CAUSAL LIFTING OF NEURAL REPRESENTATIONS: ZERO-SHOT GENERALIZATION FOR CAUSAL INFERENCES

Riccardo Cadei
ISTA
riccardo.cadei@ist.ac.at

Ilker Demirel*
MIT

Piersilvio De Bartolomeis*
ETH

Lukas Lindorfer
ISTA

Sylvia Cremer
ISTA

Cordelia Schmid
INRIA

Francesco Locatello
ISTA

ABSTRACT

A plethora of real-world scientific investigations is waiting to scale with the support of trustworthy predictive models that can reduce the need for costly data annotations. We focus on causal inferences on a target experiment with unlabeled factual outcomes, retrieved by a predictive model fine-tuned on a labeled *similar* experiment. First, we show that factual outcome estimation via Empirical Risk Minimization (ERM) may fail to yield valid causal inferences on the target population, even in a randomized controlled experiment and infinite training samples. Then, we propose to leverage the observed experimental settings during training to empower generalization to downstream interventional investigations, "*Causal Lifting*" the predictive model. We propose *Deconfounded Empirical Risk Minimization* (DERM), a new simple learning procedure minimizing the risk over a fictitious target population, preventing potential confounding effects. We validate our method on both synthetic and real-world scientific data. Notably, for the first time, we zero-shot generalize causal inferences on ISTAnt dataset (without annotation) by causal lifting a predictive model on our experiment variant.

1 INTRODUCTION

Artificial Intelligence (AI) systems hold great promise for accelerating scientific discovery by providing flexible models capable of automating complex tasks. We already depend on deep learning predictions across various applications, including biology (Jumper et al., 2021; Tunyasuvunakool et al., 2021; Elmarakeby et al., 2021; Mullowney et al., 2023), sustainability (Castello et al., 2021), and the social sciences (Jerzak et al., 2022; Daoud et al., 2023).

While these models offer transformative potential for scientific research, their black-box nature poses new challenges. They can perpetuate hidden biases, which are difficult to detect and quantify, and risk invalidating conclusions drawn from their predictions for downstream experiments. Recent efforts have focused on combining capable black-box models with partially annotated data to power valid and efficient statistical inference (Angelopoulos et al., 2023a;b). Drawing inspiration from there, we focus on enabling *causal inference* on unlabeled experimental data via *factual* predictions, developing methods that can leverage powerful AI models reliably in that endeavor, i.e., Prediction-Powered Causal Inference (PPCI). A key challenge in this setting is that small modeling biases can invalidate the causal conclusions, even in the simplest possible scenario, where the downstream experiment is a randomized controlled trial (Cadei et al., 2024). Secondly, we aim to retrieve the annotations even out-of-distribution, allowing for zero-shot generalization. Yet, manual annotation of scientific experiments is costly, requiring experts to identify subtle signals, e.g., analyzing hours of videos to detect behavioral markers in experimental ecology. Automating the annotation process with machine learning models without any further training can alleviate this burden completely, tremendously accelerating the full pipeline.

At the same time, in scientific applications, experimentalists often collect data through multiple experiments with similar designs, e.g., investigating the effect on the same outcome of interest

*Equal contribution.

under different treatment or environmental settings. While historical experiments may yield too little data to train a performant model from scratch, one can fine-tune a pre-trained foundational model to learn the patterns needed for annotating experiments. Despite being a promising direction, a critical hurdle to generalize across experiments without introducing bias remains. Fine-tuning foundational models is typically done via Empirical Risk Minimization (ERM), which tends to exploit any *statistical association* in the training data to minimize prediction error. Therefore, one risks leveraging spurious associations between experiment-specific factors (e.g., equipment artifacts) and outcomes, leading to systematic prediction errors on the target experiment. To address the problem, we propose "causal lifting" such foundation models from potential confounding effects, suppressing the application-specific spurious correlations during fine-tuning.

We first discuss the challenges and feasibility of the problem, and in agreement with Yao et al. (2024a) we show how the supervised objective has to be paired with a conditional independence constraint enforcing the model to not rely on spurious correlations in its class of experiments. We then propose a simple and tailored implementation for such constraint via a resampling approach, reweighting the samples in the empirical risk, i.e., Deconfounded Empirical Risk Minimization (DERM). We validate the full pipeline for Causal Lifting on both synthetic and real-world data. Notably, we leverage a new experiment (ours) *similar* to ISTAnt (Cadei et al., 2024) yet differing in several experimental and technical details, including lower-quality light conditions and diverse treatments. For the first time, our method enables a foundational model to retrieve valid Causal Inference on ISTAnt dataset without annotation, i.e., 0-shot generalization of causal inferences on a completely unlabelled experiment.

In broader terms, this paper emphasizes the "representation learning" aspect of "causal representation learning", which has traditionally focused on identification. In Bengio et al. (2013), good representations are defined as ones "*that make it easier to extract useful information when building classifiers or other predictors*." In a similar spirit, we focus on representations that make extracting causal information easier or at all possible with some downstream estimator. As we shall demonstrate, guaranteeing identification of the causal effect is not always possible depending on the distributional differences between the experiments and our modeling choices. Yet, we hope that our viewpoint can also offer benchmarking opportunities that are currently missing in the causal representation learning literature (Schölkopf et al., 2021) and have great potential, especially in the context of scientific discoveries.

Overall, our contributions are:

i. a **new problem** formulation, i.e., PPCI, reshaping the definition of Causal Representation Learning as Representation Learning for Causal Downstream Tasks beyond untestable identifiability results and enabling quantitative benchmarking,

ii. a **new method**, i.e., DERM, for *Causal Lifting* of foundational models unconfounding their representations from spurious correlations between the perceived experiment settings and the outcome of interest,

iii. **first** valid and efficient **0-shot generalization** for PPCI on ISTAnt, *Causal Lifting* DINOv2 on our lower-quality experiment, which **data**, i.e., recordings and annotations, we plan to **release** publicly upon acceptance (preview: https://figshare.com/s/9a490b6f6eeebd73350b).

2 Problem Formulation

Let $\mathcal{E}$ a countable index set, and consider a class of Structural Causal Models (SCM) $\mathfrak{S} := \{\mathcal{M}^e\}_{e\in\mathcal{E}}$, characterized by the following (universal) Structural Equations:

$$\begin{cases} \boldsymbol{Z} := n_{\boldsymbol{Z}} \\ Y := f_Y(\boldsymbol{Z}, n_Y) \\ \boldsymbol{X} := f_{\boldsymbol{X}}(\boldsymbol{Z}, Y, n_{\boldsymbol{X}}) \end{cases} \tag{2.1}$$

and varying the exogenous variables distribution[1]:

$$n_{\mathbf{Z}}, n_Y, n_{\mathbf{X}} \sim \mathbb{P}^e. \tag{2.2}$$

[1]Note that $\boldsymbol{Z}, Y$ and $\boldsymbol{X}$ distributions all depend on the environment e, but we omit the reference for simplicity of language by always explicit the considered distribution.

We further assume that there exists a model g^* retrieving Y from $\boldsymbol{X}$ almost surely for the whole class, i.e.,

$$\exists g^* : \mathbb{P}^e(Y = g^*(\boldsymbol{X})) = 1 \quad \forall e \in \mathcal{E}. \tag{2.3}$$

Many variants of real-world experiments can be modeled via such a class of SCM, where:

- $\boldsymbol{Z}$ is the universal set of (possible) experimental settings potentially affecting the outcome, i.e., all the ancestors of $\boldsymbol{X}$ and Y (excluding noise),
- Y is the outcome of interest,
- $\boldsymbol{X}$ is a fully informative high-dimensional observation of the experiment.

This framework is particularly suitable for Causal Inference applications[2], where the outcome of interest is commonly not observed directly but extracted from a high-dimensional observation, and some experiment settings are naturally collected and potentially controlled. In the following, we use $\boldsymbol{Z}^e$ to refer to the experiment settings actually observed in experiment $\mathcal{M}^e$, and $\boldsymbol{U}^e$ for the unobserved. We can further distinguish, within $\boldsymbol{Z}^e$, between a treatment variable T^e and observed pre-treatment variables $\boldsymbol{W}^e$. All together:

$$\boldsymbol{Z} = \underbrace{T^e \cup \boldsymbol{W}^e}_{\boldsymbol{Z}^e} \cup \boldsymbol{U}^e \quad \forall e \in \mathcal{E}. \tag{2.4}$$

Note that which variables $\boldsymbol{Z}^e \subseteq \boldsymbol{Z}$ are observed may change across experiments, in particular, the treatment of interest T^e and the observed pre-treatment variables, $\boldsymbol{W}^e$.

When a new experiment is performed, we collect observations $\boldsymbol{X}$ (and experimental conditions $T^e, \boldsymbol{W}^e$), from which the outcome Y can be extracted. Instead of annotating Y by hand for every new experiment, we wonder when we could leverage similar experiments, i.e., in the same class, to train or fine-tune a machine-learning system capable of supplying accurate predictions about the outcome of interest and obtain trustworthy confidence interval on a causal downstream task. We refer to this problem as (*factual*) Prediction-Powered Causal Inference (PPCI). In summary:

Prediction-Powered Causal Inference

Sources:

- A random sample $\mathcal{D}^{e_1} = \{(T_i^{e_1}, \boldsymbol{W}_i^{e_1}, Y_i, \boldsymbol{X}_i)\}_{i=1}^{n^{e_1}}$ from a reference experiment $\mathcal{M}^{e_1} \in \mathfrak{S}$,
- A random sample $\mathcal{D}^{e_2} = \{(T_i^{e_2}, \boldsymbol{W}_i^{e_2}, _, \boldsymbol{X}_i)\}_{i=1}^{n^{e_2}}$ from a target experiment $\mathcal{M}^{e_2} \in \mathfrak{S}$, not observing the factual outcome of interest[a].

Assumption: Existence of an invariant factual outcome model from the raw observations $\boldsymbol{X}$, i.e.,

$$\exists g^* : \mathbb{P}^e(Y = g^*(\boldsymbol{X})) = 1 \quad \forall e \in \{e_1, e_2\}. \tag{2.5}$$

Task: Learn a factual outcome model estimator $\hat{g}$ conditionally unbiased on the target population, i.e,

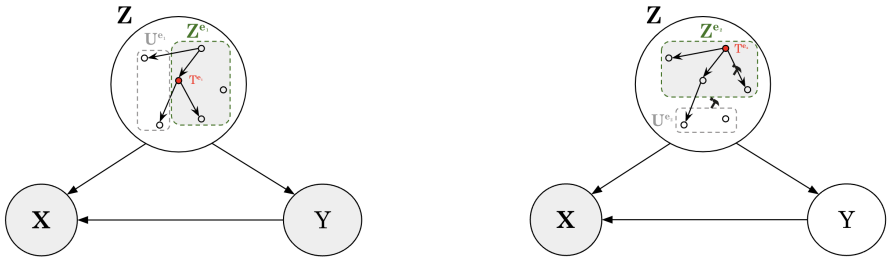
$$\mathbb{E}_{\mathbb{P}_{e_2}}[Y - \hat{g}(\boldsymbol{X}) | \boldsymbol{Z}] \overset{\text{a.s.}}{=} 0, \tag{2.6}$$

enabling different downstream causal inferences, e.g., Average Treatment Effect (ATE) estimation under conditional exchangeability.

[a]The observed experiment settings are not necessarily shared between experiments.

Figure 1 illustrates the reference and target experiment using their causal models. Condition 2.6 effectively means that the factual outcome estimator, $\hat{g}$, is unbiased under *any* experimental setting Z that can be observed in the *target* distribution, $\mathbb{P}_{e_2}$. Once we have a factual outcome estimator that satisfies Condition 2.6, we can use it to impute the missing outcome and then run the follow-up Causal Inference analysis on the prediction-powered target sample, e.g., ATE estimation via AIPW estimator (Robins et al., 1994; Robins and Rotnitzky, 1995). As Theorem 2.1 formalizes, it ensures (asymptotically) valid confidence intervals—a key requirement for scientific research—on the ATE without any factual outcome observations (assuming the causal effect is identifiable).

[2]We focus here only on *causality in mean* (Pearl and Mackenzie, 2018), ignoring counterfactual reasoning.



(a) Reference Experiment ($\mathcal{M}^{e_1}$) (b) Target Experiment ($\mathcal{M}^{e_2}$)

Figure 1: Causal Model visualization of a Reference and Target Experiment from the same SCM class $\mathfrak{S}$. The observed variables are in light gray, and the unobserved in white.

> **Theorem 2.1** (Informal)**.** *Given a PPCI problem and a factual outcome model g conditionally unbiased on the target population, i.e., satisfying Eq. 2.6. Assume that the ATE would be identifiable in the target experiment with ground-truth labels of the effect. Then, the AIPW estimator over the prediction-powered target sample provides an asymptotically valid confidence interval for the ATE.*

See the formal proposition and proof in Appendix A. Analogous results hold for interventional causal inferences on continuous treatment and heterogenous effect estimation, i.e., CATE estimation.

2.1 Zero-Shot Generalization

There are generally no guarantees for Condition 2.6 to hold while training $\hat{g}$ on the reference experiment. Indeed, due to interventions to the experimental settings $\boldsymbol{Z}$, and being $\boldsymbol{Z} \to \boldsymbol{X}$, also the high-dimensional observation $\boldsymbol{X}$ may shift out of support on target, leaving the factual outcome model not *identifiable* even in the infinite sample setting. Foundational models pre-trained on extended corpus offer a promising solution to the issue, practically enabling to consistently extract all the useful information hidden in $\boldsymbol{X}$ to predict Y (but not only). However, it is not sufficient, and it is crucial that the downstream classifier is also not overfitting on spurious correlations during fine-tuning. Indeed, even if we assume access to an oracle encoder $\phi^*(\boldsymbol{X}) = \begin{bmatrix} \phi_Y(\boldsymbol{X}) \\ \psi(\boldsymbol{X}) \end{bmatrix}$, extracting from $\boldsymbol{X}$ (among other features $\psi(\boldsymbol{X})$, e.g., related with the experimental settings):

- all the information of Y, i.e., $H_{\mathbb{P}^e}[Y|\phi_Y(\boldsymbol{X})] = 0$,
- disentangled from all the experiment setting $\boldsymbol{Z}$, i.e., $\mathrm{I}_{\mathbb{P}^e}(\phi_Y(\boldsymbol{X}), \boldsymbol{Z}|Y) = 0$,

for all possible experiments $e \in \mathcal{E}$; we may still have trouble in learning a factual outcome classifier h on top by Empirical Risk Minimization (ERM), since it could still perfectly minimize the empirical risk, but rely on spurious correlations between some experimental settings, e.g., retrievable from $\psi(\boldsymbol{X})$, and the outcome of interest. If a specific instance of observed experiment settings $\boldsymbol{z}$ is fully informative of the outcome on the reference population, e.g., $\mathrm{Var}(Y|\boldsymbol{Z}^{e_1} = \boldsymbol{z}) = 0$, while varying on target for distribution shifts of the unobserved ones, standole ERM has no criteria to privilege an invariant solution to one relying on the retrieved experiment setting spurious correlations. More generally, it is enough that the outcome support is not full on the reference experiment, conditioning on some experiment settings, that ERM may privilege a model overfitting on such spurious correlation, *stereotyping*. In real-world applications, similar issues are due to weak overlap between the conditional outcome on the experimental settings and outcome distribution, still allowing a candidate model to leverage spurious correlations, even if not perfectly solving the task.

3 Deconfounded Empirical Risk Minimization (DERM)

As motivated in Section 2, directly minimizing the empirical risk of an expressive head on top of a (pre-trained) encoder may not be sufficient in generalization for PPCI, even in the infinite sample

setting and relying on an oracle encoder. *What should we enforce then, to enable* (or at least attempt) *such desired causal generalization?*

The natural mitigation to prevent the described overfitting on spurious dependencies is to optimize for the model sufficiency while enforcing unconfoundness in the representation space, i.e.,

$$\begin{aligned} \min_{h,\phi} \quad & \mathbb{E}_{\mathbb{P}^{e_1}}\left[\mathcal{L}\left(Y, h \circ \phi(\boldsymbol{X})\right)\right] \\ \text{s.t.} \quad & \phi(\boldsymbol{X}) \perp\!\!\!\perp_{\mathbb{P}^{e_1}} \boldsymbol{Z} | Y = y \quad \forall y \in \mathcal{Y}. \end{aligned} \tag{3.1}$$

The conditional independence constraint enforces that the mutual information between the representation and the experimental settings, conditioning on the outcome of interest, is null, i.e.,

$$\mathrm{I}_{\mathbb{P}^{e_1}}\left(\phi(\boldsymbol{X}), \boldsymbol{Z} | Y\right) = 0, \tag{3.2}$$

such that it cannot be used to leverage some conditional dependencies $\mathbb{P}^{e_1}_{Y|\boldsymbol{Z}^{e_1}=\boldsymbol{z}}$ potentially breaking on the target experiment since only $\mathbb{P}^{e}_{Y|\boldsymbol{Z}=\boldsymbol{z}}$ is invariant. Note that such conditions can be enforced while learning $f = h \circ \phi$ from scratch or just fine-tuning a head h on top of a pre-trained (oracle) encoder ϕ^*. The former approach is exactly the problem described in Formulation 3.1, while in the latter case, the unconfoundness constraint, i.e., Eq. 3.2, has to be enforced on an internal representation of the head h, ideally isolating only the disentangled information of the outcome of interest (see ϕ_Y in Section 2 discussion).

Formulation 3.1 is a well-known problem in Representation Learning literature, and different approaches were developed, enforcing the unconfoundness constraint directly or indirectly. In Appendix B, we discuss a brief overview of the different paradigms. Among them we propose a resampling approach (Kirichenko et al., 2022; Li and Vasconcelos, 2019), carefully designing a fictitious auxiliary experiment $\mathcal{M}^{e_1^{\perp\!\!\!\perp}}$ to sample from during training. It is obtained by manipulating the original reference sample so that its spurious correlations between $\boldsymbol{Z}^{e_1}$ and Y are not observed, and confounding effects are prevented. In particular, we define such fictitious *unconfounded* population intervening on the joint distribution $\mathbb{P}^{e_1}_{\boldsymbol{Z}^{e_1},Y}$ and enforcing independence, i.e., $\boldsymbol{Z}^{e_1} \perp\!\!\!\perp Y$, blocking any not-causal path from $\boldsymbol{X}$ to Y during learning. Among the possible joint distribution guaranteeing independence, we define, for all $y \in \mathcal{Y}$ and $\boldsymbol{z} \in \mathcal{Z}^{e_1}$ in support of $\mathbb{P}^{e_1}_{\boldsymbol{Z}^{e_1},Y}$:

$$\mathbb{P}^{e_1^{\perp\!\!\!\perp}}\left(Y = y, \boldsymbol{Z}^{e_1} = \boldsymbol{z}\right) := \frac{\mathrm{Var}_{\mathbb{P}^{e_1}}\left(Y | \boldsymbol{Z}^{e_1} = \boldsymbol{z}\right)}{\sum_{\boldsymbol{z}' \in \mathcal{Z}^{e_1}} \mathrm{Var}_{\mathbb{P}^{e_1}}\left(Y | \boldsymbol{Z}^{e_1} = \boldsymbol{z}'\right)}, \tag{3.3}$$

weighting more the least informative experimental settings for the outcome of interest (high conditional variance) and ignoring the fully informative ones (low or null variance) over the reference population. Indeed, let's observe that the marginal outcome given the observed experimental settings is constant, i.e., uniform distribution. If, for each not fully informative observed experimental setting $\boldsymbol{z}$[3],

$$\{y \in \mathcal{Y} : \mathbb{P}^{e_1}(Y = y | \boldsymbol{Z}^{e_1} = \boldsymbol{z}) > 0\} = \{y \in \mathcal{Y} : \mathbb{P}^{e_1}(Y = y) > 0\}, \tag{3.4}$$

then the joint distribution described in Eq. 3.3 trivially implies independence, i.e., $\boldsymbol{Z}^{e_1} \perp\!\!\!\perp Y$, and an unconfounded representation is enforced. It is important to observe that in such implementation, spurious solutions may still optimize the risk, but such solutions cannot be selected and preferred since there is no signal in the task to retrieve the experimental settings.

Let $\mathcal{G}$ be an expressive factual outcome model hypothesis space (containing an invariant factual outcome model g^*). The ERM solution set over the reference sample $\mathcal{G}^{e_1}$ and the target sample $\mathcal{G}^{e_2}$ overlap, while the ERM solution set $\mathcal{G}^{e_1^{\perp\!\!\!\perp}}$ over an *unconfounded* fictitious sample from $\mathcal{M}^{e_1^{\perp\!\!\!\perp}}$ is included in such intersection and it still includes the invariant factual outcome model. In Figure 2, we illustrate the relations among these hypotheses and solution spaces.

We then propose to train/fine-tune the factual outcome model on such a fictitious population by ERM sampling from the reference population and reweighting the estimated joint distribution $\mathbb{P}^{e_1}_{\boldsymbol{Z}^{e_1},Y}$ to enforce the desired disentangled distribution described in Eq. 3.3. We refer to this approach as Deconfounded Empirical Risk Minimization (DERM). Such implementation is suitable for applications in Causal Inference where both the experimental settings and the outcome of interest are commonly low dimensional and discrete or anyway discretized for interpretability (Pearl et al., 2000; Rosenbaum et al., 2010). In Formula:

[3] i.e., $\forall \boldsymbol{z} : H_{\mathbb{P}^{e_1}}(Y | \boldsymbol{Z}^{e_1} = \boldsymbol{z}) > 0$.

Deconfounded Empirical Risk Minimization:

$$\hat{g} := \arg\min_{g \in \mathcal{G}} \sum_{i \in \mathcal{D}^{e_1}} \underbrace{w_i}_{\text{unconfoundness}} \cdot \underbrace{\mathcal{L}(g(\boldsymbol{x}_i), y_i)}_{\text{sufficiency}} \tag{3.5}$$

where:

$$w_i := \underbrace{\frac{1}{\widehat{\mathbb{P}}^{e_1}(Y = y_i, \boldsymbol{Z}^{e_1} = \boldsymbol{z}_i)}}_{\text{reference distribution}} \cdot \overbrace{\frac{\widehat{\mathrm{Var}}_{\mathbb{P}^{e_1}}(Y | \boldsymbol{Z}^{e_1} = \boldsymbol{z_i})}{\sum_{\boldsymbol{z}' \in \mathcal{Z}^{e_1}} \widehat{\mathrm{Var}}_{\mathbb{P}^{e_1}}(Y | \boldsymbol{Z}^{e_1} = \boldsymbol{z}')}}^{\text{fictitious distribution s.t. } \boldsymbol{Z} \perp\!\!\!\perp Y}$$

and the weights w_i are computed una tantum before training, the joint distribution is estimated by frequency, and the conditional variances with the sample variance.

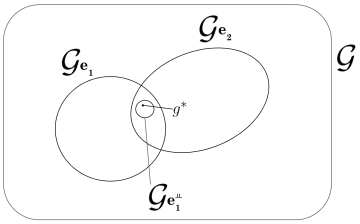


Figure 2: Illustration of the factual model hypothesis space $\mathcal{G}$, the ERM solution set over the reference $\mathcal{G}^{e_1}$ and target sample $\mathcal{G}^{e_2}$, and an *unconfounded* fictitious sample $\mathcal{G}^{e_1^{\perp\!\!\!\perp}}$.

3.1 Causal Lifting

An extensive supervised dataset to train a factual outcome model from scratch is rarely available in real-world applications. However, foundational models trained on extensive corpus may still be able to process complex data structures, e.g., images and text, preserving sufficient information for the task while having never been supervised for it directly. We can then leverage such external sources to preprocess the data and train a deconfounded head on top via DERM on the available reference sample alone. We refer to this procedure as *Causal Lifting* since enabling an expressive foundation model to filter only the invariant features for a task, thanks to a small fine-tuning on a single sample with additional supervision for the unconfoundness, i.e., the experiment settings information. Such procedure ideally leads to a conditionally unbiased factual outcome estimator, enabling efficient Causal Inference on a prediction-powered target experiment[4] according to Theorem 2.1. To summarize:

Algorithm 1 0-shot Generalization for PPCI (*Causal Lifting*)

1: **Input:** PPCI problem
2: **Output:** ATE inference on the target experiment
3: **Procedure:**
4: **Factual Outcome Model** Extract representations from experiment observations via a foundational model and fine-tune its head factual outcome estimator using DERM.
5: **Causal Inference** Via AIPW estimator on the prediction-powered target experiment.

An in-detailed description of the procedure is reported in Appendix C, while the experiment and results on both real-world and synthetic data are reported in Appendix D.

4 Conclusion

We introduced Causal Lifting, a novel paradigm enabling zero-shot generalization of foundational models for prediction-powered causal inferences. Our concrete implementation in the Deconfounded Empirical Risk Minimization (DERM) leverages a sufficiency loss paired with an unconfoundness objective in the representation space to prevent overfitting on experiment-specific spurious correlation. Additionally, we thoroughly described in which settings causal lifting can yield unbiased estimates, unlike empirical risk minimization. Our framework is widely applicable to the analysis of experimental data, which we have empirically evaluated on the ISTAnt data set. Overall, this work offers a paradigm shift from the causal representation learning literature to learning representations that enable downstream causal estimates on real-world data, which we think is a critical component of representation learning to accelerate scientific discovery. The main limitation of this work is that via PPCIs we can rarely have guarantees a priori on the Causal Estimates, being Condition 2.6 untestable without target annotations and Condition 3.4 potentially violated (on top of unobserved confounders

[4]Either a Randomized Controlled Trial or Observational Study with observed confounders.

issues). Model convergence is also not discussed, which is particularly interesting in the finite setting. At the same time, we hope that more systematic (scientifically motivated) benchmarking will lead the progress of the field, e.g., challenging and comparing Causal Representation Learning identifiability results beyond their controlled assumptions.

REFERENCES

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.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

Haitham A Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021.

Michael W Mullowney, Katherine R Duncan, Somayah S Elsayed, Neha Garg, Justin JJ van der Hooft, Nathaniel I Martin, David Meijer, Barbara R Terlouw, Friederike Biermann, Kai Blin, et al. Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery*, 22(11): 895–916, 2023.

Roberto Castello, Alina Walch, Raphaël Attias, Riccardo Cadei, Shasha Jiang, and Jean-Louis Scartezzini. Quantification of the suitable rooftop area for solar panel installation from overhead imagery using convolutional neural networks. In *Journal of Physics: Conference Series*, volume 2042, page 012002. IOP Publishing, 2021.

Connor T Jerzak, Fredrik Johansson, and Adel Daoud. Image-based treatment effect heterogeneity. *arXiv preprint arXiv:2206.06417*, 2022.

Adel Daoud, Felipe Jordán, Makkunda Sharma, Fredrik Johansson, Devdatt Dubhashi, Sourabh Paul, and Subhashis Banerjee. Using satellite images and deep learning to measure health and living standards in india. *Social Indicators Research*, 167(1):475–505, 2023.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Riccardo Cadei, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello. Smoke and mirrors in causal downstream tasks. *arXiv preprint arXiv:2405.17151*, 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*, 2024a.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Hachette UK, 2018.

James Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9572–9581, 2019.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19 (2):3, 2000.

Paul R Rosenbaum, P Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

Edward H Kennedy, Sivaraman Balakrishnan, and Max G'sell. Sharp instruments for classifying compliers and generalizing causal effects. 2020.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. *arXiv preprint arXiv:2406.02873*, 2024.

Piersilvio De Bartolomeis, Javier Abad, Guanbo Wang, Konstantin Donhauser, Raymond M. Duch, Fanny Yang, and Issa J. Dahabreh. Efficient randomized experiments using foundation models. *arXiv preprint arxiv:2502.04262*, 2025.

Pierre-Emmanuel Poulet, Maylis Tran, Sophie Tezenas du Montcel, Bruno Dubois, Stanley Durrleman, and Bruno Jedynak. Prediction-powered inference for clinical trials. *medRxiv*, pages 2025–01, 2025.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619, 2021.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Martin Q Ma, Yao-Hung Hubert Tsai, Paul Pu Liang, Han Zhao, Kun Zhang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Conditional contrastive learning for improving fairness in self-supervised learning. *arXiv preprint arXiv:2106.02866*, 2021.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7399–7407, 2022.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024.

Julius von Kügelgen. *Identifiable Causal Representation Learning: Unsupervised, Multi-View, and Multi-Environment*. PhD thesis, Cambridge University, 2024.

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

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.

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.

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.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

A PROOFS

A.1 PROOF OF THEOREM 2.1

Theorem. *Let $\hat{u}$ and $\hat{e}$ be the nuisance function estimators such that they satisfy: $||\hat{\mu} - \mu\| \cdot \|\hat{e} - e\| = o_{\mathbb{P}}(n^{-1/2})$. Further, assume that positivity holds, i.e. there exists $\epsilon > 0$ such that $\epsilon \leq e_0(W_i) \leq 1 - \epsilon$ for all W_i on the target experiment. Then, if the factual outcome model g satisfies Equation 2.6, the AIPW estimator over the prediction-powered target experiment is asymptotically normal—i.e. it holds that:*

$$\sqrt{n}(\hat{\tau} - \tau) \to \mathcal{N}(0, V),$$

where V denotes the asymptotic variance.

Proof. Our goal is to prove that the following estimator is asymptotically normal:

$$\hat{\tau} = \frac{1}{n^{e_2}} \sum_{i \in \mathcal{D}^{e_2}} \Bigg[\hat{\mu}(W_i, 1) - \hat{\mu}(W_i, 0) + \frac{T_i}{\hat{e}(W_i)}(g(X_i) - \hat{\mu}(W_i, 1)) - \frac{1 - T_i}{1 - \hat{e}(W_i)}(g(X_i) - \hat{\mu}(W_i, 0)) \Bigg]$$

where $\hat{\mu}(w, t)$ is an estimator of $\mathbb{E}_{\mathbb{P}^{e_2}}[g(X)|W = w, T = t]$ over the prediction-powered target sample $\mathcal{D}^{e_2}$, $\hat{e}(w)$ is an estimator of the propensity score $\mathbb{P}_{\mathbb{P}^{e_2}}(T = 1|W = w)$ over the target sample $\mathcal{D}^{e_2}$, and without ambiguity we dropped the superscript e_2 to the random variables to simplify the language.

Let us define the influence function of the estimator

$$\phi(O_i; \mu, e, g) = \mu(W_i, 1) - \mu(W_i, 0) + \frac{T_i}{e(W_i)}(g(X_i) - \mu(W_i, 1)) - \frac{1 - T_i}{1 - e(W_i)}(g(X_i) - \mu(W_i, 0)),$$

where $O_i = (W_i, X_i, T_i, Y_i)$. Then $\hat{\tau} = \frac{1}{n^{e_2}} \sum_{i \in \mathcal{D}^{e_2}} \phi(O_i; \hat{\mu}, \hat{e}, g)$.

We can rewrite our estimator as:

$$\hat{\tau} - \tau_0 = \frac{1}{n^{e_2}} \sum_{i \in \mathcal{D}^{e_2}} [\phi(O_i; \mu, e, g) - \tau_0] + \frac{1}{n^{e_2}} \sum_{i \in \mathcal{D}^{e_2}} \underbrace{[\phi(O_i; \hat{\mu}, \hat{e}, g) - \phi(O_i; \mu, e, g)]}_{\Delta_i}$$

Assuming that the second moment of the random variable ϕ is bounded, by a standard central limit theorem argument, the second term satisfies

$$\sqrt{n^{e_2}} \left(\frac{1}{n^{e_2}} \sum_{i \in \mathcal{D}^{e_2}} \phi(O_i; \mu_0, e_0, g_0) - \tau_0 \right) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\phi^2]).$$

It remains to show that the first term multiplied by $\sqrt{n^{e_2}}$ goes to zero in probability, i.e. it is asymptotically negligible.

To do so, observe that we can rewrite the second term as

$$\frac{1}{n^{e_2}} \sum_{i=1} \Delta_i = (\mathbb{P}_n - \mathbb{P})(\Delta_i) + \mathbb{P}(\Delta_i),$$

where $\mathbb{P}$ and $\mathbb{P}_n$ are the true and empirical target measures; $\mathbb{P}(\cdot) = \mathbb{E}[\cdot]$ as it is standard in empirical process theory. Our goal is therefore to show that both of these terms

$$\underbrace{(\mathbb{P}_n - \mathbb{P})(\Delta_i)}_{T_1} + \underbrace{\mathbb{P}(\Delta_i)}_{T_2} = o_{\mathbb{P}}(n^{-1/2}).$$

Controlling the term T_1 The first term T_1 is easy to control, as it follows directly from the following lemma.

Lemma 1. *(Kennedy et al., 2020) Let $\widehat{f}(z)$ be a function estimated from a sample $Z^N = (Z_{n+1}, \ldots, Z_N)$, and let $\mathbb{P}_n$ denote the empirical measure over $(Z_1, \ldots, Z_n)$, which is independent of Z^N. Then*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|}{\sqrt{n}}\right).$$

Since we have from assumptions that $||\phi(\cdot; \hat{\mu}, \hat{e}, g) - \phi(\cdot; \mu, e, g)||_2^2 = o_{\mathbb{P}}(1)$, it holds that $T_1 = o_{\mathbb{P}}(n^{-1/2})$.

Controlling the term T_2 The second term requires some care. We will focus on the term involving $T_i = 1$; the case for $T_i = 0$ follows by symmetry. For $T_i = 1$, after some simple calculations, we have:

$$\Delta_i = (\hat{\mu}(W_i, 1) - \mu(W_i, 1)) + \frac{1}{\hat{e}(W_i)}\left(g(X_i) - \hat{\mu}(W_i, 1)\right) - \frac{1}{e(W_i)}\left(g(X_i) - \mu(W_i, 1)\right).$$

Note that we can drop the third term since, by assumption, g and μ are equal on average. Therefore, we can write:

$$\mathbb{P}(\Delta_i) = \mathbb{E}[\Delta_i] = \mathbb{E}[\hat{\mu}(W_i, 1) - \mu(W_i, 1) + \frac{1}{\hat{e}(W_i)}\left(g(X_i) - \hat{\mu}(W_i, 1)\right)].$$

Under our assumption that $\mathbb{E}[g(X) \mid W, T] = \mu(W, T), \mathbb{P}^{e_2} -$ almost surely, we can group terms as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\hat{\mu}(W_i, 1) - \mu(W_i, 1) + \frac{1}{\hat{e}(W_i)}\left(g(X_i) - \hat{\mu}(W_i, 1)\right)\Big] \\
&\qquad = \mathbb{E}\Big[\hat{\mu}(W_i, 1) - \mu(W_i, 1) + \frac{1}{\hat{e}(W_i)}\left(\mu(W, 1) - \hat{\mu}(W_i, 1)\right)\Big] \\
&\qquad = \mathbb{E}\Big[(e(W)/\hat{e}(W) - 1)(\mu(W, 1) - \hat{\mu}(W, 1))\Big] \\
&\qquad \leq \frac{1}{\epsilon}||e - \hat{e}||_2||\mu - \hat{\mu}||_2 = o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}$$

□

B Related Works

Prediction-Powered Causal Inference The factual outcome estimation problem for causal inference from high-dimensional observation was first introduced by Cadei et al. (2024). We extended the problem to generalization to a class of SCMs, also considering observational studies motivated by practitioners desiderata, e.g., experimental ecologists. In our paper, we formalize what they describe as "encoder bias" (see our discussion on Sufficiency and Unconfoundness in Section 3), and our DERM is the first proposal to their call for "*new methodologies to mitigate this bias during adaptation*". Demirel et al. (2024) already attempted to discuss some generalization challenges in a PPCI problem but with unrealistic motivating assumptions. They ignored any high dimensional projection of the outcome of interest and assumed the experimental settings alone as sufficient for factual outcome estimation together with support overlapping, i.e., not generalization, making the model too application-specific and ignoring any connection with representation learning. Let's further observe that their framework is a special case of ours when $\boldsymbol{X} = \boldsymbol{Z}$ and low-dimensional, with the target experiment in-distribution. In contrast to the classic Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023a;b), which improves estimation efficiency by imputing unlabeled in-distribution data via a predictive model, and recent causal inference extensions (De Bartolomeis et al., 2025; Poulet et al., 2025) relying on counterfactual predictions, PPCI focuses on imputing missing *factual* outcomes to generalize across unlabeled experiments, that have different and potentially non-overlapping "covariate" distributions, agnostic of the causal estimator.

Unconfoundness Learning representations invariant to certain attributes is a challenging and widely studied problem in different machine learning communities (Moyer et al., 2018). We wish to learn useful representations of X and that can predict the outcome Y, but are invariant to the enviromental variables Z. In agreement with Yao et al. (2024a), we achieve Causal Representation Learning in DERM by combining a sufficiency objective with an invariance constraint enforcing unconfoundness in the representation space. Several alternative approaches can be considered to enforce such conditional independence constraints: (i) Conditional Mutual Information Minimization (Song et al., 2019; Cheng et al., 2020; Gupta et al., 2021) (ii) Adversarial Independence Regularization such as Louizos et al. (2015) which modifies variational autoencoder (VAE) architecture in Kingma (2014) to learn *fair* representations that are invariant sensitive attributes, by training against an adversary that tries to predict those variables (iii) Conditional Contrastive Learning such as Ma et al. (2021) whereby one learns representations invariant to certain attributes by optimizing a conditional contrastive loss (iv) Variational Information Bottleneck methods where one learns useful and sufficient representations invariant to a specific *domain* Alemi et al. (2016); Li et al. (2022).

Causal Representation Learning In the broader context of causal representation learning methods (Schölkopf et al., 2021), our proposal largely focuses on representation learning applications to causal inference: learning representations of data that make it possible to estimate causal estimands. We find this is in contrast with most recent works in causal representation learning, which uniquely focused on complete identifiability of all the variables or blocks, see Yao et al. (2024a); Varici et al. (2024); von Kügelgen (2024) for recent overviews targeting general settings. The main exceptions are Yao et al. (2024a;b). The former leverages domain generalization regularizers to debias treatment effect estimation in ISTAnt from selection bias. However, their proposal is not sufficient to prevent confounding when no data from the target experiment is given. The latter uses multi-view causal representation learning models to model confounding for adjustment in an observational climate application. In our paper, we also discuss conditions for identification, but we focus on a specific causal estimand, as opposed to block-identifiability of causal variables. Additionally, our perspective offers clear evaluation and benchmarking potential – even in theoretically underspecified setting: the accuracy of the causal estimate. As opposed to virtually all existing work in causal representation learning, this can be empirically tested in *real world* scientific experiments.

C ALGORITHM

We provide here a detailed description of our pipeline for 0-shot Generalization for PPCI, Causal Lifting a foundational model via DERM (Algorithm 1). Without loss of generality, we focus on applications with binary treatment estimating the Average Treatment Effect via AIPW estimator (as suggested by Theorem 2.1), but the same procedure with different Causal Inference estimators can be considered when the treatment is discrete or continuous.

Algorithm 1 0-shot Generalization for PPCI (*Causal Lifting*)

1: **Input:**
- a PPCI problem, i.e, $\mathcal{D}^{e_1} = \{(T_i^{e_1}, \boldsymbol{W}_i^{e_1}, Y_i, \boldsymbol{X}_i)\}_{i=1}^{n^{e_1}}$ and $\mathcal{D}^{e_2} = \{(T_i^{e_2}, \boldsymbol{W}_i^{e_2}, _, \boldsymbol{X}_i)\}_{i=1}^{n^{e_2}}$,
- a pre-trained encoder $\phi^* : \mathcal{X} \to \mathbb{R}^d$, i.e., foundational model, and an hypothesis space $\mathcal{H}$ of candidate classification heads $h : \mathbb{R}^d \to \mathcal{Y}$, i.e., model architecture,
- an optimizer, e.g. Stochastic Gradient Descent,
- a (potential) outcome and a propensity score estimator for AIPW, e.g., XGBoost (Chen and Guestrin, 2016).

2: **Output:** Average Treatment Effect Inference on the target experiment, i.e., estimate

$$\tau := \mathbb{E}_{\mathbb{P}^{e_2}}[Y|do(T^{e_2} = 1)] - \mathbb{E}_{\mathbb{P}^{e_2}}[Y|do(T^{e_2} = 0)] \tag{C.1}$$

3: **Procedure:**

4: **Factual Outcome Model** Solve

$$\hat{h} := \arg\min_{h \in \mathcal{H}} \frac{1}{n^{e_1}} \sum_{i \in \mathcal{D}^{e_1}} \underbrace{w_i}_{\text{unconfoundness}} \cdot \underbrace{\mathcal{L}^{\text{task}}(h \circ \phi^*(\boldsymbol{x}_i), y_i)}_{\text{sufficiency}} \tag{C.2}$$

using the given optimizer, where the weights:

$$w_i := \underbrace{\frac{1}{\widehat{\mathbb{P}}^{e_1}(Y = y_i, \boldsymbol{Z}^{e_1} = \boldsymbol{z}_i)}}_{\text{reference distribution}} \cdot \overbrace{\frac{\widehat{\text{Var}}_{\mathbb{P}^{e_1}}(Y|\boldsymbol{Z}^{e_1} = \boldsymbol{z_i})}{\sum_{\boldsymbol{z}' \in \mathcal{Z}^{e_1}} \widehat{\text{Var}}_{\mathbb{P}^{e_1}}(Y|\boldsymbol{Z}^{e_1} = \boldsymbol{z}')}}^{\text{fictitious distribution s.t. } \boldsymbol{Z} \perp\!\!\!\perp Y} \tag{C.3}$$

are computed una tantum before training. The (conditional) variance and joint distribution in the weights are estimated via sample variance and frequency, respectively, over the reference experiment and the experiment settings are discretized if continuous.

5: **Causal Inference** Via AIPW estimator on the prediction-powered target sample $\mathcal{D}^{e_2}$ by $\hat{g} := \hat{h} \circ \phi^*$:

$$\begin{aligned}\widehat{\tau} := \frac{1}{n^{e_2}} \sum_{i \in \mathcal{D}^{e_2}} \Bigg[& \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 1) - \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 0) + \frac{T_i^{e_2}}{\widehat{e}(\boldsymbol{W}_i^{e_2})}(\hat{g}(\boldsymbol{X}_i) - \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 1)) \\ & - \frac{1 - T_i^{e_2}}{1 - \widehat{e}(\boldsymbol{W}_i^{e_2})}(\hat{g}(\boldsymbol{X}_i) - \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 0)) \Bigg]\end{aligned} \tag{C.4}$$

where the propensity score model $\hat{e} : \mathcal{W} \to \mathcal{T}$ and the potential outcomes model $\hat{\mu} : \mathcal{W} \times \mathcal{T} \to \mathcal{Y}$ are also trained on the prediction powered target experiment. Confidence Intervals, leveraging the asymptotic normality from Thm. 2.1:

$$C_\alpha(\tau) = \left[\hat{\tau} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\text{Var}}(\hat{\tau}_i)}{n^{e_2}}} \right] \tag{C.5}$$

where $\hat{\tau}_i$ are the addends in Equation C.4.

We additionally describe the pipeline if proposing to train such a model from scratch (Algorithm 2).

Algorithm 2 0-shot Generalization for PPCI (*from scratch*)

1: **Input:**
- a PPCI problem, i.e, $\mathcal{D}^{e_1} = \{(T_i^{e_1}, \boldsymbol{W}_i^{e_1}, Y_i, \boldsymbol{X}_i)\}_{i=1}^{n^{e_1}}$ and $\mathcal{D}^{e_2} = \{(T_i^{e_2}, \boldsymbol{W}_i^{e_2}, _, \boldsymbol{X}_i)\}_{i=1}^{n^{e_2}}$,
- an hypothesis space $\mathcal{G}$ of candidate classification model $g : \mathcal{X} \to \mathcal{Y}$, i.e., model architecture,
- an optimizer, e.g. ADAM (Kingma, 2014),
- a (potential) outcome and a propensity score estimator for AIPW, e.g., XGBoost (Chen and Guestrin, 2016).

2: **Output:** Average Treatment Effect Inference on the target experiment, i.e., estimate

$$\tau := \mathbb{E}_{\mathbb{P}^{e_2}}[Y|do(T^{e_2} = 1)] - \mathbb{E}_{\mathbb{P}^{e_2}}[Y|do(T^{e_2} = 0)] \tag{C.6}$$

3: **Procedure:**

4: **Factual Outcome Model** Solve

$$\hat{g} := \arg\min_{g \in \mathcal{G}} \frac{1}{n^{e_1}} \sum_{i \in \mathcal{D}^{e_1}} \underbrace{w_i}_{\text{unconfoundness}} \cdot \underbrace{\mathcal{L}^{\text{task}}(g(\boldsymbol{x}_i), y_i)}_{\text{sufficiency}} \tag{C.7}$$

using the given optimizer, where the weights:

$$w_i := \underbrace{\frac{1}{\widehat{\mathbb{P}}^{e_1}(Y = y_i, \boldsymbol{Z}^{e_1} = \boldsymbol{z}_i)}}_{\text{reference distribution}} \cdot \overbrace{\frac{\widehat{\text{Var}}_{\mathbb{P}^{e_1}}(Y|\boldsymbol{Z}^{e_1} = \boldsymbol{z_i})}{\sum_{\boldsymbol{z}' \in \mathcal{Z}^{e_1}} \widehat{\text{Var}}_{\mathbb{P}^{e_1}}(Y|\boldsymbol{Z}^{e_1} = \boldsymbol{z}')}}^{\text{fictitious distribution s.t. } \boldsymbol{Z} \perp\!\!\!\perp Y} \tag{C.8}$$

are computed una tantum before training. The (conditional) variance and joint distribution in the weights are estimated via sample variance and frequency, respectively, over the reference experiment and the experiment settings are discretized if continuous.

5: **Causal Inference** Via AIPW estimator on the prediction-powered target sample $\mathcal{D}^{e_2}$ by $\hat{g}$:

$$\begin{aligned}\widehat{\tau} := \frac{1}{n^{e_2}} \sum_{i \in \mathcal{D}^{e_2}} \Bigg[& \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 1) - \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 0) + \frac{T_i^{e_2}}{\widehat{e}(\boldsymbol{W}_i^{e_2})}(\hat{g}(\boldsymbol{X}_i) - \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 1)) \\ & - \frac{1 - T_i^{e_2}}{1 - \widehat{e}(\boldsymbol{W}_i^{e_2})}(\hat{g}(\boldsymbol{X}_i) - \widehat{\mu}(\boldsymbol{W}_i^{e_2}, 0)) \Bigg]\end{aligned} \tag{C.9}$$

where the propensity score model $\hat{e} : \mathcal{W} \to \mathcal{T}$ and the potential outcomes model $\hat{\mu} : \mathcal{W} \times \mathcal{T} \to \mathcal{Y}$ are also trained on the prediction powered target experiment. Confidence Intervals, leveraging the asymptotic normality from Thm. 2.1:

$$C_\alpha(\tau) = \left[\hat{\tau} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\text{Var}}(\hat{\tau}_i)}{n^{e_2}}} \right] \tag{C.10}$$

where $\hat{\tau}_i$ are the addends in Equation C.9.

D EXPERIMENTS

We empirically validate our approach for 0-shot generalization of PPCI solving a real-world scientific problem from experimental Behavioural Ecology. In particular, we considered ISTAnt dataset[5] (Cadei et al., 2024), and we designed and collected a similar experimental dataset with lower filming quality and higher diversity in treatments to enable causal lifting of different pre-trained models for 0-shot generalization. We further validate our analysis on a synthetic causal manipulation of the MNIST dataset (LeCun, 1998), controlling the ground truth causal effect.

D.1 ISTANT

To empirically validate our methodology, we replicated a real-world scientific pipeline requiring generalization for PPCI and compared our performances with the existing approaches suitable for the task. We performed a similar experiment to the ISTAnt dataset with lower-quality filming conditions (in particular light conditions), slightly different ant coloring and diversified treatments with or without micro-particle application. We considered it the reference experiment and proposed to generalize to ISTAnt, the target experiment in the PPCI problem. The pipeline reflects a common desiderata in experimental research: learning a model from previously annotated experiments able to cheaply and validly annotate new, out-of-distribution experiments. The new experiments commonly consider more experimental settings and rely on better or generally different data acquisition techniques.

D.1.1 REFERENCE EXPERIMENT AND DATA RECORDING

We run an experiment very much alike the ISTAnt experiment with triplets of worker ants following the step-by-step design described in Appendix C in Cadei et al. (2024). We recorded 5 batches of 9 simultaneously run replicates, producing 45 original videos, of which one had to be excluded for experimental problems, leaving 44 analyzable videos. We used a comparable experimental setup (i.e., camera set-up, random treatment assignment, etc.) except for the following, obtaining a *similar* experiment belonging to the same class of SCMs (see Problem Formulation in Section 2).

- **Treatments:** Whereas ISTAnt used two micro-particle applications [6], our experimental treatments also constitute micro-particle application in two different treatments (n=15 each), but also one treatment completely free of micro-particles (control, n=14), all applied to the focal ant. The three treatments of the ants are visually indistinguishable, independent of micro-particle application.
- **Light conditions:** We created a lower-quality illumination of the nests by implementing a ring of light around the experiment container, resulting in more inhomogeneous lighting and a high-lux ("cold") light effect, compared to the light diffusion by a milky plexiglass sheet proposed in the original experiment. Also, our ant nests had a higher rim from the focal plane where the ants were placed, causing some obscuring of ant observation along the walls. See a comparison of the filming set-up and an example of the resulting recording in Figure 4. We also considered a slightly lower resolution, i.e., 700x700 pixels.
- **Longer Videos**: Whereas ISTAnt annotated 10 min long videos, we here annotated 30 min long videos. Ant activity generally decreases with time from the first exposure to a new environment. Our videos were recorded at 30fps, totaling 158400 annotated frames in the 44 videos.
- **Other potential distribution shifts**: Other sources of variations from the original experiment are:
  - Whereas ISTAnt used orange and blue color dots, we used yellow and blue.
  - Whereas in ISTAnt, grooming presence or absence was annotated for each frame, we here annotated a single grooming event even if the ant stopped grooming for up to one second but then kept grooming after that, with no other behaviors being performed in between. This means that intermediate frames between grooming frames were also

[5] So far the only benchmark dataset for scientifically motivated representation learning for a causal downstream task.

[6] By author correspondence.

annotated as grooming despite the ant pausing its behavior. Such less exact grooming annotations are faster to perform for the human annotator.
- The person performing annotation in this experiment was different from the annotators in the ISTAnt dataset, leading to some possible observer effects.

Despite all the experiment variants, we still rely on the assumption that an invariant factual outcome model for behavior classification g^* exists, as well as we are making when considering ground truth human experts manually annotating frame by frame. Let's observe that our work models the general pipeline in experimental ecologists, where multiple experiment variants are recorded over time, e.g., upgrading the data acquisition technique, and we aim to generalize from a lower to higher quality *similar* experiment.

D.1.2 Analysis

We considered our dataset as a Reference Experiment in a PPCI problem trying to generalize a factual outcome model to the original ISTAnt dataset, i.e., Target Experiment, causal lifting of a foundational model. For each pre-trained encoder – ViT-B (Dosovitskiy et al., 2020), ViT-L (Zhai et al., 2023), CLIP-ViT-B,-L (Radford et al., 2021), DINOv2 (Oquab et al., 2023), we fine-tuned a multi-layer perception head (2 hidden layers with 256 nodes each and ReLU activation) on top of its *class* token via Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.9, \epsilon = 10^{-8}$) for ERM, vREx (finetuning the invariance constraint in $\{0.01, 0.1, 1, 10\}$ and DERM (ours) for 15 epochs and batch size 256. So, we fine-tuned the learning rates in $[0.0005, 0.5]$, selecting the best-performing hyper-parameters for each model-method, minimizing the Treatment Effect Bias on the reference sample. We computed the ATE at the video level (aggregating the predictions per frame) via the AIPW estimator. We used XGBoost for the model outcome and estimated the propensity score via sample mean (constant) since the treatment assignments are randomized, i.e., RCT. For the outcome model, we consider the following experiment settings for controlling: experiment day, time of the day, batch, position in the batch, and annotator. For reference we considered the ATE Inference on ISTAnt by the AIPW estimator on the human-annotated factual outcomes (ground truth): on average, the treatment in ISTAnt increases the grooming time towards the focal ant by ≈ 40 seconds.

D.1.3 Results

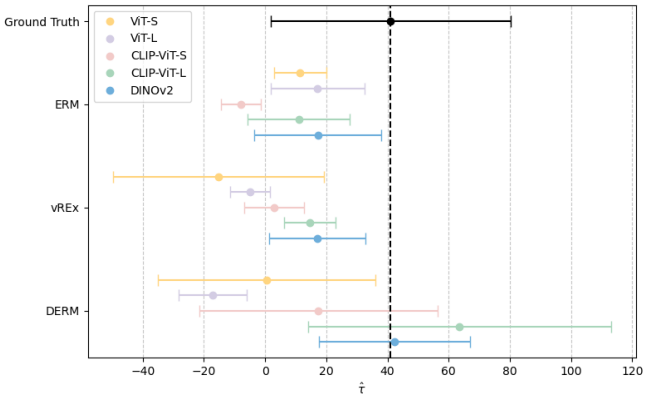


Figure 3: 0-shot ATE Inference on ISTAnt dataset from our experiment, varying method and pre-trained encoder. 95% confidence intervals estimated via AIPW asymptotic normality and baseline in black using AIPW on the ground truth outcome. Our approach applied to unsupervised backbones yield consistent estimates, unlike ERM or a general invariance regularization.

(a) Our experiment (b) ISTAnt

Figure 4: Filming box and example frame from our experiment and ISTAnt. The two datasets mainly differ in lighting quality, treatments considered, experimental nests (wall height) and color marking.

Figure 3 summarizes the results of our generalization experiment comparing the 95% confidence intervals obtained by AIPW asymptotic normality. As expected, with vanilla ERM, there are no guarantees to Causal Lift any foundational model due to potential confounding effects, and the ATE

estimates are consistently offset by underestimating/ignoring it. Similar results, using v-REx, as proposed by Yao et al. (2024a). Indeed, experiment setting performance invariance is not sufficient to prevent confounding effects on the target when certain association switches. DERM is the only method enabling 0-shot generalization for PPCI with DINOv2 and (partially) with CLIP-based vision transformers. Interestingly enough, the most supervised encoder, i.e., ViT-based (trained on ImageNet (Deng et al., 2009)), struggles in the task, underestimating the effect, as opposed to the ones trained in a fully unsupervised fashion. We hypothesize that encoders pre-trained in a supervised fashion are more inclined to extract more entangled representations, more challenging to causal lift.

D.2 CAUSALMNIST

To fully validate our method, we replicated the comparison in a controlled setting, manipulating the MNIST dataset with coloring, allowing us to (i) cheaply replicate fictitious experiments, e.g., estimating the effect of the background color or pen color on the digit value, several times and bootstrapping confidence intervals and (ii) control the underlying causal effects. While simpler, this experiment supplements the fact that obtaining ground-truth causal effect on real-world data is challenging, and one whole experiment only yields a single measurement of a target causal estimand.

D.2.1 DATA GENERATING PROCESS

We define the (universal) Structural Equations as follows:

$$\boldsymbol{Z} := [T^1, T^2, T, W, U] = n_{\boldsymbol{Z}} \tag{D.1}$$

$$Y = W \cdot \text{Unif}(\{0,1,2,3\}) + T^1 \cdot \text{Unif}(\{0,1,2,3\}) + U \cdot \text{Unif}(\{0,1,2,3\}) \tag{D.2}$$

$$X := f_{\boldsymbol{X}}(T, W, Y, U, n_{\boldsymbol{X}}) \tag{D.3}$$

where f_X is a deterministic manipulation of a random digit image $n_{\boldsymbol{X}}$ from MNIST dataset enforcing the background color W (red or green) and pen color T (black or white) and a padding size U.

For the reference experiment (RCT) we intervene on the experimental settings such that:

$$T^1 := Be(0.5) \tag{D.4}$$

$$T^2 := 0 \tag{D.5}$$

$$T := \max(T^1, T^2) \tag{D.6}$$

$$W := Be(p_W) \tag{D.7}$$

$$U := Be(p_U) \tag{D.8}$$

with $\boldsymbol{Z} := [\underbrace{\underbrace{T^2, T}_{U^{e_1}}, T^1}_{Z^{e_1}}, \underbrace{W, U}_{W^{e_1}}]$. While for the target experiment (OS):

$$T^1 := 0 \tag{D.9}$$

$$T^2 := Be(0.1) \cdot (1 - W) + Be(0.9) \cdot W \tag{D.10}$$

$$T := \max(T^1, T^2) \tag{D.11}$$

$$W := Be(p_W) \tag{D.12}$$

$$U := Be(p_U) \tag{D.13}$$

with $\boldsymbol{Z} := [\underbrace{\underbrace{T^1, T}_{U^{e_2}}, T^2}_{Z^{e_2}}, \underbrace{W, U}_{W^{e_2}}]$.

By definition, whatever the experiment setting distribution, the ATE on the reference experiment is 1.5, while on target, it is 0. We especially considered a reference sample $\mathcal{D}^{e_1}$ with $n_{e_1} = 10\,000$, $p_U = 0.02$ and $p_W = 0.5$; a target sample $\mathcal{D}^{e_2}$ with $n_{e_2} = 10\,000$, $p_U = 0.05$ and $p_W = 0.05$; and additional target sample $\mathcal{D}^{e_3}$ with $n_{e_3} = 10\,000$ and $p_U = 0.5$ and $p_W = 0.5$. Six examples of colored handwritten digits from CausalMNIST are reported in Figure 5.

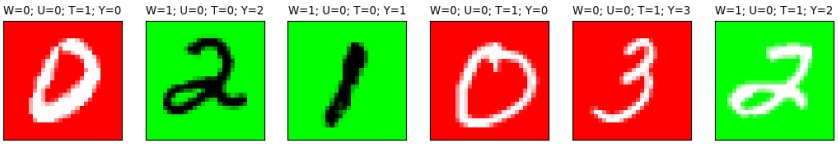


Figure 5: Random samples from a CausalMNIST sample.

D.2.2 Analysis

We fully replicated the modeling choices for CausalMNIST proposed in Cadei et al. (2024) and described in (their) Appendix E.2 (without relying on pre-trained models). We replicated the same hyper-parameter tuning (for training) and ATE inference from our experiments on ISTAnt generalization. We replicated each experiment 50 times, including resampling the data, and bootstrapped the confidence interval of the ATE estimates via AIPW.

D.2.3 Results

In Table 1, we report the ATE inference, bias and standard deviation, (i) on a target experiment $\mathcal{D}^{e_2}$ with a new treatment with the same appearance (see Example 1 in Section 2) and (ii) on a target experiment $\mathcal{D}^{e_3}$ strongly out-of-support. DERM is the unique method solving the problem on $\mathcal{D}^{e_2}$, and despite no method having guarantees on $\mathcal{D}^{e_3}$, it is still the least biased.

Method	$\mathcal{D}^{e_1}$ (ATE $= 1.5$)	$\mathcal{D}^{e_2}$ (ATE $= 0$)	$\mathcal{D}^{e_3}$ (ATE $= 0$)
ERM	$\mathbf{0.00 \pm 0.02}$	0.86 ± 0.14	1.05 ± 0.15
v-REx	0.01 ± 0.03	0.83 ± 0.15	1.05 ± 0.14
Ours	0.10 ± 0.07	$\mathbf{0.14 \pm 0.14}$	$\mathbf{0.75 \pm 0.05}$

Table 1: ATE bias and standard deviation via AIPW on a reference trial $\mathcal{D}^{e_1}$ and two target samples $\mathcal{D}^{e_2}$-$\mathcal{D}^{e_3}$ of CausalMNIST not annotated and prediction-powered by a Convolutional Neural Network trained with different objectives. Sample mean and standard deviation are computed over the same PPCI problem repeated 50 times, re-sampling both reference and target samples. ERM and v-REX yield biased estimates on the new population $\mathcal{D}^{e_2}$, unlike our approach.