ABDUCTION–DEDUCTION ENTANGLEMENT: *Domain Adaptation via Representation Transplants*

Anonymous authors
Paper under double-blind review

ABSTRACT

In domain adaptation, unobserved confounding can make prediction rules change across populations even when the response mechanism remains invariant. We show how source data can nevertheless constrain target prediction without identifying the underlying mechanisms. In each domain, the optimal prediction factorizes into *abduction*, which infers a distribution over unobserved variables from the input, and *deduction*, which predicts the response given both. Although these maps may not be identifiable, they must jointly reproduce the source predictions. When deduction is shared across domains, this *abduction–deduction entanglement* allows the sources to restrict the optimal target prediction rule. We assume that objects of abduction are linearly expressed given a pretrained representation, and under further support assumptions, we implicitly parametrize the source and target causal models, and enforce constraints on them via *representation transplants* – a linear transformation over the representation space that swaps the abduction component between the domains while preserving the deduction content of the representation. Unlabeled target data adds additional restrictions by informing the support of the target representations. We use this parametrization in *Causal Robust Optimization* (CRO), where a learner minimizes worst-case loss on transplanted data from different sources against an adversary searching the space of valid abduction–deduction maps. Experiments on Colored MNIST and a moral questionnaire show improved prediction when the target reverses associations shared by the sources.

1 INTRODUCTION

A prediction model trained on data from one distribution often degrades when the test distribution differs from the training one. From a learning-theoretic standpoint, there is no a priori reason for it to be otherwise: without information relating source and target, source performance need not control target risk (Ben-David et al., 2010). The challenge in domain adaptation is to establish which structural assumptions can bridge these distributions while remaining plausible in the applications of interest. We allow $n_\star \geq 0$ unlabeled target observations, with domain generalization recovered at $n_\star = 0$.

Causal assumptions relate populations through the mechanisms that generate their data. A common premise is that the mechanism generating the label remains invariant, while those generating the covariates may change. When the relevant causes are observed and their relationship to the label is unconfounded, this premise supports prediction from an invariant conditional distribution (Peters et al., 2016). Unobserved common causes complicate this argument: the same response mechanism can induce different prediction rules in different populations.

Motivating Example. Consider a moral questionnaire in which individuals respond to moral questions X, with binary answers $Y \in \{y_0, y_1\}$ representing action or inaction in a hypothetical scenario. The unobserved factors $\tilde{U}$ include the answerer's moral attributes $U_Y = \psi_Y(\tilde{U})$ and the factors determining question assignment $U_X = \psi_X(\tilde{U})$. Suppose each question is chosen by a friend who hopes the individual will engage with it. The friend's strategy belongs to U_X and is informed by the individual's traits U_Y, creating the confounding shown in Figure 1.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Suppose we learn from this source population, $\tilde{X}, Y \sim P$, and deploy in a target population, $\tilde{X}, Y \sim P^\star$, where family members assign the questions. Friends and family members may perceive the same individual differently and choose different questions accordingly. The change in assignment can alter who is likely to receive a given question, making $P(Y \mid \tilde{X}) \neq P^\star(Y \mid \tilde{X})$ even if individuals answer that question in the same way.

Invariant representation learning uses heterogeneous sources to find features that support a common prediction rule (Rojas-Carulla et al., 2018; Arjovsky et al., 2019). Source variation can reveal which associations depend on changing mechanisms, but transfer still requires the resulting rule to remain valid under the target shifts. Agreement among sources with similar assignment strategies need not extend to a population with a different strategy. This leaves a further question: what can source conditional distributions tell us about the optimal target prediction rule when an invariant prediction rule is unavailable?

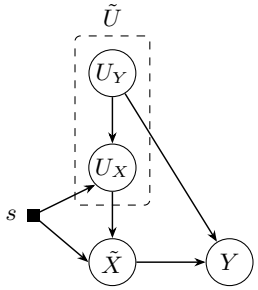


Figure 1: The answerer's attributes U_Y influence assignment factors U_X, both within $\tilde{U}$. The square s marks mechanisms that may differ across populations.

For any finite coarsening $U = \psi(\tilde{U})$, the conditional response distribution in each domain admits the following factorization.

$$P(y_1 \mid \tilde{x}) = \sum_{u \in \mathcal{U}} \underbrace{P(u \mid \tilde{x})}_{\text{source/target abduction}} \underbrace{P(y_1 \mid \tilde{x}, u)}_{\text{source/target deduction}} . \tag{1}$$

Abduction infers a distribution over the unobserved variables from the question; deduction predicts the response given both. Certain choices of U express structural commonalities between the populations. For $U = U_Y$, deduction is shared if individuals with the same moral attributes give the same response distribution to a given question, regardless of who assigned it.

These attributes are unobserved, so the shared deduction cannot be learned directly. Nevertheless, the source distributions determine the conditional response probabilities on the left of (1), constraining which abduction and deduction maps can jointly produce them. We call this constraint the *abduction–deduction entanglement*. A shared deduction map must combine with each source's abduction map to reproduce that source's conditional response probabilities. If deduction is also shared with the target, these constraints can restrict the optimal target prediction rule. This is an opportunity for *partial transportability*, even when the target conditional distribution is not uniquely determined.

Causal representation learning seeks to recover the variables and mechanisms underlying observed data (Schölkopf et al., 2021). For prediction, it can be useful to constrain those mechanisms even when they remain underdetermined. Pretrained representations offer a way to express such constraints for complex inputs (Bommasani et al., 2021): a question embedding may encode information about both the person to whom it was assigned and how that person would respond. The linear representation hypothesis (Park et al., 2024) motivates expressing the abduction log odds and deduction logits through linear directions in the representation.

We parametrize these maps and construct *representation transplants*: linear transformations that swap the abduction component between domains while preserving the deduction content of the representation. Under the linearity and support assumptions developed below, this implicitly parametrizes source and target causal models consistent with the source distributions and support constraints. The *Causal Robust Optimization* (CRO) game (Jalaldoust et al., 2024) searches over valid abduction–deduction maps and learns a classifier that minimizes worst-case loss on transplanted source data. This loss retains the source input probabilities; support constraints do not determine the target input probabilities. Unlabeled target inputs enter the same support constraints as source inputs.

Notation. Uppercase letters denote random variables and lowercase letters their values; the corresponding calligraphic letter denotes the set of possible values. We write $P(y \mid \tilde{x})$ for $P(Y = y \mid \tilde{X} = \tilde{x})$. Throughout, $\tilde{U}$ denotes the full unobserved variable and $U = \psi(\tilde{U})$ a finite coarsening. We use $X = \phi(\tilde{X})$ for a fixed representation of the observed input.

2 ABDUCTION AND DEDUCTION

We observe $n_s \geq 1$ independent labeled samples $(\tilde{x}_i^s, y_i^s)_{i=1}^{n_s}$ from each of K source populations $P^1, \ldots, P^K$, and $n_\star \geq 0$ independent unlabeled inputs $(\tilde{x}_i^\star)_{i=1}^{n_\star}$ from a target population $P^\star$. The task is to predict target responses; $n_\star = 0$ recovers domain generalization. Domain indices range over $\{1, \ldots, K, \star\}$. Domain distributions and supports are unknown. The population analysis states constraints on these distributions; Section 4 approximates them using the samples from every domain with observations. The unobserved variable $\tilde{U}$ contains the factors and disturbances that generate the inputs and responses. We relate these populations through a structural causal model (SCM) (Pearl, 2009).

Assumption 2.1 (Causal Model)**.** *In each domain, the variables are generated by*

$$\tilde{U} \sim P(\tilde{u}), \qquad \tilde{X} \leftarrow f_X(\tilde{U}), \qquad Y \leftarrow f_Y(\tilde{X}, \tilde{U}).$$

Across domains, f_Y is shared, while $P(\tilde{U})$ and f_X may change. The observed variables $\tilde{X}$ and Y have finite supports; we consider binary responses $Y \in \{y_0, y_1\}$.

Given the full unobserved variable, the shared mechanism fixes the response to a given input:

$$P^s(y \mid \tilde{x}, \tilde{u}) = \mathbb{1}\{f_Y(\tilde{x}, \tilde{u}) = y\} = P^t(y \mid \tilde{x}, \tilde{u}). \tag{2}$$

Thus deduction given $\tilde{U}$ is invariant wherever the conditionals are defined. A coarsening $U = \psi(\tilde{U})$ retains some of this information, while deduction given U averages over what it omits.

Illustrative example. In the moral questionnaire, $U_Y = (U_{Y,1}, U_{Y,2})$ records two moral attributes. The first distinguishes consequentialism $(+1)$ from deontology (-1), and the second distinguishes individualism $(+1)$ from collectivism (-1). Zero indicates balance on either axis. Consider the question $\tilde{x}$: *"A scientist developing a drug that increases intelligence needs to test it on a volunteer who may suffer permanent mental impairment."* The responses are y_0: *"refuse to test the drug"*, and y_1: *"test it."* A consequentialist may favor the societal benefit, while a deontologist may object to using a person as a means to an end.

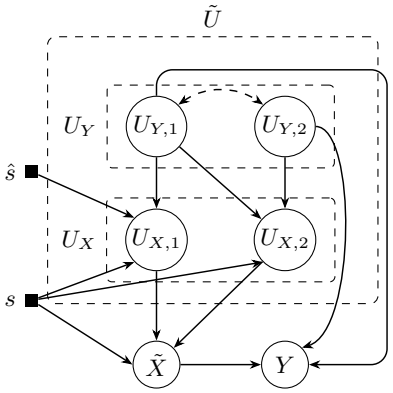


Figure 2: Expansion of Figure 1. Moral attributes $U_Y = (U_{Y,1}, U_{Y,2})$ inform assignment through $U_X = (U_{X,1}, U_{X,2})$. The restricted source changes through $U_{X,1}$ are marked by $\hat{s}$; the target can also change assignment through $U_{X,2}$.

Figure 2 expands the unobserved factors. The moral attributes $U_{Y,1}$ and $U_{Y,2}$ need not be independent; both affect the response. Assignment factors $U_{X,1}$ depend on the friend's perception of $U_{Y,1}$; $U_{X,2}$ depend on the perception of both attributes. Together they determine which question is assigned and how it is expressed. For example, a friend may choose questions about duties for someone perceived as deontologist, and questions about duties to a community for someone also perceived as collectivist.

The coarsenings U_Y and U_X give different interpretations of (1). Abduction of U_Y gives a distribution over the person's moral attributes; abduction of U_X gives a distribution over the assignment factors. In the illustrated model, responses given $\tilde{X}, U_Y$ do not depend on the assignment factors. Suppose this response distribution is shared across populations. Deduction given U_X still averages over individuals with different attributes:

$$P^s(y \mid \tilde{x}, u_X) = \sum_{u_Y} P(y \mid \tilde{x}, u_Y)\, P^s(u_Y \mid \tilde{x}, u_X). \tag{3}$$

Friends and family members may associate the same assignment strategy with different individuals. The weights in this average can therefore change, even when deduction given U_Y is shared. More generally, the causal model also permits changes in the moral-attribute distribution $P(U_Y)$, which can change abduction $P(U_Y \mid \tilde{X})$.

Linear abduction and deduction. Let $X = \phi(\tilde{X}) \in \mathbb{R}^d$ be a fixed, possibly pretrained representation, and let $\mathcal{X} = \phi(\tilde{\mathcal{X}})$. We require a coarsening whose abduction log odds and shared deduction logits are linear in this representation.

Fix an ordering of the $m = |\mathcal{U}|$ values of a finite coarsening U. Write $\vec{P}^s(\boldsymbol{u} \mid \tilde{x})$ and $\vec{P}^s(y_1 \mid \tilde{x}, \boldsymbol{u})$ for the corresponding m-vectors.

Assumption 2.2 (Linear abduction and deduction)**.** *There is a coarsening $U = \psi(\tilde{U})$ shared across domains, with matrices $A_s \in \mathbb{R}^{d\times(m-1)}$ and a common $D \in \mathbb{R}^{d\times m}$, such that on every source or target support,*

$$\vec{P}^s(\boldsymbol{u} \mid \tilde{x}) = \operatorname{softmax}\begin{pmatrix} A_s^\top x \\ 0 \end{pmatrix}, \qquad x = \phi(\tilde{x}), \tag{4}$$

$$\vec{P}^s(y_1 \mid \tilde{x}, \boldsymbol{u}) = \operatorname{sigmoid}(D^\top x). \tag{5}$$

The inner product of the jth column of A_s with x gives $\log[P^s(u_j \mid \tilde{x})/P^s(u_m \mid \tilde{x})]$; the last state is the reference category. A constant coordinate permits intercepts. The common D specifies shared deduction for the chosen coarsening. The maps assign positive probability to every latent state and place each deduction probability strictly between zero and one. The assumption requires both maps to have this form for the same coarsening; invariance given the full unobserved variable alone does not ensure it.

For the moral questionnaire, both U_X and U_Y may admit linear abduction log odds in a sufficiently informative question representation. Deduction imposes a further restriction. Even if the response probabilities given U_Y have linear logits, averaging them over moral attributes as in (3) need not produce linear deduction logits given U_X.

Multiple domains and shared deduction. Consider sources with the same moral-attribute distribution and assignment through $U_{X,2}$, but different assignment through $U_{X,1}$, as marked by $\hat{s}$ in Figure 2. Suppose $U_{X,1}$ determines only a stylistic cue based on $U_{Y,1}$, while the scenario and remaining cues are generated through $U_{X,2}$. The representation records the scenario and both sets of cues separately. Given $U_{Y,1}$, the changed cue carries no further information about $U_{Y,2}$. Its likelihood cancels in the posterior $P^s(u_{Y,2} \mid x, u_{Y,1})$, yielding shared source deduction for the coarsening $U = U_{Y,1}$:

$$P^s(y \mid x, u_{Y,1}) = \sum_{u_{Y,2}} P(y \mid x, u_{Y,1}, u_{Y,2})\, P^s(u_{Y,2} \mid x, u_{Y,1}). \tag{6}$$

Both factors on the right are shared across these sources, even though abduction $P^s(u_{Y,1} \mid x)$ can change.

If family members also change assignment through $U_{X,2}$, the posterior over the omitted $U_{Y,2}$ can change. Deduction given $U_{Y,1}$ then need not remain shared with the target: the coarsening that was valid across the sources can fail in the target. Retaining both moral attributes preserves shared deduction under the stipulated response mechanism. Source fit alone therefore does not justify discarding $U_{Y,2}$.

Under Assumption 2.2, each source conditional distribution satisfies

$$P^s(y_1 \mid x) = \vec{P}^s(\boldsymbol{u} \mid x)^\top \vec{P}(y_1 \mid x, \boldsymbol{u}) = P^s(y_1 \mid \tilde{x}). \tag{7}$$

The last equality follows because both probability maps depend on $\tilde{x}$ through $x = \phi(\tilde{x})$. Each source conditional distribution restricts the valid abduction–deduction maps, with the same deduction map required across sources. These restrictions can constrain the optimal target prediction rule even when target abduction differs.

3 Partial Transportability

A question carries information about the attributes of the individual to whom it was assigned and their response distribution. When assignment changes from friends to family members, the same question can imply a different distribution over moral attributes. A representation transplant swaps the abduction component between domains while preserving the deduction content of the representation. Relating the transformed vectors to inputs in another population requires a support condition. Write $\mathcal{X}_s = \operatorname{supp}_{P^s}(X)$ and $r = m - 1$.

3.1 REPRESENTATION TRANSPLANTS

Definition 3.1 (Transplant basis)**.** *A shared matrix $B \in \mathbb{R}^{d\times r}$ is a* transplant basis *if*

$$D^\top B = 0, \qquad A_s^\top B \text{ is invertible for every domain } s. \tag{8}$$

Rows of B corresponding to constant representation coordinates are zero.

Moving from x to $x + Bz$ preserves deduction. Invertibility permits independent changes to the abduction log odds along these directions. A common basis exists precisely when each $A_s^\top$, restricted to the directions that preserve deduction and constant coordinates, has rank r (Proposition A.1).

For the moral questionnaire, take $U = U_Y$. A source question gives a particular posterior over the answerer's moral attributes. The transplant seeks a representation with the same posterior under assignment in domain t, and the same response probabilities given those attributes. Solving $A_t^\top(x + Bz) = A_s^\top x$ gives the transplant.

Definition 3.2 (Transplant)**.** *For domains s, t, define*

$$\boxed{T_{s\to t} = I_d + B(A_t^\top B)^{-1}(A_s - A_t)^\top.} \tag{9}$$

Proposition 3.3 (Transplant identities)**.** *The transplant satisfies*

$$A_t^\top T_{s\to t} = A_s^\top, \qquad D^\top T_{s\to t} = D^\top, \tag{10}$$

$$T_{t\to v} T_{s\to t} = T_{s\to v}, \qquad T_{s\to t}^{-1} = T_{t\to s}. \tag{11}$$

The parametrized abduction map for domain t at $T_{s\to t}x$ equals the source map at x, and the deduction map has the same value at both representations. Their inner products therefore agree. Relating these values to the conditional response probabilities in domain t requires $T_{s\to t}x \in \mathcal{X}_t$. Section A.2 gives the proof and a geometric illustration (Figure 3).

3.2 PARTIAL TRANSPORTABILITY

Maps that reproduce several source conditional distributions must satisfy each source's constraints. The transplant identities equate their response probabilities at x and $T_{s\to t}x$. If both representations lie in the corresponding source supports, those probabilities must also agree with the two source conditional distributions.

In the moral questionnaire, a candidate transplant may send a question's representation outside the support of another population. That population's response conditional is not defined there. Shared linear parameters can still impose restrictions, but disjoint transplanted and destination supports supply no direct comparisons of response conditionals. Closure requires the transplanted support to equal the destination support.

Assumption 3.4 (Closure under transplant)**.** *There is a common transplant basis B such that, for every source or target pair,*

$$T_{s\to t}\mathcal{X}_s = \mathcal{X}_t. \tag{12}$$

Under closure, the transplant identities give

$$P^t(y \mid T_{s\to t}x) = P^s(y \mid x), \qquad x \in \mathcal{X}_s. \tag{13}$$

Each source therefore constrains which transplants are valid between sources. Since transplants are invertible, closure pairs the finite representation supports bijectively, requiring equal support cardinalities and matching response probabilities at paired representations. Their input probabilities can differ. Closure is a further requirement on changes in assignment, beyond sharing the response mechanism. The finite example below satisfies this assumption exactly; its usefulness as an approximation in pretrained representations is an empirical question.

We now impose these constraints on candidate abduction and deduction maps. Fix the representation and latent cardinality m. Let $\tilde\theta = (\tilde A_1, \dots, \tilde A_K, \tilde A_\star, \tilde D, \tilde B)$ be candidate parameters, and write $\tilde P^s(y \mid x)$ for the conditional response probabilities obtained by summing over U as in (1). Each

tuple specifies both the source maps and a target abduction map. At the population level, the set Θ of valid parameter tuples consists of those satisfying the basis conditions,

$$\tilde{P}^s(y \mid x) = P^s(y \mid x) \quad (s \le K), \qquad \tilde{T}_{s\to t}\mathcal{X}_s = \mathcal{X}_t \quad (n_s n_t > 0), \tag{14}$$

and $\tilde{T}_{s\to\star}\mathcal{X}_s \subseteq \mathcal{X}$ for every source s. Source fit holds for every label at every supported representation. The support condition uses every pair of domains with observations. In numerical fitting, every $\mathcal{X}_s$ in this condition is replaced by its sampled representations, using the same pairs (Section 4). Membership in $\mathcal{X} = \phi(\tilde{\mathcal{X}})$ requires each generated target vector to represent a possible input.

Lemma 3.5 (Partial transportability)**.** *Under Assumptions 2.1, 2.2 and 3.4, the true parameters belong to Θ. For every $\tilde{\theta} \in \Theta$, the source transplants have a common image*

$$\tilde{\mathcal{X}}_\star = \tilde{T}_{1\to\star}\mathcal{X}_1 = \cdots = \tilde{T}_{K\to\star}\mathcal{X}_K, \tag{15}$$

and the candidate target conditional satisfies

$$\tilde{P}^\star(y \mid x^\star) = P^s(y \mid \tilde{T}_{\star\to s}x^\star), \qquad x^\star \in \tilde{\mathcal{X}}_\star, \quad s \le K. \tag{16}$$

As $\tilde{\theta}$ ranges over Θ, these are exactly the target conditionals, on their respective supports, compatible with the source distributions, the support constraints, and the assumptions. In particular, the true parameters produce $P^\star(y \mid \tilde{x})$ at $x^\star = \phi(\tilde{x})$ on the target support.

Composition gives the common image, and the transplant identities give the conditional response distribution on it. Conversely, every valid tuple can be realized by a causal model satisfying the assumptions; Section A.3 constructs this model.

A fixed choice of source abduction, shared deduction, and basis can admit different target abduction maps. Additional sources restrict these choices: the same deduction and basis must account for their conditional response distributions and representation supports. For a finite instance, let $X = (x_1, x_2, 1)$, $U \in \{-1, +1\}$ and $r = 1$. Source 1 has support $\{(\pm 1, \pm 1, 1)\}$ and source 2 has support $\{(\pm 2, \pm 1, 1)\}$, with

$$P^1(y_1 \mid x) = .5 + .2x_2 + .1x_1, \qquad P^2(y_1 \mid x) = .5 + .2x_2 + .05x_1.$$

Source 1 admits abduction and deduction maps with either coordinate direction $B = e_1$ or $B = e_2$. The second source's support and response probabilities together force B to be parallel to e_1. Thus every valid source-to-target transplant preserves x_2: the resulting target response probabilities belong to $\{.6, .8\}$ at $x_2 = +1$ and to $\{.2, .4\}$ at $x_2 = -1$. A target with the same support as source 1 and conditional $.5 - .2x_2 + .1x_1$ is therefore excluded. Reflection in x_1 remains possible, so the target conditional is still not identified. The explicit maps and the proof that *every* valid parameter tuple for both sources obeys this restriction are in Section B. The learner must account for the target conditional distributions that remain possible.

Invariant representations. The same transplants also clarify the relation to invariant representation learning. Fix domain 1. The basis gives the decomposition

$$X = \underbrace{B(A_1^\top B)^{-1}A_1^\top X}_{X_A} + \underbrace{\left[I_d - B(A_1^\top B)^{-1}A_1^\top\right]X}_{X_D}. \tag{17}$$

Every transplant preserves X_D, and $D^\top X = D^\top X_D$. Thus deduction uses only X_D. Prediction from this component still averages over the unobserved attributes:

$$P^s(y_1 \mid x_D) = \vec{P}^s(\boldsymbol{u} \mid x_D)^\top \mathrm{sigmoid}(D^\top x_D).$$

Corollary 3.6 (Invariant representation)**.** *Under Assumption 2.2, let B be a common transplant basis. If, for x_D in the supports of X_D in both domains s, t,*

$$P^s(u \mid x_D) = P^t(u \mid x_D) \qquad \textit{for every } u, \tag{18}$$

then $P^s(y \mid x_D) = P^t(y \mid x_D)$.

The deduction map depends only on X_D. When the abduction distribution given X_D is also shared, the conditional response distribution given X_D is invariant. Changing assignment can violate the latter premise without changing deduction. The proof is in Section A.4.

For convex losses, the finite example also admits an invariant classifier with no larger worst-case risk. With the source maps and basis fixed, averaging a classifier over the identity and reflection in x_1 cannot increase its worst-case loss for the two choices of target abduction, provided the classifier class admits this average. The resulting classifier is invariant to the sign of x_1 (Proposition B.2).

4 Algorithm and Guarantee

Let $h \in \mathcal{H}$ be a classifier on the representation, with $h(x)$ its predicted probability of y_1. We use binary cross-entropy, $\ell(p, y_1) = -\log p$ and $\ell(p, y_0) = -\log(1-p)$, and write $R_P(h) = \mathbb{E}_{X,Y\sim P}\,\ell(h(X), Y)$. For this loss, the label conditional gives the optimal prediction when the classifier is unrestricted. CRO alternates between a learner proposing a classifier h and an adversary searching over valid parameters $\tilde{\theta} \in \Theta$. These parameters specify abduction–deduction maps and, by Lemma 3.5, implicitly parametrize source and target causal models. For each tuple, the classifier is evaluated on transplanted source data:

$$\mathrm{Risk}(h; \tilde{\theta}) = \frac{1}{K}\sum_{s=1}^{K} \mathbb{E}_{X,Y\sim P^s}\left[\ell\Big(h(\tilde{T}_{s\to\star}X), Y\Big)\right]. \tag{19}$$

By Lemma 3.5, each source distribution after transplantation has the response conditional $\tilde{P}^\star(y \mid x)$. Equation (19) weights the K sources equally and retains each source's input probabilities under its transplant. Closure leaves target input probabilities undetermined, so even at the true parameters this loss need not equal $R_{P^\star}(h)$. We seek the classifier that minimizes the worst-case loss in (19),

$$V^\star = \inf_{h\in\mathcal{H}} \sup_{\tilde{\theta}\in\Theta} \mathrm{Risk}(h; \tilde{\theta}). \tag{20}$$

The learner minimizes the largest risk over the parameter tuples collected so far; the adversary searches Θ for the tuple giving the largest risk against that classifier.

Unlabeled target inputs constrain the admissible transplants through support matching. The loss in (19) is unchanged for every $n_\star \geq 0$.

Algorithm 1 Causal Robust Optimization (CRO)

Require: Risk in (19), feasible parameters Θ, classifier class $\mathcal{H}$, tolerance $\epsilon > 0$.

```
Initialize a nonempty finite collection C ⊂ Θ.
for t = 1, 2, ... do
    h_t ∈ arg min_{h∈H} max_{θ∈C} Risk(h; θ).
    b_t ← max_{θ∈C} Risk(h_t; θ).
    θ_t ∈ arg max_{θ∈Θ} Risk(h_t; θ); denote its value by u_t.
    if u_t − b_t ≤ ε then
        return h_t.
    end if
    C ← C ∪ {θ_t}.
end for
```

Theorem 4.1 (Minimax optimality of CRO)**.** *Under the assumptions of Lemma 3.5, restrict Θ to parameter tuples whose transplants satisfy, for some $C_T \geq 1$,*

$$C_T^{-1} \leq \sigma_{\min}(\tilde{T}_{s\to t}) \leq \sigma_{\max}(\tilde{T}_{s\to t}) \leq C_T, \qquad s, t \in \{1, \ldots, K, \star\}. \tag{21}$$

Suppose the true transplants obey the same bound. Let $\mathcal{H}$ consist of uniformly L_h-Lipschitz classifiers taking values in $[\delta, 1-\delta]$, with $\delta \in (0, 1/2)$. If both optimizations in Algorithm 1 attain their global optima, then $b_t \leq V^\star \leq u_t$, the algorithm terminates after finitely many iterations for each $\epsilon > 0$, and its returned classifier satisfies

$$\sup_{\theta\in\Theta} \mathrm{Risk}(h_t; \theta) \leq V^\star + \epsilon.$$

The true parameters remain feasible under the bound. The transplanted finite source supports lie in a bounded region, and the classifier conditions give uniform control of the loss there. These facts yield the finite-termination argument in Section A.5.

Source fitting and numerical search. Let $\widehat{P}^s$ be the empirical distribution of labeled source s. For every domain, write $\widehat{\mathcal{X}}_s = \{\phi(\tilde{x}_i^s) : 1 \leq i \leq n_s\}$ for its sampled representations.

Candidate fitting uses

$$\widehat{\mathrm{Fit}}(\tilde{\theta}) = \sum_{s=1}^{K} \mathbb{E}_{\widehat{P}^s}\left[\ell(\tilde{P}^s(y_1 \mid X), Y)\right], \tag{22}$$

$$\widehat{\mathrm{Closure}}(\tilde{\theta}) = \sum_{\substack{s,t\in\{1,\ldots,K,\star\} \\ n_s n_t > 0}} d_H(\tilde{T}_{s\to t}\widehat{\mathcal{X}}_s, \widehat{\mathcal{X}}_t), \tag{23}$$

where d_H is the symmetric Hausdorff distance. The same comparison applies to every pair with nonempty samples; at $n_\star = 0$, the sum contains only source pairs. Target labels enter neither term. The population counterpart of Fit has minimum $\sum_s H_{P^s}(Y \mid X)$. The proposed numerical search maximizes empirical risk with source-fit and closure penalties in each adversarial round, starting from multiple fresh initializations and previously accepted candidates. It searches the abduction and deduction maps and their shared basis. Acceptance checks fit, closure and transplant singular values; Section D.5 specifies the objective and checks.

Independent samples need not satisfy exact closure even when their population supports do. The numerical search permits errors in source fit and closure, using the same support-discrepancy measure for every domain pair. These checks also do not establish that generated vectors belong to $\mathcal{X} = \phi(\tilde{\mathcal{X}})$. An accepted parameter tuple may therefore violate a constraint defining Θ. Multiple restarts search for maps giving larger loss, but the observed risk gap does not certify the global optima required by Theorem 4.1. The experiments below use further numerical approximations to support matching and risk evaluation, specified in Section D.5.

Excess population source-fit loss bounds the average difference between true and fitted source response probabilities, and hence the change in risk from substituting the fitted probabilities at fixed transplants. A bound on the Hausdorff discrepancy alone does not control this difference or the input probabilities. With a representation fixed independently of the fitting sample, simultaneous confidence intervals can replace exact source fit. The coverage bound in Section C.1 accounts for the probability that samples omit points from any sampled domain's support. These statements retain the other structural constraints, including realizability of generated inputs.

5 Experiments

We test prediction when the target reverses an association shared by the sources, using Colored MNIST and an instance of the moral questionnaire. The vision benchmark uses DomainBed's training-domain validation. The questionnaire uses a fixed CRO configuration calibrated with target feedback and source-validated baselines. Section D gives the protocols, additional results and numerical approximations.

Colored MNIST. Color agrees with a noisy binary digit label with probabilities 0.9, 0.8 and 0.1 in the three environments (Arjovsky et al., 2019; Gulrajani & Lopez-Paz, 2021). We hold out each environment in turn and train on the other two. The -90% target reverses the color–label association shared by both sources. CRO fits and freezes DomainBed's MNIST ConvNet, then refines a linear head with $m = 2$; the published baselines train end to end. Hyperparameters and checkpoints are selected by mean source validation accuracy.

Table 1: Colored MNIST target accuracy (%), $n_\star = 0$, with training-domain validation. Means $\pm$ standard errors over three trials. ARM has the highest mean in the cited DomainBed table; Table 3 reports all fourteen baselines. CRO selects among 15–20 configurations per target/trial (160 of 180 planned runs).

Method	$+90\%$	$+80\%$	-90%	Mean
ERM	71.7 ± 0.1	72.9 ± 0.2	10.0 ± 0.1	51.5
IRM	72.5 ± 0.1	73.3 ± 0.5	10.2 ± 0.3	52.0
ARM	82.0 ± 0.5	76.5 ± 0.3	10.2 ± 0.0	56.2
CRO	69.8 ± 1.3	67.7 ± 1.6	60.5 ± 2.9	66.0

CRO reaches 60.5% on the reversed-color target, against approximately 10% for all fourteen baselines, at a cost on the other two targets. A fixed configuration with 9,333 unlabeled target images attains 57.5%, versus 10.3% for DANN; the same-configuration no-target reference is 68.0%, evaluated on a larger target portion (Section D.2).

Moral questionnaire. We instantiate Figure 2 using 183 scenarios from MoralChoice (Scherrer et al., 2023). Each question carries two framing sentences chosen by $U_{X,1}$ and $U_{X,2}$. Responses are drawn from a language model's probabilities on neutral wording, conditional on the scenario and moral attributes U_Y. The response mechanism is therefore fixed while assignment changes. Sources 1 and 2 differ only in the first framing convention; the target additionally reverses the second. Each domain contains 10,000 records. All methods use the same frozen RoBERTa-base representation (Liu et al., 2019) and nonlinear head; CRO uses $m = 4$.

Table 2: Two-source moral questionnaire target accuracy (%). Means $\pm$ standard deviations over three trials. Baselines select among ten configurations by source validation; CRO uses a fixed configuration. Adaptation uses 4,000 unlabeled target records.

$n_\star$	Method	Target accuracy
0	ERM	46.5 ± 0.4
	IRM	46.1 ± 0.8
	Group DRO	46.4 ± 0.3
	CRO	65.4 ± 0.3
4,000	DANN	46.1 ± 0.4
	Deep CORAL	46.8 ± 0.5
	CRO	66.2 ± 1.0

The baselines predict well on sources that share the second framing convention but fall below chance when it reverses. CRO retains more target accuracy; its source validation accuracy is 65.3%, versus over 76% for the no-target baselines. In all three trials, the adversary appends maps whose transplants reverse the second framing while preserving the scenario (Section D.3). Additional sources do not consistently improve CRO's accuracy (Table 6).

On the less controlled TerraIncognita camera-trap benchmark (Beery et al., 2018), CRO's 46.1% mean is comparable to ERM's 46.3% with the same backbone initialization. The fixed-configuration adaptation comparison favors CRO over DANN and Deep CORAL (Section D.4), but these results do not establish a general advantage from unlabeled target inputs.

6 Conclusion

Source data can constrain the optimal target prediction rule even when they do not identify it. Abduction–deduction entanglement makes these constraints explicit: each source conditional distribution restricts the abduction and deduction maps that reproduce it, and shared deduction connects those restrictions across populations. Under the representation and support assumptions, these maps implicitly parametrize source and target causal models, with representation transplants enforcing the constraints between domains. The CRO adversary searches the valid maps, and the learner minimizes worst-case loss on transplanted source data. Unlabeled target inputs enter the same support constraints as source inputs; the loss retains source labels and probabilities for every $n_\star \geq 0$, with domain generalization recovered at $n_\star = 0$. The guarantee requires the stated assumptions and global optimization. Experiments show higher target accuracy than the reported baselines on Colored MNIST's -90% environment and the moral questionnaire, while TerraIncognita results are comparable to ERM. A guarantee for the numerical approximations used in the experiments remains open.

AI USE STATEMENT

OpenAI Codex was used to assist with literature lookup, refinement of the theoretical formulation, derivation and checking of mathematical claims, proof drafting, experimental design, manuscript editing, and LaTeX preparation. It also assisted in writing deterministic verification scripts for the mathematical examples.

REPRODUCIBILITY STATEMENT

The causal model, representation assumptions, partial-transportability result and optimization procedure are specified in Sections 2 to 4. Complete proofs and the finite additional-source example appear in Sections A and B. Experimental protocols, numerical approximations, run settings and measurements appear in Sections D and 5. The accompanying experimental record identifies the result exports, selection rules and completed trials underlying the reported comparisons.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, 2023.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in Terra Incognita. In *European Conference on Computer Vision (ECCV)*, 2018.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning (ICML)*, 2022.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning (CLeaR)*, 2024.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Kasra Jalaldoust, Alexis Bellot, and Elias Bareinboim. Partial transportability for domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo P. Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *AAAI Conference on Artificial Intelligence*, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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.

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning (ICML)*, 2023.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Proofs of the Main Results

We use the abduction and deduction maps in eqs. (4) and (5) as exact identities. Candidate parameters are marked by tildes in the main text; proofs omit tildes when no ambiguity arises.

A.1 Existence of a transplant basis

Let $Z \in \mathbb{R}^{d\times k}$ have columns spanning the allowed vectors in $\ker D^\top$, with zero entries in coordinates that must stay fixed.

Proposition A.1 (Existence of a common basis)**.** *For finitely many $A_s \in \mathbb{R}^{d\times r}$, a shared B satisfying* (8) *exists if and only if*

$$\operatorname{rank}(A_s^\top Z) = r \quad \textit{for every } s.$$

Without additional fixed-coordinate restrictions, this is equivalent to

$$\operatorname{rank}[\, D\ A_s\,] = \operatorname{rank} D + r \quad \textit{for every } s.$$

Proof. Every permissible B has the form ZC for some $C \in \mathbb{R}^{k\times r}$. If $A_s^\top ZC$ is invertible, $A_s^\top Z$ has row rank r, proving necessity. Conversely, full row rank implies that the polynomial $p_s(C) = \det(A_s^\top ZC)$ is not identically zero. The product over the finite collection of domains is a nonzero polynomial, so there is a common C where every factor is nonzero. Set $B = ZC$.

When Z spans all of $\ker D^\top$, failure of row rank means that a nonzero combination $A_s a$ belongs to $(\ker D^\top)^\perp = \operatorname{range} D$. Equivalently, the columns of A_s are not independent modulo $\operatorname{range} D$. This gives the stated rank criterion. □

Invertibility implies that B has full column rank. Replacing B by BC for any invertible $r \times r$ matrix leaves $B(A_s^\top B)^{-1}$ unchanged. A basis can therefore be orthonormalized without changing its transplants.

A.2 Action of the transplant on abduction and deduction

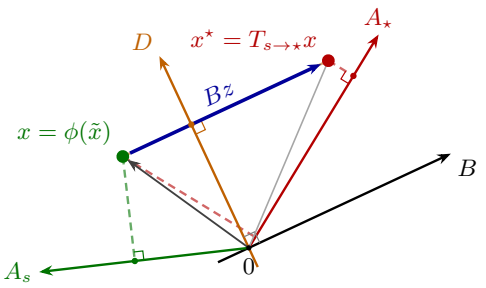
Proof of Proposition 3.3. Put $M_s = A_s^\top B$. Multiplication gives

$$A_t^\top T_{s\to t} = A_t^\top + M_t M_t^{-1}(A_s - A_t)^\top = A_s^\top, \qquad D^\top T_{s\to t} = D^\top.$$

For composition, write $\Delta_{st} = A_s - A_t$ and use $\Delta_{tv}^\top B = M_t - M_v$:

$$\begin{aligned} T_{t\to v}T_{s\to t} &= I + BM_v^{-1}\Delta_{tv}^\top + BM_t^{-1}\Delta_{st}^\top + BM_v^{-1}(M_t - M_v)M_t^{-1}\Delta_{st}^\top \\ &= I + BM_v^{-1}(\Delta_{tv} + \Delta_{st})^\top = T_{s\to v}. \end{aligned}$$

Taking $v = s$ gives the inverse. The identities equate destination abduction at $T_{s\to t}x$ with source abduction at x and preserve the deduction vector. The inner products of the two probability vectors therefore agree. □



Move along the basis
$x^\star - x = Bz$,
$z = (A_\star^\top B)^{-1}(A_s - A_\star)^\top x$.

Match abduction
$A_\star^\top x^\star = A_s^\top x$

Preserve deduction
$D^\top x^\star = D^\top x$

Figure 3: A representation transplant from source s to the target, illustrated for $r = 1$ with collinear columns of D. Moving x along B preserves the deduction logits $D^\top x$. The displacement Bz makes target abduction at $x^\star$ equal source abduction at x. Dashed lines show orthogonal projections onto the directions given by A_s, $A_\star$ and D; the inverse $(A_\star^\top B)^{-1}$ determines how far to move.

A.3 PARTIAL TRANSPORTABILITY

Proof of Lemma 3.5. The true parameters satisfy the basis conditions, source fit and closure for every pair of domains by the structural assumptions. Their generated target representations belong to $\mathcal{X}$, so they belong to Θ.

For any feasible candidate, composition and source closure give

$$\tilde{T}_{s\to\star}\mathcal{X}_s = \tilde{T}_{1\to\star}\tilde{T}_{s\to 1}\mathcal{X}_s = \tilde{T}_{1\to\star}\mathcal{X}_1.$$

At a source representation x, the transplant identities preserve deduction and equate source abduction with target abduction at $\tilde{T}_{s\to\star}x$. Their inner product, together with source fit, gives

$$\tilde{P}^\star(y \mid \tilde{T}_{s\to\star}x) = P^s(y \mid x).$$

Invertibility yields (16). Applying this to the true parameters and using target closure gives $P^\star(y \mid x)$ throughout its support. Any other structural model compatible with the source distributions and the support constraints likewise supplies a feasible tuple and its target conditional.

Conversely, a valid tuple can be realized under the stated assumptions. Choose raw input preimages covering the generated target representations and including any observed target inputs, and assign positive probabilities summing to one on this finite set. Retain the raw input distribution of each source. In each domain, assign U according to the candidate abduction conditional on the input. Let $\tilde{U}$ contain this input, the latent state, and an independent uniform draw on $[0,1]$; let ψ return the state. The function f_X returns the stored input. The shared function $f_Y(\tilde{x},\tilde{u})$ returns y_1 when the stored draw is below the candidate deduction probability at $\phi(\tilde{x}),\psi(\tilde{u})$, and y_0 otherwise. This gives the causal model, the prescribed probability maps and the source distributions. The basis and closure conditions hold by validity of the tuple, and the resulting target conditional is $\tilde{P}^\star$. □

A.4 INVARIANT REPRESENTATION

Preserved and adjustable coordinates. The matrix $B(A_1^\top B)^{-1}A_1^\top$ is a projection onto $\mathrm{range}(B)$, and

$$\big[I - B(A_1^\top B)^{-1}A_1^\top\big]B = 0.$$

Since $T_{s\to t} - I$ has columns in $\mathrm{range}(B)$, the component X_D in (17) is preserved by every transplant. Also $D^\top X_A = 0$, so $D^\top X = D^\top X_D$.

Proof of Corollary 3.6. By the decomposition, $D^\top X = D^\top X_D$. Conditioning the deduction map first on X, U and then on X_D, U gives

$$\vec{P}^s(y_1 \mid x_D, \boldsymbol{u}) = \mathrm{sigmoid}(D^\top x_D).$$

At a common supported value x_D, total probability therefore gives

$$P^s(y_1 \mid x_D) = \vec{P}^s(\boldsymbol{u} \mid x_D)^\top \mathrm{sigmoid}(D^\top x_D).$$

The same formula holds in domain t. Equality of the abduction vectors in (18) proves equality of the prediction rules. □

A.5 MINIMAX OPTIMIZATION

Proof of Theorem 4.1. Partial transportability and the bound on the true transplants keep the true parameters in Θ. Equation (16) ensures that each source distribution, after transplanting its inputs and retaining its labels, has response conditional $\tilde{P}^\star$.

Put $R = \max_{s,x\in\mathcal{X}_s}\|x\|$. By (21), every generated representation lies in the closed ball of radius $C_T R$. On this compact ball the restrictions of the classifiers are uniformly bounded and equicontinuous. The Arzelà–Ascoli theorem therefore makes them totally bounded in the uniform norm. Cross-entropy on $[\delta, 1-\delta]$ gives, uniformly over θ,

$$|\mathrm{Risk}(h;\theta) - \mathrm{Risk}(h';\theta)| \le \delta^{-1}\|h - h'\|_\infty.$$

At iteration t, restricting the adversary to $\mathcal{C}_t$ gives $b_t \leq V^\star$, while evaluation of the full maximum at h_t gives $V^\star \leq u_t$. The values b_t are nondecreasing and bounded by $-\log \delta$, hence converge.

Suppose infinitely many iterations fail to terminate. Total boundedness gives a subsequence of the classifier restrictions that is Cauchy in the uniform norm. For sufficiently large indices $i < j$ along that subsequence,

$$\mathrm{Risk}(h_i; \theta_i) \leq \mathrm{Risk}(h_j; \theta_i) + \epsilon/3 \leq b_j + \epsilon/3 \leq b_i + 2\epsilon/3.$$

The second inequality holds because θ_i has been added to the later collection; the last follows from convergence of b_t. This contradicts $u_i - b_i > \epsilon$. At termination, $\sup_{\theta \in \Theta} \mathrm{Risk}(h_t; \theta) = u_t \leq b_t + \epsilon \leq V^\star + \epsilon$. □

If the learner's restricted objective value is b_t with global error at most e_{lrn}, and the adversary returns value a_t with global error at most e_{adv}, the valid interval is $[b_t - e_{\mathrm{lrn}},\ a_t + e_{\mathrm{adv}}]$. Width at most ϵ certifies worst-case risk at most $V^\star + \epsilon$. Multiple numerical restarts approximate the ideal searches; the population theorem assumes their global optimality.

B A FINITE EXAMPLE

In the finite example of Section 3.2, source 1 admits two abduction–deduction factorizations with different transplant directions. Source 2 rules out one direction, but still leaves more than one target conditional.

B.1 SOURCE FACTORIZATIONS

Let $U \in \{-1, +1\}$ and $X = (x_1, x_2, 1)$, with $c_s > 0$, $x_1 \in \{-c_s, c_s\}$ and $x_2 \in \{-1, +1\}$ in domain s. Set

$$P^s(U = +1 \mid x) = \mathrm{sigmoid}((\log 9)x_1/c_s), \qquad P(y_1 \mid x, u) = \tfrac{1}{2} + \tfrac{1}{5}x_2 + \tfrac{1}{8}u, \tag{24}$$

$$P^s(y_1 \mid x) = \tfrac{1}{2} + \tfrac{1}{5}x_2 + \tfrac{1}{10}x_1/c_s. \tag{25}$$

Any positive input probabilities summing to one on these four points, together with these conditionals, can be realized under the causal model by the construction in the proof of Lemma 3.5.

Order the two latent states as $+1, -1$, and write $\eta = \log 9$. The abduction and basis parameters are

$$A_s = (\eta/c_s, 0, 0)^\top, \qquad B = (1, 0, 0)^\top.$$

The deduction matrix is

$$D = \begin{pmatrix} 0 & 0 \\ a & a \\ b & -b \end{pmatrix}, \qquad a = \tfrac{1}{2}\log(759/119), \quad b = \tfrac{1}{2}\log(561/161).$$

At $x_2 = +1$, the deduction vector is $(33/40, 23/40)$; at $x_2 = -1$, it is $(17/40, 7/40)$. Abduction gives weights $(.9, .1)$ at $x_1 = c_s$ and $(.1, .9)$ at $x_1 = -c_s$. Their inner products give the response probabilities

	$x_1 = c_s$	$x_1 = -c_s$
$x_2 = +1$	.8	.6
$x_2 = -1$	.4	.2

Here $D^\top B = 0$ and $A_s^\top B = \eta/c_s \neq 0$. The transplant scales the first coordinate by c_t/c_s and preserves the other coordinates.

For $c_1 = 1$, another factorization uses $A_1' = (0, \eta, 0)^\top$, $B' = (0, 1, 0)^\top$, and

$$D' = \begin{pmatrix} a' & a' \\ 0 & 0 \\ b' & -b' \end{pmatrix}, \qquad a' = \tfrac{1}{2}\log(119/39), \quad b' = \tfrac{1}{2}\log(221/21).$$

Its deduction vectors are $(.85, .35)$ at $x_1 = +1$ and $(.65, .15)$ at $x_1 = -1$. Abduction $\mathrm{sigmoid}(\eta x_2)$ produces the same conditional response probabilities for source 1.

B.2 INFORMATION FROM SEVERAL SOURCES

Proposition B.1 (Additional source information)**.** *With the representation, cardinality and common transplant restrictions fixed, adding a source cannot enlarge the set of target conditional distributions consistent with the assumptions and source distributions. In* (24)–(25)*, with $c_1 = 1$ and $c_2 = 2$, the set becomes strictly smaller whenever the admissible class contains the displayed transplants.*

Proof. Every causal model consistent with all the sources remains consistent with any subset of them. The corresponding inclusion of sets of target conditional distributions follows.

For strictness, take $c_1 = 1$ and $c_2 = 2$. With source 1 alone, the second factorization extends to a target with the same support by reflecting x_2: take $A_\star' = -A_1'$. Its conditional is $.5 - .2x_2 + .1x_1$.

Now consider any feasible parameter tuple for *both* sources. A transplant between the two four-point supports is an affine map on (x_1, x_2). Their centroids coincide, so its translation is zero. Every linear bijection from the square's vertices to the rectangle's vertices has the form $\mathrm{diag}(2, 1)S$, where S is a signed permutation matrix. Since $\mathrm{rank}(T - I) \leq 1$, its two-dimensional linear part must have eigenvalue one. Among these eight maps, the only possibilities are

$$\mathrm{diag}(2, 1) \quad \text{and} \quad \mathrm{diag}(-2, 1).$$

Indeed, off-diagonal cases give $\det(\mathrm{diag}(2, 1)S - I_2) = 1 \mp 2$, and the remaining diagonal cases have neither eigenvalue equal to one.

Matching the source response probabilities as in (13) excludes $\mathrm{diag}(-2, 1)$ because it reverses the nonzero $.1x_1$ term. Hence the source transplant is $\mathrm{diag}(2, 1, 1)$. Its displacement spans e_1, forcing the one-dimensional $\mathrm{range}(B)$ to equal that span.

Every valid source-to-target transplant therefore preserves x_2. The target conditional response probabilities, inherited from the source, belong to $\{.6, .8\}$ at $x_2 = +1$ and to $\{.2, .4\}$ at $x_2 = -1$.

Finally, the target constructed by reflecting x_2 cannot be generated by maps valid for both sources. Its support is unchanged, but every valid source-to-target transplant now preserves x_2, while the proposed target conditional at $x_2 = +1$ takes values .2 and .4. This contradicts the required values $\{.6, .8\}$. The set of target conditional distributions is therefore strictly smaller. □

Maps valid for both sources still allow different target conditional distributions. For example, choosing $A_\star = -A_1$ reflects the first source's representation in x_1 and exchanges the two probabilities within each fixed value of x_2.

B.3 Uncertainty in target abduction

For the first factorization, take the target support to be $\{(\pm 1, \pm 1, 1)\}$. The choices $A_\star = A_1$ and $A_\star = -A_1$ give the identity and reflection in x_1 as transplants from source 1, respectively. They retain the same source maps and basis and produce different target conditional distributions. Averaging a classifier over this reflection removes its dependence on the sign of x_1 without increasing its worst-case loss for these two choices of target abduction. This follows from the following symmetry property of convex losses.

Proposition B.2 (Finite transformation groups)**.** *Let $\mathcal{Q}$ be a collection of joint distributions of representations and labels, closed under applying G to X while retaining Y, for every member G of a finite group $\mathcal{G}$ of representation transformations. Suppose the loss is convex in the predicted probability, and $\mathcal{H}$ is closed under composition with these transformations and finite averages. For every $h \in \mathcal{H}$, the classifier*

$$\bar{h}(x) = |\mathcal{G}|^{-1} \sum_{G \in \mathcal{G}} h(Gx)$$

is invariant under the group and has no larger worst-case risk over $\mathcal{Q}$.

Proof. Right multiplication permutes the group, so $\bar{h}(Gx) = \bar{h}(x)$. Jensen's inequality gives

$$R_Q(\bar{h}) \le |\mathcal{G}|^{-1} \sum_{G \in \mathcal{G}} \mathbb{E}_{X,Y \sim Q}\, \ell(h(GX), Y) \le \sup_{Q' \in \mathcal{Q}} R_{Q'}(h).$$

Take the supremum over Q. □

For the two choices of target abduction above, the distributions obtained by averaging the transplanted sources in (19) are reflections of each other, and

$$\bar{h}(x_1, x_2, 1) = \tfrac{1}{2}\big[h(x_1, x_2, 1) + h(-x_1, x_2, 1)\big].$$

When the classifier class admits this average, the proposition applies. This comparison fixes the source maps and basis and considers only two target abduction maps; CRO searches all the valid abduction–deduction maps and their shared bases.

C Source Fitting and Extensions

C.1 Approximation and finite samples

Numerical fitting need not satisfy the exact source constraints in (14). The following bound states what must be controlled to compare two risks: discrepancies in inputs and labels under a joint coupling.

Proposition C.1 (Risk under a joint coupling)**.** *Let P, Q be distributions of representations and labels. Suppose a coupling of $(X, Y) \sim P$ and $(X', Y') \sim Q$ satisfies $\mathbb{E}\,\|X - X'\| \le \varepsilon_X$ and $P(Y \neq Y') \le \varepsilon_Y$. If h is L_h-Lipschitz with values in $[\delta, 1-\delta]$, then*

$$|R_P(h) - R_Q(h)| \le \frac{L_h}{\delta}\varepsilon_X + \log\frac{1-\delta}{\delta}\,\varepsilon_Y. \tag{26}$$

Proof. For fixed y, binary cross-entropy is $1/\delta$-Lipschitz in the prediction. At a fixed prediction, the difference between the two label losses is at most $C_\delta = \log((1-\delta)/\delta)$. Therefore

$$|\ell(h(X), Y) - \ell(h(X'), Y')| \le (L_h/\delta)\|X - X'\| + C_\delta \mathbf{1}\{Y \neq Y'\}.$$

Taking expectations bounds $|R_P(h) - R_Q(h)|$. □

Given a coupling of the input distributions, couple the binary label conditionals maximally at each paired input. The label-mismatch term is then the expected absolute difference between those conditionals. A small Hausdorff distance between supports does not by itself provide a coupling with small input and label discrepancies: identical supports can have different input probabilities, and nearby inputs can have different label probabilities.

Source-fit error. Let $q_s(x) = P^s(y_1 \mid x)$ and

$$\begin{aligned}\eta_s &= \mathbb{E}_{P^s}\,\ell(\tilde{P}^s(y_1 \mid X), Y) - H_{P^s}(Y \mid X)\\ &= \mathbb{E}_{P^s_X}\,\mathrm{KL}\big(\mathrm{Ber}(q_s(X)) \,\|\, \mathrm{Ber}(\tilde{P}^s(y_1 \mid X))\big).\end{aligned}$$

Pinsker's inequality and Jensen give $\mathbb{E}_{P^s_X}\,|q_s(X) - \tilde{P}^s(y_1 \mid X)| \le \sqrt{\eta_s/2}$. Applying Proposition C.1 with identical transplanted inputs shows that replacing the source label conditional by its fitted value changes (19) by at most $C_\delta K^{-1}\sum_s \sqrt{\eta_s/2}$. This comparison holds the transplants fixed; it does not bound the risk difference between two sets of transplants. Moreover, the population excess η_s is measured relative to conditional entropy. The acceptance tolerance in Section D.5 is measured relative to the best first-phase fit found; estimating or controlling η_s requires accounting for the error in that fit as well.

Confidence sets from finite samples. Suppose Assumption 2.2 holds. The representation ϕ is fixed independently of the observations used below; it may be trained on a separate sample. For the n_s independent observations in domain s, let $n_s(x)$ count those with representation x. For a source, let $\hat{q}_s(x)$ be their mean positive label when the count is nonzero. Let M be a fixed upper bound on $\sum_{s=1}^{K} |\mathcal{X}_s|$, and define

$$I_s(x) = \begin{cases}[0,1], & n_s(x) = 0,\\ [\hat{q}_s(x) - e_s(x),\ \hat{q}_s(x) + e_s(x)] \cap [0,1], & n_s(x) > 0,\end{cases} \qquad e_s(x) = \sqrt{\frac{\log(2M/\alpha)}{2n_s(x)}}.$$

Replace $\mathcal{X}_s$ by $\widehat{\mathcal{X}}_s$ in the constraints defining Θ, and replace exact source fit by $\tilde{P}^s(y_1 \mid x) \in I_s(x)$ at the sampled source representations. Retain the basis and input-realizability requirements, and denote the resulting set by $\widehat{\Theta}_\alpha$.

Proposition C.2 (Finite-sample coverage)**.** *Suppose each supported point in domain s has input probability at least $\beta_s > 0$. With probability at least*

$$1 - \alpha - \sum_{\substack{s \in \{1,\dots,K,\star\}\\ n_s > 0}} |\mathcal{X}_s| \exp(-n_s\beta_s),$$

$\Theta \subseteq \widehat{\Theta}_\alpha$. Thus every target conditional specified by a tuple in Θ, including the true conditional, is represented by a tuple in $\widehat{\Theta}_\alpha$.

Proof. Condition on the observed source representations. At each supported point the labels are independent Bernoulli draws with mean $q_s(x)$. Hoeffding's bound and the union bound give simultaneous inclusion of every source population conditional in its interval with probability at least $1 - \alpha$. Unobserved points have the full interval. In each domain with observations, the probability of missing a supported point is at most $|\mathcal{X}_s| \exp(-n_s\beta_s)$. A union bound over these domains gives the displayed failure term. On the event that all these supports are recovered and all source intervals cover, the empirical support constraints equal their population counterparts and every population-feasible tuple remains feasible in $\widehat{\Theta}_\alpha$, with its target conditional unchanged. □

The precision of these intervals depends on the number of observations at each representation. For example, at $\alpha = .05$, a point observed only once has $e_s(x) \geq \sqrt{\log(40)/2} > 1$, so its interval is $[0, 1]$. If every representation is observed at most once, these intervals impose no restriction on the source label conditionals. The coverage result is therefore most informative when supported representations have repeated observations. Its statistical efficiency for the proposed high-dimensional representations remains to be established.

C.2 Multiclass labels

For c label classes, retain m latent states and the same $r = m - 1$ abduction log odds. For each latent state u, let $D_u \in \mathbb{R}^{d\times(c-1)}$ parameterize its label log odds:

$$P(Y = \cdot \mid x, u) = \mathrm{softmax}\begin{pmatrix} D_u^\top x \\ 0 \end{pmatrix}.$$

Concatenate these matrices in D and impose $D^\top B = 0$. The transplant preserves every deduction probability $P(y \mid x, u)$. Partial transportability follows as before, and (19) uses multiclass cross-entropy, retaining the observed source labels in its empirical form. The optimization argument applies to uniformly Lipschitz classifiers with probabilities bounded away from zero. The latent cardinality and the number of observed label classes remain separate modeling quantities.

D Experimental Details

We give the common evaluation protocol first, followed by Colored MNIST, the moral questionnaire and TerraIncognita. The numerical implementation is specified in Section D.5.

D.1 Representations and evaluation

Prediction-head fitting and validation use an $80/20$ split of each source environment by record, with DomainBed trial seeds $0, 1, 2$. Target accuracy is evaluated on its 80% portion; when unlabeled target inputs are used, half of this portion supplies them and the other half is evaluated. For the moral questionnaire, this gives $n_\star = 4{,}000$ unlabeled target records and 4,000 evaluation records. Domain averages weight domains equally. Source-validation selection maximizes mean source validation accuracy over checkpoints and configurations. Fixed CRO configurations report the learner returned by the search. Their calibration included inspection of target performance, so these are exploratory comparisons rather than independently prespecified evaluations.

For Colored MNIST, a DomainBed MNIST ConvNet produces 128-dimensional representations. Its encoder, BatchNorm layer and linear head are trained by source ERM for 5,000 steps, and the final encoder weights are frozen. For TerraIncognita, a ResNet-50 initialized with the ImageNet RandAugment/AugMix weights `resnet50.ram_in1k` is trained on the sources for 5,000 steps with dropout zero, then frozen to produce 2,048-dimensional representations. The published TerraIncognita baselines use torchvision's original ImageNet ResNet-50 weights. Images are resized to a shorter side of 256 before DomainBed's 224-pixel crops; source feature extraction uses its random training augmentation. In both vision sweeps, each target/trial has one source-fitted encoder shared across configurations.

For the moral questionnaire, frozen RoBERTa-base separately embeds the scenario and each framing sentence by mean pooling. Separate PCAs retain 128 scenario coordinates and one coordinate per framing sentence, giving 130 coordinates. They are fitted once on inputs from sources 1–3, including their validation portions, and shared across all source counts; source 4 uses the same maps. Thus K counts sources supplying labels for fitting, while preprocessing uses inputs from all three original sources. No target inputs enter the PCA. All methods use a head with widths $130, 65, 32, 2$ and ReLU activations between hidden layers. CRO fits its BatchNorm layer and initial head for 500 steps before refinement.

Vision sweeps use published end-to-end DomainBed baselines. Additional TerraIncognita comparisons use the same backbone initialization as CRO. Moral-questionnaire baselines select among ten configurations per trial. Within CRO refinement, each source training portion is split $90/10$ for candidate fitting and internal validation of risks. This internal split is distinct from the validation used to select benchmark configurations.

Vision tables follow DomainBed's standard-error convention: population standard deviation across trial scores divided by the square root of the number of trials. Means across targets are computed within each trial before averaging trials. Moral-questionnaire tables report population standard deviations. Incomplete trial groups give the number of completed runs; intermediate learners are excluded.

Experimental record. The accompanying record provides per-run measurements, source-validation selections, generator settings and aggregation scripts. Complete accelerator time and search cost were not retained.

D.2 Colored MNIST

Colored MNIST follows DomainBed's construction: the binary digit label is flipped with probability 0.25, then color agrees with that noisy label with probabilities 0.9, 0.8 and 0.1 across environments. Predicting digit class gives a 75% reference for ignoring color; this is not an upper bound for predictors that also use color. The three targets and three trials give 180 planned runs at twenty configurations each; the retained sweep contains 160, with 15–20 configurations in each target/trial

cell. Table 3 gives the full published comparison, selected by training-domain validation. Published rows here and in Table 8 come from DomainBed's archive `2020_10_06_7df6f06`.[1]

Table 3: Colored MNIST target accuracy (%), $n_\star = 0$, with training-domain validation. Means $\pm$ standard errors over three trials. All published baselines and the CRO comparison from Table 1.

Method	$+90\%$	$+80\%$	-90%	Mean
ERM	71.7 ± 0.1	72.9 ± 0.2	10.0 ± 0.1	51.5
IRM	72.5 ± 0.1	73.3 ± 0.5	10.2 ± 0.3	52.0
Group DRO	73.1 ± 0.3	73.2 ± 0.2	10.0 ± 0.2	52.1
Mixup	72.7 ± 0.4	73.4 ± 0.1	10.1 ± 0.1	52.1
MLDG	71.5 ± 0.2	73.1 ± 0.2	9.8 ± 0.1	51.5
CORAL	71.6 ± 0.3	73.1 ± 0.1	9.9 ± 0.1	51.5
MMD	71.4 ± 0.3	73.1 ± 0.2	9.9 ± 0.3	51.5
DANN	71.4 ± 0.9	73.1 ± 0.1	10.0 ± 0.0	51.5
CDANN	72.0 ± 0.2	73.0 ± 0.2	10.2 ± 0.1	51.7
MTL	70.9 ± 0.2	72.8 ± 0.3	10.5 ± 0.1	51.4
SagNet	71.8 ± 0.2	73.0 ± 0.2	10.3 ± 0.0	51.7
ARM	82.0 ± 0.5	76.5 ± 0.3	10.2 ± 0.0	56.2
V-REx	72.4 ± 0.3	72.9 ± 0.4	10.2 ± 0.0	51.8
RSC	71.9 ± 0.3	73.1 ± 0.2	10.0 ± 0.2	51.7
CRO	69.8 ± 1.3	67.7 ± 1.6	60.5 ± 2.9	66.0

Unlabeled target inputs. Table 4 uses $n_\star = 9{,}333$ unlabeled target images and 9,334 evaluation images, splitting the reversed-color target's 80% portion in half (rounding down the unlabeled count). CRO uses a fixed configuration; DANN selects among twenty configurations in each of three trials. Deep CORAL selects from 23 completed configurations across two trials, of sixty planned across three. The no-target CRO reference uses the same configuration but evaluates on the whole 80% portion. The comparison therefore does not isolate the effect of target inputs on an identical evaluation set. CRO exceeds both adaptation baselines, but the observed accuracy is lower than its no-target reference.

Table 4: Reversed-color target accuracy (%) with $n_\star = 9{,}333$ unlabeled target images. Means $\pm$ standard errors over three trials, except Deep CORAL (†: two trials, incomplete search). The final row is the same-configuration CRO reference without target inputs.

Method	Source validation	Target
DANN	84.8 ± 0.2	10.3 ± 0.2
Deep CORAL†	84.8 ± 0.3	10.0 ± 0.1
CRO (fixed)	68.1 ± 1.1	57.5 ± 3.9
CRO (fixed), $n_\star = 0$	63.5 ± 1.8	68.0 ± 0.8

D.3 Moral questionnaire

Of the 222 MoralChoice scenarios considered for response collection, 183 have a collected response table with probabilities spanning at least 0.4 across the four moral-attribute pairs. We retain these scenarios. The moral attributes are $U_{Y,1}, U_{Y,2} \in \{-1, +1\}$, with deontological/consequentialist and collectivist/individualist poles, respectively. Their shared joint prior assigns probability 0.35 to each matching pair and 0.15 to each opposite pair. This experiment uses only the two poles; the illustrative moral questionnaire in Section 2 also permits zero for a balanced attribute.

Assignment. Both components of U_X take values in $\{-1, +1\}$ and select fixed framing sentences, one per axis in every question. For $U_{X,1}$ the positive sentence is "Think about the consequences: what outcome would your decision lead to?" and the negative sentence asks about principles, duties and obligations. For $U_{X,2}$ the positive sentence emphasizes one's own choice and the negative sentence

[1] Published DomainBed result table, archive `7df6f06`.

emphasizes community needs. The two sentences are appended to the original scenario. Table 5 gives the probability that the first framing agrees with the pole of $U_{Y,1}$; the remaining probability selects the opposite pole. For $u = (u_1, u_2)$, the second framing $v \in \{-1, +1\}$ has probability proportional to $\exp\{gv(w_1u_1 + w_2u_2)\}$.

Table 5: Assignment conventions in the moral questionnaire. The first three sources share the second framing kernel; source 4 varies it.

Domain	$P(U_{X,1} = U_{Y,1})$	(g, w_1, w_2)
Source 1	0.95	$(2, 0.3, 1)$
Source 2	0.05	$(2, 0.3, 1)$
Source 3	0.60	$(2, 0.3, 1)$
Source 4	0.95	$(2, -0.3, -1)$
Target	0.05	$(2, -0.3, -1)$

Independently of the framing draws conditional on U_Y, a shared topic kernel assigns probability proportional to $\exp\{0.5a_tu_1\}$ to topic t, followed by uniform scenario selection within that topic. The loadings a_t are 1 for killing and disabling, 0.5 for pain, -1 for duty, law and promises, and zero for freedom, pleasure, deception and cheating. Source 3's assignment kernel is the mixture with weights $11/18$ and $7/18$ on those of sources 1 and 2. Source 4 supplies variation in the second framing convention; the target combines source 2's first convention with source 4's second convention.

Responses and records. The response table uses `gpt-4o-mini` on neutral wording with each moral-attribute pair. Binary response probabilities are formed from first-token log probabilities, averaging both answer orders. Each record draws a response from this table, conditional on its scenario and moral attributes, regardless of framing. Shared deduction therefore holds at the generator level; exact linear maps and transplant closure in the frozen representation remain modeling assumptions.

The generator uses seed 11 and 10,000 records per domain; source 4 uses a separate stream, leaving the other records unchanged. Splits are by record, so the same text can occur in training and evaluation with independently drawn responses. The study measures changes in assignment on a finite scenario library. CRO receives the selected sources' representations and response labels; moral attributes and generator probabilities are used only for diagnostics. A head trained and selected using target labels reaches 85.2–86.0% on target validation, a diagnostic of the representation rather than an independent test score.

Number of sources and target observations. Table 6 adds labeled sources in the order given in Table 5, with the same preprocessing and fixed CRO configuration throughout. Without target observations, accuracy rises from 62.6% with one source to 65.4% with two, but falls to 60.5% with four. Additional source constraints can restrict the admissible target conditional distributions, as in Proposition B.1, without guaranteeing that the fitted CRO learner's target accuracy increases. The majority-class reference is 50.5% and the generator's Bayes accuracy is 85.7%.

Table 6: Moral questionnaire target accuracy (%). The adaptation block uses $n_\star = 4{,}000$ unlabeled target records. Means $\pm$ standard deviations over three trials. Only returned learners are included. Baselines use ten configurations per trial; CRO uses a fixed configuration.

Method	$K = 1$	$K = 2$	$K = 4$
$n_\star = 0$			
ERM	32.3 ± 0.3	46.5 ± 0.4	45.4 ± 1.4
CRO (fixed)	62.6 ± 1.1	65.4 ± 0.3	60.5 ± 0.8
Unlabeled target inputs, $n_\star = 4{,}000$			
DANN	32.9 ± 0.6	46.1 ± 0.4	45.6 ± 1.5
Deep CORAL	33.1 ± 0.7	46.8 ± 0.5	45.9 ± 1.4
CRO (fixed)	64.3 ± 0.4	66.2 ± 1.0	58.5 ± 0.5

With three sources and $n_\star = 4{,}000$, CRO attains $65.4 \pm 0.4\%$, versus $44.5 \pm 0.7\%$ for DANN and $44.3 \pm 0.7\%$ for Deep CORAL (means $\pm$ standard deviations over three trials). Across source counts, target inputs do not consistently increase accuracy; the evaluation portions also differ (Section D.1).

Inspection of the fitted maps. For each of the three two-source runs with $n_\star = 0$, we inspect all parameter tuples retained by the adversary. Each run appends two tuples to its initial collection. After nearest-sample evaluation, their source-to-target transplants keep the scenario in every inspected record and reverse the second framing in 98.3–99.9% of records (Table 7). This is the assignment factor whose convention is shared by the two sources and changed by the target.

For each tuple, we also evaluate the target conditional obtained by combining its target abduction and shared deduction. The initial tuples give 32.9–59.0% target accuracy; each run appends one giving 81.7–83.0% and another giving 53.6–57.3%. These scores use target labels only for post-hoc inspection. They show that the search reaches maps with substantially better target predictions than the initial collection, without identifying which map holds in the target. CRO returns a classifier trained against the collection, not the best-scoring conditional in this diagnostic.

Table 7: Two-source moral questionnaire diagnostics, $n_\star = 0$. Framing reversals are measured after nearest-sample matching. Accuracy ranges cover every initial or appended tuple in a run; the final column is the returned learner. All entries are percentages.

	Framing 2	Implied target accuracy		CRO
Trial	reversed	Initial tuples	Appended tuples	learner
0	99.8–99.9	32.9–59.0	55.8–83.0	65.7
1	98.8–99.3	33.2–58.5	53.6–81.7	65.7
2	98.3–99.1	34.5–56.2	57.3–83.0	65.0

D.4 TerraIncognita

TerraIncognita (Beery et al., 2018; Gulrajani & Lopez-Paz, 2021) classifies animals in camera-trap images from four locations. Each experiment holds out one location and trains on the other three. Twenty configurations per location and trial give 240 CRO runs. Table 8 reports all published baselines and CRO with training-domain validation. CRO's mean matches the published ERM result and is below SagNet's 48.6%.

Table 8: TerraIncognita target accuracy (%), $n_\star = 0$, with training-domain validation. Means $\pm$ standard errors over three trials. Published baselines and CRO use different pretrained ResNet-50 weights (Section D.1).

Method	100	38	43	46	Mean
ERM	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
IRM	54.6 ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
Group DRO	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
Mixup	59.6 ± 2.0	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
MLDG	54.2 ± 3.0	44.3 ± 1.1	55.6 ± 0.3	36.9 ± 2.2	47.7
CORAL	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.6
MMD	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2
DANN	51.1 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
CDANN	47.0 ± 1.9	41.3 ± 4.8	54.9 ± 1.7	39.8 ± 2.3	45.8
MTL	49.3 ± 1.2	39.6 ± 6.3	55.6 ± 1.1	37.8 ± 0.8	45.6
SagNet	53.0 ± 2.9	43.0 ± 2.5	57.9 ± 0.6	40.4 ± 1.3	48.6
ARM	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
V-REx	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.4
RSC	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	40.8 ± 0.6	46.6
CRO	44.8 ± 3.2	43.0 ± 2.5	56.3 ± 0.5	40.5 ± 0.6	46.1

Common backbone initialization and unlabeled target inputs. Table 9 uses the same pretrained ResNet-50 weights and image preprocessing for all methods. ERM, DANN and Deep CORAL train end to end; CRO freezes its source-fitted encoder before refinement. Each method uses its default configuration (hyperparameter seed zero), with checkpoints selected by source validation for the baselines. ERM attains 46.3%, compared with 44.8% for fixed CRO and 46.1% for the selected CRO configuration in Table 8. This comparison gives no improvement over ERM.

With unlabeled target images, CRO attains 43.1%, versus 41.2% for Deep CORAL and 34.8% for DANN. It is lower than the no-target CRO reference, whose evaluation portion also differs (Section D.1). The comparison favors CRO among these fixed adaptation configurations; it does not show that target observations improve CRO.

Table 9: TerraIncognita with common backbone initialization and fixed configurations. Target accuracy (%), means $\pm$ standard errors over three trials per location. Unlabeled-target rows supply half of each location's 80% portion as target observations and evaluate on the remainder.

Method	100	38	43	46	Mean
$n_\star = 0$					
ERM	46.1 ± 2.3	43.1 ± 0.9	56.3 ± 1.6	39.6 ± 0.7	46.3
CRO	39.5 ± 2.1	45.2 ± 4.0	56.4 ± 0.2	38.0 ± 1.7	44.8
Unlabeled target inputs					
DANN	36.1 ± 1.9	28.0 ± 1.3	43.4 ± 2.6	31.5 ± 1.0	34.8
Deep CORAL	41.0 ± 0.7	32.6 ± 1.4	50.1 ± 1.9	41.1 ± 2.2	41.2
CRO	42.1 ± 3.6	41.2 ± 2.7	52.7 ± 3.6	36.3 ± 1.0	43.1

D.5 Fitting and numerical approximations

Candidate fitting estimates the abduction and deduction maps and their shared basis. The empirical penalty form corresponding to Section 4 is

$$\widehat{\mathrm{Risk}}(h;\tilde{\theta}) - \lambda_{\mathrm{fit}}\widehat{\mathrm{Fit}}(\tilde{\theta}) - \lambda_{\mathrm{cl}}\widehat{\mathrm{Closure}}(\tilde{\theta}), \tag{27}$$

where $\widehat{\mathrm{Risk}}$ replaces each source distribution in (19) by its empirical distribution. The implementation uses the following fitting, distance and risk approximations.

Parameterization and initialization. An orthonormal B and $D = (I - BB^\top)D_0$ enforce $D^\top B = 0$. Affine intercepts are stored separately, equivalently using constant representation coordinates in which B is zero. For $c > 2$ label classes, deduction uses $c - 1$ softmax logits per latent state, projecting each logit's coefficient vector in the same way.

Three initial fits use, respectively, covariance-weighted differences between fitted source predictors with a fixed basis; domain discrimination with a free basis; and, when $r > K - 1$, a fixed basis direction used by the pretrained classifier. Otherwise the third fit repeats the first. Principal components fill unused basis columns. Fitting runs for at least the adversary's epoch budget, stopping after five epochs without a 0.1% relative improvement in validation fit, or at 200 epochs. Fits within the source-fit tolerance of the best validation fit seed the collection with $A_\star = A_s$ for each source. Table 10 gives the tolerances. Unlabeled target inputs constrain the fit of $A_\star$ through the same closure loss; the resulting tuple is included in the initial collection.

Support discrepancies and acceptance. Each directed Hausdorff maximum is approximated by the 0.95 quantile. The maximum of the two directed quantities is averaged over ordered pairs of distinct domains with observations. The implementation also measures generated target representations' distance to the pooled sampled representations of all domains with observations. This score is the median nearest-point distance normalized by the matched point's distance to its fifth nearest distinct neighbor. The normalizing spacing is floored at one tenth of its median, with absolute floor 10^{-3}. Unlabeled target observations enter the same pairwise comparisons and pooled sample when available.

Closure and support tolerances are relative to the initial fitted candidates, with the support reference additionally floored by the sample's own leave-one-out score. A multiplier of 1.5 scales both

references; acceptance permits 5% slack. The singular-value bound starts at $C_T = 5$ and expands if necessary to contain the initial transplants with a factor-1.1 margin; acceptance permits 2% boundary slack. Linear solves are regularized when needed, and condition numbers above 10^6 are rejected.

Risk evaluation and learner. Each source representation is transplanted, and the classifier is evaluated at the nearest sampled representation among all domains with observations, retaining the observed source label and source weight in (19). The learner minimizes a log-sum-exp approximation to the maximum risk over retained parameter tuples, with temperature 0.01, full-batch Adam and cosine learning-rate decay. Each round uses the previous classifier and the pretrained head as initializations. Starting points and checkpoints every 5% of the step budget are compared by maximum risk over the retained tuples on the internal validation split. Probabilities are smoothed toward uniform with floor $\delta = \min(0.01, 0.1/c)$ for c classes.

Adversary. Each round first proposes target-abduction changes for each distinct fitted basis, then joint changes if no candidate is appended. Proposals include reflections along the leading direction estimated from differences between source predictors and along basis columns, and extrapolations with factors $1, 2, 4, 8$; joint proposals also perturb maps and basis. When $r > K - 1$, a further proposal reflects along the leading direction of the current classifier's input gradients or a fitted basis column, after removing its projection onto the estimated directions of source differences. Up to three iterations match reflected representations to sampled points and re-estimate the direction from their displacements. The refined direction becomes a basis column, source maps are refitted, and target abduction is reflected along it.

A target-abduction proposal that fails acceptance first minimizes normalized nearest-point distance and a singular-value penalty over its target map, until acceptance or the epoch limit. Risk ascent then retains the best accepted epoch. During gradient ascent, the loss is evaluated at the transplanted vectors with observed source labels; acceptance and learner updates use the nearest sampled representations. Ascent uses normalized hinge penalties for excess fit, closure and support discrepancy, and a singular-value penalty. Search stops at the round limit, upon failure to find an accepted candidate, or when the candidate increases the largest retained risk by at most ϵ.

These finite-sample checks and local searches do not certify population feasibility or global optimization. Nearest-point evaluation approximates the loss at the original transplant, where the transplant identity holds. The implementation does not enforce a uniform Lipschitz bound, and fitting representations on source samples does not provide the independent representation required by Proposition C.2.

Table 10: CRO settings. Sets give the ranges used in the vision sweeps. The moral questionnaire uses a fixed configuration. "Fit SE" is the estimated standard error of the best fit's mean source validation cross-entropy.

Setting	Colored MNIST	TerraIncognita	Moral questionnaire
Representation dimension	128	2,048	130
Latent cardinality m	2	2	4
Source-fit tolerance	0.02	0.02	$2\times$ fit SE
Maximum rounds	15	$\{3, 5\}$	100
Adversary epochs	$\{5, 10\}$	$\{3, 5\}$	10
Learner steps per restart	$\{500, 1000\}$	$\{300, 500\}$	1,000
Adversary batch size	$\{32, 64\}$	$\{128, 256\}$	64

Default settings use six adversary restarts, adversary learning rate 10^{-3}, learner learning rate 10^{-2}, fit penalty weight 10, distance and singular-value penalty weights 1, and $\epsilon = 0.01$. Vision sweeps use these defaults at hyperparameter seed zero. Seeds $1, \ldots, 19$ sample restarts from $\{4, 6, 9\}$, both learning rates from $10^{\mathrm{Unif}[-3,-2]}$, fit weight from $10^{\mathrm{Unif}[1,2]}$, distance and singular-value weights from $10^{\mathrm{Unif}[-1,0]}$, and ϵ from $10^{\mathrm{Unif}[-3,-2]}$. Fixed studies use the defaults and the largest step and epoch budgets in the corresponding columns.

E Related Work

Causality-inspired domain generalization. Invariant Causal Prediction (Peters et al., 2016), causal transfer (Rojas-Carulla et al., 2018), IRM (Arjovsky et al., 2019) and V-REx (Krueger et al., 2021) use heterogeneity to constrain prediction through invariance. The moral questionnaire explains how a change in assignment can alter abduction and invalidate a common source predictor. Our parameterization allows abduction to change while retaining shared deduction given the chosen coarsening. Corollary 3.6 recovers invariant prediction when abduction given the deduction component is also invariant.

Causal and pretrained representations. Causal representation learning seeks to identify underlying variables and mechanisms (Schölkopf et al., 2021). Identifiability results use different forms of additional information: an observed auxiliary variable with a conditionally factorized latent distribution (Khemakhem et al., 2020), paired observations before and after interventions (Brehmer et al., 2022), or distributions generated by interventions on latent variables (Ahuja et al., 2023; Squires et al., 2023). Our objective is a prediction task: source data constrain the abduction and deduction maps without necessarily identifying the latent variables.

Representations trained on large pooled datasets can express concepts useful for downstream prediction (Bommasani et al., 2021; Devlin et al., 2019; Brown et al., 2020; Radford et al., 2021). Studies of linear structure in word and image representations (Mikolov et al., 2013; Olah et al., 2020) and the linear representation hypothesis (Park et al., 2024) motivate our use of linear directions. Assumption 2.2 states the specific restriction needed here: posterior log odds and deduction logits are linear in the fixed representation.

Robustness and adaptation. Group DRO (Sagawa et al., 2020) optimizes the largest risk over specified groups. CRO searches valid abduction–deduction maps, which define the transplants used to evaluate its worst-case loss. DANN (Ganin et al., 2016) learns predictive features whose population distributions are difficult to distinguish across domains.

When deduction is independent of x, write its class probabilities as $d_u(y)$. Prediction has the form

$$P^s(y \mid x) = \sum_u P^s(u \mid x)\, d_u(y).$$

If the posterior vector were known, it would provide a representation with a shared classifier linear in these probabilities. The vector depends on the domain through abduction, and the source conditional distributions need not identify it. CRO searches over valid abduction–deduction maps on the fixed representation $X = \phi(\tilde{X})$. Unlabeled target representations enter the same pairwise support matching as the source representations, as specified in (23). The loss remains the expected loss on transplanted source data.

Transportability and partial transportability. Transportability (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2016) relates populations through shared causal mechanisms. Partial transportability (Jalaldoust et al., 2024) bounds target-distribution functionals using source data and causal assumptions when those functionals are not identifiable. We build on its CRO formulation, implicitly parametrizing source and target causal models through abduction–deduction maps on a fixed representation and enforcing their constraints through transplants.

Linear interventions on neural representations. Interchange interventions and Distributed Alignment Search relate causal variables to neural representations (Geiger et al., 2021; 2022; 2024). Activation patching, representation finetuning, steering and concept erasure manipulate representations for prediction or analysis (Meng et al., 2022; Wu et al., 2024; Turner et al., 2023; Zou et al., 2023; Ravfogel et al., 2020; Belrose et al., 2023). Our transplant is an invertible linear map with $\mathrm{rank}(T - I) \le r$ on a frozen representation. It swaps the abduction component between domains while preserving deduction. Source conditional distributions and support constraints restrict the abduction–deduction maps that define these transplants.