SYNTHETIC CONTROL WITH FOUNDATION MODELS: NONLINEAR GAINS AND CONTEXT-AXIS SENSITIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic control (SC) predicts a target's untreated path using comparable unaffected units as donors. We show that classical SC with convex weights admits a restricted single-query attention form, motivating frozen foundation models (FMs) as in-context predictors. We formalize the finite-context tradeoff: nonlinear approximation gains must exceed the recovery error. We also identify context-axis sensitivity: transposing the same panel can change a checkpoint's predictions. Our two-way centering procedure reduces the mean absolute log-MSE gap between orientations by 61–85%. Our simulations show tabular FMs recover nonlinear structure better than tuned CatBoost and kernel ridge when sufficient context exists. In a real-panel study spanning 673 prediction tasks from 12 datasets, we find that centered FMs are competitive when pre-treatment history is long relative to donor-pool size, while synthetic difference-in-differences (SDID) performs best when it is short. In a semi-synthetic experiment using real U.S. traffic-fatality data, we show that tabular FMs with a suitable context-axis interface can reduce the time to reach 80% test power for a 0.7-standard-deviation effect from 12 to 9 months versus the best baseline SDID, a 25% reduction.

1 INTRODUCTION

Many policy interventions cannot be evaluated by randomized trials because only one state, country, or cohort is treated. Synthetic control (SC) reframes this as a prediction problem. Similar untreated units, called donors, often share trends and shocks with the treated target, making their outcomes useful predictors. SC learns the donor-target relationship before treatment and uses the donors' post-treatment outcomes and pre-treatment histories to predict the target's path that would have occurred without the treatment. This unobservable path is the counterfactual. The gap between the observed and predicted paths estimates how much the intervention changed the outcome (Abadie & Gardeazabal, 2003; Abadie et al., 2010; Abadie, 2021b). The challenge is to learn a donor-target mapping from limited pre-treatment history and apply it reliably beyond the treatment boundary, where the target's *untreated* label is not observable.

For example, consider the 1987 increase in rural-interstate speed limits from 55 to 65 mph. Ashenfelter & Greenstone (2004) study its effect with Fatality Analysis Reporting System (FARS) data from 1982–1993, using six annual post-treatment observations. The same outcomes are also available monthly, giving twelve times as many (though noisier) observations over the same calendar span (Figure 1). More accurate prediction can shorten the post-treatment window needed to detect an intervention effect. We show that frozen foundation models (FMs) can do so in the FARS case study and characterize when such gains appear.

In essence, synthetic control estimators are constrained regressions whose features are the donors, fitted on a pre-treatment window that is often short relative to that feature set, which is a challenge for most tabular learners. We show that canonical SC admits a single-query attention-form representation (see Section 3), making in-context learners natural candidates for the prediction task. Prior work either pretrains SC-specific models from scratch or treats FMs as off-the-shelf regressors (Dedhia et al., 2023; Richards et al., 2026), leaving open the question of why a frozen FM should outperform methods fitted on the same context. Classical SC represents the prediction as an affine rule of donor labels, forgoing nonlinearity entirely. FMs can bring in the nonlinearity. But does the gain survive the error of recovering it from a short window? And with context scarce, does how

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.

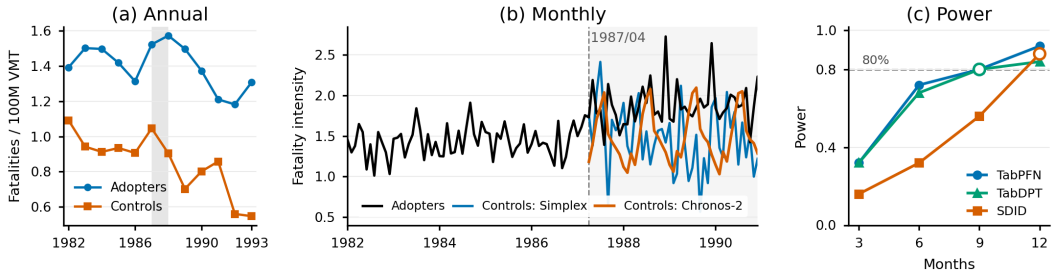


Figure 1: FARS data before and after the 1987 speed-limit change: (a) annual fatality rates; (b) monthly rates with predicted control series; (c) semi-synthetic hypothesis test power.

we present that context to the FM change what it recovers? We derive a break-even criterion for the first. For the second, we treat context-axis orientation, whether time periods or units form the rows, as part of the inference procedure and show that two-way centering reduces sensitivity to that choice.

Contributions.

- **Recoverable nonlinear headroom.** We formalize recoverable nonlinear headroom as the gain from relaxing linearity offset by finite-context recovery error (Proposition 1) and derive an exact break-even threshold. Our simulations show that frozen tabular FMs generally recover nonlinear structure more accurately than fitted nonlinear baselines (e.g., CatBoost) when context is sufficient.
- **Context-axis sensitivity and centering.** We show that frozen tabular FMs' predictions depend on whether periods or units form the context rows: transposing the same data from unit-as-row to time-as-row can inflate the MSE ratio by factors of 1.6–2.7 for the same checkpoint. We propose a two-way centering algorithm, which reduces orientation disagreement by 61% to 85% across three tabular FMs on real panels.
- **Interface-aware evaluation and application.** We score complete procedures rather than checkpoints, treating orientation and preprocessing as part of the estimator, across 673 prediction tasks from 12 published studies spanning economics, energy, public health, and politics. We show that better prediction converts into experimental power: in a semi-synthetic study on U.S. traffic-fatality data, TabPFN and TabDPT detect a 0.7-standard-deviation effect with 80% power in nine months versus twelve for SDID, a 25% reduction over the best baseline (Figure 1(c)).

2 Related Work

Synthetic control and causal panels. Classical SC uses convex donor weights (Abadie & Gardeazabal, 2003; Abadie et al., 2010; 2015). Extensions relax this weight geometry, address imperfect pre-period fit, or exploit denoising and low-rank structure (Doudchenko & Imbens, 2016; Ferman & Pinto, 2021; Ferman, 2021; Ben-Michael et al., 2021; Kellogg et al., 2021; Amjad et al., 2018; Xu, 2017; Athey et al., 2021). SDID, Time-Aware SC, and Targeted SC add time weighting, temporal structure, or learned initial estimators (Arkhangelsky et al., 2021; Rho et al., 2026; Wang et al., 2026). These studies propose alternative estimators. We use SC as a testbed for in-context learners and derive a break-even point at which nonlinear gains exceed finite-context estimation error.

Learned counterfactual prediction. CausalImpact regresses on donor series within a structural time-series model, while global forecasting and SyNBEATS pool related series (Brodersen et al., 2015; Grecov et al., 2021; Goldin et al., 2024). ArCo, tree-based SC, Synthetic Learner, and Deep SC fit flexible task-specific rules (Carvalho et al., 2018; Mühlbach & Nielsen, 2021; Viviano & Bradic, 2023; Ramachandra, 2025). CausalPFN and Do-PFN are pretrained specifically for causal estimation (Balazadeh et al., 2025; Robertson et al., 2025). We instead use frozen general-purpose FMs to make predictions, requiring no causal pretraining, fine-tuning, or pooling of related series.

Tabular and time-series foundation models. TabPFN, TabDPT, and TabFM provide in-context tabular prediction (Hollmann et al., 2025; Ma et al., 2025; Kong & Das, 2026), while Chronos, TimesFM, and related models forecast time series (Ansari et al., 2025; Das et al., 2024; Woo et al.,

2024; Moroshan et al., 2025). These models can receive donor information through different input constructions. Prior tabular-PFN work varies target order or feature encoding (Arbel et al., 2025; Ye et al., 2025). Linear panel methods distinguish horizontal and vertical regression (Athey et al., 2021; Shen et al., 2023). We show that this axis choice changes frozen tabular FM predictions and that two-way centering reduces this disagreement.

FM-based SC. SCouT pretrains a spatiotemporal Transformer for SC (Dedhia et al., 2023) rather than transferring a frozen general-purpose model. SCBench compares pretrained models with three linear baselines (Robust SC, Lasso, and Simplex) on simulated panels and includes no nonlinear baselines (Richards et al., 2026). We instead use a fitted nonlinear frontier to quantify the gain from relaxing linearity after accounting for finite-context recovery error, and characterize how context length and the inference interface shape this gain on simulated and real panels.

3 Model Framework and Setup

We follow the setup of Abadie et al. (2010). We observe $J+1$ units over periods $t = 1, \dots, T$. Unit 1, the *target*, receives an intervention after period T_0; the remaining units form the donor pool $\mathcal{C} = \{2, \dots, J+1\}$ and are never treated. Let $Y_{it}(1)$ and $Y_{it}(0)$ denote unit i's outcome at period t with and without the intervention. Assuming no anticipation and no interference, we observe $Y_{it}(0)$ everywhere except for the target after T_0, where we observe $Y_{1t} = Y_{1t}(1)$. Estimating the treatment effect $Y_{1t} - Y_{1t}(0)$ thus requires predicting the counterfactual $Y_{1t}(0)$, the target's outcome had it not been treated.

It is useful to view the untreated potential outcomes as a matrix whose rows are periods and whose columns are units:

$$\mathbf{Y}(0) = \left[\begin{array}{c|c} \mathbf{Y}_{\mathcal{C},\mathrm{pre}} & \mathbf{y}_{1,\mathrm{pre}} \\ \hline \mathbf{Y}_{\mathcal{C},\mathrm{post}} & \boxed{\mathbf{y}_{1,\mathrm{post}} = ?} \end{array}\right].$$

The three unboxed blocks are observed. Only the target's untreated post-treatment path, the counterfactual, is missing. Estimators differ in how they use the observed blocks to predict it.

Classical synthetic control. Synthetic control (SC) predicts the target at each post-treatment period as a weighted average of the donors' outcomes at that period. The weights are fitted on pre-treatment *predictors*. Let $X_1 \in \mathbb{R}^{d_X}$ hold the target's predictors and $X_0 \in \mathbb{R}^{d_X \times J}$ the donors', one column per donor. SC chooses donor weights $W = (w_2, \dots, w_{J+1})^\top$ so that $X_0 W$ matches X_1,

$$W^\star(\Omega) \in \arg\min_{W \in \mathcal{W}} (X_1 - X_0 W)^\top \Omega (X_1 - X_0 W), \qquad \mathcal{W} = \Big\{ W \in \mathbb{R}^J : w_j \geq 0,\ \textstyle\sum_{j \in \mathcal{C}} w_j = 1 \Big\}, \tag{1}$$

and reuses them at every post-treatment period. In the general formulation of Abadie et al. (2010), predictors may combine pre-treatment outcomes with covariates, and a symmetric positive semidefinite matrix Ω, typically diagonal, sets their relative importance. We use the simpler setup in which the predictors are the pre-treatment outcomes themselves, $X_1 = \mathbf{y}_{1,\mathrm{pre}}$, $X_0 = \mathbf{Y}_{\mathcal{C},\mathrm{pre}}$, and $d_X = T_0$, with $\Omega = I$. SC is then a least-squares regression of the target's pre-treatment path on the donors' paths, with no intercept and the weights restricted to the simplex. Because the weights are nonnegative and sum to one, each prediction stays within the range of donor outcomes.

Attention form. An attention rule predicts for a *query* by averaging *values*, with weights computed from the query and the *keys*. SC has this form, with each donor as one example: its key is its predictor vector (a column of $K = X_0$); its value is its outcome $v_t = \mathbf{Y}_{\mathcal{C},t}$ at $t > T_0$; and the query is the target's predictor vector $q = X_1$. For any fixed Ω,

$$\widehat{Y}_{1t}(0) = W^\star(\Omega)^\top v_t = a(q, K; \Omega)^\top v_t, \qquad a(q, K; \Omega) = W^\star(\Omega). \tag{2}$$

SC is thus a restricted attention rule: a single query whose simplex weights come from a quadratic program rather than a softmax. Lemma 1 shows that SC and softmax score donors by the same quantity, $K^\top \Omega q$, and differ in how those scores become weights. Classical SC is linear in v_t, whereas an FM given the same examples and query may learn a nonlinear rule. We next ask when such a nonlinear rule improves prediction, and whether that gain can be recovered from a short pre-treatment context.

Recoverable nonlinear headroom. Fix $t > T_0$, write $Y = Y_{1t}(0)$, $v = \mathbf{Y}_{\mathcal{C},t}$, and $m_v = \mathbb{E}[Y \mid v]$. Let $\mathcal{A}$ denote the affine donor rules. Holding the prediction input v fixed, the nonlinear approximation opportunity is

$$\Delta_{\text{lin}} := \inf_{f \in \mathcal{A}} \mathbb{E}[(m_v - f(v))^2].$$

This is the excess risk of the best affine donor rule over m_v. Appendix B.2 places this quantity in a full risk decomposition separating information loss and the additional cost of the simplex restriction.

A fitted method improves on the affine rule only if it estimates m_v accurately enough from the available pre-treatment periods.

Let $\mathcal{D}_- = \{(v_s, Y_{1s}) : s \le T_0\}$ be the labeled pre-treatment context. A learning rule L is the complete procedure, including preprocessing, fixed checkpoint, and random seed U_L, which is constant for a deterministic rule. It returns $\widehat{m}_L(v) = L(\mathcal{D}_-, v; U_L)$, with finite-context recovery error $\mathcal{E}_L = \mathbb{E}[\{\widehat{m}_L(v) - m_v\}^2]$.

We compare methods that receive the same prediction inputs and assume $\mathbb{E}[Y - m_v \mid \mathcal{D}_-, v, U_L] = 0$. Thus the remaining evaluation error has mean zero after the context, donor vector, and seed are known.

Proposition 1 (Recoverable nonlinear headroom)**.** *Under the condition above, the learning rule's expected gain over the best population affine donor rule is*

$$\inf_{f \in \mathcal{A}} \mathbb{E}[(Y - f(v))^2] - \mathbb{E}[(Y - \widehat{m}_L(v))^2] = \Delta_{\text{lin}} - \mathcal{E}_L =: G_L.$$

Proof. See Appendix B.3. □

Corollary 1 (Finite-context break-even point for OLS)**.** *In the residualized orthogonal-feature calibration, with d_ϕ nonlinear features and noise variance σ^2, Ordinary Least Squares (OLS) satisfies $\mathcal{E}_{\text{OLS}} = \sigma^2 d_\phi / T_0$ and $G_{\text{OLS}} = \Delta_{\text{lin}} - \sigma^2 d_\phi / T_0$. OLS improves on the population affine rule exactly when $\Delta_{\text{lin}} > \sigma^2 d_\phi / T_0$.*

Context-axis sensitivity. A tabular FM has no built-in roles for time periods and units, so the same panel and frozen checkpoint admit the two orientations in Figure 2. In the time-row interface, T_0 periods are labeled rows, J donors are features, and the post-period donor vector is the query. In the unit-row interface, J donors are labeled rows, T_0 periods are features, and the treated history is the query, with a separate fit at each horizon. The two orientations use the same observed outcomes but different context sizes, labels, and queries, so orientation is an inference-time choice. That choice alone can change a checkpoint's predictions substantially (EX3, Section 4.2).

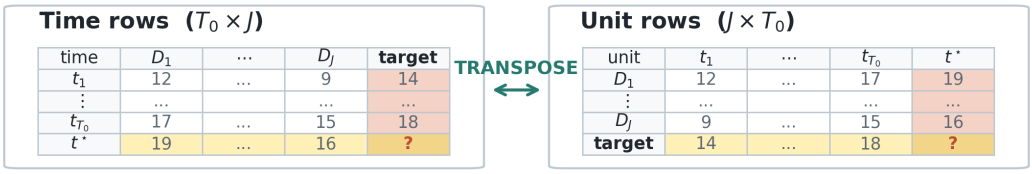


Figure 2: Two tabular interfaces for the same masked panel data. Transposition exchanges the example and feature axes, changing the context from $T_0 \times J$ to $J \times T_0$. The row in yellow marks the query with the masked target; $t^\star$ denotes one post-intervention time.

Ridge regression, by contrast, is orientation-invariant under a common penalty, no intercept, and no orientation-specific scaling (Shen et al., 2023; Appendix B.4). That result holds for any T_0 and J.

Proposition 2 (Context-range disagreement floor)**.** *Let $\widehat{\mathbf{y}}^{\text{time}}$ and $\widehat{\mathbf{y}}^{\text{unit}}$ be the paths from a fixed checkpoint and preprocessing rule, and let gap_h be the distance between the two context-label ranges at horizon h. If the predictions lie within η^{time} and η^{unit} of their respective label ranges, for every h:*

$$\left|\widehat{y}_h^{\text{time}} - \widehat{y}_h^{\text{unit}}\right| \ge \left(\text{gap}_h - \eta^{\text{time}} - \eta^{\text{unit}}\right)_+. \tag{3}$$

The time-row and unit-row interfaces expose the model to different label ranges. If the gap between those ranges exceeds the sum of the two extrapolation bounds, their predictions must differ.

Two-way centering. Motivated by the double difference in SDID (Arkhangelsky et al., 2021), we propose *two-way centering*, a model-agnostic wrapper that lets any frozen learner predict on residualized outcomes in either orientation:

$$\bar{y}_i = T_0^{-1} \sum_{s \le T_0} Y_{is}, \qquad \bar{c}_t = J^{-1} \sum_{j \in \mathcal{C}} (Y_{jt} - \bar{y}_j), \qquad \widetilde{Y}_{it} = Y_{it} - \bar{y}_i - \bar{c}_t,$$

Here $\bar{y}_i$ is unit i's pre-treatment mean and $\bar{c}_t$ the average donor deviation from its own pre-treatment mean at (t). The construction mirrors the double difference in SDID, with simple averages in place of SDID's optimized unit and time weights (Appendix B.5). The learner predicts $\widetilde{Y}_{1t}$, and the wrapper adds back $\bar{y}_1 + \bar{c}_t$, the mean donor outcome at t plus the target's average pre-treatment gap from the donors. We prove that, for any fixed base learner, adding unit levels or a common time path to the panel changes neither the prediction errors nor the disagreement between time-row and unit-row predictions (Proposition 3). The intuition is that two-way centering removes unit levels and common time movements before model estimation, so neither can account for any orientation gap that remains. We test whether centering also lowers MSE in Section 4.

4 Experiments and Results

Our empirical design uses three kinds of evaluation data. Simulations have known counterfactuals, so we can measure recovery error directly (EX1, EX2a). Masked-origin tasks hide a real unit's observed outcomes after an artificial cutoff, which gives held-out prediction error on real panel structure (EX2b-EX4). Placebo studies inject effects into untreated data to estimate the false-positive rate (FPR) in A/A tests and statistical power (EX1(c), EX5).

Evaluation protocol. All experiments share one prediction setup. At origin T_0, each method sees the target through T_0 and the donors through $T_0 + H$; the target's next H outcomes are withheld for scoring. For real masked-origin task k and method m, the loss is

$$\mathcal{L}_{km} = \frac{1}{H_k} \sum_{h=1}^{H_k} \left[Y_{k,T_0+h}(0) - \widehat{Y}_{km,T_0+h}(0) \right]^2.$$

In simulations, we instead score recovery error against the known noiseless conditional mean m_v (Proposition 1). We compare three families of models: linear SC baselines (Simplex SC, SDID, ridge), fitted nonlinear learners (kernel ridge regression (KRR) and CatBoost), and frozen FMs, both tabular (TabPFN-2.6, TabDPT, TabFM) and time-series (Chronos-2, TimesFM-2.5); Appendix D gives implementations. All comparisons are paired on tasks eligible for every method.

4.1 Nonlinear opportunity and recovery cost

EX1: Nonlinear-Headroom. EX1 uses two data-generating processes (DGPs) to test how methods differ in recovering nonlinearity as curvature is varied. Both DGPs are evaluated with varying available context in the pre-treatment periods and number of donors ((T_0, J) in $\{32, 128, 512\} \times \{20, 64\}$). In EX1(a), donors are standardized Gaussian AR(1) processes and

$$Y_{1t}(0) = W_0^\top v_t + \frac{\kappa}{\sqrt{d_\phi}} \sum_{j=1}^{d_\phi} \frac{[(v_t)_j]^2 - 1}{\sqrt{2}} + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma^2). \tag{4}$$

where W_0 assigns weight 0.2 to each of five donors. Thus Simplex SC represents the affine component exactly. The nonlinear component has variance κ^2 and is orthogonal to the affine donor span, so $\Delta_{\text{lin}} = \kappa^2$. We normalize curvature by $\kappa^\star = \sigma\sqrt{d_\phi / T_0}$, the true-dictionary OLS break-even point.

Results: EX1(a)-Nonlinear Headroom. Figure 3 shows that nonlinear opportunity yields gains only when the curvature is strong enough and there is sufficient context. At $T_0 = 32$, nearly all

methods remain below break-even; with longer histories, TabFM recovers the most headroom, followed by TabDPT, while TabPFN remains closer to the nonlinear baselines KRR and CatBoost. Ridge and simplex approach break-even from below because they cannot represent the nonlinear component. TimesFM-2.5 recovers little headroom whether it omits donors (target-only) or adds them through an OLS adjustment (XReg). Chronos-2 also remains below or near break-even despite using nonlinear group attention over donors (Ansari et al., 2025). The dashed line labeled "oracle" represents OLS given the true affine component and nonlinear dictionary, leaving only the nonlinear coefficients to estimate.

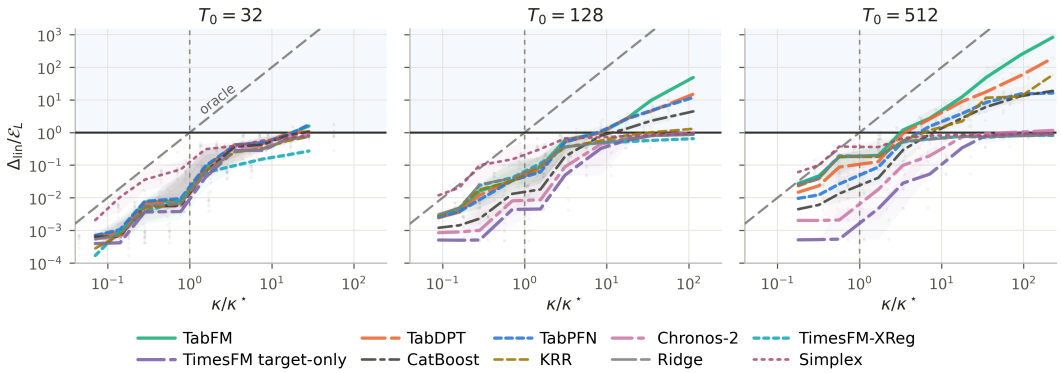


Figure 3: **Recoverable nonlinear headroom vs. curvature,** $J = 20$**.** Ratio of recovery opportunity over cost $\Delta_{\text{lin}}/\widehat{\mathcal{E}}_{L,H}$ versus normalized curvature $\kappa/\kappa^\star$. Upper shaded area indicates positive nonlinear headroom. Dashed diagonal shows OLS given the true DGP structure. Curves are cell medians and bands are interquartile ranges over 30 paired seeds per cell.

Table 1 confirms the visual ordering after exact task pairing. The patterns are similar for $J = 64$ with 44 non-informative donors added. TabFM outperforms both nonlinear baselines at both donor counts. TabPFN does the same except for one interval ($J = 20$ vs. KRR). TabDPT also improves on both baselines, although its advantage over CatBoost narrows when the donor pool grows.

Table 1: Recovery MSE relative to fitted nonlinear baselines (lower is better). Within each panel seed, entries are geometric mean MSE ratios. Confidence intervals are calculated from the 30 seed-level mean log ratios using a t-distribution.

	TabFM		TabPFN		TabDPT	
J	CatBoost	KRR	CatBoost	KRR	CatBoost	KRR
20	0.33 [0.28, 0.38]	0.55 [0.49, 0.62]	0.52 [0.48, 0.57]	0.87 [0.75, 1.02]	0.43 [0.38, 0.48]	0.71 [0.64, 0.79]
64	0.36 [0.32, 0.42]	0.34 [0.31, 0.37]	0.44 [0.40, 0.49]	0.41 [0.36, 0.47]	0.82 [0.72, 0.93]	0.76 [0.70, 0.82]

EX1(b): Nonlinear Headroom with Donor Interactions. EX1(b) changes the nonlinear component to a nonseparable sinusoidal interaction from Friedman (1991). The results are similar to EX1(a) at both donor-pool sizes. Nonlinear recovery remains limited with short histories and improves as T_0 grows (Appendix D.2).

EX1(c): Nonlinear Headroom and Experimental Power. Using the simulated data from EX1(a), we test whether better counterfactual recovery improves power at the fixed $H = 24$ horizon. After prediction, we inject effects of δ pre-treatment standard deviations and use permutation tests (Abadie (2021b)) to estimate A/A FPR and power. Power is the fraction of simulations in which the test detects the injected effect, the true positive rate (TPR). The FPR is the fraction in which it detects an effect when none is injected.

Results: EX1(c). Figure 4 shows that experimental power follows the pattern of recoverable nonlinear headroom: tabular FMs have no consistent power advantage when $\kappa = 0$, but achieve higher power at smaller effect sizes as the nonlinear component grows. Appendix Table 7 reports the A/A test results. Tabular-FM FPRs are at or below 5%, except for TabDPT and TabPFN at $\kappa = 1$, both at 5.5%.

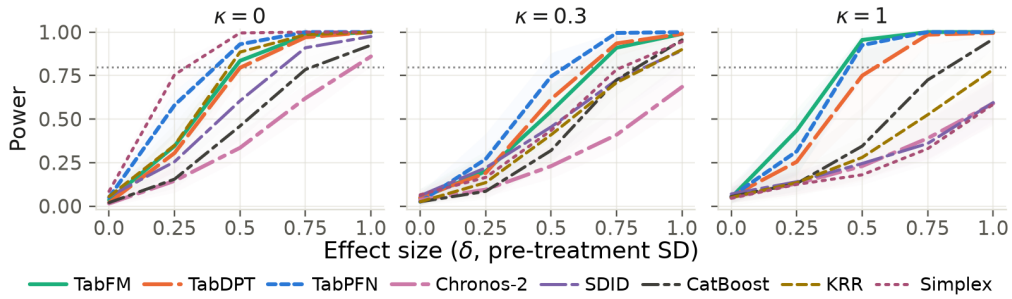


Figure 4: **EX1(c): Nonlinear headroom simulated power curve.** Power by planted effect size in pre-treatment standard deviations, for $\kappa \in \{0, 0.3, 1\}$ with the quadratic nonlinear map as in EX1(a); dotted line marks 80% power; further left is better: detecting smaller effects at a fixed 5% target FPR.

4.2 Range Response and Orientation Risk

Orientation-centering grid. For each frozen tabular checkpoint, we evaluate a 2×2 *orientation-centering grid* with four cells: $\{\text{time-row}, \text{unit-row}\} \times \{\text{raw}, \text{two-way centered}\}$. Section 3 defines the centering transform, and Algorithm 1 gives its leakage-free implementation. The time-series FMs (Chronos-2 and TimesFM variants) retain their native temporal interfaces.

EX2: Context-Label-Range. EX2 measures how far each procedure predicts beyond the range of labels in its context. The normalized out-of-range distance is zero when the prediction lies inside the range; otherwise, it is the distance to the nearest endpoint divided by the range width. We first evaluate synthetic tabular tasks whose true targets lie outside the context-label range (Appendix D.3), then real panel data to characterize model- and orientation-specific extrapolation.

Result: EX2. Appendix Table 8 shows that the extrapolation bounds and out-of-range frequencies differ substantially across procedures, increasing from CatBoost to TabDPT, TabPFN, TabFM, and ridge. On real panels, TabFM predicts outside its context-label range most often and by the largest distances among the three tabular FMs (Figure 5A), and all three stay closer to the range under unit-row. These results describe extrapolation behavior, not accuracy.

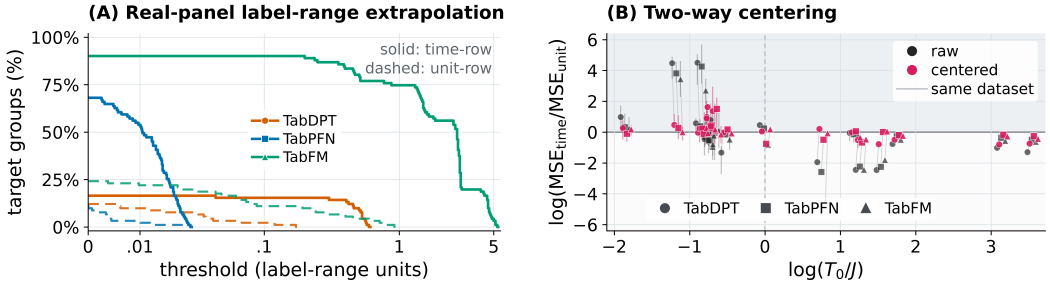


Figure 5: **(A)** Share of real-panel clusters above each out-of-range threshold (solid: time-row; dashed: unit-row). **(B)** Log MSE ratios of time-row vs. unit-row before and after two-way centering. Positive values reflect lower MSE by unit-row and vice versa; centering moves the log-ratio toward zero.

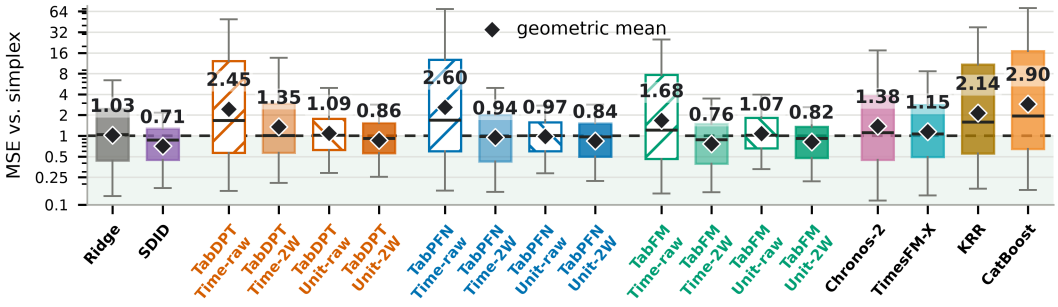
EX3: TwoWay-Center. Proposition 3 predicts that two-way centering removes additive unit levels and common time movements from the orientation comparison. We evaluate the complete orientation-centering grid on real masked origins. For each checkpoint and transform, we compare time-row and unit-row through their signed log MSE ratio. We average repeated origins within target, then report the median and interquartile range across targets for each dataset and T_0/J combination. A linear ANCOVA relates this ratio to $\log(T_0/J)$, centering, their interaction, and checkpoint indicators. Log MSE ratios near zero mean that orientation matters less. The sign identifies the better orientation. This analysis measures orientation sensitivity but cannot establish that centering lowers MSE.

Results: EX3. Figure 5B shows that raw unit-row prediction tends to perform better at low T_0/J, and time-row prediction at high T_0/J, consistent with FMs benefiting from more labeled rows. Centering moves most matched groups toward zero: mean absolute log-ratios fall from 1.42 to 0.56 for TabDPT, 1.10 to 0.36 for TabPFN, and 1.22 to 0.18 for TabFM. In the checkpoint-adjusted ANCOVA, the $\log(T_0/J)$ slope attenuates from -0.57 to -0.19, showing less dependence on input geometry after centering.

4.3 Real-panel benchmark

EX4: Real-Benchmark. EX4 compares prediction error on 673 masked-origin tasks from 12 real-panel datasets, excluding tasks with $T_0 < 5$. For each task, we compute the MSE ratio to simplex SC. We summarize these ratios by their geometric mean, with equal weighting at three nested levels: datasets, targets within datasets, and tasks within targets. Values below one favor the method. Figure 6 (top) reports all four orientation-centering cells of each tabular FM, holding the orientation fixed across tasks. The bottom table breaks the results down by history length, reporting the unit-row cell for $T_0 < 2J$ and the time-row cell for $T_0 \geq 2J$.

Results: EX4. SDID has the lowest overall geometric-mean MSE ratio (0.71). With raw inputs, orientation matters: time-row ratios are 1.6 to 2.7 times the unit-row ratios. Two-way centering lowers the ratio in all six checkpoint-orientation cells, and five of the six centered cells fall below one; centered time-row TabDPT is the exception (1.35). The better centered orientation depends on the checkpoint: unit-row for TabDPT and TabPFN, time-row for TabFM. In the descriptive breakdown, SDID has the lowest point estimate for $T_0 < 2J$, whereas centered time-row TabFM has the lowest for $T_0 \geq 2J$. Appendix D reports every cell in each group.



	SDID	TabDPT 2W	TabPFN 2W	TabFM 2W	Chronos-2	TimesFM-XReg
$T_0 < 2J$	**.66** [.44, .98]	.78 [.62, .99]	.84 [.64, 1.11]	.79 [.57, 1.11]	1.73 [.54, 5.52]	1.34 [.86, 2.10]
$T_0 \geq 2J$	.85 [.54, 1.34]	.81 [.39, 1.70]	.76 [.55, 1.06]	**.70** [.42, 1.15]	.76 [.13, 4.43]	.75 [.34, 1.69]

Figure 6: **Real-panel benchmark.** MSE ratio to simplex SC; lower is better. Top: 10th-90th percentile whiskers; diamonds are geometric means. Bottom: by history length, with 95% intervals.

4.4 Application to FARS Data

EX5: Semi-Synthetic Power Simulation. EX5 evaluates whether FMs can detect a treatment effect sooner in a semi-synthetic FARS experiment (Figure 1). We use two treated cohort designs. The first preserves the design of Ashenfelter & Greenstone (2004): the 10 states that are the real adopters in April 1987, with one deviation. Their seven comparison states admit only eight assignment ranks, so the smallest attainable p-value is 1/8 and no 5% placebo test exists, and we therefore use all 28 states not adopted by April 1987 as donors. The second design permutes the assignment itself, pooling 10 states drawn at random from the same 38-state set into one pooled pseudo-target, with 40 compositions at each of the 25 origins for 1,000 held-out assignments. The first design asks whether the gain appears in the historical adopter cohort, the second whether any FM gains are robust on a random cohort composition.

Each prediction uses $T_0 = 27$ pre-treatment periods and $J = 28$ donors ($T_0/J = 0.96$), placing EX5 on the short-history side of EX4's descriptive split. At each monthly pre-treatment origin, we

inject effects of 0.5-1.0 pre-treatment SDs and test after 3, 6, 9, and 12 months. We use a single-step max-over-horizons randomization test, calibrated on a separate placebo set to limit the probability of any false rejection across the four horizons to 5% (Westfall & Young, 1993).

Results: EX5. Table 2 summarizes eight complete model-interface procedures (see Appendix D for full results). On the historical cohort, A/A FPRs range from 2.6% to 4.6%, below 5% for every method. No method reaches 80% at $\delta = 0.5$; at $\delta = 1.0$, unit-row TabPFN, TabDPT, and SDID reach it by month 6. At $\delta = 0.7$, unit-row TabPFN and TabDPT reach the threshold at month 9, versus month 12 for SDID and other procedures that reach it, a 25% reduction in detection time.

Under permutation the model family ordering persists while the individual cell ordering does not. The best FM cells still detect earliest, reaching 80% power for a $\delta = 1.0$ effect at month 9, before other baselines reach it at month 12. Averaged over the four effect sizes, the FM frontier leads month-12 power by 9.8 points over the best linear baseline and 7.5 over the best nonlinear one. Two-way centering raises month-12 power in all 24 comparisons, by 8 points on average.

Table 2: Simulated power and A/A test in the FARS placebo study (raw inputs). Power: fraction of injected effects detected by each month across 25 artificial intervention dates. A/A statistics (1,000 held-out placebo assignments): FPR through month 12 and RMSE of the 12-month average effect. δ and effect RMSE are in pre-treatment SD units. Bold: first month reaching 80% power.

	A/A test		$\delta = 0.5$				$\delta = 0.7$				$\delta = 1.0$			
Method	FPR	RMSE	3	6	9	12	3	6	9	12	3	6	9	12
TabPFN, unit-row	.035	.352	.08	.56	.60	.68	.32	.72	**.80**	.92	.60	**.88**	.96	1.00
TabDPT, unit-row	.041	.340	.16	.44	.60	.68	.32	.68	**.80**	.84	.60	**.88**	.96	1.00
TabPFN, time-row	.026	.294	.00	.00	.04	.20	.00	.00	.48	**.84**	.08	.44	**.88**	1.00
TabFM, time-row	.044	.288	.00	.00	.04	.24	.00	.08	.48	**.84**	.12	.40	**.88**	1.00
SDID	.036	.347	.04	.08	.12	.32	.16	.32	.56	**.88**	.48	**.84**	1.00	1.00
Kernel ridge	.046	.313	.00	.00	.00	.24	.04	.04	.52	**.80**	.12	.40	**.88**	1.00
CatBoost	.041	.302	.00	.00	.12	.44	.00	.04	.52	.72	.20	.40	**.84**	1.00
Chronos-2	.040	.280	.00	.00	.16	.44	.04	.24	.56	.68	.24	.60	**.92**	1.00

5 DISCUSSION

Takeaway. We show that frozen tabular FMs can improve synthetic control when nonlinear donor–target structure is recoverable from the available context. These gains depend on the full prediction procedure: changing data orientation can substantially alter predictions, while two-way centering reduces this sensitivity. The semi-synthetic FARS study illustrates the practical benefit, showing that suitable FM procedures can shorten the experiment window needed to detect an effect and improve hypothesis test power compared to state-of-the-art baselines including SDID.

Practical procedure. Our findings suggest selecting the checkpoint, orientation, and preprocessing jointly through pre-treatment validation. For tabular FMs, this means comparing both orientations, with and without two-way centering, against classical SC learners (simplex/SDID) and fitted nonlinear baselines. Rolling holdouts should reflect the application's horizon, evaluation objective, and available information; donor placebos can provide an additional check across units. Appendix E presents one such selection procedure (Figure 11) and evaluates it on EX1(a) simulations with long pre-treatment histories (Figure 12).

Limitations. First, released checkpoints differ jointly in pretraining data, architecture, output parameterization, and interface, so our comparisons do not isolate the effect of pretraining. Second, masked-origin tasks in pre-treatment history provide observed targets for prediction-error evaluation, but the label that matters, the treated unit's untreated path, is permanently masked. Transport of pre-treatment performance to the post-treatment period is therefore assumed rather than verified. Third, the split at $T_0/J = 2$ is descriptive. Our simulations and 12-panel corpus do not support a rule for choosing the better orientation from panel geometry alone. FM-based synthetic control should therefore be evaluated as a complete procedure: checkpoint, orientation, preprocessing, and available context must be considered together.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

We used generative AI tools to assist with manuscript writing and coding, to polish prose, to turn experimental outputs into figures and illustrations, and to summarize relevant literature. We have not used generative AI tools to generate research ideas, to design the experiments, to interpret the results, or to write the original manuscript; these are the authors' own work. All code and artifacts have been checked by at least two authors.

ETHICS STATEMENT

This work uses only public, aggregate panel data and involves no human participants or person-level records. All source datasets are cited with their public repositories in Appendix C. Because synthetic-control estimates can inform consequential policy decisions, the model-generated counterfactuals here should be read with their stated uncertainty and validated against domain knowledge rather than treated as ground truth or used as the sole basis for a decision.

REPRODUCIBILITY STATEMENT

Full assumptions and proofs for all theoretical results appear in Appendix B. Appendix C documents every dataset, its provenance, and its panel construction; Appendix D gives the simulation data-generating processes, pre-processing, randomization-test procedures, and evaluation protocols, with exact model and package versions in Table 5 and tuning grids in Table 6. All foundation models are used as frozen checkpoints without fine-tuning.

REFERENCES

Alberto Abadie. Replication files for "using synthetic controls: Feasibility, data requirements, and methodological aspects". openICPSR, 2021a. URL https://doi.org/10.3886/E119933V1.

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021b.

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51, 2018.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Michael Arbel, David Salinas, and Frank Hutter. Equitabpfn: A target-permutation equivariant prior fitted network. *arXiv preprint arXiv:2502.06684*, 2025.

Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

Orley Ashenfelter and Michael Greenstone. Using mandated speed limits to measure the value of a statistical life. *Journal of Political Economy*, 112(S1):S226–S267, 2004.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536): 1716–1730, 2021.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. *arXiv preprint arXiv:2506.07918*, 2025.

Patrick Bayer. European union sectoral emissions data (eused): Data and codebook, 2019. URL https://doi.org/10.7910/DVN/DW7Y5W.

Patrick Bayer and Michaël Aklin. The european union emissions trading system reduced co2 emissions despite low prices. *Proceedings of the National Academy of Sciences*, 117(16):8804–8812, 2020. doi: 10.1073/pnas.1918128117.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021.

Konstantin Bogatyrev and Lukas F. Stoetzer. Estimating treatment effects on proportions with synthetic controls. *Political Analysis*, pp. 1–17, 2026a. doi: 10.1017/pan.2026.10046.

Konstantin Bogatyrev and Lukas F. Stoetzer. Replication data for: Estimating treatment effects on proportions with synthetic controls. Harvard Dataverse, 2026b. URL https://doi.org/10.7910/DVN/MPUEIC.

Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9(1):247–274, 2015.

Carlos Carvalho, Ricardo Masini, and Marcelo C. Medeiros. ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics*, 207(2):352–380, 2018.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10148–10167. PMLR, 2024.

Bhishma Dedhia, Roshini Balasubramanian, and Niraj K. Jha. SCouT: Synthetic counterfactuals via spatiotemporal transformers for actionable healthcare. *ACM Transactions on Computing for Healthcare*, 4(4):1–28, 2023.

Nikolay Doudchenko and Guido W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical Report 22791, National Bureau of Economic Research, 2016.

Eric Dunford. smoking: Proposition 99 and cigarette consumption (dataset in the tidysynth R package). CRAN, 2025. URL https://search.r-project.org/CRAN/refmans/tidysynth/html/smoking.html.

Bruno Ferman. On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association*, 116(536):1764–1772, 2021.

Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221, 2021.

Jared J. Finnegan, Phillip Y. Lipscy, Jonas Meckling, and Florence Metz. Replication data for: The institutional sources of economic transformation: Insulation and compensation in the politics of energy transitions. Harvard Dataverse, 2024. URL https://doi.org/10.7910/DVN/INHODY.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Jacob Goldin, Julian Nyarko, and Justin Young. Forecasting algorithms for causal inference with panel data. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 1205. ACM, 2024.

Priscila Grecov, Kasun Bandara, Christoph Bergmeir, Klaus Ackermann, Sam Campbell, Deborah Scott, and Dan I. Lubman. Causal inference using global forecasting models for counterfactual prediction. In *Advances in Knowledge Discovery and Data Mining*, pp. 282–294. Springer, 2021.

Jens Hainmueller and Alexis Diamond. basque: Panel data from spanish regions (dataset in the Synth R package). CRAN. URL https://search.r-project.org/CRAN/refmans/Synth/html/basque.html.

William B. Hankins. Replication data for: Finally, nebraska: A synthetic control analysis of legislative structure. UNC Dataverse, 2019. URL https://doi.org/10.15139/S3/N7N38V.

William B. Hankins. Finally, nebraska: A synthetic control analysis of legislative structure. *State Politics & Policy Quarterly*, 20(1):3–22, 2020. doi: 10.1177/1532440019868818.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

Maxwell Kellogg, Magne Mogstad, Guillaume A. Pouliot, and Alexander Torgovitsky. Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American Statistical Association*, 116(536):1804–1816, 2021.

Weihao Kong and Abhimanyu Das. Introducing TabFM: A zero-shot foundation model for tabular data. Google Research Blog, June 2026. URL https://research.google/blog/introducing-tabfm-a-zero-shot-foundation-model-for-tabular-data/.

David Lang, Lief Esbenshade, and Robb Willer. Replication data for: Did ohio's vaccine lottery increase vaccination rates? a pre-registered, synthetic control study. Harvard Dataverse, 2021. URL https://doi.org/10.7910/DVN/QYXN9L.

David Lang, Lief Esbenshade, and Robb Willer. Did ohio's vaccine lottery increase vaccination rates? a pre-registered, synthetic control study. *Journal of Experimental Political Science*, 10(2):242–260, 2023. doi: 10.1017/XPS.2021.32.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 172692–172722, 2025.

André F. T. Martins and Ramón Fernandez Astudillo. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pp. 1614–1623, 2016.

Vladyslav Moroshan, Julien Siems, Arber Zela, Timur Carstensen, and Frank Hutter. TempoPFN: Synthetic pre-training of linear RNNs for zero-shot time series forecasting. *arXiv preprint arXiv:2510.25502*, 2025.

Nicolaj Søndergaard Mühlbach and Mikkel Slot Nielsen. Tree-based synthetic control methods: Consequences of moving the US embassy. *arXiv preprint arXiv:1909.03968*, 2021.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

National Highway Traffic Safety Administration. Fatality analysis reporting system (fars). U.S. Department of Transportation, Data.gov. URL https://catalog.data.gov/dataset/fatality-analysis-reporting-system-fars.

Rajat Kanti Paul and Ankit Dixit. Replication data for: Regional greenhouse gas initiative as coal-displacement policy: Heterogeneity-robust difference-in-differences evidence from 2005–2023 and implications for the cost per tonne of CO2 abated. Harvard Dataverse, 2026. URL https://doi.org/10.7910/DVN/TFSSM5.

Vikas Ramachandra. Deep synthetic controls: Penalized, representation-learned, sparsity-aware counterfactual estimation. *Authorea preprint*, 2025. doi: 10.22541/au.176072431.11742213/v1. Version 1, 17 October 2025.

Saeyoung Rho, Cyrus Illick, Samhitha Narasipura, Alberto Abadie, Daniel Hsu, and Vishal Misra. Time-aware synthetic control. *arXiv preprint arXiv:2601.03099*, 2026.

Megan Richards, Saeyoung Rho, and Kyunghyun Cho. SCBench: A testbed for causal inference with time series panel data. In *Proceedings of the 2nd ICML Workshop on Foundation Models for Structured Data*, 2026. Workshop paper.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In *Advances in Neural Information Processing Systems*, 2025.

Dennis Shen, Peng Ding, Jasjeet Sekhon, and Bin Yu. Same root different leaves: Time series and cross-sectional methods in panel data. *Econometrica*, 91(6):2125–2154, 2023. doi: 10.3982/ECTA21248.

Percy Taabazuing and Jimoh Fatoki. From trade diversion to productive capacity: A synthetic control assessment of vietnam's trade war dividend, 2005–2024. Harvard Dataverse, 2026. URL https://doi.org/10.7910/DVN/SQBLUX.

United States Congress. Surface transportation and uniform relocation assistance act of 1987, 1987. URL https://www.govinfo.gov/app/details/STATUTE-101/STATUTE-101-Pg132. Public Law 100-17, 101 Stat. 132.

U.S. Department of Labor, Employment and Training Administration. Unemployment insurance weekly claims and extended benefits trigger data (eta 539). Data.gov. URL https://catalog.data.gov/dataset/unemployment-insurance-weekly-claims-extended-benefits-trigger-data-eta-539.

Davide Viviano and Jelena Bradic. Synthetic learner: Model-free inference on treatments over time. *Journal of Econometrics*, 234(2):691–713, 2023.

Yuxin Wang, Dennis Frauen, Emil Javurek, Konstantin Hess, Yuchen Ma, and Stefan Feuerriegel. Targeted synthetic control method. *arXiv preprint arXiv:2602.04611*, 2026.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley-Interscience, New York, 1993. ISBN 9780471557616.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at tabpfn v2: Understanding its strengths and extending its capabilities. *arXiv preprint arXiv:2502.17361*, 2025.

A NOTATION

Table 3: Notation used throughout the paper.

	Panel, outcomes, and indices
$T,\ T_0,\ H$	Total periods, pre-intervention periods, and post-intervention horizon
$J,\ \mathcal{C}$	Number of donors and donor pool $\mathcal{C} = \{2, \dots, J+1\}$
$i,\ j,\ t,\ s,\ h,\ k,\ m,\ o$	Unit, donor, period, pre-period or rolling-origin, horizon, task, method, and interface indices, with $o \in \{\text{time}, \text{unit}\}$
$Y_{it}(0),\ Y_{it}(1),\ Y_{it}$	Potential outcomes without and under intervention, and observed outcome
$\mathbf{Y}(0),\ X^-,\ y^-,\ X^+$	Untreated outcome matrix and its donor-pre, target-pre, and donor-post blocks, of sizes $T_0 \times J$, $T_0 \times 1$, and $H \times J$
$\mathbf{y}_{1,\text{post}},\ v_t,\ Y,\ v$	Missing target post-period path, donor vector $v_t = \mathbf{Y}_{\mathcal{C},t}$, and shorthand for $Y_{1t}(0)$ and v_t at a fixed post-intervention period
P	Observed panel $\mathsf{P} = (X^-, y^-, X^+)$
	Synthetic control and attention
$W,\ W^\star(\Omega),\ \mathcal{W}$	Generic and selected optimal donor weights, and simplex $\mathcal{W} = \{(w_j)_{j\in\mathcal{C}} : w_j \geq 0, \sum_{j\in\mathcal{C}} w_j = 1\}$
$X_1,\ X_0,\ d_X$	Target predictor vector, donor predictor matrix, and predictor dimension, with $X_1 = y^-$, $X_0 = X^-$, and $d_X = T_0$ for outcome-only predictors
$q,\ K,\ \Omega$	Attention query, key matrix, and positive-semidefinite predictor metric, with $q = X_1$ and $K = X_0$
$a(q, K; \Omega)$	Simplex attention map, $a(q, K, \Omega) = W^\star(\Omega)$
$\lambda_{\text{att}},\ \text{Ent}(W)$	Softmax temperature and entropy $\text{Ent}(W) = -\sum_{j\in\mathcal{C}} w_j \log w_j$
	Prediction classes and population risk
$\mathcal{S}$	Simplex class $\{v \mapsto w^\top v : w \in \mathcal{W}\}$
$\mathcal{A},\ \gamma_0,\ \gamma$	Affine class $\{v \mapsto \gamma_0 + \gamma^\top v\}$ and its coefficients
$\mathcal{I},\ m_{\mathcal{I}},\ m_v$	Admissible information and conditional means $m_{\mathcal{I}} = \mathbb{E}[Y \mid \mathcal{I}]$ and $m_v = \mathbb{E}[Y \mid v]$
$R^\star$	Irreducible risk given the admissible panel
$I_{\text{panel}},\ \Delta_{\text{lin}},\ \Delta_{\text{geo}}$	Costs of reducing the panel to v, restricting prediction to an affine rule, and further restricting that rule to the simplex
$s_t,\ a_0,\ a_j$	Latent state, target loading, and donor loadings in the factor-model remark
	Finite-context learning
$f,\ f^\star,\ L,\ U_L$	Prediction rule, affine projection of m_v, complete learning rule, and seed
$\mathcal{D}_-,\ \widehat{m}_L$	Pre-period sample $\mathcal{D}_- = \{(v_s, Y_{1s}) : s \leq T_0\}$ and fitted donor–target map
$\mathcal{E}_L,\ G_L$	Population recovery error and signed recovered headroom
$\phi,\ d_\phi,\ \theta,\ \Phi$	Nonlinear dictionary and dimension, coefficient vector, and pre-period design matrix
$\sigma^2,\ \varepsilon,\ \widehat{\theta},\ \mathbf{f}^\star_-$	Noise variance and vector, estimated nonlinear coefficients, and vector of pre-period affine fitted values
	Input formats and orientation risk
$A_{n,p,n_q}(C, \ell, Q)$	Tabular rule with n context rows, p features, labels ℓ, and n_q query rows
$F_{\text{time}},\ F_{\text{unit}},\ F_o$	Time-row and unit-row interfaces and the interface selected by o

x_h^+	Donor vector at horizon h, the hth row of X^+ written as a column
τ	Common ridge penalty in the orientation-equivalence result and experiments
$\widehat{\mathbf{y}}^{\text{time}}$, $\widehat{\mathbf{y}}^{\text{unit}}$	Counterfactual paths under the time-row and unit-row interfaces
gap_h, η^{time}, η^{unit}	Distance between the context-label ranges at horizon h and the respective outcome-scale extrapolation bounds
$\|u\|_H$, $\|u\|_H^2$	Horizon-normalized norm and squared norm, where $\|u\|_H = (H^{-1}\sum_{h=1}^{H} u_h^2)^{1/2}$
Two-way centering and additive shifts	
$\bar{y}_i$, $\bar{c}_t$, $\widetilde{Y}_{it}$	Unit baseline, common donor movement, and two-way-centered outcome
$\widetilde{\mathsf{P}}$, $\widehat{\widetilde{\mathbf{y}}}_1$, $\widehat{\widetilde{Y}}_{1t}$	Centered panel, predicted centered path, and its value at period t
$\widehat{Y}_{1t}(0)$, $\widehat{\mathbf{y}}_o^{\text{cent}}$	Restored counterfactual and centered-procedure path under interface o
χ_i, ψ_t, $\boldsymbol{\psi}_+$, $\bar{\psi}_-$	Additive unit and time shifts, post-period time-shift vector, and average pre-period time shift
$\mathbf{1}_H$	Length-H vector of ones
Evaluation and experiments	
$\mathcal{L}_{km}$, H_k	Mean squared prediction loss of method m on task k and that task's horizon
W_0, ξ_t	Planted affine donor weights and target noise in EX1
κ, $\kappa^\star$, SNR	Nonlinear strength, OLS break-even value $\kappa^\star = \sigma\sqrt{d_\phi/T_0}$, and signal-to-noise ratio $\text{SNR} = 0.2/\sigma^2$
$\widehat{\mathcal{E}}_{L,H}$	Empirical recovery error of procedure L over an H-period horizon
x_{hr}, y_{hr}	EX1 headroom-plot coordinates $x_{\text{hr}} = \kappa/\kappa^\star$ and $y_{\text{hr}} = \Delta_{\text{lin}}/\widehat{\mathcal{E}}_{L,H}$
ρ, d_g	Within-group donor correlation and EX1(b) design parameters
u_j, $r(\widetilde{v})$, $q_\rho(\widetilde{v})$	Gaussian-CDF donor transform, Friedman response, and standardized nonlinear component in EX1(b)
a_ρ, c_ρ, s_ρ	Centering, linear-projection, and scaling constants used to standardize q_ρ
$\mathbf{x}$, $\mathbf{b}$, z, $g(z)$, μ_z, s_z	EX2 feature vector, projection direction, latent index, nonlinear truth map, and its context center and scale
$d_{\text{out}}(\widehat{y}, \ell)$	Distance of prediction $\widehat{y}$ from the range of labels ℓ, normalized by that range's width
$\widehat{\eta}_{\max}$, $\widehat{\eta}_{0.99}$	Maximum and cellwise 99th-percentile normalized out-of-range envelopes
δ	Effect injected in EX1(c) and EX5, measured in pre-period-SD units
Method-selection recipe	
$\mathcal{M}$	Candidate method set
n_h	Number of non-overlapping rolling-backtest origins
$\mathcal{O}$	Selection objective

B THEORETICAL ANALYSES

B.1 SYNTHETIC CONTROL AS A SIMPLEX ATTENTION RULE

Queries, keys, and values. Attention aggregates *values* using weights determined from a *query* and associated *keys*. The query describes the item to be predicted, and each key describes an example whose value contributes to the prediction.

Using the notation of Section 3, SC's query is the target predictor vector $q = X_1$, its keys are the donor predictor vectors stored as columns of $K = X_0$, and its values are the donor outcomes $v_t = (Y_{2t}, \ldots, Y_{J+1,t})^\top$. We call $a(q, K; \Omega)^\top v_t$ a *simplex attention rule* when $a(q, K; \Omega) \in \mathcal{W}$: the weights are nonnegative and sum to one. The following lemma compares SC's weight construction with softmax attention; it does not identify SC with a transformer.

Lemma 1 (SC representation and comparison with softmax)**.** *Fix the predictor metric Ω and a deterministic selection $W^\star(\Omega)$ from the optimizer set in Equation 1. Then:*

(i) Classical SC has the representation

$$a(q, K; \Omega) = W^\star(\Omega), \qquad \widehat{Y}_{1t}(0) = a(q, K; \Omega)^\top v_t.$$

The weights are fixed across post-intervention periods.

(ii) Its weights solve

$$a(q, K; \Omega) \in \arg\max_{W \in \mathcal{W}} \left\{ W^\top K^\top \Omega q - \frac{1}{2} W^\top K^\top \Omega K W \right\}.$$

By comparison, for temperature $\lambda_{\text{att}} > 0$, softmax weights satisfy

$$\text{softmax}\left(K^\top \Omega q / \lambda_{\text{att}}\right) \in \arg\max_{W \in \mathcal{W}} \left\{ W^\top K^\top \Omega q + \lambda_{\text{att}} \, \text{Ent}(W) \right\},$$

where $\text{Ent}(W) = -\sum_{j \in \mathcal{C}} w_j \log w_j$, with $0 \log 0 := 0$.

(iii) If $K^\top \Omega K = I_J$, the SC optimizer is unique and

$$a(q, K; \Omega) = \text{sparsemax}(K^\top \Omega q),$$

where sparsemax is Euclidean projection onto $\mathcal{W}$ (Martins & Astudillo, 2016).

Proof. The first claim follows by substituting $q = X_1$ and $K = X_0$ into the SC optimization and prediction. For the second, expand

$$(q - KW)^\top \Omega (q - KW) = q^\top \Omega q - 2 W^\top K^\top \Omega q + W^\top K^\top \Omega K W.$$

Removing the constant and rescaling gives the SC maximization. The softmax characterization follows because its objective equals

$$\lambda_{\text{att}} \log \sum_{j \in \mathcal{C}} \exp\left((K^\top \Omega q)_j / \lambda_{\text{att}} \right) - \lambda_{\text{att}} \, \text{KL}\left(W \,\middle\|\, \text{softmax}(K^\top \Omega q / \lambda_{\text{att}}) \right).$$

This is uniquely maximized when the KL divergence is zero. Finally, if $K^\top \Omega K = I_J$, the maximization objective in (ii) equals

$$\frac{1}{2} \| K^\top \Omega q \|_2^2 - \frac{1}{2} \| W - K^\top \Omega q \|_2^2,$$

so its unique maximizer is the simplex projection. □

What the comparison establishes. SC and softmax share the query–key score $K^\top \Omega q$, but construct weights differently. Softmax adds an entropy bonus; SC subtracts a quadratic term involving the key–key Gram matrix $K^\top \Omega K$. The positive-semidefinite metric Ω weights predictor discrepancies and defines the weighted inner products above. It is neither a donor weight nor a softmax normalization.

Weighted aggregation and Gram-matrix dependence also occur in unconstrained OLS and ridge. SC differs from these methods through its simplex restriction, which implies

$$\min_{j\in\mathcal{C}} Y_{jt} \le \widehat{Y}_{1t}(0) \le \max_{j\in\mathcal{C}} Y_{jt}.$$

Thus SC stays within the unit-row context-label range, corresponding to $\eta^{\text{unit}} = 0$ in Proposition 2. Its fitted prediction remains linear in v_t. The attention representation does not itself supply nonlinear headroom.

Connection to the unit-row interface. The unit-row interface presents the data in the same roles as the SC representation above. Each donor is one labeled example: its pre-treatment history provides the features, and its outcome at the prediction period t provides the label. The target's pre-treatment history is the unlabeled query. Thus, when SC uses complete pre-treatment histories as predictors, $q = y^-$ and $K = X^-$. SC predicts by taking a convex combination of the donor labels. A tabular foundation model receives the same examples and query, but may produce a different rule using the same context.

The derivation treats Ω as fixed. If it is selected using additional pre-treatment validation data, the fitted weights also depend on those data. Once Ω is chosen, the representation above applies unchanged.

B.2 Population risk accounting

Fix a post-intervention period t and write

$$Y = Y_{1t}(0), \qquad v = v_t = \mathbf{Y}_{\mathcal{C},t}.$$

Let $\mathcal{I}$ contain all information admissible when predicting Y, including v, so that $\sigma(v) \subseteq \sigma(\mathcal{I})$. Expectations in this subsection are population expectations under the post-intervention law at period t, with the dependence on t suppressed, and Y and v are square-integrable. Define

$$m_{\mathcal{I}} = \mathbb{E}[Y \mid \mathcal{I}], \qquad m_v = \mathbb{E}[Y \mid v], \qquad R^\star = \mathbb{E}[(Y - m_{\mathcal{I}})^2].$$

Let $\mathcal{S} = \{v \mapsto w^\top v : w \in \mathcal{W}\}$ be the simplex class and $\mathcal{A} = \{v \mapsto \gamma_0 + \gamma^\top v : \gamma_0 \in \mathbb{R},\ \gamma \in \mathbb{R}^J\}$. the affine class, so that $\mathcal{S} \subseteq \mathcal{A} \subseteq L^2(\sigma(v))$, and define

$$I_{\text{panel}} = \mathbb{E}[(m_{\mathcal{I}} - m_v)^2], \qquad \Delta_{\text{lin}} = \inf_{f\in\mathcal{A}} \mathbb{E}[(m_v - f(v))^2],$$

$$\Delta_{\text{geo}} = \inf_{f\in\mathcal{S}} \mathbb{E}[(m_v - f(v))^2] - \Delta_{\text{lin}}.$$

Nested projection gives the full risk decomposition:

$$\inf_{f\in\mathcal{S}} \mathbb{E}[(Y - f(v))^2] = R^\star + I_{\text{panel}} + \Delta_{\text{lin}} + \Delta_{\text{geo}}. \tag{5}$$

Equation 5 shows why beating classical synthetic control does not establish a benefit from pretraining. A sufficiently expressive kernel or tree fitted only on the focal panel can also reduce Δ_{lin} without using anything learned before the study. So the fitted nonlinear frontier belongs in the primary comparison rather than in an ablation.

Remark 1 (When the affine gap can be positive). Let ζ_t be a latent state with $Y_{jt} = b_j^\top \zeta_t$ for each donor and $\mathbb{E}[Y_{1t}(0) \mid \zeta_t] = b_0^\top \zeta_t$. If $b_0 \in \text{span}\{b_j\}$, then m_v is linear in v and $\Delta_{\text{lin}} = 0$; if $b_0 \in \text{conv}\{b_j\}$, then also $\Delta_{\text{geo}} = 0$. Nonlinear state dynamics alone do not create nonlinear headroom when the donors span the relevant state directions. Within this model a positive gap requires donors that fail to span a relevant state direction. Nonlinear observation maps are a separate route outside it. Our simulations vary the donor–target map rather than the complexity of the paths.

B.3 Recoverable nonlinear headroom

Let $\mathcal{D}_- = \{(v_s, Y_{1s}) : s \le T_0\}$ be the labeled pre-intervention sample. By a learning rule L we mean the complete procedure, including its preprocessing, fixed checkpoint, and internal seed U_L, which is constant for a deterministic rule; it returns $\widehat{m}_L(v) = L(\mathcal{D}_-, v; U_L)$. Its finite-context recovery error and its signed recovered headroom are

$$\mathcal{E}_L = \mathbb{E}[\{\widehat{m}_L(v) - m_v\}^2], \qquad G_L = \Delta_{\text{lin}} - \mathcal{E}_L.$$

Assumption 1 (Stable donor regression)**.** All displayed second moments are finite and $\mathbb{E}[Y - m_v \mid \mathcal{D}_-, v, U_L] = 0$.

The condition says the evaluation innovation has mean zero once the sample, the query, and the seed are known. For a randomized rule it holds when the seed is conditionally independent of that innovation, and it holds in the fixed-map calibration below. It need not hold when one method receives predictive information the others do not: a native forecasting interface may use ordered target history or a future donor block that a period-row learner never sees. Such a comparison changes I_{panel} in equation 5 rather than G_L, and is not a comparison of nonlinear recovery.

Proposition 1 (Recoverable nonlinear headroom)**.** *Under the condition above, the learning rule's expected gain over the best population affine donor rule is*

$$\inf_{f \in \mathcal{A}} \mathbb{E}[(Y - f(v))^2] - \mathbb{E}[(Y - \widehat{m}_L(v))^2] = \Delta_{\text{lin}} - \mathcal{E}_L =: G_L.$$

Proof. Write $Y - \widehat{m}_L(v) = (Y - m_v) + (m_v - \widehat{m}_L(v))$. The cross term vanishes by Assumption 1, giving

$$\mathbb{E}[(Y - \widehat{m}_L(v))^2] = \mathbb{E}[(Y - m_v)^2] + \mathcal{E}_L.$$

For each fixed $f \in \mathcal{A}$ the corresponding cross term vanishes from $\mathbb{E}[Y - m_v \mid v] = 0$, which follows from the definition of m_v alone; minimizing over f gives the second. Subtracting gives the stated identity. □

Both risks contain the same irreducible term $\mathbb{E}[(Y - m_v)^2] = R^\star + I_{\text{panel}}$, which cancels, so Δ_{lin} is a ceiling and positive recovery requires $\mathcal{E}_L < \Delta_{\text{lin}}$. Proposition 1 is an accounting identity, not a claim that two rules have comparable recovery errors.

A finite-context calibration. The next result makes the price explicit in one deliberately narrow model. Let $\phi(v) \in \mathbb{R}^{d_\phi}$ be a fixed nonlinear dictionary satisfying

$$\mathbb{E}[\phi(v)\phi(v)^\top] = I_{d_\phi}, \qquad \mathbb{E}[\phi(v) f(v)] = 0 \quad \text{for every } f \in \mathcal{A},$$

so that the dictionary is normalized and carries no affine component. Let $f^\star$ be the affine projection of m_v. We represent the remaining nonlinear component as

$$m_v = f^\star(v) + \theta^\top \phi(v), \qquad \|\theta\|_2^2 = \Delta_{\text{lin}},$$

where the normalization of ϕ makes the squared magnitude of the nonlinear component equal to the affine approximation gap. Let $\Phi \in \mathbb{R}^{T_0 \times d_\phi}$ have rows $\phi(v_s)^\top$ and assume the exact orthogonal design $\Phi^\top \Phi = T_0 I_{d_\phi}$, which requires $T_0 \geq d_\phi$.

Let the sample outcomes be $Y_s = f^\star(v_s) + \theta^\top \phi(v_s) + \varepsilon_s$, with $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{T_0})^\top$ jointly independent of (Φ, v), mean zero, and covariance $\sigma^2 I_{T_0}$. Treat $f^\star$ and ϕ as known, which gives the method the affine rule and the feature list for free, so the calibration isolates only the cost of estimating the nonlinear coefficients. Define the residualized least-squares rule

$$\widehat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top (\mathbf{y}^- - \mathbf{f}_-^\star), \qquad \widehat{m}_{\text{OLS}}(v) = f^\star(v) + \widehat{\theta}^\top \phi(v).$$

Corollary 1 (Finite-context break-even point for OLS)**.** *In the residualized orthogonal-feature calibration, with d_ϕ nonlinear features and noise variance σ^2, Ordinary Least Squares (OLS) satisfies $\mathcal{E}_{\text{OLS}} = \sigma^2 d_\phi / T_0$ and $G_{\text{OLS}} = \Delta_{\text{lin}} - \sigma^2 d_\phi / T_0$. OLS improves on the population affine rule exactly when $\Delta_{\text{lin}} > \sigma^2 d_\phi / T_0$.*

Proof. Exact orthogonality and joint noise independence give $\widehat{\theta} - \theta = T_0^{-1} \Phi^\top \varepsilon$ with conditional covariance $(\sigma^2 / T_0) I_{d_\phi}$, so the conditional prediction error at the realized query is $(\sigma^2 / T_0) \|\phi(v)\|_2^2$. Taking expectations and using $\mathbb{E}[\|\phi(v)\|_2^2] = d_\phi$ gives $\mathcal{E}_{\text{OLS}}$. Subtracting it from Δ_{lin} gives G_{OLS}, and its sign gives the condition. □

Remark 2 (Reading the threshold). The threshold is exact only for the stated orthogonal-design calibration, with known affine component and nonlinear features, and compares against the population affine oracle. Increasing sampling frequency changes both context length and noise; under serial dependence, the displayed $\sigma^2 d_\phi / T_0$ formula need not apply.

Where prior information can enter. For fixed prediction inputs and a fixed post-intervention law, pretraining can change $\mathcal{E}_L$ but not Δ_{lin}. A frozen checkpoint approximates the posterior predictive distribution under its pretraining tasks (Müller et al., 2022), so its recovery error depends on how closely those tasks match the target panel. Whether it beats a rule fitted on the focal panel is an empirical question. Evidence from held-out tasks with observed targets carries over to the post-treatment period only if relative performance transports.

B.4 Input formats and transformation response

A frozen tabular checkpoint can receive periods or units as context rows. The two interfaces use the same observed outcomes but assign different roles to examples, features, labels, and queries.

Write $X^- = \mathbf{Y}_{\mathcal{C},\mathrm{pre}}$, $y^- = \mathbf{y}_{1,\mathrm{pre}}$ and $X^+ = \mathbf{Y}_{\mathcal{C},\mathrm{post}}$ for the three observed blocks of Section 3, with shapes $T_0 \times J$, $T_0 \times 1$ and $H \times J$, and collect them as the observed panel $\mathsf{P} = (X^-, y^-, X^+)$. A tabular learning rule is a family of maps $A_{n,p,n_q}(C, \ell, Q)$ taking n context rows C with p features, their labels ℓ, and n_q query rows Q, and returning n_q predictions; the dimensions are written out because the assignment changes them. Writing F_{time} and F_{unit} for the two assignments,

$$F_{\mathrm{time}}(\mathsf{P}) = A_{T_0,J,H}(X^-, y^-, X^+), \qquad F_{\mathrm{unit},h}(\mathsf{P}) = A_{J,T_0,1}\big((X^-)^\top, x_h^+, (y^-)^\top\big),$$

where x_h^+ is the hth row of X^+ as a column and $F_{\mathrm{unit}}(\mathsf{P})$ stacks the H separate fits.

The two interfaces estimate different functions. When T_0 is small relative to J, the time-row interface has fewer examples than features, while the unit-row interface has more examples than features. Horizontal regressions (unit-row) suit panels with many units and few periods, and vertical regressions (time-row) suit the reverse (Athey et al., 2021).

Ridge equivalence across orientations. For ridge with a common penalty $\tau \geq 0$, no intercept, and no orientation-specific column scaling, Shen et al. (2023, Theorem 1) gives

$$(x_h^+)^\top \big((X^-)^\top X^- + \tau I\big)^\dagger (X^-)^\top y^- = (y^-)^\top \big(X^-(X^-)^\top + \tau I\big)^\dagger X^- x_h^+.$$

The two sides are the time-row and unit-row predictions. Equality holds at every aspect ratio, including rank-deficient panels; $\tau = 0$ gives minimum-ℓ_2-norm OLS. Intercepts can break the equivalence (Shen et al., 2023, Corollary 2), while two-way centering preserves it when both interfaces use the same residualized panel and add-back.

Proposition 2 (Context-range disagreement floor)**.** *Let $\widehat{\mathbf{y}}^{\mathrm{time}}$ and $\widehat{\mathbf{y}}^{\mathrm{unit}}$ be the paths from a fixed checkpoint and preprocessing rule, and let gap_h be the distance between the two context-label ranges at horizon h. If the predictions lie within η^{time} and η^{unit} of their respective label ranges, for every h:*

$$\big|\widehat{y}_h^{\mathrm{time}} - \widehat{y}_h^{\mathrm{unit}}\big| \geq \big(\mathrm{gap}_h - \eta^{\mathrm{time}} - \eta^{\mathrm{unit}}\big)_+. \tag{3}$$

Proof. Fix a horizon h. Enlarge the time-row context-label interval by η^{time} at both endpoints, and enlarge the unit-row interval by η^{unit}. By assumption, the two predictions lie inside these enlarged intervals.

The original intervals are separated by gap_h, so the enlarged intervals are separated by

$$\big(\mathrm{gap}_h - \eta^{\mathrm{time}} - \eta^{\mathrm{unit}}\big)_+ .$$

Any two points lying in the respective intervals must be at least this far apart, which gives the pointwise inequality. Squaring these inequalities, averaging over h, and taking square roots gives the path-level floor. □

Remark 3 (No universal orientation selector). Suppose the two predictions differ on an observed panel. Completing the missing target path with the time-row prediction makes that orientation exact, and completing it with the unit-row prediction makes the other exact, while the observed panel stays the same. So no selector based only on observed inputs can be right for every completion. This rules out a universal selector, not one that works well on a restricted class of panels. Proposition 2 bounds the disagreement, not which orientation is more accurate.

B.5 TWO-WAY CENTERING ALGORITHM

Unit levels and the movement shared across donors at each date are large, well understood components of panel variation, and a frozen model should not have to rediscover them from a short context. Algorithm 1 removes both before prediction under either interface and restores them afterward. It never uses a target outcome after T_0, the masked origin in our benchmark and the intervention date in an application.

Algorithm 1 Two-way centered prediction under either interface

Require: Observed panel $\mathsf{P} = (X^-, y^-, X^+)$; frozen tabular rule A; interface map $F_o \in \{F_{\text{time}}, F_{\text{unit}}\}$

Ensure: Counterfactual path $\widehat{Y}_{1t}(0)$ for $t > T_0$

1: $\bar{y}_i \leftarrow T_0^{-1} \sum_{s \le T_0} Y_{is}$ for every unit ▷ pre-intervention level
2: $\bar{c}_t \leftarrow J^{-1} \sum_{j \in \mathcal{C}} (Y_{jt} - \bar{y}_j)$ for every period ▷ donor movement shared at t
3: $\widetilde{Y}_{it} \leftarrow Y_{it} - \bar{y}_i - \bar{c}_t$ for all donors at all periods, and for the treated unit at $t \le T_0$ ▷ no masked outcome is used
4: $\widetilde{\mathsf{P}} \leftarrow (\widetilde{X}^-, \widetilde{y}^-, \widetilde{X}^+)$ ▷ residualized panel
5: $\widehat{\widetilde{\mathbf{y}}}_1 \leftarrow F_o(\widetilde{\mathsf{P}})$ ▷ F_{unit} refits at each horizon
6: $\widehat{Y}_{1t}(0) \leftarrow \bar{y}_1 + \bar{c}_t + \widehat{\widetilde{Y}}_{1t}$ for $t > T_0$ ▷ restore level and shared movement

Proposition 3 (Additive equivariance of two-way centering)**.** *For $o \in \{\text{time}, \text{unit}\}$, let $\widehat{\mathbf{y}}_o^{\text{cent}}(\mathsf{P})$ denote the output of Algorithm 1 under a fixed base learner. For arbitrary unit shifts χ_i and common time shifts ψ_t, let P' replace every observed outcome by*

$$Y'_{it} = Y_{it} + \chi_i + \psi_t,$$

and write $\boldsymbol{\psi}_+ = (\psi_{T_0+1}, \ldots, \psi_{T_0+H})^\top$. Then

$$\widetilde{\mathsf{P}}' = \widetilde{\mathsf{P}}, \qquad \widehat{\mathbf{y}}_o^{\text{cent}}(\mathsf{P}') = \widehat{\mathbf{y}}_o^{\text{cent}}(\mathsf{P}) + \chi_1 \mathbf{1}_H + \boldsymbol{\psi}_+. \tag{6}$$

If the missing path is shifted likewise, $\mathbf{y}'_{1,\text{post}} = \mathbf{y}_{1,\text{post}} + \chi_1 \mathbf{1}_H + \boldsymbol{\psi}_+$, then

$$\left\|\widehat{\mathbf{y}}_o^{\text{cent}}(\mathsf{P}') - \mathbf{y}'_{1,\text{post}}\right\|_H = \left\|\widehat{\mathbf{y}}_o^{\text{cent}}(\mathsf{P}) - \mathbf{y}_{1,\text{post}}\right\|_H, \tag{7}$$

$$\left\|\widehat{\mathbf{y}}_{\text{time}}^{\text{cent}}(\mathsf{P}') - \widehat{\mathbf{y}}_{\text{unit}}^{\text{cent}}(\mathsf{P}')\right\|_H = \left\|\widehat{\mathbf{y}}_{\text{time}}^{\text{cent}}(\mathsf{P}) - \widehat{\mathbf{y}}_{\text{unit}}^{\text{cent}}(\mathsf{P})\right\|_H. \tag{8}$$

Proof. Let $\bar{\psi}_- = T_0^{-1} \sum_{s \le T_0} \psi_s$. The transformed means satisfy

$$\bar{y}'_i = \bar{y}_i + \chi_i + \bar{\psi}_-, \qquad \bar{c}'_t = \bar{c}_t + \psi_t - \bar{\psi}_-.$$

Hence

$$\widetilde{Y}'_{it} = Y'_{it} - \bar{y}'_i - \bar{c}'_t = Y_{it} - \bar{y}_i - \bar{c}_t = \widetilde{Y}_{it}.$$

The base learner therefore receives the same residualized panel. The add-back changes by

$$\bar{y}'_1 + \bar{c}'_t - (\bar{y}_1 + \bar{c}_t) = \chi_1 + \psi_t,$$

which proves Equation 6. The remaining claims follow by cancellation. □

Additive equivariance does not guarantee lower MSE or less orientation disagreement than uncentered prediction. The wrapper uses equal rather than estimated unit and time weights and does not inherit SDID's double-robustness or precision guarantees (Arkhangelsky et al., 2021).

C DATASETS

We use public aggregate panels in three roles. *Masked-origin benchmarks* hide the target's outcomes after an artificial cutoff, permitting direct evaluation of counterfactual prediction error. *Actual-intervention applications* reconstruct an untreated path after a real intervention; observed post-intervention outcomes are treated, not ground truth. *Structured applications* involve staggered adoption, grouped exposure, or within-unit comparisons and require an explicit estimand and comparison construction. Table 4 summarizes the sources; each entry below gives the *Data* and *Application design* (treatment or artificial origin, comparison set, window).

Table 4: Public real-world datasets before task-specific exclusions. $\max T_0$ and $\max J$ are the maximum pre-intervention length and donor-pool size.

Dataset	Unit	Freq.	Period	$\max T_0$	$\max J$	Primary outcome(s)
DOL ETA 539	Jurisdiction	Weekly	1994–2023	768	52	Initial claims
German	Country	Annual	1960–2003	30	16	GDP per capita
FARS	Pooled cohort/state	Monthly	1982–1993	63	38	Rural-Interstate fatalities per 100M miles
Proposition 99	State	Annual	1970–2000	19	38	Cigarette sales per capita
Basque	Region	Annual	1955–1997	15	16	Real GDP per capita
Ohio	State	Weekly	2021	17	50	Fully vaccinated; first-dose share
Vietnam	Country	Annual	2000–2024	13	8	FDI inflows; industrial value added
Nebraska	State	Annual	Historical	16	46	Log real expenditure per capita
RGGI	State	Annual	1960–2023	45	37	CO_2 emissions by fuel
Energy	Country	Annual	1960–1985	13	7	Gasoline tax; oil share of electricity
EUSED	Country–sector	Annual	1990–2016	15	25	Sectoral CO_2 emissions
Proportions	Municipality	Election	Varies			Vote shares and abstention

C.1 MASKED-ORIGIN BENCHMARK PANELS

DOL ETA 539 initial claims. *Data.* Weekly UI initial claims (field `c3`, aligned by `rptdate`) for 53 jurisdictions (50 states, DC, Puerto Rico, US Virgin Islands), January 1994–September 2023 (DOL ETA, 2023). *Application design.* Masked-origin reconstruction only.

C.2 ACTUAL-INTERVENTION APPLICATIONS

German reunification. *Data.* GDP per capita for West Germany and 16 OECD countries, 1960–2003 (748 country-years) (Abadie, 2021a). Donors: US, UK, Austria, Belgium, Denmark, France, Italy, Netherlands, Norway, Switzerland, Japan, Greece, Portugal, Spain, Australia, New Zealand. *Application design.* Following Abadie et al. (2015), West Germany is treated, with pre-period 1960–1989 and post-period 1990–2003 (onset 1990). The Wall opened on November 9, 1989; reunification occurred on October 3, 1990.

FARS 65-mph monthly panel. *Data.* Monthly rural-Interstate fatality counts (NHTSA FARS) combined with annual rural-Interstate VMT (FHWA VM-2) (NHTSA, 2023), for 21 adopter and 7 control states, January 1982–December 1993 (4,032 state-months; no person-level records). Outcome: fatalities per 100M miles, using each state's fixed 1982–1986 mean annual VMT$/12$. *Application design.* Adopters are pooled into one `pooled_adopters` series. Controls (CT, MD, MA, NJ, NY, PA, RI) remain separate. The origin is April 1987, with pre-period January 1982–March 1987 and horizons of 12 and 24 months. State adoption dates give the day-weighted share of baseline VMT exposed (United States Congress, 1987). We supplement the AG data with 10 additional states from the same source.

California Proposition 99. *Data.* Annual per-capita cigarette sales for California and 38 states, 1970–2000 (1,209 state-years) (Abadie et al., 2010; Dunford, 2025). *Application design.* California is treated; the other 38 states are donors. Proposition 99 was approved on November 8, 1988, and took effect on January 1, 1989, giving pre-period 1970–1988 and post-period 1989–2000.

Basque Country GDP. *Data.* Annual real GDP per capita and 13 predictors for 17 Spanish regions, 1955–1997 (Abadie & Gardeazabal, 2003; Hainmueller & Diamond). *Application design.* The Basque Country is focal and the other 16 regions are donors (the Spain average is excluded). Treatment is the onset of sustained ETA violence, with pre-period 1955–1969, post-period 1970–1997, and onset sensitivities at 1968 and 1970.

Ohio Vax-a-Million. *Data.* Weekly vaccination rates aggregated from a daily state panel, January 18–September 12, 2021 (Lang et al., 2023; 2021). Primary outcome: fully vaccinated share; secondary: first-dose share. *Application design.* Ohio is treated: announcement May 12, first drawing May 26, final drawing June 23, 2021. The design uses 17 pre-weeks, a primary post-period ending at the final drawing, and an extended window to September 12. The preregistered pool (49 states + DC) was amended to exclude 18 copycat-lottery jurisdictions, leaving 32 donors. The amended pool is primary and the original pool is a sensitivity.

Vietnam's trade-war dividend. *Data.* Net FDI inflows (current USD bn) and industrial value added (constant-2015 USD bn) for nine countries, 2000–2024 (225 obs; 180 retained for 2005–2024) (Taabazuing & Fatoki, 2026). *Application design.* Vietnam is focal; donors are Indonesia, Philippines, Thailand, India, Colombia, Peru, South Africa, and Brazil. Treatment year 2018 gives pre-period 2005–2017 and post-period 2018–2024. Given the August 2017 Section 301 investigation, 2017 is an anticipation sensitivity. Placebos use all eight donors.

Nebraska unicameral legislature. *Data.* Log real state-government expenditure per capita (1947 dollars) for Nebraska and 46 states (Hankins, 2020; 2019). *Application design.* Nebraska is treated. The amendment was approved in November 1934 and the first unicameral session began in January 1937 (treatment year 1937). The fitting window is 1915–1932 and the post-period is 1937–1947. Data from 1920-21 are missing.

C.3 Grouped, staggered, and within-unit applications

Regional Greenhouse Gas Initiative. *Data.* Balanced 49-state annual panel, 2005–2023 (931 state-years), with total, coal, natural-gas, and petroleum CO_2 emissions (Paul & Dixit, 2026). *Application design.* The 2009 cohort is CT, DE, ME, MD, MA, NH, NJ, NY, RI, and VT; VA and PA enter later. The memorandum was signed on December 20, 2005, the first auction took place in September 2008, and compliance began on January 1, 2009. Confirmatory use requires freezing the cohort or transition, the reversal rule, eligible controls, and border or market exclusions.

Institutional sources of energy transitions. *Data.* Two 18-country annual panels (Finnegan et al., 2024): demand-side gasoline tax (USD/liter), 1965–1985 (378 obs), and supply-side oil share of electricity, 1960–1985 (468 obs). *Application design.* The common shock is the 1973 oil crisis (indicator on from 1973). The demand side compares 13 institution-treated countries with 5; the supply side compares 11 with 7. The two panels are reported separately.

European Union sectoral emissions. *Data.* Annual CO_2 emissions for 33 countries and seven matched sectors, 1990–2016 (Bayer, 2019; Bayer & Aklin, 2020). *Application design.* Treatment is the EU ETS: Directive 2003/87/EC was adopted in October 2003 and Phase I began on January 1, 2005. Pre-period 1990–2004; post-period 2005–2016.

Treatment effects on proportions. *Data.* Two municipal-election applications with compositional outcomes (Bogatyrev & Stoetzer, 2026a;b). *Spain:* party vote shares in Asturias, Le'on, and Teruel, 2008–2019; coal-mining municipalities covered by the Just Transition Agreement (reached October 2018, enacted January 2019) are compared with non-mining municipalities in the same provinces at the April 2019 election. *Poland:* government share, opposition share, and abstention; municipalities adopting anti-LGBTQ resolutions (staggered from 2019) are compared with non- or not-yet-adopters. Common synthetic weights preserve compositional structure.

D EXPERIMENT DETAILS

This appendix supplies the information needed to assess and reproduce the experiments in Section 4. It records design choices and additional results omitted from the main text for space.

D.1 METHODS AND TUNING

Table 5 lists the methods used in the simulation and real-panel experiments.

Table 5: Methods used in the experiments.

Method	Method family	Exact implementation
Simplex SC	Convex synthetic control	SLSQP simplex fit
Synth	Convex synthetic control	`Synth::synth` 1.1-10
Ridge	Regularized linear regression	Standardized ridge regression
SDID	Synthetic control with difference-in-differences	`synthdid::synthdid_estimate` 0.0.9
KRR	Kernel regression	RBF kernel ridge
CatBoost	Boosted trees	CatBoost 1.2.10
FLAML	Automated model selection	FLAML 2.6.0 with LightGBM 4.6.0 and Extra Trees
TabPFN	Tabular FM	TabPFN v2.6 regressor; `tabpfn` 8.0.7
TabDPT	Tabular FM	TabDPT 1.2.0
TabFM	Tabular FM	TabFM v1.0.0 PyTorch regression checkpoint; `tabfm` 1.0.1
Chronos-2	Time-series FM	Standard checkpoint; `chronos-forecasting` 2.3.0
TimesFM-2.5	Time-series FM	200M PyTorch checkpoint; `timesfm` 2.0.2

Chronos-2 uses the target history with past and future donor outcomes as covariates. TimesFM-2.5 is evaluated target-only and with XReg.

Table 6 reports the search spaces for the fitted baselines.

Table 6: Tuning grids for the fitted baselines.

Method	Search space	Fixed settings
Ridge	$\tau \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$	Standardized predictors and target
KRR	RBF scale multiplier in $\{0.25, 1, 4\}$, applied to the median-distance scale; $\tau \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$	15 candidates
CatBoost (4)	depth in $\{3, 6\}$; `l2_leaf_reg` in $\{1, 10\}$; `learning_rate`=0.05	4 settings; 300 trees
CatBoost (32)	Five fixed anchors and 27 frozen draws: depth from $\{2, 3, 4, 6, 8, 10\}$; `l2_leaf_reg` log-uniform on $[0.1, 100]$; `learning_rate` log-uniform on $[0.02, 0.1]$	32 candidates, tree count up to 1000
FLAML	LightGBM and Extra Trees	30-second search cap; at most 32 iterations

For task-specific tuning, we use at most five expanding-window origins and a validation block with the same length as the prediction horizon. We pool squared errors over these origins, select the lowest-loss setting, and refit it using the full pre-treatment period.

D.2 NONLINEAR OPPORTUNITY AND RECOVERY COST

EX1: Nonlinear Headroom Design. We use the planted-target DGP in Section 4.1. Donors are independent stationary AR(1) processes with coefficient 0.8 and marginal variance one. The first five entries of W_0 equal 0.2, the remaining entries are zero, and $\xi_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.2/\mathrm{SNR}$. The standardized quadratic dictionary is orthogonal to the affine donor span, giving $\Delta_{\mathrm{lin}} = \kappa^2$.

The complete grid crosses

$$\begin{aligned} T_0 &\in \{32, 128, 512\}, & J &\in \{20, 64\}, \\ \mathrm{SNR} &\in \{2, 5, 20\}, & d_\phi &\in \{1, 4, 16\}, \\ \kappa &\in \{0, 0.01, 0.03, 0.1, 0.3, 1\}. \end{aligned}$$

The d_ϕ copies of the $\kappa = 0$ null are deduplicated. This leaves 144 cells per donor count and method: 135 positive-curvature cells and nine null cells. Each cell uses 30 paired seeds and a 24-period prediction horizon. Within a seed, the shorter contexts are nested in the $T_0 = 512$ path.

This experiment uses only the time-row interface and no two-way centering. TabDPT, TabPFN, TabFM, Chronos-2, and TimesFM target-only are evaluated with frozen checkpoints. CatBoost and RBF kernel ridge form the fitted nonlinear frontier, while ridge and Simplex SC are affine references. For a complete procedure L, recovery error is evaluated against the noiseless conditional mean,

$$\widehat{\mathcal{E}}_{L,H} = \frac{1}{H} \sum_{h=1}^{H} \left\{ \widehat{Y}^{N}_{1,T_0+h} - m(v_{T_0+h}) \right\}^2 .$$

This is the finite-test-block analogue of the population recovery error $\mathcal{E}_L$ in Proposition 1. The plotted coordinates are

$$x_{\mathrm{hr}} = \frac{\kappa}{\kappa^\star}, \qquad \kappa^\star = \sqrt{\frac{\sigma^2 d_\phi}{T_0}}, \qquad y_{\mathrm{hr}} = \frac{\Delta_{\mathrm{lin}}}{\widehat{\mathcal{E}}_{L,H}} = \frac{\kappa^2}{\widehat{\mathcal{E}}_{L,H}}.$$

Positive recovered headroom is $y_{\mathrm{hr}} > 1$. Curves aggregate cells in log-spaced x_{hr}-bins; lines are medians and bands are interquartile ranges.

The $J = 64$ nuisance-width row. The $J = 64$ design keeps the affine support, nonlinear dictionary, noise levels, and curvature grid fixed. Relative to $J = 20$, it adds 44 donor coordinates carrying no target signal. Its aspect ratios are $T_0/J \in \{0.5, 2, 8\}$. Figure 7 therefore tests whether the normalized opportunity remains recoverable after the donor dimension increases without adding signal.

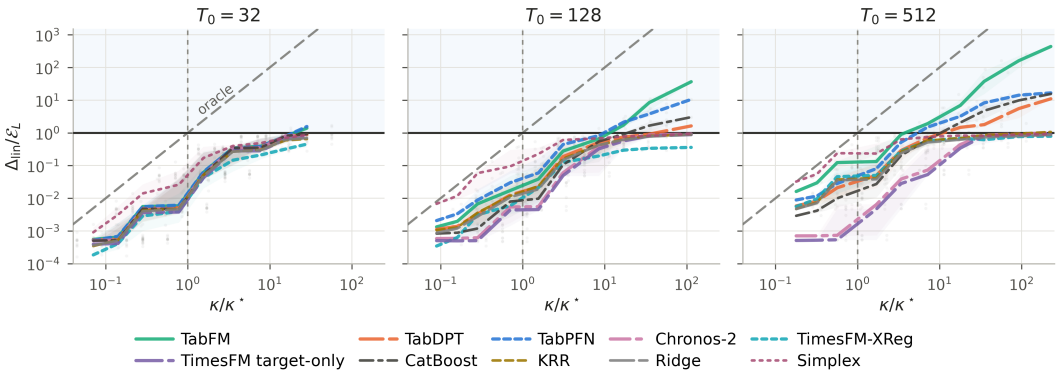


Figure 7: **Recoverable nonlinear headroom with $J = 64$.** Opportunity $\Delta_{\mathrm{lin}} = \kappa^2$ divided by recovery error, against curvature normalized by $\kappa^\star = \sqrt{\sigma^2 d_\phi / T_0}$. Panels vary T_0. The horizontal line marks break-even and the dashed diagonal is the true-dictionary OLS reference. Lines and bands are medians and interquartile ranges across cells; dots show the underlying cell summaries. Thirty paired seeds are used per cell.

EX1(b) design. We reuse the EX1(a) protocol and replace only the nonlinear component. Donor groups of five are equicorrelated at ρ within a group and independent across groups, with each coordinate following a stationary AR(1) process with coefficient 0.8 and unit marginal variance. $J = 20$ gives four groups and $J = 64$ gives twelve groups of five plus one group of four that never enters the planted map. Methods observe $v_t = \tilde{v}_t/\sqrt{1+4\rho}$, which holds the variance of the five-donor average at 0.2 for every ρ. Writing u_j for the standard normal CDF at $\tilde{v}_j$, the component is $q_\rho(\tilde{v}) = \{r(\tilde{v}) - a_\rho - c_\rho(\tilde{v}_1 + \tilde{v}_2)\}/s_\rho$ with $r(\tilde{v}) = 10\sin(\pi u_1 u_2) + 20(u_3 - \frac{1}{2})^2$; it has mean zero, unit variance, and zero covariance with every donor coordinate, giving $\Delta_{\text{lin}} = \kappa^2$ for every ρ and d_g. The primary grid crosses

$$T_0 \in \{32, 128, 512\}, \quad J \in \{20, 64\}, \quad \rho \in \{0, 0.6\},$$
$$d_g \in \{1, 4\}, \quad \kappa \in \{0, 0.01, 0.03, 0.1, 0.3, 1\}, \quad \text{SNR} = 2,$$

fixing SNR because EX1 already sweeps $\{2, 5, 20\}$.

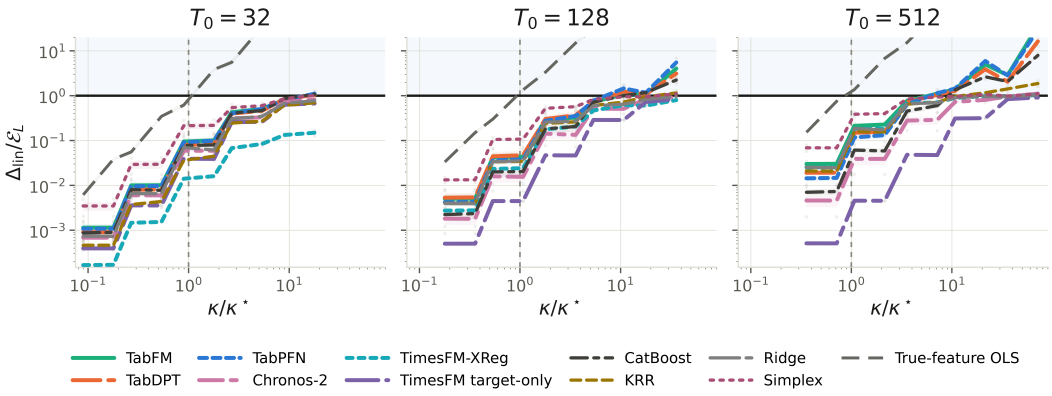


Figure 8: EX1(b) results: recoverable nonlinear headroom with $J = 20$ and a Friedman DGP.

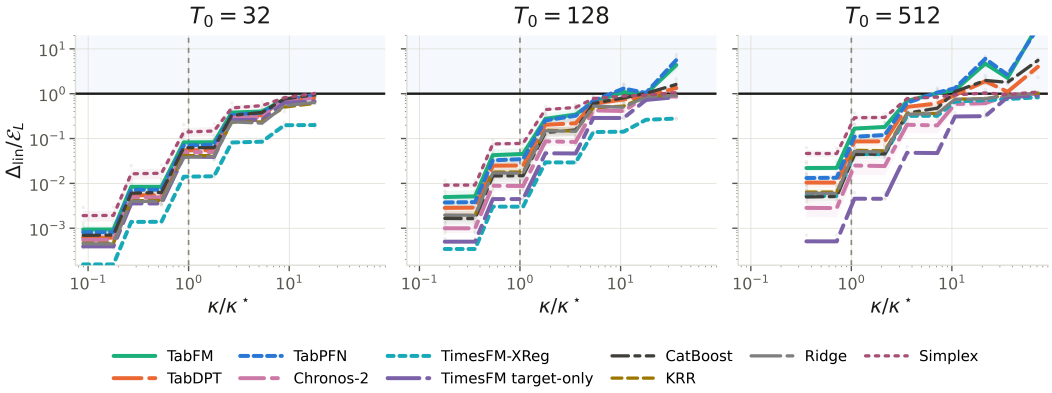


Figure 9: EX1(b) results: recoverable nonlinear headroom with $J = 64$ and a Friedman DGP.

EX1(c): Nonlinear Headroom and Experimental Power.

Table 7: EX1(c): A/A validity test at horizon=24. Point estimates use 200 held-out null panels; nominal level is 5%; bias and RMSE are in pre-treatment-SD units.

Regime	Method	A/A FPR	A/A bias	A/A RMSE
$\kappa = 0$	TabFM	0.035	0.018	0.197
$\kappa = 0$	TabDPT	0.020	-0.006	0.190
$\kappa = 0$	TabPFN	0.050	0.008	0.153
$\kappa = 0$	Chronos-2	0.015	-0.011	0.402
$\kappa = 0$	TimesFM-XReg	0.070	0.022	0.166
$\kappa = 0$	SDID	0.055	0.008	0.278
$\kappa = 0$	CatBoost	0.020	-0.009	0.299
$\kappa = 0$	KRR	0.055	0.017	0.186
$\kappa = 0$	Ridge	0.055	0.018	0.167
$\kappa = 0$	Simplex	0.085	0.005	0.112
$\kappa = 0.3$	TabFM	0.045	0.083	0.255
$\kappa = 0.3$	TabDPT	0.050	0.046	0.227
$\kappa = 0.3$	TabPFN	0.025	0.086	0.206
$\kappa = 0.3$	Chronos-2	0.035	0.039	0.428
$\kappa = 0.3$	TimesFM-XReg	0.045	0.128	0.394
$\kappa = 0.3$	SDID	0.060	0.023	0.363
$\kappa = 0.3$	CatBoost	0.025	0.080	0.277
$\kappa = 0.3$	KRR	0.025	0.128	0.343
$\kappa = 0.3$	Ridge	0.035	0.077	0.384
$\kappa = 0.3$	Simplex	0.065	-0.011	0.307
$\kappa = 1$	TabFM	0.050	0.078	0.186
$\kappa = 1$	TabDPT	0.055	0.072	0.231
$\kappa = 1$	TabPFN	0.055	0.089	0.194
$\kappa = 1$	Chronos-2	0.050	0.080	0.483
$\kappa = 1$	TimesFM-XReg	0.040	0.228	0.658
$\kappa = 1$	SDID	0.070	0.030	0.499
$\kappa = 1$	CatBoost	0.055	0.136	0.320
$\kappa = 1$	KRR	0.050	0.232	0.454
$\kappa = 1$	Ridge	0.050	0.052	0.496
$\kappa = 1$	Simplex	0.045	-0.028	0.467

D.3 Context-label range and orientation risk

Motivation. The two orientations use the same panel but expose the model to different context-label ranges: treated pre-intervention outcomes for time-row and donor outcomes at the query period for unit-row. The floor in Proposition 2 depends on procedure-specific bounds η on how far predictions can lie outside these ranges. EX2(a) calibrates these bounds in simulation, and EX2(b) measures the same normalized out-of-range distance on real panels.

EX2(a): Context-Label-Range (simulation). EX2a measures how far each procedure predicts outside its context-label range on synthetic tabular tasks with n labeled rows and p features. Features are either independent standard Gaussian draws or trending AR(1) columns with coefficient 0.7 and slopes drawn from $\mathcal{N}(0, 0.5^2)$. For $b \sim \mathcal{N}(0, I_p)/\sqrt{p}$, labels equal $g(x^\top b)$ plus Gaussian noise with standard deviation $0.3\,\mathrm{sd}(x^\top b)$. The truth map is linear or

$$g(z) = \mu_z + s_z \tanh\left(\frac{z - \mu_z}{s_z}\right), \qquad s_z = 4\,\mathrm{sd}(z),$$

where μ_z and s_z are computed from the context.

The *along* probe moves the latent index beyond the upper or lower context edge by 0.5, 1, 2, 4, 8 latent-index standard deviations. The *orthogonal* probe moves the query by 1, 4, 16 feature standard deviations in a direction orthogonal to b, leaving the noiseless label near the context center.

The grid crosses

$$n \in \{8, 32, 128, 512\}, \qquad p \in \{8, 32, 128\},$$

the three feature–truth combinations described above, both orientations, and 30 replicates. Each design is packed into both time-row and unit-row layouts, then evaluated raw and with two-way centering. The learner roster is CatBoost, ridge with a fixed penalty of one, TabDPT, TabPFN v2.6, and TabFM. The completed experiment contains 2,160 configurations and 280,800 prediction rows.

Table 8: **Empirical extrapolation-bound calibration.** $\widehat{\eta}_{\max}$ is the maximum distance outside the context-label range, normalized by its width; $\widehat{\eta}_{0.99}$ is the corresponding 99th-percentile envelope. "Out-of-range cells" counts the along-probe cells with positive mean distance at the most extreme probe, out of 72.

Learner	Orientation	Input	$\widehat{\eta}_{\max}$	$\widehat{\eta}_{0.99}$	Out-of-range cells
CatBoost	time	raw	0.193	0.183	20/72
CatBoost	time	centered	0.194	0.176	20/72
CatBoost	unit	raw	0.258	0.189	24/72
CatBoost	unit	centered	0.177	0.155	21/72
TabDPT	time	raw	0.340	0.334	19/72
TabDPT	time	centered	0.340	0.287	15/72
TabDPT	unit	raw	0.344	0.306	23/72
TabDPT	unit	centered	0.326	0.284	16/72
TabPFN v2.6	time	raw	0.857	0.771	44/72
TabPFN v2.6	time	centered	0.778	0.727	38/72
TabPFN v2.6	unit	raw	0.918	0.756	50/72
TabPFN v2.6	unit	centered	0.807	0.734	47/72
TabFM	time	raw	1.439	1.311	57/72
TabFM	time	centered	1.696	1.267	55/72
TabFM	unit	raw	1.666	1.330	57/72
TabFM	unit	centered	1.324	1.242	59/72
Ridge, $\tau = 1$	time	raw	2.100	1.554	64/72
Ridge, $\tau = 1$	time	centered	2.504	1.542	64/72
Ridge, $\tau = 1$	unit	raw	1.708	1.435	58/72
Ridge, $\tau = 1$	unit	centered	1.843	1.536	58/72

Extrapolation bounds. For prediction $\widehat{y}$ and context labels ℓ, define the normalized out-of-range distance

$$d_{\text{out}}(\widehat{y}; \ell) = \frac{\text{dist}\left(\widehat{y}, [\min(\ell), \max(\ell)]\right)}{\max(\ell) - \min(\ell)}.$$

$\widehat{\eta}_{\max}$ is the maximum value of d_{out} over the declared along-probe grid, and $\widehat{\eta}_{0.99}$ is the largest cellwise 99th-percentile value. These are dimensionless empirical bounds over the calibration grid, not universal checkpoint constants. When applying Proposition 2, each bound is multiplied by the width of its corresponding context-label range to return it to outcome units.

CatBoost has the smallest and ridge the largest normalized out-of-range distances, with the frozen checkpoints between them. For TabDPT, this distance does not increase monotonically with the imposed displacement in most cells where it is nonzero. Its $\widehat{\eta}_{\max}$ should therefore be read as an empirical bound over the tested grid rather than a smooth extrapolation limit.

The orthogonal control moves the query while leaving its noiseless label near the center of the context-label range. The displaced query lies outside the context's observed range along that direction in 9.2%, 97.8%, and 100% of cases at displacements of 1, 4, and 16 feature standard deviations, respectively. These orthogonal probes do not enter $\widehat{\eta}$.

EX2b: Context-Label-Range (real-panel). EX2b computes $d_{\text{out}}(\widehat{y}, \ell)$ on the same 673 masked-origin tasks used for the real-panel benchmark. Context labels are rebuilt separately for every checkpoint, orientation, centering choice, task, and horizon. For each dataset and target, we take the largest normalized out-of-range distance over all horizons and repeated origins. Figure 5(A) shows the percentage above each threshold.

We also compute the gap between the two context-label intervals at each horizon and evaluate the disagreement floor over a grid of η values. We do not apply one bound from EX2(a) to every real

panel. The normalized out-of-range distance is used to describe extrapolation and does not use hold-out target information.

EX4: Real-Benchmark.

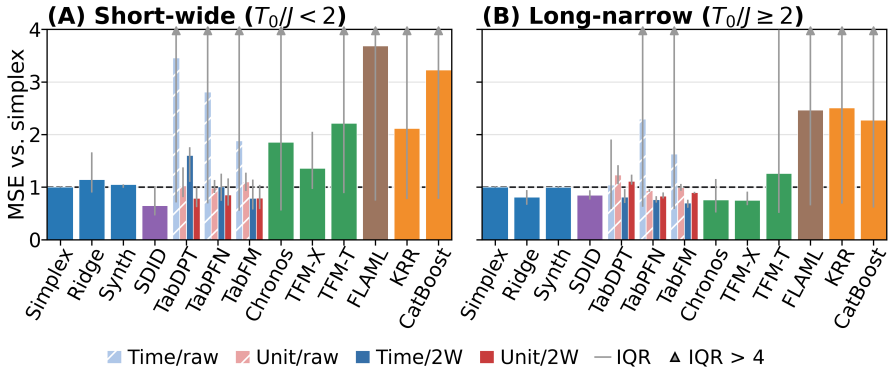


Figure 10: Equal-dataset MSE ratios to Simplex SC (bars: geometric mean; whiskers: dataset IQR) for short-wide ($T_0/J < 2$; left) and long-narrow ($T_0/J \geq 2$; right) panels. Tabular FMs are evaluated by the complete orientation–centering grid.

D.4 EX5: FARS SEMI-SYNTHETIC POWER DESIGN

In EX5, we preserve the pooled-state design of Ashenfelter & Greenstone (2004), but use monthly data to measure detection power at finer horizons. We pool the 10 states in their sample that adopted in April 1987 into one VMT-weighted target and evaluate detection after 3, 6, 9, and 12 months. Their seven comparison states (donors) cannot support a 5% unit-placebo test: with eight possible ranks, the minimum p-value is $1/8$. We therefore use all 28 states that had not adopted by April 1987.

We use the 63 months from January 1982 through March 1987, so the actual intervention period never enters the experiment. A 27-month fit window and a maximum 12-month test horizon leave 25 rolling origins, from April 1984 through April 1986. At each origin, the procedure receives the preceding 27 months of target and donor outcomes and the donor outcomes through the test horizon; the future target is withheld. Thus $T_0 = 27$, $J = 28$, and $T_0/J = 0.96$.

We follow the max-over-horizons construction in Westfall & Young (1993) and execute the simulation in three steps. In all three steps, each placebo assignment randomly partitions the same 38-state universe into a 10-state pseudo-treated cohort and 28 donors. These Monte Carlo draws approximate the randomization distribution while preserving the 10-versus-28 geometry of the fixed application.

First, 500 placebo assignments estimate the null center and scale for each procedure and horizon, placing the horizon-specific statistics on a common scale. Second, a separate 500 assignments construct the null distribution of the largest standardized statistic across the four horizons. Its finite-sample 5% cutoff becomes the common rejection boundary. Third, we freeze this boundary and apply it to 1,000 held-out A/A assignments. The A/A false-positive rate is the fraction of assignments that cross the boundary at least once. Only after these steps do we apply the test to the fixed AG-10 cohort, with and without injected effects.

For the power simulation, we use the fixed 10 adopted states from Ashenfelter & Greenstone (2004) as a single target. For power, we add a constant effect δ, with $\delta \in \{0.5, 0.6, 0.7, 0.8, 1.0\}$, to each post-origin target outcome. Because these outcomes are withheld, predictions remain fixed and the test statistic increases by exactly $\sqrt{h}\delta$. Power through month h is the fraction of the 25 origins detected at that or an earlier horizon. Power is averaged over 25 rolling pre-treatment origins. Because their windows overlap, these are repeated evaluation dates rather than independent trials. The numbers of nonoverlapping windows are 12, 6, 4, and 3 at the four horizons.

Table 9 reports the full results of all studied procedures.

Table 9: EX5 power for the fixed AG-10 FARS cohort. A/A FPR and null-effect RMSE use 1,000 held-out placebo assignments. T/U denote time-/unit-row input, 2W denotes two-way normalization, and bold marks the first month among the four horizons at which power reaches 80%.

			A/A test		$\delta = 0.5$				$\delta = 0.7$				$\delta = 0.8$				$\delta = 1.0$			
Method	Orient.	Prep.	FPR	$RMSE_{12}$	3	6	9	12	3	6	9	12	3	6	9	12	3	6	9	12
TabPFN v2.6	T	raw	.026	.294	.00	.00	.04	.20	.00	.00	.48	**.84**	.04	.08	.72	**.92**	.08	.44	**.88**	1.00
TabPFN v2.6	T	2W	.040	.285	.00	.00	.00	.60	.00	.08	.60	**.88**	.08	.12	.68	**.96**	.16	.40	**.84**	1.00
TabPFN v2.6	U	raw	.035	.352	.08	.56	.60	.68	.32	.72	**.80**	.92	.40	.76	**.84**	1.00	.60	**.88**	.96	1.00
TabPFN v2.6	U	2W	.040	.290	.08	.16	.40	.56	.24	.44	.68	**.84**	.24	.48	.76	**.84**	.28	.72	**.88**	1.00
TabDPT v1.2	T	raw	.036	.278	.00	.00	.12	.24	.04	.12	.52	**.84**	.08	.24	.68	**.92**	.16	.40	**.92**	1.00
TabDPT v1.2	T	2W	.037	.261	.00	.00	.28	.56	.08	.16	.40	.76	.16	.28	.52	**.96**	.20	.36	.76	**1.00**
TabDPT v1.2	U	raw	.041	.340	.16	.44	.60	.68	.32	.68	**.80**	.84	.44	.72	**.80**	.96	.60	**.88**	.96	1.00
TabDPT v1.2	U	2W	.029	.287	.08	.16	.16	.24	.20	.36	.56	**.84**	.24	.44	**.80**	.92	.36	.68	**.92**	1.00
TabFM v1.0	T	raw	.044	.288	.00	.00	.04	.24	.00	.08	.48	**.84**	.04	.16	.72	**.96**	.12	.40	**.88**	1.00
TabFM v1.0	T	2W	.034	.283	.00	.00	.08	.32	.04	.08	.40	.76	.12	.20	.56	**.84**	.20	.32	.76	**1.00**
TabFM v1.0	U	raw	.039	.378	.16	.28	.44	.52	.24	.56	.64	.68	.28	.60	.68	**.80**	.40	**.84**	.96	1.00
TabFM v1.0	U	2W	.022	.298	.04	.04	.04	.08	.12	.32	.48	.64	.16	.44	.60	**.80**	.24	.52	**.84**	1.00
Chronos-2	series	raw	.040	.280	.00	.00	.16	.44	.04	.24	.56	.68	.08	.40	.60	**.80**	.24	.60	**.92**	1.00
TimesFM v2.5	target	raw	.042	.359	.04	.04	.16	.24	.08	.08	.40	.48	.08	.24	.44	.64	.16	.40	.64	**.80**
TimesFM v2.5	XReg	raw	.032	1.206	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
SDID (R)	native	raw	.036	.347	.04	.08	.12	.32	.16	.32	.56	**.88**	.24	.40	**.80**	.96	.48	**.84**	1.00	1.00
Ridge SC ($\tau = 1$)	native	raw	.044	.294	.00	.00	.00	.16	.00	.00	.52	**.80**	.08	.24	.60	**.88**	.12	.44	**.84**	1.00
Simplex SC	native	raw	.029	.358	.04	.08	.28	.44	.08	.32	.56	.68	.16	.40	.72	**.80**	.28	**.84**	.88	1.00
Simplex SC + offset	native	raw	.027	.346	.04	.08	.20	.36	.08	.20	.48	.60	.16	.36	.68	.76	.24	.76	**.88**	.96
KRR (63)	native	raw	.046	.313	.00	.00	.00	.24	.04	.04	.52	**.80**	.08	.08	.68	**.92**	.12	.40	**.88**	1.00
KRR (15)	native	raw	.045	.316	.00	.00	.04	.32	.08	.08	.52	**.80**	.12	.16	.72	**.88**	.12	.40	**.84**	1.00
RFF ridge	native	raw	.040	.315	.00	.00	.00	.12	.00	.00	.36	.56	.00	.00	.56	**.80**	.12	.28	**.84**	.96
CatBoost (32)	native	raw	.041	.302	.00	.00	.12	.44	.00	.04	.52	.72	.04	.16	.72	**.92**	.20	.40	**.84**	1.00
CatBoost (4)	native	raw	.034	.308	.00	.04	.16	.36	.00	.16	.52	.76	.08	.28	.56	**.80**	.20	.40	**.84**	1.00

To assess robustness of the power simulation results, we also compute power on the 1,000 held-out placebo assignments, which differ from the fixed 10-state design of Ashenfelter & Greenstone (2004). Each assignment pools 10 randomly selected states from the same 38-state universe into a VMT-weighted pseudo-target and uses the remaining 28 as donors. We allocate 40 random 10-state pseudo-treated compositions to each of the 25 origins, totaling 1,000 held-out assignments. Table 10 shows the results.

The randomized any-10-state check does not show a universal time-to-detection lead, but leading FM procedures retain higher power: at $\delta = 0.7$ and month 12, the three centered time-row tabular FMs average (0.65) power, about 6 percentage points above the best baseline, CatBoost(32) at (0.59). Two-way centering improves 12-month power over raw time-row inputs for all three tabular FMs by 4–13 points, averaging 8 points. A timing advantage remains for stronger effects: centered TabDPT alone reaches 80% power at $\delta = 0.8$ by month 12, while at $\delta = 1.0$, Chronos-2 and centered unit-row TabDPT and TabPFN reach it by month 9, before any baseline.

E Method Selection Algorithm

We provide an algorithm to select the method $L \in \mathcal{M}$, where $\mathcal{M}$ is the set of available model×interface×preprocessing combinations. Each method uses a supported interface, which may be time-row, unit-row, or model-native, and optional two-way centering when supported. We evaluate each method by rolling back n_h non-overlapping H-period origins (default $n_h = 3$), fitting through each origin, and scoring the subsequent H periods under objective $\mathcal{O}$. The method that performs best at the origin closest to treatment is selected if it remains within relative tolerance (default 10%) of the best method at every origin. Otherwise, we evaluate all methods on cross-sectional donor placebos at the closest origin and select the method with the best aggregate placebo performance if it remains within ϵ of the best on a majority of placebos; if neither test is stable, we fall back to classical SC or SDID, which are generally more reliable when context is limited.

Table 10: EX5 power for random 10-of-38 pseudo-treated cohorts. A/A FPR and null-effect RMSE use 1,000 held-out assignments. T/U denote time-/unit-row input, 2W denotes two-way normalization, and bold marks the first month among the four horizons at which power reaches 80%.

			A/A test		$\delta = 0.5$				$\delta = 0.7$				$\delta = 0.8$				$\delta = 1.0$			
Method	Orient.	Prep.	FPR	$RMSE_{12}$	3	6	9	12	3	6	9	12	3	6	9	12	3	6	9	12
TabPFN v2.6	T	raw	.026	.294	.07	.11	.19	.28	.11	.20	.37	.52	.14	.25	.48	.65	.20	.39	.68	**.85**
TabPFN v2.6	T	2W	.040	.285	.08	.12	.24	.36	.12	.22	.40	.65	.15	.27	.49	.76	.20	.39	.70	**.93**
TabPFN v2.6	U	raw	.035	.352	.09	.17	.24	.28	.15	.29	.41	.51	.20	.38	.52	.63	.32	.56	.73	**.81**
TabPFN v2.6	U	2W	.040	.290	.11	.17	.26	.35	.17	.31	.48	.60	.22	.40	.60	.74	.33	.59	**.81**	.92
TabDPT v1.2	T	raw	.036	.278	.08	.13	.21	.33	.12	.21	.41	.60	.15	.28	.54	.74	.21	.39	.74	**.92**
TabDPT v1.2	T	2W	.037	.261	.07	.12	.22	.38	.11	.20	.38	.68	.13	.25	.48	**.81**	.18	.36	.68	**.95**
TabDPT v1.2	U	raw	.041	.340	.10	.16	.22	.27	.17	.28	.40	.50	.21	.37	.51	.61	.32	.56	.72	**.84**
TabDPT v1.2	U	2W	.029	.287	.10	.17	.25	.32	.17	.32	.47	.59	.22	.40	.61	.74	.34	.62	**.82**	.92
TabFM v1.0	T	raw	.044	.288	.07	.12	.20	.31	.12	.22	.39	.59	.15	.27	.50	.71	.21	.38	.70	**.89**
TabFM v1.0	T	2W	.034	.283	.07	.12	.22	.35	.11	.20	.38	.63	.13	.25	.47	.76	.18	.35	.66	**.92**
TabFM v1.0	U	raw	.039	.378	.09	.14	.19	.23	.13	.25	.33	.40	.17	.32	.43	.50	.27	.48	.63	.72
TabFM v1.0	U	2W	.022	.298	.09	.15	.21	.28	.14	.26	.41	.52	.18	.34	.52	.65	.29	.55	.75	**.86**
Chronos-2	series	raw	.040	.280	.08	.14	.25	.35	.12	.26	.49	.64	.16	.33	.59	.77	.24	.50	**.81**	.94
TimesFM v2.5	target	raw	.042	.359	.08	.14	.18	.25	.12	.21	.34	.43	.15	.27	.42	.56	.22	.37	.63	.78
TimesFM v2.5	XReg	raw	.032	1.206	.02	.04	.05	.05	.02	.04	.05	.06	.03	.04	.06	.06	.03	.05	.07	.08
SDID (R)	native	raw	.036	.347	.09	.15	.23	.28	.17	.28	.39	.49	.21	.36	.48	.59	.34	.55	.71	**.81**
Ridge SC ($\tau = 1$)	native	raw	.044	.294	.09	.13	.20	.29	.13	.22	.40	.56	.16	.28	.51	.69	.22	.41	.73	**.89**
Simplex SC	native	raw	.029	.358	.08	.13	.18	.22	.13	.22	.34	.41	.17	.29	.44	.52	.28	.47	.64	.74
Simplex SC + offset	native	raw	.027	.346	.09	.14	.21	.27	.15	.27	.41	.49	.20	.35	.51	.60	.30	.53	.71	.80
KRR (63)	native	raw	.046	.313	.08	.13	.21	.29	.13	.22	.39	.54	.15	.27	.51	.68	.21	.40	.72	**.87**
KRR (15)	native	raw	.045	.316	.09	.14	.23	.32	.14	.23	.42	.58	.17	.30	.53	.71	.23	.42	.75	**.89**
RFF ridge	native	raw	.040	.315	.07	.11	.18	.25	.10	.19	.34	.48	.13	.24	.46	.63	.19	.35	.67	**.83**
CatBoost (32)	native	raw	.041	.302	.07	.12	.20	.30	.10	.20	.38	.58	.13	.26	.49	.72	.19	.37	.70	**.90**
CatBoost (4)	native	raw	.034	.308	.07	.11	.17	.28	.10	.21	.37	.55	.13	.26	.48	.68	.18	.36	.67	**.88**

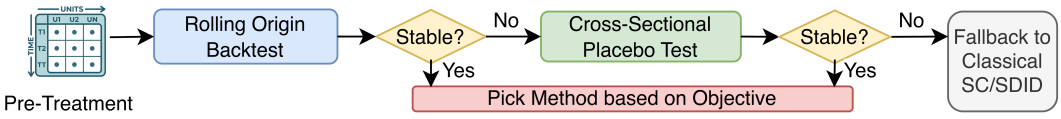


Figure 11: The model×interface×preprocessing selection recipe based on the pre-treatment panel. *Stable* indicates that a model's relative performance is consistent across the backtest or cross-sectional placebo test. Instability coincides with limited context, in which classical SC/SDID outperforms FMs. See Appendix E for details and Figure 12 for recipe performance on EX1(a) simulations.

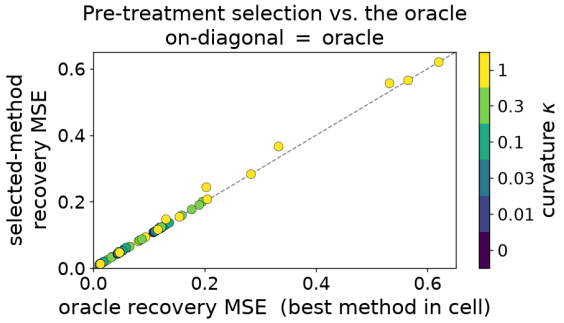


Figure 12: On the EX1(a) simulations, the recovery MSE of the recipe-selected method tracks the oracle (mean regret (Recipe MSE - Oracle MSE), 0.002; on-diagonal = oracle), colored by curvature κ. Note that EX1(a) simulations have sufficiently long T_0, which enhances the recipe performance.