PACE: PROXIMAL CAUSAL CONCEPT EXPLANATIONS WITH UNOBSERVED CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept-based explanations can reproduce a model's predictions yet yield biased causal attributions when unobserved concepts confound the target concept and model output. We introduce Proximal Causal Concept Explanations (PACE), which adapts proximal causal learning to estimate conditional mean concept effects without reconstructing unobserved concepts. PACE assigns two proxy measurements distinct roles: one enters an outcome bridge, while the other supplies conditional moments for learning it from factual observations. Under explicit causal, proxy, and bridge assumptions, averaged bridge contrasts identify these mean effects. We develop finite-feature moment estimators with optional predictive stabilization and characterize, for fixed features, how approximation, moment conditioning, and regularization affect population effect contrasts. Structural simulations demonstrate correction of confounding bias that predictive adjustment can retain. Analytical factor examples show that regression can recover a particular reference mean without recovering the bridge, while exact bridges can retain noise in individual contrasts. On CEBaB human-edited reviews, we evaluate intervention predictions for review-rating models using lexical features and pretrained language-model representations, including large language model embeddings. These same-review views have unverified proxy validity, and the comparisons show no consistent advantage over feature-matched regression. Together, these results distinguish bridge recovery, mean-effect accuracy, and individual-intervention prediction, highlighting the importance of measurement design and evaluation targets in causal concept explanation.

1 INTRODUCTION

Concept-based explanations express model behavior through human-understandable attributes such as food quality in a review or wing shape in an image. A causal concept explanation asks how a fixed model's output would change if one concept were altered while relevant context were preserved; query access alone does not construct this semantic intervention. For a review predictor built on pretrained language representations, including large language model (LLM) embeddings, this question concerns how its score would change if the expressed sentiment about food were altered while service and ambiance were held fixed. The causal concept effect (CaCE) averages this change; the individual causal concept effect (ICaCE) describes an input-specific change (Goyal et al., 2019; Abraham et al., 2022). However, even an explanation that accurately reproduces the model's factual predictions can still yield biased causal attributions.

Such bias can arise naturally when the concept vocabulary is incomplete. Concept discovery and label-free bottlenecks expand the available coordinates (Ghorbani et al., 2019; Oikarinen et al., 2023), but a predictive concept set need not contain every common cause. For example, unobserved service quality may affect both how food is described and a review model's score. Comparing food labels then mixes the intended intervention with changes in service. This is omitted-concept confounding, distinct from the information leakage and shortcut reliance studied in concept predictors (Mahinpei et al., 2021; Wang et al., 2022). Improving concept coverage or predictive completeness alone is generally insufficient to identify the causal effect under omitted-concept confounding.

We ask whether indirect measurements of omitted concepts can identify the required adjustment. Proximal causal learning uses an outcome-oriented measurement inside a bridge and a treatment-oriented

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.

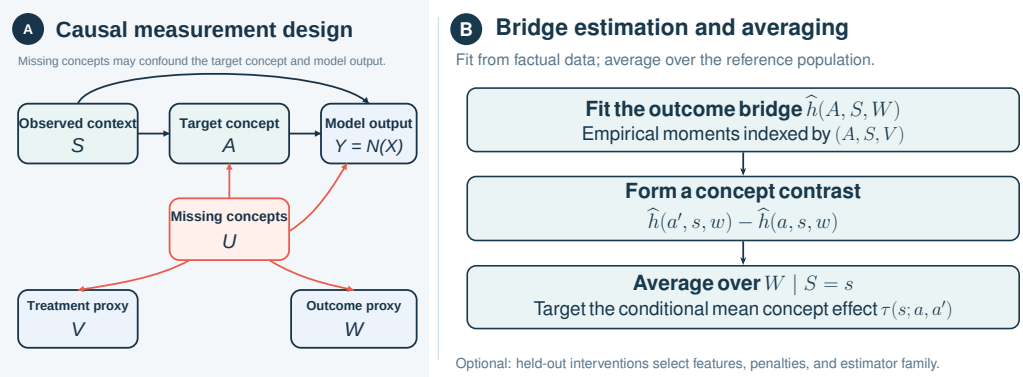


Figure 1: **PACE's measurement and estimation workflow.** An omitted concept U may confound A and Y. The outcome proxy W enters the bridge; V indexes its moments. Averaging contrasts over $W \mid S = s$ targets the mean effect. Context-to-proxy and optional A-to-V arrows are omitted for clarity. Intervention selection is separate from identification.

measurement to identify it through conditional moments (Miao et al., 2018; Tchetgen Tchetgen et al., 2024). Neither measurement must reconstruct the unobserved factor. Their validity depends on the measurement process: representations and metadata provide candidates, while separate pre-treatment records offer a stronger design argument than two transformations of one input (Chen et al., 2024).

We introduce PACE, which specializes this strategy to post-hoc concept explanation (Fig. 1). A bridge fitted to factual model outputs yields conditional mean concept effects through averaging. Finite-feature estimation permits predictive stabilization, while a joint counterfactual selector (Joint-CF) compares PACE with regressions using the same views when validation interventions are available. The resulting identification formula leaves a separate estimation question: whether fitting an outcome bridge yields accurate concept-effect contrasts under finite features and intervention-based selection. We investigate this question by separating bridge recovery, conditional-mean estimation, and individual-edit prediction.

Our contributions are threefold. **(i)** We formulate omitted-concept confounding as a proximal measurement problem and specialize outcome-bridge identification to conditional mean concept effects. **(ii)** We develop a finite-feature implementation and analyze how moment approximation, weakly constrained directions, and regularization affect estimated contrasts, using feature-matched regression to distinguish moment fitting from predictive adjustment. **(iii)** Through structural simulations, designed factor measurements, and human-edited reviews, we examine when bridge recovery translates into effect accuracy and when including proximal candidates improves intervention-based selection.

2 Related Work

Causal concept effects and intervention-based explanations. CaCE formalizes the effect of a semantic intervention on a model's prediction and uses generative modeling to approximate interventions that cannot be directly simulated (Goyal et al., 2019). CEBaB supports evaluation with human-written concept edits (Abraham et al., 2022). For language models, CausaLM constructs counterfactual representations through graph-informed adversarial tasks (Feder et al., 2021), whereas Causal Proxy Models learn intervenable surrogates using approximate counterfactual supervision (Wu et al., 2023). PACE shares the goal of explaining a fixed model's response to a semantic change, but obtains its identification argument from auxiliary measurements rather than a learned counterfactual generator or intervenable surrogate. PACE identifies conditional mean concept effects from factual observations under proximal assumptions; held-out edits, when available, support estimator selection and individual-edit evaluation.

Incomplete concepts and predictive representations. An incomplete concept vocabulary raises both a representation problem and an identification problem. Concept discovery, bottleneck models,

and post-hoc concept representations expand or organize the semantic information available to an explanation (Ghorbani et al., 2019; Koh et al., 2020; Yuksekgonul et al., 2023; Espinosa Zarlenga et al., 2022; Oikarinen et al., 2023). Completeness-aware discovery assesses how well concept scores capture a model's predictive information (Yeh et al., 2020). MCCE addresses unobserved concepts by augmenting observed annotations with representation-derived pseudo-concepts that are orthogonal to the observed concept matrix, and fitting a predictive surrogate (Gao & Chen, 2024). PACE instead assigns two measurements distinct roles: one enters an outcome bridge, while the other supplies conditional moments for estimating it. The distinction concerns the adjustment strategy and the assumptions identifying the averaged effect, rather than the semantic labeling of auxiliary coordinates.

Proximal identification and text measurements. Proxy-variable identification and proximal causal learning establish conditions under which auxiliary measurements identify effects despite unobserved confounding (Miao et al., 2018; Tchetgen Tchetgen et al., 2024). PACE specializes the outcome-bridge argument to a fixed model output and a specified concept transition, with completeness imposed within the stated bridge class. For text, Chen et al. (2024) infer proxies from distinct pre-treatment text instances and give identification conditions on their measurement process. This provides a direct connection between language representations and proximal inference: the relevant information must be accompanied by justified measurement restrictions. Our CEBaB analysis uses same-review views to evaluate intervention prediction; these views do not inherit the identification argument for separately collected text measurements.

Bridge estimation and effect-targeted learning. Conditional-moment estimation includes min-imax procedures, kernel proximal estimators, and neural bridge estimators (Dikkala et al., 2020; Mastouri et al., 2021; Xu et al., 2021; Kompa et al., 2022). PACE implements a finite-feature moment criterion with optional predictive stabilization. Holding bridge features fixed across moment estimation and regression allows the experiments to separate fitting-objective differences from representation capacity. Semiparametric and doubly robust proximal procedures develop effect estimation and inference (Cui et al., 2024; Ghassami et al., 2022), while the P-learner fits conditional average treatment effects through a tailored two-stage loss (Sverdrup & Cui, 2023). PACE uses an outcome-bridge plug-in estimator: it fits the bridge and averages its concept contrasts over a specified reference population. Our analysis examines how bridge approximation, moment stability, and the validation objective affect the resulting mean-effect estimates and individual-edit predictions.

3 Problem Setup

Let X be an input, $\mathcal{N}$ a fixed model, and $Y = \mathcal{N}(X) \in \mathbb{R}^L$ its factual output. The analyst observes a target concept A and other annotated concepts S; U denotes omitted common causes. We treat S and U as pre-intervention variables. Let X_a be the potential input under the specified semantic intervention and $Y_a = \mathcal{N}(X_a)$, with consistency $X = X_A$ and hence $Y = Y_A$. The intervention specifies how non-target input variation is preserved or generated. Latent ignorability is an assumption about this family, not a consequence of $\mathcal{N}$ being fixed. For levels a, a', the estimand is

$$\tau(s; a, a') = \mathbb{E}\{Y_{a'} - Y_a \mid S = s\}. \tag{1}$$

This estimand concerns the fixed model's response, not a real-world outcome. Its intervention and averaging population must be specified.

Let W denote an auxiliary outcome proxy, whose measurement role is specified in Section 4.1. For a fitted bridge h, we also evaluate the pointwise prediction at the factual context (S_x, W_x),

$$d_h(x; a, a') = h(a', S_x, W_x) - h(a, S_x, W_x). \tag{2}$$

Comparing this prediction with a paired edit response is an ICaCE-style prediction task. The theorem below identifies Eq. (1) after averaging over $W \mid S = s$; it does not identify a unit-level effect from Eq. (2).

4 METHOD

4.1 PROXIMAL MEASUREMENT DESIGN

The outcome-inducing proxy W enters the bridge $h(A, S, W)$; the treatment-inducing proxy V indexes its conditional moments without entering the reported contrast. The bridge residual must average to zero given (A, S, V), supplying restrictions beyond predictive adjustment through W.

The proxy restrictions concern the joint distribution of factual variables, including latent U:

$$Y \perp\!\!\!\perp V \mid A, S, U; \qquad W \perp\!\!\!\perp (A, V) \mid S, U. \tag{3}$$

Both proxies may depend on S, and V may depend on A. A sufficient illustrative design is $W = f_W(S, U, \varepsilon_W)$, $V = f_V(A, S, U, \varepsilon_V)$, and $Y = f_Y(A, S, U, \varepsilon_Y)$, with errors conditionally independent of each other and jointly independent of A given (S, U). The measurement process must justify these restrictions; different encoders or modalities alone do not. Direct target information in W can violate its exclusion, but prediction of A through U need not. Observable predictability or moment residuals cannot certify exclusions conditional on unobserved U.

Setting $V = W$ reduces the population restriction to $\mathbb{E}\{Y - h(A, S, W) \mid A, S, W\} = 0$. Its solution is $\mathbb{E}(Y \mid A, S, W)$ when this conditional mean belongs to the chosen class. Duplicating W, or using an invertible transformation as V, therefore adds no measurement information.[1] Structural duplication and permutation controls examine the contribution of distinct information in V (Appendix B.2).

4.2 IDENTIFICATION

Consistency and latent ignorability connect factual and potential outputs. With the proxy and bridge conditions below, they identify $\mathbb{E}(Y_a \mid S = s)$. Potential outputs and evaluated bridge sections are integrable under their reference laws at the levels considered.

Assumption 1 (Intervention and positivity)**.** *Consistency, $Y = Y_A$, holds. On the common treatment domain, the conditional mass or density of $A \mid S, U$ is positive almost everywhere for almost every (S, U).*

Assumption 2 (Latent ignorability)**.** *$\{Y_a\}_a \perp\!\!\!\perp A \mid S, U$ for the specified intervention family.*

Assumption 3 (Proxy exclusions)**.** *The measurements satisfy Eq. (3).*

Assumption 4 (Latent outcome bridge)**.** *There is an integrable h_0 in the specified bridge class such that*

$$\mathbb{E}(Y \mid A, S, U) = \mathbb{E}\{h_0(A, S, W) \mid A, S, U\}. \tag{4}$$

These conditions imply the observable restriction

$$\mathbb{E}\{Y - h_0(A, S, W) \mid A, S, V\} = 0. \tag{5}$$

The bridge generally differs from $\mathbb{E}(Y \mid A, S, W)$: a noisy W can leave residual confounding. Equation (4) instead requires its measurement-averaged value to reproduce the outcome mean at each latent state; V supplies observable restrictions for learning this correction without U.

Assumption 5 (Observable completeness)**.** *For almost every (a, s), $g \mapsto \mathbb{E}\{g(W) \mid A = a, S = s, V\}$ is injective on differences of bridge sections $h(a, s, \cdot)$.*

Theorem 1 (Conditional mean identification)**.** *Under Assumptions 1–5, Eq. (5) identifies h_0 within the bridge class up to $P_{A,S,W}$-almost-sure equality, and, for $P_{A,S}$-almost every (a, s),*

$$\mathbb{E}(Y_a \mid S = s) = \mathbb{E}\{h_0(a, s, W) \mid S = s\}. \tag{6}$$

Consequently, reference-averaged bridge contrasts identify Eq. (1) at these levels.

Appendix A.1 gives the proof and the regularity needed for pointwise evaluation at continuous treatment levels. Bridge existence concerns representability; completeness concerns uniqueness within the class. As in nonparametric instrumental-variable models, identification and a stable inverse are distinct (Newey & Powell, 2003).

[1] Single-proxy identification relies on additional structure (Park et al., 2024; Xu & Gretton, 2025); determinism of $\mathcal{N}(X)$ alone does not establish the required concept-level measurement conditions.

A linear illustration. In the linear model of Appendix B.2, let $t(s) = 0.5 + 0.5s$, with

$$A = 0.5S + U + \varepsilon_A, \qquad Y = t(S)A + U + \varepsilon_Y, \qquad W = U + \varepsilon_W, \qquad V = U + \varepsilon_V.$$

The errors are centered, mutually independent, and independent of (S, U). A unit increase in A has causal effect $t(S)$. Under the Gaussian specification in Appendix B.2, the predictive regressions are

$$\mathbb{E}(Y \mid A, S) = \{t(S) + 0.8\}A - 0.4S,$$
$$\mathbb{E}(Y \mid A, S, W) = \{t(S) + \tfrac{2}{3}\}A - \tfrac{1}{3}S + \tfrac{1}{6}W.$$

Thus, both observed-only regression and adjustment through the noisy proxy W overstate the target effect. In contrast,

$$h_0(a, s, w) = t(s)a + w, \qquad Y - h_0(A, S, W) = \varepsilon_Y - \varepsilon_W.$$

This residual has zero conditional mean given (A, S, V), although generally not given (A, S, W) because W contains ε_W. The moments indexed by V identify the correction through W, yielding the bridge contrast $t(s)(a' - a)$ without estimating each input's U.

4.3 Finite-feature estimation and matched controls

Let $b_i = b(A_i, S_i, W_i)$ be bridge features, $q_i = \phi(A_i, S_i, V_i)$ be moment-test features, and $R_i(\theta) = Y_i - h_\theta(A_i, S_i, W_i)$. Using a finite-feature generalized method of moments objective (Hansen, 1982), we minimize

$$\widehat{\theta} = \arg\min_\theta \left\| \frac{1}{n} \sum_i q_i R_i(\theta)^\top \right\|_F^2 + \lambda_{\text{pred}} \frac{1}{n} \sum_i \|R_i(\theta)\|_2^2 + \lambda_{\text{ridge}} \|\theta\|_D^2. \tag{7}$$

The first term fits feature–residual moments and is a finite-feature kernel criterion (Appendix A.2); for scalar outputs its Frobenius norm reduces to the Euclidean norm. The other terms penalize factual prediction error and non-intercept coefficients: $D = \text{diag}(0, 1, \ldots, 1)$, with matrix penalty $\text{tr}(\theta^\top D\theta)$. PACE-Moment sets $\lambda_{\text{pred}} = 0$; PACE-Stabilized uses $\lambda_{\text{pred}} > 0$; both allow ridge. PACE-CF selects their features and penalties using validation interventions. Nonvanishing penalties can displace coefficients and mean contrasts. Fixed-feature consistency assumes both penalties vanish and does not automatically cover intervention-selected fits (Appendices A.2–A.4).

For $h_\theta = b^\top\theta$, estimation solves the linear system in Appendix A.2, coordinatewise for vector outputs. Our random moment map uses standardized encoded (A, S, V) coordinates z and M joint cosine coordinates: $q = (1, z^\top, \sqrt{2/M}\cos(\Omega^\top z + \xi)^\top)^\top$. Appendix B.1 specifies the construction and separate analytic maps. Full column rank of $G = \mathbb{E}(qb^\top) \in \mathbb{R}^{m\times p}$ requires $m \geq p$, but that inequality ensures neither rank, completeness, nor stability.

For a discrete reference stratum s, let $\mathcal{R}_s$ index a nonempty reference sample drawn independently of fitting from $W \mid S = s$. With $\widehat{h} = h_{\widehat{\theta}}$, we estimate the conditional mean effect by

$$\widehat{\tau}(s; a, a') = \frac{1}{|\mathcal{R}_s|} \sum_{j\in\mathcal{R}_s} \{\widehat{h}(a', s, W_j) - \widehat{h}(a, s, W_j)\}.$$

Reference observations represent the averaging population, not a subset selected by factual treatment. Coefficient error affects this mean only through its projection onto the averaged feature contrast (Appendix A.3).

We study linear and generalized additive model (GAM) feature templates. With nonconstant random Fourier features (RFFs) $\psi(W)$ (Rahimi & Recht, 2007), the scalar-target GAM is

$$h_\theta = \theta_0 + \theta_A A + \theta_S^\top S + \theta_W^\top \psi(W) + \theta_{AW}^\top \{A\psi(W)\}. \tag{8}$$

For a J-level target, let $c_j(A)$ be indicators for the $J - 1$ non-reference categories. The categorical template is

$$h(A, S, W) = \alpha + \beta^\top c(A) + r(S) + f(W) + \sum_{j=1}^{J-1} c_j(A) g_j(W). \tag{9}$$

The functions f and g_j share an RFF basis with separate coefficients. These templates omit direct target–context interactions: context main effects cancel in contrasts, although reference means can depend on S through $P(W \mid S)$. Here, "complete" denotes target–proxy interactions for all non-reference categories, not all possible interactions. The structural polynomial span instead includes AS (Appendices B.2, B.4.2).

Reg-W minimizes mean squared prediction loss with PACE's bridge matrix and intercept treatment. Reg-WV uses both views, with full per-view capacity or the one-view bridge's total RFF budget. Observed-only uses (A, S). Appendix B.1 gives the mean-loss penalties, grids, and matched candidate counts.

4.4 Intervention-based selection

Coefficients are fitted to factual observations; mean paired ICaCE L_2 loss (mean Euclidean error, or MAE for scalar outputs) selects features and penalties. PACE-CF selects among PACE candidates. Joint-CF compares the selected PACE, Reg-W, full Reg-WV, and observed-only candidates. Choices are fixed before testing.

All tuning and selection use the same K intervention pairs. Joint-CF-noPACE (NoPACE) excludes only PACE; predictive candidates, budget draws, and tie-breaking match Joint-CF. Their paired test-loss difference measures the value of including PACE. Calibration adds the mean or median validation effect residual to the selected observed-only predictor (Appendix B.5). K excludes model and proxy-transformation training data (Appendix B.4).

Pair loss assesses individual intervention prediction, not the reference-mean target in Theorem 1. In the factor examples, proxy noise raises exact-bridge pair error but cancels under reference averaging, so pair-optimal tuning need not be mean-optimal. A separate diagnostic selects by squared reference-group mean error on 1,000 independently labeled reference-validation rows, changing both loss and population and using additional intervention supervision.

5 Experiments

Structural models test recovery under confounding; factor simulations examine specification, stability, and mean estimation; human-edited reviews assess intervention prediction. Comparisons use feature-matched regressions, not external explainers (controls: Appendix B.6). Table 4 distinguishes analytic support from the known generators, finite-class numerical evidence, and assumptions unverified for the text views.

Structural simulations. A Gaussian confounder U influences a continuous target A and outcome Y, while $W = U + \varepsilon_W$ and $V = U + \varepsilon_V$ provide measurements with independent errors. With context $S \sim \text{Bernoulli}(1/2)$, the outcome is $Y = (0.5 + 0.5S)A + f(U) + \varepsilon_Y$, where f is linear or cubic. A one-unit increase in A has true effect $0.5 + 0.5S$. Bridge features contain the true bridge. Each model uses 50 replicates per training size ($n = 1{,}000$ or $5{,}000$). Appendix B.2 gives the generators, fitted spans, and fixed-penalty mechanism controls.

Factor simulations with designed proxies. For the dSprites and 3D Shapes rendering-factor grids (Matthey et al., 2017; Burgess & Kim, 2018), outcomes are analytic functions of factors, not image-model outputs. The target is x-position in dSprites and orientation in Shapes; assignment depends on orientation and object hue, respectively. Shape and scale form context. Both target levels have assignment probability at least 0.1 and occur in training. The proxies measure non-target factors with independent errors. The confounding parameter is $c \in \{0, 2\}$, with $c = 0$ providing an unconfounded control, and proxy noise is $\sigma \in \{0.2, 0.8\}$. Reference evaluation fixes a low-to-high transition under a product-uniform distribution of non-target factors.

Two independent 20-replicate studies use $n = 4{,}000$. One compares correct bridges with analytic or RFF moments and RFF bridges with RFF moments; the other evaluates RFF estimators selected by pair loss on sampled and exact reference populations. Known spans use structural knowledge; coefficients are fitted to factual data without latent covariates (Appendix B.3).

Text intervention prediction. CEBaB provides human-edited restaurant reviews (Abraham et al., 2022). Food is the target concept, restaurant-noise annotations are withheld, and ambiance and service form context. Views use term frequency–inverse document frequency (TF–IDF) word and character decompositions or distinct pooling rules from a frozen BERT, RoBERTa, or Qwen3-Embedding encoder (Devlin et al., 2019; Liu et al., 2019; Zhang et al., 2025). The explained output is a scalar ridge prediction of the 1–5 overall review rating, without rescaling or clipping; errors are in rating points. Models are fixed across explainers, and comparisons are within representation. Score-model fitting uses training edits, whereas explainer fitting uses factual rows. Restricted and complete categorical bridges share the data, views, and 512 random moment features. Five seeds per representation vary the estimator, not the reviews. The official split keeps edits of each original review together and was examined during method development (Appendix B.4).

Evaluation. Pair error is mean Euclidean error of predicted edit responses, or mean absolute error (MAE) for scalar outputs. Mean-effect evaluation reports weighted root mean squared error (RMSE) of group-averaged effects for a fixed transition, with groups defined by S. We report finite-reference and exact-population errors. Methods are paired within replicate; text intervals resample original reviews and condition on fitted candidates. Appendices B.3–B.4 specify reference evaluation and review-cluster uncertainty.

6 Results

6.1 Controlled recovery of concept effects

PACE-Moment improves with training size. At $n = 5{,}000$, its pair MAE is 0.029 (linear) and 0.059 (cubic), versus 0.077 and 0.092 for Reg-W (Table 1). These controls use intervention-selected regularization, so contrasts need not match the unpenalized population regressions in Section 4.2. Fixed-penalty cubic controls reproduce Reg-W with $V = W$ but have large errors when V is permuted (Table 6; Appendix B.2).

At $n = 1{,}000$ in the cubic model, full Reg-WV outperforms PACE-Moment. PACE-CF and Joint-CF reduce prediction error relative to PACE-Moment, whereas the larger sample favors moment estimation. Thus, identification does not determine which finite-sample procedure is most accurate.

Table 1: **Structural-model pair prediction.** Mean test MAE over 50 seeds. PACE-Moment tunes ridge; PACE-CF also allows predictive stabilization. Table 5 reports uncertainty for the PACE variants, Reg-W, and Reg-WV; Table 6 gives fixed-penalty controls.

Setting	Observed	Reg-W	Reg-WV	PACE-Moment	PACE-CF	Joint-CF
Linear, $n = 1{,}000$	0.080	0.078	0.064	0.060	0.054	0.043
Linear, $n = 5{,}000$	0.078	0.077	0.063	0.029	0.025	0.023
Cubic, $n = 1{,}000$	0.101	0.096	0.079	0.101	0.084	0.068
Cubic, $n = 5{,}000$	0.109	0.092	0.082	0.059	0.055	0.053

6.2 Mean effects, bridge specification, and moment stability

The factor simulations distinguish recovery of an outcome bridge from recovery of a reference mean. Reference means are constant across contexts, although individual effects vary with the assignment-driving factor. Observational contrasts are unbiased at $c = 0$; at $c = 2$, bias is small and constant in dSprites but context-dependent in Shapes. Full Reg-WV has the lowest population mean-effect error among pair-selected RFF estimators (Table 2); Tables 7–8 report reference-mean and pair errors.

Mean-effect accuracy need not track bridge recovery: measurement noise in an exact bridge's individual contrasts averages to zero under the reference law, and symmetry in Shapes allows unpenalized regression in the analytic feature class to recover the reference mean without satisfying the bridge equation (Appendix B.3.2).

With correct analytic features, unpenalized moment estimators attain low-noise population RMSEs of 0.0086 and 0.0070 in dSprites and Shapes. At high noise, sample moment inversion is unstable

despite full population rank: Shapes RMSE falls from 4.48 at $n = 4{,}000$ to 0.122 at $n = 16{,}000$ (Table 10). The RFF bridge leaves a nonzero approximation residual for the latent outcome mean, indicating error beyond finite-sample instability.

With correctly specified analytic features at high noise, mean-targeted rather than pair-targeted validation reduces ridge-regularized PACE-Moment RMSE from 0.106 to 0.023 in dSprites and 0.095 to 0.031 in Shapes (Fig. 2); Reg-W remains competitive or better. This diagnostic uses the additional reference-intervention labels described in Section 4.4.

Table 2: **Exact-population conditional-mean comparison.** Weighted group RMSE (mean ± standard error, 20 replicates) for RFF estimators selected by pair loss at $c = 2$, integrated over the known reference law. Table 7 gives finite-reference errors and $c = 0$ controls for the same fits.

Setting	Observed	PACE-Moment	Reg-W	Reg-WV	PACE-CF
dSprites, $\sigma = 0.2$	0.024 ± 0.002	0.026 ± 0.002	0.017 ± 0.001	0.013 ± 0.001	0.017 ± 0.001
dSprites, $\sigma = 0.8$	0.024 ± 0.002	0.037 ± 0.005	0.019 ± 0.001	0.018 ± 0.001	0.023 ± 0.002
Shapes, $\sigma = 0.2$	0.097 ± 0.003	0.037 ± 0.003	0.022 ± 0.001	0.017 ± 0.001	0.022 ± 0.001
Shapes, $\sigma = 0.8$	0.097 ± 0.003	0.059 ± 0.006	0.037 ± 0.002	0.034 ± 0.002	0.041 ± 0.002

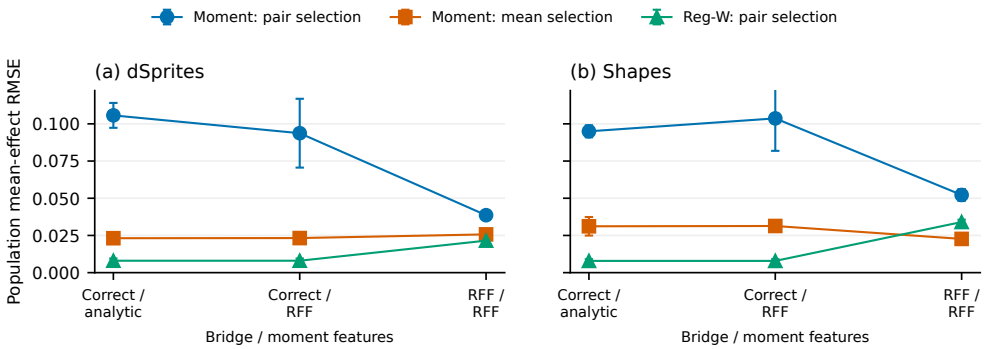


Figure 2: **Bridge specification and selection target.** Population mean-effect RMSE ($c = 2$, $\sigma = 0.8$, $n = 4{,}000$; 20 seeds; bars: one standard error). Moment denotes PACE-Moment. Mean selection uses 1,000 additional labeled reference-validation rows, changing both loss and population. Table 9 reports the complete bridge–moment comparison at both noise levels.

6.3 Text prediction with complete categorical target–proxy interactions

Table 3 compares complete categorical target–proxy bridges. Paired PACE-CF–Reg-W intervals contain zero for all four representations. The PACE-CF–Reg-W ranking also varies for signed errors averaged within context and directed-transition groups (Appendix B.4.3). These same-review views lack an established proxy-identification argument; neither predictive comparison isolates deconfounding from approximation, regularization, or prediction-target differences. Reg-WV feature-budget controls assess additional predictive capacity (Table 13).

Table 3: **Text intervention prediction.** Test MAE in rating points (five estimator seeds), with complete categorical target–proxy interactions and 512 random moment coordinates. Comparisons are within each explained model; paired PACE-CF–Reg-W intervals: Table 12.

Setting	Observed	Reg-W	Reg-WV	PACE-Moment	PACE-CF	Joint-CF
CEBaB: TF–IDF	0.521	0.494	0.494	0.485	0.485	0.494
CEBaB: BERT	0.667	0.669	0.665	0.670	0.670	0.665
CEBaB: RoBERTa	0.499	0.490	0.496	0.491	0.491	0.496
CEBaB: Qwen	0.626	0.652	0.654	0.652	0.652	0.626

We compare the complete class with a nested bridge restricted to food-Negative–proxy interactions, making Positive–Unknown contrasts proxy-independent; these transitions account for 128 of 449

test pairs (Table 11). Adding the food-Positive interaction block lowers PACE-CF's mean MAE by 0.012–0.023 across the four representations, using the same moment map (Table 12). Paired original-review cluster intervals exclude zero for BERT and RoBERTa, but not TF–IDF or Qwen.

6.4 When does including PACE improve selection?

Figure 3 compares selection with and without PACE as the intervention-validation budget K varies. In the cubic model ($n = 5{,}000$), Joint-CF reduces test MAE relative to NoPACE by 0.0304 at $K = 10$ and 0.0287 at $K = 100$. In contrast, at $K = 10$, including PACE increases error across all four CEBaB representations; the TF–IDF increase is 0.0281. With full validation, Joint-CF and NoPACE select the same predictive candidates in all text fits. For TF–IDF, median residual calibration also performs worse than NoPACE at $K = 10$ and 100 (Appendix B.5).

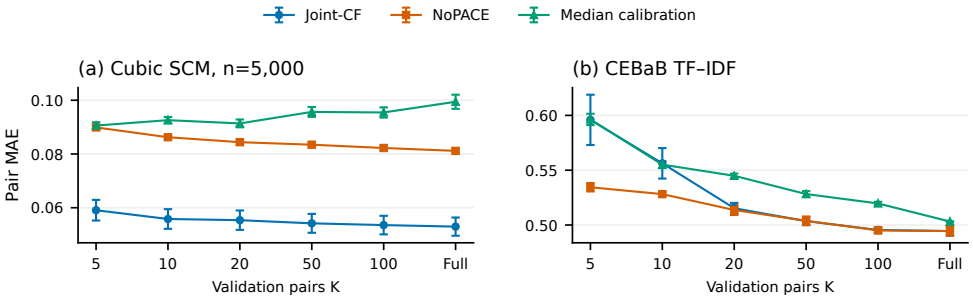


Figure 3: **Including PACE in intervention-based selection.** Test MAE versus validation-pair budget K. Joint-CF includes PACE; NoPACE selects only among predictive regressions. Median calibration adjusts observed-only effect predictions by the median validation residual. Points and bars show means and ± 1 standard error across 50 simulation replicates (a) or five estimator seeds on the same reviews (b), after averaging 20 shared validation subsets per non-full budget within each seed. "Full" uses 1,000 and 431 validation pairs, respectively. Appendix B.5 reports paired comparisons and review-cluster intervals.

7 Discussion

Under the stated assumptions, PACE identifies mean concept effects from auxiliary measurements without semantically labeling each missing factor. Identification and intervention-prediction accuracy require separate evaluation: a predictor can recover a particular reference mean without recovering the bridge, while a valid bridge can retain measurement noise in individual contrasts. Accurate edit prediction therefore does not establish removal of omitted-concept confounding. Averaging can cancel individual-edit errors, so accuracy need not transfer across reference distributions. Effects should therefore be reported with their intervention and averaging population. Validation should match the target: the mean-targeted diagnostic changes both the objective and reference supervision. Finite-sample estimation balances flexibility, moment strength, and validation resources. Bridge specification determines which contrasts can depend on the proxy, as the categorical ablation illustrates; tuning cannot create excluded interactions. More estimators can improve attainable accuracy yet worsen selection under noisy validation (Cawley & Talbot, 2010).

Limitations and future directions. Our estimator theory assumes fixed feature maps and does not cover learned or growing feature classes or post-selection inference. Some simulations use known bridge spans; the factor studies rely on designed proxies and analytic outcomes whose reference mean effects are constant across contexts. Our CEBaB evaluation considers one target–withheld-aspect configuration on a public split examined during development. Its same-review views lack an established proxy-identification argument, so these results assess intervention prediction rather than validate causal identification. Each representation also defines a different explained model, so cross-representation comparisons do not isolate proxy quality. Future work should evaluate separately collected proxies with justified measurement roles, compare effect-targeted proximal learners (Sverdrup & Cui, 2023), and develop uncertainty estimates that account for feature learning and estimator selection.

Reproducibility statement

Appendix A provides the theoretical results and proofs. Appendix B specifies the experimental designs, estimation and evaluation procedures, and supporting analyses. The accompanying code provides experiment configurations, numerical tests, and scripts for running the experiments. A separate archive contains saved numerical outputs for regenerating numerical tables and figures without refitting models. The README documents dependencies, execution commands, and the scope of verified reproduction.

Ethics statement

Experiments use public research benchmarks and involve no new human-subject data collection. Concept explanations can be overinterpreted when intervention semantics, support, or measurement assumptions are weak. Causal interpretations require a justified identification argument; otherwise, intervention predictions should be described as predictive rather than identified causal effects.

AI use statement

Generative AI tools including Research Pro and Codex were used to assist us with the formulation and refinement of mathematical claims related to the main theorem, the development and verification of intermediate derivations and proof steps, and the implementation of parts of the proposed methodology. They were also used for software-code development and debugging, LaTeX formatting, and language editing. All AI-assisted mathematical arguments, proof steps, and implementations were independently checked by the authors. The authors made the final decisions regarding the theorem statements, assumptions, methodology, interpretation of results, and scientific conclusions, and independently verified the theoretical claims, experimental results, citations, and code used in the paper. The authors take full responsibility for the final content of the work.

References

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney K. Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(3):998–1028, 2026. doi: 10.1093/jrsssb/qkaf075.

Chris Burgess and Hyunjik Kim. 3D Shapes Dataset. https://github.com/google-deepmind/3d-shapes, 2018.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010.

Jacob M. Chen, Rohit Bhattacharya, and Katherine A. Keith. Proximal causal inference with text data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric J. Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024. doi: 10.1080/01621459.2023.2191817.

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*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Nishanth Dikkala, Greg Lewis, Lester W. Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021. doi: 10.1162/coli_a_00404.

Jifan Gao and Guanhua Chen. MCCE: Missingness-aware causal concept explainer. *arXiv preprint arXiv:2411.09639*, 2024.

Amiremad Ghassami, Andrew Ying, Ilya Shpitser, and Eric Tchetgen Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7210–7239, 2022.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (CaCE). *arXiv preprint arXiv:1907.07165*, 2019.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. doi: 10.2307/1912775.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020.

Benjamin Kompa, David R. Bellamy, Thomas Kolokotrones, James M. Robins, and Andrew L. Beam. Deep learning methods for proximal inference via maximum moment restriction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11189–11201, 2022.

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.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7512–7523, 2021.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. https://github.com/google-deepmind/dsprites-dataset, 2017.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018. doi: 10.1093/biomet/asy038.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Chan Park, David B. Richardson, and Eric J. Tchetgen Tchetgen. Single proxy control. *Biometrics*, 80(2):ujae027, 2024. doi: 10.1093/biomtc/ujae027.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Erik Sverdrup and Yifan Cui. Proximal causal learning of conditional average treatment effects. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33285–33298, 2023.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024. doi: 10.1214/23-STS911.

Jiaxuan Wang, Sarah Jabbour, Maggie Makar, Michael Sjoding, and Jenna Wiens. Learning concept credible models for mitigating shortcuts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33343–33356, 2022.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. Causal proxy models for concept-based model explanations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37313–37334, 2023.

Liyuan Xu and Arthur Gretton. Kernel single proxy control for deterministic confounding. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3736–3744. PMLR, 2025.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26264–26275, 2021.

Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565, 2020.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Supplement to

"PACE: Proximal Causal Concept Explanations with Unobserved Concepts"

A Theoretical results

A.1 Proof of conditional-mean identification

We use regular conditional distributions and interpret the intervention family as a random element of a function space with jointly measurable evaluation. Positivity is with respect to counting measure for a discrete target and Lebesgue measure on the common treatment domain for a continuous target. At the stated levels, we assume $\mathbb{E}(\|Y_a\|_2 \mid S = s) < \infty$ and $\mathbb{E}\{\|h(a, s, W)\|_2 \mid S = s\} < \infty$ for admissible bridge sections; reference integrability is not implied by null-set domination alone. By iterated expectation,

$$\begin{aligned}&\mathbb{E}\{Y - h_0(A, S, W) \mid A, S, V\}\\ &\quad = \mathbb{E}\left[\mathbb{E}\{Y - h_0(A, S, W) \mid A, S, U, V\} \mid A, S, V\right]. \end{aligned} \tag{10}$$

The first exclusion gives $\mathbb{E}(Y \mid A, S, U, V) = \mathbb{E}(Y \mid A, S, U)$. The second gives $\mathbb{E}\{h_0(A, S, W) \mid A, S, U, V\} = \mathbb{E}\{h_0(A, S, W) \mid A, S, U\}$. The latent bridge makes their difference zero, proving Eq. (5). If two bridge functions solve this moment, then for almost every (a, s) their difference has zero conditional expectation given $(A = a, S = s, V)$. Completeness on differences of bridge sections implies equality up to $P_{A,S,W}$-null sets within the class.

For $P_{A,S}$-almost every (a, s), write p for conditional masses or densities. Positivity gives

$$\frac{dP_{U|A=a,S=s}}{dP_{U|S=s}}(u) = \frac{p(a \mid s, u)}{p(a \mid s)} > 0 \quad P_{U|S=s}\text{-almost surely.}$$

Thus these latent laws have the same null sets. The shared kernel $P_{W|S,U}$ implied by $W \perp\!\!\!\perp A \mid S, U$ gives $P_{W|S=s} \ll P_{W|A=a,S=s}$. Observationally equal bridge sections therefore have equal, integrable reference averages. This applies to both discrete and continuous treatments within the stated almost-everywhere scope.

Consistency, joint latent ignorability, positivity, and the latent bridge now give, at these levels,

$$\begin{aligned}\mathbb{E}(Y_a \mid S = s) &= \int \mathbb{E}(Y \mid A = a, S = s, U = u)\, dP(u \mid S = s)\\ &= \int \mathbb{E}\{h_0(a, s, W) \mid A = a, S = s, U = u\}\, dP(u \mid S = s). \end{aligned} \tag{11}$$

Because $W \perp\!\!\!\perp A \mid S, U$, the integrand is $\mathbb{E}\{h_0(a, s, W) \mid S = s, U = u\}$. Integrating yields Eq. (6); subtracting the expressions for a' and a gives the target contrast. These conclusions initially apply only at the almost-everywhere identified levels.

Continuous treatment levels. For continuous A, observable moments alone cannot distinguish changes confined to a single treatment value. A sufficient pointwise extension is to require, for almost every s, that both $a \mapsto \mathbb{E}(Y_a \mid S = s)$ and $a \mapsto \mathbb{E}\{h(a, s, W) \mid S = s\}$ are continuous on the topological support of $P_{A|S=s}$ for every admissible bridge h. A continuous difference nonzero at a support point is nonzero on a relative neighborhood of positive probability, contradicting almost-everywhere equality. Neither compactness nor connectedness is required. This identifies the reference average, not a pointwise bridge value for every w, and does not extrapolate beyond support. The polynomial structural models in Appendix B.2, with their polynomial bridge classes and finite moments, satisfy these conditions.

A.2 Fixed-feature consistency

For scalar Y, let $b = b(A, S, W) \in \mathbb{R}^p$ and $q = \phi(A, S, V) \in \mathbb{R}^m$ be column vectors of fixed measurable features, and write $h_\theta = b^\top \theta$. The first bridge coordinate is an intercept, so $D = \mathrm{diag}(0, 1, \ldots, 1)$. Define

$$G = \mathbb{E}(qb^\top) \in \mathbb{R}^{m\times p}, \qquad r = \mathbb{E}(qY) \in \mathbb{R}^m, \qquad B = \mathbb{E}(bb^\top) \in \mathbb{R}^{p\times p}, \qquad c = \mathbb{E}(bY) \in \mathbb{R}^p.$$

Proposition 1 (Consistency on a fixed nonredundant feature span). *Suppose* $(q_i, b_i, Y_i)_{i=1}^n$ *are independent and identically distributed,* p *and* m *are fixed, and* $\mathbb{E}(\|q\|_2^2 + \|b\|_2^2 + Y^2) < \infty$. *Assume* $\operatorname{rank}(G) = p$, *which requires* $m \geq p$. *Let* $\lambda_{\mathrm{pred},n}, \lambda_{\mathrm{ridge},n} \geq 0$ *converge to zero in probability. Any measurable minimizer* $\widehat{\theta}$ *of Eq. (7) over* $\mathbb{R}^p$ *then satisfies*

$$\widehat{\theta} \xrightarrow{p} \theta_* = (G^\top G)^{-1} G^\top r.$$

Consequently, $\ell^\top \widehat{\theta} \xrightarrow{p} \ell^\top \theta_*$ *for every fixed* $\ell \in \mathbb{R}^p$. *If* $r = G\theta_0$, *then* $\theta_* = \theta_0$.

Proof. Define

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^n q_i b_i^\top, \quad \widehat{r} = \frac{1}{n}\sum_{i=1}^n q_i Y_i, \quad \widehat{B} = \frac{1}{n}\sum_{i=1}^n b_i b_i^\top, \quad \widehat{c} = \frac{1}{n}\sum_{i=1}^n b_i Y_i.$$

For a scalar outcome, Eq. (7) is

$$L_n(\theta) = \|\widehat{r} - \widehat{G}\theta\|_2^2 + \lambda_{\mathrm{pred},n}\frac{1}{n}\sum_{i=1}^n (Y_i - b_i^\top\theta)^2 + \lambda_{\mathrm{ridge},n}\theta^\top D\theta.$$

Its normal equation is

$$\underbrace{\left(\widehat{G}^\top\widehat{G} + \lambda_{\mathrm{pred},n}\widehat{B} + \lambda_{\mathrm{ridge},n}D\right)}_{\widehat{H}_n}\widehat{\theta} = \widehat{G}^\top\widehat{r} + \lambda_{\mathrm{pred},n}\widehat{c}. \tag{12}$$

By Cauchy–Schwarz, the second-moment assumptions make every entry of $qb^\top$, qY, $bb^\top$, and bY integrable. The law of large numbers gives $(\widehat{G}, \widehat{r}, \widehat{B}, \widehat{c}) \xrightarrow{p} (G, r, B, c)$. Thus $\widehat{H}_n \xrightarrow{p} G^\top G$ and the right-hand side converges to $G^\top r$. Since $G^\top G$ is positive definite, $\widehat{H}_n$ is positive definite with probability tending to one; on this event the minimizer is unique. Continuity of matrix inversion gives consistency, and the continuous mapping theorem gives the contrast result.

For fixed-dimensional vector outputs $Y \in \mathbb{R}^L$, write $h_\Theta = \Theta^\top b$ with $\Theta \in \mathbb{R}^{p\times L}$ and replace $\widehat{r}, \widehat{c}$ by $n^{-1}\sum_i q_i Y_i^\top \in \mathbb{R}^{m\times L}$ and $n^{-1}\sum_i b_i Y_i^\top \in \mathbb{R}^{p\times L}$. The Frobenius objective separates over output coordinates, so the same proof applies coordinatewise under the corresponding moment and integrability conditions. Independent random-feature draws may be held fixed in this argument; feature maps estimated from the fitting sample or growing with sample size require additional convergence arguments.

Kernel-moment interpretation. With $z_i = (A_i, S_i, V_i)$ and $k_\phi(z_i, z_j) = \phi(z_i)^\top\phi(z_j) = q_i^\top q_j$, the moment term in Eq. (7) satisfies

$$\left\|\frac{1}{n}\sum_i q_i R_i(\theta)^\top\right\|_F^2 = \frac{1}{n^2}\sum_{i,j} k_\phi(z_i, z_j) R_i(\theta)^\top R_j(\theta).$$

It is therefore a finite-feature kernel moment criterion as well as a finite-dimensional generalized method of moments criterion. Predictive stabilization and coefficient regularization add the other two terms in Eq. (7).

A.3 Contrast error, approximation, and weak moments

Proposition 1 targets the population moment minimizer even under moment misspecification; this is not by itself causal consistency. The finite moment condition $r = G\theta_0$ alone does not establish that $b^\top\theta_0$ is a valid outcome bridge. For its causal interpretation, suppose $h_0 = b^\top\theta_0$ satisfies the identification assumptions of Theorem 1. For levels a, a' and context s within the theorem's almost-everywhere scope, or its pointwise extension in Appendix A.1, assume $b(a, s, W)$ and $b(a', s, W)$ are integrable under $W \mid S = s$. Define

$$\ell_s = \mathbb{E}\{b(a', s, W) - b(a, s, W) \mid S = s\} \in \mathbb{R}^p.$$

Then $\tau(s; a, a') = \ell_s^\top\theta_0$, and the sample estimator's contrast error under the population reference law is $\ell_s^\top(\widehat{\theta} - \theta_0)$. Errors orthogonal to ℓ_s cancel. The Shapes calculation in Appendix B.3.2 illustrates this distinction. If reference averaging replaces ℓ_s by $\widehat{\ell}_s$, consistency additionally requires $\widehat{\ell}_s \xrightarrow{p} \ell_s$; coefficient consistency alone does not remove error from a fixed finite reference sample.

Approximation and weak moments. Allow moment misspecification, while retaining $\mathrm{rank}(G) = p$. Define the unpenalized population fit and the left pseudoinverse by

$$\theta_* = \arg\min_{\theta \in \mathbb{R}^p} \|r - G\theta\|_2^2 = G^\dagger r, \qquad G^\dagger = (G^\top G)^{-1} G^\top.$$

For any candidate approximation θ_m, let $\delta_m = \|r - G\theta_m\|_2$. Since $G^\dagger G = I_p$,

$$\theta_* - \theta_m = G^\dagger (r - G\theta_m), \qquad \|\theta_* - \theta_m\|_2 \le \|G^\dagger\|_{\mathrm{op}} \delta_m = \frac{\delta_m}{\sigma_{\min}(G)}.$$

Here $\|\cdot\|_{\mathrm{op}}$ is the induced Euclidean norm and $\sigma_{\min}(G) > 0$ is the smallest singular value of G. The contrast-specific bound is

$$|\ell_s^\top (\theta_* - \theta_m)| \le \|\ell_s^\top G^\dagger\|_2 \, \delta_m.$$

Writing $G = U_G \Sigma V_G^\top$, with right singular vectors v_j,

$$\|\ell_s^\top G^\dagger\|_2^2 = \ell_s^\top (G^\top G)^{-1} \ell_s = \sum_{j=1}^{p} \frac{(v_j^\top \ell_s)^2}{\sigma_j(G)^2}.$$

Weak moment directions matter according to the contrast's projection onto them. This is population sensitivity to moment discrepancies, not a sampling-variance formula. These bounds concern a moment residual, not the full error relative to a true bridge. For a causal target $\tau(s; a, a')$, the total population contrast error also includes approximation error:

$$|\ell_s^\top \theta_* - \tau(s; a, a')| \le \|\ell_s^\top G^\dagger\|_2 \, \delta_m + |\ell_s^\top \theta_m - \tau(s; a, a')|.$$

Small singular values can amplify moment discrepancies in the specified feature coordinates. Removing exact duplicate bridge coordinates is necessary to avoid their induced rank deficiency, but nonredundancy alone does not ensure full rank or well-conditioned moments.

More generally, a linear functional can be estimable even when its nuisance functions are weakly identified (Bennett et al., 2026). The contrast-specific bound above makes this distinction explicit for the finite-feature moment system.

A.4 Population effects of predictive and ridge penalties

Return to $r = G\theta_0$ and fix $\lambda = \lambda_{\mathrm{pred}} \ge 0$ and $\rho = \lambda_{\mathrm{ridge}} \ge 0$. The population objective is

$$L_{\lambda,\rho}(\theta) = \|r - G\theta\|_2^2 + \lambda \mathbb{E}\{(Y - b^\top \theta)^2\} + \rho \theta^\top D \theta.$$

Let $e_0 = Y - b^\top \theta_0$ and $M_{\lambda,\rho} = G^\top G + \lambda B + \rho D$. Under the full-column-rank condition, $M_{\lambda,\rho}$ is positive definite. Its unique minimizer satisfies $M_{\lambda,\rho}\theta_{\lambda,\rho} = G^\top r + \lambda c$. Subtracting $M_{\lambda,\rho}\theta_0$ yields

$$\theta_{\lambda,\rho} - \theta_0 = M_{\lambda,\rho}^{-1}\{\lambda \mathbb{E}(b e_0) - \rho D \theta_0\}. \tag{13}$$

The moment restriction implies $\mathbb{E}(q e_0) = \mathbf{0}_m$, but does not generally imply $\mathbb{E}(b e_0) = \mathbf{0}_p$. Fixed predictive or ridge penalties can therefore change the population coefficients. The displacement of a particular effect is $\ell_s^\top(\theta_{\lambda,\rho} - \theta_0)$, which may vanish even when the coefficient displacement is nonzero. Positive predictive stabilization need not produce nonzero displacement in every model, nor guarantee lower finite-sample variance or risk. Vanishing penalties suffice for Proposition 1; they are not necessary for every correctly estimated contrast.

B Experimental details and supporting analyses

B.1 Shared estimation and evaluation procedures

Compared methods share factual outcomes, proxy realizations, validation pairs, and test pairs within each replicate.

Loss normalization and features. Predictive regression minimizes mean squared prediction loss plus $\lambda\|\theta\|_D^2$. A solver based on summed cross-products implements this objective using $X^\top X + \alpha_{\text{sum}} D$, with $\alpha_{\text{sum}} = n\lambda$. Supplying λ directly instead gives an effective mean-loss penalty of λ/n.

Order-preserving column selection uses two reorthogonalization passes to detect dependencies at tolerance $10^{-12}\max\{\max_j \|X_j\|_2, 1\}$, using training rows only. The fitted matrix retains the original selected columns, not the orthogonalized residuals; this is not whitening. Retained indices are reused for all evaluation rows. The structural polynomial span includes AS, whereas the scalar GAM and categorical bridges omit target–context interactions. Observed-only is a separate predictive template with first-target-coordinate–context products, not the bridge with its proxy block simply removed.

Random-feature construction. For random moment features, we first center and scale W and V coordinatewise on factual training rows. The conditioning vector concatenates encoded (A, S), in that order, followed by standardized V; it is again centered and scaled on those rows, including indicator coordinates. Denote the result by z. Scales are empirical standard deviations with divisor n, with values below 10^{-8} replaced by one. The map is

$$q(z) = \left(1, z^\top, \sqrt{2/M}\cos(\Omega^\top z + \xi)^\top\right)^\top, \quad \omega_j \sim N(0, \ell^{-2} I), \quad \xi_j \sim \text{Unif}(0, 2\pi),$$

where ω_j is column j of Ω and M counts random coordinates only. The bandwidth ℓ is the median Euclidean distance over distinct row pairs among at most 1,000 training rows, sampled without replacement with a separate seed-zero generator if needed. A nonfinite or nonpositive median is replaced by one, and the bandwidth is floored at 10^{-6}. Frequencies and phases use the estimator seed. The raw z block and intercept are retained without moment-column selection.

Bridge RFFs use the same distribution, normalization, and bandwidth rule on standardized W, but omit raw W coordinates and, in the nonredundant fits, an internal intercept; the bridge has its own intercept. The W and additional Reg-WV V maps use seed offsets 101 and 211. Maps and fitted transformations are reused across penalty candidates, validation, test, and intervention rows. Widths 32 and 64 are separately drawn with the same seed, not nested draws. The restricted and complete text classes share the moment map. Analytic studies instead use the explicit moment coordinates below. Total dimensions include all deterministic terms; $m \geq p$ alone does not guarantee full rank, completeness, or numerical stability.

Hyperparameter grids. The main comparisons use mean-loss penalties and nonredundant features, with $\lambda_{\text{pred}} \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and nine ridge coefficients,

$$\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{2}\}.$$

Predictive families use 54 logarithmically spaced mean-loss penalties from 10^{-10} to 10^2 per feature template, matching PACE's 54-candidate budget. Observed-only has a single feature specification. Reg-WV (full) allocates the per-view RFF count to each view; Reg-WV (equal) splits it across the views. Equal-budget controls match the nominal nonconstant RFF allocation; retained ranks can differ after redundant columns are removed. Analytic SCMs use one declared polynomial template per family.

Primary error measures and replication. For effect prediction, the loss is $n^{-1}\sum_i \|\widehat{\tau}_i - \tau_i\|_2$. For a fixed transition on reference observations partitioned into predefined groups g, let $e_g = \overline{\widehat{\tau}}_g - \overline{\tau}_g$. The weighted group RMSE is $(\sum_g n_g \|e_g\|_2^2 / \sum_g n_g)^{1/2}$. The studies below specify their reference populations, groups, sample sizes, and units of replication.

Simulation results informed method development; the reported summaries exclude preliminary runs and retain all weak-moment outliers.

Scope of the identification assumptions. We assess Assumptions 1–5 for each measurement construction and bridge class. The known generators permit analytic checks, whereas the text application provides only partial observable diagnostics. Table 4 separates analytic support, structural violations, finite-class numerical evidence, and unresolved assumptions.

Table 4: **Status of identification assumptions across experimental constructions.** Each row specifies a measurement construction and bridge class. The entries distinguish analytic justification, structural violations, finite-class numerical evidence, and unresolved assumptions.

Construction and bridge class	A1	A2	A3	A4	A5
Linear / cubic structural models Polynomial bridges	Supported	Supported	Supported	Supported	Supported
Designed factor simulations Analytic bridges[a]	Supported	Supported	Supported	Supported	Supported
Designed factor simulations RFF32 / RFF64 classes[b]	Supported	Supported	Supported	Misfit	Numerical only
Target-containing factor views RFF32 / RFF64 classes[b,c]	Supported	Supported	Violated	Misfit	Numerical only
Cubic duplication control $V = W$; polynomial bridges	Supported	Supported	Violated	Supported	Supported
Cubic independent-V control Polynomial bridges[d]	Supported	Supported	Supported	Supported	Violated
CEBaB same-review views TF–IDF, BERT, RoBERTa, Qwen[e]	Unverified	Unverified	Unverified	Unverified	Unverified

Assumptions. A1: intervention consistency and positivity; A2: latent ignorability; A3: proxy exclusions; A4: latent outcome-bridge existence in the specified class; A5: observable completeness within that class.

Evidence categories. Supported: justified by the declared generator and analytic arguments. *Violated*: contradicted by the construction. *Misfit*: numerical evidence of failure of exact bridge representation in the examined finite class. *Numerical only*: numerical support for injectivity on the examined finite-dimensional bridge sections. *Unverified*: not established for the full latent-state model.

[a]The designed factor simulations are dSprites x-position and 3D Shapes orientation, each with confounding parameter $c \in \{0, 2\}$ and proxy-noise standard deviation $\sigma \in \{0.2, 0.8\}$. A4 and A5 refer to the stated analytic classes; unrestricted completeness fails in these noisy finite-factor models.

[b]RFF32 and RFF64 denote the examined random Fourier feature classes. Numerical conclusions cover the seed-100 maps and settings examined in the diagnostic study, not every fitted map or experimental seed. Finite-class misfit does not rule out a bridge in a richer class, and numerical injectivity does not imply a stable inverse.

[c]The target-containing views are those used in the four factor-prediction benchmarks: dSprites x-position and 3D Shapes orientation, object hue, and scale. Their W distributions depend on A after conditioning on the full (S, U), violating A3.

[d]Here V is drawn independently from its marginal law, the population counterpart of the finite-sample permutation control. The latent bridge is unchanged, but A5 fails. By contrast, duplication violates A3 while preserving A5 because conditioning on $V = W$ reveals W.

[e]CEBaB diagnostics find residual W–V dependence conditional on ambiance, service, and restaurant-noise annotations. Under the diagnostic sampling assumptions, this is evidence against the annotation-only proxy-independence model, not a rejection of the exclusions conditional on a richer latent U. Observed empty target cells do not prove population positivity failure, and pooled sample rank does not establish A5.

B.2 Structural demonstrations and simulations

B.2.1 Data-generating models and exact bridges

The cubic generator uses independent $S \sim \text{Bernoulli}(1/2)$, $U \sim N(0,1)$, and mutually independent Gaussian errors:

$$\begin{aligned} A &= 0.5S + U + 0.35(U^2 - 1) + \varepsilon_A, \\ W &= U + \varepsilon_W, \qquad V = U + \varepsilon_V, \\ Y &= (0.5 + 0.5S)A + 0.8U + 0.6U^2 + 0.25U^3 + \varepsilon_Y. \end{aligned} \tag{14}$$

The error standard deviations for (A, W, V, Y) are $(0.5, 1, 0.5, 0.5)$. The linear control removes the quadratic treatment term and uses latent outcome coefficients $(1, 0, 0)$. Both have exact $A \mapsto A + 1$ effect $0.5 + 0.5S$. Factual training uses the realized A; reference evaluation fixes $A = 0$ and $A' = 1$ on a sample of (S, U, W) not selected by treatment. Because these SCM effects depend only on S, the two transitions have the same true contrast here, though the distinction matters for general models.

For the cubic outcome a valid bridge is

$$h_0(A, S, W) = (0.5 + 0.5S)A + 0.8W + 0.6(W^2 - 1) + 0.25(W^3 - 3W).$$

Conditional Gaussian moments give $\mathbb{E}(W \mid U) = U$, $\mathbb{E}(W^2 - 1 \mid U) = U^2$, and $\mathbb{E}(W^3 - 3W \mid U) = U^3$. These identities verify the latent bridge. The implemented bridge and moment coordinates are

$$b = (1, S, A, AS, W, W^2 - 1, W^3 - 3W)^\top, \qquad q = (1, S, A, AS, V, V^2 - \tfrac{1}{4}, V^3 - \tfrac{3}{4}V)^\top.$$

The V coordinates use its measurement-error variance $1/4$. Both maps are unstandardized in the validation-selected comparisons. Cubic PACE has $p = m = 7$; the linear control keeps only the first-order proxy coordinate, giving $p = m = 5$. Reg-W shares PACE's bridge design; full Reg-WV retains ten and six columns, respectively. Moment tests need only be measurable in (A, S, V); unbiased latent-power estimation is not required. Although raw and corrected polynomials can span the same space, changing their coordinates can change the unweighted objective, penalties, and conditioning.

B.2.2 Estimation and evaluation

The validation-selected simulations use 50 independent replicates per model and sample size, with $n = 1{,}000$ or $5{,}000$ training observations, 1,000 validation observations, and 3,000 test observations. They use the common grids in Appendix B.1. Reference outcomes are known at both target levels, and structural groups are the two S strata.

Table 5: **Validation-selected structural estimators.** Pair MAE, mean ± seed standard error over 50 replicates. PACE-Stabilized restricts λ_{pred} to positive values.

Setting	PACE-Moment	PACE-Stabilized	PACE-CF	Reg-W	Reg-WV
Linear, $n = 1{,}000$	0.060 ± 0.005	0.056 ± 0.005	0.054 ± 0.005	0.078 ± 0.001	0.064 ± 0.001
Linear, $n = 5{,}000$	0.029 ± 0.003	0.027 ± 0.003	0.025 ± 0.003	0.077 ± 0.001	0.063 ± 0.001
Cubic, $n = 1{,}000$	0.101 ± 0.009	0.084 ± 0.005	0.084 ± 0.005	0.096 ± 0.002	0.079 ± 0.002
Cubic, $n = 5{,}000$	0.059 ± 0.004	0.055 ± 0.004	0.055 ± 0.004	0.092 ± 0.001	0.082 ± 0.001

B.2.3 Fixed-penalty proxy controls

Table 6 examines the role of the treatment proxy in the cubic model using fixed procedures at $n = 10{,}000$. Duplicating W reduces the moment construction to a single-view fit, whereas permuting V produces weak-moment outliers. These controls use fixed penalties rather than the validation-selected penalties in Table 1.

The controls use training-standardized nonconstant polynomial features. Reg-W and Reg-WV are unpenalized least-squares fits; PACE-Moment and its permutation and duplication controls set

both penalties to zero. PACE-Stabilized uses the penalties in Table 6, multiplying mean squared prediction loss and the squared non-intercept coefficient norm, respectively. Moment cross-products are normalized by sample size as in Eq. (7), and the intercept is unpenalized. The validation-selected comparisons instead use the unstandardized polynomial coordinates specified above.

Table 6: **Fixed-penalty cubic mechanism controls.** At $n = 10{,}000$, 200 seeds. Bias and RMSE average the two stratum summaries; factual mean squared error (MSE) is separate. PACE-Stabilized uses $\lambda_{\text{pred}} = 0.05$ and $\lambda_{\text{ridge}} = 10^{-4}$.

Fixed procedure	Mean bias	Stratum RMSE	Factual MSE
Reg-W cubic	1.164	1.164	1.059
Reg-WV cubic	0.793	0.794	0.746
PACE-Moment	-0.005	0.136	7.614
PACE-Stabilized	1.025	1.027	1.663
PACE-Moment, permuted V	1.554	11.079	737.308
PACE-Moment, $V = W$	1.164	1.164	1.059

B.3 Factor simulations with designed proxies

B.3.1 Data-generating models

For a factor with discrete value f_j and L_j levels, define $z_j = 2f_j/(L_j - 1) - 1$. The dSprites teacher is

$$Y = z_x + 0.6z_{\text{shape}} - 0.4z_{\text{scale}} + 0.7z_x z_{\text{orientation}} + 0.5\sin(\pi z_y) + 0.4z_{\text{shape}}z_{\text{scale}}.$$

The Shapes teacher is

$$\begin{aligned} Y = {} & z_{\text{orientation}} + 0.45z_{\text{shape}} - 0.35z_{\text{scale}} + 0.65z_{\text{orientation}}z_{\text{object hue}} \\ & + 0.45\sin(\pi z_{\text{wall hue}}) + 0.35z_{\text{floor hue}}z_{\text{scale}} + 0.25z_{\text{shape}}z_{\text{object hue}}. \end{aligned}$$

In the dSprites simulation, the target is x-position, at raw levels 8 and 24; context is shape and scale, and the confounding driver is orientation. For Shapes the target is orientation, at raw levels 3 and 11; context is shape and scale, and the driver is object hue. Other non-target factors belong to the declared latent state. After sampling non-target factors uniformly, high-level assignment has probability

$$p(A = \text{high} \mid U) = 0.1 + 0.8\,\text{expit}(c\,z_{\text{driver}}), \qquad c \in \{0, 2\}.$$

Both target levels have assignment probability at least 0.1 and occur in each factual training sample. Let z_{-A} include every nonconstant non-target factor, including the observed context. Measurements are

$$W = z_{-A} + \sigma\epsilon_W, \qquad V = z_{-A} + \tfrac{\sigma}{2}\epsilon_V, \qquad \sigma \in \{0.2, 0.8\},$$

with independent standard Gaussian errors. Changing only A leaves W unchanged; both exclusions hold conditional on the full declared factor state. Reference factors and measurement errors are generated independently of training and assignment, using the same factors across confounding and noise conditions. Evaluation fixes the low-to-high transition. Each setting has 4,000 training observations, 1,000 validation observations, and 3,000 observations in each of the test and reference samples.

B.3.2 Population effects and bridges

Let a be the normalized target, Z the normalized assignment-driving factor (measured by W_Z), and s_1, s_2 normalized shape and scale. All non-target factors are sampled independently on their finite grids. The teacher has the form

$$Y = a + \gamma aZ + \ell(S)Z + B(S, R),$$

where R collects the remaining factors. For dSprites, $\gamma = 0.7$ and $\ell = 0$; for Shapes, $\gamma = 0.65$ and $\ell(S) = 0.25s_1$. The target levels are $(-15/31, 17/31)$ for dSprites and $(-4/7, 4/7)$ for Shapes. Their differences are $32/31$ and $8/7$.

Mean effects and observational bias. Write $\Delta a = a_{\mathrm{hi}} - a_{\mathrm{lo}}$ and $p_c(d) = 0.1 + 0.8\,\mathrm{expit}(cd)$. Under the product-uniform reference law,

$$\tau(s) = \Delta a, \qquad \mu_c := \mathbb{E}(Z \mid A = a_{\mathrm{hi}}) = 2\mathbb{E}\{Z p_c(Z)\}.$$

The observational high-minus-low contrast is

$$\Delta a + \gamma(a_{\mathrm{hi}} + a_{\mathrm{lo}})\mu_c + 2\ell(s)\mu_c.$$

To obtain these expressions, note that the paired teacher difference is $\Delta a(1 + \gamma Z)$ and $\mathbb{E}(Z \mid S) = 0$. Symmetry gives $p_c(d) + p_c(-d) = 1$, hence $P(A = a_{\mathrm{hi}}) = 1/2$ and $\mathbb{E}(Z \mid A = a_{\mathrm{lo}}) = -\mu_c$. The remaining factor term has the same conditional mean at both target levels. Subtracting the two conditional teacher means gives the formula. Finite-grid enumeration gives the same expression.

For $c = 0$ the bias is zero. For $c = 2$, the dSprites driver has 40 levels and $\mu_c = 0.23477$, giving constant bias 0.01060. Shapes object hue has ten levels and $\mu_c = 0.26690$, giving bias $0.13345 s_1$. Its conditional-mean bias RMSE is 0.09947, although its population average bias vanishes. Population integration of the specified observed-only regression span yields the same limiting contrasts.

Known bridge spans. For $W_j = z_j + \sigma\epsilon_j$ with independent standard Gaussian errors,

$$\mathbb{E}\{\sin(\pi W_j) \mid z_j\} = e^{-\pi^2\sigma^2/2}\sin(\pi z_j).$$

Consequently, exact bridges in canonical factor units are

$$\begin{aligned} h_d(a, s, w) &= a + 0.6 s_1 - 0.4 s_2 + 0.7 a w_{\mathrm{orientation}} + 0.5 e^{\pi^2\sigma^2/2}\sin(\pi w_y) + 0.4 s_1 s_2, \\ h_s(a, s, w) &= a + 0.45 s_1 - 0.35 s_2 + 0.65 a w_{\mathrm{object\ hue}} + 0.45 e^{\pi^2\sigma^2/2}\sin(\pi w_{\mathrm{wall\ hue}}) \\ &\quad + 0.35 w_{\mathrm{floor\ hue}} s_2 + 0.25 s_1 w_{\mathrm{object\ hue}}. \end{aligned}$$

The seven- and eight-dimensional fitting spans contain the displayed terms without their known coefficients. Moment functions replace W by V and retain exact S. Since the raw sine coordinate already spans its rescaled version, fitting requires neither the true coefficient nor σ. Reg-W shares the exact bridge matrix; full Reg-WV adds corresponding terms from both views. All coefficients are estimated from factual training data. The choice of these spans uses structural knowledge of the generators.

Exact finite-grid integration gives full population cross-moment rank, seven for dSprites and eight for Shapes. At $c = 2$, the smallest singular values for analytic moments are 0.07773 and 0.10698 at $\sigma = 0.2$, and 0.00935 and 0.00868 at $\sigma = 0.8$, respectively. These values refer to the declared canonical basis; singular values are not invariant to feature rescaling. Gaussian attenuation of the sine coordinate increases the bridge coefficient and its measurement variability at high noise.

Pair and mean targets. The known bridge contrast is $\Delta a(1 + \gamma W_Z)$. Its discrepancy from the individual teacher effect is $\gamma\Delta a\sigma\epsilon_W$, so its expected pair MAE is $|\gamma\Delta a|\sigma\sqrt{2/\pi}$. At $\sigma = 0.8$, this is 0.4612 for dSprites and 0.4742 for Shapes, although the exact-population mean error is zero. This calculation concerns the known bridge; other effect predictors can have smaller pair error. Finite reference groups retain measurement noise, so their realized errors are evaluated separately from the population mean.

Mean-effect recovery by regression in the Shapes design. For the Shapes teacher and measurement law above, let b be the declared eight-dimensional correct span and let θ^{LS} minimize $\mathbb{E}(Y - b^\top\theta)^2$ without penalties. If $\mathbb{E}(bb^\top)$ is nonsingular, then its coefficient on A is one and

$$\mathbb{E}\{b(a_{\mathrm{hi}}, s, W)^\top\theta^{\mathrm{LS}} - b(a_{\mathrm{lo}}, s, W)^\top\theta^{\mathrm{LS}} \mid S = s\} = \Delta a.$$

This contrast identity does not imply the latent outcome-bridge equation.

To see this, note that $A \in \{-a_0, a_0\}$ with $a_0 = 4/7$, and the joint law of (A, Z) is invariant under simultaneous sign reversal. Context variables are independent of (A, Z) and centered. In the canonical span

$$b = (1, A, S_1, S_2, AW_Z, \sin(\pi W_{\mathrm{wall}}), S_2 W_{\mathrm{floor}}, S_1 W_Z)^\top,$$

$\mathbb{E}(Ab_j) = 0$ for every coordinate except A: in particular, $\mathbb{E}(A^2 W_Z) = a_0^2 \mathbb{E}(W_Z) = 0$, and the terms containing S_1 or S_2 vanish by centering and independence. The teacher similarly gives $\mathbb{E}(AY) = \mathbb{E}(A^2) = a_0^2$. The target-coordinate normal equation therefore yields $\theta_A^{\mathrm{LS}} = 1$. Only A and AW_Z change under the intervention, and $\mathbb{E}(W_Z \mid S = s) = 0$ under the reference law, giving the stated mean contrast.

Finite-grid integration, including the within-measurement covariance in $\mathbb{E}(bb^\top)$, verifies the identity at both confounding and noise levels. At $c = 2$, the mean-contrast error is below 10^{-12} while the root mean squared latent bridge residual is 0.1137 at $\sigma = 0.2$ and 0.3522 at $\sigma = 0.8$. The result is specific to the teacher, feature span, and symmetric reference law; finite sampling or positive penalties need not preserve it. It explains how the low Reg-W mean errors in Table 9 can coexist with failure to recover the bridge.

RFF approximation and exact integration. For a cosine coordinate and Gaussian measurement covariance Σ,

$$\mathbb{E}\{\cos(\omega^\top W + b) \mid z\} = e^{-\omega^\top \Sigma \omega / 2} \cos(\omega^\top z + b).$$

We transform Σ through each training-fitted standardizer before applying this identity. Finite-grid integration then gives the population cross-moment matrix, the exact context-specific mean contrast, and a least-squares projection of the latent outcome mean onto the expected bridge span. The projection is used only to assess approximation error, not to fit or select candidates. Across the confounded configurations, RFF best-span residuals are nonzero, unlike the analytic spans.

B.3.3 Estimation and evaluation

The reference-comparison study uses 20 replicates. A separate set of 20 replicates is shared across the three bridge–moment specifications: a correctly specified bridge with analytic moments, the same bridge with RFF moments, and an RFF bridge with RFF moments. In both studies, latent factors are not supplied as additional fitting covariates. The predefined context grid gives the reference groups, and outcomes are known from the teacher at both target levels.

The analytic dSprites columns, in order, are $(1, A, S_1, S_2, AW_{\mathrm{orientation}}, \sin(\pi W_y), S_1 S_2)$; Shapes uses the eight columns displayed in Appendix B.3.2. Analytic moments replace W by V, without additional coordinate standardization, giving $p = m = 7$ or 8. Proxy coordinates follow the original factor order after removing the target and constant factors. Flexible moments use 160 random coordinates with total $m = 168$ for dSprites and 169 for Shapes. Scalar RFF bridges retain $p = 68$ or 132 for 32 or 64 proxy coordinates; full Reg-WV retains 132 or 260 and equal-budget Reg-WV 68 or 132. With analytic bridges, Reg-WV instead retains nine dSprites or twelve Shapes columns. Reg-W always shares PACE's retained bridge design.

For mean-targeted selection in the factor simulations, candidates are compared by squared reference-group mean error on an independent 1,000-row reference-validation sample. Its outcomes are known at both target levels. This sample is independent of the 3,000-row final reference sample and is additional to the pair-validation data; the diagnostic changes both the validation loss and its population.

B.3.4 Additional comparisons

Table 7 compares finite-reference and exact-population errors for the same pair-selected RFF estimators. Fitted coefficients and hyperparameters are fixed across these evaluations. Sample errors use the realized reference-group means; population errors integrate over the factor and measurement distributions.

Table 7: **Finite-reference and exact-population mean errors.** Mean ± seed standard error over 20 paired replicates. Both evaluations use the same RFF estimators selected by pair loss. NoPACE denotes Joint-CF-noPACE, using the same predictive representatives as Joint-CF.

Target	Observed	PACE-Moment	Reg-W	Reg-WV	PACE-CF	NoPACE
dSprites, $c = 0$**,** $\sigma = 0.2$						
Sample	0.041 ± 0.002	0.027 ± 0.001	0.019 ± 0.001	0.012 ± 0.001	0.019 ± 0.001	0.012 ± 0.001
Population	0.024 ± 0.003	0.025 ± 0.001	0.017 ± 0.001	0.011 ± 0.001	0.016 ± 0.001	0.011 ± 0.001
dSprites, $c = 0$**,** $\sigma = 0.8$						
Sample	0.041 ± 0.002	0.052 ± 0.005	0.033 ± 0.001	0.025 ± 0.001	0.034 ± 0.001	0.025 ± 0.001
Population	0.024 ± 0.003	0.040 ± 0.005	0.018 ± 0.001	0.017 ± 0.001	0.019 ± 0.001	0.017 ± 0.001
dSprites, $c = 2$**,** $\sigma = 0.2$						
Sample	0.041 ± 0.002	0.029 ± 0.002	0.020 ± 0.001	0.014 ± 0.001	0.019 ± 0.001	0.014 ± 0.001
Population	0.024 ± 0.002	0.026 ± 0.002	0.017 ± 0.001	0.013 ± 0.001	0.017 ± 0.001	0.013 ± 0.001
dSprites, $c = 2$**,** $\sigma = 0.8$						
Sample	0.041 ± 0.002	0.051 ± 0.005	0.034 ± 0.001	0.025 ± 0.001	0.037 ± 0.002	0.025 ± 0.001
Population	0.024 ± 0.002	0.037 ± 0.005	0.019 ± 0.001	0.018 ± 0.001	0.023 ± 0.002	0.018 ± 0.001
Shapes, $c = 0$**,** $\sigma = 0.2$						
Sample	0.058 ± 0.004	0.033 ± 0.001	0.022 ± 0.001	0.016 ± 0.001	0.022 ± 0.001	0.016 ± 0.001
Population	0.027 ± 0.004	0.028 ± 0.002	0.016 ± 0.001	0.014 ± 0.001	0.017 ± 0.001	0.014 ± 0.001
Shapes, $c = 0$**,** $\sigma = 0.8$						
Sample	0.058 ± 0.004	0.053 ± 0.002	0.040 ± 0.001	0.029 ± 0.001	0.042 ± 0.001	0.029 ± 0.001
Population	0.027 ± 0.004	0.028 ± 0.002	0.016 ± 0.001	0.017 ± 0.001	0.018 ± 0.002	0.017 ± 0.001
Shapes, $c = 2$**,** $\sigma = 0.2$						
Sample	0.110 ± 0.003	0.041 ± 0.003	0.026 ± 0.001	0.019 ± 0.001	0.026 ± 0.001	0.019 ± 0.001
Population	0.097 ± 0.003	0.037 ± 0.003	0.022 ± 0.001	0.017 ± 0.001	0.022 ± 0.001	0.017 ± 0.001
Shapes, $c = 2$**,** $\sigma = 0.8$						
Sample	0.110 ± 0.003	0.075 ± 0.005	0.051 ± 0.002	0.042 ± 0.002	0.055 ± 0.002	0.042 ± 0.002
Population	0.097 ± 0.003	0.059 ± 0.006	0.037 ± 0.002	0.034 ± 0.002	0.041 ± 0.002	0.034 ± 0.002

Table 8: **Pair prediction in the factor simulations.** MAE, mean ± seed standard error over 20 replicates. $c = 0$ removes designed confounding and $c = 2$ introduces it. The teacher and reference factor population are fixed across c and noise scale σ.

Setting	Observed	PACE-CF	Reg-W	Reg-WV	Joint-CF
dSprites, $c = 0,\ \sigma = 0.2$	0.371 ± 0.001	0.110 ± 0.001	0.110 ± 0.001	0.056 ± 0.001	0.056 ± 0.001
dSprites, $c = 0,\ \sigma = 0.8$	0.371 ± 0.001	0.285 ± 0.001	0.285 ± 0.001	0.178 ± 0.001	0.178 ± 0.001
dSprites, $c = 2,\ \sigma = 0.2$	0.371 ± 0.001	0.111 ± 0.001	0.111 ± 0.001	0.059 ± 0.001	0.059 ± 0.001
dSprites, $c = 2,\ \sigma = 0.8$	0.371 ± 0.001	0.290 ± 0.001	0.288 ± 0.001	0.180 ± 0.001	0.180 ± 0.001
Shapes, $c = 0,\ \sigma = 0.2$	0.413 ± 0.001	0.116 ± 0.001	0.116 ± 0.001	0.063 ± 0.001	0.063 ± 0.001
Shapes, $c = 0,\ \sigma = 0.8$	0.413 ± 0.001	0.306 ± 0.001	0.306 ± 0.001	0.187 ± 0.001	0.187 ± 0.001
Shapes, $c = 2,\ \sigma = 0.2$	0.417 ± 0.001	0.119 ± 0.001	0.119 ± 0.001	0.067 ± 0.001	0.067 ± 0.001
Shapes, $c = 2,\ \sigma = 0.8$	0.417 ± 0.001	0.312 ± 0.001	0.311 ± 0.001	0.191 ± 0.001	0.191 ± 0.001

Table 9 compares bridge and moment specifications on matched data. Mean-targeted selection uses the independent reference-validation sample defined in Appendix B.3.3. Appendix B.3.2 explains how Reg-W can recover the Shapes reference mean without recovering the bridge.

Table 9: **Exact-population mean-effect error under confounding.** At $c = 2$, $n = 4{,}000$; mean $\pm$ seed standard error over 20 replicates. Moment-pair and Moment-mean denote PACE-Moment tuned by pair loss and independent reference-group loss, respectively; other estimator columns use pair selection. Bridge / moments specifies the function classes. The direct-label baseline uses the same additional reference labels as Moment-mean; its results are identical across noise levels and feature classes.

Bridge / moments	Moment-pair	Moment-mean	PACE-CF	Reg-W	Reg-WV
dSprites, $\sigma = 0.2$					
Correct / analytic	0.015 ± 0.002	0.008 ± 0.001	0.007 ± 0.001	0.006 ± 0.001	0.003 ± 0.000
Correct / RFF	0.009 ± 0.001	0.009 ± 0.001	0.008 ± 0.001	0.006 ± 0.001	0.003 ± 0.000
RFF / RFF	0.028 ± 0.002	0.025 ± 0.001	0.018 ± 0.001	0.017 ± 0.001	0.013 ± 0.001
dSprites, $\sigma = 0.8$					
Correct / analytic	0.106 ± 0.008	0.023 ± 0.003	0.012 ± 0.002	0.008 ± 0.002	0.008 ± 0.001
Correct / RFF	0.094 ± 0.023	0.023 ± 0.002	0.013 ± 0.002	0.008 ± 0.002	0.008 ± 0.001
RFF / RFF	0.039 ± 0.003	0.026 ± 0.002	0.024 ± 0.002	0.022 ± 0.002	0.022 ± 0.001
Shapes, $\sigma = 0.2$					
Correct / analytic	0.014 ± 0.002	0.008 ± 0.001	0.005 ± 0.001	0.005 ± 0.001	0.003 ± 0.001
Correct / RFF	0.007 ± 0.001	0.007 ± 0.001	0.006 ± 0.001	0.005 ± 0.001	0.003 ± 0.001
RFF / RFF	0.035 ± 0.002	0.027 ± 0.002	0.021 ± 0.002	0.021 ± 0.001	0.016 ± 0.001
Shapes, $\sigma = 0.8$					
Correct / analytic	0.095 ± 0.004	0.031 ± 0.006	0.017 ± 0.005	0.008 ± 0.001	0.006 ± 0.001
Correct / RFF	0.104 ± 0.022	0.031 ± 0.002	0.011 ± 0.002	0.008 ± 0.001	0.006 ± 0.001
RFF / RFF	0.052 ± 0.004	0.023 ± 0.002	0.035 ± 0.002	0.034 ± 0.002	0.035 ± 0.002
Direct reference-label mean baseline					
dSprites					0.058 ± 0.003
Shapes					0.079 ± 0.002

Direct reference-label averaging. To assess the contribution beyond additional intervention supervision, we estimate each context's effect by directly averaging $Y_j(a') - Y_j(a)$ over the same 1,000 reference-validation rows used for mean-targeted selection. The 18 dSprites and 32 Shapes contexts are estimated separately, without pooling; no test or final reference labels enter estimation. Exact-population RMSE weights the contexts uniformly, as in the other population comparisons. All validation groups are populated (16–78 rows); an empty group would remain undefined, making full-grid RMSE undefined rather than inserting a zero or merging groups. The resulting errors, 0.058 ± 0.003 and 0.079 ± 0.002 (Table 9), exceed the mean-selected bridge errors in these settings.

High-noise sample-size analysis. High-noise instability motivated an additional analysis using nested training samples of sizes 1,000, 4,000, and 16,000, with shared validation, test, and reference samples. The analysis uses $c = 2$, $\sigma = 0.8$, and 20 replicates. One unpenalized Shapes fit at $n = 4{,}000$ has RMSE 84.45, compared with a median of 0.238 across replicates. The nonmonotone mean at smaller sizes and its reduction at $n = 16{,}000$ characterize finite-sample instability under the same population model (Table 10).

Table 10: **High-noise nested sample-size comparison.** Exact-population mean-effect RMSE, mean ± seed standard error. Unpenalized denotes PACE-Moment with both penalties fixed at zero; Moment-pair and Moment-mean select ridge using their respective validation objectives. The lower panel gives unpenalized medians [first, third quartiles], using linearly interpolated empirical quantiles. All 20 replicates, including extreme values, enter both panels.

Bridge / moments	Procedure	$n = 1{,}000$	$n = 4{,}000$	$n = 16{,}000$
dSprites				
Correct / analytic	Unpenalized	0.294 ± 0.048	0.441 ± 0.275	0.094 ± 0.013
Correct / analytic	Moment-pair	0.098 ± 0.011	0.106 ± 0.008	0.116 ± 0.005
Correct / analytic	Moment-mean	0.047 ± 0.008	0.023 ± 0.003	0.019 ± 0.002
Correct / RFF	Unpenalized	0.222 ± 0.052	0.133 ± 0.027	0.096 ± 0.019
Correct / RFF	Moment-pair	0.118 ± 0.023	0.094 ± 0.023	0.051 ± 0.016
Correct / RFF	Moment-mean	0.035 ± 0.005	0.023 ± 0.002	0.017 ± 0.002
RFF / RFF	Moment-pair	0.055 ± 0.005	0.039 ± 0.003	0.034 ± 0.002
RFF / RFF	Moment-mean	0.035 ± 0.002	0.026 ± 0.002	0.024 ± 0.002
Shapes				
Correct / analytic	Unpenalized	0.753 ± 0.238	4.476 ± 4.210	0.122 ± 0.029
Correct / analytic	Moment-pair	0.102 ± 0.009	0.095 ± 0.004	0.097 ± 0.003
Correct / analytic	Moment-mean	0.055 ± 0.009	0.031 ± 0.006	0.024 ± 0.005
Correct / RFF	Unpenalized	0.168 ± 0.025	0.107 ± 0.019	0.080 ± 0.015
Correct / RFF	Moment-pair	0.132 ± 0.021	0.104 ± 0.022	0.076 ± 0.019
Correct / RFF	Moment-mean	0.049 ± 0.007	0.031 ± 0.002	0.026 ± 0.002
RFF / RFF	Moment-pair	0.060 ± 0.005	0.052 ± 0.004	0.054 ± 0.004
RFF / RFF	Moment-mean	0.030 ± 0.003	0.023 ± 0.002	0.022 ± 0.002
Unpenalized median [Q1, Q3]				
dSprites, correct / analytic		$0.252\ [0.132, 0.435]$	$0.099\ [0.049, 0.230]$	$0.083\ [0.041, 0.133]$
dSprites, correct / RFF		$0.141\ [0.084, 0.311]$	$0.113\ [0.042, 0.163]$	$0.075\ [0.025, 0.150]$
Shapes, correct / analytic		$0.265\ [0.169, 0.973]$	$0.238\ [0.111, 0.311]$	$0.073\ [0.034, 0.159]$
Shapes, correct / RFF		$0.149\ [0.080, 0.224]$	$0.075\ [0.041, 0.180]$	$0.061\ [0.036, 0.109]$

B.4 Text intervention prediction

B.4.1 Data, views, and partitions

We explain food edits while withholding restaurant-noise annotations. Food, ambiance, and service are each encoded in the order Negative, Unknown, Positive; food occupies the first three of nine columns. Null, unrecognized, and no-majority aspect annotations map to Unknown. These same-review views are candidate representations for intervention prediction; their construction does not establish Eq. (3).

The supervised target is the numeric overall-review rating, `review_majority`, on its original 1–5 scale, not an aspect label. Each score model has a scalar ridge head ($L = 1$), with no output rescaling or clipping. Explainers fit these predicted scores; intervention targets are edited-minus-original score differences, so text MAE is measured in rating points. The teacher training set contains 9,848 valid-rated reviews (1,495 originals and 8,353 edits); rows with missing or nonnumeric ratings are excluded. The ridge penalty coefficients are 3 for TF–IDF and 10 for each pretrained-encoder model, using summed squared-error loss with an unpenalized intercept. These teacher penalties are distinct from the explainer penalties in Appendix B.1. Qwen's last-token embedding is L2-normalized before the ridge head; this does not rescale the rating target or predicted score.

Explainers use all 1,755 observational-training originals: the 1,495 rated originals and 260 additional originals whose overall rating has no majority. The latter remain usable because explainer responses are teacher predictions, not human ratings. Word/character transformations and their reducers fit the teacher's 9,848-row population; neural view reducers fit all 1,755 factual originals. Teacher-training edits are separate from the intervention pairs counted in K. Original-review identifiers are disjoint across CEBaB's `train_inclusive`, validation, and test partitions, and all edits remain with

their original. Proxy construction uses neither an edit-derived masking lexicon nor paired outcome differences.

The lexical teacher uses word TF–IDF; W and V use word and character views. Word features are unigrams/bigrams and character features are within-word 3–5 grams, each with minimum document frequency two and at most 5,000 features. Both lowercase, strip Unicode accents, use smoothed inverse document frequency and L2 row normalization, and do not use sublinear term frequency. Separate uncentered truncated singular-value decompositions retain 64 coordinates per view, followed by coordinate standardization; W and V use seed offsets 11 and 23. The dimension cap is $\min(64, \min(n, d) - 1)$, yielding 64 coordinates for each lexical view.

BERT and RoBERTa use 768-dimensional classification-token and attention-masked mean pooling for W and V, without L2 normalization. Qwen uses 1,024-dimensional last-unmasked-token and mean pooling, both L2-normalized. Mean pooling includes nonpadding special tokens. Maximum lengths are 192 for BERT/RoBERTa and 64 for Qwen, with left padding for Qwen. The fixed scalar teacher fits the raw classification/last-token embeddings, before view reduction. Each neural view then uses training-coordinate standardization, centered randomized principal component analysis with 128 components and no whitening, and another coordinate standardization; reducer seed offsets are 101 for W and 211 for V. All transformations are reused for validation, test, and edited inputs. Explanations use the factual pair's proxy values, while edited responses come from applying the fixed teacher to the edited input. No evaluation rows fit a transformation.

CEBaB uses five estimator seeds on the same reviews, shared between the two categorical classes. The public text split was examined during method development; these evaluations should not be interpreted as independent confirmatory studies.

B.4.2 Categorical features and transition support

The complete text bridge uses Eq. (9). Its nested restricted class includes only food-Negative interactions with the proxy RFF basis; the complete class additionally includes food-Positive interactions. Unknown is the interaction reference. Column removal retains Negative and Unknown main-effect coordinates, making Positive the main-effect reference. This parameterization preserves nesting and defines the coefficient ridge penalty, even though a different reference coding can represent the same unpenalized function class. Neither class adds a target–context block.

The paired categorical comparison uses the same 512 random moment coordinates in both classes, the next power of two at least twice the maximum nominal bridge dimension of 202. Including nine encoded concepts, reduced V, and the intercept gives total moment dimension $m = 586$ for TF–IDF and 650 for neural views. With 32 or 64 bridge RFFs, restricted designs have 74 or 138 nominal columns and retain 71 or 135; complete target–proxy designs have 106 or 202 and retain 103 or 199. Recorded PACE design and cross-moment ranks equal its retained bridge dimensions. Full Reg-WV retains 199 or 391 columns for the complete class, versus 103 or 199 for equal-budget Reg-WV. The separate numerical ablation in Appendix B.6 uses restricted interactions and 160 random coordinates, not this moment map.

CEBaB's 1,755 factual training rows contain 749 Negative, 370 Unknown, and 636 Positive food labels. We retain food-edit pairs whose cleaned food label changes and whose original is available in the same split. The resulting 431 validation pairs (235 originals) and 449 test pairs (247 originals) preserve ambiance and service labels. All six directed transitions occur (Table 11); opposite directions are separate edits of different originals.

Table 11: **CEBaB transition support.** Counts use the same official split for all four representations. All edits of an original review remain in the same partition.

Food transition	Validation pairs	Test pairs
Negative → Unknown	72	87
Negative → Positive	87	94
Unknown → Negative	19	24
Unknown → Positive	22	27
Positive → Negative	124	116
Positive → Unknown	107	101

B.4.3 Evaluation and additional comparisons

Text uncertainty resamples the 247 original test-example clusters underlying 449 food pairs, using 1,000 shared bootstrap draws. All edits from an original input receive the same weight. Intervals condition on fitted candidates and average seeds and, where applicable, budget draws; they do not include retraining or constitute post-selection causal confidence intervals.

Table 12 compares the nested categorical classes defined above. The classes share the explained model, pooling rules, proxy maps, split, loss scale, penalty grid, and 512 random moment features. Overall MAE weights all 449 pairs equally.

Table 12: **Categorical target–proxy interaction sensitivity.** Restricted and Complete give mean PACE-CF pair MAE. Paired differences have 95% original-review cluster intervals; negative favors the first named method. Both classes use 512 random moment coordinates. Displayed zeros may reflect rounding.

Representation	Restricted	Complete	Complete − restricted	Complete PACE-CF − Reg-W
TF–IDF	0.502	0.485	-0.017 [-0.035, 0.001]	-0.009 [-0.023, 0.005]
BERT	0.687	0.670	-0.017 [-0.030, -0.003]	0.001 [-0.007, 0.010]
RoBERTa	0.513	0.491	-0.023 [-0.035, -0.009]	0.000 [-0.005, 0.005]
Qwen	0.664	0.652	-0.012 [-0.024, 0.000]	0.000 [-0.003, 0.003]

Table 13: **Text predictive controls and feature budgets.** Pair MAE, mean ± seed standard error. Reg-WV (equal) matches the total nonconstant RFF budget of the one-view bridge; Reg-WV (full) has that budget in each view. All rows use complete categorical interactions. Seed errors describe estimator variability on the same reviews.

Representation	Observed	Reg-W	Reg-WV (equal)	Reg-WV (full)	PACE-CF
TF–IDF	0.521 ± 0.000	0.494 ± 0.008	0.482 ± 0.005	0.494 ± 0.004	0.485 ± 0.002
BERT	0.667 ± 0.000	0.669 ± 0.001	0.667 ± 0.000	0.665 ± 0.002	0.670 ± 0.000
RoBERTa	0.499 ± 0.000	0.490 ± 0.000	0.491 ± 0.001	0.496 ± 0.004	0.491 ± 0.001
Qwen	0.626 ± 0.000	0.652 ± 0.001	0.654 ± 0.001	0.654 ± 0.001	0.652 ± 0.001

Grouped edit-prediction errors. Using the same fitted complete-class estimators as Table 3, we average signed prediction errors within groups $g = (S, A, A')$ defined jointly by ambiance, service, and the directed food transition. If e_g is the group's mean signed error and n_g its pair count, grouped edit-prediction RMSE is $\{\sum_{g:n_g>0}(n_g/449)e_g^2\}^{1/2}$. We compute this separately for each seed and then average the five RMSEs (Table 14); we do not average predictions before computing error. Of 54 declared groups, 46 are supported; eight remain undefined, not zero. Among supported groups, 28 have fewer than five originals, including nine singletons.

Table 14: **Grouped edit-prediction RMSE.** Rating-point error, mean of five per-seed metrics, using the edit-pair distribution and full-validation estimators from Table 3. Moment denotes PACE-Moment; Reg-WV uses the full per-view budget. This is a predictive diagnostic, not a test of proximal identification.

Representation	Observed	Reg-W	Reg-WV	Moment	PACE-CF	Joint-CF
TF–IDF	0.364	0.267	0.223	0.273	0.273	0.223
BERT	0.359	0.487	0.440	0.453	0.453	0.437
RoBERTa	0.343	0.341	0.297	0.331	0.331	0.298
Qwen	0.358	0.443	0.460	0.465	0.453	0.358

PACE-CF has lower grouped error than Reg-W for BERT and RoBERTa, but higher error for TF–IDF and Qwen. The paired PACE-CF-minus-Reg-W differences are respectively -0.0344 $[-0.0447, -0.0189]$, -0.0097 $[-0.0155, -0.0032]$, 0.0059 $[-0.0212, 0.0281]$, and 0.0101 $[0.0055, 0.0133]$. These conditional 95% reweighting intervals use 1,000 shared original-review multiplier draws: independent positive exponential weights with mean one, shared by all pairs of an original, with group means and pair shares recomputed in each draw. Unlike the multinomial cluster bootstrap above, positive weighting preserves observed group support. It does not resolve sparse-group uncertainty; the intervals condition on fitted models, selection, seeds, and observed support, and omit fitting, selection, and unsupported-group uncertainty. Within these edit groups, the distribution of W need not be $P(W \mid S = s)$ from Theorem 1; aggregation neither establishes valid proxies nor verifies causal debiasing.

B.5 Intervention-budget selection

B.5.1 Budget sampling, selection, and calibration

The budget K counts held-out intervention pairs used for candidate tuning, estimator selection, and residual calibration. The explained models and proxy transformations are held fixed as K varies; their training data, including edited training reviews, are specified in Appendix B.4.

Budgets are $K \in \{5, 10, 20, 50, 100\}$ and full validation. Each small-budget setting uses 20 common draws per seed without replacement. Every family tunes on those labels only; the selected estimators are then evaluated on test data. The reference for all-family oracle regret is

$$L_{\text{test}}(\widehat{f}_K) - \min_{j \in \{\text{PACE,W,WV,observed}\}} L_{\text{test}}(\widehat{f}_{j,K}),$$

where each family representative $\widehat{f}_{j,K}$ is selected using the same K validation pairs. Test losses define only the evaluation reference; they are not used to select candidates. A selected member of these four families has nonnegative regret. Effect calibration is an additional procedure, so its difference from this reference can be negative.

Mean calibration adds $K^{-1} \sum_{i \in \mathcal{V}_K} (\tau_i - \widehat{\tau}_i^{\text{obs}})$ to the K-selected observed-only predictor; median calibration replaces this average with the pooled scalar median. Neither has further tuned parameters. Repeated budget draws are averaged within seed before seed variability is computed. Review-cluster uncertainty follows Appendix B.4 and conditions on the fitted candidates.

B.5.2 Selection increments and calibration results

Table 15 compares Joint-CF and Joint-CF-noPACE using identical K-tuned predictive representatives, validation pairs, and tie-breaking. Negative differences favor including PACE. Figure 4 reports regret relative to the best test-performing representative of the same four families. This reference is conditional on within-K tuning, so it isolates cross-family selection error rather than all tuning error. Figure 3 reports the corresponding absolute losses.

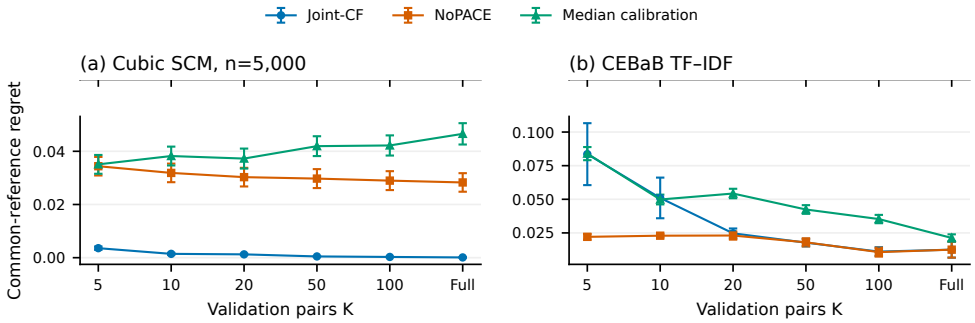


Figure 4: **Regret relative to the common four-family reference.** The reference is the best test-performing family representative after within-K tuning (Appendix B.5). NoPACE denotes Joint-CF-noPACE; median calibration is outside this family set. Shared budget draws are averaged within seed; bars show seed standard errors. The text panel uses complete categorical interactions.

Table 15: **Joint-CF minus predictive-only selection.** Paired MAE differences, mean ± seed standard error after averaging budget draws within seed. Full uses all available validation pairs. Text seed errors describe estimator variability; review-sampling intervals are in Table 16. Zeros are rounded, except that full-validation text differences are exactly zero.

Setting	Δ_{10}	Δ_{100}	Δ_{full}
CEBaB: BERT	0.012 ± 0.005	0.000 ± 0.000	0.000 ± 0.000
CEBaB: Qwen	0.021 ± 0.009	0.000 ± 0.000	0.000 ± 0.000
CEBaB: RoBERTa	0.013 ± 0.004	0.000 ± 0.000	0.000 ± 0.000
CEBaB: TF–IDF	0.028 ± 0.015	0.000 ± 0.000	0.000 ± 0.000
Cubic, $n = 1{,}000$	-0.010 ± 0.002	-0.010 ± 0.002	-0.010 ± 0.002
Cubic, $n = 5{,}000$	-0.030 ± 0.004	-0.029 ± 0.004	-0.028 ± 0.004
dSprites, c =0, σ =0.2	0.001 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
dSprites, c =0, σ =0.8	0.002 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
dSprites, c =2, σ =0.2	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
dSprites, c =2, σ =0.8	0.005 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
Shapes, c =0, σ =0.2	0.001 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Shapes, c =0, σ =0.8	0.004 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
Shapes, c =2, σ =0.2	0.001 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Shapes, c =2, σ =0.8	0.005 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
dSprites: position	0.002 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
Linear, $n = 1{,}000$	-0.023 ± 0.003	-0.021 ± 0.003	-0.021 ± 0.003
Linear, $n = 5{,}000$	-0.043 ± 0.003	-0.040 ± 0.003	-0.040 ± 0.003
Shapes: object hue	0.005 ± 0.001	0.000 ± 0.000	-0.001 ± 0.001
Shapes: orientation	0.005 ± 0.001	0.000 ± 0.001	0.000 ± 0.000
Shapes: scale	0.003 ± 0.001	0.000 ± 0.000	0.000 ± 0.000

Table 16 compares predictive-only selection with the mean and median residual-calibration rules in Appendix B.5. Each rule uses the same K pairs as selection. At $K = 10$, including PACE increases prediction error in all four text representations, with positive original-review cluster intervals. For TF–IDF, median calibration has MAE 0.555 and 0.520 at $K = 10$ and 100, respectively: lower than mean calibration at $K = 10$, but higher than predictive-only selection at both budgets.

Table 16: **Text selection and residual calibration.** Mean loss or paired difference with 95% original-review cluster intervals. NoPACE denotes Joint-CF-noPACE. All tuning and calibration share the stated pair budget. Intervals condition on fitted candidates; very small differences and limits can round to zero.

Representation, K	NoPACE	Joint-CF − NoPACE	Mean calibr.	Median calibr.
TF–IDF, 10	0.528 [0.491, 0.567]	0.028 [0.025, 0.031]	0.567 [0.530, 0.608]	0.555 [0.517, 0.597]
TF–IDF, 100	0.495 [0.457, 0.538]	0.000 [0.000, 0.001]	0.520 [0.480, 0.563]	0.520 [0.479, 0.564]
BERT, 10	0.698 [0.657, 0.741]	0.012 [0.009, 0.015]	0.710 [0.666, 0.757]	0.710 [0.665, 0.758]
BERT, 100	0.672 [0.629, 0.719]	0.000 [0.000, 0.000]	0.670 [0.625, 0.718]	0.671 [0.626, 0.718]
RoBERTa, 10	0.518 [0.485, 0.553]	0.013 [0.010, 0.015]	0.537 [0.505, 0.572]	0.532 [0.499, 0.567]
RoBERTa, 100	0.499 [0.464, 0.536]	0.000 [0.000, 0.000]	0.495 [0.462, 0.532]	0.497 [0.464, 0.533]
Qwen, 10	0.664 [0.626, 0.705]	0.021 [0.018, 0.025]	0.672 [0.630, 0.716]	0.675 [0.632, 0.719]
Qwen, 100	0.635 [0.593, 0.679]	0.000 [-0.001, 0.001]	0.631 [0.585, 0.678]	0.635 [0.589, 0.682]

B.6 Sensitivity to numerical conventions

This analysis tests whether loss scaling, redundant columns, and search range change the comparison with Reg-W while the measurements and explained models are fixed.

Factor-prediction benchmarks. The numerical ablation uses 6,000 sampled dSprites factor rows and 8,000 Shapes rows, with 60/20/20 train/validation/test fractions. These benchmarks evaluate deterministic analytic teachers and nonlinear random maps of the full standardized factor vector. The target itself enters W; uniform factor sampling does not deliberately create omitted-common-cause confounding. Their grouped results summarize the edit-pair population, not the fixed-transition reference estimand.

The numerical ablation uses the same GAM templates, with 32 or 64 RFFs per W view and 160 random moment features, in all four conditions. It crosses summed-loss versus mean-loss predictive penalties with redundant versus nonredundant bridge features. The reduced grid uses $\lambda_{\text{pred}} \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and ridge values $\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 10^{-1}\}$. The reference condition retains redundant columns and uses summed-loss predictive penalties. The other three conditions change the penalty scale, remove redundant columns, or apply both changes.

The Expanded condition uses mean-loss penalties, nonredundant features, and the larger grid in Appendix B.1. The text rows use restricted interactions and 160 random moment coordinates, not the main comparison's complete target–proxy class and 512 coordinates. Total moment dimensions are 196 (dSprites), 212 (Shapes), 234 (TF–IDF), and 298 (neural text), reflecting their different view dimensions. With 32 or 64 bridge RFFs, redundant factor designs have 70 or 134 columns but ranks 68 or 132; redundant text designs have 76 or 140 columns but ranks 71 or 135. Column selection retains those original nonredundant coordinates. Full rank does not imply a well-conditioned moment matrix.

Table 17 separates these changes using paired comparisons across settings and seeds. The factor-prediction benchmarks use ten replicates per setting; the text comparisons use five estimator seeds. Together, the changes to penalty scaling, feature redundancy, and the search grid reverse some method rankings in the factor benchmarks. Removing redundant columns alone leaves all four between-method differences unchanged at the reported precision; penalty scaling and the expanded grid change these differences. Changes in a between-method difference do not by themselves identify which method improved.

Table 17: **Loss-normalization and feature-parameterization ablation.** Mean pair error of PACE-CF minus Reg-W; negative favors PACE. Reference uses summed-loss predictive penalties and redundant columns. Mean loss changes the penalty scale; Columns removes redundant columns; Both applies both changes on the reduced grid. Expanded applies both changes with the larger grid. Text rows use the restricted categorical class and 160 random moment coordinates.

Setting	Reference	Mean loss	Columns	Both	Expanded
dSprites: position	-0.015	-0.006	-0.015	-0.005	+0.005
Shapes: orientation	-0.003	-0.001	-0.003	-0.001	+0.001
Shapes: object hue	-0.010	+0.004	-0.010	+0.005	+0.011
Shapes: scale	-0.042	-0.006	-0.042	-0.006	+0.001
CEBaB: TF–IDF	+0.010	+0.011	+0.010	+0.019	+0.011
CEBaB: BERT	+0.021	+0.019	+0.021	+0.019	+0.016
CEBaB: RoBERTa	+0.018	+0.021	+0.018	+0.029	+0.017
CEBaB: Qwen	+0.021	+0.022	+0.029	+0.024	+0.013