MODEL-AGNOSTIC INVARIANT CAUSAL PREDICTION TESTING IN LEARNED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying invariant causal features in learned representations by modern machine learning models is central to robust generalization across environments and to the development of trustworthy AI systems. Based on invariant prediction, we develop a model-agnostic test of whether the conditional distribution of a response given a candidate feature subset remains invariant across environments. Under conditional invariance, the adjusted predictiveness scores coincide, even when the working prediction class is misspecified. We establish asymptotic normality of the component statistics and a chi-square null limit for their squared cyclic aggregate, providing asymptotic Type I error control. Power depends on whether the chosen working class and score distinguish changes in the conditional mechanism. Our approach can be readily integrated with pretrained foundation models. Simulation studies show that the proposed test reliably identifies causal features in complex settings and remains robust to severe model misspecification. Experiments using the Contrastive Language-Image Pretraining (CLIP) model further illustrate that our approach uncovers meaningful invariant features and improves generalization in downstream tasks.

1 INTRODUCTION

Modern machine learning models support scientific and predictive tasks across many fields (Wang et al., 2023; Moor et al., 2023; Davies et al., 2021). However, features learned from observational data may encode spurious associations that change across environments, limiting generalization (Beery et al., 2018; Ovadia et al., 2019; Xu et al., 2024; Bayat et al., 2025). This raises a practical question: *which features in a learned representation retain a stable relationship with the response?*

Causal mechanisms can remain stable under changes elsewhere in a system. Identifying these mechanisms is a goal in applications of machine learning (Chen et al., 2022; Zhang et al., 2023a; Chen et al., 2024). Invariant causal prediction (ICP) uses stability across environments to obtain confidence statements about causal features under structural assumptions (Peters et al., 2016). This connection also motivates work on trustworthy AI (Rawal et al., 2025), scientific interpretability (Nussberger et al., 2022; Wu et al., 2023), and causal fairness (Nilforoshan et al., 2022; Plecko & Bareinboim, 2023). For learned features, statistical invariance and causal interpretation are closely related but distinct: a stable concept score may be a proxy for a cause.

Embedding models (Darrin et al., 2024) map inputs such as text (Devlin et al., 2019) and images (Dosovitskiy et al., 2020) into numerical representations. Pretrained models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) also connect these representations to human-specified concepts. Their predictive performance alone does not establish which coordinates have invariant conditional relationships with a response. The size and overparameterization of modern models further complicate feature interpretation (Nakkiran et al., 2021; Kaplan et al., 2020; Brown et al., 2020). We therefore study invariance within a supplied representation. Candidate coordinates, generated by human experts and/or modern AI models, are fixed before statistical inference, and observations from different environments allow us to test their relationship with the response.

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.

1.1 Our Contributions

We propose the **M**odel-**A**gnostic **I**nvariant **C**ausal **P**rediction (MAICP) test. For a fixed candidate subset $\mathcal{S}$, the statistical question is

$$H_0(\mathcal{S}): \quad P^{\mathsf{e}}(Y \mid \boldsymbol{X}_{\mathcal{S}}) \text{ is the same across } \mathsf{e} \in \mathscr{E}. \tag{1}$$

Following Peters et al. (2016), this null supports a causal interpretation under additional assumptions on the target features and environments. We develop a test that allows the conditional response model to be unknown and the working prediction class to be misspecified. It extends ICP testing (Peters et al., 2016) to flexible or black-box predictors (Guidotti et al., 2018; Murdoch et al., 2019), subject to estimation and regularity conditions.

The main difficulty is that predictive performance can vary with the covariate distribution even when the conditional response mechanism is fixed. To account for this variation, MAICP adjusts each source environment to the covariate distribution of a reference chosen from $\mathscr{E}$. It fits a predictor in each adjusted environment and evaluates the predictors on separate held-out reference samples. Under conditional invariance, their optimal population scores coincide. We use this score contrast to test each candidate subset and intersect the nonrejected subsets for feature selection with valid Type I error control.

Our contributions are threefold:

New framework for invariant feature selection. We formulate invariant feature selection through conditional invariance testing on a fixed dictionary (concept scores, human-specified, model-derived, etc.) derived from a learned representation. The test assesses the relationship between candidate features and a specified response across different environments through predictiveness.

Model-agnostic null validity. We instantiate the framework using density-adjusted, working-class predictors to construct a held-out predictiveness contrast. Under overlap and mild regularity conditions, the proposed test has asymptotic Type I error control even when the response model is unknown and misspecified. Exact feature recovery additionally requires the score contrast to distinguish candidate subsets.

Application to learned concept scores. Experiments examine the validity of statistical inference under misspecified working models. Applications to concept dictionaries from modern machine learning models, such as CLIP, illustrate the use of MAICP for selecting invariant features. The prediction results show that the selected invariant features help achieve more stable generalization performance under distribution shift.

1.2 Related Work

Invariant causal prediction. ICP (Peters et al., 2016) tests candidate subsets and returns the intersection of those not rejected. The nonlinear extensions of Heinze-Deml et al. (2018); Rojas-Carulla et al. (2018) are the closest inferential comparisons. Invariant risk minimization (IRM), introduced by Arjovsky et al. (2019), learns a predictor through an invariance objective. Extensions by Ahuja et al. (2021); Jiang & Veitch (2022); Yin et al. (2024) and theoretical analyses by Fan et al. (2024); Wang et al. (2023); Gu et al. (2024; 2025) study related questions about stable prediction. MAICP retains the ICP subset-testing target and uses density adjustment to accommodate a misspecified prediction class. The resulting inference guarantees differ from guarantees for predictive performance.

Causal representation learning. CRL studies the recovery of latent causal variables from observed representations under assumptions about mixing and environmental variation (Schölkopf et al., 2021; Schölkopf & von Kügelgen, 2022; Ahuja et al., 2023; Sun et al., 2025; Wang & Jordan, 2024; Zhang et al., 2024). For example, Squires et al. (2023) consider linear mixing, Buchholz et al. (2023) study nonlinear mixing, Varici et al. (2024) consider more general nonlinear models, and Jin & Syrgkanis (2024) allow broader environmental conditions. We start with a supplied representation and test the conditional relationship between candidate coordinates and a specified response.

Explainable artificial intelligence. Our work also relates to the study of explanatory factors (Watson et al., 2021). For example, Geiger et al. (2022) study causal structure in neural networks; Wu et al. (2023) examine interpretable structures in language models; Zhang et al. (2023a) develop structural models of rationalization; and Chen et al. (2024) study causal video understanding. Concept-based

approaches by Rajendran et al. (2024); Marks & Tegmark (2024); Burns et al. (2023) localize interpretable features within networks, although associations in hidden activations require care in causal interpretation (Sharkey et al., 2025). MAICP tests cross-environment invariance within a supplied concept dictionary.

2 PRELIMINARIES

We observe a response Y and a fixed representation $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ in environments $\mathsf{e} \in \mathscr{E}$. The coordinates may be supplied semantic features extracted from a pretrained model. Our target is a subset of this dictionary that is stable across the environments, rather than a latent representation or an equivalence class of causal graphs (Varici et al., 2024; Jin & Syrgkanis, 2024).

Assumption 2.1. The $|\mathscr{E}|$ environments provide mutually independent samples, with n_e observations drawn i.i.d. from P^e within each environment $\mathsf{e} \in \mathscr{E}$. There is a designated set $\mathcal{S}^* \subseteq \{1, \ldots, d\}$ for which $P^\mathsf{e}(Y \mid \boldsymbol{X}_{\mathcal{S}^*})$ is the same wherever the environments' covariate supports overlap.

This is the statistical invariance assumption used in ICP (Peters et al., 2016; Fan et al., 2024). The overlap condition is needed as we do not impose any structural assumption (such as linearity) on $P^\mathsf{e}(Y \mid \boldsymbol{X}_{\mathcal{S}^*})$. Under additional causal assumptions (structural causal model, causal sufficiency, sufficient heterogeneity, etc.), the set $\mathcal{S}^*$ becomes a useful proxy for the parent set of Y in the underlying causal graph.

For each candidate subset $\mathcal{S} \neq \varnothing$, we test

$$H_0(\mathcal{S}): \quad P^\mathsf{e}(Y \mid \boldsymbol{X}_\mathcal{S}) \text{ is the same for all } \mathsf{e} \in \mathscr{E}. \tag{2}$$

Following Peters et al. (2016), if $H_0(\mathcal{S})$ is not rejected, then $\mathcal{S}$ is viewed as a *plausible* invariance set. Let $\mathcal{A}_\alpha$ be the family of nonrejected subsets and define

$$\widehat{\mathcal{S}}^* = \cap_{\mathcal{S} \in \mathcal{A}_\alpha} \mathcal{S}, \qquad \widehat{\mathcal{S}}^* = \varnothing \text{ if } \mathcal{A}_\alpha = \varnothing. \tag{3}$$

Consequently, if $\mathcal{S}^*$ is tested at asymptotic level α, then

$$\liminf_{n \to \infty} \mathbb{P}(\widehat{\mathcal{S}}^* \subseteq \mathcal{S}^*) \geq 1 - \alpha.$$

Notably, this controls false inclusion of features relative to $\mathcal{S}^*$, while it does not control missed features or guarantee exact recovery.

Unlike the linear tests of Peters et al. (2016) and the nonparametric conditional tests of Heinze-Deml et al. (2018), we use a working prediction class for invariance testing, whose validity is *model-agnostic* in the sense that it does not require a known, correct specification of the conditional response model $P^\mathsf{e}(Y \mid \boldsymbol{X}_{\mathcal{S}^*})$.

3 MAICP: MODEL-AGNOSTIC INVARIANT CAUSAL PREDICTION TEST

3.1 POPULATION-LEVEL INVARIANCE TEST

We quantify the invariance of a given subset of features $\mathcal{S}$ by assessing how its predictiveness for the response variable Y changes across different environments e. Fix a candidate $\mathcal{S}$, a working class $\mathcal{F}_\mathcal{S}$, and a predictiveness measure $\mathbb{V}_P(f) = \mathbb{E}_P[v(f(\boldsymbol{X}_\mathcal{S}), Y)]$, with larger values indicating better prediction (Williamson et al., 2023). Examples of predictiveness measures include:

- In regression, the negative squared error is $\mathbb{V}_P(f) = -\mathbb{E}_P[(Y - f(\boldsymbol{X}_\mathcal{S}))^2]$.
- In classification, the accuracy is $\mathbb{V}_P(f) = \mathbb{P}_P\{Y = f(\boldsymbol{X}_\mathcal{S})\}$; it measures how often $f(\boldsymbol{X}_\mathcal{S})$ correctly predicts Y under P.
- Higher-order scores (Hoeffding, 1992), such as the area under the ROC curve (AUC), are discussed in Appendix E under additional regularity conditions.

Predictiveness also underlies statistical variable importance (Boileau et al., 2024; Williamson & Feng, 2020; Williamson et al., 2021; 2023). In invariance learning, we expect the "best model", using invariant features, to exhibit similar predictiveness across environments.

To formulate this idea, at the population level, the class-specific oracle under a distribution P is

$$f_*^P = \arg\max_{f \in \mathcal{F}_\mathcal{S}} \mathbb{V}_P(f).$$

For squared error it equals the conditional mean only when that mean belongs to the working class. Our approach instead allows, for example, a linear class when the conditional mean is nonlinear.

A direct comparison across environments mixes changes in the response mechanism with changes in the feature distribution. We adjust to a reference $\mathsf{e}_0 \in \mathscr{E}$ chosen in advance. For $\mathscr{E} = \{\mathsf{e}_1, \mathsf{e}_2\}$, this requires $\mathsf{e}_0 = \mathsf{e}_1$ or $\mathsf{e}_0 = \mathsf{e}_2$. Let $p^{\mathsf{e}}(x_\mathcal{S})$ be the density of its marginal law $P^{\mathsf{e}}(\boldsymbol{X}_\mathcal{S})$, $\mathbb{V}_\mathsf{e} \equiv \mathbb{V}_{P^\mathsf{e}}$, and $\mathbb{E}_\mathsf{e} \equiv \mathbb{E}_{P^\mathsf{e}}$. For a generic weight ω, define

$$\omega^{\mathsf{e}_1}(x_\mathcal{S}) = \frac{p^{\mathsf{e}_0}(x_\mathcal{S})}{p^{\mathsf{e}_1}(x_\mathcal{S})}, \qquad \mathbb{V}_{\mathsf{e}_1}(f;\omega) = \mathbb{E}_{\mathsf{e}_1}\Big[\omega(\boldsymbol{X}_\mathcal{S}) v(f(\boldsymbol{X}_\mathcal{S}), Y)\Big]. \tag{4}$$

We require $p^{\mathsf{e}_1}(x_\mathcal{S}) > 0$ and $p^{\mathsf{e}_2}(x_\mathcal{S}) > 0$ wherever $p^{\mathsf{e}_0}(x_\mathcal{S}) > 0$. Weighting replaces the source covariate marginal by the reference marginal while retaining the source conditional law:

$$\mathbb{V}_{\mathsf{e}_1}(f;\omega^{\mathsf{e}_1}) = \int \mathbb{E}_{\mathsf{e}_1}\Big[v(f(x_\mathcal{S}), Y) \mid \boldsymbol{X}_\mathcal{S} = x_\mathcal{S}\Big]\, dP^{\mathsf{e}_0}(x_\mathcal{S}). \tag{5}$$

Under $H_0(\mathcal{S})$, the sources and internal reference share a conditional law, so this objective equals $\mathbb{V}_{\mathsf{e}_0}(f)$. Define the weighted optimizers and their reference contrast by

$$f_w^{\mathsf{e}_1} = \arg\max_{f \in \mathcal{F}_\mathcal{S}} \mathbb{V}_{\mathsf{e}_1}(f;\omega^{\mathsf{e}_1}), \qquad f_w^{\mathsf{e}_2} = \arg\max_{f \in \mathcal{F}_\mathcal{S}} \mathbb{V}_{\mathsf{e}_2}(f;\omega^{\mathsf{e}_2}),$$
$$\psi_{\mathsf{e}_0}^{\mathsf{e}_1,\mathsf{e}_2}(\mathcal{S}) = \tfrac{1}{2}\{\mathbb{V}_{\mathsf{e}_0}(f_w^{\mathsf{e}_1}) - \mathbb{V}_{\mathsf{e}_0}(f_w^{\mathsf{e}_2})\}. \tag{6}$$

Here f_w^e denotes a weighted source optimizer, whereas f_*^P is the oracle under the unweighted law P; write $f_*^{\mathsf{e}_0} \equiv f_*^{P^{\mathsf{e}_0}}$ for the reference oracle.

Under $H_0(\mathcal{S})$ the objectives coincide, and hence $\psi_{\mathsf{e}_0}^\mathscr{E}(\mathcal{S}) = 0$, including for a misspecified class. We expect the converse often holds but not always: different conditional mechanisms can sometimes induce the same optimizer or the same reference score. Thus, nonrejection of $H_0(\mathcal{S})$ suggests *plausible* invariance but is not a proof for $\mathcal{S}$.

3.2 Sample-Level Invariance Test: $\mathscr{E} = \{\mathsf{e}_1, \mathsf{e}_2\}$

We propose a model-agnostic test of $H_0(\mathcal{S})$ based on $\psi_{\mathsf{e}_0}^\mathscr{E}(\mathcal{S})$ using observed data. We estimate the density ratios and weighted optimizers, then compare their performance on held-out reference data. Density-ratio approaches to covariate shift are common, including those of Tibshirani et al. (2019); Zhang et al. (2023b). Working classes may include regularized linear models such as Lasso (Tibshirani, 1996), trees, or neural networks.

For each environment e, including the one chosen as the reference, reserve a fitting pool and an evaluation pool before any fitting or tuning. Split the reference evaluation pool into disjoint sets $\mathcal{D}_1^{\mathsf{e}_0}, \mathcal{D}_2^{\mathsf{e}_0}$, each of size $n/2$; here n denotes the retained evaluation sample size. Density fitting, model fitting, and tuning use fitting pools only. Any data-dependent dictionary or screening step uses independent data. This held-out design follows the sample-separation principle used in cross-fitting (Chernozhukov et al., 2018).

For each source $\mathsf{e} \in \{\mathsf{e}_1, \mathsf{e}_2\}$, let $\mathcal{D}_{\text{fit}}^\mathsf{e}$ denote its fitting pool and $\widehat{\omega}^\mathsf{e}$ estimate ω^e using the fitting pools. With exact optimization, define the fitted estimate of f_w^e by

$$\widehat{f}^\mathsf{e} \in \arg\max_{f \in \mathcal{F}_\mathcal{S}} \left\{ \frac{1}{|\mathcal{D}_{\text{fit}}^\mathsf{e}|} \sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{D}_{\text{fit}}^\mathsf{e}} \widehat{\omega}^\mathsf{e}(\boldsymbol{X}_{i,\mathcal{S}}) v(f(\boldsymbol{X}_{i,\mathcal{S}}), Y_i) - \text{pen}_\mathsf{e}(f) \right\}.$$

Here $\text{pen}_\mathsf{e}(f)$ is an optional regularization penalty chosen using fitting pools only, and is zero without regularization. Approximate maximization is allowed provided the complete learner, including weight estimation, regularization, tuning, and optimization error, satisfies Condition (C3).

For $\mathsf{e} \in \{\mathsf{e}_1, \mathsf{e}_2\}$, let $\widehat{V}_\mathsf{e}$ and $\widehat{\eta}_\mathsf{e}$ be the empirical mean and variance of $v(\widehat{f}^\mathsf{e}(\boldsymbol{X}_\mathcal{S}), Y)$ on $\mathcal{D}_2^{\mathsf{e}_0}$ for $\mathsf{e} = \mathsf{e}_1$ and $\mathcal{D}_1^{\mathsf{e}_0}$ for $\mathsf{e} = \mathsf{e}_2$. Then

$$\widehat{\psi}_{\mathsf{e}_0}^\mathscr{E}(\mathcal{S}) = \frac{\widehat{V}_{\mathsf{e}_1} - \widehat{V}_{\mathsf{e}_2}}{2}, \qquad \widehat{\sigma}_{\mathsf{e}_0}^2 = \frac{\widehat{\eta}_{\mathsf{e}_1} + \widehat{\eta}_{\mathsf{e}_2}}{2}, \qquad T_{\mathsf{e}_0}^\mathscr{E}(\mathcal{S}) = \frac{\sqrt{n}\, \widehat{\psi}_{\mathsf{e}_0}^\mathscr{E}(\mathcal{S})}{\widehat{\sigma}_{\mathsf{e}_0}}. \tag{7}$$

The source pair $(\mathsf{e}_1, \mathsf{e}_2)$ is implicit in $\widehat{\sigma}_{\mathsf{e}_0}$. Separate evaluation halves avoid cancellation of the leading evaluation terms under the null. Appendix F gives the unequal-fold variance and requires a finite, positive limiting variance.

For theoretical analysis, we impose the following conditions, separating null validity from power. C1–C4 are imposed for the null subsets to which the normal approximation is applied, in each relevant reference/source comparison. For general scores, let $\phi_{f,P}$ denote their centered influence function; for a mean score it is $v(f(\boldsymbol{X}_{\mathcal{S}}), Y) - \mathbb{V}_P(f)$. All limits below concern the fixed working class and fixed data-generating distributions.

(C1) *Local quadratic loss.* There are $C, \delta > 0$ such that, for $\|f - f_*^{\mathsf{e}_0}\|_{\mathcal{F},P^{\mathsf{e}_0}} < \delta$,

$$|\mathbb{V}_{\mathsf{e}_0}(f) - \mathbb{V}_{\mathsf{e}_0}(f_*^{\mathsf{e}_0})| \le C\|f - f_*^{\mathsf{e}_0}\|^2_{\mathcal{F},P^{\mathsf{e}_0}}.$$

The class-specific reference optimizer, rather than the true conditional response function, is the target in this condition.

(C2) *Held-out first-order expansion.* For the fixed oracle $f_*^{\mathsf{e}_0}$ and the fitted predictor $\widehat{f}$ evaluated on an independent reference empirical distribution $\widehat{P}_m$, with m proportional to n,

$$\mathbb{V}_{\widehat{P}_m}(f) - \mathbb{V}_{\mathsf{e}_0}(f) = (\widehat{P}_m - \mathbb{P}_{\mathsf{e}_0})\phi_{f,P^{\mathsf{e}_0}} + o_p(n^{-1/2}), \qquad f \in \{f_*^{\mathsf{e}_0}, \widehat{f}\}.$$

Here $(\widehat{P}_m - \mathbb{P}_P)h$ denotes the difference between empirical and population averages of h. The remainder is identically zero for mean scores. For higher-order or normalized measures this condition includes the required control of their empirical expansion remainder.

(C3) *Rate of the weighted learner.* For $\mathsf{e} \in \{\mathsf{e}_1, \mathsf{e}_2\}$, $\|\widehat{f}^{\mathsf{e}} - f_w^{\mathsf{e}}\|_{\mathcal{F},P^{\mathsf{e}_0}} = o_p(n^{-1/4})$.

(C4) *Score and variance stability.* For each fitted predictor and its oracle, the influence functions obey

$$\mathbb{P}_{\mathsf{e}_0}\{(\phi_{\widehat{f},P^{\mathsf{e}_0}} - \phi_{f_*^{\mathsf{e}_0},P^{\mathsf{e}_0}})^2\} = o_p(1), \quad \widehat{P}_m\{\phi^2_{\widehat{f},\widehat{P}_m}\} \xrightarrow{p} \mathbb{P}_{\mathsf{e}_0}\{\phi^2_{f_*^{\mathsf{e}_0},P^{\mathsf{e}_0}}\} < \infty.$$

For a mean score, it suffices that $\|v(\widehat{f}(\boldsymbol{X}_{\mathcal{S}}), Y) - v(f_*^{\mathsf{e}_0}(\boldsymbol{X}_{\mathcal{S}}), Y)\|_{L_2(P^{\mathsf{e}_0})} = o_p(1)$; sample independence then gives consistent empirical score variances. The limiting variance used to studentize the contrast is strictly positive.

(C5) *Two-source separation for recovery.* The candidate family is fixed and finite, contains $\mathcal{S}^*$, and for every candidate $\mathcal{S}$ omitting at least one coordinate of $\mathcal{S}^*$,

$$|\psi^{\mathscr{E}}_{\mathsf{e}_0}(\mathcal{S})| \ge c > 0.$$

On these alternatives we separately require $\widehat{\psi}^{\mathscr{E}}_{\mathsf{e}_0}(\mathcal{S}) \xrightarrow{p} \psi^{\mathscr{E}}_{\mathsf{e}_0}(\mathcal{S})$ and $\widehat{\sigma}_{\mathsf{e}_0,\mathcal{S}} \xrightarrow{p} \sigma_{\mathsf{e}_0,\mathcal{S}} \in (0, \infty)$. Consistent held-out scores and variances suffice; a null-specific quadratic expansion is not assumed under an alternative.

Remark 3.1 (Discussion of Conditions). Condition (C3) is a high-level requirement on the complete learner with estimated weights, regularization, and optimization error, which makes the quadratic contribution in Condition (C1) negligible at the $n^{-1/2}$ scale, as in orthogonal machine learning (Chernozhukov et al., 2018). (C3) concerns convergence to the optimizer of the working class and therefore permits model misspecification. Simpler working classes or regularization may help estimation, but the required rate must still hold for the resulting learner. (C5) is not a consequence of Assumption 2.1. For example, an omitted-target subset can have zero contrast because the score does not detect the changed aspect of the conditional law. (C5) imposes conditions for identification and power relative to the chosen score and working class; failure of separation does not by itself invalidate the null or false-inclusion guarantees.

Theorem 3.2. *Let $\mathscr{E} = \{\mathsf{e}_1, \mathsf{e}_2\}$ and choose $\mathsf{e}_0 \in \mathscr{E}$ in advance. Suppose Assumption 2.1 and C1–C4 hold for a fixed candidate $\mathcal{S}$, with finite positive limiting variance. Under $H_0(\mathcal{S})$, $T^{(\mathsf{e}_1,\mathsf{e}_2)}_{\mathsf{e}_0}(\mathcal{S}) \xrightarrow{d} \mathcal{N}(0,1)$. Consequently the two-sided test is asymptotically level α for fixed $\alpha \in (0,1)$, and (3) has asymptotic false-inclusion probability at most α if $\mathcal{S}^*$ is among the candidates. For a fixed finite candidate family, suppose also C5 and consistent contrast and variance estimation under the relevant alternatives. If $\alpha_n \to 0$ and $z_{1-\alpha_n/2} = o(\sqrt{n})$, then $\mathbb{P}(\widehat{\mathcal{S}}^* = \mathcal{S}^*) \to 1$. One valid choice is $\alpha_n = 1/\log n$.*

Algorithm 1 MAICP for Two Sources and an Internal Reference

Require: Prespecified sources $\mathsf{e}_1, \mathsf{e}_2$ and reference e_0 in $\mathscr{E}$; independent fitting/evaluation pools; fixed candidates, working classes, and score; level α.

1: Split the reference evaluation pool into two equal halves; exclude both halves from every fitting and tuning step.
2: **for** each candidate subset $\mathcal{S}$ **do**
3: Estimate $p^{\mathsf{e}_0}(x_\mathcal{S})/p^{\mathsf{e}_1}(x_\mathcal{S})$ and $p^{\mathsf{e}_0}(x_\mathcal{S})/p^{\mathsf{e}_2}(x_\mathcal{S})$ from fitting pools; use ratio 1 for a coincident source/reference.
4: Fit each predictor by maximizing its source empirical weighted score, with regularization and tuning confined to fitting pools.
5: Evaluate $\widehat{f}^{\mathsf{e}_1}$ on $\mathcal{D}_2^{\mathsf{e}_0}$ and $\widehat{f}^{\mathsf{e}_2}$ on $\mathcal{D}_1^{\mathsf{e}_0}$; compute (7).
6: Reject when $|T_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)}| \geq z_{1-\alpha/2}$; if the variance estimate is zero, record the test as unresolved and do not reject.
7: **end for**

Ensure: The intersection in (3).

3.3 Extension to $|\mathscr{E}| \geq 3$

Let $K = |\mathscr{E}| \geq 3$ be fixed. We use a cyclic comparison scheme related to Hu & Lei (2024), with an order fixed before testing. Comparison i uses internal reference $\mathsf{e}_i \in \mathscr{E}$ and sources $\mathsf{e}_{i+1}, \mathsf{e}_{i+2}$, with indices modulo K. The null over $\mathscr{E}$ includes all three environments. Write $T_{i,\mathcal{S}}^{(i+1,i+2)} \equiv T_{\mathsf{e}_i,\mathcal{S}}^{(\mathsf{e}_{i+1},\mathsf{e}_{i+2})}$, recomputing weights and predictors for each comparison. Using equal retained reference evaluation sizes n, we aggregate the squared component statistics:

$$Q_\mathcal{S} = \sum_{i=1}^{K} \left[T_{i,\mathcal{S}}^{(i+1,i+2)}\right]^2. \tag{8}$$

Reject $H_0(\mathcal{S})$ when $Q_\mathcal{S} > \chi^2_{K,1-\alpha}$, the $(1-\alpha)$ quantile of the chi-square distribution with K degrees of freedom. Squaring before summation prevents cancellation between positive and negative component contrasts.

(C6) *Cyclic separation for recovery.* The fixed finite candidate family contains $\mathcal{S}^*$. For the chosen order, write $\psi_{i,\mathcal{S}} \equiv \psi_{\mathsf{e}_i,\mathcal{S}}^{(\mathsf{e}_{i+1},\mathsf{e}_{i+2})}$ and let $\sigma_{i,\mathcal{S}} \in (0,\infty)$ be the probability limit of its standardizer. Every candidate omitting a target coordinate satisfies

$$\Delta_\mathcal{S} := \sum_{i=1}^{K} \left(\frac{\psi_{i,\mathcal{S}}}{\sigma_{i,\mathcal{S}}}\right)^2 \geq c > 0,$$

with consistent contrasts and standardizers for every component. Thus at least one component must detect each such candidate; a conditional-distribution change invisible to every chosen score contrast remains undetectable.

Proposition 3.3. *For fixed $K \geq 3$, suppose $H_0(\mathcal{S})$ holds over $\mathscr{E}$, Assumption 2.1 holds, and every cyclic comparison satisfies overlap, sample separation, C1–C4, and finite positive limiting variance. Then $Q_\mathcal{S} \xrightarrow{d} \chi^2_K$. If these null conditions hold at the tested candidate $\mathcal{S}^*$, the intersection estimator has asymptotic false-inclusion probability at most α for fixed $\alpha \in (0,1)$. Under these target-null conditions and C6, if $\alpha_n \to 0$ and $\chi^2_{K,1-\alpha_n} = o(n)$, then $\mathbb{P}(\widehat{\mathcal{S}}^* = \mathcal{S}^*) \to 1$. One valid choice is $\alpha_n = 1/\log n$.*

The chi-square calibration uses independent reference evaluation samples across environments; shared fitting pools contribute only asymptotically negligible remainders under the stated conditions. For $K > 3$, changing the order can change the selected source pairs, so C6 must hold for the chosen comparisons.

Computational complexity. The task of selecting invariant features is known to be NP-hard (Gu et al., 2025). In our case, the search enumerates $2^{d_*} - 1$ subsets when an independently screened dictionary retains $d_* \leq d$ coordinates. Its worst-case cost is $O(2^{d_*} C_{\mathrm{fit}})$ for one comparison and

Algorithm 2 MAICP with Squared Cyclic Aggregation

Require: $K \geq 3$ environments in a prespecified order; separated fitting and evaluation pools; equal retained reference evaluation sizes; level α.

1: **for** each candidate subset $\mathcal{S}$ **do**
2: **for** $i = 1, \ldots, K$ **do**
3: Compute the component statistic from Algorithm 1 with internal reference $\mathsf{e}_i \in \mathscr{E}$ and sources $\mathsf{e}_{i+1}, \mathsf{e}_{i+2}$, using indices modulo K.
4: **end for**
5: Form $Q_{\mathcal{S}}$ in (8) and reject when $Q_{\mathcal{S}} > \chi^2_{K,1-\alpha}$; if a required variance is zero, record the test as unresolved and do not reject.
6: **end for**

Ensure: The intersection in (3).

$O(|\mathscr{E}|\, 2^{d_*} C_{\mathrm{fit}})$ for cyclic aggregation, where C_{fit} includes density fitting, prediction fitting, and evaluation. The coverage guarantee requires that the designated target remains a tested candidate; however, screening does not provide this automatically. We view this as a practical limitation of the method.

4 Experiments and Results

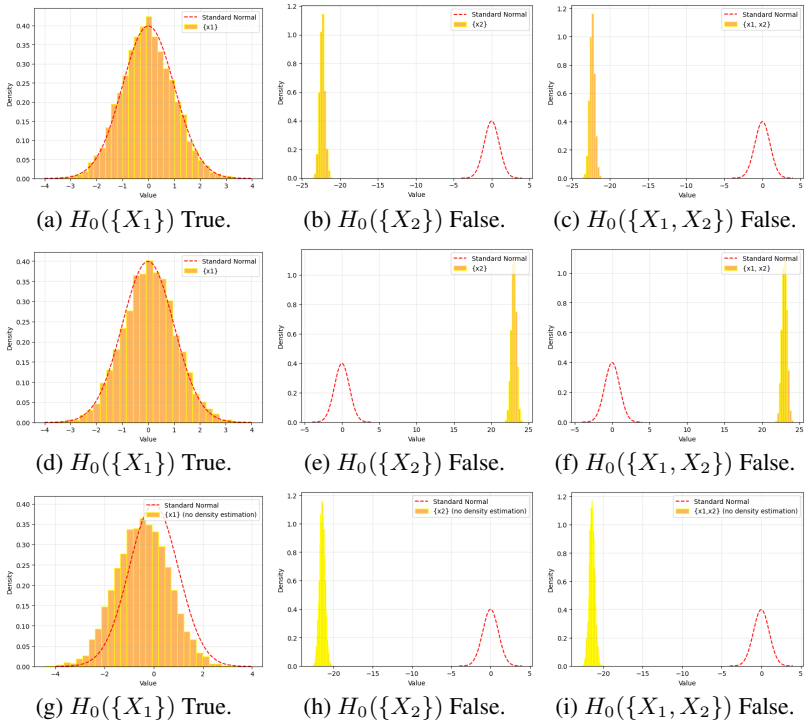
We examine null calibration under working-model misspecification, sensitivity to spurious correlations, and prediction using selected features.

Working-model misspecification and the role of covariate adjustment. We conduct a simulation using the nonlinear mechanism $X_1 \to Y \to X_2$, with an invariant response mechanism and an environment-dependent child mechanism; see Appendix B for details. Linear and first-order logistic working models both omit the nonlinear terms. With $n = 1,000$, the null histograms are approximately standard normal, whereas the alternative histograms are separated from zero (Figure 1, top two rows). Removing density adjustment shifts the null histogram away from zero (Figure 1, bottom row) and inflates Type I error. These results illustrate the desired performance of the MAICP test under misspecification and the role of covariate adjustment for valid inference.

Appendix B further compares MAICP with a linear ICP implementation and examines feature importance under linear and neural-network IRM. The comparison reports null rejection, alternative rejection, and exact recovery for ICP, and illustrates how IRM feature rankings can depend on the working model.

Invariant semantic feature selection with confidence. We use CelebA (Liu et al., 2015), which contains 40 binary facial attributes, and CLIP with a ViT-B/32 backbone. Candidate coordinates are projections of image embeddings onto text-prompt embeddings. Environments e_1 and e_2 reverse the association between the target and selected attributes. Such spurious correlations can affect downstream prediction (Wang & Jordan, 2024). When the response is the dataset's gender label, the `Male`/`Female` directions are semantic proxies for that label. Their scores are computed from image embeddings rather than test labels. Figure 2 reports statistics over 200 subsets of the constructed environments for linear working models (panels a–b) and logistic working models (panels c–d). Under both working models, the target-aligned coordinates have approximately normal histograms, while the eyeglasses and smiling coordinates, and the joint sets containing them, have shifted histograms. The histograms show how the test responds to the constructed changes in these supplied coordinates.

Prediction after selection. Table 1 reports the linear working-model results for three four-feature CelebA dictionaries. Each attribute serves in turn as the response. Training uses 180 images from e_1 and 20 from e_2; testing uses the remaining images from these environments. At $\alpha = 0.05$, all displayed target-coordinate statistics have absolute values below 1.96, while those of the other coordinates exceed this threshold. Selected-feature probes have higher accuracy in every reported row. For example, accuracy changes from 0.478 to 0.697 for Male in panel (a) and from 0.970 to 0.973 for Eyeglasses in panel (b). Appendix C details the protocol and gives logistic working-model results with the same qualitative pattern on different random splits.

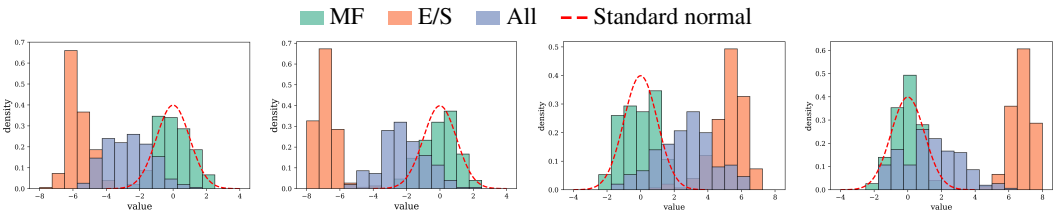


(a) $H_0(\{X_1\})$ True. (b) $H_0(\{X_2\})$ False. (c) $H_0(\{X_1, X_2\})$ False.

(d) $H_0(\{X_1\})$ True. (e) $H_0(\{X_2\})$ False. (f) $H_0(\{X_1, X_2\})$ False.

(g) $H_0(\{X_1\})$ True. (h) $H_0(\{X_2\})$ False. (i) $H_0(\{X_1, X_2\})$ False.

Figure 1: Test distributions under working-model misspecification and a density-adjustment ablation. Top and middle rows: density-adjusted linear and first-order logistic working models, respectively ($n = 1{,}000$). Bottom row: logistic working-model ablation without density adjustment. Red dashed curves show the standard-normal density. The invariant set $\mathcal{S}^* = \{X_1\}$ has approximately standard-normal histograms with adjustment, but its histogram shifts away from zero and inflates Type I errors without adjustment. Subsets containing $\{X_2\}$ remain separated from zero in all three rows.



MF	E	All
0.02	−5.71	−2.65

(a) Linear: eyeglasses

MF	S	All
−0.05	−6.86	−2.19

(b) Linear: smiling

MF	E	All
−0.11	5.24	2.69

(c) Logistic: eyeglasses

MF	S	All
−0.09	6.72	1.36

(d) Logistic: smiling

Figure 2: Distributions of CelebA test statistics under linear (a)–(b) and logistic (c)–(d) working models, with means listed below each plot. Panels (a) and (c) use eyeglasses as the spurious attribute; panels (b) and (d) use smiling. MF denotes the target-aligned gender features, E denotes "Wearing Eyeglasses", S denotes "Smiling", and All denotes the joint set of gender and spurious features. Red dashed curves show the standard-normal density.

The synthetic-image color task changes from 33.4% with all features to 87.6% with selected features, while both attain 100% on cat/dog classification (Appendix A). These studies assess prediction under constructed changes in environment composition, not transfer to a separately collected unseen environment.

Table 1: CelebA coordinate-test statistics and linear-probe accuracies under linear working models. Panels (a)–(c) use distinct four-feature dictionaries. Training uses 180 images from e_1 and 20 from e_2; testing uses the remaining images from these environments. Boldface marks the higher accuracy within each row.

	Test statistic				Test accuracy	
(a) Hair, age, and gender						
Target Feature	Straight Hair	Black Hair	Male	Young	All	Selected
Straight Hair	0.32	−11.16	−10.49	−9.87	0.241	**0.577**
Black Hair	−8.70	−1.54	−10.96	−10.19	0.545	**0.657**
Male	−4.57	−10.12	1.52	−8.72	0.478	**0.697**
Young	−3.12	−11.02	−10.51	−0.43	0.448	**0.771**
(b) Nose, eyeglasses, age, and gender						
Target Feature	Big Nose	Eyeglasses	Young	Male	All	Selected
Eyeglasses	2.27	1.03	−9.91	−9.68	0.970	**0.973**
Male	−8.13	−12.14	−7.26	0.95	0.191	**0.721**
Young	−10.39	−11.21	−0.63	−8.78	0.117	**0.714**
Big Nose	−0.83	−12.86	−8.99	−9.87	0.107	**0.495**
(c) Expression and gender						
Target Feature	Cheekbone	Smile	Mouth Open	Male	All	Selected
Mouth Open	3.08	−10.19	−0.36	−8.61	0.276	**0.595**
Smile	−4.18	1.54	−4.76	−8.11	0.560	**0.687**
Cheekbone	−0.89	−2.81	−2.31	−8.27	0.249	**0.378**
Male	−4.28	−11.47	−2.78	−0.69	0.420	**0.686**

5 Discussion and limitations

MAICP tests conditional invariance through density-adjusted predictiveness in a supplied representation. Under overlap and very mild regularity conditions, it offers valid Type I error control $\liminf_{n\to\infty} \mathbb{P}(\widehat{\mathcal{S}}^* \subseteq \mathcal{S}^*) \geq 1-\alpha$ without assuming a specific model, permitting a flexible working prediction class, including a misspecified class. Power, or exact identification $\mathbb{P}(\widehat{S}^* = \mathcal{S}^*)$, requires the working class and score to expose a difference between mechanisms. The MAICP-identified invariant features are closely related to causal features, but they do not by themselves establish causality; their causal interpretation requires additional structural assumptions. Invariant features need not maximize predictive accuracy, but they can be useful for understanding the underlying mechanisms and for robust generalization.

MAICP has several limitations. First, its success depends on the availability of sufficiently heterogeneous environments to reveal invariant structure. In settings where environmental variation is weak or uninformative, the proposed test's power may be limited. Second, poor overlap can also affect estimation and calibration. Third, the effectiveness of the proposed method in embedding space depends critically on the quality and expressiveness of the pre-trained model. Additionally, the technique requires pre-specified feature representations, which may not be readily available for models such as autoencoders. Finally, the task of selecting invariant features is fundamentally NP hard (Gu et al., 2025). MAICP enumerates candidate subsets, which can be computationally intensive, as a trade-off between inferential guarantees and computational complexity. Practically, MAICP can be combined with screening techniques to mitigate this issue. It remains open whether the method can be extended to end-to-end differentiable learning while maintaining its statistical inference guarantees, such as controlled Type I error.

AI USE STATEMENT

We used Stable Diffusion v1-5 to generate the synthetic cat and dog images and ChatGPT-4 to summarize visual characteristics for defining candidate features, as described in Appendix A. Generative AI assistance was also used to revise manuscript language and adapt the LaTeX source to the ICLR 2027 template. The authors reviewed the AI-assisted text, citations, code, figures, candidate concepts, and summaries, verified the reported results, and take responsibility for the final content of this work.

ETHICS STATEMENT

MAICP can help assess which predictive associations persist across specified environments. Statistical nonrejection should not be used to label semantic attributes as causal or to justify decisions about individuals. The CelebA study uses dataset annotations in a controlled evaluation; its conclusions do not establish fairness or the suitability of demographic prediction for deployment.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pp. 372–407. PMLR, 2023.

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

Reza Bayat, Mohammad Pezeshki, Elvis Dohmatob, David Lopez-Paz, and Pascal Vincent. The pitfalls of memorization: When memorization hurts generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vVhZh9ZpIM.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Philippe Boileau, Ning Leng, Nima S Hejazi, Mark van der Laan, and Sandrine Dudoit. A nonparametric framework for treatment effect modifier discovery in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkae084, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36:45419–45462, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, et al. Mecd: Unlocking multi-event causal discovery in video reasoning. *Advances in Neural Information Processing Systems*, 37:92554–92580, 2024.

Wei Chen, Jibin Chen, Ruichu Cai, Yuequn Liu, and Zhifeng Hao. Learning granger causality for non-stationary hawkes processes. *Neurocomputing*, 468:22–32, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1), 2018.

Maxime Darrin, Philippe Formont, Ismail Ayed, Jackie CK Cheung, and Pablo Piantanida. When is an embedding model more promising than another? *Advances in Neural Information Processing Systems*, 37:68330–68379, 2024.

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jianqing Fan, Cong Fang, Yihong Gu, and Tong Zhang. Environment invariant linear least squares. *The Annals of Statistics*, 52(5):2268–2292, 2024.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pp. 7324–7338. PMLR, 2022.

Yihong Gu, Cong Fang, Peter Bühlmann, and Jianqing Fan. Causality pursuit from heterogeneous environments via neural adversarial invariance learning. *arXiv preprint arXiv:2405.04715*, 2024.

Yihong Gu, Cong Fang, Yang Xu, Zijian Guo, and Jianqing Fan. Fundamental computational limits in pursuing invariant causal prediction and invariance-guided regularization. *arXiv preprint arXiv:2501.17354*, 2025.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), August 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334, 1992.

Xiaoyu Hu and Jing Lei. A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, 119(546):1136–1154, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. *Advances in Neural Information Processing Systems*, 35:20782–20794, 2022.

Jikai Jin and Vasilis Syrgkanis. Learning linear causal representations from general environments: Identifiability and intrinsic ambiguity. *Advances in Neural Information Processing Systems*, 37: 63466–63509, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pp. 16848–16887. PMLR, 2022.

Anne-Marie Nussberger, Lan Luo, L Elisa Celis, and Molly J Crockett. Public attitudes value interpretability but prioritize accuracy in artificial intelligence. *Nature communications*, 13(1): 5821, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

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: Statistical Methodology*, 78(5):947–1012, 2016.

Drago Plecko and Elias Bareinboim. Causal fairness for outcome control. *Advances in Neural Information Processing Systems*, 36:47575–47597, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296, 2024.

Atul Rawal, Adrienne Raglin, Danda B Rawat, Brian M Sadler, and James McCoy. Causality for trustworthy artificial intelligence: status, challenges and perspectives. *ACM Computing Surveys*, 57(6):1–30, 2025.

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.

Bernhard Schölkopf and Julius von Kügelgen. From statistical to causal learning. In *Proceedings of the International Congress of Mathematicians*, volume 1, 2022.

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.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International conference on machine learning*, pp. 32540–32560. PMLR, 2023.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, et al. Causal representation learning from multimodal biomedical observations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

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

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *Journal of Machine Learning Research*, 25(275):1–65, 2024.

David S Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice. In *Uncertainty in Artificial Intelligence*, pp. 1382–1392. PMLR, 2021.

Brian Williamson and Jean Feng. Efficient nonparametric statistical inference on population feature importance using shapley values. In *International conference on machine learning*, pp. 10282–10291. PMLR, 2020.

Brian D Williamson, Peter B Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36:78205–78226, 2023.

Xingcheng Xu, Zihao Pan, Haipeng Zhang, and Yanqing Yang. It ain't that bad: understanding the mysterious performance drop in ood generalization for generative transformer models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/727. URL https://doi.org/10.24963/ijcai.2024/727.

Mingzhang Yin, Yixin Wang, and David M Blei. Optimization-based causal estimation from heterogeneous environments. *Journal of Machine Learning Research*, 25(168):1–44, 2024.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *International Conference on Machine Learning*, pp. 60057–60075. PMLR, 2024.

Wenbo Zhang, Tong Wu, Yunlong Wang, Yong Cai, and Hengrui Cai. Towards trustworthy explanation: On causal rationalization. In *International Conference on Machine Learning*, pp. 41715–41736. PMLR, 2023a.

Yu-Jie Zhang, Zhen-Yu Zhang, Peng Zhao, and Masashi Sugiyama. Adapting to continuous covariate shift via online density ratio estimation. *Advances in Neural Information Processing Systems*, 36: 29074–29113, 2023b.

APPENDIX

Appendix A describes the synthetic-image study. Appendices B and C give simulation results and details of the CelebA protocol, including interpretation of Table 1, logistic working-model results, the weight ablation, and ICP and IRM comparisons. Appendix D collects notation, Appendix E treats general predictiveness measures, and Appendix F provides assumptions and proofs.

A ILLUSTRATION EXPERIMENTS USING SYNTHETIC DATA

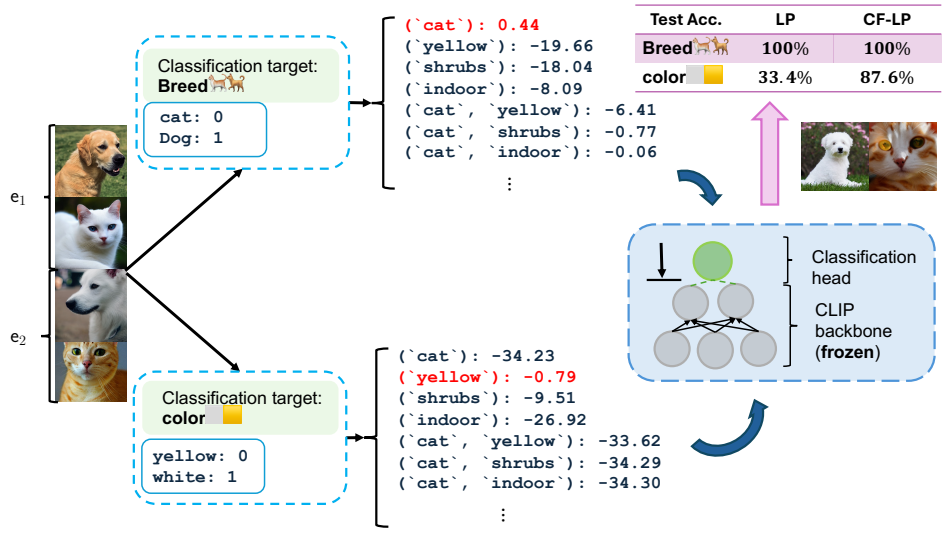


Figure 3: Illustration of MAICP with linear probing results on synthetic cat and dog image data generated using Stable Diffusion v1-5.

We apply MAICP to synthetic images to study supplied-feature testing under a controlled reversal of color–label association. Two environments, e_1 and e_2, are generated using Stable Diffusion v1-5 on an A100 GPU. Each contains 2000 images, with 1000 cats and 1000 dogs. In e_1, all dogs are yellow and all cats are white; this association is reversed in e_2.

The two prediction targets are animal class and animal color. For each task, we randomly select 800 images from e_1 and 200 from e_2 for training and use the remaining images for testing. To define the dictionary, 10 images are sampled from each environment and ChatGPT-4 is asked to summarize their common visual characteristics using single-word descriptors. The resulting dictionary is `['cat', 'yellow', 'shrubs', 'indoor']`. Coordinates are obtained by projecting image embeddings onto the corresponding CLIP text embeddings. We compare linear probing using all features (LP) with probing using the selected subset, denoted CF-LP in Figure 3. The red entries indicate the selected features.

For animal-class prediction, MAICP selects `cat`, and both LP and CF-LP attain 100% test accuracy. For color prediction, LP attains 33.4% and the selected-feature probe attains 87.6%. Thus, selection improves prediction in the color task, while all-feature probing already succeeds in the animal-class task. These results concern this supplied dictionary and the two constructed environments.

B ADDITIONAL SIMULATION RESULTS

We examine the empirical distribution of the test statistic under the following two structural equation models:

$$\text{Environment } \mathsf{e}_1\text{:} \begin{cases} x_1 &= \epsilon_1 \\ y &= \mathbb{1}\{\text{expit}((x_1 + x_1^2/4 + x_1^3/3 - 0.5)/2) \le u\}, \\ x_2 &= y + \epsilon, \end{cases} \tag{9}$$

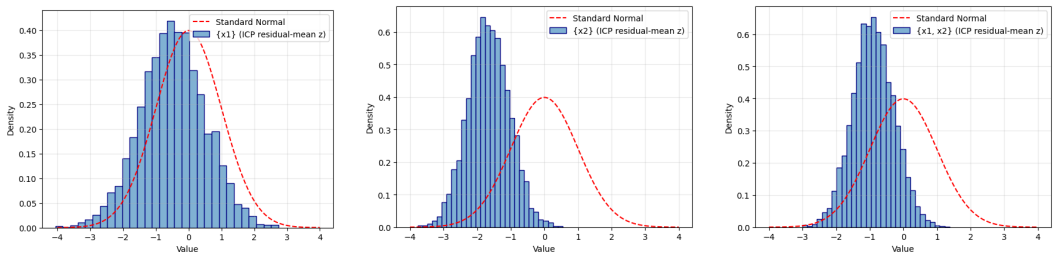
and

$$\text{Environment } \mathsf{e}_2\text{:} \begin{cases} x_1 &= \epsilon_2 \\ y &= \mathbb{1}\{\text{expit}((x_1 + x_1^2/4 + x_1^3/3 - 0.5)/2) \le u\}, \\ x_2 &= -y + \epsilon, \end{cases} \tag{10}$$

Here $\text{expit}(t) = (1+\exp(-t))^{-1}$. With the displayed inequality, the Bernoulli success probability is one minus this map in both environments. We take $\epsilon_1 \sim \mathcal{N}(0,1)$ truncated to $(-1,2)$, $\epsilon_2 \sim \mathcal{U}(-1,2)$, $u \sim \mathcal{U}(0,1)$, and $\epsilon \sim \mathcal{N}(-1, 0.25)$. We use $\mathcal{U}(a,b)$ to denote the uniform distribution within (a,b) for $a < b$ and $\mathcal{N}(a,b)$ to denote the normal distribution with mean a and variance b for $b > 0$. The target response variable y follows a Bernoulli distribution with probability depending only on the feature x_1. The corresponding causal DAG is $x_1 \to y \to x_2$, and the directed edge $y \to x_2$ is subject to a soft intervention. Thus, the subset $\{x_1\}$ satisfies conditional invariance. The environment-dependent child mechanism supplies an alternative whose detectability still depends on the working class and score. The distributions of x_1 differ between the two environments, creating a covariate shift in the features. We use a kernel-based method to estimate the density ratio and consider two working models: a linear model and a logistic regression model with only first-order terms, excluding interactions. Note that the true model is a logistic regression with up to third-order terms, so both working models are misspecified. The sample size is $n = 1,000$, and the histogram is based on $5,000$ replications.

Comparison with ICP and IRM

We extend the same nonlinear simulation to compare MAICP with an ICP implementation based on linear regression (Peters et al., 2016) and to examine feature importance under IRM (Arjovsky et al., 2019). The working linear model is misspecified for the response mechanism. Figure 4 shows the reported ICP residual-mean statistics for the invariant subset $\{X_1\}$ and the non-invariant subsets $\{X_2\}$ and $\{X_1, X_2\}$.

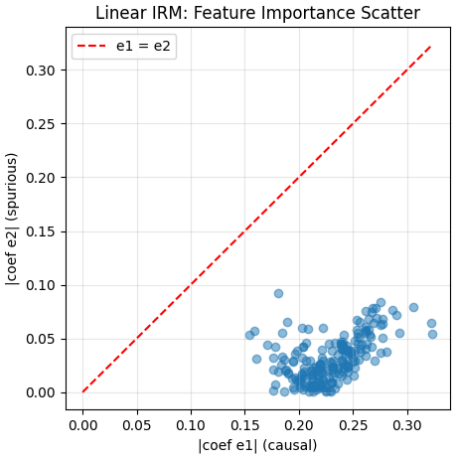


(a) $\{x_1\}$: invariant. (b) $\{x_2\}$: non-invariant. (c) $\{x_1, x_2\}$: non-invariant.

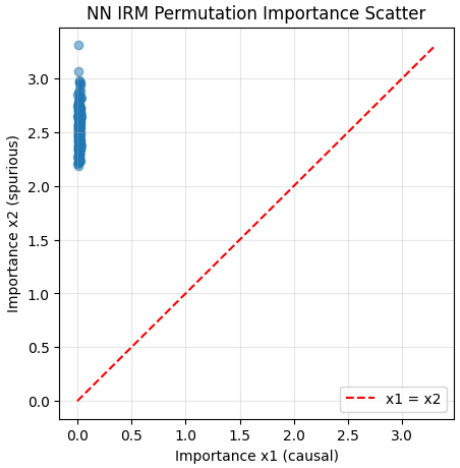
Figure 4: Empirical distributions of the reported linear-ICP residual-mean test statistics under the nonlinear simulation mechanism. Red dashed curves show the standard-normal density.

The reported null rejection rates are 0.053 for MAICP and 0.0586 for ICP on $\{X_1\}$. The ICP rejection rates are 0.34 on $\{X_2\}$ and 0.0276 on $\{X_1, X_2\}$, with exact recovery 0.2908. Thus, the reported null rejection rates are close to the nominal 0.05 level, but ICP has limited power against the displayed alternatives.

We evaluate a linear model trained with the IRMv1 objective and a two-layer multilayer perceptron. Feature importance is measured by coefficient magnitude for the linear model and by permutation importance for the neural network. Figure 5 shows that linear IRM assigns higher importance to X_1, whereas neural-network IRM assigns higher importance to X_2. This pattern is consistent with the neural network exploiting nonlinear spurious associations that the linear model cannot represent, even while approximately satisfying the IRM penalty. The experiment does not isolate this explanation from optimization effects. In this setting, the IRM objective alone does not prevent reliance on the spurious feature, and the result depends on model capacity and optimization.



(a) Linear IRM: coefficient magnitude.



(b) Neural-network IRM: permutation importance.

Figure 5: Feature importance for the invariant feature X_1 and spurious feature X_2 under two IRM implementations. Each panel uses its indicated importance measure; magnitudes are not directly comparable across panels.

C Additional Experiment Results

CelebA features and constructed environments

We extract 512-dimensional image embeddings with the pretrained CLIP ViT-B/32 encoder. The tested concept dictionaries contain fewer than ten coordinates; subset search operates on these coordinates, not on all 512 embedding components. Text prompts include "`A photo of a {} person`", "`A photo of a man`", and "`A photo of a man wearing {}`". The text encoder supplies directions for projecting each image embedding into semantic coordinates. Figure 6 shows example images. In environment e_1, a spurious attribute and the response have matching binary labels; in e_2, their association is reversed. The dictionary therefore includes a target-aligned semantic coordinate as well as other correlated coordinates. The target-aligned coordinate is computed from the image, whereas the response is the dataset annotation.

For the gender-label histograms, the spurious coordinate is either wearing eyeglasses or smiling. We randomly select up to $20,000$ images per class in each environment and proportionally divide the dataset into 200 subsets according to the environment-specific class counts. Figure 2 reports the linear working-model results in panels a–b and the logistic working-model results in panels c–d. This repeated subdivision describes the available finite-sample histograms; it does not supply independent evidence of causal ground truth.

For downstream prediction, four highly correlated attributes are chosen using pairwise correlations in the dataset. Each is used in turn as the response, with the other three defining spurious correlations. The reported training sample has 200 images: 180 from e_1 (90 per class) and 20 from e_2 (10 per class). Remaining images from these constructed environments form the test set. MAICP is applied to the training data, followed by linear probing on either all four semantic coordinates or the selected subset.

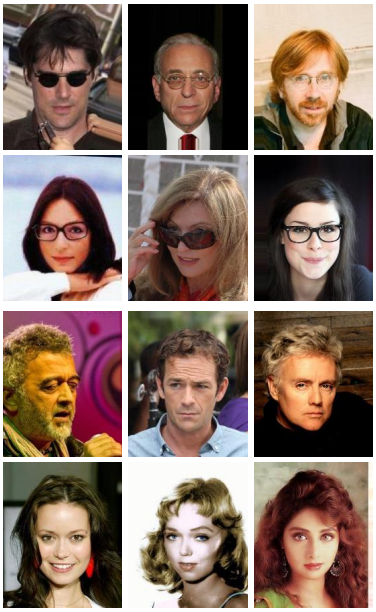


Figure 6: Examples of semantic dimensions in the CelebA dataset. First row: male celebrities with straight hair and wearing eyeglasses; second row: female celebrities with straight hair and wearing eyeglasses; third row: male celebrities with wavy hair and not wearing eyeglasses; last row: female celebrities with wavy hair and not wearing eyeglasses.

RESULTS USING LOGISTIC WORKING MODELS

We repeat the CelebA analysis using logistic regression as the working model. The gender-label histograms use the same two settings, up to $20,000$ images per class in each environment, and 200 proportionally formed subsets. The logistic results are shown alongside the linear results in Figure 2. In panels c–d, the target-coordinate histograms are approximately normal, while the spurious-coordinate and joint-set histograms are shifted.

The downstream tasks use the same feature groups and training composition as the linear working-model analysis in Table 1, with different random train/test splits. At $\alpha = 0.05$, the absolute target-coordinate statistics remain below 1.96 and the absolute other-coordinate statistics exceed it. The selected-feature probe has higher accuracy in every reported row of Table 2. The different splits confound a direct comparison of the two working classes. For example, the all-feature Eyeglasses accuracy is 0.970 in Table 1(b) and 0.751 in Table 2(b), whereas the selected-feature accuracies differ by 0.048.

Table 2: CelebA coordinate-test statistics and linear-probe accuracies under logistic working models. Panels (a)–(c) use distinct four-feature dictionaries. Training uses 180 images from e_1 and 20 from e_2; testing uses the remaining images from these environments. Boldface marks the higher accuracy within each row.

	Test statistic				Test accuracy	
(a) Hair, age, and gender						
Target Feature	Straight Hair	Black Hair	Male	Young	All	Selected
Straight Hair	−0.60	−11.10	8.89	8.83	0.216	**0.580**
Black Hair	−7.37	−0.136	−10.66	−9.12	0.490	**0.638**
Male	3.80	9.70	0.85	9.43	0.619	**0.714**
Young	−3.34	−11.43	11.28	−0.55	0.450	**0.759**
(b) Nose, eyeglasses, age, and gender						
Target Feature	Big Nose	Eyeglasses	Young	Male	All	Selected
Eyeglasses	−2.63	0.54	5.00	10.13	0.751	**0.925**
Male	8.14	12.12	4.69	0.21	0.361	**0.719**
Young	3.06	11.68	−0.39	9.35	0.311	**0.711**
Big Nose	1.92	11.94	4.87	8.76	0.122	**0.502**
(c) Expression and gender						
Target Feature	Cheekbones	Smiling	Mouth Open	Male	All	Selected
Mouth Open	3.73	9.22	−0.44	8.71	0.323	**0.592**
Smiling	4.03	−0.86	3.07	8.42	0.565	**0.682**
Cheekbones	−1.92	7.02	5.28	8.17	0.213	**0.390**
Male	13.29	−10.99	−2.31	1.25	0.612	**0.704**

D Notations and Abbreviations

Write $Z = (\boldsymbol{X}, Y)$, and let P^{e} be its distribution in environment e. Every reference is a prespecified member of the observed family $\mathscr{E}$; it may be distinct from both sources only when $|\mathscr{E}| \geq 3$. For a coordinate set $\mathcal{S} \subseteq \{1, \ldots, d\}$, $P^{\mathsf{e}}(\boldsymbol{X}_{\mathcal{S}})$ and $P^{\mathsf{e}}(Y \mid \boldsymbol{X}_{\mathcal{S}})$ denote its covariate marginal and conditional response law. The working class $\mathcal{F}_{\mathcal{S}}$ is fixed across environments and is equipped with a norm $\|\cdot\|_{\mathcal{F}, P^{\mathsf{e}_0}}$, for example the $L_2(P^{\mathsf{e}_0}(\boldsymbol{X}_{\mathcal{S}}))$ prediction norm. The oracle $f_*^{\mathsf{e}_0} \equiv f_*^{P^{\mathsf{e}_0}}$ maximizes $\mathbb{V}_{\mathsf{e}_0}(f)$ within this class. We assume either a unique optimizer, up to equality on the reference support, or a common deterministic selection rule among maximizers.

For source $\mathsf{e} \in \{\mathsf{e}_1, \mathsf{e}_2\}$, the ratio $\omega^{\mathsf{e}}(x_{\mathcal{S}}) = p^{\mathsf{e}_0}(x_{\mathcal{S}})/p^{\mathsf{e}}(x_{\mathcal{S}})$ is defined wherever the reference density is positive, with value zero outside the reference support where needed. The support requirement is $p^{\mathsf{e}_0}(x_{\mathcal{S}}) > 0 \Rightarrow p^{\mathsf{e}}(x_{\mathcal{S}}) > 0$. A sufficient moment condition is that these ratios are bounded and the score has finite second moments under the corresponding reweighted source laws. These conditions ensure that the relevant weighted expectations exist; they do not themselves establish the learner rate in C3. Support inclusion also does not ensure good effective overlap in a finite sample. Large weights can make estimation unstable, and weighting provides no justification for extrapolation beyond source support.

For an integrable function h of a joint observation, write

$$\widehat{P}_m h = \frac{1}{m} \sum_{j=1}^{m} h(Z_j), \qquad \mathbb{P}_P h = \mathbb{E}_P[h(Z)], \qquad \mathbb{P}_{\mathsf{e}} h \equiv \mathbb{P}_{P^{\mathsf{e}}} h.$$

Here $\widehat{P}_m$ is the empirical distribution of the designated reference evaluation half; braces may enclose the integrand. Thus $(\widehat{P}_m - \mathbb{P}_P)h$ is an empirical-minus-population average. The law P remains the argument of $\mathbb{V}_P$ and $\phi_{f,P}$, while $\mathbb{P}(\text{event})$ denotes sampling probability. In a marginal integral, $dP^{\mathsf{e}}(x_{\mathcal{S}})$ means integration with respect to $P^{\mathsf{e}}(\boldsymbol{X}_{\mathcal{S}})$.

The weighted source oracle and population contrast are

$$f_w^{\mathsf{e}} = \arg\max_{f \in \mathcal{F}_{\mathcal{S}}} \mathbb{V}_{\mathsf{e}}(f; \omega^{\mathsf{e}}), \qquad \psi_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)} = \tfrac{1}{2}\{\mathbb{V}_{\mathsf{e}_0}(f_w^{\mathsf{e}_1}) - \mathbb{V}_{\mathsf{e}_0}(f_w^{\mathsf{e}_2})\}.$$

The factor $1/2$ is used in both the population and sample definitions throughout. Each physical environment has a fitting pool and an independent evaluation pool. Fitting pools may be reused across comparisons, but no evaluation observation enters any ratio estimate, predictor fit, or tuning decision. Reference evaluation halves have sizes m_1, m_2, with $n = m_1 + m_2$; the main algorithms use $m_1 = m_2 = n/2$. Fitting sample sizes grow at rates comparable to n, or otherwise satisfy C3 at the stated scale. Source sample sizes need not be equal.

E Extension to General Predictiveness Measures

The average score in the main text is a special case of a fixed-degree V-measure Williamson et al. (2023). For a fixed candidate $\mathcal{S}$, each kernel may depend on an observation only through its selected covariates and response. Specifically, for $z_j = (x_j, y_j)$, require

$$G_f(z_1, \ldots, z_\ell) = \widetilde{G}_f\big((x_{1,\mathcal{S}}, y_1), \ldots, (x_{\ell,\mathcal{S}}, y_\ell)\big), \tag{11}$$

where, for each $f \in \mathcal{F}_{\mathcal{S}}$, $\widetilde{G}_f$ is a fixed measurable symmetric kernel, common across environments and independent of the distribution P. All displayed expectations are assumed to exist. For independent copies $Z_1, \ldots, Z_\ell$, define

$$\mathbb{V}_P(f) = \mathbb{E}_{P^{\otimes \ell}}[G_f(Z_1, \ldots, Z_\ell)]. \tag{12}$$

For example, $G_f(Z) = v(f(\boldsymbol{X}_{\mathcal{S}}), Y)$ when $\ell = 1$. The transported objective is

$$\mathbb{V}_{\mathsf{e}_1}(f; \omega) = \mathbb{E}_{(P^{\mathsf{e}_1})^{\otimes \ell}}\left[G_f(Z_1, \ldots, Z_\ell) \prod_{j=1}^{\ell} \omega(\boldsymbol{X}_{j,\mathcal{S}})\right].$$

The fitted source predictor maximizes the corresponding empirical objective over its fitting pool, with any regularization included in the complete learner subject to C3. Evaluation uses $\mathbb{V}_{\widehat{P}_m}(\widehat{f})$ on its designated reference half. For these general measures, $\widehat{V}_{\mathsf{e}} = \mathbb{V}_{\widehat{P}_m}(\widehat{f}^{\mathsf{e}})$ replaces the empirical mean score in (7), with $\widehat{P}_m$ taken from the half assigned to source e. The influence function for a fixed f is

$$\phi_{f,P}(z) = \ell\left\{\mathbb{E}_{P^{\otimes(\ell-1)}}[G_f(z, Z_2, \ldots, Z_\ell)] - \mathbb{V}_P(f)\right\}. \tag{13}$$

It is the first-order contribution of an evaluation observation and has mean zero under P. The variance estimate on that half is

$$\widehat{\eta}_{\mathsf{e}} = \widehat{P}_m\{\phi^2_{\widehat{f}^{\mathsf{e}},\widehat{P}_m}\}.$$

For mean scores this reduces to the empirical score variance in (7). C2 and C4 explicitly require the first-order remainder and this variance estimate to be valid; mere differentiability is not enough.

Some familiar measures also estimate a normalization. We treat normalized measures as fixed smooth functions of finitely many moments, using the same functions across environments and their corresponding chain-rule influence functions. Each moment kernel satisfies (11), and any normalizing denominator is nonzero. Let $P_{\mathcal{S}}^{\mathsf{e}}$ denote the law of $(\boldsymbol{X}_{\mathcal{S}}, Y)$ and define the transported law by $dQ_{\mathcal{S}}^{\mathsf{e}} = \omega^{\mathsf{e}}(x_{\mathcal{S}})\, dP_{\mathcal{S}}^{\mathsf{e}}$. Every moment, including those in a normalizing denominator, is evaluated under $Q_{\mathcal{S}}^{\mathsf{e}}$ in the transported objective. In particular, for the population coefficient $\mathbb{R}^2_P(f)$ and its sample version R^2, let $M = \mathbb{E}_P[(Y - f(\boldsymbol{X}_{\mathcal{S}}))^2]$, $\mu = \mathbb{E}_P Y$, and $D = \mathrm{Var}_P(Y) > 0$. Its fixed-predictor influence function is

$$\phi_{f,P}(\boldsymbol{X}, Y) = -\frac{(Y - f(\boldsymbol{X}_{\mathcal{S}}))^2 - M}{D} + \frac{M}{D^2}\{(Y - \mu)^2 - D\}.$$

Its variance, rather than the variance of squared residuals alone, is required when D is estimated on the evaluation sample. AUC similarly combines a pairwise ranking score with class-probability normalization. Its application requires nonvanishing class probabilities and conditions ensuring the local quadratic bound, score stability, and a valid first-order expansion despite ranking indicators. Accuracy and AUC are therefore potential scores, not examples for which C1–C4 hold without further conditions.

Theorem E.1 (General predictiveness measures)**.** *Choose sources* $\mathsf{e}_1, \mathsf{e}_2$ *and a reference* e_0 *from* $\mathscr{E}$ *in advance. For a fixed candidate* $\mathcal{S}$ *satisfying* $H_0(\mathcal{S})$ *over* $\mathscr{E}$*, suppose overlap, sample separation, C1–C4, and finite positive limiting variance hold. Assume also* (11)*, applied to each constituent moment for a normalized measure, with normalization under the transported law as specified above. Then the statistic constructed with the influence-function variance above satisfies* $T_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)} \xrightarrow{d} \mathcal{N}(0,1)$*. The asymptotic coverage and exact-recovery conclusions of Theorem 3.2 hold with the same additional assumptions for the candidate family and alternatives. Under the conditions of Proposition 3.3, the squared cyclic aggregate satisfies* $Q_{\mathcal{S}} \xrightarrow{d} \chi^2_K$ *and has the stated false-inclusion and recovery guarantees.*

F Technical Proofs

Lemma F.1. *Suppose* $H_0(\mathcal{S})$ *holds over* $\mathscr{E}$*, the sources* $\mathsf{e}_1, \mathsf{e}_2$ *and reference* e_0 *belong to* $\mathscr{E}$*, and the reference covariate distribution has support contained in each source support. For general measures, assume* (11) *and the transported-law normalization specified in Appendix E, when applicable. Then* $\mathbb{V}_{\mathsf{e}_1}(f;\omega^{\mathsf{e}_1}) = \mathbb{V}_{\mathsf{e}_2}(f;\omega^{\mathsf{e}_2}) = \mathbb{V}_{\mathsf{e}_0}(f)$ *for every admissible* f*. With the common rule for choosing maximizers,*

$$f_w^{\mathsf{e}_1} = f_w^{\mathsf{e}_2} = f_*^{\mathsf{e}_0} \tag{14}$$

on the reference support, and $\psi_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)} = 0$*.*

F.1 Proof of Lemma F.1

For a mean score, conditioning on $\boldsymbol{X}_{\mathcal{S}} = x_{\mathcal{S}}$ gives

$$\begin{aligned}\mathbb{V}_{\mathsf{e}_1}(f;\omega^{\mathsf{e}_1}) &= \int \mathbb{E}_{\mathsf{e}_1}[v(f(x_{\mathcal{S}}),Y) \mid \boldsymbol{X}_{\mathcal{S}} = x_{\mathcal{S}}]\, \frac{p^{\mathsf{e}_0}(x_{\mathcal{S}})}{p^{\mathsf{e}_1}(x_{\mathcal{S}})} p^{\mathsf{e}_1}(x_{\mathcal{S}})\, dx_{\mathcal{S}} \\ &= \int \mathbb{E}_{\mathsf{e}_0}[v(f(x_{\mathcal{S}}),Y) \mid \boldsymbol{X}_{\mathcal{S}} = x_{\mathcal{S}}]\, p^{\mathsf{e}_0}(x_{\mathcal{S}})\, dx_{\mathcal{S}} = \mathbb{V}_{\mathsf{e}_0}(f).\end{aligned}$$

Here integrals are with respect to the common dominating measure, and the second equality uses $H_0(\mathcal{S})$ together with $\mathsf{e}_1, \mathsf{e}_0 \in \mathscr{E}$ to equate the source and reference conditional laws. The same argument applies to e_2. Under the null and overlap, the transported reduced law satisfies $Q_{\mathcal{S}}^{\mathsf{e}} = P_{\mathcal{S}}^{\mathsf{e}_0}$. By (11), each kernel argument uses only this reduced observation. Applying the change of law to each of the ℓ independent arguments therefore gives the reference expectation, using the product of the ℓ density ratios. For normalized measures, equality of all constituent moments also gives equality of the normalized objectives. Thus the optimization objectives agree for every f, which proves the claim with the common maximizer rule. Without that rule, the sets of maximizers and their optimal values still agree, but equality of chosen functions would not follow automatically.

F.2 Proof of Theorems E.1 and 3.2

We first establish the held-out expansion underlying both tests. It is the score expansion used in predictiveness inference Williamson et al. (2023), stated here with the sample separation required by our comparison.

Theorem F.2. *For prespecified sources and reference in* $\mathscr{E}$*, under* $H_0(\mathcal{S})$ *over* $\mathscr{E}$*, overlap, and C1–C4, let* $\widehat{P}_m$ *be the independent reference empirical distribution used to evaluate source predictor* $\widehat{f}^{\mathsf{e}}$*. For general measures, assume also the kernel and normalization conditions of Theorem E.1. Then*

$$\mathbb{V}_{\widehat{P}_m}(\widehat{f}^{\mathsf{e}}) - \mathbb{V}_{\mathsf{e}_0}(f_w^{\mathsf{e}}) = \frac{1}{m}\sum_{j=1}^{m} \phi_{f_*^{\mathsf{e}_0},P^{\mathsf{e}_0}}(Z_j) + o_p(n^{-1/2}).$$

Proof. Write $P = P^{\mathsf{e}_0}$, $f_w^{\mathsf{e}} = f_*^{\mathsf{e}_0}$ by Lemma F.1, and $\widehat{f} = \widehat{f}^{\mathsf{e}}$. C2 gives

$$\begin{aligned}\mathbb{V}_{\widehat{P}_m}(\widehat{f}) - \mathbb{V}_P(f_*^{\mathsf{e}_0}) = (\widehat{P}_m - \mathbb{P}_P)\phi_{f_*^{\mathsf{e}_0},P} + \{\mathbb{V}_P(\widehat{f}) - \mathbb{V}_P(f_*^{\mathsf{e}_0})\} \\ + (\widehat{P}_m - \mathbb{P}_P)(\phi_{\widehat{f},P} - \phi_{f_*^{\mathsf{e}_0},P}) + o_p(n^{-1/2}).\end{aligned} \tag{15}$$

By C1 and C3, the population-score difference is $o_p(n^{-1/2})$. Conditional on all fitting data, the remaining empirical-average term has mean zero and variance at most

$$\frac{1}{m}\mathbb{P}_P\{(\phi_{\widehat{f},P} - \phi_{f_*^{\mathsf{e}_0},P})^2\} = o_p(n^{-1}). \tag{16}$$

Conditional Chebyshev's inequality and C4 make this term $o_p(n^{-1/2})$. This proves the expansion. For mean scores, C2 is exact, so no empirical-process differentiability argument is needed. □

Let $\phi_0 = \phi_{f_*^{\mathsf{e}_0},P^{\mathsf{e}_0}}$, with variance $\eta_0 \in (0,\infty)$. On the two independent reference halves, the preceding theorem gives

$$\widehat{\psi}_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)} = \frac{1}{2m_2}\sum_{Z\in\mathcal{D}_2^{\mathsf{e}_0}}\phi_0(Z) - \frac{1}{2m_1}\sum_{Z\in\mathcal{D}_1^{\mathsf{e}_0}}\phi_0(Z) + o_p(n^{-1/2}).$$

For $m_1 = m_2 = n/2$, the central limit theorem yields $\sqrt{n}\,\widehat{\psi}_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)} \xrightarrow{d} \mathcal{N}(0,\eta_0)$. C4 gives $(\widehat{\eta}_{\mathsf{e}_1} + \widehat{\eta}_{\mathsf{e}_2})/2 \xrightarrow{p} \eta_0$, and studentization proves the null limit. The positive variance assumption excludes degenerate or perfectly constant evaluation scores; the stated nonrejection convention does not assign a normal approximation to such cases.

For unequal evaluation halves with $m_j/n \to \rho_j \in (0,1)$ and $\rho_1 + \rho_2 = 1$, use instead

$$\widehat{\sigma}^2_{\mathsf{e}_0,\mathcal{S}} = \frac{n}{4}\left(\frac{\widehat{\eta}_{\mathsf{e}_1}}{m_2} + \frac{\widehat{\eta}_{\mathsf{e}_2}}{m_1}\right). \tag{17}$$

Its null limit is $\eta_0(\rho_2^{-1} + \rho_1^{-1})/4$. The equal-half formula follows by substitution. This distinction concerns evaluation sample sizes; unequal fitting-pool sizes are permitted when the required rates hold.

At fixed α, the normal limit gives asymptotic rejection probability α under a nondegenerate null. Whenever $H_0(\mathcal{S}^*)$ is not rejected, the intersection is a subset of $\mathcal{S}^*$; if the accepted family is empty our convention also returns a subset. Thus

$$\limsup_{n\to\infty}\mathbb{P}(\widehat{\mathcal{S}}^* \not\subseteq \mathcal{S}^*) \le \limsup_{n\to\infty}\mathbb{P}\{H_0(\mathcal{S}^*) \text{ is rejected}\} \le \alpha.$$

This argument requires no multiplicity adjustment across candidate subsets for this particular false-inclusion guarantee. It is distinct from simultaneous control of all subset-test errors.

For exact recovery, take $\alpha_n \to 0$ and $t_n = z_{1-\alpha_n/2} = o(\sqrt{n})$. Under the target null, $T_{\mathsf{e}_0,\mathcal{S}^*}^{(\mathsf{e}_1,\mathsf{e}_2)} = O_p(1)$ while $t_n \to \infty$, so the target is accepted with probability tending to one. For any candidate omitting a target coordinate, C5 and the assumed alternative consistency give

$$\frac{T_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)}}{\sqrt{n}} \xrightarrow{p} \frac{\psi_{\mathsf{e}_0,\mathcal{S}}^{(\mathsf{e}_1,\mathsf{e}_2)}}{\sigma_{\mathsf{e}_0,\mathcal{S}}} \neq 0.$$

Such a candidate is rejected with probability tending to one. A union bound over the fixed finite family shows that, with probability tending to one, every accepted subset contains $\mathcal{S}^*$ and $\mathcal{S}^*$ itself is accepted. Their intersection equals $\mathcal{S}^*$. For $\alpha_n = 1/\log n$, $t_n = O(\sqrt{\log\log n}) = o(\sqrt{n})$, as required. Fixed α supports the coverage statement but does not make acceptance of the target tend to one.

Theorem 3.2 follows by taking a mean score, for which $\phi_{f,P}(Z) = v(f(\boldsymbol{X}_\mathcal{S}),Y) - \mathbb{V}_P(f)$ and C2 has zero remainder. The same argument proves Theorem E.1 when its additional expansion and variance conditions hold. No null-only stationarity argument has been used to derive power under an alternative.

F.3 Proof of Proposition 3.3

Let $K = |\mathscr{E}|$ be fixed. Under $H_0(\mathcal{S})$ over $\mathscr{E}$, each internal reference e_i has the same conditional response law as its sources $\mathsf{e}_{i+1}, \mathsf{e}_{i+2}$. Lemma F.1 and Theorem F.2 apply to each comparison. After studentization, write $T_{i,\mathcal{S}}^{(i+1,i+2)} = L_{i,n} + o_p(1)$, where $L_{i,n}$ is the $\sqrt{n}$-scaled leading oracle

score contrast on the two evaluation halves of reference environment e_i, divided by $\sigma_{i,\mathcal{S}}$. Each $L_{i,n} \xrightarrow{d} \mathcal{N}(0,1)$. These leading terms are independent across environments by Assumption 2.1; no evaluation observation enters any nuisance fit or tuning step. Since K is fixed, the vector of remainders is $o_p(1)$ even when comparisons share fitting pools. Joint convergence and Slutsky's theorem therefore give

$$\left(T_{1,\mathcal{S}}^{(2,3)}, \ldots, T_{K,\mathcal{S}}^{(1,2)}\right) \xrightarrow{d} \mathcal{N}_K(0, I_K).$$

The continuous mapping theorem yields $Q_{\mathcal{S}} \xrightarrow{d} \chi^2_K$. The chi-square law is continuous, so the rejection probability at the fixed threshold $\chi^2_{K,1-\alpha}$ tends to α. Positive limiting variances make the probability of a zero estimated component variance tend to zero; the unresolved-test convention does not change this limit.

Suppose the null conditions hold at the tested candidate $\mathcal{S}^*$. Whenever that candidate is accepted, the intersection in (3) is contained in $\mathcal{S}^*$. Hence

$$\limsup_{n\to\infty} \mathbb{P}(\widehat{\mathcal{S}}^* \not\subseteq \mathcal{S}^*) \le \limsup_{n\to\infty} \mathbb{P}\{H_0(\mathcal{S}^*) \text{ is rejected}\} \le \alpha.$$

No multiplicity adjustment across candidate subsets is needed for this false-inclusion guarantee.

For each candidate omitting a target coordinate, consistency of the component contrasts and standardizers in C6 gives

$$\frac{T_{i,\mathcal{S}}^{(i+1,i+2)}}{\sqrt{n}} \xrightarrow{p} \frac{\psi_{i,\mathcal{S}}}{\sigma_{i,\mathcal{S}}}, \qquad \frac{Q_{\mathcal{S}}}{n} \xrightarrow{p} \Delta_{\mathcal{S}} = \sum_{i=1}^{K} \left(\frac{\psi_{i,\mathcal{S}}}{\sigma_{i,\mathcal{S}}}\right)^2 \ge c > 0.$$

Take $\alpha_n \to 0$ with $q_n = \chi^2_{K,1-\alpha_n} = o(n)$. Under the target null, $Q_{\mathcal{S}^*} = O_p(1)$ and $q_n \to \infty$, so the target is accepted with probability tending to one. Each candidate omitting a target coordinate is rejected with probability tending to one because $q_n/n \to 0$ and $Q_{\mathcal{S}}/n \xrightarrow{p} \Delta_{\mathcal{S}} > 0$. A union bound over the fixed finite family then shows that every accepted candidate contains $\mathcal{S}^*$ while $\mathcal{S}^*$ itself is accepted, with probability tending to one. Their intersection equals $\mathcal{S}^*$. For $\alpha_n = 1/\log n$ and $n > e$, Markov's inequality for a χ^2_K variable gives $q_n \le K/\alpha_n = K \log n = o(n)$. Fixed α yields false-inclusion control but does not ensure exact recovery.

For unequal retained reference evaluation sizes n_i with $n_i/n_0 \to \lambda_i \in (0,\infty)$, impose the component conditions at the corresponding n_i scales and use the appropriate standardizers, including (17) if the halves are unequal. The same joint null argument applies, while under a relevant alternative

$$\frac{Q_{\mathcal{S}}}{n_0} \xrightarrow{p} \sum_{i=1}^{K} \lambda_i \left(\frac{\psi_{i,\mathcal{S}}}{\sigma_{i,\mathcal{S}}}\right)^2 \ge \left(\min_i \lambda_i\right) c > 0.$$

Thus C6 still ensures separation and recovery when $q_n = o(n_0)$. Algorithm 2 uses equal retained sizes. Squaring eliminates sign cancellation but cannot create a nonzero contrast when all component contrasts vanish. For $K > 3$, changing the order can change the chosen source pairs and hence whether C6 holds; no universal power claim over orders or conditional alternatives follows.