

Appendix

Table of Contents

478	A Proofs	13
479	A.1 Proof of Lemma 2.1	13
480	A.2 Proof of Lemma 2.2	14
481	A.3 Proof of Theorem 3.1	16
482	B ISTAnt	17
483	B.1 Similar experiment and data recording (ours)	17
484	B.2 Zero-shot PPCI	17
485	C CausalMNIST	19
486	C.1 Data Generating Process	19
487	C.2 Analysis	20

A Proofs

Note: In the rest of the manuscript we use the expression "under standard causal identification assumptions" to refer to the canonical conditions for identifiability in a observational study (and so a randomized controlled trial too) [Rubin, 1974], i.e.,:

- *Stable Unit Treatment Value Assumption (SUTVA)*, i.e., not interference and no hidden version of the treatment,
- *Overlap Assumption*, i.e., $0 < \mathbb{E}[T|W = w] < 1$ for all $w \in \mathcal{W}$,
- *Conditional Exchangeability (or Unconfoundness) Assumption*, i.e., $\mathbb{P}(Y|do(T = t), W) = \mathbb{P}(Y|T = t, W)^4$.

A.1 Proof of Lemma 2.1

Lemma (Prediction-Powered Identification). *Given a PPCI problem \mathcal{P} with standard causal identification assumptions. If an outcome model $g : \mathcal{X} \rightarrow \mathcal{Y}$ is conditionally calibrated with respect to the treatment and a valid adjustment set, then it is (causally) valid for \mathcal{P} .*

Proof. Given the identification of the causal estimand, the thesis follows directly by the tower rule over a valid adjustment set \tilde{W} and leveraging the conditional calibration:

$$\tau_Y(t) = \mathbb{E}[Y|do(T = t)] = \tag{11}$$

$$= \mathbb{E}_{\tilde{W}}[\mathbb{E}_Y[Y|do(T = t), \tilde{W}]] = \tag{12}$$

$$= \mathbb{E}_{\tilde{W}}[\mathbb{E}_Y[Y|T = t, \tilde{W}]] = \tag{13}$$

$$= \mathbb{E}_{\tilde{W}}[\mathbb{E}_X[g(X)|T = t, \tilde{W}]] = \tau_{g(X)}(t) \quad \forall t \in \mathcal{T}. \tag{14}$$

⁴Re-framing it in *do*-calculus notation.

504 **A.2 Proof of Lemma 2.2**

Lemma (Prediction-Powered Estimation). *Given a PPCI problem \mathcal{P} with standard causal identification assumptions. If an outcome model $g : \mathcal{X} \rightarrow \mathcal{Y}$ is conditionally calibrated with respect to the treatment and a valid adjustment set, then the AIPW estimator [Robins et al., 1994] over the prediction-powered sample with nuisance function estimators satisfying $\|\hat{\mu} - \mu\| \cdot \|\hat{e} - e\| = o_{\mathbb{P}}(n^{-1/2})$, preserves the asymptotically valid confidence interval for the true ATE, i.e.,*

$$\sqrt{n}(\hat{\tau}_{g(X)} - \tau_Y) \rightarrow \mathcal{N}(0, V),$$

where V the asymptotic variance.

505 *Proof.* Our goal is to prove that the following estimator for τ_Y is asymptotically normal:

$$\hat{\tau}_{g(X)} = \frac{1}{n} \sum_{i=1}^n \left[\hat{\mu}_{g(X)}(W_i, 1) - \hat{\mu}_{g(X)}(W_i, 0) + \frac{T_i}{\hat{e}(W_i)} (g(X_i) - \hat{\mu}_{g(X)}(W_i, 1)) - \frac{1 - T_i}{1 - \hat{e}(W_i)} (g(X_i) - \hat{\mu}_{g(X)}(W_i, 0)) \right] \quad (15)$$

506 where $\hat{\mu}_{g(X)}(w, t)$ is an estimator of the true predicted-outcome model $\mu_{g(X)}(w, t) = \mathbb{E}[g(X)|W = w, T = t]$, and $\hat{e}(w)$ is an estimator of the true propensity score $e(w) = \mathbb{P}(T = 1|W = w)$.

508 Given a generic outcome model μ and a propensity score e , let us define the influence function of the estimator:

$$\begin{aligned} \phi(O_i; \mu, e, g) &= \mu(W_i, 1) - \mu(W_i, 0) + \frac{T_i}{e(W_i)} (g(X_i) - \mu(W_i, 1)) \\ &\quad - \frac{1 - T_i}{1 - e(W_i)} (g(X_i) - \mu(W_i, 0)), \end{aligned} \quad (16)$$

510 where $O_i = (T_i, W_i, Y_i, X_i)$. Then

$$\hat{\tau}_{g(X)} = \frac{1}{n} \sum_{i=1}^n \phi(O_i; \hat{\mu}_{g(X)}, \hat{e}, g). \quad (17)$$

511 We can rewrite our estimator as:

$$\hat{\tau}_{g(X)} - \tau_Y = \frac{1}{n} \sum_{i=1}^n [\phi(O_i; \mu, e, g) - \tau_Y] + \frac{1}{n} \sum_{i=1}^n \underbrace{[\phi(O_i; \hat{\mu}_{g(X)}, \hat{e}, g) - \phi(O_i; \mu, e, g)]}_{\Delta_i}, \quad (18)$$

512 where $\mu(w, t) = \mathbb{E}[Y|W = w, T = t]$ is the true outcome model and by conditional calibration:

$$\mu(w, t) = \mathbb{E}[Y|W = w, T = t] = \mathbb{E}[g(X)|W = w, T = t] \quad \forall w \in \mathcal{W}, t \in \mathcal{T}. \quad (19)$$

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

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \phi(O_i; \mu, e, g) - \tau_Y \right) \xrightarrow{d} \mathcal{N}(0, \underbrace{\mathbb{E}[\phi^2]}_V). \quad (20)$$

515 It remains to show that the second term multiplied by \sqrt{n} goes to zero in probability, i.e. it is asymptotically negligible. To do so, observe that we can rewrite the second term as

$$\frac{1}{n} \sum_{i=1}^n \Delta_i = (\mathbb{P}_n - \mathbb{P})(\Delta_i) + \mathbb{P}(\Delta_i), \quad (21)$$

517 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

$$\underbrace{(\mathbb{P}_n - \mathbb{P})(\Delta_i)}_{T_1} + \underbrace{\mathbb{P}(\Delta_i)}_{T_2} = o_{\mathbb{P}}(n^{-1/2}). \quad (22)$$

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

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

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}} \left(\frac{\|\hat{f} - f\|}{\sqrt{n}} \right). \quad (23)$$

521 Since we have from assumptions that $\|\phi(\cdot; \hat{\mu}_{g(X)}, \hat{e}, g) - \phi(\cdot; \mu, e, g)\|_2^2 = o_{\mathbb{P}}(1)$, it holds that
 522 $T_1 = o_{\mathbb{P}}(n^{-1/2})$.

523 **Controlling the term T_2** The second term requires some care. We will focus on the term involving
 524 $T_i = 1$; the case for $T_i = 0$ follows by symmetry. For $T_i = 1$, after some simple calculations, we
 525 have:

$$\begin{aligned} \Delta_i &= (\hat{\mu}_{g(X)}(W_i, 1) - \mu(W_i, 1)) + \frac{T_i}{\hat{e}(W_i)} (g(X_i) - \hat{\mu}(W_i, 1)) \\ &\quad - \frac{T_i}{e(W_i)} (g(X_i) - \mu(W_i, 1)). \end{aligned} \quad (24)$$

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

$$\mathbb{E}[\Delta_i] = \mathbb{E}[\hat{\mu}_{g(X)}(W_i, 1) - \mu(W_i, 1) + \frac{1}{\hat{e}(W_i)} (g(X_i) - \hat{\mu}(W_i, 1))] \quad (25)$$

528 By conditional calibration, we can substitute $g(X_i)$ with $\mu(W_i, 1)$ and group:

$$\mathbb{E}[\Delta_i] = \mathbb{E} \left[(\hat{\mu}_{g(X)}(W_i, 1) - \mu(W_i, 1)) + \frac{T_i}{\hat{e}(W_i)} (\mu(W_i, 1) - \hat{\mu}_{g(X)}(W_i, 1)) \right] \quad (26)$$

$$= \mathbb{E} \left[(\hat{\mu}_{g(X)}(W, 1) - \mu(W, 1)) \left(1 - \frac{T_i}{\hat{e}(W)} \right) \right] \quad (27)$$

$$= \mathbb{E} \left[(\hat{\mu}_{g(X)}(W, 1) - \mu(W, 1)) \frac{\hat{e}(W) - T_i}{\hat{e}(W)} \right]. \quad (28)$$

529 Conditioning on W_i we obtain:

$$\mathbb{E}[\Delta_i] = \mathbb{E} \left[\left(\frac{e(W)}{\hat{e}(W)} - 1 \right) (\mu(W, 1) - \hat{\mu}_{g(X)}(W, 1)) \right]. \quad (29)$$

530 To bound this term, we use the positivity assumption that $\hat{e}(W) \geq \epsilon > 0$ for some constant ϵ :

$$|\mathbb{E}[\Delta_i]| \leq \mathbb{E} \left[\left| \frac{e(W) - \hat{e}(W)}{\hat{e}(W)} \right| \cdot |\mu(W, 1) - \hat{\mu}_{g(X)}(W, 1)| \right] \quad (30)$$

$$\leq \frac{1}{\epsilon} \mathbb{E} [|e(W) - \hat{e}(W)| \cdot |\mu(W, 1) - \hat{\mu}_{g(X)}(W, 1)|]. \quad (31)$$

531 Applying Cauchy-Schwarz inequality:

$$\frac{1}{\epsilon} \mathbb{E} [|e(W) - \hat{e}(W)| \cdot |\mu(W, 1) - \hat{\mu}_{g(X)}(W, 1)|] \leq \frac{1}{\epsilon} \|e - \hat{e}\|_2 \|\mu(\cdot, 1) - \hat{\mu}_{g(X)}(\cdot, 1)\|_2. \quad (32)$$

532 If the estimators \hat{e} and $\hat{\mu}$ achieve suitable convergence rates such that their product of L_2 norms is
 533 $o_{\mathbb{P}}(n^{-1/2})$, then:

$$\frac{1}{\epsilon} \|e - \hat{e}\|_2 \|\mu(\cdot, 1) - \hat{\mu}_{g(X)}(\cdot, 1)\|_2 = o_{\mathbb{P}}(n^{-1/2}). \quad (33)$$

534 This completes the proof. \square

535 A.3 Proof of Theorem 3.1

536 **Theorem** (Causal Lifting transfers causal validity). *Given two similar PPCI problems \mathcal{P} and \mathcal{P}'*
 537 *with standard causal identification assumptions. Let $g = h \circ \phi : \mathcal{X} \rightarrow \mathcal{Y}$ be an outcome model for \mathcal{P}'*
 538 *with h Bayes-optimal, and assume that the representation transfers, i.e., $\mathbb{P}(\phi(X)|Y) = \mathbb{P}'(\phi(X)|Y)$.*
 539 *Then the Causal Lifting constraint on \mathcal{P} implies g is valid on \mathcal{P} .*

540 *Intuition:* The theorem shows that Causal Lifting, supported by other transferability assumptions,
 541 *guarantees to transfer validity among PPCI problems, regardless of potential shifts in the joint*
 542 *distribution of (Z, Y) , e.g., different treatment effect. The core of the proof relies on establishing*
 543 *a hard representation transferability, i.e., $\mathbb{P}(\phi(X)|Y, Z) = \mathbb{P}'(\phi(X)|Y, Z)$, given by the assumed*
 544 *representation transferability (on standalone Y) and the causal lifting constraint. This is done*
 545 *within the expectation of the conditional calibration, which we have already shown implies validity.*
 546 *Note that, indeed, the (marginal) representation transferability, conditioning over Y alone, does not*
 547 *guarantee transferability conditioning over Z too, potentially due by systematic biases within the*
 548 *encoder (e.g., fully retrieving the outcome signal for a certain subgroup and partially missing it for*
 549 *others).*

550 *Proof.* First, let us remark that if $\mathcal{P} = \mathcal{P}'$, then Bayes optimality is sufficient for validity. Therefore,
 551 we now only focus on the true generalization setting with $\mathcal{P} \neq \mathcal{P}'$. By Lemma 2.1 it is sufficient to
 552 show conditional calibration to show the outcome model causal validity. Then, by linearity of the
 553 expected value we aim to show:

$$\mathbb{E}_Y[Y | Z] \stackrel{a.s.}{=} \mathbb{E}_X[h(\phi(X)) | Z], \quad (34)$$

554 where $Z = [T, \tilde{W}]$, and \tilde{W} is a valid adjustment set. By the tower rule, we can expand the RHS:

$$\mathbb{E}_X[h(\phi(X)) | Z] \stackrel{a.s.}{=} \mathbb{E}_Y \left[\underbrace{\mathbb{E}_X[h(\phi(X)) | Y, Z]}_{\zeta(Y, Z)} | Z \right]. \quad (35)$$

555 By assumptions $\forall y \in \mathcal{Y}, z \in \mathcal{Z}$:

$$\zeta(y, z) := \mathbb{E}_X[h(\phi(X)) | Y = y, Z = z] = \quad (36)$$

$$= \mathbb{E}_X[h(\phi(X)) | Y = y] = \quad (\text{Casual Lifting constraint}) \quad (37)$$

$$= \mathbb{E}'_X[h(\phi(X)) | Y = y] = \quad (\text{representation transfers}) \quad (38)$$

$$= \mathbb{E}'_X[\mathbb{E}'_Y[Y | \phi(X)] | Y = y] = \quad (\text{Bayes-optimal predictor}) \quad (39)$$

$$= \mathbb{E}_X[\underbrace{\mathbb{E}'_Y[Y | \phi(X)]}_{\text{function of } \phi(X)} | Y = y] = \quad (\text{representation transfers}) \quad (40)$$

$$= \mathbb{E}_X[\mathbb{E}'_Y[Y | X] | Y = y] = \quad (\text{sufficiency}) \quad (41)$$

$$= \mathbb{E}_X[\mathbb{E}_Y[Y | X] | Y = y] = \quad (\text{similarity}) \quad (42)$$

$$= y \quad (\text{law of iterated expectations}) \quad (43)$$

556 Where with the expected value with superscript prime we refer to the expected value over the data
 557 distribution of the problem \mathcal{P}' . Substituting back $\zeta(Y, Z)$ we have the thesis. \square

B ISTAnt

In this Section, we describe in detail our procedure to test zero-shot PPCI on ISTAnt dataset, by designing and recording a similar experiment for finetuning.

B.1 Similar experiment and data recording (ours)

We run an experiment very much alike the ISTAnt experiment with triplets of worker ants (one treated focal ant, and two nest ants), following the step-by-step design described their Appendix C [Cadei et al., 2024]. We recorded 5 batches of 9 simultaneously run replicates with similar background pen marking for the dish palettes positioning, producing 45 original videos, of which one had to be excluded for experimental problems, leaving 44 analyzable videos. Then, for each video, grooming events from the nest ants to the focal ant were annotated by a single domain expert, and we focus on the ‘or’ events, i.e., “Is one of the nest ants grooming the focal one?”. We used a comparable experimental setup (i.e., camera set-up, random treatment assignment, etc.) except for the following, guaranteeing invariant annotation mechanism, i.e., *similar* experiment.

- **Treatments:** Whereas ISTAnt used two micro-particle applications⁵, 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 3. 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 158 400 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 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.

Let’s observe that our work models the general pipeline in experimental ecology, 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.

B.2 Zero-shot PPCI

We considered the full dataset so recorded for finetuning a pre-trained Vision Transformer, i.e., 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], as proposed by Cadei et al. [2024] and we left ISTAnt for testing causal estimation performances relying on artificial predictions. For each pre-trained encoder, we fine-tuned a multi-layer perception head (2 hidden layers with 256 nodes each and ReLU activation) on top of

⁵By author correspondence.

607 its *class* token via Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.9, \epsilon = 10^{-8}$) for ERM, vREx (finetuning the
 608 invariance constraint in $\{0.01, 0.1, 1, 10\}$) and DERM (ours) for 15 epochs and batch size 256. So,
 609 we fine-tuned the learning rates in $[0.0005, 0.5]$, selecting the best-performing hyper-parameters for
 610 each model-method, minimizing the Treatment Effect Bias on the training sample, while guaranteeing
 611 good predictive performances, i.e., accuracy greater than 0.8, on a small validation set (1 000 random
 612 frames). We computed the ATE at the video level (aggregating the predictions per frame) via the
 613 AIPW estimator. We used XGBoost for the model outcome and estimated the propensity score via
 614 sample mean (constant) since the treatment assignments are randomized, i.e., RCT. For the outcome
 615 model, we consider the following experiment settings for controlling: experiment day, time of the
 616 day, batch, position in the batch, and annotator. We run all the analyses using 48GB of RAM, 20
 617 CPU cores, and a single node GPU (NVIDIA GeForce RTX2080Ti). The main bottleneck in the
 618 analysis is the feature extraction from the pre-trained Vision Transformers. We estimate 72 GPU
 619 hours to run the full analysis, despite given a candidate outcome model g already finetuned the
 620 standalone prediction-powered causal inference component, excluding feature extraction on the target
 621 experiment takes less than a GPU minute.

622 C CausalMNIST

623 To exhaustively validate our method, we replicated the comparison between DERM (our) enforcing
 624 the Causal Lifting constraint and ERM (baseline), vREx and IRM (invariant trainings) in a controlled
 625 setting, manipulating the MNIST dataset with coloring, allowing us to (i) cheaply replicate fictitious
 626 experiments several times, bootstrapping confidence intervals, and (ii) control the underlying causal
 627 effects (only empirically estimated in real-world experiments).

628 C.1 Data Generating Process

629 We considered the following training data distribution \mathbb{P}^A :

$$W = Be(0.5) \quad (44)$$

$$U = Be(0.02) \quad (45)$$

$$T = Be(0.5) \quad (46)$$

$$Y = W \cdot \text{Unif}(\{0, 1, 2, 3\}) + T \cdot \text{Unif}(\{0, 1, 2, 3\}) + U \cdot \text{Unif}(\{0, 1, 2, 3\}) \quad (47)$$

$$X := f_X(T, W, Y, U, n_X) \quad (48)$$

630 representing a Randomized Controlled Trial where f_X is a deterministic manipulation of a random
 631 digit image n_X from MNIST dataset enforcing the background color W (red or green) and pen color
 632 T (black or white) and padding size U (0 or 8). By exchangeability assumption:

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)] = \\ &= \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0] \\ &= 1.5 \end{aligned} \quad (49)$$

633 Six examples of colored handwritten digits from CausalMNIST are reported in Figure 4.



Figure 4: Random samples from a CausalMNIST sample.

634 Then, we considered the generic *similar* target Observational Study, with distribution:

$$W = Be(p_W) \quad (50)$$

$$U = Be(p_U) \quad (51)$$

$$T = Be(0.1) \cdot (1 - W) + Be(0.9) \cdot W \quad (52)$$

$$X := f_X(T, W, Y, U, n_X) \quad (53)$$

635 with *linear* outcome/effect (null):

$$Y = W \cdot \text{Unif}(\{0, 1, 2, 3\}) + \text{Unif}(\{0, 1, 2, 3\}) + U \cdot \text{Unif}(\{0, 1, 2, 3\}) \quad (54)$$

636 where simply:

$$\text{ATE} = \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)] = 0 \quad (55)$$

637 and *non-linear* outcome/effect:

$$Y = (T \vee U) \cdot \text{Unif}(\{0, 1, 2, 3\}) + \text{Unif}(\{0, 1, 2, 3, 4, 5, 6\}) \quad (56)$$

638 with respect to the experimental settings, where, by adjustment formula:

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)] = \\ &= P(U = 0) \cdot (\mathbb{E}[Y|T = 1, U = 0] - \mathbb{E}[Y|T = 0, U = 0]) = \\ &= (1 - p_U) \cdot 1.5 \end{aligned} \quad (57)$$

and we considered 4 different instances \mathbb{P}^B , \mathbb{P}^C , \mathbb{P}^D and \mathbb{P}^E varying the outcome model and experimental settings parameters. \mathbb{P}^B and \mathbb{P}^C have null effect, while \mathbb{P}^D and \mathbb{P}^E have heterogeneous effect. \mathbb{P}^B and \mathbb{P}^D are closer in distribution to the training since $p_U \ll 1$, while \mathbb{P}^C and \mathbb{P}^E are more out-of distribution since the padding variable U is balanced, i.e., $p_U = 0.5$, but during training it is rarely observed activated. For simplicity we refer to \mathbb{P}^B as the distribution of the the target experiment population with linear effect and “soft” (experimental settings) shift, \mathbb{P}^C with linear effect and “hard” shift, \mathbb{P}^D with non-linear effect and “soft” shift, \mathbb{P}^E with non-linear effect and “hard” shift. In Table 2 we summarize the different distribution parameters, also reporting the corresponding (exact) ATE value.

Table 2: Summary of the training and target experiment distributions.

In-Distribution		OoD (<i>linear effect</i>)		OoD (<i>non-linear effect</i>)	
		Soft shift	Hard shift	Soft shift	Hard shift
Distribution	\mathbb{P}^A	\mathbb{P}^B	\mathbb{P}^C	\mathbb{P}^D	\mathbb{P}^E
p_W	0.5	0.05	0.5	0.2	0.5
p_U	0.02	0.05	0.5	0.2	0.5
Randomized	True	False	False	False	False
Effect	Linear	Null	Null	Non Linear	Non Linear
ATE	1.5	0	0	1.2	0.75

These data generating processes represent fictitious experiments where we aim to quantify the effect of the pen color on the number to draw (if asked to pick one), relying on a machine learning model for handwritten-digits classification. Particularly, the shift in the treatment effect between training (ATE= 1.5) and target population (ATE $\in \{0, 0.75, 1.2\}$), may be interpreted as (i) testing a different, still homogeneous group of individuals, or (ii) varying the effect by changing the brand of the pens, unobserved variable (perfectly retrievable by the pen color on the training distribution), while keeping the same colors, reflecting the crucial challenges described in Section 2.1.

C.2 Analysis

We sampled 10 000 observations from \mathbb{P}^A to train a digits classifier (a Convolutional Neural Network) and tested it in PPCI in-distribution (10 000 more sample from \mathbb{P}^A) and out-of-distribution (zero-shot) on 10 000 observations for each \mathbb{P}^B , \mathbb{P}^C , \mathbb{P}^D , \mathbb{P}^E . We 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). Particularly, the proposed network 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 ten logits (one per digit) on which we apply a softmax activation to model the probabilities directly. Table 3 reports a full description of the training details for such network. Particularly we tuned the number of epochs in $\{10, 11, \dots, 50\}$ and learning rate in $\{0.01, 0.001, 0.0001, 0.00001\}$ by minimizing the Mean Squared Error on a small validation set (1 000 images) and finally retraining the model on the full training sample with the optimal parameters.

Table 3: Training details for the Convolutaional Neural Network training on CausalMNIST.

Hyper-parameters	Value
Loss	Cross Entropy
Learning Rate	0.0001
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.9, \epsilon = 10^{-8}$)
Batch Size	32
Epochs	40

670 For each learning objective, i.e., DERM (ours), ERM, vREx and IRM, we trained a model on the
671 training sample, imputed the outcome in each target sample and estimated the ATE via AIPW. Within
672 the AIPW estimator we used the XGBoost regressor for the outcome regression and the XGBoost
673 classifier to estimate the propensity score on observational studies (or vanilla sample mean on
674 randomized controlled trials since constant). For both IRM and vREx we replicated the same hyper-
675 parameter tuning of the invariant coefficient from our experiments on ISTAnt generalization, selecting
676 in both case the invariant coefficient $\lambda = 0.1$. We repeated each experiment 50 times, including
677 resampling the data, and bootstrapped the confidence interval of the ATE estimates. We run all the
678 analysis using 10GB of RAM, 8 CPU cores, and a single node GPU (NVIDIA GeForce RTX2080Ti).
679 The main bottleneck of each experiment is re-generating a new version of CausalMNIST from MNIST
680 dataset, and then training the model. The Prediction-Powered ATE estimation is significantly faster.
681 We estimate a total of 12 GPU hours to reproduce all the experiments described in this section.