WHEN TIME SERIES FOUNDATION MODELS FAIL UNDER CROSS-CHANNEL DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariate time series foundation models (TSFMs) can access related series and covariates, but whether their forecasts reliably reflect the predictive value of cross-channel information remains unclear. We examine how Chronos-2, TimesFM-3, and TiRex-2 use cross-channel information through controlled synthetic experiments and real-world forecasting tasks. Our analysis covers cross-channel temporal relationships, dependence structures among observed channels, and the effects of known-future conditioning and covariate reliability. Useful cross-channel relationships are not consistently reflected in forecasts, while redundant or conditionally irrelevant inputs can still influence predictions. Statistical baselines fitted to the same observed contexts recover the relationships that the models underuse, indicating that these shortfalls reflect limited use of available information rather than its absence. On real-world tasks, conditioning on each model's own median covariate forecasts increases quantile loss and reduces interval coverage. Propagating covariate uncertainty recovers much of this loss, but not all. Together, these findings reveal model-dependent limitations in exploiting temporal dependencies, combining information across channels, and producing reliable forecasts under covariate uncertainty.

1 INTRODUCTION

Related time series can provide predictive information beyond what is available in a target's own history, motivating the use of multivariate forecasting. Yet it remains unclear when and how this information should be exploited. Some approaches explicitly model cross-channel dependencies (Zhang & Yan, 2023; Liu et al., 2024a; Wang et al., 2024), while channel-independent approaches that process each series separately with shared parameters have also achieved strong forecasting performance (Nie et al., 2023). Prior analyses suggest a trade-off between modeling capacity and robustness to distribution shifts (Han et al., 2023). These lines of work reflect a common challenge in multivariate forecasting: how to translate the information available across related series into improved predictions.

Early time series foundation models (TSFMs) were largely developed around univariate forecasting (Garza et al., 2024; Goswami et al., 2024; Liu et al., 2024b; 2025; Ansari et al., 2024; Das et al., 2024). Other approaches subsequently incorporated covariates into pretrained forecasters through modular adaptation, in-context learning or cross-channel attention (Arango et al., 2025; Auer et al., 2025; Potosnak et al., 2026). More recent TSFMs have expanded native support for multivariate forecasting and historical or known-future covariates (Ansari et al., 2025; Jain & Sen, 2026; Podest et al., 2026; Meyer et al., 2026; Yamaguchi et al., 2026). This expansion raises the prospect of using pretrained models not only to forecast individual series, but also to exploit relationships among them in new tasks. Yet the ability to accept cross-channel inputs does not mean that a model uses them in accordance with their predictive value.

Assessing this alignment is challenging because cross-channel relationships vary in both their temporal structure and their predictive value. A signal in one series may anticipate a later change in the target (Zhao & Shen, 2024), while several series may provide complementary evidence about its future (Taga et al., 2025; Stock & Watson, 2002). Other inputs may instead repeat information already available or contribute no additional predictive information (Guyon & Elisseeff, 2003; Lim et al., 2020). Moreover, future covariate values may themselves be estimates, so their availability does not

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.

guarantee their reliability (Hyndman & Athanasopoulos, 2021). These cases make it necessary to distinguish the predictive information provided by an input from the influence that input has on a model's forecast: useful information may be underused, while redundant, irrelevant, or unreliable inputs may still affect predictions. We therefore ask: *does a model's use of cross-channel inputs align with the predictive value of those inputs?*

Forecasting benchmarks provide broad assessments of model performance across diverse tasks (Aksu et al., 2024; Cohen et al., 2025; Shchur et al., 2026), but aggregate accuracy alone offers limited insight into how specific cross-channel relationships contribute to a forecast. The empirical benefits of cross-channel modeling also depend on the evaluation setting: TSMixer reports limited gains from feature mixing on common long-term forecasting benchmarks, but larger gains from cross-variate information on the M5 demand forecasting task (Chen et al., 2023). More fundamentally, an improvement after adding related series does not establish which predictive relationships the model exploited. Conversely, an absence of improvement does not distinguish a lack of useful information from a failure to use the information available. These ambiguities motivate evaluations that examine how model forecasts respond to cross-channel inputs under explicitly specified predictive relationships, complementing comparisons of overall forecasting performance.

Prior diagnostics examine simple target–covariate mappings (Berthelier et al., 2026), as well as leading indicators and robustness to additional independent channels (Ghoroghchian et al., 2026). Our analysis instead centers on an alignment-based diagnostic: whether the influence of a cross-channel input on the forecast matches the predictive information it contributes. This diagnostic distinguishes underuse of informative inputs from sensitivity to inputs that add no predictive information.

We examine four complementary aspects of cross-channel information use. First, responding to a temporally aligned covariate does not establish that a model can connect an earlier signal to the appropriate later target response (Zhao & Shen, 2024). This motivates **(RQ1)**: *How do temporal relationships affect TSFMs' use of cross-channel information?* Second, multiple observations may provide complementary evidence about a shared predictive state (Stock & Watson, 2002), whereas duplicates or irrelevant channels add no predictive information beyond the informative observations. This distinction leads to **(RQ2)**: *How do TSFMs use cross-channel information under different dependence structures?* A third consideration is that models may use historical information differently across input and conditioning configurations, even when the available predictive information is unchanged. Accordingly, **(RQ3)** asks: *How does the conditioning interface affect TSFMs' use of cross-channel information?* Finally, future covariates may themselves vary in reliability, particularly when their future values are estimated rather than observed (Hyndman & Athanasopoulos, 2021). This motivates **(RQ4)**: *How does covariate reliability affect TSFMs' use of cross-channel information?*

We evaluate frozen Chronos-2, TimesFM-3, and TiRex-2 on controlled synthetic processes with explicitly specified cross-channel relationships. The generating processes establish the predictive value of the supplied inputs, while matched input comparisons measure their influence on forecasts. Oracle references use knowledge of the data-generating process, whereas statistical baselines fitted separately to the same observed contexts assess achievable performance under explicit model-class assumptions (Appendix A.2). On three real-world tasks, we compare direct multivariate forecasting with conditioning on each model's own covariate forecasts.

2 Temporal covariate event–target response underutilization

A pattern in a covariate series can signal a corresponding change in a target series. For example, a brief burst of rainfall can signal a streamflow response that rises, peaks, and recedes (Weaver, 2003). To address **RQ1**, we examine whether TSFMs use such cues to produce the corresponding target response across different event–response patterns, and how this use changes as the temporal separation between the trigger and response increases. We first evaluate responses across different event–response patterns, and then vary the trigger–response lag while holding the target-response position fixed within each sweep.

2.1 Event–Response Pattern Combinations

Trigger Patterns, Target Responses, and Historical Examples. We cross three auxiliary-trigger patterns—a positive spike, a negative dip, and a biphasic (positive–negative) pulse—with positive,

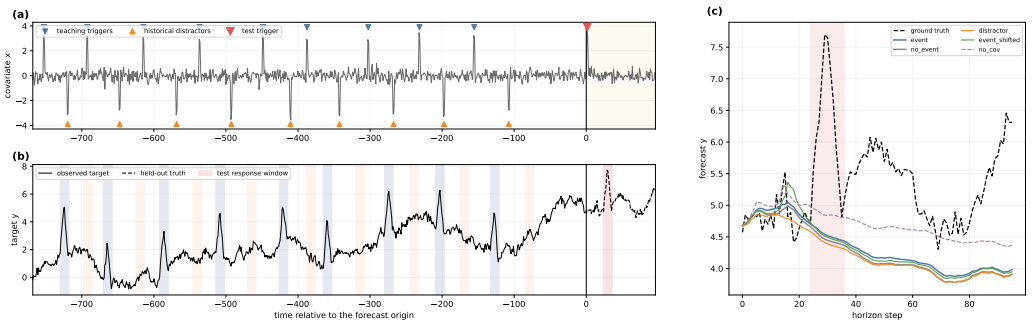

Figure 1: A spike trigger with a positive target response at $\tau = 24$. (a) Covariate triggers and matched distractors. (b) Delayed target responses, with the test response fixed in the forecast horizon and the test trigger anchored at the forecast origin. (c) Counterfactual TimesFM-3 forecasts for a single window. Aggregate response gains are reported in Figure 2.

negative, and biphasic target responses, yielding nine event–response combinations. Each condition contains nine historical teaching pairs that repeat its single trigger–response mapping, together with sign-reversed distractor patterns that produce no target response. Figure 1 illustrates a spike trigger with a positive target response. The kernels and generation settings are specified in Appendix B.1.

Varying the Delay between Events and Responses. We denote the target-response anchor by t_{resp}, measured relative to the forecast origin $t = 0$, and place the trigger at $t_{\text{trig}} = t_{\text{resp}} - \tau$. We hold t_{resp} fixed within each lag sweep, so that changing the lag does not move the target response within the forecast horizon. At $\tau = 0$, the trigger and response anchors coincide. The covariate history and future continuation are supplied through the same conditioning interface at every lag. The trigger lies in the known-future segment for $\tau < t_{\text{resp}}$, at the forecast origin for $\tau = t_{\text{resp}}$, and in history for $\tau > t_{\text{resp}}$. We compare lags within the future and historical segments separately to keep the trigger's input segment fixed. The main experiments use $t_{\text{resp}} = 24$, with additional response-position controls reported in Appendix B.4.

Measuring Event-Induced Target Responses. We subtract the no-event forecast from the event forecast for paired inputs that differ only in the presence of the test trigger. For each combination c and seed s, averaging these differences over windows gives the response profile $\overline{D}_{c,s}$. With Q_c denoting the true injected response at the fixed anchor, we measure its signed strength using projection gain:

$$G_{c,s} = \frac{\langle \overline{D}_{c,s}, Q_c \rangle}{\langle Q_c, Q_c \rangle}. \tag{1}$$

The ideal response has unit gain; negative gain indicates a component opposite to the expected response pattern. We compute gain within each combination and seed, then average equally across the nine combinations before reporting the mean and SEM across five seeds (Figure 2). Timing and profile-shape measurements are provided as supplementary diagnostics in Appendix B.2.

2.2 Target Responses across Patterns and Delays

Baseline Responses across Event–Response Patterns. At $\tau = 0$, Chronos-2 and TimesFM-3 produce positive projection gains across all nine combinations, indicating a response component aligned with the expected target pattern. Their responses are attenuated, and their strength depends on the event–response mapping. Both models respond more weakly to biphasic triggers across the target-response families, while producing pronounced biphasic target responses when the trigger is a spike or dip. TiRex-2 produces weak event-induced responses across all nine combinations even at zero lag. Chronos-2 and TimesFM-3 respond preferentially to event triggers, with consistently weaker absolute projection gains for sign-reversed distractors (Appendix B.2).

Increasing Delay within the Known-Future Segment. TimesFM-3's gain drops sharply at short nonzero lags in all nine combinations. This attenuation also recurs across the tested response anchors

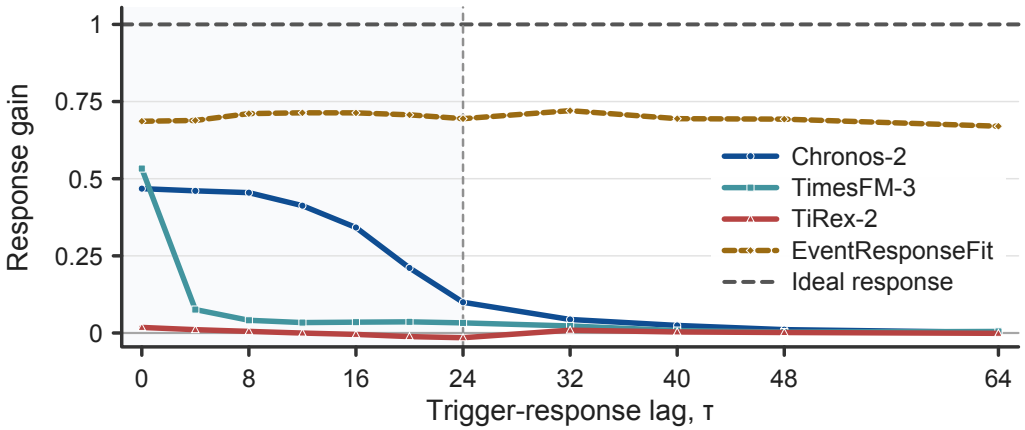

Figure 2: Response gain, averaged equally over nine event–response combinations (mean ± SEM across five seeds). Black dashed lines mark unit gain and the trigger at the forecast origin ($\tau = 24$).

in the spike-to-positive control (Appendix B.4). Chronos-2 largely maintains its average gain over the shortest lags, but its gain decreases as the lag grows further.

Increasing Delay within Observed History. At larger lags within the historical segment, the responses of Chronos-2 and TimesFM-3 become weak despite the continued availability of predictive information. A simple event-response regression fitted to each window's observed target and covariate histories (EventResponseFit; Appendix B.3) retains substantial gains across lags. Responsiveness at short lags therefore does not ensure that a model preserves its response to the same predictive relationship at larger lags. Combination-specific results and controls are reported in Appendix B.2.

3 Structural Dependency Failures

To investigate **RQ2**, we examine how TSFMs use predictive information under different cross-channel dependence structures. We study a controlled single-factor setting inspired by dynamic factor models, which describe multiple series through common dynamic factors and series-specific components (Stock & Watson, 2016). Factor-based forecasting uses these observations to estimate predictive factors (Stock & Watson, 2002). We first provide the underlying state directly and then replace it with noisy observations.

3.1 Direct and Noisy Observations of a Shared Latent State

We generate a shared latent process and a target series as

$$z_t = \rho z_{t-1} + \sqrt{1-\rho^2}\, u_t, \tag{2}$$

$$Y_t = \kappa z_t + \sigma_y \epsilon_t, \tag{3}$$

where u_t and ϵ_t are mutually independent standard Gaussian white-noise processes. We vary latent persistence ρ while holding the target loading and noise level fixed and preserving unit stationary variance for z_t. All comparisons use the same multivariate history interface and matched target windows. Full settings and numerical results are provided in Appendices C.1 and C.2.

Providing the Latent State Directly. We compare target-only forecasting with forecasts that additionally receive the true latent history $z_{1:T}$, without supplying future latent values. We measure the benefit as the MAE reduction relative to the target-only baseline. This condition provides a reference for interpreting performance when the state must instead be inferred from noisy observations.

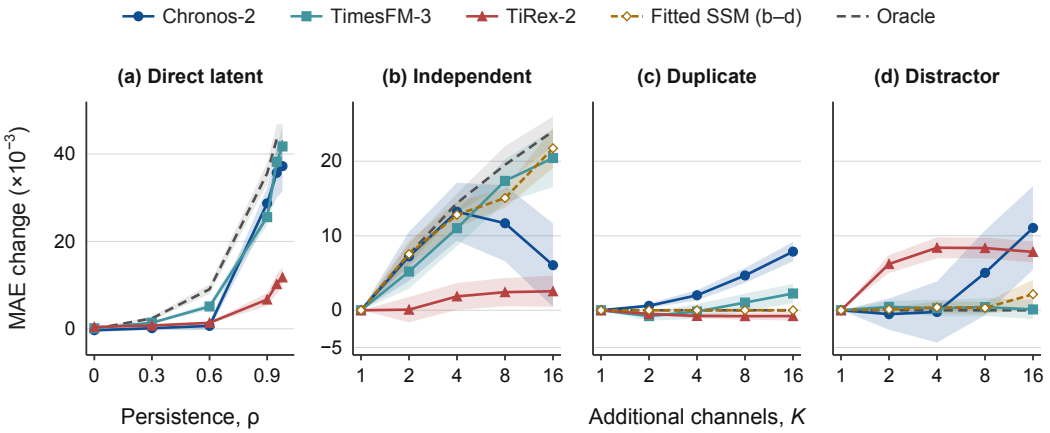

Figure 3: MAE reductions with (a) true latent history, relative to target-only input, and (b) independent observations, relative to noisy $K = 1$; (c–d) MAE increases with duplicates and distractors, relative to noisy $K = 1$. Each method uses its own reference. Panel (a) varies ρ; (b–d) use $\rho = 0.95$ and $\lambda = 0.75$. Curves show means, with shading for $\pm$ paired SEM over five seeds; dashed gray lines denote the oracle. Fitted SSM is shown in (b–d), with zero duplicate penalty by construction.

Adding Noisy Observation Channels of the Same Latent State. We replace direct latent access with auxiliary observations

$$X_{j,t} = \lambda z_t + \sqrt{1 - \lambda^2}\, \eta_{j,t}, \tag{4}$$

where the η_j are mutually independent AR(1) processes, independent of z and the target noise, with the same persistence and unit stationary variance as z. The fixed loading λ controls the contribution of the shared state, while the normalization preserves unit variance for each observation. The *independent* condition varies the number K of observations, each with its own noise realization. This supplies complementary evidence about the same latent state.

Adding Duplicate or Irrelevant Channels. The *duplicate* condition repeats one noisy observation, while the *distractor* condition retains one informative channel and adds target-irrelevant series with matched temporal statistics. At each K, the three channel compositions have matched channel counts and individual-channel marginal statistics. Neither duplicates nor distractors add predictive information beyond the single informative observation. For all channel-composition comparisons, we measure paired MAE changes relative to each model's noisy single-channel baseline ($K = 1$). The oracle conditions on exactly the histories available in each condition, providing a reference for the predictive value of additional observations.

3.2 Forecast Benefits and Channel Interference

Benefits of Direct Latent Access. The predictive value of the latent history grows with persistence, and so do the models' gains from direct access (Figure 3a). At $\rho \geq 0.9$, Chronos-2 and TimesFM-3 obtain most of the oracle gain, whereas TiRex-2 obtains about a quarter of it or less. At intermediate persistence ($\rho = 0.6$), TimesFM-3 already obtains part of the available gain, while Chronos-2's gain remains near zero (Table 15a). Because the oracle gain itself varies with ρ, model gains across persistence levels must be interpreted alongside the corresponding oracle gains.

Gains from Additional Noisy Observation Channels. In the persistent setting shown in Figure 3b, TimesFM-3 continues to benefit from additional independent observations, whereas Chronos-2's mean gain declines beyond an intermediate channel count despite increasing oracle utility (Appendix C.4). As a reference for what is achievable from the same finite contexts, we fit a Gaussian state-space model to each window's observed histories (Appendix C.3). Its model class contains the generating process, but its parameters, including the persistence, are estimated from each context alone. This baseline continues to improve as independent observations are added and

approaches the oracle, so Chronos-2's decline is not a necessary consequence of estimation from a 512-step context. TiRex-2 gains little from additional noisy observations, and its modest benefit from direct latent access indicates that this limitation is already visible before multiple observations must be combined.

Interference from Duplicate and Irrelevant Channels. Distractor channels increase the error of TiRex-2 as soon as one is added and that of Chronos-2 at larger distractor counts, although the latter varies considerably across seeds; TimesFM-3 remains nearly unchanged (Figure 3d). The fitted state-space model, which is not told which channels are irrelevant, incurs almost no distractor penalty. Avoiding this interference is therefore achievable from the same finite contexts when the model class matches the generating structure. Exact duplicates add no information beyond the single informative observation, yet Chronos-2's error increases steadily with their number, while the other two models change little (Figure 3c).

Supplementary target-side probes of latent-state accessibility are reported in Appendix G.1.

4 When Known-Future Conditioning Misleads

This section addresses **RQ3** and **RQ4**. To study **RQ3**, we first examine how TSFM forecasts change when cross-channel information is presented through different conditioning interfaces, using settings where the supplied future covariate adds no predictive information beyond the observed histories. We then turn to **RQ4** by examining how forecast behavior changes as future-covariate reliability deteriorates, first in controlled settings and then in real-world tasks where future covariates are themselves predicted. Detailed numerical comparisons for the controlled settings are provided in Appendices D and E.

4.1 When Future Covariates Have No Predictive Information

Lagged Copies of the Covariate and Conditioning Interfaces. We generate the target as

$$Y_t = \beta X_{t-L} + \sqrt{1-\beta^2}\,\epsilon_t, \tag{5}$$

where X_t and ϵ_t are independent standard Gaussian white-noise processes. We vary the relation strength over $\beta \in \{0.4, 0.6, 0.8, 1\}$. At $\beta = 1$, the target is an exact lagged copy of the covariate; smaller positive values introduce target noise while preserving unit marginal target variance. With horizon $H = 8$ and lags $L \in \{8, 16\}$, all source values needed to predict the target horizon lie in the observed history; the future continuation adds no predictive information. We compare MV, which supplies both histories as multivariate channels; PastCov, which supplies the covariate history as a past covariate; and KC, which additionally provides its future continuation. A channel-independent (CI) baseline uses only the target history. The protocols are detailed in Appendix D.1.

Forecasting an Exact Lagged Copy of the Covariate. The noise-free case tests whether models exploit a fully predictive historical relationship without irreducible target uncertainty. Chronos-2 benefits strongly from the covariate history under MV/PastCov, while TimesFM-3 and TiRex-2 remain close to their target-only baselines (Table 1). KC improves the latter two models despite adding no predictive information, whereas its benefit for Chronos-2 is not consistent across lags. We measure whether changing a predictive historical value produces the expected change in the forecast. Let $J_{h,j}$ denote the average forecast change at step h per unit perturbation of the historical source value for target step j. Ideally, only step j responds, with sensitivity β, giving the response matrix $J^\star = \beta I_H$. We measure the mismatch as

$$E_J = \frac{\|J - \beta I_H\|_F}{\|\beta I_H\|_F}. \tag{6}$$

This error captures both incorrect response strength and changes at forecast steps that should remain unaffected. It is zero when the average response matches the ideal pattern; lower values indicate closer agreement (Appendix D.1).

For TimesFM-3 and TiRex-2, KC lowers both MAE and E_J relative to PastCov, improving forecast accuracy alongside the response to predictive historical values. Thus, historical information can be used differently across input and conditioning configurations even when the available predictive information is unchanged. Figure 5 illustrates these differences for a single window.

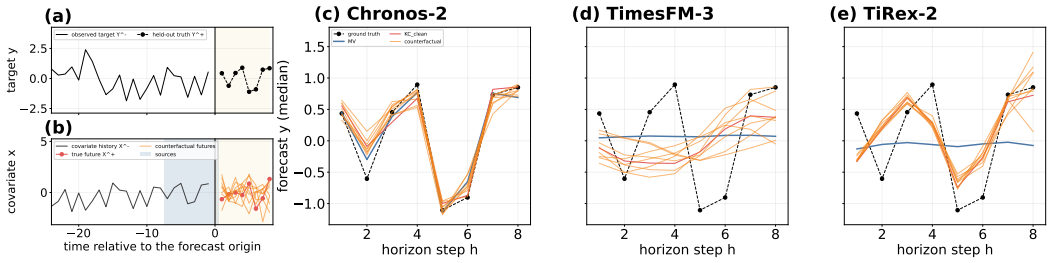

Figure 5: A single-window exact-copy example ($\beta = 1$, $L = 8$). (a–b) Target and covariate series; blue shading marks the historical source values for the forecast horizon. (c–e) Median forecasts under MV, KC with the original future, and KC with eight resampled futures, holding the histories fixed.

Interface Dependence across Relation Strengths. A linear regression whose lag is selected from each window's observed target and covariate histories identifies the true lag in every tested window for $\beta \in \{0.8, 1\}$ and $L \in \{8, 16\}$. It attains near-oracle accuracy and produces identical forecasts with or without the future continuation (Appendix D.4). TSFM accuracy varies with the conditioning interface. Across the noisy-copy settings, KC yields lower mean MAE than PastCov for all three models at both lags, although some gains are small relative to seed-level variation. For TimesFM-3 and TiRex-2, mean gains increase as the relationship strengthens, and KC also lowers E_J relative to PastCov (Appendix D.3). These improvements occur without additional predictive information from future covariates and are accompanied by responses to predictive historical values that more closely match the oracle.

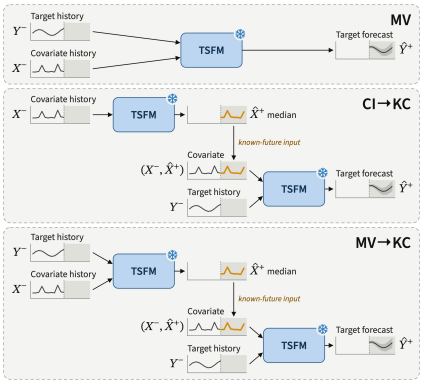

Figure 4: Direct forecasting and median-path reuse. Orange paths are predicted covariates.

Sensitivity to Irrelevant Future Covariate Values. We hold the histories and KC interface fixed and resample the future continuation. Forecast change under future resampling, S_{future}, is the mean pairwise absolute difference between the resulting median forecasts. The oracle is invariant to this resampling, as is the fitted regression in the tested $\beta \in \{0.8, 1\}$, $L \in \{8, 16\}$ settings. All three TSFMs remain sensitive at every tested positive relation strength. At $L = 8$, their mean sensitivity exceeds that of the no-dependence control (Table 20). For TimesFM-3 and TiRex-2, mean sensitivity increases as the historical relationship strengthens. Known-future conditioning can therefore improve accuracy while making forecasts sensitive to inputs that leave the oracle predictive distribution unchanged.

4.2 Changes in Covariate Reliability at Forecast Time

Predictive Covariates and Reliability Degradation. Future covariates supplied through a known-future interface may themselves be forecasts. For example, electricity-price tasks in fev-bench include day-ahead load and generation forecasts as known covariates (Shchur et al., 2026; Lago et al., 2021). Availability at forecast time therefore does not imply error-free values. We study a controlled setting in which the clean covariate and target share a persistent latent state, making clean future covariates predictive of the target. Holding the target process fixed, we degrade the supplied covariate by mixing it with an independent realization having matched temporal statistics. The mixing weight λ controls reliability, from the clean covariate at $\lambda = 1$ to an independent series at $\lambda = 0$.

Degrading Covariate History, Future Values, or Both. Stationary degrades both history and future, Future-Only degrades only the future, and History-Only degrades only the history. Stationary and Future-Only receive exactly the same degraded future values, differing only

Table 1: Known-future conditioning for exact ($\beta = 1$) and noisy ($\beta = 0.8$) lagged copies ($H = 8$). CI uses only the target history. MV/PastCov MAEs match at the displayed precision. S_{future} is mean $\pm$ SEM across five seeds.

		MAE ↓			E_J ↓		
Model	L	CI	MV/PastCov	KC	PastCov	KC	S_{future} ↓
Exact lagged copy ($\beta = 1$)							
Chronos-2	8	0.8037	0.1713	0.1846	0.084	0.092	0.1480 ± 0.0006
	16	0.8014	0.3234	0.3058	0.256	0.222	0.1462 ± 0.0004
TimesFM-3	8	0.8039	0.7900	0.7024	0.981	0.855	0.1645 ± 0.0008
	16	0.8008	0.7898	0.7375	0.985	0.894	0.0851 ± 0.0007
TiRex-2	8	0.8040	0.8013	0.5507	0.997	0.666	0.1626 ± 0.0005
	16	0.8014	0.7999	0.6421	0.999	0.769	0.0746 ± 0.0006
Noisy lagged copy ($\beta = 0.8$)							
Chronos-2	8	0.8014	0.5395	0.5328	0.144	0.062	0.1619 ± 0.0009
	16	0.8027	0.6147	0.5865	0.390	0.247	0.1545 ± 0.0005
TimesFM-3	8	0.8015	0.7965	0.7555	0.991	0.890	0.1141 ± 0.0014
	16	0.8023	0.7990	0.7785	0.994	0.939	0.0531 ± 0.0008
TiRex-2	8	0.8012	0.7992	0.6679	0.997	0.673	0.1285 ± 0.0007
	16	0.8030	0.8022	0.7297	0.999	0.801	0.0613 ± 0.0007

Table 2: Forecast-time covariate reliability shift. Stationary and Future-only share identical degraded future inputs. Δ_{shift} reports their MAE difference (Future-only minus Stationary; mean $\pm$ paired SEM). Clean denotes $\lambda = 1$.

λ	Stationary	Future-only	History-only	Δ_{shift}
Chronos-2 CI: 0.802 Clean: 0.589				
0.75	0.723	0.812	0.679	0.089 ± 0.008
0.50	0.778	0.985	0.739	0.207 ± 0.013
0.25	0.797	1.129	0.778	0.332 ± 0.017
0.00	0.800	1.259	0.808	0.459 ± 0.020
TimesFM-3 CI: 0.797 Clean: 0.580				
0.75	0.707	0.814	0.676	0.107 ± 0.008
0.50	0.765	0.992	0.720	0.227 ± 0.013
0.25	0.792	1.141	0.757	0.349 ± 0.017
0.00	0.796	1.271	0.792	0.475 ± 0.020
TiRex-2 CI: 0.796 Clean: 0.585				
0.75	0.711	0.809	0.678	0.098 ± 0.008
0.50	0.770	0.976	0.721	0.206 ± 0.012
0.25	0.795	1.113	0.765	0.317 ± 0.016
0.00	0.797	1.232	0.798	0.435 ± 0.018

in whether the historical covariate has also been degraded. Their comparison measures how the effect of those future values depends on the preceding history. HISTORY-ONLY evaluates the reverse mismatch. The generation process and degradation conditions are specified in Appendix E.1.

Forecast Accuracy and Coverage under Reliability Mismatch. Under STATIONARY degradation, forecast error approaches the channel-independent baseline as covariate reliability decreases. The same future values produce larger errors when they follow a clean history, while the reverse mismatch is less harmful (Table 2). At matched degraded future values, FUTURE-ONLY also produces narrower prediction intervals and lower empirical coverage than STATIONARY (Table 24). A relationship supported by history can therefore become a source of forecasting risk when future input reliability changes. This comparison characterizes risk under an unannounced change without assuming that the change is identifiable from the available inputs.

4.3 USING THE MODEL'S OWN COVARIATE FORECASTS AS KNOWN-FUTURE INPUTS

Real-world Tasks with Unobserved Future Covariates. Forecasting pipelines may estimate future covariate values and supply these estimates to a target forecaster (Hyndman & Athanasopoulos, 2021). The reliability of these future inputs may differ from that of the observed histories. Related approaches in load forecasting pass temperature scenarios through a demand model (Dimoulkas et al., 2019). We examine this practice on three paired-channel tasks from fev-bench (Shchur et al., 2026): air temperature → electricity load (GFC17), wind speed → active power (KDD), and customer count → sales (Rossmann). We retain the benchmark horizons and rolling forecast origins, restrict inputs to the selected source X and target Y, and cap the context at 512 observations. All routes in the main comparison withhold realized future source values. Dataset selection and task adaptations are detailed in Appendix F.1.

Direct Target Forecasting versus Conditioning on Predicted Covariates. Direct MV forecasting predicts both channels from their observed histories (Figure 4). In median-path MV→KC, we supply the covariate median from this forecast as a known-future input and predict the target again with the same frozen model and unchanged histories, so the second stage receives no new observations. The CI→KC variant instead predicts the covariate from its own history only (Appendix F.2). We also use quantile-path propagation to examine whether any loss of probabilistic quality persists beyond single-path conditioning. We condition separately on nine MV covariate quantile paths, each holding one quantile level across the horizon, and obtain target quantiles by inverting the equally weighted mixture of the resulting target CDFs. This finite-scenario approximation and its sensitivity checks are detailed in Appendix F.3.

Table 3: Conditioning on predicted covariates with a median path (Median) or nine source quantile paths (Paths), compared with direct MV. Bold columns show the Median-to-Paths change: relative SQL change (%) and coverage change in percentage points (pp), computed from unrounded scores.

		SQL ↓				Coverage (%; nominal 80%)			
Dataset	Model	MV	Median	Paths	Δ (%)	MV	Median	Paths	Δ (pp)
GFC17	Chronos-2	0.968	1.062	0.998	**−6.1**	69.4	46.2	59.6	**+13.5**
	TimesFM-3	0.954	0.996	0.954	**−4.2**	71.2	53.2	64.2	**+10.9**
	TiRex-2	0.930	0.942	0.933	**−1.0**	71.4	60.9	62.9	**+2.0**
KDD	Chronos-2	0.828	0.925	0.863	**−6.7**	71.8	46.1	68.7	**+22.6**
	TimesFM-3	0.806	0.931	0.812	**−12.8**	71.7	25.5	65.7	**+40.2**
	TiRex-2	0.735	0.839	0.737	**−12.1**	84.0	30.4	72.1	**+41.6**
Rossmann	Chronos-2	0.492	0.554	0.503	**−9.3**	76.2	36.6	65.3	**+28.8**
	TimesFM-3	0.528	0.596	0.537	**−9.9**	82.3	36.1	72.0	**+35.9**
	TiRex-2	0.570	0.613	0.574	**−6.3**	79.7	46.8	75.3	**+28.5**

Quantile Loss and Interval Coverage with Predicted Covariates. We evaluate scaled quantile loss (SQL) and empirical coverage of the nominal 80% prediction interval. SQL normalizes quantile loss by the seasonal naive error of the target context; scoring and aggregation details are given in Appendix F.1. Conditioning on predicted covariate medians increases SQL and lowers coverage relative to direct MV across all evaluated dataset–model combinations (Table 3). The same deterioration occurs when the source median is obtained through CI (Appendix F.2). These losses in probabilistic quality can occur even when point forecasts improve: on KDD, both median-path routes reduce target MAE for every model while producing narrower intervals with lower coverage. Using the same model in both stages therefore does not by itself preserve probabilistic forecast quality.

Compared with median-path conditioning, quantile-path propagation reduces SQL and moves coverage closer to the nominal level in every evaluated dataset–model combination. Nevertheless, coverage remains below both the nominal level and that of direct MV forecasting. The same qualitative pattern holds under the alternative tail reconstruction (Appendix F.3). As a diagnostic, this comparison shows that accounting for covariate uncertainty recovers a substantial portion of the probabilistic quality lost under single-path conditioning, while a gap relative to direct multivariate forecasting remains.

5 Conclusion

Across the four research questions, we find that cross-channel information can be available without being effectively or reliably used; statistical baselines fitted to the same observed contexts recover several relationships that the models underuse. Depending on the model, TSFMs underuse delayed signals or complementary observations while remaining sensitive to redundant or conditionally irrelevant inputs. The use of historical information can change across conditioning interfaces even without additional predictive information, and forecast-time changes in covariate reliability can compromise both accuracy and probabilistic forecast quality. On real-world tasks, supplying predicted covariates as known-future inputs degrades probabilistic forecasts; accounting for covariate uncertainty recovers a substantial portion of this loss, but not all of it.

These findings motivate evaluating not only whether cross-channel inputs improve forecasts, but also what predictive information they contribute beyond the observations already available. Promising directions include connecting these behavioral findings to architectural and pretraining choices, and developing training and inference procedures that selectively exploit cross-channel information while accounting for uncertainty in predicted covariates.

AI USE STATEMENT

We used generative AI tools to assist with experimental design, code development, experiment execution, interpretation of results, literature search, and the preparation of manuscript text, figures, and tables. Generative AI was not used to generate the initial research idea. The authors reviewed intermediate experimental results at multiple stages of the study and take responsibility for the final content, including all AI-assisted text, code, analyses, and other artifacts.

ETHICS STATEMENT

This study uses controlled synthetic data and existing forecasting benchmarks to examine the reliability of cross-channel forecasting. A potential downstream risk is overconfidence when predicted covariates are supplied as known-future inputs without accounting for their uncertainty. Our analysis documents this risk and shows that propagating covariate uncertainty recovers a substantial portion of the lost probabilistic quality but does not fully close the gap, motivating careful evaluation of probabilistic forecasts in downstream applications.

REPRODUCIBILITY STATEMENT

We document the model versions, experimental settings, random seeds, and evaluation protocols in Appendix A and the experiment-specific appendices. The synthetic data-generating processes, oracle references, controls, baseline-fitting procedures, and detailed numerical results are provided in Appendices B–E. Data preparation and evaluation protocols for the real-data experiments are described in Appendix F.1, and representation-probe procedures are detailed in Appendix G. We plan to release the experimental code during the rebuttal period.

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, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Senetaire, Caner Turkmen, Oleksandr Shchur, Danielle C. Maddix, Michael Bohlke-Schneider, Yuyang Wang, and Syama Sundar Rangapuram. Chronosx: Adapting pretrained time series models with exogenous variables, 2025. URL https://arxiv.org/abs/2503.12107.

Andreas Auer, Raghul Parthipan, Pedro Mercado, Abdul Fatir Ansari, Lorenzo Stella, Bernie Wang, Michael Bohlke-Schneider, and Syama Sundar Rangapuram. Zero-shot time series forecasting with covariates via in-context learning, 2025. URL https://arxiv.org/abs/2506.03128.

Gaspard Berthelier, Mariia Baranova, Andrei-Tiberiu Pantea, Etienne Le Naour, Adrien Petralia, Tahar Nabil, and Themis Palpanas. Investigating simple target-covariate relationships for chronos-2 and tabpfn-ts, 2026. URL https://arxiv.org/abs/2605.12200.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *TMLR*, 2023.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models, 2025. URL https://arxiv.org/abs/2505.14766.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Ilias Dimoulkas, Peyman Mazidi, and Lars Herre. Neural networks for GEFCom2017 probabilistic load forecasting. *International Journal of Forecasting*, 35(4):1409–1423, 2019. doi: 10.1016/j.ijforecast.2018.09.007.

FlorianKnauer and Will Cukierski. Rossmann Store Sales. Kaggle competition, 2015. URL https://www.kaggle.com/competitions/rossmann-store-sales.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1, 2024. URL https://arxiv.org/abs/2310.03589.

Nafiseh Ghoroghchian, Haipeng Zhang, Shuyi Han, Alex Labach, and George Stein. Foundations without fundamentals: Zero-shot blind spots in time series FMs. In *ICML 2026 Workshop on Foundation Models for Structured Data (FMSD)*, Seoul, South Korea, 2026. URL https://openreview.net/forum?id=iIRdd86Xkr.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003. ISSN 1532-4435.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting, 2023. URL https://arxiv.org/abs/2304.05206.

Tao Hong, Jingrui Xie, and Jonathan Black. Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 35(4):1389–1399, 2019. doi: 10.1016/j.ijforecast.2019.02.006.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3 edition, 2021. URL https://otexts.com/fpp3/. Section 7.6: Forecasting with regression.

Ayush Jain and Rajat Sen. Timesfm-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, August 2026. URL https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. ISSN 0306-2619. doi: 10.1016/j.apenergy.2021.116983. URL http://dx.doi.org/10.1016/j.apenergy.2021.116983.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020. URL https://arxiv.org/abs/1912.09363.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting, 2024a. URL https://arxiv.org/abs/2310.06625.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models, 2024b. URL https://arxiv.org/abs/2402.02368.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025.

Lucas Meyer, Claudio Sole, Huikan Xiang, Nicolas Li, Lucas Franceschino, Arnau Quera-Bofarull, Maarten P. Scholl, Joachim Fainberg, and Geoffrey Négiar. t_0: A time-series foundation model for forecasting with context, 2026. URL https://arxiv.org/abs/2609.24559.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex-2: Generalizing tirex to multivariate data and streaming, 2026. URL https://arxiv.org/abs/2607.01204.

Willa Potosnak, Nina Żukowska, Michał Wiliński, Dan Howarth, Ignacy Stępka, Mononito Goswami, and Artur Dubrawski. Mica: Multivariate infini compressive attention for time series forecasting, 2026. URL https://arxiv.org/abs/2604.06473.

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, 2026. URL https://arxiv.org/abs/2509.26468.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002. doi: 10.1198/016214502388618960.

James H. Stock and Mark W. Watson. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In John B. Taylor and Harald Uhlig (eds.), *Handbook of Macroeconomics*, volume 2, chapter 8, pp. 415–525. Elsevier, 2016. doi: 10.1016/bs.hesmac.2016.04.002.

Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Timepfn: Effective multivariate time series forecasting with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20761–20769, 2025.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. *arXiv preprint arXiv:2410.16032*, 2024.

J. Curtis Weaver. Methods for estimating peak discharges and unit hydrographs for streams in the city of Charlotte and Mecklenburg County, North Carolina. Water-Resources Investigations Report 03-4108, U.S. Geological Survey, 2003. URL https://pubs.usgs.gov/wri/wri034108/.

Yosuke Yamaguchi, Issei Suemitsu, Yuki Kajihara, and Wenpeng Wei. Citras-fm: Tiny time series foundation model for covariate-informed zero-shot forecasting, 2026. URL https://arxiv.org/abs/2606.10798.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Lifan Zhao and Yanyan Shen. Rethinking channel dependence for multivariate time series forecasting: Learning from leading indicators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JiTVtCUOpS.

Jingbo Zhou, Xinjiang Lu, Yixiong Xiao, Jiantao Su, Junfu Lyu, Yanjun Ma, and Dejing Dou. SDWPF: A dataset for spatial dynamic wind power forecasting challenge at KDD Cup 2022, 2022. URL https://arxiv.org/abs/2208.04360.

A Experimental Protocol and Reporting

The appendices follow the behavioral experiments in the main text: temporal event–response diagnostics (Appendix B), cross-channel dependency structure (Appendix C), conditionally irrelevant future covariates (Appendix D), reliability shifts (Appendix E), and conditioning on predicted covariates in real data (Appendix F). The separate representation experiments are collected in Appendix G.

A.1 Models and Input Conditions

We evaluate frozen Chronos-2, TimesFM-3, and TiRex-2 checkpoints, identified as `amazon/chronos-2`, `google/timesfm-3.0-pytorch`, and `NX-AI/TiRex-2`, respectively. No backbone is fitted to the synthetic tasks or the real-data evaluation cohorts. Representation probes are fitted separately, using the splits described in Appendix G.

CI uses only the scored target's history; MV supplies the histories of the target and other channels as a multivariate input. PastCov supplies the covariate history through the past-covariate interface, and KC additionally supplies future covariate values. The information supplied by each condition is specified within its experiment. These names describe inputs rather than assuming that every backbone implements the interfaces with the same internal computation. Point-forecast measurements use the native median forecast, or the available quantile closest to 0.5 as specified for the temporal task.

A.2 Baselines Fitted to the Observed Context

We distinguish three types of reference: an analytic oracle that knows the data-generating process, a statistical model fitted to the available observations under a specified model class, and a frozen pretrained TSFM. The fitted baselines assess performance achievable from the finite observed context with explicit modeling assumptions. Each fit uses only one window's history; training examples are not pooled across windows, conditions, or seeds. Supplied future covariates may be used for prediction when allowed by the corresponding input condition, but target values in the final forecast horizon are never used for fitting or model selection. Validation targets used for hyperparameter selection lie entirely inside the observed context. Fitting objectives and selection criteria are specified separately for each estimator below.

Generating parameters, true lags, event labels, kernels, and event annotations are excluded from estimator inputs. The known-lag OLS reference is an explicitly privileged exception: it receives the true lag, while estimating its regression coefficients from the observed history. In the structural direct-access condition, the true latent history is available because it is part of that condition's observed input. All other structural conditions use only their specified target and covariate histories. These distinctions separate a model-class assumption from knowledge of the realized generating parameters.

The baselines use the same windows and five generation seeds as their corresponding behavioral experiments. Lagged-copy and structural tasks use a 512-step context, an eight-step horizon, and 256 windows per seed; event-response and position tasks use a 768-step context, a 96-step horizon, and 300 windows per seed. TSFM forecasts are taken from the existing experiment records. Comparisons are formed within matched windows and summarized within each seed before reporting the mean and SEM across the five seeds. For response profiles, the original equal-combination aggregation is retained. Additional model fits, lags, anchors, and channel counts are not treated as independent repetitions.

These baselines produce point forecasts only. We report MAE, event-response metrics, and input-sensitivity diagnostics as appropriate; no baseline quantile loss, interval coverage, or calibration claim is made. Successful fitted baselines demonstrate achievable performance under their stated assumptions, while unsuccessful fits do not establish that predictive information is absent from the context. Method details and results follow the corresponding behavioral experiments in Appendices B.3, C.3, and D.4.

Table 4: Evaluation populations. Counts refer to each synthetic setting, unless stated otherwise. Probe experiments are separate from the behavioral sweeps.

Experiment	Context	Horizon	Seeds	Evaluation units per seed
Temporal responses	768	96	5	300 windows per combination/lag
Structural dependency	512	8	5	256 paired windows
Irrelevant future	512	8	5	256 windows; 64 for E_J
KC context control	1024/2048	8	5	256 windows; $\beta = 0.8$
Reliability shift	512	8	5	256 paired windows
Latent-state probes	512	8	3	256 windows
Lag-class probes	512	8	5	128 base draws $\times$ 6 classes

A.3 Evaluation Populations and Uncertainty

Table 4 distinguishes the independent generation seeds from the windows and probe folds used within each experiment. Synthetic comparisons share realizations between their compared conditions. For the temporal and structural behavioral experiments, uncertainty is computed from seed-level quantities; paired differences are formed before averaging over seeds. The reliability-shift table uses a different, paired-window SEM convention, which is stated explicitly in Appendix E.1. Metric-specific definitions and aggregation order are given with each experiment.

All five-seed experiments, including temporal responses, use generation seeds 0–4. Probe uncertainties summarize held-out-seed folds whose training sets overlap; the folds are not additional independently trained foundation models. Real-data cohorts follow the dataset-specific rolling origins and series-level aggregation in Appendix F.1, rather than the synthetic generation-seed protocol.

B Temporal Event–Response Diagnostics

B.1 Temporal Cross-Channel Utilization Task

Data generation. For condition $c = (p, r)$, comprising trigger family p and response family r, the covariate and target channels follow

$$X_t^{(p)} = \sigma_x \epsilon_t^x + A \sum_k P_p(t - t_k) + A \sum_j P_{\mathrm{dis},p}(t - u_j), \tag{7}$$

$$Y_t^{(r)} = L_t + B \sum_k R_r(t - t_k - \tau) + \sigma_y \epsilon_t^y, \tag{8}$$

$$L_t = L_{t-1} + \sigma_{\mathrm{level}} \epsilon_t^L, \tag{9}$$

where the noise sequences are independent standard Gaussian sequences, t_k are trigger anchors, and u_j are distractor anchors. The target level is initialized at zero immediately before the generated context. The covariate background is white noise, and the target background is an independent random walk. The amplitudes are $A = B = 3$, with noise scales $\sigma_x = 0.3$, $\sigma_y = 0.1$, and $\sigma_{\mathrm{level}} = 0.15$.

Trigger and response families. The width-four trigger kernels are

$$P_{\mathrm{spike}} = [0, 1, 1, 0], \tag{10}$$

$$P_{\mathrm{dip}} = [0, -1, -1, 0], \tag{11}$$

$$P_{\mathrm{biphasic}} = [1, 1, -1, -1]. \tag{12}$$

Each family uses the sign-reversed distractor $P_{\mathrm{dis},p} = -P_p$, which produces no target response. We cross the three triggers with three width-twelve response kernels:

$$R_{\mathrm{positive}} = [0, 0.2, 0.4, 0.6, 0.8, 1, 1, 0.8, 0.6, 0.4, 0.2, 0], \tag{13}$$

$$R_{\mathrm{negative}} = -R_{\mathrm{positive}}, \tag{14}$$

$$R_{\mathrm{biphasic}} = [0, 0.5, 1, 1, 0.5, 0, 0, -0.5, -1, -1, -0.5, 0]. \tag{15}$$

Every kernel is zero outside its listed support. The biphasic response concatenates a width-six positive triangle with its negative counterpart. Each of the nine combinations is a separate condition: all teaching pairs and the test event within one window use the same trigger–response mapping. The nine conditions are not mixed within a window. The injected trigger energies $A^2\|P_p\|_2^2$ are 18, 18, and 36 for spike, dip, and biphasic triggers; the response energies $B^2\|R_r\|_2^2$ are 39.6, 39.6, and 45 for positive, negative, and biphasic responses. Within each set, absolute peak amplitudes and nominal widths agree, but total energies and nonzero supports differ.

Windows and event placement. We generate 300 windows for each seed $s \in \{0, 1, 2, 3, 4\}$, trigger–response combination, and lag, using context length 768 and forecast horizon 96. Time is indexed relative to the forecast origin: history occupies $t = -768, \ldots, -1$ and the forecast horizon occupies $t = 0, \ldots, 95$. We denote the test-response anchor by t_{resp} and its trigger anchor by $t_{\text{trig}} = t_{\text{resp}} - \tau$. The main nine-combination sweep fixes $t_{\text{resp}} = 24$, the 25th forecast sample. The additional spike-to-positive position control varies $t_{\text{resp}} \in \{16, 24, 32, 40\}$ at short lags $\tau \in \{0, 4, 8, 16\}$ (Appendix B.4). The main lag grid is

$$\tau \in \{0, 4, 8, 12, 16, 20, 24, 32, 40, 48, 64, 96, 128\}. \tag{16}$$

The trigger is in the known-future segment for $\tau < t_{\text{resp}}$, at its first sample for $\tau = t_{\text{resp}}$, and in history for $\tau > t_{\text{resp}}$. Lag refers to the separation between injection anchors; the first nonzero samples and peaks need not coincide at $\tau = 0$.

Each context contains nine teaching triggers on a grid with spacing 76, starting at -762, with independent integer jitter drawn uniformly from $\{-8, \ldots, 8\}$. Distractor anchors use the midpoint grid, offset by 38 steps, with independently drawn jitter. Distractor grid points later than $-1 - \tau - 12$ are excluded, giving nine distractors through $\tau = 96$ and eight at $\tau = 128$. For a fixed seed, lag, and window, background realizations and event placements are shared across all nine combinations. Holding the response family fixed gives identical target arrays across triggers; holding the trigger fixed gives identical covariate arrays across responses. Changing the response modifies both historical teaching responses and the test response. Across lags, teaching responses move to $t_k + \tau$, so the target history also changes.

The original placement rules are retained for every combination. For seeds 0 through 4, respectively, 44, 33, 41, 36, and 32 of the 2,700 scheduled teaching triggers extend beyond the left context boundary at each lag. The retained teaching-trigger energy is approximately 99.61–99.72% for spike and dip, and 99.40–99.56% for biphasic triggers. At zero lag, teaching responses at these clipped anchors can also be truncated; this placement rule is shared across response families. The test trigger, shifted trigger, and test response are fully contained in the supplied window.

Conditioning and counterfactual inputs. Each frozen model receives the target history together with covariate history and known-future values through the same conditioning interface at every lag. We use the native forecast quantile closest to 0.5 and each backbone's internal scaling, without additional experiment-level normalization. The event and no-event inputs differ only by removal of the test trigger; the target history, teaching examples, distractors, and background values remain fixed within this pair. For combination c, seed s, and window w, define

$$D_{csw,\tau}(h) = \widehat{Y}_{\text{event},csw,\tau}(h) - \widehat{Y}_{\text{no-event},csw,\tau}(h). \tag{17}$$

The primary response profile is the within-seed window average,

$$\overline{D}_{cs,\tau}(h) = \frac{1}{300} \sum_{w=1}^{300} D_{csw,\tau}(h). \tag{18}$$

Gain, onset, and profile measurement. For a fixed response anchor, let $Q_c(h) = BR_r(h - t_{\text{resp}})$ be the true injected response over the full forecast horizon. We measure the signed projection gain

$$g_{cs,\tau} = \frac{\langle \overline{D}_{cs,\tau}, Q_c \rangle}{\langle Q_c, Q_c \rangle}. \tag{19}$$

This definition applies to positive, negative, and zero-sum biphasic responses. The gain is 1 for the exact response, 0 for a zero response, and -1 for its exact sign reversal. All nine combinations

use this definition, including the reused positive-response forecasts. For each valid response anchor $o \in \{0, \ldots, 84\}$, the response-specific matched-filter score is

$$a_{cs,\tau}(o) = \frac{\sum_{j=0}^{11} \overline{D}_{cs,\tau}(o+j) R_r(j)}{\|R_r\|_2}. \tag{20}$$

We select $\widehat{t}_{\mathrm{resp},cs,\tau} = \arg\max_o |a_{cs,\tau}(o)|$, resolving ties at the earliest anchor, and define signed onset error $e_{cs,\tau} = \widehat{t}_{\mathrm{resp},cs,\tau} - t_{\mathrm{resp}}$. Positive errors indicate a late response and negative errors an early response. The absolute onset error is $|e_{cs,\tau}|$. These metrics localize each combination's window-averaged profile; timing is not averaged over separately localized windows. No detection threshold filters the reported model results. An exactly zero response has no meaningful onset, even though an unmasked argmax returns a numerical index. To capture deviations in response shape as well as amplitude, we also report

$$E_{cs,\tau} = \frac{\|\overline{D}_{cs,\tau} - Q_c\|_2}{\|Q_c\|_2}, \tag{21}$$

with norms taken over the full horizon. The exact response has $E = 0$ and a zero response has $E = 1$.

Aggregation and uncertainty. We use equal-combination averages within each seed,

$$G_{s,\tau} = \frac{1}{9} \sum_c g_{cs,\tau}, \qquad T_{s,\tau} = \frac{1}{9} \sum_c |e_{cs,\tau}|. \tag{22}$$

Figure 2 reports the across-seed mean and SEM of $G_{s,\tau}$. Timing is a supplementary diagnostic: Table 5 reports the across-seed mean of $T_{s,\tau}$ alongside gain. For any seed-level quantity q_s, SEM is

$$\mathrm{SEM}(q) = \sqrt{\frac{\sum_s (q_s - \overline{q})^2}{S(S-1)}}, \qquad S = 5. \tag{23}$$

Absolute timing errors are taken before averaging over combinations or seeds, so early and late errors do not cancel. Profiles with different response signs are never averaged together before computing their metrics. Figure 6 shows combination-specific gain means. Table 5 summarizes gain and mean absolute onset error at selected lags; Tables 8 and 9 compare macro gain and profile error with the fitted baseline. This uncertainty describes five seeds of a fixed nine-combination evaluation. Within-segment lag differences use paired seed-level quantities, with reference lag 0 for the known-future segment and 32 for history; lag 24 is reported separately.

Control arms and analytic references. The distractor arm replaces the test trigger with $-P_p$. The shifted-trigger arm moves it eight steps later while retaining the original historical demonstrations, implying an ideal response anchor of $t_{\mathrm{resp}}+8$ (32 in the main setting). Each control response is measured relative to the same no-event input. We distinguish the shifted response's onset error relative to $t_{\mathrm{resp}} + 8$ from its onset shift relative to the unshifted forecast. The no-covariate arm removes both covariate history and covariate future, including all covariate teaching evidence; it is a global ablation rather than a test-trigger intervention. The analytic event response is $Q_c(h) = BR_r(h - t_{\mathrm{resp}})$, with gain 1 and onset error 0. For the shifted analytic response $Q_c^{\mathrm{shift}}(h) = BR_r(h - t_{\mathrm{resp}} - 8)$, gain relative to the original template is $\langle Q_c^{\mathrm{shift}}, Q_c \rangle / \|Q_c\|_2^2$. This equals approximately 0.03636 for the positive and negative responses and -0.2 for the biphasic response. Scoring against its own shifted template instead gives gain 1 and onset error 0 relative to anchor $t_{\mathrm{resp}} + 8$. The distractor has zero analytic response and no defined onset.

Scope. Lag variation changes the trigger's absolute position and the temporal arrangement of teaching responses, and traverses the future–history boundary. Within-segment comparisons keep the trigger's input segment fixed but do not isolate temporal distance from absolute position. The trigger and response families also differ in energy and support, so their comparison does not isolate shape at equal signal energy. Detailed results are reported in Appendix B.2.

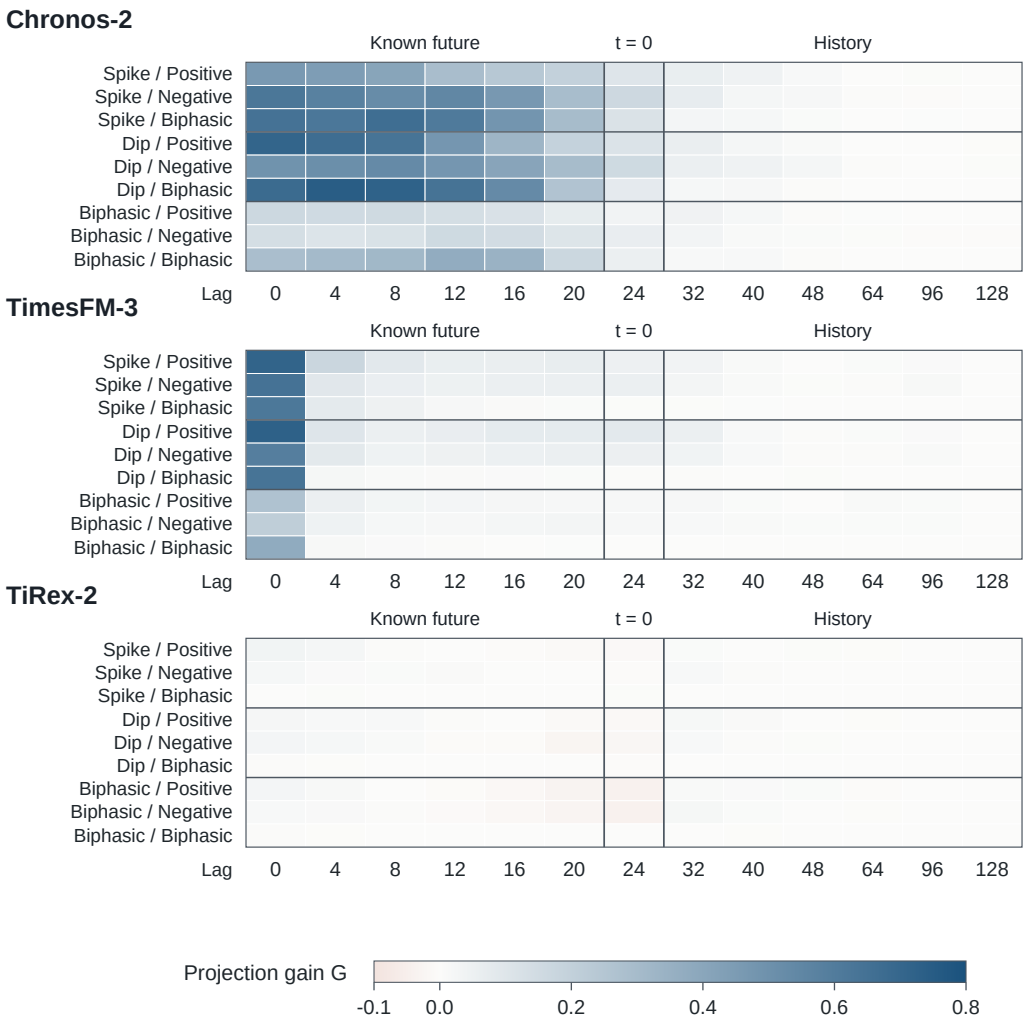

Figure 6: Signed projection gain across all nine trigger–response combinations and 13 lags. Rows identify trigger / target-response families; cells are means across five seeds on a shared color scale. Columns show discrete lag settings. With $t_{\text{resp}} = 24$, the trigger is in the known future for $\tau < 24$, at the forecast origin for $\tau = 24$, and in history for $\tau > 24$.

B.2 Detailed Responses and Controls

Figure 6 shows all nine trigger–response combinations across the 13 tested lags. Metrics are computed from each seed's window-averaged response profile before aggregation across five seeds. Table 5 reports equal-combination gains and mean absolute onset errors at selected lags. Definitions are given in Appendix B.1.

Combination-specific patterns. At zero lag, Chronos-2 and TimesFM-3 respond less strongly to biphasic triggers than to spikes or dips for each target-response family, even though both can produce substantial biphasic target responses. This distinguishes trigger and response roles without isolating shape from energy or support. TimesFM-3's reduction at $\tau = 4$ occurs in every combination and seed. The paired macro-gain changes from $\tau = 0$ to 8 are -0.01299 ± 0.00209 for Chronos-2 and -0.49167 ± 0.00507 for TimesFM-3. Chronos-2's attenuation varies across combinations and is not uniformly monotonic; its onset error remains zero through $\tau = 20$ in every combination and seed. When gain becomes small, a matched-filter onset estimate does not independently establish a meaningful target response.

Table 5: Temporal utilization averaged equally across nine event–response combinations. Gain is reported as mean ± SEM over five seeds; Timing is the mean absolute error in the response anchor, in steps.

	Chronos-2		TimesFM-3		TiRex-2	
τ	Gain	Timing	Gain	Timing	Gain	Timing
Trigger in the known-future segment						
0	0.4678 ± 0.0032	0.0	0.5328 ± 0.0050	1.0	0.0180 ± 0.0010	5.3
8	0.4548 ± 0.0030	0.0	0.0412 ± 0.0014	7.0	0.0052 ± 0.0007	15.1
16	0.3418 ± 0.0020	0.0	0.0351 ± 0.0008	8.1	-0.0046 ± 0.0006	16.4
Trigger at the forecast origin						
24	0.0996 ± 0.0021	0.6	0.0327 ± 0.0006	4.7	-0.0154 ± 0.0005	15.4
Trigger in the observed history						
32	0.0438 ± 0.0013	0.0	0.0228 ± 0.0009	5.4	0.0090 ± 0.0004	8.0
40	0.0245 ± 0.0002	0.9	0.0078 ± 0.0003	13.3	0.0034 ± 0.0001	7.6
48	0.0109 ± 0.0001	6.3	0.0030 ± 0.0002	16.9	0.0015 ± 0.0001	18.9
64	0.0019 ± 0.0001	38.8	0.0050 ± 0.0002	25.6	-0.0011 ± 0.0001	31.1

Distractor and shifted-trigger controls. Tables 6 and 7 retain the seed-level control results. At zero lag, absolute distractor gain is smaller than event gain for Chronos-2 and TimesFM-3 in every combination and seed. TiRex-2's distractor gain exceeds its event gain in the biphasic-to-negative condition in every seed. Shifting the trigger eight steps later shifts Chronos-2's estimated response anchor by eight steps in every combination and seed; TimesFM-3's shifted onset error is zero or -1. These controls distinguish responsiveness from selective use of the trigger. The reported full grid contains all 8,775 planned model–combination–lag–seed–arm records, with no missing or nonfinite forecasts.

B.3 Event-Response Regression Fitted to the Observed Context

Method and prior. We fit a polarity-gated event-response regression to each window's target and covariate histories. The model class assumes sparse, large covariate events, transient target responses, and a local-level background. Detected events can induce separate response kernels for positive and negative onset polarities. The estimator is not given the generating lag, trigger or response family, kernels, event annotations, teaching-event locations, or noiseless target background.

The event detector uses a strict threshold of four on the standardized covariate and a median-absolute-deviation scaling constant of 1.4826. Bursts separated by at most two non-flagged samples are merged, and the onset's first polarity is retained. The onset rule reads only the current and preceding three samples, so extending the supplied path into the future does not revise previously fixed onsets.

Fitting the response kernels. Let $N_q(t)$ be the detected onset sequence for polarity $q \in \{+, -\}$, and let $k_{q,\ell}$ denote its response kernel at lag $\ell \in \{0, \dots, 160\}$. The event contribution is

$$g_t = \sum_{q \in \{+,-\}} \sum_{\ell=0}^{160} k_{q,\ell} N_q(t-\ell). \tag{24}$$

We fit the stacked kernel vector k to the target's first differences $\Delta Y_t = Y_t - Y_{t-1}$. The design has entries $D_{t,(q,\ell)} = N_q(t-\ell) - N_q(t-1-\ell)$, and the fitting objective is

$$\min_k \ \|\Delta Y - Dk\|_2^2 + d\lambda_{\mathrm{sm}} \|D_2 k\|_2^2 + d\, 10^{-4} \|k\|_2^2. \tag{25}$$

Here D_2 applies separate second differences to the two polarity kernels, with zero extension at their boundaries. The scale d is the mean of the positive diagonal entries of $D^\top D$ on the retained design, or one if none are positive; it does not use a generating amplitude or noise scale. The first 161 difference rows are removed in each fit, leaving 606 rows for the full 768-step context.

Table 6: Distractor projection gain at $\tau = 0$ for seeds 0–4 (column headings). The ideal value is zero; gain uses the condition's signed target response.

		Chronos-2					TimesFM-3					TiRex-2				
Trigger	Response	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Spike	Positive	−0.0172	−0.0180	−0.0187	−0.0263	−0.0193	−0.0274	−0.0246	−0.0253	−0.0292	−0.0274	0.0091	0.0065	0.0083	0.0048	0.0007
Spike	Negative	−0.0405	−0.0385	−0.0368	−0.0397	−0.0379	−0.0277	−0.0266	−0.0285	−0.0280	−0.0272	0.0065	0.0044	0.0068	0.0057	0.0004
Spike	Biphasic	−0.0194	−0.0209	−0.0213	−0.0224	−0.0222	−0.0005	0.0010	0.0013	−0.0017	0.0007	0.0007	0.0004	0.0004	0.0003	0.0012
Dip	Positive	−0.0277	−0.0320	−0.0345	−0.0334	−0.0286	−0.0078	−0.0140	−0.0125	−0.0144	−0.0097	0.0057	0.0047	0.0059	0.0045	0.0023
Dip	Negative	−0.0273	−0.0258	−0.0243	−0.0186	−0.0253	−0.0222	−0.0238	−0.0187	−0.0167	−0.0249	0.0026	−0.0012	0.0015	0.0016	−0.0013
Dip	Biphasic	−0.0373	−0.0398	−0.0398	−0.0337	−0.0376	−0.0063	−0.0068	−0.0060	−0.0058	−0.0061	−0.0024	−0.0026	−0.0027	−0.0029	−0.0025
Biphasic	Positive	0.0353	0.0340	0.0353	0.0328	0.0384	0.0142	0.0080	0.0135	0.0140	0.0127	0.0250	0.0254	0.0272	0.0260	0.0256
Biphasic	Negative	0.0258	0.0272	0.0297	0.0306	0.0259	0.0185	0.0171	0.0172	0.0177	0.0166	0.0271	0.0239	0.0245	0.0260	0.0239
Biphasic	Biphasic	−0.0179	−0.0186	−0.0227	−0.0184	−0.0201	−0.1164	−0.1264	−0.1232	−0.1243	−0.1109	−0.0094	−0.0080	−0.0138	−0.0085	−0.0091

Table 7: Shifted-event onset error at $\tau = 0$ for all nine trigger–response combinations, reported separately for seeds 0–4. The error $\widehat{t}_{\mathrm{resp,shifted}} - (t_{\mathrm{resp}} + 8)$ is measured in steps relative to the ideal shifted anchor $t_{\mathrm{resp}} + 8 = 32$ in the main setting; its ideal value is zero. Column headings 0–4 identify the seeds.

		Chronos-2					TimesFM-3					TiRex-2				
Trigger	Response	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Spike	Positive	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
Spike	Negative	0	0	0	0	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
Spike	Biphasic	0	0	0	0	0	0	0	0	0	0	5	4	4	4	5
Dip	Positive	0	0	0	0	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
Dip	Negative	0	0	0	0	0	−1	−1	−1	−1	−1	0	0	1	−28	0
Dip	Biphasic	0	0	0	0	0	0	0	0	0	0	−7	−7	−6	−6	−6
Biphasic	Positive	0	0	0	0	0	−1	−1	−1	−1	−1	−32	−31	−32	−32	−32
Biphasic	Negative	0	0	0	0	0	−1	−1	−1	−1	−1	−1	−1	1	1	−1
Biphasic	Biphasic	0	0	0	0	0	0	0	0	0	0	−2	−2	1	10	−2

Selecting regularization and forecasting. The smoothness weight λ_{sm} is selected from $\{0, 0.1, 1, 10\}$ using history prefixes of length 480, 576, and 672, each followed by a 96-step validation interval within the observed context. For each prefix, covariate centering and scaling are estimated from that prefix only. The detector uses these statistics with the fixed threshold, and response kernels are fitted to the same prefix. The following covariate values are supplied as known-future inputs to the validation forecast, without entering the fitted scaler or kernels. Selection minimizes the unweighted mean of the three folds' 96-step *level-forecast MSE in raw target units*, rather than the penalized first-difference objective used for fitting. Ties favor the larger smoothness weight, with comparison tolerance 10^{-12} relative to score magnitude and a unit floor. The final held-out target horizon is never used in this selection.

After refitting at the selected weight on the full observed history, the level forecast is

$$\widehat{Y}_h = Y_{-1} - \widehat{g}_{-1} + \widehat{g}_h. \tag{26}$$

Future covariate values enter the detected-event sequence used for prediction. Counterfactual changes confined to the future share the historical fit; changes to the covariate history recompute detection, regularization selection, and the kernels. If no burst is detected, the procedure reduces to a local-level forecast. No such fallback occurred in the reported temporal grid. The no-covariate arm directly repeats the last observed target value.

Response gains and profile errors. We evaluate the same nine event–response combinations, 13 lags, five seeds, and 300 windows per combination and seed as the frozen-model experiment. Gains and profile errors are computed from the window-averaged event-minus-no-event response, using the original signed template. Equal-combination mean gain remains between 0.670 and 0.780 over the full lag grid, while mean profile error ranges from 0.256 to 0.358 (Tables 8 and 9). At each lag, its mean gain is higher and its mean profile error is lower than those of each TSFM. These are comparisons of the nine-combination averages, not claims that every individual combination favors the baseline. For example, Chronos-2 has lower profile error for the dip-to-biphasic combination at $\tau = 4$.

Table 10 summarizes the gain and profile-error ranges for each combination over the lag grid. All 117 combination–lag mean gains are positive. Matched-filter onset error is zero for every seed's window-averaged profile, which does not imply exact timing in every individual window. At zero lag, mean distractor gains range from -0.0027 to 0.0019; shifting the trigger by eight steps shifts the baseline's averaged response by eight steps in every combination and seed. The results demonstrate substantial recoverability from the supplied contexts under an explicit event-response model class.

B.4 Short-Lag Responses across Forecast Positions

Design. For the spike-to-positive condition, we vary the target-response anchor over $t_{\mathrm{resp}} \in \{16, 24, 32, 40\}$ and the lag over $\tau \in \{0, 4, 8, 16\}$, placing the trigger at $t_{\mathrm{trig}} = t_{\mathrm{resp}} - \tau$. The context, horizon, noise settings, and sample counts match the temporal experiment. Event, no-event, and distractor inputs are evaluated for each setting. The full grid contains 16 anchor–lag cells; at $t_{\mathrm{resp}} = 16$, $\tau = 16$, the trigger is at the first forecast sample. We retain and flag this boundary cell in the full results. For a balanced comparison with strictly future triggers at every lag, we average within-anchor changes over $t_{\mathrm{resp}} \in \{24, 32, 40\}$.

For each method and seed, we first compute $G_s(t_{\mathrm{resp}}, \tau) - G_s(t_{\mathrm{resp}}, 0)$ and the corresponding profile-error change, then average equally over the three common-interior anchors. Means and SEMs are computed over the resulting five seed-level contrasts, not by averaging separate SEMs. The additional positions reuse the archived TSFM forecasts; they are not new independently trained models or additional generation seeds.

Results. The fitted event regression has mean gain between 0.708 and 0.722 across all 16 cells (Table 11). Its mean profile errors range from 0.299 to 0.343 (Table 12). TimesFM-3's response is strongly attenuated at nonzero lags at every tested anchor, while Chronos-2's response strength also varies with the anchor. For the common-interior comparison from $\tau = 0$ to 16, mean gain changes are 0.0050 ± 0.0046 for the fitted regression, -0.2501 ± 0.0041 for Chronos-2, -0.7721 ± 0.0033 for TimesFM-3, and -0.0318 ± 0.0031 for TiRex-2 (Table 13). The short-lag attenuation is therefore not confined to the main response anchor $t_{\mathrm{resp}} = 24$. This control covers one event–response

Table 8: Projection gain over all nine event–response combinations. Each combination metric is formed before equal-weight averaging within a seed; entries are mean ± SEM across five seeds. Event regression is fitted to each observed context; TSFM results are the archived frozen-model forecasts.

τ	Event regression	Chronos-2	TimesFM-3	TiRex-2
0	0.686 ± 0.003	0.468 ± 0.003	0.533 ± 0.005	0.018 ± 0.001
4	0.689 ± 0.005	0.461 ± 0.003	0.076 ± 0.002	0.011 ± 0.001
8	0.711 ± 0.002	0.455 ± 0.003	0.041 ± 0.001	0.005 ± 0.001
12	0.714 ± 0.002	0.413 ± 0.002	0.034 ± 0.000	0.000 ± 0.001
16	0.713 ± 0.004	0.342 ± 0.002	0.035 ± 0.001	-0.005 ± 0.001
20	0.707 ± 0.004	0.211 ± 0.003	0.036 ± 0.001	-0.012 ± 0.000
24	0.695 ± 0.004	0.100 ± 0.002	0.033 ± 0.001	-0.015 ± 0.000
32	0.720 ± 0.002	0.044 ± 0.001	0.023 ± 0.001	0.009 ± 0.000
40	0.695 ± 0.005	0.024 ± 0.000	0.008 ± 0.000	0.003 ± 0.000
48	0.693 ± 0.003	0.011 ± 0.000	0.003 ± 0.000	0.002 ± 0.000
64	0.670 ± 0.004	0.002 ± 0.000	0.005 ± 0.000	-0.001 ± 0.000
96	0.743 ± 0.004	0.001 ± 0.001	0.007 ± 0.000	0.000 ± 0.000
128	0.780 ± 0.005	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 9: Profile error over all nine event–response combinations. Each combination metric is formed before equal-weight averaging within a seed; entries are mean ± SEM across five seeds. Event regression is fitted to each observed context; TSFM results are the archived frozen-model forecasts.

τ	Event regression	Chronos-2	TimesFM-3	TiRex-2
0	0.358 ± 0.004	0.562 ± 0.002	0.577 ± 0.004	0.983 ± 0.001
4	0.330 ± 0.005	0.559 ± 0.003	0.929 ± 0.002	0.990 ± 0.001
8	0.308 ± 0.002	0.561 ± 0.003	0.962 ± 0.001	0.996 ± 0.001
12	0.306 ± 0.002	0.601 ± 0.002	0.970 ± 0.001	1.001 ± 0.001
16	0.309 ± 0.004	0.668 ± 0.002	0.967 ± 0.001	1.005 ± 0.001
20	0.317 ± 0.003	0.794 ± 0.003	0.966 ± 0.001	1.012 ± 0.000
24	0.330 ± 0.004	0.902 ± 0.002	0.969 ± 0.001	1.016 ± 0.000
32	0.298 ± 0.002	0.957 ± 0.001	0.978 ± 0.001	0.991 ± 0.000
40	0.324 ± 0.005	0.976 ± 0.000	0.993 ± 0.000	0.997 ± 0.000
48	0.328 ± 0.003	0.990 ± 0.000	0.998 ± 0.000	0.999 ± 0.000
64	0.350 ± 0.005	0.998 ± 0.000	0.995 ± 0.000	1.001 ± 0.000
96	0.283 ± 0.003	1.000 ± 0.001	0.993 ± 0.000	1.000 ± 0.000
128	0.256 ± 0.004	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

Table 10: EventResponseFit across the nine combinations. Each range is the minimum–maximum of the five-seed means over the 13 lags, not an uncertainty interval.

Trigger	Response	Gain G	Profile error E
Spike	Positive	0.654–0.774	0.265–0.369
Spike	Negative	0.644–0.769	0.280–0.374
Spike	Biphasic	0.651–0.794	0.226–0.382
Dip	Positive	0.653–0.771	0.267–0.369
Dip	Negative	0.642–0.768	0.281–0.376
Dip	Biphasic	0.650–0.794	0.227–0.383
Biphasic	Positive	0.652–0.780	0.260–0.370
Biphasic	Negative	0.644–0.774	0.277–0.374
Biphasic	Biphasic	0.640–0.797	0.224–0.377

combination and the tested short lags; because $t_{\text{trig}} = t_{\text{resp}} - \tau$, it does not independently identify effects of lag, trigger position, and response position.

C Cross-Channel Dependency Structure

C.1 Shared-Latent Construction and Measurements

Data generation. We generate a shared latent AR(1) process and a target channel:

$$z_t = \rho z_{t-1} + \sqrt{1-\rho^2}\,\epsilon_t^z, \tag{27}$$

$$Y_t = \kappa z_t + \sigma_y \epsilon_t^y. \tag{28}$$

The innovations ϵ_t^z and ϵ_t^y are independent standard Gaussian sequences. We use $\kappa = 1$, $\sigma_y = 1/\sqrt{2}$, and $\rho \in \{0, 0.3, 0.6, 0.9, 0.95, 0.98\}$. Each AR(1) path starts from its stationary $\mathcal{N}(0,1)$ distribution, with 512 burn-in steps discarded before the 512-step context and eight-step horizon. Thus $\mathrm{Var}(z_t) = 1$ at every persistence level and $\mathrm{Var}(Y_t) = 1.5$. An informative auxiliary channel is given by

$$X_{j,t} = \lambda z_t + \sqrt{1-\lambda^2}\eta_{j,t}, \tag{29}$$

where the η_j are mutually independent AR(1) processes, independent of z, with the same marginal variance and autocorrelation as z. The parameter λ controls the strength of the shared latent component. This construction preserves the marginal statistics of individual auxiliary channels while allowing additional observations to provide complementary evidence about z_t.

We set $\lambda = 0.75$ for the noisy-observation sweep. The idiosyncratic innovations have variance $1-\rho^2$, so that each η_j also has stationary variance one. Before varying the noisy channels, we compare target-only input with direct access to the true latent history $z_{1:T}$. Neither condition supplies future latent values. There are 256 independent windows per seed, with seeds 0–4. Conditions and models share the same realizations, and smaller channel sets are prefixes of larger sets. Across ρ, the standard-normal innovation draws are shared within each seed.

Channel configurations. For each $K \in \{1, 2, 4, 8, 16\}$, we construct three families of inputs:

- **Independent:** All K auxiliary channels observe the shared latent process through independent noise realizations. Increasing K therefore increases the available information about the target.
- **Duplicate:** All K auxiliary channels are identical copies of a single informative observation. Increasing K adds no information beyond the $K = 1$ condition.
- **Distractor:** One auxiliary channel observes the shared latent process, while the remaining $K - 1$ channels observe independent, target-irrelevant latent processes with matched temporal statistics. Increasing K adds no target-relevant information.

The three configurations coincide at $K = 1$.

Aggregation gains. Let $R_M^f(K)$ denote model MAE for configuration f with K auxiliary channels, and let $R_M(1)$ denote the common single-channel baseline. For the independent configuration, we define

$$\Delta_M(K) = R_M(1) - R_M^{\text{ind}}(K). \tag{30}$$

The corresponding exact-oracle gain is

$$\Delta^\star(K) = R^\star(1) - R^{\star,\text{ind}}(K). \tag{31}$$

Positive gains indicate improved forecasting accuracy. Each gain is measured relative to its own $K = 1$ baseline. The oracle gain quantifies the additional predictive utility available at each K, without assuming a linear relationship between channel count and forecast improvement.

Interference penalties. For the duplicate and distractor configurations, we define

$$I_f(K) = R_M^f(K) - R_M(1), \qquad f \in \{\text{dup}, \text{dis}\}. \tag{32}$$

Positive penalties indicate that additional channels increase forecast error. Because these configurations provide no additional target-relevant information, their exact-oracle risk remains unchanged from $K = 1$. The behavioral sweep reports paired seed-level differences over five seeds.

Table 11: Response gain for the spike-to-positive condition across response anchors t_{resp} and short lags. Entries are mean ± SEM across five seeds. The trigger is at $t_{\mathrm{trig}} = t_{\mathrm{resp}} - \tau$; † marks the boundary cell $t_{\mathrm{resp}} = 16, \tau = 16$. The balanced common-interior comparison uses $t_{\mathrm{resp}} \in \{24, 32, 40\}$ only.

t_{resp}	τ	Event regression	Chronos-2	TimesFM-3	TiRex-2
16	0	0.717 ± 0.005	0.536 ± 0.006	0.765 ± 0.007	0.029 ± 0.004
16	4	0.709 ± 0.003	0.466 ± 0.008	0.287 ± 0.006	0.026 ± 0.003
16	8	0.722 ± 0.002	0.510 ± 0.010	0.190 ± 0.002	0.011 ± 0.002
16	$16^{\dagger}$	0.722 ± 0.006	0.434 ± 0.002	0.121 ± 0.003	0.000 ± 0.001
24	0	0.717 ± 0.004	0.462 ± 0.005	0.713 ± 0.008	0.036 ± 0.003
24	4	0.708 ± 0.003	0.443 ± 0.005	0.176 ± 0.004	0.025 ± 0.002
24	8	0.721 ± 0.002	0.408 ± 0.008	0.089 ± 0.003	0.004 ± 0.001
24	16	0.722 ± 0.006	0.243 ± 0.007	0.062 ± 0.002	-0.004 ± 0.001
32	0	0.717 ± 0.004	0.366 ± 0.006	0.859 ± 0.003	0.028 ± 0.004
32	4	0.709 ± 0.003	0.223 ± 0.006	0.146 ± 0.003	0.017 ± 0.001
32	8	0.722 ± 0.002	0.233 ± 0.007	0.061 ± 0.001	0.011 ± 0.000
32	16	0.721 ± 0.006	0.173 ± 0.006	0.032 ± 0.002	-0.006 ± 0.001
40	0	0.717 ± 0.004	0.413 ± 0.008	0.866 ± 0.005	0.024 ± 0.004
40	4	0.709 ± 0.003	0.377 ± 0.009	0.283 ± 0.008	0.015 ± 0.003
40	8	0.722 ± 0.002	0.264 ± 0.009	0.119 ± 0.004	0.011 ± 0.002
40	16	0.722 ± 0.006	0.074 ± 0.002	0.029 ± 0.001	0.003 ± 0.001

Table 12: Profile error for the spike-to-positive condition across response anchors t_{resp} and short lags. Entries are mean ± SEM across five seeds. The trigger is at $t_{\mathrm{trig}} = t_{\mathrm{resp}} - \tau$; † marks the boundary cell $t_{\mathrm{resp}} = 16, \tau = 16$. The balanced common-interior comparison uses $t_{\mathrm{resp}} \in \{24, 32, 40\}$ only.

t_{resp}	τ	Event regression	Chronos-2	TimesFM-3	TiRex-2
16	0	0.343 ± 0.006	0.470 ± 0.006	0.316 ± 0.005	0.973 ± 0.004
16	4	0.312 ± 0.003	0.541 ± 0.008	0.732 ± 0.005	0.974 ± 0.003
16	8	0.299 ± 0.002	0.498 ± 0.010	0.825 ± 0.002	0.989 ± 0.002
16	$16^{\dagger}$	0.307 ± 0.006	0.579 ± 0.002	0.883 ± 0.003	1.000 ± 0.001
24	0	0.341 ± 0.006	0.551 ± 0.005	0.384 ± 0.004	0.965 ± 0.003
24	4	0.313 ± 0.004	0.567 ± 0.005	0.838 ± 0.005	0.976 ± 0.002
24	8	0.300 ± 0.002	0.598 ± 0.008	0.920 ± 0.003	0.996 ± 0.001
24	16	0.307 ± 0.006	0.760 ± 0.007	0.946 ± 0.002	1.004 ± 0.001
32	0	0.339 ± 0.006	0.636 ± 0.006	0.287 ± 0.002	0.973 ± 0.003
32	4	0.312 ± 0.003	0.779 ± 0.006	0.864 ± 0.003	0.983 ± 0.001
32	8	0.299 ± 0.002	0.769 ± 0.007	0.947 ± 0.001	0.990 ± 0.000
32	16	0.307 ± 0.006	0.828 ± 0.006	0.974 ± 0.002	1.007 ± 0.001
40	0	0.338 ± 0.006	0.596 ± 0.007	0.264 ± 0.001	0.977 ± 0.004
40	4	0.312 ± 0.003	0.629 ± 0.009	0.736 ± 0.006	0.985 ± 0.003
40	8	0.299 ± 0.002	0.738 ± 0.009	0.891 ± 0.003	0.990 ± 0.002
40	16	0.305 ± 0.006	0.929 ± 0.002	0.977 ± 0.001	0.998 ± 0.001

Conditional oracle. The oracle conditions only on the histories supplied in the corresponding arm. For noisy informative channels, a Gaussian state-space filter retains $(z_t, \eta_{1,t}, \ldots, \eta_{K,t})$, whose transition matrix is ρI and innovation covariance is $(1-\rho^2)I$. Its observations are Y_t with observation-noise variance σ_y^2 and the noiseless linear combinations $X_{j,t} = \lambda z_t + \sqrt{1-\lambda^2}\eta_{j,t}$. The conditional forecast is $\kappa\rho^h \mathbb{E}[z_T \mid \mathcal{I}_T]$; under direct latent access it reduces to $\kappa\rho^h z_T$. Gaussian conditional means and medians coincide, making this the MAE-optimal forecast under the specified generating process. Duplicate rows retain the single-observation information set; independent distractors leave the target posterior unchanged. Reported oracle MAEs are evaluated on the same finite windows as the models.

C.2 Structural Dependency Experiments: Numerical Results

Evaluation and reporting. These results use five seeds (0–4), with 256 paired target windows per seed, context length 512, and forecast horizon 8. All auxiliary observations are supplied as historical rows through the same multivariate interface; no future latent values are supplied. We score the target's median forecast. For each comparison, we first average paired MAE differences within each seed, then report their mean and sample standard deviation divided by $\sqrt{5}$. Oracle entries are empirical MAEs of the analytic conditional forecasts on the same windows, rather than population expected MAEs. Each oracle uses only the observations available in its condition and accounts for the AR(1) idiosyncratic components. The latent-state representation probes in Appendix G.1 are a separate three-seed experiment and are not pooled with this sweep.

Direct latent access. Table 15(a) reports the MAE reduction from adding the true latent history to the target-only input. At $\rho = 0.95$, the reductions are 0.03570 ± 0.00570 for Chronos-2, 0.03823 ± 0.00310 for TimesFM-3, and 0.01022 ± 0.00230 for TiRex-2; the corresponding oracle reduction is 0.04342 ± 0.00330. All three models have positive mean reductions at $\rho \in \{0.9, 0.95, 0.98\}$. The oracle gain is exactly zero at $\rho = 0$, where latent history contains no information about future latent values. The oracle gains also vary across nonzero persistence levels, so model gains across ρ must be read alongside these changes in available predictive value.

Independent observations. Table 14(a) gives the channel-count results at $\rho = 0.95$ relative to each method's noisy single-channel baseline. Chronos-2's mean gain rises to 0.01322 ± 0.00390 at $K = 4$ and then declines to 0.00605 ± 0.00560 at $K = 16$. The latter estimate remains positive; it does not show an average deterioration relative to $K = 1$. TimesFM-3 improves through the tested channel counts, reaching 0.02043 ± 0.00390 at $K = 16$. TiRex-2's corresponding gain is 0.00255 ± 0.00210, while the oracle gain is 0.02402 ± 0.00200. At $\rho = 0.98$, Chronos-2's mean gain again declines from 0.01092 ± 0.00290 at $K = 4$ to 0.00329 ± 0.00440 at $K = 16$. Table 15(b) reports the $K = 16$ comparison at every tested ρ, including the smaller oracle gains at low persistence. These estimates describe forecasting behavior; they do not identify a particular architectural bottleneck.

Duplicate and distractor controls. At $\rho = 0.95$ and $K = 16$, duplicates and distractors increase Chronos-2's MAE by 0.00787 ± 0.00130 and 0.01105 ± 0.00560, respectively. TimesFM-3's corresponding changes are 0.00225 ± 0.00130 and 0.00011 ± 0.00130. TiRex-2's duplicate change is -0.00078 ± 0.00061, while its distractor penalty is 0.00783 ± 0.00140. The full channel-count comparisons appear in Table 14(b–c). The oracle changes are exactly zero under both controls, because neither changes the predictive information set.

C.3 A Fitted State-Space Baseline for Shared-Latent Dependencies

Input preprocessing. The state-space baseline receives the target history and the covariate histories of the corresponding condition, without future covariates. Within each observed context, a covariate is removed as constant when its population standard deviation is at most $10^{-12}\max(1, \max_t |X_t|)$. Remaining covariates are deduplicated only when their historical columns are bitwise identical, and are ordered by a digest of their own historical values. The target is always retained. The retained channels are centered and scaled using their context means and population standard deviations; a constant target uses a scaling denominator of one. No generating parameter or condition label enters this preprocessing.

Fitted state-space model: assumptions and observations. The preprocessing above is applied to the full observed context, without cross-validation. Let C be the number of retained channels, including the target. In standardized units, define the state and observation vectors as

$$\mathbf{s}_t = (m_t, \widetilde{X}_{1,t}, \ldots, \widetilde{X}_{C-1,t})^\top, \tag{33}$$

$$\mathbf{v}_t = (\widetilde{Y}_t, \widetilde{X}_{1,t}, \ldots, \widetilde{X}_{C-1,t})^\top. \tag{34}$$

The Gaussian state-space model is

$$\mathbf{s}_t = r\mathbf{s}_{t-1} + \mathbf{u}_t, \qquad \mathbf{u}_t \sim \mathcal{N}(0, Q), \tag{35}$$

$$\mathbf{v}_t = \mathbf{s}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim \mathcal{N}(0, R), \tag{36}$$

$$R = \mathrm{diag}(R_y, 0, \ldots, 0), \qquad \mathbf{s}_1 \sim \mathcal{N}\left(0, \frac{Q}{1-r^2}\right). \tag{37}$$

The observation matrix is fixed to the identity and Q is a full innovation covariance. Within the observed context, only the target signal m_t is hidden; the covariates are observed directly as components of the state. Shared scalar persistence and white target measurement noise are explicit model-class assumptions. The true generating persistence ρ is not supplied to the fit.

State-space fitting and start selection. Expectation–maximization (EM) estimates r, Q, and R_y, including the stationary initial law in the complete-data likelihood. For smoothed moments conditioned on the observed history, write

$$S_{11} = \mathbb{E}[\mathbf{s}_1\mathbf{s}_1^\top], \tag{38}$$

$$S_{++} = \sum_{t=2}^{T} \mathbb{E}[\mathbf{s}_t\mathbf{s}_t^\top], \tag{39}$$

$$S_{+-} = \sum_{t=2}^{T} \mathbb{E}[\mathbf{s}_t\mathbf{s}_{t-1}^\top], \tag{40}$$

$$S_{--} = \sum_{t=2}^{T} \mathbb{E}[\mathbf{s}_{t-1}\mathbf{s}_{t-1}^\top]. \tag{41}$$

The persistence-dependent quantity is

$$M(r) = (1-r^2)S_{11} + S_{++} - r(S_{+-} + S_{+-}^\top) + r^2 S_{--}. \tag{42}$$

The M-step uses the update $Q = M(r)/T$ and maximizes

$$-\frac{T}{2}\log\det M(r) + \frac{C}{2}\log(1-r^2) \tag{43}$$

numerically over $r \in [-0.995, 0.995]$. The target measurement variance is updated in closed form. Both the eigenvalue floor for Q and the floor for R_y are 10^{-8} in standardized units. The zero covariate measurement variances are retained exactly; covariance updates use the Joseph form and symmetrization.

Three deterministic starts initialize R_y at 0.1, 0.5, and 0.9 of the standardized target variance. Each permits at most 100 EM iterations, with relative tolerance 10^{-5} for three consecutive iterations. Likelihood-decreasing updates are damped by factors 0.5, 0.25, and 0.125, or rejected if they still decrease the observed-history likelihood. Among converged starts, the one with highest history likelihood is selected. If no start converges, the best finite iterate across all starts is retained and marked nonconverged. No target values beyond the final forecast origin enter these decisions. The standardized target forecast is $r^h\mathbb{E}[m_T \mid \mathbf{v}_{1:T}]$, which is returned to raw units using the target mean and scale from the same context. This baseline is evaluated only at $\rho = 0.95$, matching the prespecified persistence level of the main structural figure.

Evaluation and convergence. The state-space baseline covers all 15 channel conditions at $\rho = 0.95$. It uses the original 512-step contexts, eight-step horizons, 256 windows per seed, and five seeds. All 19,200 logical state-space forecast requests have a converged selected solution and finite forecasts, although not every initialization converges. Results are reported over all finite forecasts without excluding a difficult window based on convergence.

Accuracy and sensitivity to channel composition. The fitted state-space model has lower MAE than each TSFM in every structural condition at $\rho = 0.95$, with the same direction in all five seeds (Table 17). At independent $K = 16$, its MAE is 0.7404 ± 0.0095 and its paired gap to the same-information oracle is 0.0045 ± 0.0018 (Table 16). Its improvement over its own noisy $K = 1$ baseline is 0.0217 ± 0.0027. These results provide an achievable reference under the specified fitted state-space class.

The state-space baseline's distractor-$K = 16$ penalty relative to its own noisy-$K = 1$ condition is 0.0021 ± 0.0019. Its duplicate penalties are exactly zero because of explicit duplicate removal, not learned robustness to redundancy.

C.4 Paired Changes in Channel Count

We additionally analyze paired changes between channel counts using the archived forecasts. The focal comparison uses $\rho = 0.95$, $\lambda = 0.75$, independent observations, and $K = 4$ versus $K = 16$. For each method, we first compute each window's MAE difference between the two channel counts, then average within each seed:

$$d_s = \frac{1}{N} \sum_{w=1}^{N} \left[e_{s,w}(16) - e_{s,w}(4) \right]. \tag{44}$$

Positive values indicate worse performance with more channels. All methods share the same 256 valid windows per seed. We report the mean, SEM, and descriptive 95% Student-t interval over five seed-level differences. The SEM is computed from these paired differences rather than from two separately estimated gain SEMs. This is an exploratory analysis of existing forecasts.

Increasing K from 4 to 16 raises Chronos-2's MAE by 0.00717 ± 0.00200 and lowers TimesFM-3's MAE by 0.00943 ± 0.00196; both directions hold in all five seeds. TiRex-2's change is -0.00069 ± 0.00043, and its descriptive 95% interval includes zero. The oracle and the fitted state-space baseline improve in the same comparison (Table 18). This comparison quantifies the loss of Chronos-2's gain beyond an intermediate channel count; it does not imply deterioration relative to $K = 1$.

D Conditionally Irrelevant Future Covariates

This section supplements the lagged-copy experiment in Section 4.1. Reliability mismatch is reported separately in Appendix E, and lag-class representation probes in Appendix G.2.

D.1 Lagged-Copy Construction and Measurements

Data generation and conditional irrelevance. We generate

$$X_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \tag{45}$$

$$Y_t = \beta X_{t-L} + \sqrt{1 - \beta^2}\, \epsilon_t, \tag{46}$$

where $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ is independent of X. We evaluate relation strengths $\beta \in \{0.4, 0.6, 0.8, 1\}$, with forecast horizon $H = 8$ and lags $L \in \{8, 16\}$. The case $\beta = 1$ is an exact lagged copy; smaller positive values introduce target noise while preserving unit marginal target variance. Each setting uses a context of 512 steps and seeds $0, 1, 2, 3, 4$, with 256 windows per seed for forecast metrics and future sensitivity. At $\beta = 1$, the target is the noiseless copy $Y_t = X_{t-L}$, so the conditional oracle MAE is zero when $L \geq H$.

Let X^-, Y^- denote the observed histories and X^+, Y^+ their future segments. For $L \geq H$, the source index for every target horizon step satisfies

$$T + h - L \leq T, \qquad h = 1, \ldots, H. \tag{47}$$

All target-relevant source values therefore lie in the observed history, implying

$$p(Y^+ \mid X^-, Y^-, X^+) = p(Y^+ \mid X^-, Y^-). \tag{48}$$

Equivalently, $I(Y^+; X^+ \mid X^-, Y^-) = 0$.

Input interfaces. A CI baseline uses only Y^-. We compare three additional conditions using the same target and covariate histories:

- **MV:** X^- and Y^- are supplied as ordinary multivariate channels.
- **PastCov:** Y^- is the target history, and X^- is supplied through the past-covariate interface.
- **KC:** Y^- is the target history, and (X^-, X^+) is supplied through the known-covariate interface.

We summarize the accuracy change associated with supplying the future continuation as

$$\Delta_{\mathrm{KC}} = \mathrm{MAE}_{\mathrm{PastCov}} - \mathrm{MAE}_{\mathrm{KC}}. \tag{49}$$

This is a descriptive interface comparison: it changes the supplied future values and how the covariate series is presented. It does not identify an internal mechanism or establish that token counts or positions change.

Historical-source response. For each forecast step h, the target-relevant source location is

$$s_h = T + h - L. \tag{50}$$

We perturb these historical source values individually and estimate the local response of the median forecast using central finite differences:

$$J_{h,j} = \frac{\hat{q}_{0.5,h}(X_{s_j} + \delta) - \hat{q}_{0.5,h}(X_{s_j} - \delta)}{2\delta}. \tag{51}$$

All other inputs are held fixed. We use $\delta = 0.5$ and the first 64 windows of each seed, a deterministic subset of the 256 windows used for forecast metrics. We average the estimated response operators across these windows and the five seeds before computing E_J. The oracle response is $J^\star = \beta I_H$. We measure deviation from this response by

$$E_J = \frac{\|J - \beta I_H\|_F}{\|\beta I_H\|_F}, \tag{52}$$

where lower values indicate closer agreement with the oracle.

Future-continuation sensitivity. For each fixed history pair (X^-, Y^-), we independently sample eight continuations,

$$X^{+,(m)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_H), \qquad m = 1, \ldots, 8, \tag{53}$$

while holding the target realization fixed. These continuations are all on-distribution and conditionally irrelevant when $L \geq H$. We measure their effect on the median forecast using

$$S_{\mathrm{future}} = \frac{1}{\binom{8}{2}} \sum_{m<m'} \frac{1}{H} \sum_{h=1}^{H} \left| \hat{q}^{(m)}_{0.5,h} - \hat{q}^{(m')}_{0.5,h} \right|. \tag{54}$$

The oracle satisfies $S^\star_{\mathrm{future}} = 0$. Because the interface and histories remain fixed, this comparison isolates sensitivity to the future values themselves. We compute sensitivity within each window, average over the 256 windows of each seed, and report the mean and SEM across the five seed-level means. The SEM is the sample standard deviation of these means divided by $\sqrt{5}$.

Additional controls. We include a negative control with $\beta = 0$ and $L = 8$, where the covariate is unrelated to the target; the $\beta = 0$, $L = 16$ condition was not evaluated. We also include relevant-future conditions with $\beta = 0.8$ and $L < H$. For the latter, the fraction of target steps whose source lies in the future is $r_{\mathrm{future}} = (H - L)/H$. The corresponding oracle sensitivity is

$$S^\star_{\mathrm{future}} = r_{\mathrm{future}} \beta \frac{2}{\sqrt{\pi}}. \tag{55}$$

These controls distinguish sensitivity to irrelevant values from the response expected when future values are predictive.

D.2 Accuracy and Future-Value Sensitivity

Table 1 reports accuracy and historical-source response error for exact and noisy copies. The main exception to the accuracy benefit of KC is Chronos-2 at $\beta = 1$, $L = 8$, where both MAE and E_J increase. The relation-strength sweep and its seed-level variation are reported in Appendix D.3. Table 19 retains the no-dependence and relevant-future controls. Future sensitivity remains nonzero in the no-dependence control, but is lower than in the positive-β settings at $L = 8$. Accuracy gains and sensitivity to irrelevant values are therefore distinct diagnostics.

D.3 Sensitivity to Relation Strength

Design and reporting. We extend the lagged-copy experiment to weaker relationships with $\beta \in \{0.4, 0.6\}$, alongside $\beta \in \{0.8, 1\}$. All four positive values are evaluated at $L \in \{8, 16\}$ for all three models, using a 512-step context, an eight-step horizon, and five seeds with 256 windows per seed. The $\beta = 0$ negative control is available only at $L = 8$. At every relation strength, $L \geq H$, so the future covariate remains conditionally irrelevant.

Table 20 reports the paired MAE change Δ_{KC} and future-value sensitivity S_{future} as mean $\pm$ SEM across five seeds. Comparisons are paired within each $(\beta, L, \mathrm{seed})$ setting; windows at different lags are not paired. We also report E_J from the mean response operator, following Appendix D.1. This differs from averaging the per-seed normalized errors; the latter's SEM is not attached to the reported E_J. At $\beta = 0$, E_J is undefined because its denominator is zero.

The theoretical oracle MAE is $\sqrt{2/\pi}\sqrt{1-\beta^2}$, giving approximately 0.7313, 0.6383, 0.4787, and 0 at $\beta = 0.4, 0.6, 0.8, 1$, respectively. Thus, changes in raw MAE across β also reflect changes in irreducible target noise. We assess the conditioning effect through comparisons at the same relation strength and lag.

Accuracy and historical-source response. At $\beta = 0.6$, KC lowers MAE relative to PastCov for every model, lag, and seed. At $\beta = 0.4$, all mean MAE reductions remain positive, but the changes are small relative to seed-level variation for TimesFM-3 at both lags and for Chronos-2 and TiRex-2 at $L = 16$. For example, at $L = 16$ these reductions are 0.0009 ± 0.0010, 0.0003 ± 0.0003, and 0.0008 ± 0.0013 for Chronos-2, TimesFM-3, and TiRex-2. TimesFM-3 and TiRex-2 show increasing mean gains over the four positive β values at both lags. Chronos-2 does not: at $L = 8$, its mean gain changes from 0.0254 at $\beta = 0.6$ to 0.0067 at $\beta = 0.8$ and -0.0133 at $\beta = 1$. The exact-copy deterioration occurs in all five seeds. KC lowers E_J in all 12 added model–lag–β settings; across the full positive-β grid, the only exception is again Chronos-2 at $\beta = 1$, $L = 8$.

Sensitivity to conditionally irrelevant future values. Future-value sensitivity is nonzero for every model and lag at both added relation strengths, despite zero oracle sensitivity. For TimesFM-3 and TiRex-2, mean sensitivity increases across the four positive β values at each lag. Chronos-2's mean sensitivity peaks at $\beta = 0.8$ and decreases at $\beta = 1$. The sweep therefore extends the observation that sensitivity to irrelevant future values can coexist with improved accuracy. It also shows that this sensitivity persists in weak-relation settings where the accuracy benefit is small or inconsistent across seeds.

D.4 Linear Baselines Fitted to the Observed History

Methods and input conditions. The target-mean baseline predicts the arithmetic mean of all observed target values, repeated over the forecast horizon, without model selection. The lag-selected ordinary least squares (OLS) baseline fits

$$Y_t = a + bX_{t-\ell} + e_t \tag{56}$$

for candidate lags $\ell \in \{0, \ldots, 64\}$, together with an intercept-only candidate. All candidates use the same one-based response times $t = 65, \ldots, T$. The intercept and slope are fitted by ordinary least squares using QR decomposition, without coefficient clipping or residual rounding. Selection minimizes the equally weighted mean validation MSE of two blocked folds wholly inside the 512-step history. The first fold fits response times 65–256 and validates on 257–384; the second fits 65–384 and validates on 385–512. Tied scores favor the intercept-only candidate, then the smaller

lag, using a relative comparison tolerance of 10^{-12} with a unit floor. The selected model is refitted on response times $65, \ldots, T$. Its intercept-only candidate therefore differs from the target-mean baseline, which uses all T target observations. The known-lag OLS reference receives the true L, performs no lag selection, and fits the intercept and slope on the same response support; its privileged input is kept separate from the lag-selected baseline.

The history-only route uses observed source values when the selected lag places them in history. If a prediction requires an unobserved source value, an AR(1) model with an intercept, fitted to every adjacent pair in the observed source history, supplies that value. Its slope is constrained to $[-0.995, 0.995]$, and the intercept is refitted conditional on the constrained slope. The estimator does not use the source's true white-noise law to set this slope to zero. The known-covariate route instead uses supplied future source values when required by the selected lag. Thus, an input's availability and the selected regression determine which source values enter a forecast. For a selected lag model, the forecast is $\widehat{Y}_{T+h} = \widehat{a} + \widehat{b} X_{T+h-\widehat{\ell}}$, with an estimated source value substituted only when required by the history-only route. The intercept-only candidate repeats its fitted intercept. MV and PASTCOV provide identical numerical inputs to these regressions and are treated as one history-only route.

We evaluate $\beta \in \{0.8, 1\}$ at $L \in \{8, 16\}$, with the $\beta = 0$, $L = 8$ negative control and the relevant-future controls $\beta = 0.8$, $L \in \{0, 4\}$. Known-lag OLS is evaluated only in the four primary conditions. These fitted-baseline results do not cover the additional $\beta = 0.4, 0.6$ conditions in Appendix D.3.

Forecast accuracy and lag selection. In each primary condition, lag-selected OLS selects the true lag in all 1,280 windows across five seeds. Its MAE is below 10^{-12} for the exact copies and is 0.4825 ± 0.0035 and 0.4833 ± 0.0037 for the noisy copies at $L = 8$ and 16, respectively (Table 21). The corresponding paired differences from the empirical same-information oracle are 0.0007 ± 0.0006 and 0.0015 ± 0.0007 (Table 22). The historical relationship is therefore recoverable from these finite contexts under a fitted lagged-linear model. Known-lag OLS produces identical forecasts in the primary conditions; providing the true lag does not improve this particular baseline there.

In the no-dependence control, lag-selected OLS chooses spurious lagged relationships and has higher mean MAE than the target-mean baseline. This illustrates the limits of selection from a finite context. For the relevant-future controls, supplying future covariates improves MAE, as expected when some or all required source values lie in the forecast horizon.

Future sensitivity and historical-source response. At $\beta \in \{0.8, 1\}$ and $L \in \{8, 16\}$, both OLS routes give identical forecasts and $S_{\text{future}} = 0$ in every seed. The selected lag places all required source values in observed history. This invariance is not universal: at $\beta = 0$, future-resampling sensitivity is 0.0091 ± 0.0005 because lag selection can introduce dependence on irrelevant source values. At $\beta = 0.8$, $L = 0$ and 4, the known-covariate route is sensitive to future values that are genuinely predictive.

For the response-operator diagnostic, each perturbed history triggers lag reselection and coefficient refitting; the finite difference therefore measures the whole estimator, including its adaptation. We use the same $\delta = 0.5$ perturbation and leading 64 windows per seed as in the TSFM experiment. The reported E_J is computed from the mean operator, not from the mean of seed-level errors. It is below 10^{-8} for exact copies and 0.00131 and 0.00271 for noisy copies at $L = 8$ and 16.

D.5 SENSITIVITY TO CONTEXT LENGTH

We additionally evaluate the noisy lagged-copy construction at context lengths 1,024 and 2,048, alongside the 512-step setting. Each setting uses $\beta = 0.8$, $H = 8$, $L \in \{8, 16\}$, and five generation seeds with 256 windows per seed. Within each context setting, the interface comparisons use matched histories and targets, and future sensitivity uses eight independent continuations per window. Table 23 reports the paired MAE reduction from PASTCOV to KC and the forecast change under future resampling. Uncertainty for the MAE reduction is computed across seed-level paired differences. Sensitivity entries report means only; the window-level SEMs in the corresponding experiment reports are not used as seed-level uncertainties.

Chronos-2's positive mean KC benefit at context length 512 does not persist at the longer contexts, where the mean differences are negative at both tested lags. Its sensitivity to irrelevant future values decreases with context length but remains nonzero. TimesFM-3 and TiRex-2 retain positive mean KC benefits and nonzero future sensitivity at all three context lengths. These results distinguish context-dependent accuracy changes from persistent sensitivity to conditionally irrelevant inputs; they do not support a context-independent accuracy benefit for Chronos-2.

E Forecast-Time Covariate Reliability Shift

E.1 Predictive Process and Matched Degradation

Clean predictive process. Unlike the conditionally irrelevant future in the lagged-copy task, the clean covariate here provides information about the target's future through a shared dynamic state:

$$z_t = \rho z_{t-1} + \sqrt{1-\rho^2}\, u_t, \tag{57}$$

$$X_t = \frac{z_t + \sigma_x \epsilon_t^x}{\sqrt{1+\sigma_x^2}}, \tag{58}$$

$$Y_t = \kappa z_t + \sigma_y \epsilon_t^y. \tag{59}$$

The innovations u_t, ϵ_t^x, and ϵ_t^y are mutually independent standard Gaussian white-noise processes. We use $\rho = 0.95$, $\kappa = 1$, $\sigma_x = 0.15$, and $\sigma_y = 1/\sqrt{2}$, with 512 burn-in steps. The stationary variances of z and X are one, and that of Y is 1.5. Each of five seeds (0–4) supplies 256 windows with a 512-step context and an eight-step forecast horizon. The target process and realization remain fixed across all reliability levels and matched conditions.

Covariate degradation. Given a clean covariate X, we construct

$$X_\lambda = \lambda X + \sqrt{1-\lambda^2}\, \xi, \tag{60}$$

where ξ is an independent realization with the same marginal variance and temporal dependence as X. We evaluate $\lambda \in \{1, 0.75, 0.5, 0.25, 0\}$. The endpoint $\lambda = 1$ recovers the clean covariate, whereas $\lambda = 0$ produces a covariate independent of it.

Applied to a complete segment, this transformation preserves marginal variance and within-segment autocorrelation. Applying it only on one side of the forecast boundary can also alter cross-boundary dependence. We therefore interpret the construction as a forecast-time reliability mismatch rather than a change in dependence strength alone.

Matched conditions. For each λ, we hold the target realization fixed and use the same clean and degraded covariate realizations to construct

$$\textsc{Stationary} : (X_\lambda^-, X_\lambda^+), \tag{61}$$

$$\textsc{Future-Only} : (X^-, X_\lambda^+), \tag{62}$$

$$\textsc{History-Only} : (X_\lambda^-, X^+), \tag{63}$$

where superscripts $-$ and $+$ denote historical and future segments. All conditions use the known-covariate interface. At $\lambda = 1$, the three inputs coincide.

The Stationary and Future-Only conditions share exactly the same future covariate X_λ^+. They differ only in whether the historical covariate has also been degraded. The History-Only condition evaluates the reverse mismatch, in which a degraded history precedes a clean future.

Reliability-shift penalty. Let R_M denote model MAE, with the unchanged target history omitted from the notation. We define

$$\Delta_{\text{shift}}(\lambda) = R_M(X^-, X_\lambda^+) - R_M(X_\lambda^-, X_\lambda^+). \tag{64}$$

Positive values indicate additional error when the same degraded future covariate follows a clean rather than equally degraded history. This is a behavioral comparison; it does not identify how the model internally represents or estimates covariate reliability.

Reporting unit for the shift penalty. The displayed Δ_{shift} uncertainties in Table 2 use paired window-level MAE differences across the five seeds (1,280 windows), rather than the SEM of five seed means. The two conditions are paired on the same target and degraded future realization. This uncertainty convention differs from the seed-level SEM used for the structural sweep and should be distinguished when reading the tables.

Predictive uncertainty. In addition to MAE, we evaluate empirical coverage and mean width of the 80% prediction intervals. Table 24 compares STATIONARY and FUTURE-ONLY at every tested reliability level. The clean endpoint coincides; at each degraded level, all three models show larger point-forecast errors, narrower intervals, and lower coverage under the future-only mismatch.

E.2 ACCURACY AND INTERVAL BEHAVIOR

Reliability mismatch and predictive uncertainty. At zero reliability ($\lambda = 0$), Chronos-2's MAE is 0.800 under STATIONARY and 1.259 under FUTURE-ONLY, compared with a CI MAE of 0.802 (Table 2). The corresponding STATIONARY and FUTURE-ONLY MAEs are 0.796 and 1.271 for TimesFM-3, and 0.797 and 1.232 for TiRex-2. Their mismatch penalties are 0.459 ± 0.020, 0.475 ± 0.020, and 0.435 ± 0.018 for Chronos-2, TimesFM-3, and TiRex-2, respectively, with paired SEMs. The reverse HISTORY-ONLY condition yields MAEs of 0.808, 0.792, and 0.798, respectively.

For Chronos-2, changing from STATIONARY to FUTURE-ONLY at $\lambda = 0$ reduces empirical 80% coverage from 0.790 to 0.452 and interval width from 2.524 to 1.879 (Table 24). The corresponding coverage and width changes are $0.797 \rightarrow 0.452$ and $2.541 \rightarrow 1.950$ for TimesFM-3, and $0.786 \rightarrow 0.453$ and $2.484 \rightarrow 1.898$ for TiRex-2. Thus, in each model, the increase in point-forecast error coincides with narrower intervals and lower empirical coverage.

F CONDITIONING ON PREDICTED COVARIATES

F.1 DATASETS AND EVALUATION PROTOCOL

We select three datasets from fev-bench (Shchur et al., 2026) with interpretable source–target relationships and aligned observations of both channels. The selection targets settings in which a relevant covariate is observed historically, but using its future values requires an estimate at forecast time.

GFC17 (air temperature → electricity load) provides a setting for weather-dependent demand forecasting (Hong et al., 2019). Prior work on GEFCom2017 generates temperature scenarios from historical observations and supplies them to load forecasting models to construct probabilistic forecasts (Dimoulkas et al., 2019). We use this setting to compare conditioning on a single predicted temperature path with propagation of multiple predicted paths through the known-future interface.

KDD (wind speed → active power) provides a physically motivated relationship in turbine-level measurements: the source is wind speed recorded by an anemometer, and the target is turbine active power. The original SDWPF challenge requires power forecasts without access to future meteorological observations (Zhou et al., 2022). Predicting wind speed before supplying it to a power forecaster therefore provides a natural test of conditioning on predicted covariates when future observations are unavailable.

Rossmann (customer count → sales) extends the analysis to retail, pairing daily customer counts with store turnover (FlorianKnauer & Cukierski, 2015). Customer count is available only as a historical covariate in the fev-bench task (Shchur et al., 2026). We use this pair to examine whether forecasting future customer counts and supplying these predictions to the sales forecaster preserves probabilistic forecast quality. The three tasks cover regional demand, turbine-level generation, and store-level revenue, with channel directions motivated by their documented roles.

Task adaptations. We retain the benchmark horizons and rolling forecast origins, restrict inputs to the selected source X and target Y, and cap the context at 512 observations. These are paired-channel adaptations of the benchmark tasks. All primary routes withhold realized future source values, including the temperature continuation exposed as a known covariate in the original fev-

bench GFC17 task. Other dynamic and static covariates are excluded. The experiments evaluate conditioning on predicted covariates under matched historical information.

Evaluation cohorts. Table 25 summarizes the task geometry and evaluated cohorts. KDD has 128,628 valid target horizon values. For Rossmann, 540 series–origin pairs have no finite target value in the forecast horizon and are excluded. Each primary route is scored on the same target mask and series–origin cohort within each dataset.

Scoring and aggregation. For the comparisons using native quantiles below, target point forecasts and supplied source paths are the models' native median forecasts. SQL averages twice the pinball loss over the common quantile grid $\{0.1, 0.2, \ldots, 0.9\}$ and valid horizon steps, then divides by the mean absolute seasonal difference in the raw, capped target context. The seasonal periods are given in Table 25; the denominator is identical across routes at a given series and origin. These native comparisons use the 0.1 and 0.9 quantiles without the reconstruction introduced below for path propagation. For each metric, we average origins within a series and then average series with equal weight within a dataset. These diagnostic SQL values therefore use a different context denominator and aggregation from the official benchmark score. We do not form SEMs from the dependent rolling forecasts or pool raw MAEs across datasets with different target units.

F.2 Point Accuracy and Predictive Intervals

Across the nine dataset–model combinations, both routes that condition on predicted source medians increase mean SQL relative to MV, with relative increases of 1.3–15.5% (Table 26). Empirical 80% interval coverage ranges from 69.4–84.0% under MV and from 25.5–60.9% when conditioning on predicted source medians. On KDD, both conditioning routes lower median-forecast MAE for every model, while their intervals are narrower and cover fewer observations (Table 27). Point-forecast improvements thus coexist with worse SQL and lower interval coverage in this dataset.

F.3 Quantile-Path Propagation

Source scenarios. We extend the same frozen models and evaluation cohorts by replacing the single multivariate source median with nine source paths, indexed by $\mathcal{A} = \{0.1, 0.2, \ldots, 0.9\}$. For each $a \in \mathcal{A}$, we hold the rank fixed across the horizon and concatenate its native quantile predictions to form $\widehat{\mathbf{x}}^{(a)}$. Each path is supplied separately to the same target KC interface, using the same observed histories. The source values are not sorted, clipped, rounded, smoothed, or otherwise repaired; possible source quantile crossings are retained. All nine scenario weights are fixed at $1/9$, with no fitting or selection using the evaluation data.

Target distributions and mixing. For each scenario and horizon step, we select the nine target quantiles on $\mathcal{A}$ and sort their values in ascending order, obtaining $q_1 \le \cdots \le q_9$. The primary reconstruction linearly interpolates the quantile function between consecutive ranks and holds it constant outside $[0.1, 0.9]$. Equivalently, each endpoint carries mass 0.1, with mass 0.1 on each of the eight interior segments; ties can add point mass. If $\widehat{F}^{(a)}_{Y,h}(\cdot \mid \mathcal{I})$ is the resulting conditional CDF for scenario a and observed histories $\mathcal{I}$, we compute

$$\widetilde{F}_{Y,h}(y \mid \mathcal{I}) = \frac{1}{9} \sum_{a \in \mathcal{A}} \widehat{F}^{(a)}_{Y,h}(y \mid \mathcal{I}), \tag{65}$$

$$\widetilde{Q}_{Y,h}(p \mid \mathcal{I}) = \inf\{y : \widetilde{F}_{Y,h}(y \mid \mathcal{I}) \ge p\}. \tag{66}$$

We invert the mixture CDF rather than average the component quantiles, and score the resulting quantiles on $\mathcal{A}$. The reconstructed median is used for MAE. Direct MV, median-path KC, and the target-history-only CI baseline receive the same target sorting and reconstruction, so comparisons isolate propagation from this postprocessing.

Reconstruction checks. All 36 available raw-native baseline dataset–model–route scores reproduce the original experiment exactly before postprocessing. Common-grid selection and sorting change baseline SQL by at most 1.05×10^{-5} and coverage by at most 0.154 percentage points. Thus the small differences between the native values in Table 26 and the reconstructed baselines in

Table 3 are intentional. Analytic mixture tests, nine-identical-component checks, component-order checks, and an independent bisection inversion all pass the specified numerical tolerances. The same target masks, seasonal denominators, and within-series then across-series aggregation are retained.

Primary comparison and tail sensitivity. Under the primary endpoint-clamped reconstruction, propagation reduces SQL by 1.02–12.75% relative to conditioning on the predicted source median across all nine dataset–model combinations. Coverage rises by 2.0–41.6 percentage points, but remains below the nominal 80% and direct MV coverage in every combination. These improvements need not coincide with better point forecasts: on KDD, TiRex-2 MAE changes from 205.354 to 205.591 while its coverage rises from 30.4% to 72.1%. As a prespecified sensitivity check, we instead add the knots $Q(0) = 2q_1 - q_2$ and $Q(1) = 2q_9 - q_8$, interpolating linearly through both tails without new inference. Table 28 reports both conventions. The improvements over conditioning on the predicted source median and remaining undercoverage persist, although small SQL differences from direct MV can change sign: for GFC17 TiRex-2, the propagated SQL changes from 0.932558 to 0.929137, compared with 0.929746 for direct MV.

Source forecast quality. Table 29 evaluates the unmodified multivariate source forecasts on available source observations, using the source's own seasonal scale s_X for SQL. Missing source truth restricts this diagnostic mask only; it does not remove cases from the primary target evaluation. Source coverage is itself imperfect, including undercoverage for all three models on GFC17 and KDD. The remaining target error therefore cannot be attributed solely to the KC interface or solely to path propagation.

Realized-source reference. A separate privileged condition supplies the realized future source to KC. It is evaluated only where the entire future source is available: 160 GFC17, 1,339 KDD, and 10,250 Rossmann series–origin pairs. Table 30 reaggregates direct MV and quantile-path propagation on those same subsets. The realized-source condition has lower SQL than either route in all nine comparisons, but is unavailable at forecast time and is neither a deployable method nor a guaranteed upper bound. Its coverage also varies around the nominal level.

Scope of the approximation. Holding one source rank across all horizon steps fixes a same-rank temporal coupling; it does not estimate the source's joint trajectory distribution. The nine equally weighted paths are prescribed scenarios, not verified samples or an exact quadrature of that distribution. Likewise, the target reconstruction estimates a marginal CDF at each horizon step and does not recover a joint target path law. The source forecast and KC output are not guaranteed to be conditionals of a shared joint model, so the mixture need not coincide with direct MV. The measured gains support uncertainty propagation as a partial mitigation of the deterioration caused by conditioning on predicted source medians under this protocol. They do not identify a single mechanism responsible for this deterioration.

G Representation Diagnostics

These experiments assess linearly accessible information in frozen target-side representations. They are separate from the behavioral sweeps and do not by themselves identify the mechanisms behind forecast changes.

G.1 Representation Diagnostics for Distributed Evidence

Data and comparison conditions. We use the shared-latent process described in Appendix C.1, with $\rho = 0.95$ and $\lambda = 0.75$. This separate experiment uses seeds $0, 1, 2$ and 256 windows per seed. The main diagnostic fixes the total number of auxiliary channels at $D = 16$ and varies

$$K_{\text{info}} \in \{1, 2, 4, 8, 16\}. \tag{67}$$

The remaining $D - K_{\text{info}}$ channels are matched distractors. All gains in Table 31 compare $K_{\text{info}} = 16$ with $K_{\text{info}} = 1$ at this fixed dimensionality. Unlike the channel-count sweep in Section 3, the baseline therefore contains one informative channel and 15 distractors.

We additionally examine the original duplicate and distractor configurations as their total auxiliary-channel count increases from $K = 1$ to $K = 16$.

Target-side latent-state probes. For each layer ℓ, we extract the hidden representation $h_Y^{(\ell)}$ associated with the scored target channel. At the final layer, this is the representation consumed by the forecasting head. We fit a linear ridge probe

$$\widehat{z}_T = W h_Y^{(\ell)} + b \tag{68}$$

to recover the latent state at the forecast origin. The probe never receives auxiliary-channel representations directly. For each model, we retain one target token without pooling: Chronos-2 uses zero-based token 33 (the future patch), TiRex-2 token 63 (the token producing the first requested forecast steps), and TimesFM-3 token 15 (the last context patch). At the head input, the vectors have dimensions 768, 512, and 1,280, respectively. Block outputs and the input to the forecast head are extracted separately, including any transformation between the final block and the head.

Evaluation uses leave-one-seed-out splits. Feature normalization, regularization selection, and probe fitting use training seeds only. Each outer split has 512 training windows and 256 held-out windows. The ridge probe includes an intercept, uses feature means and standard deviations estimated from the training windows, and centers z_T by its training mean. Inner leave-one-training-seed-out validation chooses the regularization coefficient from $\alpha \in \{10^{j/2} : j = -8, -7, \ldots, 16\}$ by mean validation MSE; the probe is then refitted on both training seeds.

Probe gains are reported as mean $\pm$ SEM over three held-out-seed folds; forecast gains are reported as mean $\pm$ SEM over three seeds.

Oracle reference. We compute the exact Gaussian posterior mean

$$\mu_K^\star = \mathbb{E}\left[z_T \mid Y_{1:T}, X_{1,1:T}, \ldots, X_{K,1:T}\right]. \tag{69}$$

For independent informative channels, the population oracle R^2 increases from 0.889 at $K = 1$ to 0.971 at $K = 16$. Duplicate and distractor channels add no target-relevant information, so their oracle R^2 remains unchanged.

Oracle and probe ΔR^2 values are not interpreted as directly comparable fractions of captured information. The oracle is an exact conditional expectation, whereas the probe is an out-of-sample linear estimator fitted to a pretrained representation. Likewise, probe R^2 and forecast MAE assess different quantities and are not compared on a common numerical scale.

Representation and forecast changes. Table 31 reports changes when informative channels increase from 1 to 16 at a fixed total of $D = 16$ auxiliary channels. Latent-state probe R^2 improves in all three models, with gains ranging from 0.0806 to 0.1042. Forecast MAE reductions range from 0.0086 to 0.0149, with all point estimates below the corresponding oracle gain of 0.0219. This comparison combines increased predictive evidence with reduced distraction, so it does not isolate the contribution of either change. Improved linear accessibility of z_T accompanies forecast improvement.

Duplicate and distractor controls. For Chronos-2, increasing distractors from $K = 1$ to $K = 16$ reduces head-input probe R^2 from 0.791 to 0.768 and increases MAE by 0.0163. In the duplicate condition, latent-state decodability remains nearly unchanged while MAE increases by 0.0082.

G.2 Target-Side Lag-Class Diagnostics

Data generation. We generate $X_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and, for each lag class L, construct

$$Y_t = \beta X_{t-L} + \sqrt{1-\beta^2}\,\epsilon_t, \tag{70}$$

where $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ is independent of X. For the no-dependence class, we instead use

$$Y_t = \beta X'_t + \sqrt{1-\beta^2}\,\epsilon_t, \tag{71}$$

with X' an independent standard Gaussian series. We set $\beta = 0.8$, use 512 context steps, and request an eight-step forecast from each model. Forecasts are used to check the extraction path, not to score predictive accuracy in this experiment.

For each base realization, all classes share the same X, ϵ, and X' arrays; lag classes differ only in the alignment of X used to generate Y. A fixed 64-step source prefix supports every lag without changing the supplied X history across classes. The predefined lag grids are

$$\mathcal{L}_{\text{primary}} = \{0, 5, 10, 15, 20\}, \tag{72}$$

$$\mathcal{L}_{\text{fine}} = \{0, 1, 2, 4, 8\}, \tag{73}$$

$$\mathcal{L}_{\text{coarse}} = \{0, 8, 16, 32, 64\}. \tag{74}$$

Each grid includes the no-dependence class, giving six balanced classes and chance accuracy $1/6$. We use five seeds with 128 base realizations per seed, yielding 768 windows per seed and 3,840 windows per grid. Results are evaluated separately for each grid; the grids are not treated as independent replications because they share base realizations and overlapping classes.

Input conditions and representation extraction. Under MV, the model receives $[X^-, Y^-]$; under CI, it receives the identical Y^- history alone. Neither condition receives observed future covariate values. All pretrained checkpoints remain frozen. At each block and at the forecasting-head input, we retain only the target row and one token, with no pooling or concatenation of covariate rows. The selected zero-based token positions are 33 for Chronos-2 (the future-patch token), 63 for TiRex-2 (the token producing the first requested forecast steps), and 15 for TimesFM-3 (the last context patch). The corresponding hidden dimensions are 768, 512, and 1,280. At the head input, we use the representation after any intervening final transformation, rather than substituting the last block output. The reported extraction check finds Spearman correlation one between the head median path at the selected token and the returned median forecast for every window in each model, mode, and grid.

Probe training and evaluation. We standardize features using training data and fit a multinomial logistic regression with an L2 penalty, using the `lbfgs` solver and a maximum of 2,000 iterations. Outer evaluation holds out one generation seed at a time. All class variants of a base realization remain in the same split. Within the four training seeds, inner leave-one-seed-out validation selects

$$C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\} \tag{75}$$

by mean validation accuracy, breaking ties in favor of the smaller C. The selected probe is refitted on all four training seeds and evaluated once on the held-out seed. We report mean accuracy with SEM over the five outer folds. Paired MV–CI differences are computed within each held-out seed before aggregation.

Controls. The target-only model representations provide a CI control. Additional probes use raw X^-, raw Y^-, or a fixed random nonlinear map of the concatenated histories:

$$\phi(X^-, Y^-) = \tanh\left([X^-; Y^-]W/\sqrt{1024}\right), \tag{76}$$

where the entries of W are independently standard Gaussian and its output width matches the respective model representation. All controls use the same splitting, normalization, and regularization-selection procedure. Low accuracy for these particular probes is not evidence that the raw joint histories lack lag information.

Results and interpretation. Table 32 summarizes head-input accuracy and paired MV–CI differences across grids. Raw and random-feature controls, coarse-grid class recall, and confusion counts are retained in Tables 33, 34, and 35. For Chronos-2, coarse-grid recall decreases at longer lags, with confusion between long-lag and no-dependence classes. This concerns the fitted linear probe, not the absence of information in the full representation.

Across depth, TimesFM-3's coarse-grid accuracy peaks at block 16 before decreasing toward the head input, whereas Chronos-2 largely maintains its late-layer accuracy (Figure 7).

These diagnostics assess linear accessibility of lag classes, not recovery of predictive source values or their causal use by the forecast head. The probe experiment uses the same lagged Gaussian family as Section 4.1, but separate realizations.

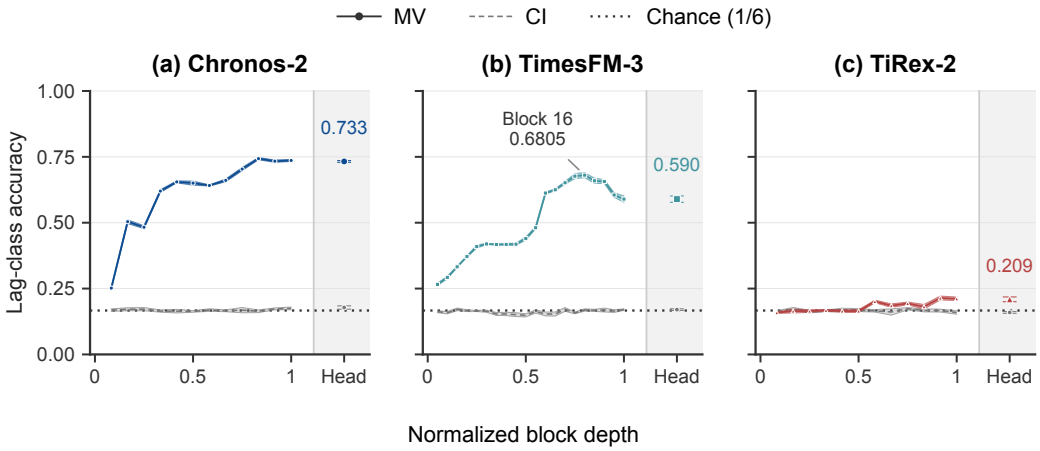

Figure 7: Layer-wise linear decodability of cross-channel lag relationships. Each layer is evaluated with a separate probe using only the target token at the forecast-readout position, under multivariate (MV) or target-only (CI) input. The coarse grid contains $L \in \{0, 8, 16, 32, 64\}$ and a no-dependence class; chance accuracy is $1/6$. Curves show mean accuracy with $\pm$ SEM over five held-out-seed folds. Block depth is normalized by the number of blocks (12, 20, and 12). The shaded *Head* region reports the actual head input separately, with SEM error bars; it does not represent an additional block.

Table 13: Paired changes from zero lag at the same response anchor, averaged equally over $t_{\text{resp}} \in \{24, 32, 40\}$ within each seed before reporting mean $\pm$ seed SEM. Negative ΔG indicates attenuation; positive ΔE indicates larger profile error. The lag, trigger position, and response position are linked by $t_{\text{trig}} = t_{\text{resp}} - \tau$.

τ	Event regression	Chronos-2	TimesFM-3	TiRex-2
ΔG				
4	-0.0083 ± 0.0024	-0.0660 ± 0.0012	-0.6112 ± 0.0044	-0.0103 ± 0.0017
8	0.0049 ± 0.0035	-0.1117 ± 0.0059	-0.7231 ± 0.0036	-0.0208 ± 0.0025
16	0.0050 ± 0.0046	-0.2501 ± 0.0041	-0.7721 ± 0.0033	-0.0318 ± 0.0031
ΔE				
4	-0.0272 ± 0.0030	0.0637 ± 0.0010	0.5014 ± 0.0044	0.0099 ± 0.0015
8	-0.0401 ± 0.0056	0.1072 ± 0.0057	0.6080 ± 0.0034	0.0205 ± 0.0023
16	-0.0328 ± 0.0073	0.2448 ± 0.0039	0.6543 ± 0.0034	0.0313 ± 0.0029

Table 14: MAE changes corresponding to Figure 3 at $\rho = 0.95$ and $\lambda = 0.75$. Values are mean $\pm$ paired SEM over five seeds relative to each method's noisy $K = 1$ baseline. Positive values indicate improvement in (a) and degradation in (b–c). Oracle changes in (b–c) are exactly zero.

Model	$K = 2$	$K = 4$	$K = 8$	$K = 16$
(a) Independent channels: MAE reduction				
Oracle	0.00751 ± 0.00140	0.01439 ± 0.00130	0.01952 ± 0.00250	0.02402 ± 0.00200
Chronos-2	0.00723 ± 0.00330	0.01322 ± 0.00390	0.01168 ± 0.00510	0.00605 ± 0.00560
TimesFM-3	0.00520 ± 0.00230	0.01101 ± 0.00240	0.01736 ± 0.00290	0.02043 ± 0.00390
TiRex-2	0.00008 ± 0.00170	0.00186 ± 0.00180	0.00243 ± 0.00190	0.00255 ± 0.00210
(b) Duplicate channels: MAE increase				
Chronos-2	0.00060 ± 0.00031	0.00201 ± 0.00070	0.00467 ± 0.00100	0.00787 ± 0.00130
TimesFM-3	-0.00079 ± 0.00060	-0.00017 ± 0.00100	0.00103 ± 0.00120	0.00225 ± 0.00130
TiRex-2	-0.00049 ± 0.00018	-0.00076 ± 0.00034	-0.00079 ± 0.00047	-0.00078 ± 0.00061
(c) Distractor channels: MAE increase				
Chronos-2	-0.00052 ± 0.00210	-0.00023 ± 0.00410	0.00502 ± 0.00570	0.01105 ± 0.00560
TimesFM-3	0.00045 ± 0.00083	0.00027 ± 0.00110	0.00043 ± 0.00120	0.00011 ± 0.00130
TiRex-2	0.00621 ± 0.00120	0.00839 ± 0.00140	0.00835 ± 0.00140	0.00783 ± 0.00140

Table 15: MAE reductions across latent persistence levels. (a) True latent history versus target-only input. (b) Independent $K = 16$ observations versus noisy $K = 1$. Values are mean $\pm$ paired SEM over five seeds in raw MAE units; each method uses its own baseline.

ρ	Chronos-2	TimesFM-3	TiRex-2	Oracle
(a) Direct latent access				
0	-0.00033 ± 0.00042	0.00011 ± 0.00013	0.00042 ± 0.00014	0.00000 ± 0.00000
0.3	0.00013 ± 0.00046	0.00138 ± 0.00026	0.00075 ± 0.00019	0.00237 ± 0.00049
0.6	0.00066 ± 0.00095	0.00510 ± 0.00076	0.00136 ± 0.00018	0.00905 ± 0.00110
0.9	0.02871 ± 0.00350	0.02555 ± 0.00190	0.00669 ± 0.00150	0.03558 ± 0.00240
0.95	0.03570 ± 0.00570	0.03823 ± 0.00310	0.01022 ± 0.00230	0.04342 ± 0.00330
0.98	0.03723 ± 0.00580	0.04174 ± 0.00460	0.01176 ± 0.00220	0.04287 ± 0.00420
(b) Independent observations, $K = 16$				
0	0.00102 ± 0.00038	-0.00035 ± 0.00039	0.00015 ± 0.00019	0.00000 ± 0.00000
0.3	-0.00194 ± 0.00095	-0.00153 ± 0.00140	0.00011 ± 0.00009	0.00166 ± 0.00063
0.6	-0.00090 ± 0.00140	0.00118 ± 0.00220	0.00028 ± 0.00021	0.00577 ± 0.00140
0.9	0.00780 ± 0.00520	0.01736 ± 0.00440	0.00218 ± 0.00075	0.02111 ± 0.00150
0.95	0.00605 ± 0.00560	0.02043 ± 0.00390	0.00255 ± 0.00210	0.02402 ± 0.00200
0.98	0.00329 ± 0.00440	0.02077 ± 0.00240	0.00358 ± 0.00140	0.02145 ± 0.00210

Table 16: State-space baseline at $\rho = 0.95$: MAE and paired difference from the same-information oracle (baseline minus oracle). Entries are mean $\pm$ SEM across five seeds. This is the only persistence value evaluated for this baseline. Exact duplicate columns are removed before fitting.

Condition	Fitted SSM MAE	MAE $-$ oracle
Target only	0.7811 ± 0.0118	0.0058 ± 0.0017
True latent history	0.7358 ± 0.0083	0.0040 ± 0.0018
Noisy $K = 1$	0.7621 ± 0.0116	0.0022 ± 0.0014
Independent $K = 2$	0.7546 ± 0.0098	0.0022 ± 0.0011
Independent $K = 4$	0.7493 ± 0.0104	0.0038 ± 0.0005
Independent $K = 8$	0.7470 ± 0.0106	0.0067 ± 0.0023
Independent $K = 16$	0.7404 ± 0.0095	0.0045 ± 0.0018
Duplicate $K = 2$	0.7621 ± 0.0116	0.0022 ± 0.0014
Duplicate $K = 4$	0.7621 ± 0.0116	0.0022 ± 0.0014
Duplicate $K = 8$	0.7621 ± 0.0116	0.0022 ± 0.0014
Duplicate $K = 16$	0.7621 ± 0.0116	0.0022 ± 0.0014
Distractor $K = 2$	0.7622 ± 0.0118	0.0023 ± 0.0016
Distractor $K = 4$	0.7625 ± 0.0120	0.0026 ± 0.0022
Distractor $K = 8$	0.7624 ± 0.0116	0.0025 ± 0.0019
Distractor $K = 16$	0.7642 ± 0.0100	0.0043 ± 0.0016

Table 17: Paired MAE differences between the fitted state-space model and each TSFM at $\rho = 0.95$, using the same observations in each condition. Negative values favor the fitted baseline. Entries are mean $\pm$ seed SEM.

Condition	SSM $-$ Chronos-2	SSM $-$ TimesFM-3	SSM $-$ TiRex-2
Target only	-0.0146 ± 0.0041	-0.0106 ± 0.0025	-0.0116 ± 0.0032
True latent history	-0.0242 ± 0.0060	-0.0176 ± 0.0032	-0.0467 ± 0.0058
Noisy $K = 1$	-0.0165 ± 0.0037	-0.0170 ± 0.0033	-0.0224 ± 0.0044
Independent $K = 2$	-0.0168 ± 0.0032	-0.0193 ± 0.0031	-0.0298 ± 0.0036
Independent $K = 4$	-0.0161 ± 0.0035	-0.0188 ± 0.0043	-0.0333 ± 0.0043
Independent $K = 8$	-0.0199 ± 0.0050	-0.0147 ± 0.0055	-0.0350 ± 0.0045
Independent $K = 16$	-0.0322 ± 0.0049	-0.0183 ± 0.0051	-0.0416 ± 0.0055
Duplicate $K = 2$	-0.0171 ± 0.0037	-0.0162 ± 0.0029	-0.0219 ± 0.0044
Duplicate $K = 4$	-0.0185 ± 0.0038	-0.0168 ± 0.0028	-0.0216 ± 0.0045
Duplicate $K = 8$	-0.0212 ± 0.0040	-0.0180 ± 0.0027	-0.0216 ± 0.0045
Duplicate $K = 16$	-0.0244 ± 0.0042	-0.0192 ± 0.0025	-0.0216 ± 0.0046
Distractor $K = 2$	-0.0159 ± 0.0028	-0.0173 ± 0.0028	-0.0285 ± 0.0041
Distractor $K = 4$	-0.0159 ± 0.0030	-0.0169 ± 0.0023	-0.0304 ± 0.0030
Distractor $K = 8$	-0.0212 ± 0.0032	-0.0171 ± 0.0022	-0.0304 ± 0.0033
Distractor $K = 16$	-0.0254 ± 0.0028	-0.0149 ± 0.0029	-0.0281 ± 0.0043

Table 18: Exploratory paired comparison of independent $K = 16$ versus $K = 4$ channels at $\rho = 0.95$, $\lambda = 0.75$. Positive differences mean that the larger channel set has higher MAE. All methods share 256 valid windows per seed. Uncertainty is over five paired seed means; intervals are descriptive 95% Student-t intervals with four degrees of freedom.

Method	MAE(16)$-$MAE(4)	95% interval
Chronos-2	0.00717 ± 0.00200	$[0.00162, 0.01271]$
TimesFM-3	-0.00943 ± 0.00196	$[-0.01487, -0.00398]$
TiRex-2	-0.00069 ± 0.00043	$[-0.00188, 0.00050]$
Oracle	-0.00963 ± 0.00119	$[-0.01293, -0.00632]$
Fitted SSM	-0.00892 ± 0.00201	$[-0.01449, -0.00335]$

Table 19: **Future-covariate sensitivity across irrelevant and relevant regimes.** S_{future} measures the mean pairwise absolute change in the median forecast across eight independently sampled future-covariate continuations while holding the history fixed. The $\beta = 0$ condition is a negative control in which the covariate is unrelated to the target. For $\beta = 0.8$, r_{future} is the fraction of forecast steps whose target-relevant source lies after the forecast origin. When $r_{\text{future}} = 0$, the future covariate is conditionally irrelevant and the oracle sensitivity is zero. For $L < H$, the oracle values are analytically obtained from the data-generating process as $S^{\star}_{\text{future}} = r_{\text{future}}\beta\, 2/\sqrt{\pi}$. Uncertainties denote SEMs across five seed-level means.

Model	β	L	r_{future}	S_{future}	Oracle
Chronos-2	0	8	0	0.0152 ± 0.0006	0
Chronos-2	0.8	16	0	0.1545 ± 0.0005	0
Chronos-2	0.8	8	0	0.1619 ± 0.0009	0
Chronos-2	0.8	4	0.5	0.5429 ± 0.0018	0.451
Chronos-2	0.8	0	1	0.9359 ± 0.0020	0.903
TimesFM-3	0	8	0	0.0218 ± 0.0006	0
TimesFM-3	0.8	16	0	0.0531 ± 0.0008	0
TimesFM-3	0.8	8	0	0.1141 ± 0.0014	0
TimesFM-3	0.8	4	0.5	0.2912 ± 0.0019	0.451
TimesFM-3	0.8	0	1	0.9130 ± 0.0015	0.903
TiRex-2	0	8	0	0.0126 ± 0.0006	0
TiRex-2	0.8	16	0	0.0613 ± 0.0007	0
TiRex-2	0.8	8	0	0.1285 ± 0.0007	0
TiRex-2	0.8	4	0.5	0.3953 ± 0.0006	0.451
TiRex-2	0.8	0	1	0.9124 ± 0.0032	0.903

Table 20: Known-future conditioning across relation strengths at $H = 8$. $\Delta_{\mathrm{KC}} = \mathrm{MAE}_{\textsc{PastCov}} - \mathrm{MAE}_{\mathrm{KC}}$; positive values indicate lower MAE under KC. Δ_{KC} and S_{future} are mean $\pm$ SEM over five seeds; MAE differences are paired within each seed. E_J is computed from the mean response operator, not averaged over seed-level errors. Future covariates are conditionally irrelevant in every row, so oracle sensitivity is zero. The $\beta = 0$, $L = 16$ control was not run; E_J is undefined at $\beta = 0$ (—).

β	L	Δ_{KC}	S_{future}	E_J (PastCov)	E_J (KC)
Chronos-2					
0	8	-0.0002 ± 0.0003	0.0152 ± 0.0006	—	—
0.4	8	0.0074 ± 0.0019	0.0667 ± 0.0014	0.868	0.700
0.4	16	0.0009 ± 0.0010	0.0499 ± 0.0013	0.895	0.803
0.6	8	0.0254 ± 0.0027	0.1504 ± 0.0006	0.530	0.193
0.6	16	0.0226 ± 0.0027	0.1213 ± 0.0013	0.688	0.446
0.8	8	0.0067 ± 0.0037	0.1619 ± 0.0009	0.144	0.062
0.8	16	0.0282 ± 0.0037	0.1545 ± 0.0005	0.390	0.247
1	8	-0.0133 ± 0.0007	0.1480 ± 0.0006	0.084	0.092
1	16	0.0176 ± 0.0030	0.1462 ± 0.0004	0.256	0.222
TimesFM-3					
0	8	-0.0002 ± 0.0004	0.0218 ± 0.0006	—	—
0.4	8	0.0002 ± 0.0002	0.0277 ± 0.0003	0.996	0.989
0.4	16	0.0003 ± 0.0003	0.0234 ± 0.0005	0.997	0.994
0.6	8	0.0076 ± 0.0016	0.0524 ± 0.0006	0.995	0.960
0.6	16	0.0026 ± 0.0010	0.0295 ± 0.0007	0.997	0.983
0.8	8	0.0410 ± 0.0026	0.1141 ± 0.0014	0.991	0.890
0.8	16	0.0205 ± 0.0016	0.0531 ± 0.0008	0.994	0.939
1	8	0.0876 ± 0.0039	0.1645 ± 0.0008	0.981	0.855
1	16	0.0523 ± 0.0022	0.0851 ± 0.0007	0.985	0.894
TiRex-2					
0	8	0.0000 ± 0.0002	0.0126 ± 0.0006	—	—
0.4	8	0.0074 ± 0.0023	0.0342 ± 0.0011	0.996	0.892
0.4	16	0.0008 ± 0.0013	0.0194 ± 0.0003	0.998	0.978
0.6	8	0.0479 ± 0.0063	0.0797 ± 0.0015	0.997	0.733
0.6	16	0.0154 ± 0.0015	0.0351 ± 0.0007	0.999	0.908
0.8	8	0.1313 ± 0.0063	0.1285 ± 0.0007	0.997	0.673
0.8	16	0.0725 ± 0.0013	0.0613 ± 0.0007	0.999	0.801
1	8	0.2505 ± 0.0041	0.1626 ± 0.0005	0.997	0.666
1	16	0.1578 ± 0.0019	0.0746 ± 0.0006	0.999	0.769

Table 21: Lagged-copy baselines fitted separately to each observed context. Entries are mean $\pm$ seed SEM over five seeds. Lag-selected OLS is evaluated with history-only inputs (Hist.) or known-future covariates (KC). S_{future} is measured under the KC route. The privileged known-lag OLS has identical forecasts in the four primary rows ($\beta \in \{0.8, 1\}$, $L \in \{8, 16\}$); it was not run for the controls. At $\beta = 1$, numerical MAEs are below 10^{-12}.

β	L	Target mean	OLS (Hist.)	OLS (KC)	S_{future}
1	8	0.8030 ± 0.0075	$< 10^{-12}$	$< 10^{-12}$	0.0000 ± 0.0000
1	16	0.8003 ± 0.0048	$< 10^{-12}$	$< 10^{-12}$	0.0000 ± 0.0000
0.8	8	0.8002 ± 0.0047	0.4825 ± 0.0035	0.4825 ± 0.0035	0.0000 ± 0.0000
0.8	16	0.8017 ± 0.0049	0.4833 ± 0.0037	0.4833 ± 0.0037	0.0000 ± 0.0000
0	8	0.8036 ± 0.0059	0.8100 ± 0.0061	0.8102 ± 0.0062	0.0091 ± 0.0005
0.8	0	0.7912 ± 0.0041	0.7913 ± 0.0042	0.4831 ± 0.0038	0.8998 ± 0.0020
0.8	4	0.8012 ± 0.0050	0.6420 ± 0.0034	0.4826 ± 0.0036	0.4492 ± 0.0009

Table 22: Lag-selected OLS: paired MAE difference from the same-information oracle (baseline minus oracle; mean ± seed SEM) and historical-source response error. E_J is computed from the mean operator after lag reselection and refitting at each perturbed history. It is not the mean of seed-level errors. Hist. and KC give the same results in these four conditions.

β	L	MAE − oracle	E_J
1	8	$< 10^{-12}$	$< 10^{-8}$
1	16	$< 10^{-12}$	$< 10^{-8}$
0.8	8	0.0007 ± 0.0006	0.00131
0.8	16	0.0015 ± 0.0007	0.00271

Table 23: Context-length sensitivity at $\beta = 0.8$ and $H = 8$. $\Delta_{\mathrm{KC}} = \mathrm{MAE}_{\mathrm{PastCov}} - \mathrm{MAE}_{\mathrm{KC}}$ is mean ± paired SEM over five seeds; positive values favor KC. S_{future} entries report means only.

	$L = 8$		$L = 16$	
Context	Δ_{KC}	S_{future}	Δ_{KC}	S_{future}
Chronos-2				
512	0.0067 ± 0.0037	0.1619	0.0282 ± 0.0037	0.1545
1024	-0.0012 ± 0.0013	0.1197	-0.0007 ± 0.0021	0.1246
2048	-0.0048 ± 0.0006	0.0946	-0.0090 ± 0.0023	0.1002
TimesFM-3				
512	0.0410 ± 0.0026	0.1141	0.0205 ± 0.0016	0.0531
1024	0.0493 ± 0.0045	0.1243	0.0302 ± 0.0027	0.0657
2048	0.0507 ± 0.0045	0.1315	0.0339 ± 0.0025	0.0653
TiRex-2				
512	0.1313 ± 0.0063	0.1285	0.0725 ± 0.0013	0.0613
1024	0.1325 ± 0.0056	0.1346	0.0882 ± 0.0049	0.0668
2048	0.1342 ± 0.0031	0.1379	0.1029 ± 0.0030	0.0711

Table 24: Predictive uncertainty across covariate reliability levels. Each pair shows Stationary → Future-only with identical degraded future values. Entries are means; coverage and width refer to the nominal 80% interval.

λ	MAE	$\mathrm{Coverage}_{80}$	Width_{80}
Chronos-2			
1.00	$0.589 \to 0.589$	$0.778 \to 0.778$	$1.808 \to 1.808$
0.75	$0.723 \to 0.812$	$0.794 \to 0.630$	$2.294 \to 1.823$
0.50	$0.778 \to 0.985$	$0.791 \to 0.548$	$2.448 \to 1.841$
0.25	$0.797 \to 1.129$	$0.790 \to 0.491$	$2.502 \to 1.860$
0.00	$0.800 \to 1.259$	$0.790 \to 0.452$	$2.524 \to 1.879$
TimesFM-3			
1.00	$0.580 \to 0.580$	$0.792 \to 0.792$	$1.847 \to 1.847$
0.75	$0.707 \to 0.814$	$0.800 \to 0.641$	$2.271 \to 1.874$
0.50	$0.765 \to 0.992$	$0.802 \to 0.554$	$2.461 \to 1.900$
0.25	$0.792 \to 1.141$	$0.797 \to 0.493$	$2.531 \to 1.925$
0.00	$0.796 \to 1.271$	$0.797 \to 0.452$	$2.541 \to 1.950$
TiRex-2			
1.00	$0.585 \to 0.585$	$0.798 \to 0.798$	$1.872 \to 1.872$
0.75	$0.711 \to 0.809$	$0.775 \to 0.641$	$2.174 \to 1.880$
0.50	$0.770 \to 0.976$	$0.773 \to 0.555$	$2.328 \to 1.886$
0.25	$0.795 \to 1.113$	$0.781 \to 0.496$	$2.446 \to 1.892$
0.00	$0.797 \to 1.232$	$0.786 \to 0.453$	$2.484 \to 1.898$

Table 25: Task geometry for the real-data experiments. Origins per series are the scheduled benchmark origins; evaluated pairs exclude origins with no finite future target value.

Dataset	H	Series	Origins per series	Evaluated pairs	Seasonality
GFC17	168	8	20	160	24
KDD	96	134	10	1,340	48
Rossmann	48	1,115	10	10,610	7

Table 26: Conditioning on predicted source medians, evaluated using native target quantiles. CI→KC predicts the source from its own history; MV→KC uses both channel histories. Both supply one source median path. Coverage is for a nominal 80% interval.

		SQL ↓			Coverage (%)		
Dataset	Model	MV	CI→KC	MV→KC	MV	CI→KC	MV→KC
GFC17	Chronos-2	0.968	1.027	1.062	69.4	48.6	46.2
	TimesFM-3	0.954	0.977	0.996	71.2	55.7	53.2
	TiRex-2	0.930	0.943	0.942	71.4	60.8	60.9
KDD	Chronos-2	0.828	0.911	0.925	71.7	46.0	46.1
	TimesFM-3	0.806	0.863	0.931	71.7	27.8	25.5
	TiRex-2	0.735	0.834	0.839	84.0	30.3	30.4
Rossmann	Chronos-2	0.492	0.556	0.554	76.1	36.4	36.6
	TimesFM-3	0.528	0.597	0.596	82.3	36.1	36.1
	TiRex-2	0.570	0.614	0.613	79.7	46.6	46.8

Table 27: KDD point accuracy and interval behavior when conditioning on predicted covariates. Each entry shows the matched MV value followed by the value obtained by conditioning on the predicted source median. MAE and width are in the target's units; coverage is for the nominal 80% interval.

Model	Conditioning route	MAE	Width	Coverage (%)
Chronos-2	CI→KC	$251.262 \to 244.408$	$788.900 \to 262.052$	$71.7 \to 46.0$
	MV→KC	$251.262 \to 249.111$	$788.900 \to 261.743$	$71.7 \to 46.1$
TimesFM-3	CI→KC	$243.878 \to 216.073$	$728.253 \to 97.513$	$71.7 \to 27.8$
	MV→KC	$243.878 \to 239.434$	$728.253 \to 108.338$	$71.7 \to 25.5$
TiRex-2	CI→KC	$211.746 \to 204.183$	$620.007 \to 68.824$	$84.0 \to 30.3$
	MV→KC	$211.746 \to 205.354$	$620.007 \to 71.448$	$84.0 \to 30.4$

Table 28: Tail sensitivity of quantile-path propagation. Clamped is the primary reconstruction; Linear extends the two endpoint quantile segments. Both use identical predictions and cohorts.

		SQL ↓		Coverage (%)	
Dataset	Model	Clamped	Linear	Clamped	Linear
GFC17	Chronos-2	0.9978	0.9932	59.6	61.8
	TimesFM-3	0.9536	0.9503	64.2	66.5
	TiRex-2	0.9326	0.9291	62.9	65.4
KDD	Chronos-2	0.8631	0.8623	68.7	70.4
	TimesFM-3	0.8123	0.8116	65.7	66.3
	TiRex-2	0.7369	0.7367	72.1	72.5
Rossmann	Chronos-2	0.5031	0.5021	65.3	66.0
	TimesFM-3	0.5369	0.5366	72.0	72.6
	TiRex-2	0.5741	0.5742	75.3	75.9

Table 29: Native multivariate source forecast quality. Coverage and the lower/upper miss rates refer to the nominal 80% source interval. MAE is in source units; SQL uses s_X.

Dataset	Model	MAE	SQL	Coverage (%)	Lower (%)	Upper (%)
GFC17	Chronos-2	8.120	1.079	68.9	22.3	8.8
	TimesFM-3	7.792	1.028	68.2	23.2	8.6
	TiRex-2	7.905	1.053	70.6	20.7	8.7
KDD	Chronos-2	2.008	0.805	73.0	15.2	11.8
	TimesFM-3	1.896	0.756	72.9	15.3	11.8
	TiRex-2	1.712	0.692	77.7	10.2	12.1
Rossmann	Chronos-2	99.897	0.542	78.7	8.9	12.3
	TimesFM-3	103.433	0.570	83.4	6.9	9.7
	TiRex-2	112.718	0.635	82.6	8.6	8.8

Table 30: Matched realized-source reference under the primary target reconstruction. Paths denotes quantile-path MV→KC; True-X denotes KC given realized future source values. Cohorts differ from the main table where future source observations are missing.

		SQL ↓			Coverage (%)		
Dataset	Model	Direct MV	Paths	True-X	Direct MV	Paths	True-X
GFC17	Chronos-2	0.968	0.998	0.619	69.4	59.6	79.9
	TimesFM-3	0.954	0.954	0.648	71.2	64.2	80.3
	TiRex-2	0.930	0.933	0.753	71.4	62.9	75.0
KDD	Chronos-2	0.828	0.863	0.184	71.9	68.7	85.3
	TimesFM-3	0.807	0.813	0.150	71.8	65.7	81.7
	TiRex-2	0.735	0.737	0.142	84.1	72.1	90.4
Rossmann	Chronos-2	0.459	0.470	0.161	76.5	65.5	76.1
	TimesFM-3	0.490	0.497	0.149	83.0	72.6	78.3
	TiRex-2	0.520	0.523	0.194	81.8	77.4	87.7

Table 31: Representation and forecast gains when informative channels increase from 1 to 16 at a fixed total of $D = 16$ auxiliary channels. ΔR_z^2 measures improved target-side latent-state recovery; Δ_M measures forecast MAE reduction. Oracle gains are $\Delta R_z^2 = 0.0825$ and $\Delta^\star = 0.0219$. Values are mean $\pm$ SEM over three held-out-seed folds for probes and three seeds for forecasts.

Model	Probe gain ΔR_z^2	Forecast gain Δ_M
Chronos-2	0.1042 ± 0.0015	0.0142 ± 0.0065
TimesFM-3	0.1009 ± 0.0087	0.0149 ± 0.0038
TiRex-2	0.0806 ± 0.0062	0.0086 ± 0.0024

Table 32: Head-input lag-class accuracy across six-class grids (chance $1/6$). Values are mean $\pm$ SEM over five held-out-seed folds; MV–CI uses paired differences. Grids share base realizations and are not independent replications.

Model	MV	CI	MV–CI
Primary: $\{0, 5, 10, 15, 20\}$, plus no dependence			
Chronos-2	0.9612 ± 0.0042	0.1693 ± 0.0056	0.7919 ± 0.0091
TimesFM-3	0.5398 ± 0.0032	0.1651 ± 0.005	0.3747 ± 0.0046
TiRex-2	0.1974 ± 0.005	0.1674 ± 0.0048	0.02995 ± 0.0072
Fine: $\{0, 1, 2, 4, 8\}$, plus no dependence			
Chronos-2	0.9359 ± 0.0052	0.1664 ± 0.0082	0.7695 ± 0.0078
TimesFM-3	0.5552 ± 0.0075	0.1581 ± 0.0025	0.3971 ± 0.0087
TiRex-2	0.1958 ± 0.0069	0.1714 ± 0.0065	0.02448 ± 0.012
Coarse: $\{0, 8, 16, 32, 64\}$, plus no dependence			
Chronos-2	0.7326 ± 0.0032	0.1781 ± 0.0062	0.5544 ± 0.0045
TimesFM-3	0.5896 ± 0.012	0.1701 ± 0.0031	0.4195 ± 0.015
TiRex-2	0.2091 ± 0.0091	0.1594 ± 0.0032	0.04974 ± 0.0093

Table 33: Raw-history and fixed random-feature control accuracies (mean $\pm$ SEM over five folds). Random features use both histories, with output widths matching Chronos-2 (768), TimesFM-3 (1,280), and TiRex-2 (512). Chance accuracy is $1/6$.

Probe input	Width	Primary	Fine	Coarse
Raw X^-	512	0.1667 ± 0	0.1667 ± 0	0.1667 ± 0
Raw Y^-	512	0.1747 ± 0.006	0.1737 ± 0.0043	0.1669 ± 0.0043
Random features	768	0.1622 ± 0.0036	0.1628 ± 0.0034	0.1672 ± 0.0062
Random features	1,280	0.1651 ± 0.0062	0.168 ± 0.007	0.1609 ± 0.0053
Random features	512	0.1674 ± 0.0057	0.1643 ± 0.0064	0.1682 ± 0.0043

Table 34: Per-class recall at the head input for the coarse grid, aggregated over the held-out predictions. None denotes no cross-channel dependence. These are class-specific recalls, not six-class accuracies.

Model	Input	$L = 0$	$L = 8$	$L = 16$	$L = 32$	$L = 64$	None
Chronos-2	MV	0.997	0.973	0.936	0.605	0.419	0.466
	CI	0.217	0.211	0.128	0.113	0.183	0.217
TimesFM-3	MV	0.989	0.498	0.469	0.534	0.48	0.567
	CI	0.145	0.186	0.178	0.161	0.158	0.192
TiRex-2	MV	0.223	0.216	0.206	0.178	0.227	0.205
	CI	0.133	0.173	0.152	0.159	0.155	0.184

Table 35: Coarse-grid head-input confusion matrices under MV input. Rows are true classes and columns are predicted classes. Counts are summed across five held-out-seed folds, giving 640 examples per true class. None denotes no cross-channel dependence. Bold entries are correct predictions.

True / predicted	$L = 0$	$L = 8$	$L = 16$	$L = 32$	$L = 64$	None
Chronos-2						
$L = 0$	**638**	0	0	0	0	2
$L = 8$	1	**623**	9	4	1	2
$L = 16$	0	2	**599**	36	2	1
$L = 32$	0	1	15	**387**	131	106
$L = 64$	1	0	3	116	**268**	252
None	1	1	2	85	253	**298**
TimesFM-3						
$L = 0$	**633**	2	1	0	2	2
$L = 8$	2	**319**	126	46	64	83
$L = 16$	1	133	**300**	102	62	42
$L = 32$	0	66	85	**342**	98	49
$L = 64$	5	55	41	74	**307**	158
None	5	57	39	45	131	**363**
TiRex-2						
$L = 0$	**143**	127	93	77	109	91
$L = 8$	105	**138**	125	89	95	88
$L = 16$	113	109	**132**	108	91	87
$L = 32$	98	90	109	**114**	117	112
$L = 64$	85	94	77	111	**145**	128
None	98	98	83	103	127	**131**