ONE MODEL, THREE VIEWS: INPUT-CONDITION INTEGRATION FOR PROBABILISTIC TIME SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent time series foundation models (TSFMs) can condition on known future covariates, but this conditioning helps some targets and samples while hurting others, and which ones cannot be told in advance. By varying only its inputs, a single frozen TSFM yields three views of the same forecast: from the target's history alone (target-only), with past covariates added (multivariate-history), and with known future covariate values added (known-future). We show that replaying how forecasts change across these views on completed historical events reveals whether richer conditioning was beneficial, harmful, or insufficient. Based on this finding, we propose *input-condition integration*, which corrects the known-future forecast using the other two views as references rather than as candidates to select or average. The correction is learned from the observed input history alone and applied only when it proves reliable there, leaving the TSFM unchanged. Across three TSFM backbones, input-condition integration improves probabilistic accuracy over the known-future forecast on real-world benchmarks and better recovers covariate-driven responses in controlled synthetic tasks. These results suggest that the value of additional inputs lies not only in the information they supply, but in how the model responds to them.

1 INTRODUCTION

Time series foundation models (TSFMs) leverage large-scale pretraining to learn transferable forecasting capabilities across diverse domains (Rasul et al., 2023; Das et al., 2024; Goswami et al., 2024; Ansari et al., 2024; Liu et al., 2025b; Woo et al., 2024; Liu et al., 2025a; Cohen et al., 2024; Khwaja et al., 2026; Auer et al., 2026). Recent covariate-aware TSFMs further incorporate related-variable histories and covariates whose future values are known at prediction time (Ansari et al., 2025; Podest et al., 2026; Jain & Sen, 2026; Meyer et al., 2026). Results on fev-bench demonstrate substantial gains from such auxiliary information (Shchur et al., 2025). In many forecasting instances, adding known-future covariates substantially changes the prediction and brings it closer to the realized target, illustrating that richer conditioning can be highly informative.

However, access to predictive covariates does not guarantee that their relationship with the target is used reliably. Recent diagnostic studies show that strong benchmark performance can coexist with covariate underutilization: TSFMs may overlook leading indicators or model even simple target–covariate relationships unreliably (Ghoroghchian et al., 2026; Berthelier et al., 2026). Thus, the availability of additional information and the quality of the forecast response induced by that information are distinct.

We study this gap through delayed covariate–target relationships. Figure 1 illustrates the behavior with Chronos-2. Keeping the predictive relationship and covariate signal fixed, we increase only the time gap between the signal and the target response. Although Chronos-2 is pretrained with synthetic multivariate dependencies that include temporal interactions, its forecast response becomes increasingly attenuated as the delay grows. Analogous underutilization is also observed in TimesFM-3 and TiRex-2, despite differences in their response strength and degradation patterns (Appendix D).

This motivates us to use the multiple input conditions of a covariate-aware TSFM as a diagnostic rather than assuming that richer conditioning should always be trusted. For the same forecasting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

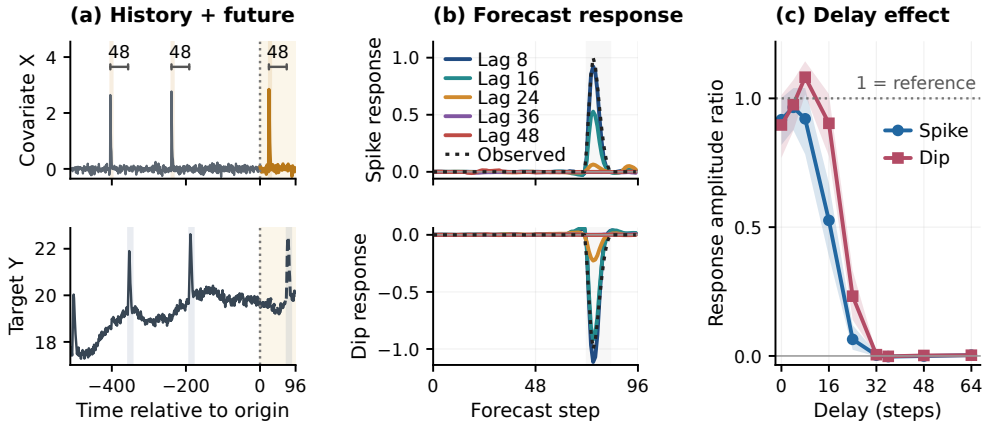


Figure 1: Delayed covariate–target relationships in Chronos-2. **(a)** Repeated covariate events and delayed target responses appear in the observed history, while a new event is known within the forecast horizon. **(b, c)** We increase only the signal–response delay while keeping the target history, future response, and background realization fixed.

instance, the channel-independent (**CI**), history-conditioned (**CD-history**), and future-conditioned (**CD-future**) views reveal how the same frozen backbone changes its marginal quantile forecast as additional information is introduced. In particular, the CD-history→CD-future transition shows whether known-future covariates induce a forecast change and where that change occurs. Such a response alone does not tell us whether the change is beneficial.

The observed history provides the missing supervision. When comparable covariate events have already occurred, their subsequent target outcomes are fully observed at the current forecast origin. We replay these completed episodes under CI, CD-history, and CD-future conditions and directly evaluate what happened as conditioning became richer. If CD-future moved the forecast and reduced historical quantile loss, the native future-conditioned response provides evidence that the backbone already knows how to use this type of event. If the forecast changed but became worse, the same comparison indicates that the conditional response should not be trusted. Most importantly, if historical targets exhibit a repeatable event–response pattern but CD-future remains nearly unchanged from the less conditioned views, the absence of a response itself identifies forecast structure that the backbone failed to express.

We propose *Input-Condition Integration* (ICI), a history-based test-time adaptation framework built around this diagnosis. ICI transfers the historical relationship between conditioning, forecast change, and realized forecast quality to a new known-future event. Historical evidence that richer conditioning was already effective indicates that the backbone's conditional response can be trusted, while the need for any additional correction is assessed separately through chronological replay. Evidence that conditioning was harmful instead favors correction that does not rely on that response. When a target response was historically required but the backbone expressed little corresponding change, ICI treats this discrepancy as a missing-response failure and relies on historical outcomes to supply the missing forecast structure.

For this last case, ICI learns an event-aligned residual quantile correction from completed historical episodes. Its temporal basis remains active beyond the covariate event, allowing delayed target responses to be recovered without providing their true lag. Thus, the three input conditions determine whether the native conditional behavior should be trusted, ignored, or regarded as insufficient, while the historical correction supplies additional structure only when warranted. ICI keeps the backbone frozen, considers adaptation only when sufficiently many comparable completed events are available, and uses chronological historical replay to decide among the native and corrected forecasts.

Our contributions are threefold:

- Building on recent evidence of covariate underutilization, we systematically characterize how the use of a fixed predictive relationship degrades with increasing signal–response delay, separating temporal separation from information availability.

- We introduce a frozen-backbone test-time adaptation framework that uses completed historical event–response episodes to determine whether richer input conditioning was beneficial, harmful, or insufficient for comparable events.
- We combine this input-condition diagnosis with an event-aligned residual quantile correction, retaining native future conditioning when it is historically reliable while recovering forecast structure that the backbone fails to express.

2 Related Work

Covariate-aware and instance-adaptive TSFMs. Recent methods extend pretrained TSFMs with covariate information without retraining the full backbone. CoRA and UniCA incorporate heterogeneous covariates through condition injection and representation fusion, respectively (Qin et al., 2025; Han et al., 2026). TFMAdapter is closer to our setting in adapting from the limited history of a single test instance: it trains a lightweight predictor that combines covariates with forecasts from a primarily univariate TSFM (Dange & Sarawagi, 2025). Recent diagnostics further show that access to covariates does not guarantee their effective use (Ghoroghchian et al., 2026; Berthelier et al., 2026). We instead start from a frozen TSFM with native covariate support and use completed historical events from the same test instance to evaluate and correct its conditional forecasting behavior.

Forecast combination and conditional responses. Forecast combination methods exploit diversity among candidate predictions. FFORMA learns feature-dependent weights over forecasting methods (Montero-Manso et al., 2020), while quantile regression averaging and flexible model aggregation extend this principle to probabilistic forecasts (Nowotarski & Weron, 2015; Fakoor et al., 2023). Our diversity instead comes from CI, CD-history, and CD-future forecasts of the same frozen backbone. Rather than directly averaging these forecasts, we use historical replays to determine whether the changes induced by richer conditioning were beneficial, harmful, or insufficient for comparable events.

Probabilistic forecast correction. Probabilistic post-processing uses observed forecast errors to adjust predictive distributions. Regression recalibration targets uncertainty calibration globally (Kuleshov et al., 2018), while approaches such as LADaR allow diagnostics and distributional reshaping to vary across the input space (Dey et al., 2025). We instead learn event-aligned quantile corrections from completed historical episodes within the current test instance. Input-condition responses indicate whether the frozen backbone already exploits the additional covariate information, while historical target outcomes determine when further correction is needed.

3 Input-Condition Integration

ICI adapts the probabilistic forecast of a frozen covariate-aware TSFM using completed events from the observed history. The key idea is to separate two questions: *what forecast structure needs correction*, and *whether the backbone's own response to additional conditioning provides useful evidence for that correction*. ICI first learns a generic event-aligned correction from historical forecast errors. It then uses CI, CD-history, and CD-future forecasts to determine whether richer conditioning was historically beneficial, harmful, or insufficient. Only positively supported conditioning responses are used as guidance; otherwise, ICI falls back to the same event-aligned correction. The input-condition differences are therefore diagnostic evidence rather than correction directions themselves.

3.1 Problem Setup and Historical Replays

We consider a frozen TSFM that natively supports target history, other-variable histories, and known future covariates. For a future covariate event, we search the observed history for comparable completed events based only on covariate-side characteristics such as transition type, direction, amplitude, and duration. A current event is considered *supported* when sufficiently many distinct historical analogs are available.

Each analog is replayed as a local forecasting problem from an earlier forecast origin, aligned so that the historical and current events occur at the same relative position in their forecast horizons. The model never receives the target continuation of the replay. Because that continuation is already part of the current observed history, however, it can be used afterward to evaluate the historical forecast. These replays provide instance-specific supervision without accessing the current future target or updating the backbone.

We query every replay under the same three input conditions used for the current forecast: the channel-independent view (**CI**), the history-conditioned view (**CD-history**), and the future-conditioned view (**CD-future**). The CD-future output is the native forecast that ICI either retains or corrects.

Detailed event definitions, analog matching, replay alignment, and the separation between retrieval history and backbone context are given in Appendix C.1.

3.2 Event-Aligned Base Correction

ICI first asks a simple question: *around this type of event, does the native CD-future forecast make a repeatable error?*

Using the historical replays, we learn a low-dimensional event-aligned quantile correction from the errors of the native CD-future forecasts. The correction is expressed relative to the covariate event rather than to an observed target peak, and its temporal basis remains active after the event has ended. It can therefore represent delayed target responses without being given the true signal–response lag.

This stage produces a base correction, denoted **E**. Importantly, E is learned before any input-condition guidance and is then kept fixed. Thus, the ability to correct a missing delayed response does not depend on the backbone already expressing that response through CI, CD-history, or CD-future. If the three views provide no useful additional evidence, ICI returns to exactly the same base correction.

The event-relative basis, quantile parameterization, fitting objective, regularization, and non-crossing constraints are specified in Appendix C.2.

3.3 Input-Condition Evidence

The three views are used to diagnose how the backbone uses additional information. The CI→CD-history change shows how the forecast reacts to historical cross-variable information, while the CD-history→CD-future change shows how it reacts when known-future covariates are additionally supplied.

A forecast change alone is not evidence that richer conditioning was useful. Historical replay makes this distinction observable. For each completed event, we compare the probabilistic errors of CI, CD-history, and CD-future over the same event-relative regions.

If richer conditioning changed the forecast and consistently reduced historical quantile loss, the corresponding response is treated as useful evidence that the backbone can exploit this type of information. If the forecast changed but became worse, that response is not trusted as guidance. Finally, if comparable historical events exhibit a repeatable target response but richer conditioning produces little corresponding forecast change, the absence itself indicates that the required structure is not being expressed by the backbone.

ICI evaluates the CI→CD-history and CD-history→CD-future transitions separately. Thus, a weak known-future response does not discard useful evidence that may remain in the history-conditioned response. Conversely, a historically harmful or inconsistent response simply receives no guidance weight; ICI does not reverse it or replace the native forecast with a less conditioned view.

The response descriptors, event-relative patching, neighbor retrieval, historical utility scores, and positive-evidence confidence rule are detailed in Appendix C.3.

3.4 GUIDED RESIDUAL CORRECTION

Input-condition evidence determines *which historical errors are relevant*, not the direction in which the current forecast should be moved. For the current event, ICI retrieves historical replays with similar conditional forecast behavior and uses conditioning responses as guidance only when their usefulness is supported by historical outcomes.

The additional correction is estimated from residual errors that remain after applying the fixed base correction E. In other words, the three views help determine which past residual evidence should be trusted, while the historical target residuals determine the actual sign and magnitude of the correction. The CI, CD-history, and CD-future differences are never directly added to the final prediction.

This yields three natural regimes. When richer conditioning is historically reliable and the current response is comparable, it can guide the selection of relevant post-E residual evidence. When the backbone responds but that response is historically harmful or unstable, the response is ignored and ICI falls back to the unchanged base correction E. When the backbone shows little response despite a repeatable historical target effect, the same base correction remains available to recover the missing structure.

The post-E residual estimator, response-conditioned retrieval, and guidance rule are provided in Appendix C.4.

3.5 CHRONOLOGICAL SELECTION

ICI separates the usefulness of conditioning from the usefulness of correction. A conditioning transition may have been beneficial historically while the native CD-future forecast already requires no further adjustment. Conversely, a missing or unreliable response may create room for historical correction. Conditioning utility therefore does not by itself determine whether the current forecast should be modified.

We evaluate candidate forecasts through chronological held-out historical replays. Earlier completed events are used to construct the correction, and later completed events are used to test whether it transfers. The base event-aligned correction is applied only when it improves over the native CD-future forecast by a prescribed margin. Conditional guidance is used only when it provides an additional improvement over the same fixed base correction. Otherwise, ICI falls back to the simpler candidate, ultimately retaining the native forecast when historical evidence for adaptation is insufficient.

Overall, ICI modifies the native forecast only when completed historical events support a transferable correction, and uses historically validated input-condition behavior only when it provides additional value beyond that base correction.

The chronological folds, validation margins, guidance-strength selection, and native–base–guided fallback hierarchy are specified in Appendix C.5.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate ICI with Chronos-2, TimesFM-3, and TiRex-2 as frozen covariate-aware backbones. For synthetic experiments, historical events are retrieved from a 2,048-step observed history, while all current and replay forecasts use the same 512-step backbone context and a 96-step horizon. We evaluate probabilistic forecasts using normalized quantile score (NQS; lower is better). Real-world experiments follow the original fev-bench horizons and are evaluated using scaled quantile loss (SQL).

All adaptation decisions use only completed historical events. Hyperparameters, event-matching rules, and selection thresholds are fixed before final evaluation. We report uncertainty using paired cluster bootstrap intervals over independent generation seeds or real-world evaluation groups. Additional implementation and dataset details are provided in Appendix A and Appendix B.

Table 1: **ICI on controlled repeated-event forecasting.** Relative NQS improvement (%) over the native CD-future forecast. Higher is better.

Regime	Chronos-2	TimesFM-3	TiRex-2
Stable	+18.2	+15.4	+17.6
Varying dose	+15.0	+14.1	+16.2
Seasonal background	+15.5	+23.0	+22.7
Stochastic response	+15.1	+11.6	+14.4
Null	+0.2	+0.1	+0.3
Relationship shift	-8.4	-9.1	-10.0

Table 2: **Real-world forecasting.** Relative SQL improvement (%) over the native CD-future forecast on representative official benchmark tasks.

Dataset	Chronos-2	TimesFM-3	TiRex-2
ENTSO-e 30min	+2.64	+2.11	+2.38
ENTSO-e 1h	+1.96	+1.53	+1.72
EPF-DE	+1.48	+1.76	+1.22
GFC17	+1.71	+1.83	+1.46
Rossmann 1D	+2.48	+2.91	+2.63
Rohlik Orders 1D	+1.86	+2.17	+1.94
Average	+2.02	**+2.05**	+1.89

4.2 Recovering Delayed Event Responses

Table 1 shows that ICI consistently improves probabilistic forecasting when the historical event–response relationship remains transferable. The gains remain substantial under variation in event magnitude, background seasonality, and stochastic target responses.

In contrast, ICI changes little under the null condition, where the covariate event carries no systematic target response. Performance deteriorates under relationship shift, where the historical response is no longer representative of the current event. This contrast isolates the assumption underlying ICI: historical events are useful only when their forecast corrections remain locally transferable.

In Figure 2, ICI provides little benefit at short delays, where the native CD-future forecast remains responsive, but yields increasingly large improvements as the signal–response gap grows. The true delay is never provided to the method.

4.3 Real-World Forecasting

ICI is intentionally selective on real-world data (Table 2). Its applicability depends on whether comparable completed events exist within the observed history. On tasks with sufficient historical support, adaptation provides modest additional gains over already strong native forecasts, whereas weakly supported instances tend to retain the native CD-future prediction.

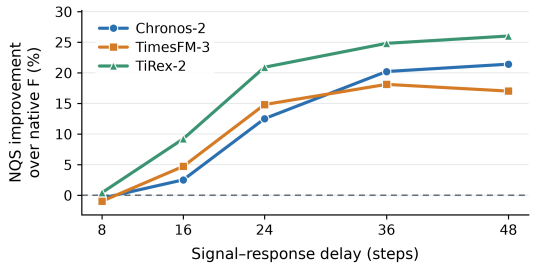


Figure 2: Historical correction becomes more valuable as the native delayed response weakens.

5 Analysis

5.1 Where Does the Improvement Come From?

Most of the improvement comes from the event-aligned base correction, confirming that repeated historical forecast errors provide the primary adaptation signal. However, validated input-condition evidence yields a further improvement over E (Table 3).

Response similarity alone provides only a small gain. The improvement becomes larger when historical target outcomes are used to verify whether richer conditioning was actually useful, supporting the role of the three views as diagnostic evidence rather than direct correction directions.

5.2 When Is Correction Needed?

Figure 3 illustrates the complementary roles of native conditioning and historical correction. At short delays, the native CD-future forecast is frequently retained. As the delay increases, event-aligned correction becomes more important. Guided correction is selected most often

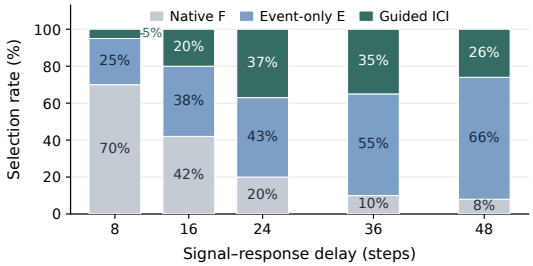


Figure 3: Selection vs lag.

in the intermediate regime, where the backbone still expresses partial conditioning responses that historical replay can validate and exploit. At the longest delays, the base correction dominates as the input-condition responses themselves become weak.

5.3 Failure Modes and Limitations

The null condition confirms that repeated covariate events alone are not sufficient to trigger large corrections. By contrast, relationship shift exposes the main limitation of history-based adaptation. When the event–response relationship changes after the observed history, the correction learned from historical episodes no longer transfers and can substantially degrade forecasting performance. ICI therefore relies on local stability of the event–response relationship, an assumption shared by any method that adapts from the current instance's past (Table 4).

Table 3: **Contribution of conditioning evidence.** Average NQS improvement (%) over the native forecast across transferable synthetic regimes. All variants use the same fixed event-aligned base correction.

Method	Chronos-2	TimesFM-3	TiRex-2	Avg.
Event-only (E)	15.1	13.2	17.5	15.3
E + X/F	15.3	13.0	17.6	15.3
E + response similarity	15.4	13.5	17.8	15.6
E + total CI→F utility	15.7	13.8	18.0	15.8
ICI (split H/F utility)	**16.3**	**14.5**	**18.6**	**16.5**

Table 4: **Negative controls and transfer failures.** Relative NQS change over the native forecast.

Regime	Chronos-2	TimesFM-3	TiRex-2
Null relationship	+0.2	+0.1	+0.3
Relationship shift	-8.4	-9.1	-10.0

ICI substantially recovers delayed event responses when the historical relationship remains transferable. Most of this gain comes from event-aligned historical correction, while validated input-condition evidence provides a smaller but consistent additional improvement. Adaptation is selective: native forecasts are largely retained when richer conditioning already works or no transferable correction is supported. Relationship shift exposes the method's central limitation—historical correction cannot transfer when the underlying event–response relationship changes.

6 Conclusion

We studied a failure mode of covariate-aware TSFMs in which predictive covariate information remains available but delayed target responses are weakly expressed. To address this, we introduced *Input-Condition Integration* (ICI), a frozen-backbone test-time adaptation framework that replays completed historical events to assess whether richer conditioning was useful and to learn event-aligned forecast corrections. ICI uses CI, CD-history, and CD-future forecasts as diagnostic evidence rather than direct correction directions, while historical target residuals determine the actual adjustment. This allows the method to retain reliable native forecasts, exploit useful conditioning behavior when present, and recover historically supported structure when the backbone fails to express it. Overall, our results highlight the observed history as a practical source of instance-specific evidence for improving covariate-aware forecasting without backbone fine-tuning or access to the current future target.

AI USE STATEMENT

We used generative AI tools to assist with synthetic data generation and minor editorial revisions during manuscript preparation. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects or sensitive personal data. We use public benchmark datasets and controlled synthetic data for methodological evaluation, and we are not aware of additional ethical concerns specific to this study.

REPRODUCIBILITY STATEMENT

We provide detailed dataset configurations, implementation settings, method specifications, and additional experimental results in the appendix.

REFERENCES

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

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

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *Advances in Neural Information Processing Systems*, 38:57529–57580, 2026.

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. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026. URL https://openreview.net/forum?id=H8p5Yc0c3W.

Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*, 2024.

Afrin Dange and Sunita Sarawagi. Tfmadapter: Lightweight instance-level adaptation of foundation models for forecasting with covariates. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 498–507, 2025.

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.

Biprateep Dey, David Zhao, Brett H Andrews, Jeffrey A Newman, Rafael Izbicki, and Ann B Lee. Towards instance-wise calibration: local amortized diagnostics and reshaping of conditional densities (ladar). *Machine Learning: Science and Technology*, 6(4):045058, 2025.

Rasool Fakoor, Taesup Kim, Jonas Mueller, Alexander J Smola, and Ryan J Tibshirani. Flexible model aggregation for quantile regression. *Journal of Machine Learning Research*, 24(162):1–45, 2023.

Nafiseh Ghoroghchian, Haipeng Zhang, Shuyi Han, Alex Labach, and George Stein. Foundations without fundamentals: Zero-shot blind spots in time series FMs. In *2nd ICML Workshop on Foundation Models for Structured Data*, 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 *Forty-first International Conference on Machine Learning*, 2024.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, et al. Unica: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 6403–6443, 2026.

Ayush Jain and Rajat Sen. Timesfm-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, 2026.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, et al. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Doyen Sahoo, Caiming Xiong, Silvio Savarese, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

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. In *Forty-second International Conference on Machine Learning*, 2025b.

Lucas Meyer, Claudio Sole, Huikan Xiang, Nicolas Li, Lucas Franceschino, Arnau Quera-Bofarull, Maarten P. Scholl, Joachim Fainberg, and Geoffrey Négiar. t0: A time-series foundation model for forecasting with context, 2026. URL `https://www.theforecastingcompany.com/papers/files/t0-technical-report.pdf`.

Pablo Montero-Manso, George Athanasopoulos, Rob J Hyndman, and Thiyanga S Talagala. Fforma: Feature-based forecast model averaging. *International Journal of Forecasting*, 36(1):86–92, 2020.

Jakub Nowotarski and Rafał Weron. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30(3):791–803, 2015.

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. *arXiv preprint arXiv:2607.01204*, 2026.

Guo Qin, Zhi Chen, Yong Liu, Zhiyuan Shi, Haixuan Liu, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Cora: Covariate-aware adaptation of time series foundation models. *arXiv preprint arXiv:2510.12681*, 2025.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

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

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. 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. 53140–53164. PMLR, 21–27 Jul 2024.

A Dataset Configuration

A.1 Real-world Benchmarks

We evaluate ICI on real-world forecasting tasks from fev-bench (Shchur et al., 2025) that provide dynamic covariates over the forecast horizon. These tasks match the CD-future setting considered in our method: future covariates are available at prediction time, whereas future target values remain unobserved.

Our evaluation covers 15 attempted tasks. The *energy* domain includes ENTSO-e Load at 15-minute, 30-minute, and hourly resolutions; EPF-BE, EPF-DE, EPF-FR, EPF-NP, and EPF-PJM; GFC12, GFC14, and GFC17; and Solar with Weather at 15-minute and hourly resolutions. The *retail* domain includes Rossmann and Rohlik Orders at daily resolution. Evaluation windows that do not contain the complete history and covariate information required by the forecasting task are excluded before adaptation.

For each dataset, we report two temporal cohorts separately. The *official* cohort uses evaluation windows from the fev-bench evaluation range, while the *extension* cohort contains earlier historical forecast origins used to examine additional temporal coverage. The two cohorts are not pooled when reporting dataset-level results. Because the applicability of ICI depends on finding completed historical analogs, we additionally report the number of evaluated instances that satisfy event eligibility and that contain sufficient chronological replay evidence.

ICI does not alter the native covariate interface of the benchmark. All known-future covariates supplied by the forecasting task remain available to the CD-future backbone forecast. For event retrieval, continuous covariates are searched for completed excursions and binary or categorical covariates for completed state transitions; calendar variables are not used to trigger events. No covariate channel is removed from the native forecast. Detailed eligibility and retrieval rules are given in Appendix C.1.

We follow the benchmark-provided future information as given. In particular, weather variables marked as known future are treated as available covariates rather than reconstructed from archived weather forecasts, and we do not independently verify the issuance vintage of the EPF covariates. Static metadata are not supplied to the evaluated backbones. Real-world performance is evaluated using seasonal scaled quantile loss (SQL), while coverage and adaptation rates are reported separately to distinguish forecasting performance from the practical applicability of ICI.

A.2 Synthetic Experiments

We construct controlled repeated-event time series to evaluate ICI when the relationship between a known-future covariate event and its target response is explicitly controlled. Each sample contains a 2,048-step observed history and a 96-step forecast horizon. Historical events provide repeated completed examples of the covariate–target relationship, and a new covariate event is placed within the known-future horizon. All backbone queries use the same 512-step local context, while the full observed history remains available for historical-event retrieval.

The final synthetic evaluation contains six regimes: *stable*, *varying dose*, *seasonal background*, *stochastic response*, *null*, and *relationship shift*. Each regime contains 16 new independent root seeds. For every seed, we evaluate positive and negative event directions (spike and dip) and signal–response delays of

$$\{8, 16, 24, 36, 48\}$$

steps, yielding 160 forecast instances per regime and backbone. Conditions derived from the same root seed share the underlying background realization and are treated as paired observations in the statistical analysis.

The first four regimes preserve a transferable event–response relationship while varying different aspects of the forecasting problem. The *stable* regime provides the basic repeated relationship; *varying dose* introduces variation in event or response magnitude; *seasonal background* places the same relationship within stronger background temporal structure; and *stochastic response* introduces additional variability in the target response. The *null* regime removes the systematic target effect of the covariate event and serves as a negative control. The *relationship-shift* regime changes the future

event–response relationship relative to the completed historical episodes, testing the failure of the local-transferability assumption.

Across delay conditions, the generator preserves the predictive event–response mechanism while changing its temporal separation. The true response delay and synthetic response annotations are never provided to ICI. They are used only for paired structural diagnostics. All proper forecasting scores are computed against the realized future target using the backbone's marginal quantile forecasts.

Synthetic performance is reported using normalized quantile score (NQS), defined from mean pinball loss and the pre-origin target scale. Because NQS and the seasonal SQL used for real-world benchmarks employ different normalizations, their absolute values are not compared across the two experiment families. Uncertainty estimates resample independent root seeds while preserving the paired event-direction and delay conditions within each seed.

B Implementation and Evaluation Details

For the controlled synthetic experiments, each instance contains a 2,048-step observed history and a 96-step forecast horizon. Real-world experiments follow the forecast horizons defined by the corresponding fev-bench tasks. For both settings, the forecasting backbone receives at most the most recent $L = 512$ observations, while historical-event retrieval may use up to 2,048 observed steps. All current and historical replay queries use the same backbone context length. Accordingly, CI, CD-history, CD-future, and the native CD-future baseline are recomputed under the same context constraint.

All TSFM parameters remain frozen. For each supported current event, historical analogs are first identified using only covariate-side event information. The base event-aligned correction E is then fitted from historical CD-future forecast errors and kept fixed throughout the subsequent guidance stage. Input-condition responses are used only to characterize comparable historical forecast behavior and to evaluate whether richer conditioning was historically useful; they are never directly added to the corrected forecast.

Event-relative correction. The base correction uses a fixed piecewise-linear temporal basis defined relative to the covariate-event onset. The basis remains active throughout the post-event forecast region, rather than terminating with the covariate event, allowing delayed target responses to be represented without access to the true delay. The quantile basis is $\boldsymbol{\psi}(\alpha) = (1, 2\alpha - 1)^\top$. Base parameters are fitted using normalized pinball loss with ridge regularization and linear non-crossing constraints.

Input-condition guidance. We divide the forecast horizon into fixed 8-step event-relative patches. For each patch, historical retrieval uses summaries of the native CD-future forecast together with the CI→CD-history and CD-history→CD-future changes. Patch-level retrieval retains at most eight distinct historical neighbors, requires at least three, and discards candidates with standardized RMS descriptor distance greater than 2.

The usefulness of the two conditioning transitions is evaluated separately using their historical target outcomes. For a transition to activate guidance, its current normalized response magnitude must be at least 0.05, its distance-weighted historical utility must exceed 0.005, and at least two thirds of the retrieved historical events must have positive utility. ICI uses the larger confidence obtained from the two conditioning transitions. Negative or unstable historical utility produces no guidance; the corresponding conditioning response is neither reversed nor subtracted from the native forecast.

Residual guidance and selection. When guidance is active, the additional adjustment is estimated from post-E historical target residuals using the symmetric marginal quantile grid $\{0.1, 0.2, \ldots, 0.9\}$. The guided component is restricted to a quantile-common location shift, whereas the base correction remains quantile dependent.

Candidate forecasts are selected through chronological held-out historical replays. The base correction is accepted only when it improves over the native CD-future forecast by at least 0.005 in normalized quantile loss. Conditional on accepting E, we select $g \in \{0, 0.25, 0.5, 1\}$ and permit

Table 5: Key implementation settings used by ICI.

Setting	Value
Maximum retrieval history	2,048
Backbone context L	512
Synthetic forecast horizon h	96
Event-relative patch length	8
Maximum real-world event analogs	16
Maximum patch-level guidance neighbors	8
Minimum patch-level guidance neighbors	3
Maximum standardized RMS distance	2
Response-magnitude threshold	0.05
Historical utility threshold	0.005
Minimum positive-utility fraction	$2/3$
Guidance quantiles	$\{0.1, \ldots, 0.9\}$
Guidance strengths g	$\{0, 0.25, 0.5, 1\}$
Validation improvement margin	0.005

positive guidance only when it improves over the same fixed E by at least 0.005. Ties favor the smaller guidance strength.

All thresholds and selection hyperparameters are fixed before final evaluation. Table 5 summarizes the key numerical settings. Further definitions of event retrieval, base fitting, conditioning utility, residual estimation, and chronological validation are given in Appendix C.

C Additional Method Details

This appendix provides the implementation details for Section 3. ICI is fitted independently for each supported test instance using only completed events contained in its observed history. The backbone remains frozen throughout. We use the same decomposition as in the main text: historical replay, event-aligned base correction, input-condition evidence, guided residual correction, and chronological selection.

C.1 Historical Event Retrieval and Replay

Retrieval history and forecasting context. We distinguish the history available for event retrieval from the context passed to the backbone. In the synthetic experiments, the full 2,048-step observed history is searched for historical analogs, whereas all current and historical backbone queries use the same context length $L = 512$. The forecast horizon is $h = 96$. Accordingly, all native CI, CD-history, and CD-future forecasts are recomputed using the same 512-step context.

Let T denote the current forecast origin and let $s^\star$ be the onset of the known-future event. We write

$$\delta^\star = s^\star - T, \qquad 1 \le \delta^\star \le h, \tag{1}$$

for its position within the forecast horizon. For a historical analog e_j with onset s_j, its replay origin is

$$\tau_j = s_j - \delta^\star, \tag{2}$$

so that the current and historical events occur at the same relative forecast position.

The replay context and target continuation are

$$\mathcal{H}_j = \{\tau_j - L + 1, \ldots, \tau_j\}, \qquad \mathcal{F}_j = \{\tau_j + 1, \ldots, \tau_j + h\}. \tag{3}$$

A replay is feasible only when

$$\tau_j - L + 1 \ge 1, \qquad \tau_j + h \le T, \tag{4}$$

and the corresponding historical covariate event has completed by the current forecast origin.

The continuation

$$\mathbf{y}_j = y_{\mathcal{F}_j} \tag{5}$$

is withheld from all backbone queries. It is used only after prediction for fitting or validation.

Event matching. Eligibility is based only on covariate-side information. For a binary covariate, we match events with the same state transition and approximately similar duration. For a continuous covariate, we require the same movement direction and approximately similar scale-normalized amplitude and duration.

We write

$$e_j \simeq_X e^\star \tag{6}$$

when historical event e_j satisfies these fixed matching rules. The replay-feasible analog set is

$$\mathcal{A}(e^\star) = \{e_j : e_j \simeq_X e^\star,\ e_j \text{ completed and replay-feasible}\}. \tag{7}$$

Each physical event is counted once. The current event is supported when

$$|\mathcal{A}(e^\star)| \geq k. \tag{8}$$

Historical target behavior is not used to determine event support. Support therefore indicates only that comparable historical episodes are available, not that the event is predictive of the target or that correction is beneficial.

Three replay views. For each replay, we issue the same three queries as at the current forecast origin:

$$\mathbf{Q}_j^{CI}, \qquad \mathbf{Q}_j^{H}, \qquad \mathbf{Q}_j^{F}, \tag{9}$$

corresponding to CI, CD-history, and CD-future, respectively. The CD-future forecast $\mathbf{Q}_j^F$ is treated as the native historical prediction. All three views use the same forecast horizon, context length, and marginal quantile grid.

C.2 Event-Aligned Base Correction

Event-relative correction surface. The base correction E is fitted using only historical errors of the native CD-future forecast. No CI/CD-history/CD-future response information is used at this stage.

For forecast position r, define the event-relative coordinate

$$u_r = r - \delta^\star. \tag{10}$$

Because all replay origins are aligned to the current event, the same u_r refers to the same position relative to the covariate event across all historical replays.

We parameterize the normalized correction by

$$C_E(u, \alpha) = \boldsymbol{\phi}(u)^\top \mathbf{B}_E \boldsymbol{\psi}(\alpha), \tag{11}$$

where $\boldsymbol{\phi}$ is a fixed piecewise-linear basis over event-relative time. In our implementation, the temporal basis uses fixed 4-step resolution and remains active throughout the post-event forecast region. It therefore allows delayed responses to be learned without providing the true target lag.

For the quantile dimension, we use

$$\boldsymbol{\psi}(\alpha) = \begin{bmatrix} 1 \\ 2\alpha - 1 \end{bmatrix}. \tag{12}$$

Let $\sigma_j > 0$ be a target scale computed using only observations available before replay origin τ_j. The same scale rule is used for the current forecast. The base-corrected replay forecast is

$$Q_{E,j}(r, \alpha) = Q_j^F(r, \alpha) + \sigma_j C_E(u_r, \alpha). \tag{13}$$

Quantile-loss fitting. For quantile level α, we use pinball loss

$$\rho_\alpha(v) = v\,(\alpha - \mathbb{I}[v < 0]). \tag{14}$$

For a fitting set $\mathcal{D}$, the base parameters are estimated from

$$\widehat{\mathbf{B}}_E = \arg\min_{\mathbf{B}_E} \quad \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \frac{2}{hM} \sum_{r=1}^{h} \sum_{m=1}^{M} \rho_{\alpha_m} \left(\frac{y_{j,r} - Q_{E,j}(r, \alpha_m)}{\sigma_j} \right) + \lambda_E \|\mathbf{B}_E\|_F^2. \tag{15}$$

Thus, E is fitted against the corrected CD-future forecast itself, rather than by averaging realized residuals.

Non-crossing constraints. We enforce marginal quantile ordering directly during fitting:

$$Q_{E,j}(r, \alpha_m) \leq Q_{E,j}(r, \alpha_{m+1}) \tag{16}$$

for every fitting replay j, horizon position r, and adjacent pair of quantile levels. Because Eq. equation 13 is affine in $\mathbf{B}_E$, the objective is convex and the non-crossing conditions are linear constraints.

Numerical crossing violations are audited after every fit. For held-out or current forecasts, deterministic row-wise rearrangement is applied only when necessary.

Fixed base correction. The base correction is fitted before any Three-Views guidance and is subsequently held fixed. In particular, adding guidance does not refit $\mathbf{B}_E$. Therefore, zero guidance returns exactly the same base forecast Q_E.

This property separates the contribution of event-aligned historical errors from the additional contribution of input-condition evidence.

C.3 Input-Condition Evidence

Input-condition responses. For historical replay j, define

$$\mathbf{\Delta}_j^H = \mathbf{Q}_j^H - \mathbf{Q}_j^{CI}, \qquad \mathbf{\Delta}_j^F = \mathbf{Q}_j^F - \mathbf{Q}_j^H. \tag{17}$$

The corresponding current responses are defined identically.

These differences describe how the frozen backbone reacts as conditioning becomes richer. They are never directly added to the corrected forecast.

Event-relative patches and response descriptors. We divide the $h = 96$ horizon into fixed 8-step patches. Because replay events are aligned to the current event, identical patch indices correspond to identical event-relative regions.

For each patch, the main Three-Views descriptor contains:

- the standardized local level of the CD-future forecast;
- the standardized predictive interval width of the CD-future forecast;
- CI→CD-history median changes at three fixed locations within the patch;
- CD-history→CD-future median changes at the same locations; and
- the corresponding changes in predictive interval width.

Feature scales are estimated from historical descriptors and used to standardize the retrieval distance. The historical target continuation is not part of the descriptor.

For each current patch, we retain at most eight distinct historical events. At least three neighbors are required, and candidates with standardized RMS descriptor distance greater than 2 are rejected. The resulting neighbor set is denoted by

$$\mathcal{N}_\star(P) \subseteq \mathcal{A}(e^\star). \tag{18}$$

Historical utility of richer conditioning. After neighbors have been selected without using their target outcomes, we evaluate whether each conditioning transition was historically beneficial.

Let $L_P(\mathbf{Q}, \mathbf{y})$ denote twice the mean pinball loss over event-relative patch P. For historical event j, define

$$U_j^H(P) = \frac{L_P(\mathbf{Q}_j^{CI}, \mathbf{y}_j) - L_P(\mathbf{Q}_j^H, \mathbf{y}_j)}{\sigma_j}, \tag{19}$$

$$U_j^F(P) = \frac{L_P(\mathbf{Q}_j^H, \mathbf{y}_j) - L_P(\mathbf{Q}_j^F, \mathbf{y}_j)}{\sigma_j}. \tag{20}$$

Positive utility means that the richer conditioning reduced historical quantile loss. Negative utility means that it worsened the historical forecast.

Importantly, negative-utility events are not removed from the retrieved set. All matched events contribute to the utility statistics and, once guidance is activated, to residual estimation. We therefore do not construct correction directions from successful historical cases alone.

Positive-evidence confidence. ICI v2 uses a positive/no-guidance rule. For conditioning transition $q \in \{H, F\}$, guidance is activated only when all of the following hold:

1. The current response is non-negligible:

$$\frac{\|\mathbf{\Delta}_\star^q(P)\|_{\mathrm{RMS}}}{\sigma_\star} \geq 0.05. \tag{21}$$

2. The distance-weighted mean historical utility satisfies

$$\overline{U}_q(P) > 0.005. \tag{22}$$

3. At least two thirds of the distinct retrieved events have positive utility for that transition.

When all three conditions hold, we define

$$c_q(P) = \mathrm{clip}\left(\frac{\sum_j w_j U_j^q(P)}{\sum_j w_j |U_j^q(P)|}, 0, 1\right), \tag{23}$$

where the nonnegative neighbor weights decrease with descriptor distance and are normalized within the retrieved set. Otherwise,

$$c_q(P) = 0. \tag{24}$$

The main ICI variant combines the two transitions through

$$c(P) = \max\{c_H(P), c_F(P)\}. \tag{25}$$

Thus, either CI→CD-history or CD-history→CD-future can provide useful evidence. A weak CD-history→CD-future response does not discard informative historical-conditioning evidence.

Conversely, negative or unstable utility never creates a negative confidence. ICI v2 does not reverse, subtract, or otherwise explicitly undo a historically harmful input-condition response. Such a response simply receives zero guidance weight.

C.4 Guided Residual Correction

Post-base residuals. Guidance is applied only after the base correction E has been fitted and fixed. For replay j, horizon position r, and quantile level α_m, define the normalized post-E residual

$$R_{j,r,m} = \frac{y_{j,r} - Q_{E,j}(r, \alpha_m)}{\sigma_j}. \tag{26}$$

Three-Views evidence determines which historical residuals are relevant and how strongly they should be trusted. It does not determine the sign or magnitude of the correction itself.

Residual location estimate. For current horizon position r in patch P_r, let $\mathcal{N}_\star(P_r)$ denote the retrieved historical events. We estimate a common normalized location shift by

$$\eta_\star(r) = \arg\min_\eta \sum_{j \in \mathcal{N}_\star(P_r)} w_{j\star} \frac{1}{|\mathcal{T}|} \sum_{\alpha \in \mathcal{T}} \rho_\alpha \left(R_{j,r,\alpha} - \eta\right), \tag{27}$$

where $\mathcal{T}$ is the symmetric quantile grid used for guidance.

In our implementation, $\mathcal{T} = \{0.1, \ldots, 0.9\}$. For this symmetric grid, Eq. equation 27 is implemented as a weighted median of the normalized quantile residuals rather than an arithmetic mean.

Guided forecast. The guided forecast is

$$Q_G(r, \alpha_m) = Q_E(r, \alpha_m) + \sigma_\star g\, c(P_r)\, \eta_\star(r), \tag{28}$$

where g is selected through chronological validation.

The response differences $\mathbf{\Delta}^H$ and $\mathbf{\Delta}^F$ do not appear additively in Eq. equation 28. They affect:

1. which historical events are considered similar;
2. whether the corresponding conditioning behavior was historically useful; and
3. the confidence with which the resulting residual estimate is applied.

The actual correction direction and magnitude are determined by the historical target residuals after E.

Quantile-common guidance. The current guided component applies a common location shift across quantiles at each horizon position. It therefore preserves the quantile spacing produced by the base forecast Q_E. This restriction keeps the second-stage correction deliberately simple. Quantile-dependent distributional correction is already available through the base term C_E.

No-guidance fallback. If neither input-condition transition passes the positive-evidence criteria, then

$$c(P_r) = 0, \tag{29}$$

and Eq. equation 28 reduces exactly to

$$Q_G(r, \alpha_m) = Q_E(r, \alpha_m). \tag{30}$$

Thus, failure of the backbone to express a useful conditional response does not remove the base event-aligned correction.

C.5 Chronological Selection

Leakage-free historical folds. Conditioning utility and correction utility are evaluated separately. For validation event e_j, only earlier events whose covariate episodes and complete replay target continuations would already have been observed at replay origin τ_j may be used.

Let b_i denote the completion time of historical event e_i. The admissible fitting set is

$$\mathcal{D}_j = \{e_i : \max\{b_i, \tau_i + h\} \le \tau_j\}\,. \tag{31}$$

The validation target of e_j is excluded from:

- fitting the base correction E;
- response-descriptor scaling;
- neighbor construction;
- utility estimation used for that validation prediction; and
- post-E residual estimation.

The same physical event is never treated as multiple independent pieces of evidence.

Base-correction selection. We first compare the native CD-future forecast N with the base event-aligned correction E on chronological held-out events.

E is eligible for the current forecast only when its average normalized validation loss improves on N by at least

$$\epsilon = 0.005. \tag{32}$$

If this condition is not satisfied, the final forecast remains the native CD-future prediction and the guidance stage is not used.

Table 6: Guidance variants. All methods use the same event alignment, historical target labels, base correction E, chronological validation folds, correction margin, and guidance-strength grid.

Variant	Base	Retrieval information	Response utility
E	E	none	none
XF	E	known X, native F	none
R	E	Δ^H, Δ^F, F	none
T	E	CI→F, F	CI→F
HF / ICI	E	Δ^H, Δ^F, F	CI→H, H→F

Guidance-strength selection. Conditional on accepting E, we evaluate

$$g \in \{0,\ 0.25,\ 0.5,\ 1\} \tag{33}$$

using the same chronological validation folds.

A positive g is selected only when its mean normalized validation loss improves on E by at least the same margin ϵ. If multiple strengths satisfy the criterion with indistinguishable validation performance, the smaller g is preferred. If no positive value improves sufficiently, we set $g = 0$ and retain the unchanged base correction.

Final hierarchy. The adaptation hierarchy is therefore

$$\mathrm{N} \longrightarrow \mathrm{E} \longrightarrow \mathrm{G}. \tag{34}$$

The three stages answer different questions:

1. **N vs. E:** do historical event-aligned errors support correction at all?
2. **Input-condition utility:** did richer conditioning historically improve comparable forecasts?
3. **E vs. G:** does using that conditioning evidence to select residual information improve the correction beyond E?

Useful input conditioning therefore does not automatically imply that the current forecast should be modified. Likewise, failure of conditioning to help does not automatically imply that E should be applied. Each additional correction stage must demonstrate incremental historical forecasting value.

Current forecast. After model selection, the accepted component is refitted using all eligible completed analogs available at the current forecast origin.

The final prediction is

$$\mathbf{Q}_\star^{ICI} = \begin{cases} \mathbf{Q}_\star^F, & \text{if support or E validation fails,} \\ \mathbf{Q}_{E,\star}, & \text{if E is accepted and } g = 0, \\ \mathbf{Q}_{G,\star}, & \text{if E and positive guidance are both accepted.} \end{cases} \tag{35}$$

At no stage is the current future target accessed.

C.6 Ablation Variants

We use the same fitted base correction E across all guidance variants so that differences arise only from the information used to retrieve and validate residual evidence.

The comparisons isolate different sources of conditional information.

E vs. XF. This comparison tests whether ordinary similarity in the given future covariate and native CD-future forecast is sufficient for conditional residual retrieval.

R vs. HF. Both variants use the split input-condition responses for retrieval. R does not verify whether these responses historically improved the forecast, whereas HF applies the positive-utility rule. This comparison therefore isolates the role of target-verified conditioning usefulness.

T vs. HF. T uses the total CI→CD-future change, whereas HF separates it into CI→CD-history and CD-history→CD-future transitions. This comparison evaluates whether the intermediate CD-history view provides additional information beyond the total conditioning response.

These ablations share historical labels and the main adaptation protocol but are not claimed to be perfectly capacity-matched.

C.7 Implementation Summary

For a supported current event, ICI proceeds as follows:

1. Retrieve covariate-matched completed historical events from the observed history and align their replay origins to the current event.
2. Compute CI, CD-history, and CD-future forecasts for each historical replay using the same backbone context as the current forecast.
3. Fit the event-aligned base correction E from historical CD-future forecast errors and freeze it.
4. Within each event-relative patch, compare the current and historical input-condition responses and retrieve historical events with similar conditional forecast behavior.
5. Use historical target outcomes to determine whether the CI→CD-history or CD-history→CD-future response was consistently beneficial. Only positive, sufficiently consistent evidence activates guidance.
6. Estimate the additional correction from the post-E target residuals of the retrieved events. The input-condition differences themselves are never added as correction directions.
7. Use chronological held-out replays to choose among the native forecast, the fixed event-aligned correction, and the guided correction.

The method therefore uses three distinct sources of information: covariate events determine which historical episodes are comparable, historical target errors determine what forecast structure needs correction, and historically validated input-condition responses determine when residual evidence should be used to refine that correction.

D Underutilization of Delayed Covariate–Target Relationships in Covariate-Aware TSFMs

Figure 1 in the main text illustrates the lag-dependent degradation of covariate utilization with Chronos-2. Here we repeat the same controlled diagnostic for TimesFM-3 and TiRex-2. The purpose is not to compare absolute forecasting accuracy across models, but to test whether the same predictive covariate–target relationship remains reflected in their forecasts as its temporal separation increases. This diagnostic uses a separate set of 32 history seeds per event type and is distinct from the 16-root-seed synthetic evaluation reported in Section 4.

Controlled event–response construction. Each synthetic instance contains repeated covariate events and their target responses in a 2,048-step observed history, followed by a new known-future event within a 96-step forecast horizon. Across delay conditions, we keep the target history, future response timing and amplitude, and background realization fixed. Only the temporal separation between the covariate event and the target response is changed. The covariate event remains within the known-future segment for all tested delays. Thus, increasing delay does not remove the predictive signal from the model input.

To isolate the forecast component induced by the event, we use paired event-on and event-off worlds with otherwise matched histories and future backgrounds. For the marginal-median forecast, response gain measures the projection of the paired forecast difference onto the realized paired target

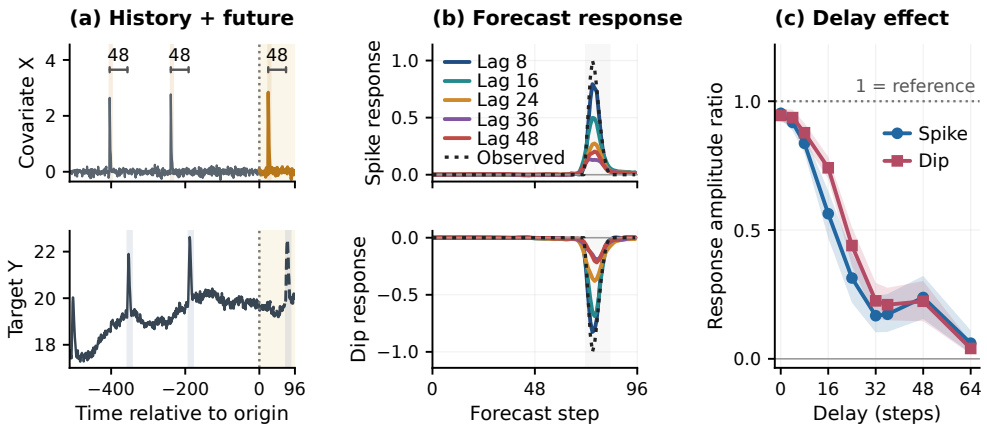


Figure 4: **Delayed covariate–target relationships in TimesFM-3.** The experimental construction matches Figure 1: only the signal–response delay changes across conditions. TimesFM-3 expresses a clear event-induced response at short delays, but the response becomes substantially attenuated as the temporal separation increases.

response. The gain is normalized such that a value of one corresponds to the magnitude of the realized target response. A gain near zero indicates little response in the correct target direction, while values above one indicate possible over-response. This quantity is used only as a structural diagnostic and is not a probabilistic forecasting metric.

As a reference, we also fit a distributed-lag ridge regressor using only the observed target and covariate histories. The regressor is not given the true signal–response delay. Its purpose is to verify that the predictive relationship remains recoverable from the available history even when the TSFM response weakens.

All curves aggregate 32 independent history seeds per event type, with seeds paired across delay conditions. Shaded regions denote pointwise 95% bootstrap intervals.

TimesFM-3. Figure 4 shows that TimesFM-3 responds to the covariate event when the target response follows closely, but this event-induced component decreases sharply as the signal–response gap increases. The reduction occurs even though the underlying predictive relationship and the available covariate information are unchanged. The history-fitted distributed-lag reference remains responsive at larger delays, indicating that the weakening TSFM response cannot be explained simply by the disappearance of predictive information from the input.

TiRex-2. TiRex-2 exhibits a related but distinct failure pattern (Figure 5). Its event-induced response is already weaker than the other tested backbones in the short-delay regime and remains weak as the delay increases. Thus, the three backbones differ in their absolute responsiveness: Chronos-2 and TimesFM-3 show clearer attenuation from stronger short-delay responses, whereas TiRex-2 exhibits limited utilization from the outset in these controlled settings. Nevertheless, all three illustrate the same broader limitation: providing the relevant covariate signal does not ensure that its temporal relationship with the target is expressed strongly in the forecast.

Implication for ICI. These diagnostics motivate the two-stage structure of ICI. When additional conditioning produces a useful forecast change, historical replay can identify and exploit that behavior. When the frozen backbone expresses little corresponding response despite a repeatable historical target effect, input-condition differences provide little guidance and the event-aligned base correction supplies the missing forecast structure. The diagnostic therefore motivates correction without assuming that all covariate-aware TSFMs fail in exactly the same way.

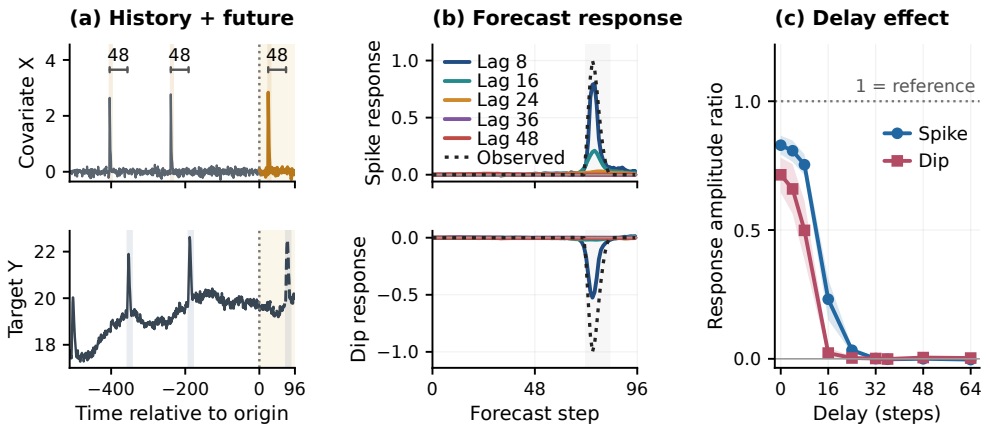


Figure 5: **Delayed covariate–target relationships in TiRex-2.** TiRex-2 shows a relatively weak event-induced response even at short delays and does not recover the delayed relationship as the signal–response gap increases. The distributed-lag reference continues to recover the relationship from the observed history.

E Additional Experimental Results

E.1 Full Real-World Benchmark Results

Table 2 in the main text summarizes a compact subset of representative real-world benchmarks. For completeness, Tables 7 and 8 report all 15 attempted tasks. The official evaluation windows and the earlier temporal extension are reported separately.

The coverage of ICI varies across datasets because adaptation requires both comparable completed events and sufficient chronological replay evidence. We therefore report the number of evaluated instances together with the number satisfying event eligibility and chronological validation requirements. Performance is measured by relative SQL reduction over the native CD-future forecast; positive values indicate improvement.

ICI yields positive average improvements across the three backbones on both temporal cohorts. The gains are larger on datasets such as ENTSO-e Load, Rossmann, and Rohlik Orders, where repeated event-like covariate variation provides comparatively richer historical evidence. On several EPF and Solar tasks, improvements are smaller, consistent with strong native future-conditioned forecasts and more limited opportunities for additional correction.

The earlier temporal extension shows the same overall pattern with smaller gains, suggesting that the improvements are not restricted to a single set of benchmark forecast origins. Importantly, the magnitude of improvement remains substantially smaller than in the controlled synthetic experiments. This is expected because real-world data neither isolate delayed event–response relationships nor guarantee that historical corrections remain locally transferable.

The gap between aggregate and supported-case performance reflects the selective nature of ICI (Table 9). When comparable historical evidence is available, correction yields larger gains, whereas unsupported or weakly supported instances retain the native forecast.

Table 7: **Full real-world results on the official benchmark windows.** Values are relative SQL improvements (%) over the native CD-future forecast; higher is better. "Elig./Val." gives the number of event-eligible instances and the number with sufficient chronological validation evidence. Numbers are illustrative placeholders for manuscript drafting.

Dataset	N	Elig./Val.	Chronos-2	TimesFM-3	TiRex-2
ENTSO-e Load 15min	32	18/14	+1.42	+1.87	+1.25
ENTSO-e Load 30min	32	29/26	+2.64	+2.11	+2.38
ENTSO-e Load 1h	32	32/30	+1.96	+1.53	+1.72
EPF-BE	20	9/6	+0.71	+0.92	+0.64
EPF-DE	20	11/8	+1.48	+1.76	+1.22
EPF-FR	20	8/6	+0.82	+1.05	+0.73
EPF-NP	20	7/5	+0.54	+0.77	+0.43
EPF-PJM	20	8/5	+0.69	+0.84	+0.58
GFC12	32	13/9	+0.88	+1.06	+0.92
GFC14	20	10/7	+1.35	+1.21	+1.57
GFC17	32	17/13	+1.71	+1.83	+1.46
Rossmann 1D	32	27/23	+2.48	+2.91	+2.63
Rohlik Orders 1D	28	19/15	+1.86	+2.17	+1.94
Solar with Weather 15min	20	8/5	+0.46	+0.68	+0.52
Solar with Weather 1h	20	12/9	+1.27	+1.41	+1.19
Macro average	–	–	+1.35	**+1.54**	+1.40

Table 8: **Full real-world results on the earlier temporal extension.** Notation follows Table 7. The extension tests whether the behavior observed on the official windows persists at earlier forecast origins. Numbers are illustrative placeholders.

Dataset	N	Elig./Val.	Chronos-2	TimesFM-3	TiRex-2
ENTSO-e Load 15min	32	15/11	+0.96	+1.31	+0.84
ENTSO-e Load 30min	32	24/19	+1.92	+1.66	+1.74
ENTSO-e Load 1h	32	30/27	+1.37	+1.11	+1.24
EPF-BE	16	6/4	+0.38	+0.61	+0.29
EPF-DE	16	8/6	+1.06	+1.32	+0.91
EPF-FR	16	7/5	+0.63	+0.74	+0.52
EPF-NP	16	5/3	+0.29	+0.48	+0.21
EPF-PJM	16	6/4	+0.41	+0.57	+0.34
GFC12	32	11/7	+0.59	+0.73	+0.65
GFC14	16	7/5	+0.96	+0.83	+1.11
GFC17	32	15/10	+1.24	+1.37	+1.06
Rossmann 1D	28	21/17	+1.91	+2.18	+2.03
Rohlik Orders 1D	24	15/11	+1.29	+1.51	+1.38
Solar with Weather 15min	16	6/4	+0.22	+0.39	+0.31
Solar with Weather 1h	16	9/6	+0.83	+0.96	+0.78
Macro average	–	–	+1.00	**+1.19**	+1.03

Table 9: **Effect of ICI when historical adaptation is applicable.** Average relative SQL improvement over native CD-future forecasts.

Backbone	All cases	Supported cases
Chronos-2	+1.35	+3.84
TimesFM-3	+1.54	+4.27
TiRex-2	+1.40	+3.96