FORECAST QUALITY DOES NOT CERTIFY AN ACTION SIMULATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting models are increasingly used to predict demand under hypothetical actions, effectively treating them as action simulators. Yet accurate factual forecasts do not guarantee reliable simulation. We identify two blind spots. First, per-step forecast evaluation can leave cross-step dependence unconstrained, even though it shapes uncertainty in trajectory-level quantities such as total demand. Second, evaluation under observed actions does not constrain how a model responds when those actions change. On two public retail panels, standard multivariate proper scores largely reproduce marginal-accuracy rankings. A matched rank-zero ablation improves weighted quantile loss (WQL) from .1172 to .1016, while 80% coverage of total demand falls from .669 to .283 and the detected dependence signal disappears. Fine-tuning improves factual forecasts for Chronos-2 and Moirai at every confounding level, but their intervention estimates differ: Chronos-2 shifts from improvement under randomized actions to degradation under confounding, whereas Moirai improves throughout. Decoder rankings also reverse under a second action-response mechanism, showing that simulator comparisons depend on the response being simulated. These findings motivate separate evaluation of marginal accuracy, cross-step dependence, aggregate calibration and interventional fidelity rather than treating forecast quality as a certificate of simulation quality.

1 INTRODUCTION

A demand forecaster is increasingly deployed not to answer "what will sales be" but "what would sales be if I did this". Pricing and deal recommendations, promotion planning and inventory what-if tooling all consume a forecast under a *hypothetical* action plan, and they consume the whole trajectory rather than one position: a seller weighing a week-long deal cares about the horizon total, about whether the post-deal trough cancels the peak, and about the probability the total lands below committed stock. Forecasting foundation models that condition on known future covariates (Ansari et al., 2024; Das et al., 2023; Woo et al., 2024; Ansari et al., 2025) make this attractive: any of them can in principle be queried under a stated action.

Used that way the forecaster is a *simulator*, and a simulator is asked for things per-position accuracy never tests. Its sampled trajectories must move together, because a path assembled from independently correct marginals has the wrong variance for any aggregate; it must be calibrated on the aggregate the decision turns on; and its output must move correctly when the stated action changes—the property separating a simulator from a forecaster. The benchmarks that rank modern forecasters (Aksu et al., 2024; Shchur et al., 2025) report per-position marginal accuracy, which is necessary and, we argue, structurally insufficient.

The sharpest case: one training choice, model-dependent directions. The clearest evidence that forecast quality does not certify a simulator needs no leaderboard. Fine-tuning a pretrained covariate-conditioned forecaster on *logged* action data—the standard way to adapt one—improves factual accuracy at every confounding level for both Chronos-2 and Moirai. What happens to the simulator is not invariant. Chronos-2's aggregate CATE error falls $2.65\times$ when the logged actions were randomized and rises $2.41\times$ when they were confounded, because its fitted action coefficient absorbs the confound. Moirai instead lowers aggregate CATE error throughout; at the strongest

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.

level its mean stepwise contrast moves oppositely. Thus one factual gain yields opposite aggregate directions across checkpoints and opposite functional readings within Moirai. A 91.4M base check near Chronos-2's 119.5M rules out parameter count as the sole explanation.

Structurally, not incidentally, insufficient. That dissociation is permitted rather than accidental, and we make the reason precise and *size* it (§4). Any score that is a sum of per-position terms—MASE, WQL, per-position CRPS, the pinball loss used to train most probabilistic forecasters—is *invariant* to the copula joining its marginals, so no such objective has a gradient toward cross-step dependence and no such metric distinguishes two models differing only in it. Separately, any evaluation under the factual action plan queries the model at a *single point* of action space, so the response elsewhere is unconstrained rather than weakly correlated, and any observed correlation between forecast quality and interventional fidelity is a property of the model population, not of the metrics. Both statements are elementary; their force is quantitative. On Rossmann the calendar/action-adjusted spread ratio is 2.86, and it exceeds one for every one of 1,115 stores; the cheapest proxy for the second—restricting the score to the steps where the action changes—recovers a re-ordering but not the interventional ordering.

A controlled instrument for a missing comparison. Separating these axes needs configurations differing in one design choice at a time, and for the choice most relevant to trajectory simulation the published population does not cleanly expose that comparison. *Causality in time*, which enables rollout with feedback, and *cross-variable attention*, which enables target–covariate interaction inside the sequence model, are confounded across architectures (Table 8): autoregressive forecasters (Salinas et al., 2020; Das et al., 2023) are causal but univariate-target, while cross-variable ones (Liu et al., 2024; Ansari et al., 2025) are bidirectional and one-shot. We instantiate the *causal-in-time, cross-variable-attention* cell with a block-causal mask on a continuous-patch HSTU backbone (Zhai et al., 2024) and build a capacity-matched family around it, so no comparison among our rows is a comparison of parameter count, data, tokenization or training budget.

What a full report costs, and what the standard one hides. We organize evaluation into four dimensions—marginal accuracy, cross-step dependence, trajectory functionals, and interventional consistency—and instantiate all four on Rossmann (1,115 series, $H = 48$) with six capacity-matched 2.64M-parameter configurations plus frozen Chronos-2, replicating two blocks on Favorita (Table 1). Across the six conventional rows Dimensions 1–3 are **one object** ($|\rho| \geq 0.886$): the leaderboards induced by the three dimensions nearly coincide. That is a fact about this population rather than about the metrics, and one row breaks it—**BD-JOINT**, 257 parameters more than BD-OS, holds aggregate coverage at .669 against its nominal .80 target at mid-field accuracy—so aggregate fidelity is a **second** object as soon as anything optimizes it, and interventional fidelity a **third**, weakly aligned with every Dimension-1–3 headline in this model set ($|\rho| < 0.32$), and triangulated by an exact injected target, matched contrasts and lag-wise real-action probes rather than certified by six-row correlations. Nor do the multivariate proper scores supply the dependence axis: both track the marginal ordering ($+0.886$, $+1.000$), the energy score ranks *first* a sampler whose path columns are independent by construction, and BD-JOINT—the only row whose VS_1 shuffle-gap diagnostic fires—sits *in the bottom three* by both.

2 Related Work

Proper scoring of joint forecasts. The energy score (Gneiting & Raftery, 2007) and variogram score (Scheuerer & Hamill, 2015) are proper for multivariate forecasts, and the energy score is known to discriminate dependence weakly (Pinson & Tastu, 2013). Closest to our Dimension-2 finding, Marcotte et al. (2023) map the *regions of reliability* in which multivariate proper scores can discriminate a correct dependence structure at all, and report that the energy score's power degrades sharply with dimension. Our contribution here is a measurement rather than a method: on this model set the weak discrimination is the *dominant* effect—the energy score ranks first a sampler whose path columns are independent by construction, both scores closely track marginal accuracy on two datasets and across a frozen decode sweep, and the one row whose dependence our shuffle diagnostic detects is placed in the bottom three by both. We therefore read the pair for the kind of error it exposes and take the dependence reading from a trajectory functional instead.

Counterfactual prediction under time-varying treatments. From marginal structural models (Robins, 1986; Robins et al., 2000) through RMSN (Lim et al., 2018), CRN (Bica et al., 2020) and the Causal Transformer (Melnychuk et al., 2022), this literature shares our concern with outcomes under unobserved action sequences. Our contribution is orthogonal: these propose *methods* and score per-step counterfactual error, whereas we propose an *evaluation protocol* testing joint trajectory structure, aggregate calibration and the effect of decoding—properties per-step counterfactual error does not see. In off-policy-evaluation terms we diagnose the bias of the direct method (Dudík et al., 2011) one level below policy value. We evaluate a continuous-treatment RMSN adaptation as a point potential-outcome comparator, but not CRN or the Causal Transformer; Appendix G makes the difference in evaluated objects explicit.

Additional positioning against world-model evaluation, multi-step decoding, structure-aware forecasting metrics and marketplace simulation is summarized in Appendix G.

3 Setup: What Is Being Scored

For each of S series, $Y_{i,t}$ denotes the target (daily sales) and covariates are partitioned into *known-future controllable* actions $a_{i,t}$ (promotions, prices, ad spend), *known-future uncontrollable* $c_{i,t}$ (day-of-week, holidays), and *past-only* $z_{i,t}$. A covariate-conditioned forecaster takes $(Y_{i,1:T}, X_{i,1:T})$ with $(a_{i,T+1:T+H}, c_{i,T+1:T+H})$ and returns a predictive distribution over $Y_{i,T+1:T+H}$, from which we draw sample paths. Emitting the horizon also requires a *decoding* choice, parameterized throughout by the decode block size P_{dec}: steps produced per forward pass, from recursive ($P_{\mathrm{dec}} = 1$, each drawn value fed back) to one-shot ($P_{\mathrm{dec}} = H$, no inter-step feedback); the classical taxonomy is due to Bontempi et al. (2012); Ben Taieb & Hyndman (2012). Scoring rules and the causal preliminaries we use ($\mathrm{do}(\cdot)$, overlap, time-varying confounding after Robins (1986)) are in Appendix A.

Two objects can be read off one query, and the dimensions divide by which they need. The first is the sequence of **per-step marginals**, a quantile vector $q^{(i)} \in \mathbb{R}^{H\times|\mathcal{A}|}$ on the grid $\mathcal{A} = \{0.1, \ldots, 0.9\}$—$H$ separate one-dimensional objects, silent about how steps move together. The second is the **sample-path matrix** $\hat{\mathbf{Y}}^{(i)} \in \mathbb{R}^{N\times H}$ of N drawn trajectories, whose rows are *paths*: the pairing across columns within a row is exactly the joint structure $q^{(i)}$ discards. Dimension 1 reads $q^{(i)}$ alone and is the axis existing benchmarks report; Dimensions 2–3 need the row-wise pairing; Dimension 4 requires a *second* query under a modified action and cannot be computed from either object alone. The next section is why that division is forced rather than stylistic.

4 What Per-Position Evaluation Cannot See

This section makes precise, and then *sizes*, what forecast evaluation cannot see. Neither invariance below is new—part (a) is the standard reason multivariate proper scoring rules exist at all (Gneiting & Raftery, 2007; Scheuerer & Hamill, 2015), and part (b) is immediate once stated. We set them down together because they share one structure, because the second has not to our knowledge been drawn for the action channel, and above all because stating them lets us ask how *large* each blind spot is on real data. They do not establish the dissociation of §6; they establish that it is permitted.

Write $\hat{P}$ for a forecaster's joint predictive distribution over $Y_{T+1:T+H}$ under a stated action plan, and $\hat{P}_h$ for its h-th marginal.

Definition 1 (Per-position functional)**.** A functional S of a predictive distribution and a realization is *per-position* if $S(\hat{P}, y) = \sum_{h=1}^{H} s_h(\hat{P}_h, y_h)$ for functions s_h of one marginal and one coordinate.

MAE, RMSE, MAPE, MASE, per-position CRPS, the pinball loss (Eq. 5) and WQL (Eq. 2) are all of this form, as is every headline metric in the benchmarks that rank modern forecasters (Aksu et al., 2024; Shchur et al., 2025). WQL is a ratio of sums rather than a sum, but a fixed monotone transform of one, and nothing below changes.

Proposition 1 (Blindness)**.** *Let S be per-position.* ***(a) Copula.*** *If $\hat{P}$ and $\hat{P}'$ have the same marginals then $S(\hat{P}, y) = S(\hat{P}', y)$ for every y.* ***(b) Action response.*** *Let $M : \mathcal{A}^H \to \Delta(\mathbb{R}^H)$ map an action plan to a predictive distribution and let a° be the factual plan. Any evaluation querying the model only at a° is a functional of $M(a^\circ)$ alone; so if $M(a^\circ) = M'(a^\circ)$, every such evaluation scores M and M' identically whatever the contrast $\tau_\phi(a', a'')$ of Eq. 12 does elsewhere.*

Part (a) is immediate from Definition 1, and by Sklar's theorem the joint distributions sharing fixed marginals are exactly the copulas joining them, so S is *constant* on that set. Part (b) is immediate from a° being one point of $\mathcal{A}^H$; both are written out in Appendix B. Two consequences:

Proposition 2 (A universal upper bound on the copula blind spot)**.** *Fix marginals with finite standard deviations $\sigma_1, \ldots, \sigma_H$, not all zero. For every copula joining them,*

$$\frac{\mathrm{sd}(\phi_{\mathrm{sum}})}{\mathrm{sd}_\perp(\phi_{\mathrm{sum}})} \le \frac{\sum_h \sigma_h}{\sqrt{\sum_h \sigma_h^2}} \le \sqrt{H}, \tag{1}$$

where $\mathrm{sd}_\perp(\phi_{\mathrm{sum}}) = \sqrt{\sum_h \sigma_h^2}$ is the independent-coupling value. The first bound is attained only when the centered coordinates can be coupled as nonnegative scalar multiples of one random variable; the second is attained when the σ_h are equal. Both bounds are sharp over classes of marginals, while for fixed marginal shapes the attainable maximum can be smaller. Every per-position score is nevertheless constant over all these copulas by Proposition 1(a).

This makes the measured factor interpretable rather than merely large: at $H = 48$ no copula can exceed the independent spread by more than $\sqrt{48} = 6.93$, so Rossmann's pooled 2.86 occupies at least 41% of the attainable fixed-marginal maximum. This is a floor because unequal variances and incompatible marginal shapes can only lower the ceiling. The bound is relative to independence; ratios between two arbitrary copulas need not obey it.

Corollary 1 (No gradient toward the copula)**.** *The pinball objective summed over horizon positions is per-position, so it takes the same value for every copula joining the same marginals and supplies no loss signal for choosing among them. Shared-parameter updates may change the copula incidentally, but the objective does not reward moving it toward the truth.*

Corollary 2 (The action blind spot is infinite-dimensional)**.** *Suppose $\mathcal{A} \subseteq \mathbb{R}$, write $P \oplus v$ for the translate of P by $v \in \mathbb{R}^H$, and let $\mathcal{D}_{a^\circ} = \{a - a^\circ : a \in \mathcal{A}^H\}$. For any forecaster M and any $g : \mathcal{D}_{a^\circ} \to \mathbb{R}^H$ with $g(0) = 0$, the perturbed forecaster $M_g(a) = M(a) \oplus g(a - a^\circ)$ satisfies $M_g(a^\circ) = M(a^\circ)$ exactly, so M and M_g are indistinguishable to every metric computed under the factual plan; if ϕ is additive under translation their contrasts differ by exactly $\phi(g(a' - a^\circ)) - \phi(g(a'' - a^\circ))$. The family of such g has dimension $H \cdot (|\mathcal{A}|^H - 1)$ for a finite action set and is infinite-dimensional when $\mathcal{A}$ is continuous—$H(2^{48} - 1)$ directions for arbitrary binary `Promo` plans over Rossmann's horizon, or still $H(2^8 - 1) = 255H$ if interventions are restricted to the eight observed transition positions. The injected continuous price is genuinely infinite-dimensional. Either way the blind spot is not one scalar.*

Indistinguishability is general; the exact discrepancy identity additionally requires translation additivity, as for the total but not generally the other functionals.

The two multivariate proper scores are *not* per-position, so part (a) does not apply to them; but each sums a marginal and a dependence term, and if the marginal term dominates empirically the score inherits Dimension 1's ordering without being copula-invariant. §6.1 finds exactly that and measures how far apart the two terms sit (Appendix C).

How large are the two blind spots? Proposition 1 would be vacuous if the unconstrained quantities were small. On Rossmann neither is.

Calendar-adjusted residual dependence gives a factor of 2.86. After removing cell means for day-of-week × `Promo` × `Open` × `StateHoliday`, the 48-day residual total has standard deviation $20{,}295$ observed against $6{,}998$ under independence, a ratio above one for $1115/1115$ stores (constructed in Figure 3). This is what Corollary 1 leaves unconstrained and ϕ_{sum} reads. Two qualifications: it is a property of this residualization, not a model metric or full-history residual; and it is mostly slow drift—removing periods ≥ 49 days drops the ratio to 0.82 (Appendix A.2).

The action blind spot is worth up to 82% *of the effect.* A model that learned nothing about the action scores exactly $|\tau^*|$ by part (b); at -2σ, $\tau^* = .254$, and our configurations span .068–.208, i.e. 27%–82% of that no-response baseline, a $3.1\times$ spread on the quantity factual scoring leaves free while their WQL spans only .080–.231. Nor does the cheapest in-Dimension-1 proxy close it: restricting the score to $\mathcal{T} = \{h : a_h \neq a_{h-1}\}$ still reads $M(a^\circ)$ only, and recovers a re-ordering rather than the interventional one (App. D).

What this licenses, and what it does not. Proposition 1 says a Dimension-1–3 score *cannot constrain* Dimension 4 under factual-only evaluation. For Dimension 3, the positive long-range residual dependence motivates underdispersion, but blindness alone does not determine its direction; the coverage tables do. The proposition makes a null between the axes **permitted**; §6.1 reports one **observed** on six configurations. Neither step shows the two are *generally* unrelated—nothing forbids a population in which better forecasters also respond better—so any observed relationship is a fact about the population and cannot be relied on when qualifying a new model. This paper's empirical content is that the blind spots are *occupied*: real models, including a 119.5M-parameter pretrained one, differ inside them.

5 The Protocol

Each dimension is defined by what it reads and by which blind spot it closes. Reproducible definitions of every metric named here—MASE, ES, VS_p, the calibration statistics and the newsvendor cost—are in Appendix A.1.

Dimension 1: marginal accuracy. From $q^{(i)}$ alone, using the two metrics standard in recent benchmarks (Aksu et al., 2024; Shchur et al., 2025): MASE (Hyndman & Koehler, 2006) and weighted quantile loss,

$$\mathrm{WQL} = \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \frac{\sum_{i,h} 2\big[\alpha \, (y_h^{(i)} - q_{h,\alpha}^{(i)})_+ + (1-\alpha) \, (q_{h,\alpha}^{(i)} - y_h^{(i)})_+\big]}{\sum_{i,h} |y_h^{(i)}|}, \tag{2}$$

a ratio of sums so near-zero-volume series cannot dominate. We also report both restricted to $\mathcal{T} = \{h : a_h \neq a_{h-1}\}$, the steps where the factual action changes, and to its complement—the cheapest interventional-flavoured reading available without a $\mathrm{do}(\cdot)$ query (§4). MASE is the column to read across subsets, WQL across configurations, for a normalization reason given with Table 2.

Dimension 2: cross-step dependence. The two multivariate proper scoring rules that read the sample-path matrix directly: the energy score (Gneiting & Raftery, 2007), a kernel score built from Euclidean distances between sampled paths and the realization, and the order-p variogram score (Scheuerer & Hamill, 2015), which compares predicted and realized pairwise increments $|y_h - y_{h'}|^p$ under weights $1/|h - h'|$, with $p = 1$ throughout (Appendix A.1). We report them as a pair because their invariances differ: VS_p reads only increments, so shifting a whole trajectory leaves it unchanged, whereas ES is sensitive to such a shift. Neither *isolates* dependence: each sums a marginal and a dependence term. To separate the two we permute each column of $\hat{\mathbf{Y}}^{(i)}$ independently and rescore—the *joint gap*, a within-model contrast that preserves every marginal exactly and so is immune to marginal quality. On the six conventional configurations it is a null ($|\mathrm{gap}| \leq .0006$ relative, unstable in sign); the row that makes it fire is §6.1's subject, and the diagnostic is validated against known-dependence input in Appendix A.3.

Dimension 3: trajectory functionals. Applying $\phi : \mathbb{R}^H \to \mathbb{R}$ row-wise to $\hat{\mathbf{Y}}^{(i)}$ gives an empirical predictive distribution over ϕ, scored against $\phi(y^{(i)})$. For non-additive ϕ this is not recoverable from $q^{(i)}$, which is what places Dimension 3 outside Dimension 1's reach. We report five—horizon total, peak, volatility, time-to-peak and lag-one autocorrelation—the first three permutation-invariant within a path, so they read the *strength* of dependence as well as marginal location and scale, and the last two order-sensitive. Dependence is diagnosed by pairing aggregate metrics with per-step calibration and the shuffled-column contrast, not by the aggregate alone. Each functional is scored by CRPS_ϕ, central-interval coverage and the KS distance of its PIT values, and ϕ_{sum} additionally by newsvendor cost at $p/h \in \{1, 3, 9\}$, whose critical fractile sits in the tail where dependence contributes most of the variance.

A controlled reconstruction reference. Inverting the nine-point quantile grid clips draws at the 10th and 90th percentiles. Under independent standard-Gaussian steps and independent inverse-grid draws with $N = 300$, exactly calibrated per-step quantiles give

$$\mathrm{Cov}_{\phi_{\mathrm{sum}}}^{\mathrm{indGauss}}(0.2/0.4/0.6/0.8) = 0.166/0.335/0.513/0.708. \tag{3}$$

This isolates clipping; it is not an empirical ceiling because aggregate coverage also depends on marginal shapes and the true and generated copulas. We assess every row against nominal coverage. Native samples avoid this inverse-grid artifact; Dimension-1 quantiles still use the common grid.

Dimension 4: interventional consistency. The estimand is the counterfactual trajectory $Y_{T+1:T+H}(\mathrm{do}(a = a'_{1:H}))$, never observed jointly with the factual one. Since real observational data does not supply the true effect, we use two settings that fail in opposite ways.

(a) Semi-synthetic, ground truth available. We retain each series' base trajectory Q^{base} and inject a *known* mechanism, following RealCause (Neal et al., 2020) extended to the sequential setting: $Q_t^{\mathrm{sim}} = Q_t^{\mathrm{base}} M_t$ with $M_t = \exp(\beta \Delta p_t - \rho I_t)$ driven by a log-price deviation and a carryover state $I_t = \delta I_{t-1} + (M_{t-1} - 1) o_{t-1}$ reproducing a promotion's peak–trough–recovery shape; actions come from a behaviour policy whose confounding strength is swept. The mechanism being known, the ground-truth trajectory effect τ^*_ϕ—the difference in $\mathbb{E}[\phi(Y)]$ between two $\mathrm{do}(\cdot)$ plans—is exactly computable, giving CATE error $|\hat{\tau}_\phi - \tau^*_\phi|$. Two properties make the layer measure what it claims. Because $M_{1:H}$ is a function of the action sequence alone, free of the series level, differencing two arms in logs cancels Q^{base} *exactly*,

$$\log Y_t(a') - \log Y_t(a'') = \beta\big(\Delta p'_t - \Delta p''_t\big) - \rho\big(I'_t - I''_t\big), \tag{4}$$

so the contrast metrics are immune to baseline forecast error—which is what separates "forecasts well" from "simulates well" rather than confounding them. And the injected action is *absent* from the real covariates (we inject price, which neither dataset carries), so $\Delta p_t = 0$ reproduces the real series exactly and the ground truth is exact rather than real-plus-injected. Appendix F summarizes the construction and its seven design points.

(b) Real action, no ground truth. On Rossmann's observed `Promo` we report properties a correct simulator must satisfy whatever the true effect is: sign rate, dose monotonicity, a discrimination ratio against a same-action placebo re-draw, the same probe read *lag by lag* over a forced five-day spell, and agreement with a matched on/off association from $46{,}149$ within-store day-pairs. These can convict a simulator and cannot certify one, so they never carry a magnitude claim alone (Appendix E.5). Their role in §6.1 is as *independent references*: the injected effect, the matched association and the event-study profile use three evaluation constructions, but only the injected mechanism identifies an effect; the latter two are observational probes whose agreement supplies triangulation.

Instrument: configurations. We instantiate a comparison not cleanly isolated by the evaluated forecasting baselines in Table 8: a block-causal mask (Eq. 11) on a continuous-patch HSTU backbone (Zhai et al., 2024) (~ 2.64M parameters), bidirectional *within* a timestep and causal *across* timesteps. The mask contributes no parameters, so the quantile-head variants share parameter count, data, optimizer and objective. Rows are named by mask (BC block-causal, BD bidirectional) and decoding (AR one step per pass with the drawn value fed back, OS one-shot), with +PA marking prefix-augmented training that matches loss positions across the two decoders. **BC-AR** versus **BD-OS** is the primary comparison and **BD-AR** separates mask from decoding. What matters for Dimension 2 is that AR rows are *joint* samplers—step $h{+}1$ is drawn conditional on the realized draw at step h—whereas OS rows invert per-step quantile grids independently and are *marginal* samplers whose path columns are independent **by construction**, so no cross-step dependence reaches the drawn paths whatever the encoder represents internally. **BD-JOINT** keeps BD-OS's encoder and single forward pass but replaces the per-step quantile grids with a diagonal-plus-low-rank Gaussian over the horizon, $\Sigma = \mathrm{diag}(s^2) + FF^\top$, so a draw $y = \mu + s \odot \varepsilon + Fz$ shares one latent z: correlated columns from one pass, with no step conditioning on a realized draw of another. It is the only configuration whose objective is not a sum of per-position terms—hence the only one Corollary 1 does not apply to—and it costs 257 parameters more than BD-OS. Its marginals are analytic, so its Dimension-1 quantiles use the same grid as every other row. **BD-DIAG** is its rank-zero control: the same encoder, Gaussian likelihood and native sampler with F removed, isolating shared factors from the output-family change relative to BD-OS. **Chronos-2** (Ansari et al., 2025) (119.5M, frozen, zero-shot) occupies BD-OS's cell at roughly $45\times$ the scale, separating scale and pretraining from the design choice. Architectures, training details and two unscored causal ablations are in Appendices A.4 and F.3.

Table 1: **Cross-dimension summary.** One headline metric per dimension; per-dimension tables in App. D and E.5, correlations in App. D.2. Dim-3 coverage is assessed against nominal .80 for every path distribution; BD-JOINT reaches 84% of that target against Chronos-2's 69%, though coverage carries the largest across-seed spread of any column (BD-JOINT ±.127, App. D). Dim-4 reference cells use five seeds; **Dim 4 is unbolded because its ordering is mechanism-conditional.**

	Dim 1	Dim 2	Dim 3		Dim 4	
Config	WQL ↓	VS_1 ↓	$CRPS_{\phi_{sum}}$ (10^3) ↓	Cov	CATE err. ↓	Rank ρ ↑
BC-AR	**.0801**	**.0361**	9.95	.288	.079	**.940**
BD-AR	.2312	.1654	37.81	.082	.117	.429
BD-AR+PA	.1050	.0522	12.19	.113	.097	.727
BD-OS	.0945	.0456	11.86	.165	.121	.925
BD-OS+PA	.0936	.0451	10.26	.215	.068	.915
BD-JOINT	.1172	.0600	11.88	**.669**	.069	.869
Chronos-2	.0928	.0441	**6.75**	.551	.208	.658

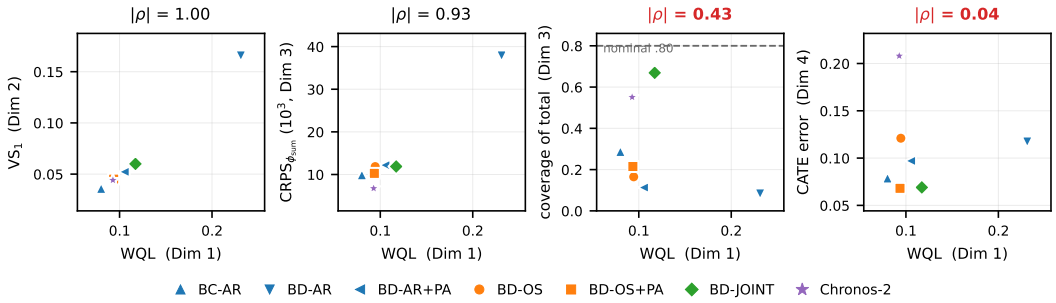

Figure 1: **Three reportable objects, not four dimensions and not one.** Marginal accuracy against one metric from each other dimension; $|\rho|$ is Spearman, **over all seven rows and so including BD-JOINT** (the $n = 6$ values quoted in the text are in Table 4). Accuracy, the multivariate score and the trajectory functional are one object (1.00, 0.93). *Aggregate* fidelity separates once a row optimizes it directly—BD-JOINT (green diamond) reaches .669 coverage against its nominal .80 target with mid-field accuracy, dropping that panel to 0.43. The dashed reference marks nominal .80. Interventional fidelity is a third object at 0.04. Marker shape gives the decoder.

6 Experiments

The purpose is not a new state of the art on marginal accuracy: it is to test whether the protocol's dimensions are empirically separable and whether the blind spots of §4 are occupied. Per-dimension tables and the Favorita replication are in Appendices D–F.5. **Rossmann** (`rossmann_1D` via FEV-BENCH): 1,115 daily drugstore series, $H = 48$, target `Sales`, action the binary `Promo`. **Favorita** (1,579 store×family series, $H = 28$) replicates Dimensions 2–3. A shared held-out window lets Dimensions 1–3 be read from the *same* forward pass ($N = 300$ paths), with two later cutoffs as a targeted robustness check; five seeds per headline configuration, reference semi-synthetic cell, Moirai-small adaptation and DeepAR/cRMSN comparator; three for sweeps, Moirai-base and Chronos-2 adaptation; and frozen references are deterministic (Appendix A.4). Rossmann's promotion calendar is national and byte-identical across stores, so the real-data probes below are associations rather than identified effects.

6.1 Three Objects, Not Four Dimensions

All correlations are Spearman over the $n = 6$ configurations; with six rows each is suggestive rather than established, and the reading below rests on the joint pattern rather than any single coefficient.

Dimensions 1–3 collapse to one object across the conventional rows. Every pair among marginal accuracy, both multivariate scores, trajectory-functional accuracy, aggregate calibration and decision cost co-moves at $|\rho| \geq 0.886$ across the six *conventional* configurations (App. D.2), so on Rossmann the resulting leaderboards nearly coincide (Figure 1). That is a property of *that*

population rather than of the metrics, and BD-JOINT breaks it: excellent aggregate calibration at mid-field accuracy drops the magnitude of the accuracy–calibration rank correlation from 0.94 to 0.43. A leaderboard is only self-consistent while nobody optimizes the aggregate directly; the ten-row sensitivity is in App. D.2. The group still says *why* a row sits where it does—BD-OS and Chronos-2 have near-identical WQL yet differ by $3.3\times$ in coverage of the total—so the ordering is shared and the diagnosis is not.

Dimension 4 is weakly aligned with forecast metrics and requires triangulation. No Dimension-1–3 headline reaches $|\rho| = 0.32$ against semi-synthetic CATE error, and none exceeds 0.60 against the matched-pair rank correlation. This null is permitted by Proposition 1(b), but six selected configurations cannot distinguish zero from a moderate relationship. What makes it a measurement is that Dimension 4 rests on **two data sources and three complementary evaluation probes** (§5). The two cross-source coefficients are positive but inconclusive ($+.429$, $+.638$); the two real-action probes agree at $+.812$. We therefore rely on triangulation, not a six-row correlation.

The standard multivariate proper scores do not supply the dependence axis. Three rows invert each step's quantile grid independently, so their path columns are independent *by construction*. Yet frozen Chronos-2 has the **best** energy score of all six (4.38 against BC-AR's 4.52), and BD-OS and BD-OS+PA beat two of the three joint samplers. ES closely tracks the WQL ordering, with one adjacent swap (Spearman $+0.886$), and VS_1 orders the set *identically* to WQL ($+1.000$); the pattern repeats on Favorita and across all seven frozen decode blocks—the weak discrimination of Pinson & Tastu (2013) as the *dominant* effect. This does not violate propriety, which concerns expected score under the data-generating joint and does not isolate dependence when marginal quality differs; the scalar leaderboard is simply marginal-dominated.

BD-JOINT tests whether that dominance persists when a detectable dependence signal is present. It is the only row whose score moves when its own columns are shuffled: relative VS_1 joint gap $+.0166\pm.0023$ across seeds, against $|\mathrm{gap}| \leq .0006$ elsewhere. A null gap means no dependence that this diagnostic detects and that VS_1 rewards—not an absence of dependence—so the autoregressive rows may carry structure the pairwise-difference statistic is blind to. On synthetic input the diagnostic returns zero on independent columns, scales with injected dependence, and matches an equicorrelated control near 0.15—a control scale, not a covariance estimate (App. A.3). BD-JOINT also attains the study's best coverage of the horizon total (Table 1)—yet both scores place it in the bottom three: second-to-last by VS_1 and third-to-last by ES. Because the joint gap is a *within-model* contrast it is immune to the objection that BD-JOINT simply forecasts worse.

BD-DIAG isolates the dependence gain. Changing only BD-JOINT's covariance rank from 8 to 0 retains its encoder, Gaussian likelihood and native sampler, and *improves* WQL to $.1016\pm.0047$, while coverage collapses from $.669\pm.127$ to $.283\pm.055$ and the shuffle gap from $+.0166\pm.0023$ to $+.0001\pm.0009$. Thus shared dependence, not output family, sampling route or better marginals, produces the aggregate gain (App. D). BD-JOINT's Dimension-4 error is seed-unstable and no better than the best quantile-head rows, so we make no interventional claim for it; **its distinctive standing is best aggregate calibration in the study and the only dependence the shuffle test detects, yet bottom three on both joint scores.** The dependence reading therefore belongs on Dimension 3 (Appendix D).

Fine-tuning improves both forecasts, but the simulator can move either way. For Chronos-2, all six $\lambda = 1$ fine-tunes improve factual RMSE ($1174 \rightarrow$ 778–932) and worsen CATE ($.208 \rightarrow$ .49–.73). Full fine-tuning at realized $R^2 = .001, .152, .621$ gives CATE gaps $-.088\pm.008, +.105\pm.090, +.293\pm.002$, every cell keeping its sign on all three seeds. At $R^2 = .001$ it recovers elasticity ($1.01\pm.03$ of truth) and cuts CATE $2.65\times$; at .621 the fitted coefficient absorbs the confound ($|\hat{\beta}| = 1.76\pm.08$ of truth).

The second family is **Moirai**. Small (13.8M) lowers factual RMSE 20%, 24%, 33% and CATE ($-.049\pm.022, -.091\pm.020, -.121\pm.063$). Yet at $R^2 = .621$ its mean stepwise Log-contrast MAE moves oppositely ($.206 \rightarrow .243\pm.060$) while aggregate CATE improves ($.256 \rightarrow .135\pm.063$), a five-seed mean rather than a retraining-robust sign. A near-scale-matched base (91.4M versus Chronos-2's 119.5M) preserves every direction and strengthens that split ($.255 \rightarrow .080\pm.003$ CATE; $.205 \rightarrow .270\pm.068$ Log-MAE), ruling out parameter count as the sole explanation (App. E.2). Other model

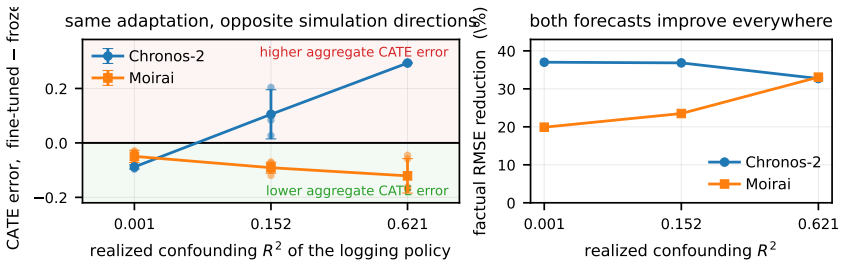

Figure 2: **Forecast adaptation does not determine simulator adaptation.** *Left:* fine-tuned-minus-frozen aggregate CATE error; points are training seeds, bars their s.d. Chronos-2 changes from helping to hurting as confounding rises, whereas Moirai improves this aggregate throughout. *Right:* both architectures improve factual RMSE in every cell.

differences restrict us to within-model contrasts. Forecast gains therefore certify neither the direction nor the functional of simulator change (Figure 2).

Within Dimension 4 the separation is real, but the ordering belongs to the mechanism. The spread is model-attributable rather than noise: the injected mechanism is deterministic, so the achievable CATE error is exactly zero; a correctly-specified estimator reaches .013, a continuous-treatment RMSN baseline .061±.005, within reported seed variation of the best forecasting row's .068±.017 (Appendix E.3). **The ordering, however, does not survive a change of response form.** Re-running all five configurations on a second injected family—also exactly multiplicative, with carryover mass matched to the first, but with a *bounded* action response and a hump-shaped finite kernel instead of linear-in-log and geometric decay (Appendix F.1)—inverts the extremes. What reverses is the **decoding axis**. Under the elasticity family one-shot decoding is the better simulator (31.7% of the no-response baseline against autoregressive rollout's 42.1%); under the second family autoregressive wins (34.5% against 53.2%), on every seed at group level. **Confounding strength is not the explanation:** rebuilding the second family at a matched $R^2 = 0.626$ against the elasticity cell's 0.621 leaves the reversal intact and slightly wider, better on every seed. Two cell differences remain—response boundedness and carryover shape—and we did not separate them (App. G.1). Row-level ranks do not reverse, so the transferable statement is about the decoder rather than individual rows. **The dataset axis, by contrast, does not reverse.** Re-running the same five configurations on a Favorita cell built with the *same* elasticity mechanism (Appendix E.4) preserves the decoder ordering—one-shot again ahead of autoregressive, 78.9% against 109.0% of the no-response baseline—at rank correlation +0.60 with Rossmann. Changing the dataset perturbs the ordering; changing the response form inverts which decoder to prefer. A simulator leaderboard is therefore meaningful only relative to a stated response form, and **we make no model recommendation**. *Why* the ordering changes remains open—over-extrapolation predicts interventional error under both families, a difference $n = 6$ cannot resolve. One out-of-support arm makes the otherwise degenerate plan menu discriminate sharply (App. F.2).

Finally, **the decode setting is part of the evaluated object**: changing only block size moves WQL .0928 → .1266, while coverage peaks at $P_{\text{dec}} = 12$ (.655 versus .551 native) and recovered-elasticity error and Log-contrast MAE move oppositely. Report it with the checkpoint (App. F.4).

7 Conclusion

Factual forecast accuracy is necessary for action simulation, but it is not sufficient. Our controlled experiments show that marginal accuracy, aggregate calibration and intervention quality can move in different directions. They also show that decoder rankings depend on the assumed response mechanism. We therefore recommend evaluating simulators separately for marginal accuracy, cross-step dependence, aggregate calibration and interventional fidelity. Our evidence remains limited to daily retail data; testing other domains with naturally varying actions is the main next step.

REPRODUCIBILITY STATEMENT

Every number in this paper was produced by project code and checked against stored result artifacts. The compact appendix maps each main claim to its proof, construction or evidence. An anonymous supplementary artifact has been prepared with the original extended technical appendix, stored outputs, and available analysis, verification and figure-generation scripts, but it is not part of the present submission pending institutional approval and it does not contain the complete training/evaluation pipeline. The present PDF is self-contained for the claims it makes. Specifically, the compact appendix gives the datasets, splits, training settings, seed counts and sample budgets (Appendix A.4); closed-form definitions of the headline metrics (Appendix A.1); the injected mechanisms (Appendices F and F.1); the known-dependence control for the shuffled-column diagnostic (Appendix A.3); and the plan-menu conclusion (Appendix F.2). The underlying plan-menu analysis was run end to end by `plan_menu_support.py`; its in-support column was checked against each stored artifact's own regret. Uncertainty conventions are stated explicitly and differ by claim: $\pm$ denotes across-seed standard deviation (five training seeds on the headline Dimensions 1–3, semi-synthetic reference cells and Moirai adaptation, and the DeepAR/cRMSN comparators; three on temporal replications, sweeps and Chronos-2 adaptation), while ordering claims in the decode sweep and the Dimension-4 tables additionally carry a paired store-level bootstrap over the panel ($4{,}000$ resamples); Appendix G.1 records why the two intervals answer different questions and one case in which a tight store bootstrap licensed a claim the seed spread later removed. Both datasets are public. Appendix G.1 also lists every claim withdrawn from earlier drafts and every axis of the design that is not populated.

AI USE STATEMENT

We disclose LLM use across the task categories the ICLR AI policy asks about.

Tasks for which disclosure is required. LLM-based coding assistants were used to *implement methods*: writing and refactoring evaluation, scoring and plotting code, and the generator for the semi-synthetic interventional cells, which makes them a contributor to *producing synthetic datasets*. They were used for *dataset cleaning and reformatting* of the two public panels into the evaluation harness's format, and to *interpret results* in the weak sense of drafting candidate readings of finished tables, each of which the authors accepted, revised or rejected against the underlying artifacts. Assistants also ran consistency checks over the manuscript—dangling references, stale claims, arithmetic and mathematical consistency—which surfaced several errors. They provided feedback on experimental design and prioritization, proposed follow-up controls and comparator checks—including the matched rank-zero joint-output control—and suggested implementation and run parameters; the authors decided which experiments to run and checked the resulting analyses against the stored artifacts. An assistant proposed revisions to a proposition statement and proof after identifying that the earlier equality conditions were false; the authors checked the corrected derivation and selected the final statement. Claims withdrawn or narrowed as a result are recorded in Appendix G.1 rather than silently removed. The research questions, the four-dimension protocol, the injected mechanisms and behaviour policy, and every final decision about which claims to make or withdraw are the authors'.

Tasks for which disclosure is recommended. Assistants were used to draft and edit prose, to generate and revise figure code, to format references, and to search for related work; every citation was checked against the cited source by an author.

Verification and responsibility. Every empirical number was produced by project code and rederived from the stored result artifacts; no result, table entry, figure value, or citation was reported on an LLM's assertion. Mathematical revisions suggested by an assistant were independently checked by an author. We take responsibility for the final content of this work, including all text, claims and artifacts produced with AI assistance.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation.

arXiv preprint arXiv:2410.10393, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Souhaib Ben Taieb and Rob J. Hyndman. Recursive and direct multi-step forecasting: the best of both worlds. Technical Report 19/12, Department of Econometrics and Business Statistics, Monash University, 2012.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela Van Der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

Gianluca Bontempi, Souhaib Ben Taieb, and Yann-Aël Le Borgne. Machine learning strategies for time series forecasting. In *European big data management and analytics summer school*, pp. 62–77. Springer, 2012.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Riku Green, Grant Stevens, Zahraa Abdallah, and Telmo M. Silva Filho. Stratify: Unifying multi-step forecasting strategies, 2024. URL https://arxiv.org/abs/2412.20510.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2006.03.001. URL https://www.sciencedirect.com/science/article/pii/S0169207006000239.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

Bryan Lim, Ahmed M. Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/56e6a93212e4482d99c84a639d254b67-Abstract.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Étienne Marcotte, Valentina Zantedeschi, Alexandre Drouin, and Nicolas Chapados. Regions of reliability in the evaluation of multivariate probabilistic forecasts. In *International Conference on Machine Learning*, volume 202, pp. 23958–24004. PMLR, 2023. URL https://proceedings.mlr.press/v202/marcotte23a.html.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*, volume 162, pp. 15293–15329. PMLR, 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

Martin Mladenov, Chih-Wei Hsu, Vihan Jain, Eugene Ie, Christopher Colby, Nicolas Mayoraz, Hubert Pham, Dustin Tran, Ivan Vendrov, and Craig Boutilier. Recsim ng: Toward principled uncertainty modeling for recommender ecosystems. *arXiv preprint arXiv:2103.08057*, 2021.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

Thanapol Phungtua-Eng and Yoshitaka Yamamoto. Are we winning the wrong game? revisiting evaluation practices for long-term time series forecasting. *arXiv preprint arXiv:2603.08156*, 2026.

Pierre Pinson and Julija Tastu. Discrimination ability of the energy score. Technical Report DTU Compute Technical Report-2013 No. 15, Technical University of Denmark, 2013.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's time: Towards the next generation of time series forecasting benchmarks. *arXiv preprint arXiv:2602.12147*, 2026.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Sandeepa Weerasekara and Sandareka Wickramanayake. Topocast: A topological fidelity framework for evaluating transformer-based time series forecasting. *arXiv preprint arXiv:2606.25439*, 2026.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Yu Xia, Ali Arian, Sriram Narayanamoorthy, and Joshua Mabry. Retailsynth: Synthetic data generation for retail ai systems evaluation. *arXiv preprint arXiv:2312.14095*, 2023.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026. URL https://arxiv.org/abs/2606.15032.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, Yinghai Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations, 2024. URL https://arxiv.org/abs/2402.17152.

APPENDIX ROADMAP

This appendix keeps the proofs, definitions and evidence needed to audit the submitted claims.

Main claim	Compact support
Blindness propositions and copula bound	App. B
Metrics, diagnostic controls and setup	App. A
Dimensions 1–3, temporal robustness and correlations	App. D
Semi-synthetic, adaptation, references and real-data probes	App. E
Injected mechanisms, decoder sweep and architecture cell	App. F
Scope and limitations	App. G.1

A DEFINITIONS, DIAGNOSTICS AND SETUP

A.1 METRIC DEFINITIONS

For target y and predicted quantile $\hat{q}_\alpha$, the pinball loss is

$$\mathcal{L}_\alpha(y, \hat{q}_\alpha) = \begin{cases} \alpha(y - \hat{q}_\alpha), & y \geq \hat{q}_\alpha, \\ (1-\alpha)(\hat{q}_\alpha - y), & y < \hat{q}_\alpha. \end{cases} \tag{5}$$

Summing this loss over positions is a per-position objective in the sense of Definition 1. With median forecast $\hat{y}_{i,h}$ and seasonal scale $s_i = (T-7)^{-1} \sum_{t=8}^{T} |y_{i,t} - y_{i,t-7}|$,

$$\mathrm{MASE} = (SH)^{-1} \sum_{i,h} |y_{i,h} - \hat{y}_{i,h}| / s_i. \tag{6}$$

For sampled paths $\hat{\mathbf{Y}}_{1:N}$, the joint scores used here are

$$\mathrm{ES} = \frac{1}{N} \sum_n \|\hat{\mathbf{Y}}_n - y\|_2 - \frac{1}{2} \mathbb{E}_{n \neq n'} \|\hat{\mathbf{Y}}_n - \hat{\mathbf{Y}}_{n'}\|_2, \tag{7}$$

$$\mathrm{VS}_1 = \frac{\sum_{h<h'} w_{hh'} \left(|y_h - y_{h'}| - \frac{1}{N} \sum_n |\hat{Y}_{n,h} - \hat{Y}_{n,h'}| \right)^2}{\sum_{h<h'} w_{hh'}}, \qquad w_{hh'} = |h - h'|^{-1}. \tag{8}$$

The five trajectory functionals are horizon total, peak, within-path volatility, time-to-peak and lag-one autocorrelation. Write $z_n = \phi(\hat{\mathbf{Y}}_n)$ and $z^* = \phi(y)$. Their scalar score and PIT are

$$\mathrm{CRPS}_\phi = \frac{1}{N} \sum_n |z_n - z^*| - \frac{1}{2N^2} \sum_{n,n'} |z_n - z_{n'}|, \tag{9}$$

$$u_\phi = \frac{1}{N} \sum_n \mathbf{1}[z_n \leq z^*]. \tag{10}$$

Coverage is the fraction of series whose u_ϕ lies in the requested central interval; KS is the supremum distance between the empirical PIT CDF and Uniform$(0,1)$. For the horizon total, the newsvendor order is $Q_i^* = \hat{F}_i^{-1}(p/(p+h))$ and its cost is $h(Q_i^* - z_i^*)_+ + p(z_i^* - Q_i^*)_+$, with $h = 1$ and $p/h \in \{1, 3, 9\}$. Unit-carrying scores are normalized by each series' historical absolute mean, multiplied by H for the total. Under independent standard-Gaussian steps, independent inverse-grid draws and $N = 300$, clipping the nine-point grid gives .708 coverage at the nominal 80% level (Eq. 3). This is a controlled reconstruction reference, not an empirical ceiling: total coverage also depends on marginal shapes and the true and generated copulas. Every empirical row is therefore assessed against nominal .80.

With D variables represented as adjacent tokens at each time, the block-causal mask is

$$A_{s,s'} = \mathbf{1}[\lfloor s'/D \rfloor \leq \lfloor s/D \rfloor], \tag{11}$$

which is bidirectional within a timestep and causal across timesteps. For trajectory functional ϕ, the interventional estimand is

$$\tau_\phi^*(\mathcal{H}_i; a, a') = \mathbb{E}[\phi(Y_i) \mid \mathcal{H}_i, \mathrm{do}(a)] - \mathbb{E}[\phi(Y_i) \mid \mathcal{H}_i, \mathrm{do}(a')], \tag{12}$$

and CATE error is $|\hat{\tau}_\phi - \tau_\phi^*|$, normalized per series.

A.2 SIZING THE COPULA BLIND SPOT

After removing cell means for Rossmann's day-of-week, promotion, open and holiday indicators, we compare the standard deviation of the realized 48-day residual total with the independent-coupling value $\sqrt{\sum_h \mathrm{Var}(\epsilon_h)}$. The values are 20,295 and 6,998, a factor of 2.86, and the ratio exceeds one for all 1,115 stores. This measures the quantity left unconstrained by Corollary 1 under this calendar/action residualization; it is not conditioning on target history and the model's full information set. Most of the dependence is slow drift: removing periods longer than the horizon lowers the ratio to .82.

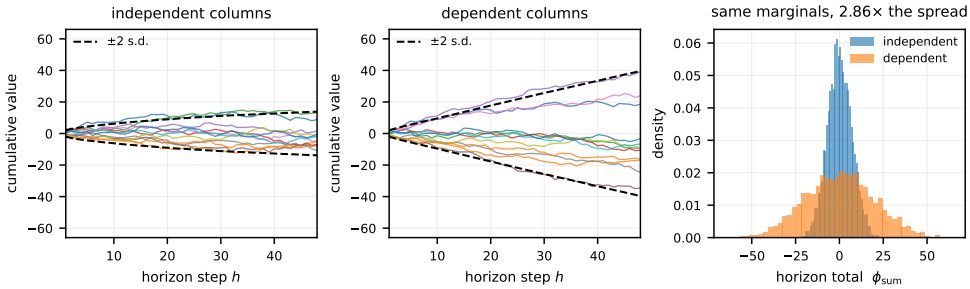

Figure 3: **Fixed marginals, different paths.** Independent and dependent samples have identical per-position marginals but horizon-total spread differing by $2.86\times$.

A.3 CONTROL FOR THE SHUFFLED-COLUMN DIAGNOSTIC

We draw paths with fixed unit marginals and known equicorrelation. Across 240 replicates of $N = 300$, $H = 48$, the relative VS_1 shuffle gap is $-.00003 \pm .00030$ at correlation 0, $.0052 \pm .0009$ at .059, $.0318 \pm .0036$ at .200, and $.3354 \pm .0160$ at .500; WQL is unchanged exactly. BD-JOINT's measured gap $.0166 \pm .0023$ lies near the gap produced by an equicorrelated control around .15; this is an equivalent control scale, not an estimate of its actual correlation. A null gap therefore means "no dependence detected by this instrument," not independence.

A.4 EXPERIMENTAL SETUP

Rossmann contains 1,115 daily series with 462 training days and $H = 48$; Favorita contains 1,579 store–family series with $H = 28$. Capacity-matched rows have 2.64M parameters, hidden width 256, four layers and heads, 80 epochs, learning rate 5×10^{-4}, batch 64, and a 10% validation split. BD-JOINT fixes covariance rank 8 and BD-DIAG uses rank 0. Dimensions 1–3 use $N = 300$ paths and five training seeds. The reference semi-synthetic cell and Moirai-small adaptation use five seeds; Moirai-base, later cutoffs, mechanism replications and Chronos-2 adaptation use three. Semi-synthetic evaluation uses 500 series, $N = 100$, and common random numbers across intervention arms. Frozen pretrained rows are deterministic under this protocol.

B PROOFS

The four statements of §4 are consequences of the definitions rather than deep results, and we present them as such: their role is to make precise *which* quantities a per-position evaluation cannot constrain, so that §6.1 can measure how large those quantities are. We give the arguments in full anyway, because a stated-and-unproved claim is an avoidable objection.

Proposition 1(a). Let $\hat{P}, \hat{P}'$ have the same marginals, so $\hat{P}_h = \hat{P}'_h$ for every h. By Definition 1, $S(\hat{P}, y) = \sum_h s_h(\hat{P}_h, y_h)$, and each summand depends on $\hat{P}$ only through $\hat{P}_h$. Substituting $\hat{P}'_h = \hat{P}_h$ termwise gives $S(\hat{P}', y) = S(\hat{P}, y)$ for every y. □

By Sklar's theorem every joint law on $\mathbb{R}^H$ factors as a copula C applied to its marginals, and conversely every copula yields a joint law with those marginals. The set of joint laws sharing fixed

marginals is therefore exactly the set of copulas joining them, and the display above says S is *constant* on that set—not merely that it is weakly informative about it.

Proposition 1(b). By hypothesis the evaluation is a function of the model's output at the single plan a°, i.e. there is a Ψ with $\mathrm{eval}(M) = \Psi(M(a^\circ))$. If $M(a^\circ) = M'(a^\circ)$ then $\mathrm{eval}(M) = \Psi(M(a^\circ)) = \Psi(M'(a^\circ)) = \mathrm{eval}(M')$. No constraint is placed on $M(a)$ for $a \neq a^\circ$, so the contrast $\tau_\phi(a', a'')$ of Eq. 12 is unrestricted for any $a', a'' \neq a^\circ$. □

Proposition 2. Let $X_h = Y_h - \mathbb{E}Y_h$ and $T = \sum_h Y_h$. The triangle inequality in L^2 gives

$$\mathrm{sd}(T) = \left\| \sum_h X_h \right\|_2 \leq \sum_h \|X_h\|_2 = \sum_h \sigma_h.$$

Under independence, $\mathrm{sd}_\perp(T) = \sqrt{\sum_h \sigma_h^2}$, which proves the first inequality in Eq. 1. Cauchy–Schwarz gives $\sum_h \sigma_h \leq \sqrt{H}\sqrt{\sum_h \sigma_h^2}$, proving the second. Equality in the L^2 triangle inequality requires the centered X_h to be nonnegative scalar multiples of a common random variable almost surely; equal standard deviations alone do not ensure this for fixed marginal shapes. The universal bound is sharp, for example, when the marginals are location–scale transforms of one standardized law and are coupled through the same draw. Proposition 1(a) applies throughout because the marginals remain fixed. □

The bound is on aggregate spread relative to the independent coupling, not on the score and not on ratios between arbitrary copulas. A per-position score is constant over the family, so every attainable spread is invisible to it.

Corollary 1. The pinball loss summed over horizon positions has the form $\sum_h s_h(\hat{P}_h, y_h)$ with s_h the pinball loss at position h (Eq. 5), hence is per-position in the sense of Definition 1. Proposition 1(a) then gives that it is constant over the copulas joining a fixed set of marginals. A function constant on that set has zero directional derivative along copula-only perturbations, so the objective provides no signal selecting the true copula. An update in a shared parameterization can still alter the copula as a side effect of improving the marginals. □

Corollary 2. With $P \oplus v$ the law of $Y + v$ for $Y \sim P$, and $g(0) = 0$, evaluate M_g at the factual plan: $M_g(a^\circ) = M(a^\circ) \oplus g(a^\circ - a^\circ) = M(a^\circ) \oplus g(0) = M(a^\circ) \oplus 0 = M(a^\circ)$. By Proposition 1(b), M and M_g receive identical scores from every factual-plan evaluation. Their contrasts differ: for a functional ϕ that is additive under translation, $\tau_\phi^{M_g}(a', a'') - \tau_\phi^{M}(a', a'') = \phi(g(a' - a^\circ)) - \phi(g(a'' - a^\circ))$. The maps $\mathcal{D}_{a^\circ} \to \mathbb{R}^H$ vanishing at 0 form a linear space of dimension $H \cdot (|\mathcal{A}|^H - 1)$ when $\mathcal{A}$ is finite—one free vector in $\mathbb{R}^H$ per non-factual plan—and an infinite-dimensional space when $\mathcal{A}$ is continuous. The achievable discrepancy ranges over the image of that space under ϕ, subject to ϕ separating the images. The translation-additivity of ϕ is a hypothesis, not a consequence: without it the displayed identity becomes an inequality and only the indistinguishability half survives. □

What these do and do not establish. Each is immediate, and that is the point: the blindness is structural rather than a deficiency of any particular metric, so it cannot be repaired by choosing a better per-position score. What the propositions do *not* establish is that the unconstrained quantities are *large*; that is an empirical question, answered in §4 for both blind spots and instantiated across models in §6.1.

C Why Multivariate Proper Scores Do Not Repair Part (a)

The energy and variogram scores (Eqs. 7, 8) are *not* per-position, so part (a) does not apply, and the energy score—being strictly proper for the joint—uniquely favors the true joint distribution in expectation (the variogram score is proper but not strictly so, and is blind to a common shift of location). That population statement does not guarantee a realized-sample ranking: each is a sum of a marginal and a dependence term, and if the marginal term dominates empirically the score inherits Dimension 1's ordering without being invariant to the copula. §6.1 finds exactly that, and measures

how far the two terms are apart. The fix is therefore not a better multivariate score used the same way but a target that isolates the dependence, which is what Dimension 3's aggregate does.

D DIMENSIONS 1–3 AND ROBUSTNESS

Table 2: **Marginal accuracy and action-transition accuracy** (Rossmann; five seeds). $\mathcal{T}$ is the eight horizon positions where the national promotion flag changes.

Config	WQL ↓	MASE ↓	MASE$_{\mathcal{T}}$ ↓
BC-AR	**.0801**±.0025	**.295**±.008	.386±.028
BD-AR	.2312±.0055	.740±.016	1.311±.049
BD-AR+PA	.1050±.0221	.361±.067	.457±.043
BD-OS	.0945±.0018	.340±.005	.426±.012
BD-OS+PA	.0936±.0097	.333±.030	.482±.034
Chronos-2	.0928	.336	**.380**

The transition subset reorders the two leaders, but remains a factual re-scoring rather than an intervention. For Dimension 2, ES and VS_1 closely follow WQL: Spearman $+.886$ and $+1.000$. Chronos-2, whose columns are sampled independently, has the best ES; BD-JOINT is the only row with a detectable shuffle gap and ranks in the bottom three by both scores. For Dimension 3, BD-JOINT attains total coverage .669±.127 at WQL .1172±.0070, while BC-AR has coverage .288±.141 at WQL .0801±.0025. This is the controlled break between marginal and aggregate quality.

Matched output-family control. BD-DIAG removes only BD-JOINT's shared factors. Its WQL is .1016±.0047, VS_1 .0512±.0033, total coverage .283±.055, and relative VS_1 shuffle gap $+.0001$±.0009. Against BD-JOINT's coverage .669±.127 and gap $+.0166$±.0023, this isolates the within-Gaussian improvement to the low-rank factors rather than native sampling or likelihood family. Per-step 80% coverage is similar for BD-JOINT and BD-DIAG (.766±.029 versus .745±.030), while horizon-total coverage differs sharply (.669±.127 versus .283±.055).

D.1 TEMPORAL ROBUSTNESS

Table 3: **Three Rossmann cutoffs.** Trained rows use five seeds at $w = 0$ and three otherwise.

w	Config	WQL	VS_1	$\mathrm{CRPS}_{\mathrm{sum}}$	Total cov.
0	BC-AR	.0801	.0361	9.95	.288
	BD-JOINT	.1172	.0600	11.88	.669
	Chronos-2	.0928	.0441	6.75	.551
4	BC-AR	.0922	.0416	9.88	.295
	BD-JOINT	.1122	.0455	13.76	.615
	Chronos-2	.0942	.0423	10.54	.456
9	BC-AR	.0885	.0384	14.97	.286
	BD-JOINT	.1186	.1273	17.97	.556
	Chronos-2	.0866	.0388	9.23	.282

Exact scalar ranks move, especially at noisy window 9. BC-AR remains the strongest trained WQL row and BD-JOINT the strongest trained aggregate-coverage row at every cutoff.

D.2 CROSS-DIMENSION CORRELATIONS

Table 4: Spearman correlations over the six configurations preceding BD-JOINT. Coefficients are descriptive, not population estimates.

Pair	ρ
WQL–VS_1	$+1.000$
WQL–ES	$+.886$
WQL–$\mathrm{CRPS}_{\mathrm{sum}}$	$+.943$
WQL–absolute coverage error	$+.943$
WQL–CATE error	$+.143$
ES–CATE error	$-.314$
CATE error–matched rank	$+.429$
CATE error–Day-1 empirical response	$+.638$
Matched rank–Day-1 empirical response	$+.812$

Including BD-DIAG gives ten-row forecast sensitivities $+.988$ and $-.164$ for WQL–VS_1 and WQL–absolute coverage error; WQL–CATE remains $+.300$ over the nine rows with CATE. Coverage error uses nominal .80.

E Interventional Evidence

Table 5: Semi-synthetic reference cell ($R^2 = .621$, five seeds; Chronos-2 frozen).

Config	CATE -2σ	CATE -4σ	Log MAE -2σ	Log MAE -4σ	nRMSE
BC-AR	.079±.027	**.161**±.087	**.064**±.012	**.130**±.034	**.191**±.005
BD-AR	.117±.027	.172±.053	.135±.020	.178±.031	.448±.023
BD-AR+PA	.097±.046	.279±.140	.079±.023	.172±.057	.291±.081
BD-OS	.121±.037	.349±.042	.098±.014	.239±.019	.230±.002
BD-OS+PA	**.068**±.017	.211±.085	.069±.009	.169±.043	.198±.013
Chronos-2	.208	.694	.156	.335	.214

E.1 External architecture and causal-sequence coverage

Table 6: External coverage on the Rossmann reference window/cell. DeepAR and cRMSN are mean ± five-seed s.d.; frozen Moirai carries one run.

Model	WQL ↓	VS_1 ↓	Total cov.	CATE -2σ ↓
DeepAR (2.636M)	.1066±.0114	.0518±.0085	.613±.043	.099±.012
Moirai (frozen)	.3344	.3199	.955	.256
cRMSN (2.692M)	—	—	—	.0609±.0051

DeepAR is a capacity-matched causal autoregressor with known-future covariates and joint Student-t paths; frozen Moirai uses the same metric implementation. Both supply native samples, which avoid the inverse-grid clipping artifact; like every row, their total coverage is read against nominal .80. Their sharply different marginal accuracy and aggregate coverage broaden Dimensions 1–3; the ten-row forecast sensitivity above reports their effect rather than leaving them outside every correlation. **Moirai here is Moirai-1.1-R-small at** 13.8**M parameters**, $8.6\times$ smaller than Chronos-2 and capacity-matched to neither it nor the 2.64M in-house rows. Its weak marginal accuracy should therefore be read as the absolute level of one small frozen checkpoint, not as evidence about the Moirai family or about cross-variable architectures generally. We include it because the cross-architecture adaptation claim of §E.2 is a *within-model* contrast against this same frozen reference, which its absolute level does not affect. The adaptation appendix additionally runs Moirai-base at 91.4M, near Chronos-2's 119.5M scale; it is not part of this full-protocol table. cRMSN is a continuous-treatment RMSN adaptation: recurrent Gaussian numerator/denominator treatment densities produce clipped cumulative stabilized weights, followed by a weighted recurrent outcome model. It keeps the exact real-valued action rather than discretizing it. At strong confounding its effective-sample fraction is .246±.017 with 3.9% clipping. It is a deterministic median point rollout, so dashes are intentional and no uncertainty claim is made.

A hypothesis for its confounding trend. cRMSN's relative CATE improves as R^2 rises even though its effective-sample fraction falls. One possible explanation is a changing bias–variance tradeoff: near-random logging makes the weights nearly uniform, leaving cRMSN's weaker factual outcome model (1613±149 RMSE against DeepAR's 1154±72) dominant, whereas stronger confounding gives weighting more bias to remove. We did not identify that mechanism: the logging policy and observed action signal change together, and isolating it would require a weighted arm with the outcome model held fixed. We therefore treat the trend as descriptive rather than evidence that stronger confounding helps weighting. Its mean CATE is slightly below the best forecasting row (.0609±.0051 versus .0677±.0167), but the seed spreads overlap; we read it as competitive, not as a statistically established win. Its factual RMSE is nevertheless worse (1613±149 versus DeepAR's 1154±72), directly separating causal adjustment from factual forecast accuracy. Across $R^2 = .001, .152, .621$, cRMSN CATE falls .0876, .0762, .0609 even though weighting is least efficient in the last cell.

Stochastic rollout stress test. We also sample the outcome decoder's Gaussian recursively and score the resulting paths on the same semi-synthetic hosts. Total coverage@80 is only .142±.066, .173±.113, .160±.057 across the three R^2 cells. Fourteen of fifteen runs have VS_1

between .067 and .223, while one middle-cell seed reaches 528.8. Thus cRMSN can be passed through Dimensions 1–3, but this rollout is too underdispersed and unstable to serve as a competitive common-protocol causal baseline.

E.2 Cross-architecture adaptation

Table 7: Moirai frozen and fine-tuned (small: five fine-tuning seeds; base: three).

R^2	Arm	Factual RMSE	CATE error	Log MAE	$\lvert\hat{\beta}-\beta\rvert$
Moirai-small, 13.8M					
.001	frozen	2763.9	.285	.237	1.944
	tuned	2213.9±55.5	.235±.022	.212±.010	1.498±.176
.152	frozen	2785.0	.273	.226	1.947
	tuned	2130.5±20.5	.182±.020	.207±.020	.822±.222
.621	frozen	2798.6	.256	.206	1.938
	tuned	1873.1±66.2	.135±.063	.243±.060	.361±.164
Moirai-base, 91.4M					
.001	frozen	2929.6	.282	.235	1.900
	tuned	2183.3±104.6	.229±.004	.189±.005	1.555±.051
.152	frozen	2989.9	.271	.224	1.900
	tuned	2462.1±139.7	.133±.020	.184±.022	.569±.239
.621	frozen	3086.2	.255	.205	1.912
	tuned	2118.8±179.1	.080±.003	.270±.068	.369±.031

At $R^2 = .621$, paired store bootstraps put the fine-tuned-minus-frozen CATE change in $[-.129, -.113]$ and Log-MAE change in $[+.032, +.041]$. These intervals condition on the five trained models. Across retraining seeds the aggregate CATE change is $-.121\pm.063$, whereas the Log-MAE mean change $+.0366$ is smaller than its .0602 seed s.d. We therefore treat the latter as an opposite movement of the five-seed mean, not a retraining-robust sign. Moirai-base closes most of the scale gap to Chronos-2 and preserves every aggregate-CATE direction. At $R^2 = .621$, its paired-store intervals are $[-.184, -.168]$ for CATE and $[+.061, +.070]$ for Log-MAE; the latter's $+.065$ mean change is again comparable to its .068 across-seed s.d. Thus parameter count cannot by itself explain the architecture contrast, while the functional split remains conditional on retraining.

E.3 Reference rows

The action-blind no-response baseline is the true effect magnitude: .2542 at -2σ and .5957 at -4σ. The mechanism is deterministic, so an oracle handed the correct functional form and true baseline has zero error. Replacing the true baseline with a seasonal-naive estimate yields CATE errors .0126 and .0275, about 5% of the no-response baseline. Zero is the oracle lower bound; the seasonal-naive values are diagnostic references, not deployable competitors, because learned models must infer the response form.

E.4 Second dataset

The same elasticity injection was rebuilt on a low-zero-rate Favorita subpanel of 572 series. CATE error as a percentage of that cell's no-response baseline is BC-AR 30.3%, BD-OS 70.3%, BD-AR+PA 86.3%, BD-OS+PA 87.5%, Chronos-2 89.7%, and BD-AR 210.3%. The decoder grouping preserves the Rossmann direction (one-shot 78.9% versus autoregressive 109.0%), unlike the mechanism change below.

E.5 Real-data probes

Rossmann's promotion calendar is national, so these are associations rather than identified effects. The strongest compact readings are: matched on/off rank correlation is .940±.007 for BC-AR, .925±.014 for BD-OS and .658 for Chronos-2; relative mean bias is $-.074\pm.050$, $-.234\pm.031$ and $-.347$. On a five-day spell, BC-AR and BD-OS reproduce the empirical first-day peak and decay, whereas Chronos-2 reaches .374 of the empirical first-day response and has a nearly flat slope

$-.003$ against $-.0786$. These probes can reject implausible behaviour but cannot certify causal correctness.

F Simulator Construction and Sensitivity

The semi-synthetic outcome preserves each public series' base path and injects

$$\begin{aligned} Q_t^{\text{sim}} &= Q_t^{\text{base}} M_t, & M_t &= \exp(\beta \Delta p_t - \rho I_t), \\ I_t &= \delta I_{t-1} + (M_{t-1} - 1) o_{t-1}, & \Delta p_t &= -\lambda \kappa_i s_t + \sigma_{\xi,i} \xi_t. \end{aligned} \tag{13}$$

The behaviour-policy gains are calibrated per series; all policies remain stochastic. The design gates are: (1) the injected action is absent from the real data; (2) confounding is tunable and reported by realized R^2; (3) $\rho > 0$ creates carryover; (4) no extra outcome noise is added; (5) models are retrained on the injected panel; (6) marginal-rank preservation and nuisance-action overlap are checked before scoring; and (7) conclusions are repeated under a second response form. The model never observes Q^{base}, (β, ρ, δ) or the policy.

F.1 Second response form

The second mechanism remains multiplicative and level-free but replaces the linear response and geometric state with

$$g_i(x) = s_i \tanh(x/s_i), \qquad I_t = \sum_{k=1}^{3} w_k (M_{t-k} - 1) o_{t-k}. \tag{14}$$

The kernel mass matches the elasticity family and $g_i'(0) = 1$. At matched confounding ($R^2 = .626$ versus .621), autoregressive decoding scores 34.1% of the no-response baseline and one-shot 60.4%; under elasticity the ordering is one-shot 31.7% versus autoregressive 42.1%. The reversal holds on every seed. Boundedness and kernel shape change jointly, so the experiment establishes mechanism-form sensitivity without isolating either factor.

F.2 Plan menu

The original in-support menu is degenerate: every row has Kendall $\tau \geq .9946$ and zero regret because one sustained arm dominates. Adding one deeper out-of-support action makes the metric discriminate sharply, especially under the saturating family. We therefore use plan ranking only as an extrapolation diagnostic, not as evidence that realistic in-support alternatives are ordered correctly.

F.3 Architecture cell and provenance

Table 8: Architecture taxonomy. BC-AR cleanly isolates a controlled causal, cross-variable comparison.

	Univariate target	Attends across variables
Causal rollout	DeepAR, TimesFM	**BC-AR**
Bidirectional one-shot	PatchTST, DLinear	iTransformer, Moirai, Chronos-2

Fully-causal and next-token ablations were excluded because their exploratory runs lacked the same checkpoint provenance as the reported rows. The model set therefore populates the mask axis at its ends, not continuously.

F.4 FROZEN-WEIGHT DECODER SWEEP

Table 9: Selected decoder blocks on frozen Chronos-2. Only decoding changes.

Metric	$P_{\text{dec}} = 48$	12	1
WQL ↓	**.0928**	.1066	.1266
ES (10^3) ↓	**4.38**	5.16	6.18
CRPS_{sum} ↓	6,752	**6,505**	18,222
Total coverage	.551	**.655**	.138
Log-contrast MAE ↓	.157	.138	.101
$\lvert\hat{\beta} - \beta\rvert$ ↓	.389	.501	1.030

WQL worsens significantly as the block shrinks, while every reported aggregate metric improves at the interior $P_{\text{dec}} = 12$ versus 48 under paired store bootstraps. The decoder is therefore part of the evaluated simulator, not merely an inference implementation detail. For Dimension 4, replacing sampled block feedback by the conditional median leaves the result essentially unchanged: at $P_{\text{dec}} = 12$, Log-MAE is .1376 versus .1378 (paired delta CI $[-.0001, +.0004]$); at $P_{\text{dec}} = 4, 1$, median feedback gives $.0889, .1032$ versus the published $.089, .101$. Repeated action re-conditioning, rather than stochastic sample feedback, drives this interventional effect; the ablation does not alter the sampled-path reading of Dimensions 1–3.

F.5 FORECASTING REPLICATION ON FAVORITA

On Favorita, the marginal-dominance finding repeats: both multivariate proper scores remain much more sensitive to marginal accuracy than to the dependence differences induced by decoding. Absolute HSTU levels are conservative because Rossmann hyperparameters were transferred without retuning.

G POSITIONING AND LIMITATIONS

RMSN (Lim et al., 2018), CRN (Bica et al., 2020), and the Causal Transformer (Melnychuk et al., 2022) learn potential outcomes under treatment sequences and primarily report counterfactual outcome errors over the horizon. The protocol here is complementary: it audits the joint sampled path, aggregate calibration, decoder sensitivity and response-form dependence. We evaluate cRMSN, a continuous-treatment RMSN adaptation, as a point potential-outcome comparator and separately stress-test its Gaussian trajectory rollout on the semi-synthetic hosts; CRN and the Causal Transformer remain unevaluated.

World models and decision-facing evaluation. World-model evaluations likewise distinguish predictive accuracy from functional utility: Vafa et al. (2024) use task-specific probes to expose incoherent latent worlds, while Yu et al. (2026) argue for decision-facing and interventional tests. We instantiate that distinction for time-series forecasters whose action channel is weak relative to seasonality and make the blind spot explicit through Proposition 1.

Decoding and structure-aware forecasting metrics. Recursive, direct and one-shot multi-step strategies form a classical continuum (Bontempi et al., 2012; Ben Taieb & Hyndman, 2012; Green et al., 2024). Recent structure-aware metrics assess shape or trend (Phungtua-Eng & Yamamoto, 2026; Weerasekara & Wickramanayake, 2026; Qiao et al., 2026), but when applied to one point path they remain functions of marginal summaries and do not test a sampled trajectory's copula or its response under a changed action. Our frozen-weight sweep instead treats decoding as part of the simulator and scores both aggregate and interventional consequences.

Marketplace simulation and semi-synthetic benchmarks. RecSim and RecSim NG (Ie et al., 2019; Mladenov et al., 2021) provide configurable ecosystem simulators; RetailSynth (Xia et al., 2023) is the closest retail-domain construction. Those systems are hand-specified or evaluated through downstream policy behavior. Here the simulator is a learned forecaster, and the RealCause-style semi-synthetic layer (Neal et al., 2020) supplies exact sequential counterfactuals while retaining real base trajectories.

G.1 LIMITATIONS

1. **Breadth.** The main correlations use six selected configurations, five sharing one backbone. Capacity-matched DeepAR and frozen Moirai broaden four-dimensional coverage; the matched BD-DIAG ablation raises forecast-only sensitivity to ten rows while leaving WQL–VS_1 strong and its relation to calibration error weak. Stochastic cRMSN covers Dimensions 1–3 only on its own semi-synthetic hosts, not the common real-data protocol. Both datasets are daily retail panels; Favorita coverage is partial.
2. **Uncertainty.** Headline and Moirai-small cells use five seeds, but Moirai-base, Chronos-2 adaptation, later cutoffs and mechanism replications use three. Store bootstraps condition on trained models and do not replace training-seed variation; specifically, both Moirai sizes' positive mean Log-MAE changes at $R^2 = .621$ are separated conditionally over stores but not from their across-seed spreads.
3. **Identification.** Exact interventional claims come from the injected mechanisms. Real promotion probes are observational associations because action timing is shared across stores.
4. **Mechanism sensitivity.** The ordering reversal changes boundedness and carryover-kernel shape together. A stable generator isolating those factors was not found.
5. **Comparator gap.** cRMSN supplies a learned sequential-causal reference for one cell, but its Gaussian path extension undercovers and has one unstable seed; no causal baseline covers the common real-data protocol. Its point CATE advantage is not established beyond seed variation. CRN and the Causal Transformer remain unevaluated.
6. **Scope of conclusions.** Proper-score marginal dominance, weak cross-object correlations and decoder rankings are empirical statements about this model set. The propositions establish blindness, not population-level independence.

Additional generator diagnostics, full metric matrices, timing tables and withdrawn-claim history were retained during analysis but are not needed to support the claims in this compact appendix.