( $P = 1$ ) CATE, and retrieving the annotations on  $\mathcal{D}_u$  becomes mandatory. Unfortunately, a model could misclassify the new images under this covariate shift or hallucinate just for a specific treatment group (e.g., green background and white pen), leading to a biased estimate of the ATE. In Table 5, we summarize the 4 different annotation sampling proposed. For validation, we consider a random subsample of  $\mathcal{D}_u$  as large as  $\mathcal{D}_s$ . Please observe that for the biased subsampling, not all the images with black pen ( $P=0$ ) are allocated  $\mathcal{D}_s$ . Indeed, since  $\mathbb{P}(P = 0) = 0.5 > \frac{n_s}{n_s+n_u}$  (in both few and many annotations regime), then  $\mathcal{D}_u$  contains both images of hand-written digits in white and black.

Annotations	Criteria	$\mathcal{D}_s$	$n_s$	$\mathcal{D}_u$	$n_u$
Many	Random	random	12 000	random	48 000
	Biased	only black (P=0)	12 000	the remaining (mixed)	48 000
Few	Random	random	1 800	random	58 200
	Biased	only black (P=0)	1 800	the remaining (mixed)	58 200

Table 5: Annotation splitting criteria details for CausalMNIST experiments.

**Modeling** Since the vision task is relatively simple, i.e., extracting features from a pre-trained VisionTransformer is unnecessary, we don't replicate the comparison among different backbones, but we directly model the outcome through a simple Convolutional Neural Network. On the other hand, since we have control over the data-generating process, we generated CausalMNIST 100 times for each annotation sampling criteria using different random seeds, and we trained a Convolutional Neural Network (ConvNet) for each of them (i.e., Monte Carlo simulations). This way, comparing the different models, we still replicated the results for (i) data bias, (ii) discretization bias, and (iii) evaluation metrics already obtained for ISTAnt. The proposed ConvNet architecture consists of two convolutional layers followed by two fully connected layers. The first convolutional layer applies 20 filters of size 5x5 with ReLU activation, followed by a 2x2 max-pooling layer. The second convolutional layer applies 50 filters of size 5x5 with ReLU activation, followed by another 2x2 max-pooling layer. The output feature maps are flattened and passed to a fully connected layer with 500 neurons and ReLU activation. The final fully connected layer reduces the output to a single logit for binary classification (mapped to a probability through the sigmoid function). Table 6 reports a full description of the training details for such a ConvNet.

**Evaluation** We collected the same evaluation metrics for each training on both validation and the full dataset as described in Section 5.

Hyper-parameters	Value(s)
Pre-Processing	Normalization
Dropout	No
Regularization	No
Loss	BCELoss
Loss: Positive Weight	No
Learning Rates	0.001
Optimizer	Adam ( $\beta_1 = 0.9, \beta_2 = 0.9, \epsilon = 10^{-8}$ )
Batch Size	64
Epochs	6
Seeds	{0,1, ..., 99}

Table 6: Training details for the ConvNets training on CausalMNIST.

### E.3 Results

We run all the analysis using 10GB of RAM, 8 CPU cores, and a single node GPU (NVIDIA GeForce RTX2080Ti). The main bottleneck of each experiment is re-generating a new version of CausalMNIST from MNIST dataset. We estimate a total of 6 GPU hours to reproduce all the experiments described in this section.

**Annotation criteria matter** Theory suggests that biased annotating criteria (i.e., depending on the experimental settings) can lead to biased treatment effect estimation, wrongly retrieving the conditional treatment effect on unseen experimental settings. Figure 9 validates this observation, and the results are validated via the  $t$ -tests reported in Table 7. Overall, the results perfectly align with the analogous discussion on IStAnt.

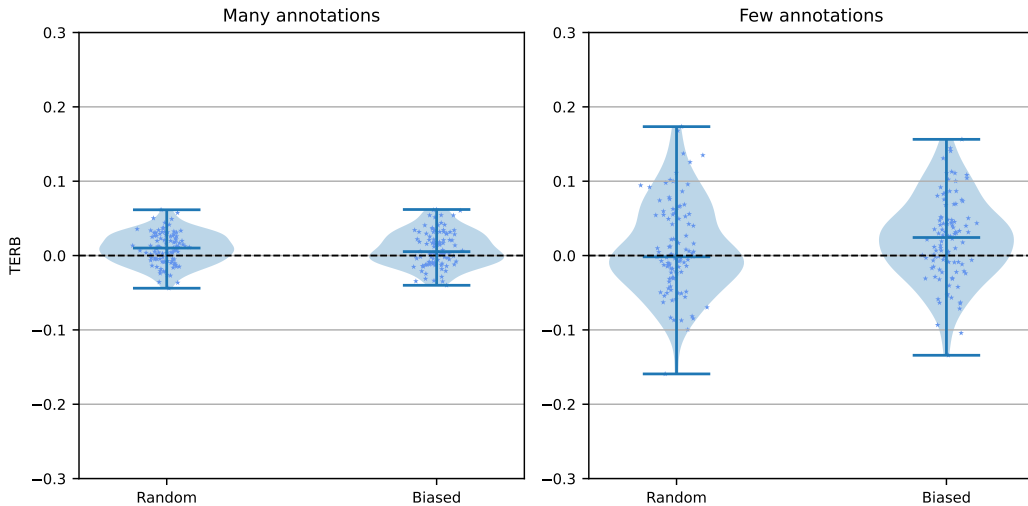


Figure 9: Violin plots comparing the Treatment Effect Relative Bias (TERB) per annotation criteria in few and many annotations regime varying the seeds. Biased annotations lead to biased ATE estimation (i.e.,  $TERB \neq 0$ ) and random annotation should be preferred.

**Discretization bias** We considered the absolute value of the TEB over all the 400 experiments, and we tested ( $t$ -test):

$$\mathcal{H}_0 : \mathbb{E}[|\text{TEB}(f)|] = \mathbb{E}[|\text{TEB}(\mathbf{1}_{[0.5,1]}(f))|] \quad vs \quad \mathcal{H}_1 : \mathbb{E}[|\text{TEB}(f)|] \neq \mathbb{E}[|\text{TEB}(\mathbf{1}_{[0.5,1]}(f))|] \quad (57)$$

Annot.	Criteria	$t$	$p$ -value
Many	Random	4.421	$2.5 \cdot 10^{-5}$
	biased	4.030	$1.1 \cdot 10^{-4}$
Few	Random	1.607	0.111
	Biased	3.911	$1.7 \cdot 10^{-4}$

Table 7: Two-sided  $t$ -test for  $\mathcal{H}_0 : \mathbb{E}[\text{TEB}(f)] = 0$ . We found statistical evidence to reject the hypothesis that  $f$  is unbiased for (almost) each annotation criterion.

We found no statistical evidence to reject the null hypothesis ( $t$  statistic=1.188,  $p$ -value=0.235). Still, this result doesn't contradict Theorem 3.1, where we show that predictions, discretized and not, generally differ in expectation, but they can still be close (by chance). Some evidence of this undesired discretization effect can still be observed in the distribution of the  $\text{TEB}(f)$  and  $\text{TEB}(\mathbb{1}_{[0.5,1]}(f))$  as illustrated in Figure 10 for both random and biased sampling. In random sampling, in particular  $\text{TEB}(\mathbb{1}_{[0.5,1]}(f))$  mean in random sampling is positive and 72.5% higher than  $\text{TEB}(f)$  mean.

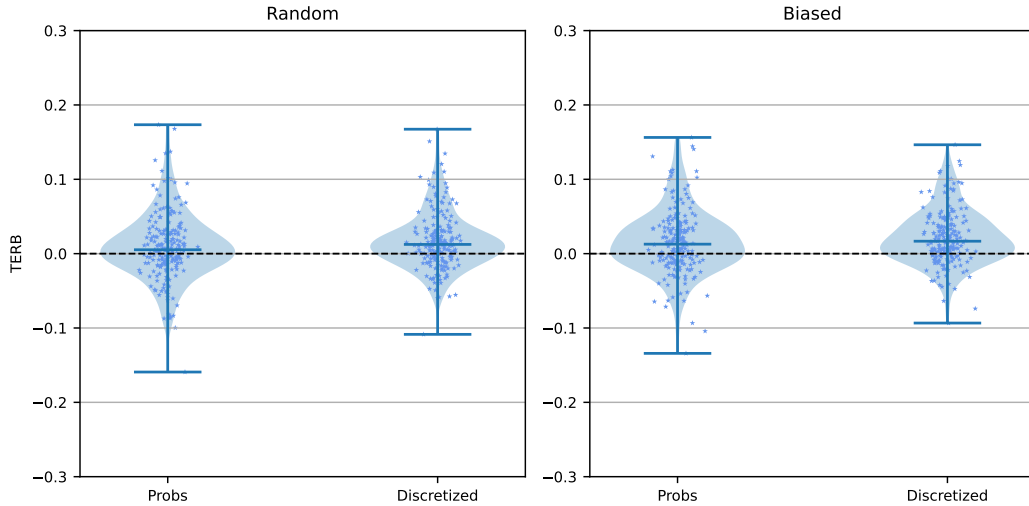


Figure 10: Violin plots of the TERB of the model (discretized or not) for both random and biased annotation sampling, varying number of annotations (few/many) and seeds.

**Prediction is not Causal Estimation** Distinct statistical and causal objectives cannot be used as a proxy for one another. We already formalized this in Lemma 3.1 and discussed it for IStAnt dataset. In Figure 11 and 12, we systematically show it again for our new synthetic benchmark by comparing the rank-correlation among 200 ConvNets using random and biased sampling, respectively. Both matrices fully align with the discussion presented for the IStAnt dataset in Section 6.

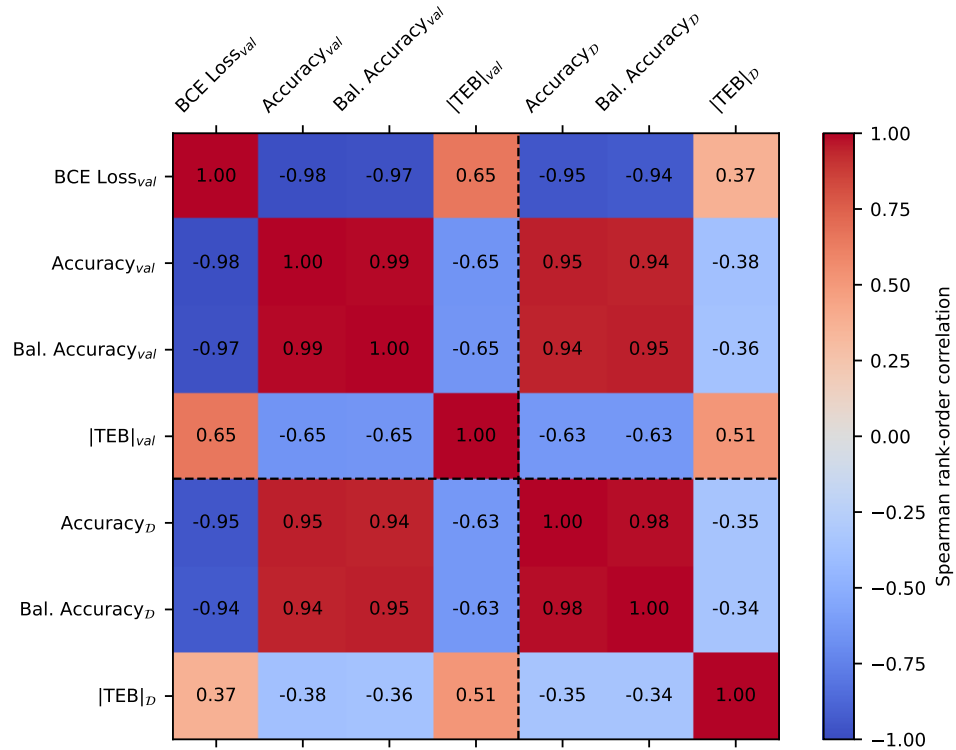


Figure 11: Spearman rank-order correlation matrix comparing different metrics for model selection on validation (subscript  $val$ ) and over the full dataset (subscript  $\mathcal{D}$ ). We considered all the 200 models trained with **random sampling**, varying the number of annotations (few and many) and seeds. Standard prediction metrics on validation strongly correlate, but they are less associated with  $|TEB|_{val}$ . Similarly, they correlate with the prediction metrics on the full dataset but poorly predict the  $|TEB|_{\mathcal{D}}$ . On the other hand,  $|TEB|_{val}$  is the most correlated metric with  $|TEB|_{\mathcal{D}}$ , unlike even the prediction metrics on the full dataset.

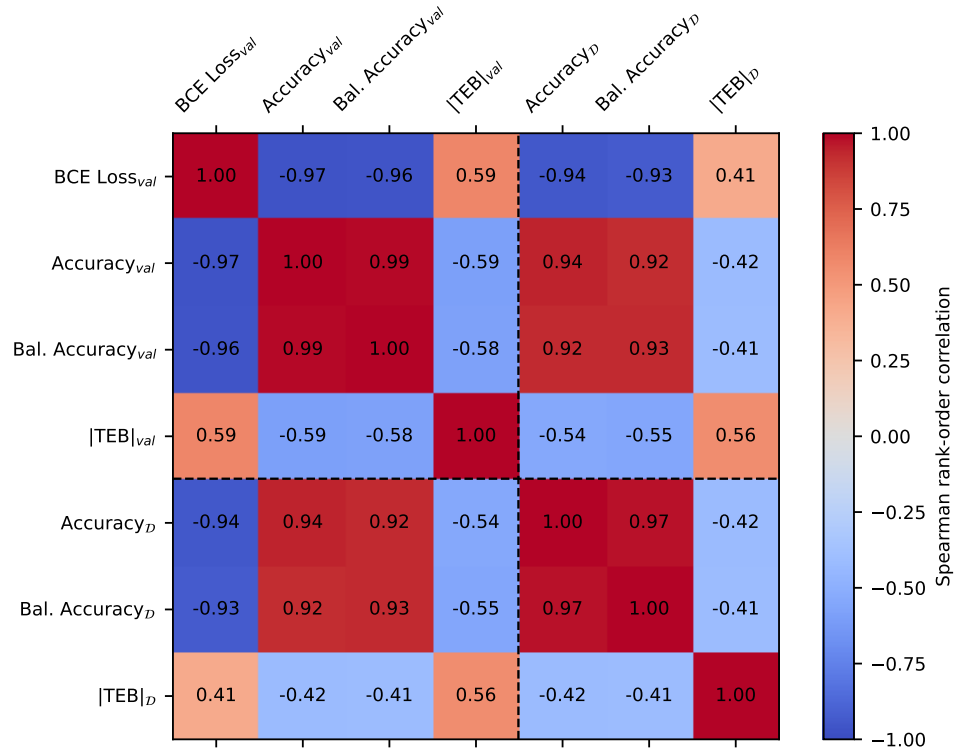


Figure 12: Spearman rank-order correlation matrix comparing different metrics for model selection on validation (subscript  $val$ ) and over the full dataset (subscript  $\mathcal{D}$ ). We considered all the 200 models trained with **biased sampling**, varying the number of annotations (few and many) and seeds. Standard prediction metrics on validation strongly correlate, but they are less associated with  $|TEB|_{val}$ . Similarly, they correlate with the prediction metrics on the full dataset but poorly predict the  $|TEB|_{\mathcal{D}}$ . On the other hand,  $|TEB|_{val}$  is the most correlated metric with  $|TEB|_{\mathcal{D}}$ , unlike even the prediction metrics on the full dataset.



## NeurIPS Paper Checklist

### 1. Claims

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

Answer: [Yes]

Justification: In the abstract, we claim that we theoretically discuss potential sources of biases in downstream treatment effect estimation from Machine Learning pipelines in randomized controlled trials. In the paper, we discuss it in Section 3 (proofs in Appendix A). In the abstract, we claim that we validate our results experimentally, introducing a new real-world dataset and a new synthetic one. In the paper, we introduce the new benchmarks in Section 2, Appendix E and C, and we report the results in Section 6 and Appendix E.

Guidelines:

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

### 2. Limitations

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

Answer: [Yes]

Justification: Limitations are discussed in Section 7.

Guidelines:

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

### 3. Theory Assumptions and Proofs

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

Answer: [Yes]

Justification: Theoretical results are introduced in Section 3 and the corresponding proofs are reported in Appendix A.

Guidelines:

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

### 4. Experimental Result Reproducibility

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

Answer: [Yes]

Justification: Full description of the data generating process, experiment settings and analysis are reported in Section 5 and 6 and Appendix E, C, and D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

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

Answer: [Yes]

Justification: We provide open access to the code on GitHub (<https://github.com/CausalLearningAI/ISTAnt>) and the data on FigShare (<https://doi.org/10.6084/m9.figshare.26484934.v2>).

Guidelines:

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

## 6. Experimental Setting/Details

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

Answer: [Yes]

Justification: We specify all the experimental settings for the real-world experiment in Appendix D, and the settings for the synthetic experiments in Appendix E.

Guidelines:

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

## 7. Experiment Statistical Significance

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

Answer: [Yes]

Justification: We support all our results (both synthetic and real-world) with confidence intervals and statistical tests (see Example 1, Section 6 and Section E) specifying the factor of variabilities.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments Compute Resources

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

Answer: [Yes]

Justification: Computer resource requirements are discussed in Appendix D and E.

Guidelines:

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

#### 9. Code Of Ethics

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

Answer: [Yes]

Justification: Data collection and all experimental work were performed in compliance with international, national and institutional regulations and ethical guidelines.

Guidelines:

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

#### 10. Broader Impacts

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

Answer: [No]

Justification: We believe this work has no specific positive or negative societal impacts except to potentially accelerate science on an important topic for biodiversity (which is discussed in the paper).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

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

Answer: [NA]

Justification: Both data and analysis don't poses such risks.

Guidelines:

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

#### 12. Licenses for existing assets

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

Answer: [Yes]

Justification: We use our proprietary assets and publicly available public assets (pre-trained Hugging Face models and MNIST dataset), which we correctly acknowledge and cite (see Appendix E and D).

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. New Assets

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

Answer: [Yes]

Justification: The main new asset is the real-world dataset, which we release and describe at <https://doi.org/10.6084/m9.figshare.26484934.v2>. A more extended description of the dataset is also included in Appendix C. We also submit all the code on GitHub at <https://github.com/CausalLearningAI/ISTant> and in the supplementary material to reproduce the results fully. The supplementary materials consist of 3 folders, `ISTant` containing the code for the analysis on the real-word dataset, `CausalMNIIST` for the analysis of the synthetic dataset and `Example 1` to reproduce the simulations in Example 1.

Guidelines:

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

### 14. Crowdsourcing and Research with Human Subjects

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

Answer: [NA]

Justification: We do not work with human subjects or crowdsourcing.

Guidelines:

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

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

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

Answer: [NA]

Justification: We do not work with human subjects or crowdsourcing.

Guidelines:

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