DELAYCAST: EVENT-GATED LOW-RANK RESIDUAL ADAPTATION FOR TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models achieve strong forecasting performance from historical numerical observations, while external event notifications may provide additional information about future changes. Existing approaches incorporate external information through numerical covariates or multimodal fusion, yet effectively utilizing sparsely released textual notifications remains challenging when their availability and subsequent effects are temporally uncertain. To address this challenge, we propose DelayCast, an event-gated low-rank residual adaptation framework that incorporates structured information extracted from event notifications into frozen foundation-model forecasts. DelayCast combines historical numerical states with cutoff-available event attributes to generate constrained multi-horizon residual corrections, while preserving the original backbone forecast when no eligible event information is available. On the SpaceWeather dataset, DelayCast reduces the 72-hour forecasting mean absolute error of frozen Chronos-2 by 8.97% on the 2025 evaluation and 3.50% on a partial-year 2026 cohort. Beyond forecasting performance, we employ hierarchical comparisons to distinguish numerical calibration from the incremental contributions of event information, and further investigate the gap between the predictive value of retrospectively known response timing and the ability to recover such temporal structure from sparse event supervision.

1 INTRODUCTION

Recently, time-series foundation models (TSFMs) have demonstrated strong cross-dataset transferability and zero-shot forecasting capabilities through large-scale pretraining. Representative models such as Chronos Ansari et al. (2024), TimesFM Das et al. (2024), and Moirai Woo et al. (2024) have advanced general-purpose forecasting, while Chronos-2 Ansari et al. (2025) further supports multivariate and covariate-informed prediction. However, external information in real-world applications is not always available as regularly sampled and temporally aligned numerical covariates. Instead, it may arrive as sparse, asynchronously released event notifications whose potential effects have yet to appear in historical observations. Incorporating such event context into existing TSFMs while respecting information availability at each forecasting cutoff therefore remains an important challenge.

Prior work has incorporated exogenous variables, textual information, and event context into time-series forecasting through covariate encoding, attention, and multimodal fusion Lim et al. (2021); Wang et al. (2024); Xu et al. (2024a); Kurisinkel et al. (2024). Unlike regularly sampled covariates, however, event notifications are often sparse, asynchronous, and temporally uncertain: events occur intermittently, notification release times are not necessarily aligned with numerical observations, and the onset and duration of their subsequent effects may be unknown at the forecasting cutoff. Multiple events may also overlap, while cutoff-available information may differ from retrospectively confirmed records. Consequently, event information must not only satisfy availability constraints but may also affect only a subset of future forecasting steps. Explicitly modeling response onset, duration, and magnitude could capture such temporal dependencies, yet these parameters are difficult to estimate reliably from sparse and overlapping event observations.

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.

To address these challenges, we propose DelayCast, an event-gated low-rank residual adaptation framework for TSFMs. Rather than retraining the forecasting backbone or explicitly estimating a complete event-response kernel, DelayCast learns constrained multi-horizon residuals on top of frozen Chronos-2 forecasts. A deterministic eligibility rule first selects event records available at the forecasting cutoff. Historical numerical states and structured attributes of eligible events are then mapped to a small set of discrete cosine transform (DCT) coefficients, producing a smooth residual trajectory across the forecast horizon. Global shrinkage and residual magnitude constraints further limit the correction. When no event is eligible, the residual becomes exactly zero and the prediction falls back to the frozen backbone. This design enables event-conditioned forecast adaptation without explicitly recovering uncertain event-response mechanisms.

We evaluate DelayCast across multiple years on the SpaceWeather dataset. At the 72-hour forecasting horizon, DelayCast achieves an 8.97% reduction in mean absolute error relative to frozen Chronos-2 on the 2025 test set and a 3.50% reduction on the partial-year 2026 evaluation. We further investigate the recoverability of event-response timing structures under sparse event supervision to examine the relationship between the predictive value of event information and its underlying temporal response mechanisms.

Our main contributions are threefold:

- **Event-gated low-rank residual adaptation.** We propose DelayCast, which learns constrained low-rank corrections to frozen TSFM forecasts using cutoff-available structured event context, with exact fallback when no event is eligible.
- **Hierarchical attribution of event information.** We disentangle the incremental contributions of numerical calibration, event presence, and structured event attributes through hierarchical controls, complemented by cross-year evaluation and episode-clustered statistical analysis.
- **Event-response timing recoverability.** We investigate the recoverability and potential predictive value of event-response timing structures under sparse supervision through response detection, conditional onset prediction, and hindsight temporal-kernel experiments.

2 Related Work

2.1 Time-Series Foundation Models and Exogenous Forecasting

Recent advances in time-series foundation models have substantially improved forecasting performance across diverse tasks. Chronos Ansari et al. (2024) transforms continuous time series into discrete tokens through scaling and quantization and employs language-model architectures for probabilistic forecasting. TimesFM Das et al. (2024) adopts a patch-based decoder-only architecture for cross-domain zero-shot prediction, whereas Moirai Woo et al. (2024) employs a masked encoder and a unified training strategy to accommodate diverse sampling frequencies and numbers of variates.

For multivariate and exogenous forecasting, Chronos-2 Ansari et al. (2025) introduces group attention to support multivariate and covariate-informed prediction. Temporal Fusion Transformer Lim et al. (2021) combines variable selection networks, gating, and attention to handle heterogeneous inputs; TimeXer Wang et al. (2024) employs cross-attention to model interactions between target series and exogenous variables; and UniCA Han et al. (2026) uses covariate homogenization and unified attention-based fusion to incorporate heterogeneous covariates into pretrained TSFMs.

2.2 Lightweight Adaptation and Event-Response Modeling

For lightweight forecasting adaptation, Time-LLM Jin et al. (2024) reprograms time-series inputs into the representation space of a frozen large language model using input reprogramming and prompting, while ELF Lee et al. (2025) adapts frozen foundation-model forecasts using an online-trained auxiliary forecaster and dynamic forecast combination. Regarding compact forecast representations, N-BEATS Oreshkin et al. (2020) constructs forecasts through residual stacking and basis expansion, whereas FITS Xu et al. (2024b) performs interpolation in the complex frequency domain with a small parameter budget. Distributed lag nonlinear models (DLNMs) Gasparrini et al.

(2010) further employ cross-basis functions to jointly characterize nonlinear response relationships and their lagged effects.

In contrast, DelayCast uses event records available at the forecasting cutoff to perform event-conditioned residual correction on a frozen TSFM.

3 METHOD

3.1 PROBLEM FORMULATION

Given a forecasting decision time τ, let t denote the start timestamp of the latest completed numerical sampling interval, $\mathbf{x}_{\leq t}$ the numerical history indexed up to t, H the forecasting horizon, and $\mathcal{E}_\tau$ the candidate event records released by τ. Each record contains its release time and cutoff-available structured attributes, excluding retrospectively determined information such as realized response onset and duration.

Let F denote the frozen Chronos-2 model, and let $\hat{\mathbf{y}}^{\,0}$ be its H-step median (0.5-quantile) forecast based on the numerical history. DelayCast estimates a multi-horizon residual $\boldsymbol{\delta} \in \mathbb{R}^H$ conditioned on numerical states and structured attributes of eligible events. The baseline and final forecasts are defined as

$$\hat{\mathbf{y}}^{\,0} = Q_{0.5}(F(\mathbf{x}_{\leq t}))\,, \qquad \hat{\mathbf{y}} = \hat{\mathbf{y}}^{\,0} + g_\tau \boldsymbol{\delta}, \tag{1}$$

where $g_\tau \in \{0, 1\}$ is a deterministic indicator of eligible event-context availability. When $g_\tau = 0$, the correction is disabled, and the final forecast exactly matches the frozen backbone.

3.2 FRAMEWORK OVERVIEW

As illustrated in Fig. 1, DelayCast consists of three components: a frozen forecasting backbone, a deterministic event-eligibility gate, and a residual adapter. The backbone generates baseline forecasts from numerical history, while the gate selects event records available at the forecasting decision time. The adapter combines numerical-state features with structured attributes of eligible events to produce multi-horizon forecast corrections. During training, it learns from residuals between observed future values and baseline forecasts. At inference, the estimated correction is added to the baseline forecast, with exact fallback when no event is eligible.

3.3 FROZEN FORECASTING BACKBONE

We employ Chronos-2 as a frozen forecasting backbone and use its median forecasts as baseline point predictions. We extract numerical-state features from historical observations and baseline forecasts to characterize recent dynamics and predicted future trends. The backbone remains unchanged throughout training, while only the residual adapter is fitted.

3.4 EVENT ELIGIBILITY AND FEATURE CONSTRUCTION

We construct eligible event samples from notifications available at the forecasting decision time. For each materialized sample, the deterministic gate enables residual correction only when the sample represents an event window with an available notification:

$$g_\tau = \mathbb{I}\left[\text{event window} \wedge \text{notification available}\right]. \tag{2}$$

For eligible events, we extract cutoff-available timing, strength, and geometry attributes, together with missingness indicators for unavailable fields. These structured event features are combined with numerical-state features as input to the residual adapter. The gate indicates the availability of eligible event context rather than estimating the probability of an actual event response.

3.5 LOW-RANK RESIDUAL ADAPTATION

To restrict the degrees of freedom of forecast correction, we represent the H-step residual using a truncated discrete cosine transform (DCT) basis. Let $\mathbf{B}_H \in \mathbb{R}^{H \times K}$ denote the orthonormalized cosine basis and $\mathbf{z}_\tau$ the combined numerical and structured event features. A Ridge regressor predicts

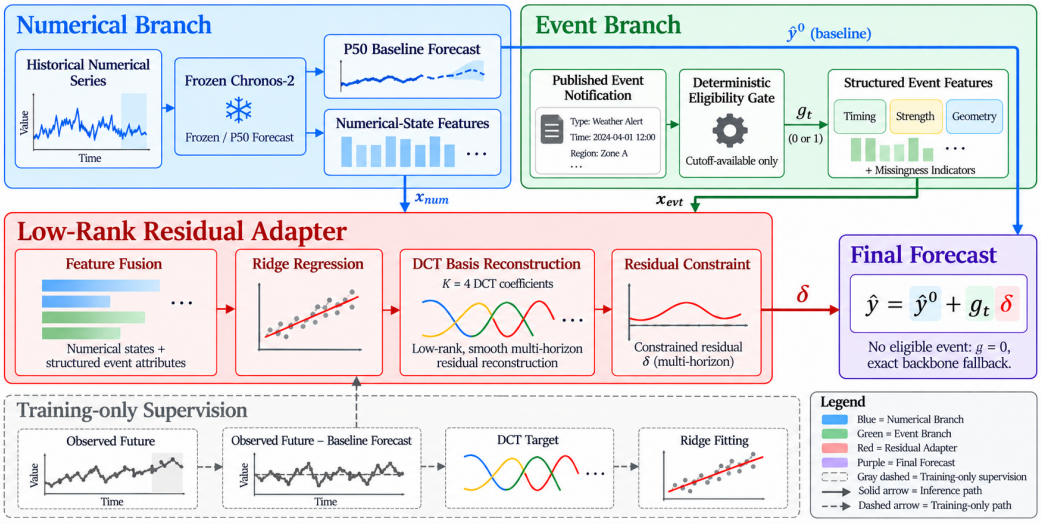

Figure 1: Overview of DelayCast. Historical numerical observations are processed by frozen Chronos-2 to obtain baseline forecasts and numerical-state features. A deterministic gate selects cutoff-available event information, which is combined with numerical features to predict low-dimensional DCT coefficients. The reconstructed residual is constrained by shrinkage and magnitude clipping before being added to the baseline forecast. The dashed path indicates training-only supervision; without an eligible event, the forecast exactly matches the backbone.

the low-dimensional coefficient vector $\hat{\mathbf{a}}_\tau \in \mathbb{R}^K$, from which the residual trajectory is reconstructed:

$$\hat{\mathbf{a}}_\tau = f_\theta(\mathbf{z}_\tau), \qquad \tilde{\mathbf{r}}_\tau = \mathbf{B}_H \hat{\mathbf{a}}_\tau. \tag{3}$$

During training, we construct residual targets from the differences between observed future values and frozen backbone forecasts. For each training sample i, the target coefficients are obtained by projecting the residual onto the orthonormalized DCT basis:

$$\mathbf{a}_i^* = \mathbf{B}_H^\top \left(\mathbf{y}_i - \hat{\mathbf{y}}_i^0\right). \tag{4}$$

We standardize the input features and fit a multi-output Ridge regressor by minimizing

$$\min_{\mathbf{W}} \left\| \mathbf{Z}\mathbf{W} - \left(\mathbf{A}^* - \mathbf{1}\bar{\mathbf{a}}^{*\top}\right) \right\|_F^2 + \alpha \|\mathbf{W}\|_F^2, \tag{5}$$

where $\mathbf{Z}$ is the standardized training feature matrix, $\mathbf{A}^*$ contains the target coefficient vectors, and $\bar{\mathbf{a}}^*$ is their training-set mean. The mean coefficients are added back to the regressor output at inference.

To limit excessive forecast corrections, we apply global shrinkage and a smooth magnitude constraint to the reconstructed residual:

$$\boldsymbol{\delta} = s\, c_H \tanh\left(\frac{\tilde{\mathbf{r}}_\tau}{c_H}\right), \tag{6}$$

where s denotes the shrinkage factor and c_H is a horizon-specific magnitude threshold determined from the absolute training residuals. The resulting correction is applied through the deterministic gate in Eq. (1).

4 Experiments

4.1 Experimental Setup

Dataset and evaluation protocol. We evaluate DelayCast on the SpaceWeather dataset, comprising half-hourly geomagnetic activity measurements and associated event notifications. Each forecasting sample contains 672 historical observations spanning 14 days. We consider forecasting

horizons of 48, 72, and 96 hours, corresponding to $H \in \{96, 144, 192\}$, with 72 hours as the primary setting. Statistical analyses use 72-hour event episodes as clustering units. Numerical histories are restricted to completed sampling intervals, and event features are extracted only from notifications released by the forecasting decision time.

Following rolling-origin evaluation practices Tashman (2000); Cerqueira et al. (2020), we adopt an expanding-window, rolling-year development protocol. We train on 2019–2021, 2019–2022, and 2019–2023 and validate on the respective subsequent years. After model selection, we refit the adapter on the complete 2019–2024 development set, followed by a legacy replication on 2025 data and an evaluation on a partial-year 2026 cohort.

Baselines and ablations. To better understand the sources of forecasting gains, we construct four hierarchical comparisons: Base, Numeric, Presence, and Structured. Base uses frozen Chronos-2 median forecasts. Numeric applies low-rank residual calibration using numerical-state features alone. Presence additionally incorporates event availability, while Structured further includes timing, strength, and geometry attributes, representing the full DelayCast model. All residual variants share the same training samples, model family, and hyperparameter search budget. We additionally perform matched feature-group ablations and shuffled-attribute controls to examine the contributions of structured event information.

Implementation details. We freeze all Chronos-2 parameters and train a separate residual adapter for each forecasting horizon. The backbone receives only univariate numerical histories, while event information is processed by the post-hoc adapter. We search over the number of DCT components $K \in \{4, 8\}$, Ridge regularization $\alpha \in \{1, 10, 100\}$, shrinkage $s \in \{0.5, 0.75, 1\}$, and magnitude-cap quantile $q \in \{0.95, 0.99\}$. The quantile q determines the threshold c_H from the absolute training residuals for each horizon. Hyperparameters are selected using pooled 72-hour event-message MAE across the three development folds and reused across horizons, with separate final models fitted for each horizon. The selected configuration is $K = 4$, $\alpha = 100$, $s = 0.75$, and $q = 0.99$.

Evaluation metrics. We assess forecasting performance using mean absolute error (MAE) and report relative MAE reductions against frozen Chronos-2. Primary comparisons are conducted on event-message forecasts, while ordinary windows are used to verify exact fallback when no eligible event context is available. Statistical uncertainty is quantified using 10,000 episode-clustered bootstrap resamples with 95% confidence intervals. All messages within each 72-hour event episode are retained together during resampling.

4.2 Main Forecasting Results

Table 1 reports forecasting performance across horizons and evaluation periods, while Fig. 2 illustrates the relative gains and episode-level variation at the primary 72-hour horizon. DelayCast achieves lower MAE than frozen Chronos-2 across all forecasting horizons and evaluation periods. During development, it reduces MAE by 2.84%, 3.30%, and 3.08% at the 48-, 72-, and 96-hour horizons, respectively.

In the 2025 legacy replication, 72-hour MAE decreases from 1.376 to 1.253, corresponding to an 8.97% reduction, alongside improvements of 7.34% and 8.54% at the other two horizons. On the partial-year 2026 cohort, DelayCast maintains positive point estimates across horizons, with MAE reductions of 3.50% and 4.80% at 72 and 96 hours, respectively. Figure 2(a) further shows the episode-clustered uncertainty associated with these cross-period comparisons, while Fig. 2(b) presents the variation in forecasting gains across the independent 2026 event episodes.

4.3 Hierarchical Gain Attribution

To further examine the sources of forecasting gains, we compare the four hierarchical variants: Base, Numeric, Presence, and Structured. As shown in Table 2, MAE decreases progressively across the four variants at all three forecasting horizons. For the primary 72-hour task, Numeric reduces MAE from 1.2825 to 1.2561, corresponding to a 2.06% improvement over Base. Incorporating event availability further reduces MAE to 1.2475, while adding structured event attributes yields an MAE of 1.2403, representing a total reduction of 3.30% relative to Base. This hierarchy separates

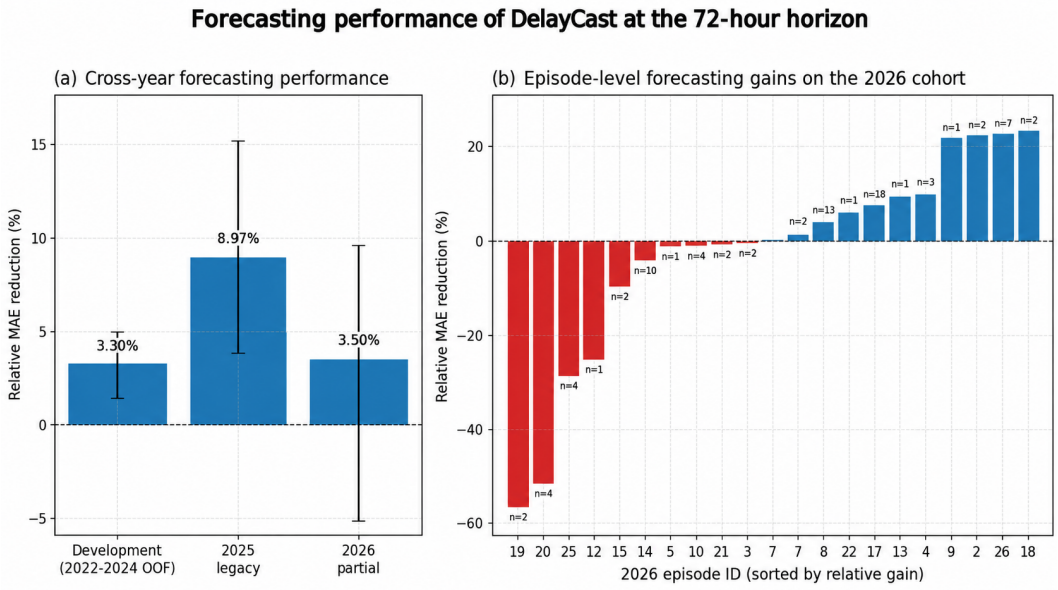

Figure 2: Forecasting performance of DelayCast at the primary 72-hour horizon. (a) Relative MAE reduction compared with frozen Chronos-2 across development, 2025 legacy replication, and partial-year 2026 evaluation. Error bars indicate episode-clustered 95% confidence intervals. (b) Episode-level relative MAE reductions for the 20 independent event episodes in the 2026 cohort, sorted by improvement.

Table 1: Forecasting performance across evaluation periods. All values are event-message MAE; lower is better. Relative reductions are computed against frozen Chronos-2 within each period. At the primary 72-hour horizon, episode-clustered 95% confidence intervals for relative MAE reduction are [1.41%, 4.92%] for development, [3.83%, 15.10%] for the 2025 legacy replication, and [-5.14%, 9.57%] for the partial-year 2026 evaluation.

Period	Horizon	Chronos-2	DelayCast	Reduction
Development	48 h	1.2096	**1.1752**	2.84%
	72 h	1.2825	**1.2403**	3.30%
	96 h	1.2722	**1.2330**	3.08%
2025	48 h	1.258	**1.165**	7.34%
	72 h	1.376	**1.253**	8.97%
	96 h	1.391	**1.272**	8.54%
2026 partial	48 h	1.2930	**1.2792**	1.07%
	72 h	1.3833	**1.3349**	3.50%
	96 h	1.3673	**1.3016**	4.80%

the observed forecasting gains into numerical calibration, event availability, and structured event attributes, providing a basis for subsequent feature ablations. Episode-clustered confidence intervals confirm the overall reduction relative to Base, while the incremental gains from event availability and structured attributes remain less certain (Appendix C).

4.4 Event-Feature Ablation

We conduct feature-group ablations at the primary 72-hour horizon to examine the contributions of different event attributes. As shown in Table 3, removing strength attributes increases MAE from 1.2403 to 1.2490, whereas removing timing or geometry attributes yields MAEs of 1.2396 and 1.2401, respectively. The strength-only variant achieves an MAE of 1.2396, comparable to the full model. Furthermore, shuffling structured event attributes during training increases MAE to 1.2519.

Table 2: Hierarchical comparison on the development folds. Values are pooled out-of-fold event-message MAE; lower is better.

Model	48 h	72 h	96 h
Base	1.2096	1.2825	1.2722
Numeric	1.1804	1.2561	1.2489
Presence	1.1770	1.2475	1.2395
Structured	**1.1752**	**1.2403**	**1.2330**

Table 3: Event-feature ablation on the development folds at the 72-hour forecasting horizon. All variants are refitted using matched training protocols. Lower MAE is better.

Variant	MAE	Reduction vs. Base
Full Structured	1.2403	3.30%
w/o Timing	**1.2396**	3.34%
w/o Strength	1.2490	2.61%
w/o Geometry	1.2401	3.31%
Strength Only	**1.2396**	3.34%
Shuffled Structured	1.2519	2.39%

These results suggest that strength attributes provide the main predictive signal among the structured event features in the current experiments.

4.5 Event-Response Timing Analysis

The preceding feature ablations reveal differences in the predictive contributions of structured event attributes, raising two further questions: whether event-response timing offers additional forecasting value and whether such temporal structure can be reliably recovered using information available at prediction time. To distinguish its potential value from practical predictability, we investigate response detection, conditional onset prediction, and forecasting with retrospectively known temporal information. These analyses use development-period event data and are separate from the deployed residual adapter.

To assess the recoverability of event-response timing, we examine both response detection and conditional onset prediction. On the 2024 validation set, logistic regression incorporating structured event attributes achieves an AUPRC of 0.426 and an AUROC of 0.570, compared with 0.404 and 0.529 using numerical features alone. Its Brier score, however, remains at 0.245, slightly above the prevalence baseline of 0.240. Among events with detected responses, a structured ordinal model reduces episode-equal onset MAE from 11.679 to 11.347 hours, although it does not improve prediction accuracy within either ± 6 or ± 12 hours. These results reveal some predictive information about response occurrence and timing, while highlighting the difficulty of precisely recovering response onset.

The limited precision of predicted onset raises a complementary question: how much forecasting value could accurate temporal information provide if it were known? We investigate this through a hindsight experiment on 35 uncensored detected messages from 17 episodes. As shown in Table 4, frozen Chronos-2 achieves an MAE of 1.584, which decreases to 1.511 with direct DCT residual adaptation. In contrast, an oracle incorporating retrospectively known response information reduces MAE to 1.068. At the episode level, the oracle improves absolute MAE over the frozen backbone by 0.317 (95% CI: [0.111, 0.570]). By using realized response information unavailable at the forecasting cutoff, this experiment isolates the potential predictive value of temporal structure.

Together, these experiments reveal a gap between the potential forecasting value of accurate response timing and its recoverability from sparse, potentially overlapping event observations. This motivates the use of direct low-rank residual adaptation in DelayCast rather than an explicitly predicted event-response kernel.

Table 4: Event-response timing analysis. Conditional onset is evaluated using episode-equal MAE in hours, whereas hindsight forecasting uses message-level MAE. The two tasks involve different evaluation subsets and metrics.

Task	Model	MAE
Conditional onset	Median baseline	11.679 h
	Structured ordinal	**11.347 h**
Hindsight forecasting	Frozen Chronos-2	1.584
	Direct DCT	1.511
	Oracle	**1.068**

5 Discussion

Our findings demonstrate the potential of lightweight residual adaptation for incorporating sparse event information into frozen time-series foundation models. However, hierarchical comparisons reveal that a substantial portion of the forecasting gains arises from numerical calibration rather than structured event attributes alone. The timing analysis further highlights a distinction between the predictive value of accurate temporal information and the ability to recover it from cutoff-available observations. Under sparse and overlapping event supervision, these findings motivate direct, constrained residual adaptation without relying on an explicitly predicted event-response kernel.

6 Conclusion

In this paper, we introduced DelayCast, an event-gated low-rank residual adaptation framework for frozen time-series foundation models. By combining numerical states with cutoff-available structured event information, DelayCast generates constrained multi-horizon corrections while preserving the backbone forecast when no eligible event is available. Cross-year evaluations demonstrate its forecasting performance, while hierarchical comparisons and response-timing analyses distinguish the contributions of numerical calibration, event information, and temporal recoverability. These findings provide empirical insights into adapting foundation models under sparse event supervision.

AI use statement

During the preparation of this manuscript, we used generative AI tools, such as ChatGPT, solely for language polishing, grammar checking, and improving clarity of expression. These tools were not used to formulate the research questions, design the methodology, conduct experiments, analyze data, or draw conclusions. The authors take full responsibility for the content, technical contributions, experimental results, accuracy, completeness, and originality of this work.

Ethics statement

This study does not involve human participants or personal data. The experiments use historical geomagnetic observations and associated event notifications. We distinguish information available at the forecasting cutoff from retrospectively determined response labels and explicitly report the limitations of retrospective data versions. The findings are intended to support research on time-series forecasting rather than establish causal relationships between individual notifications and observed responses.

Reproducibility statement

Section 3 describes the proposed framework, including event eligibility, feature construction, residual parameterization, and optimization. Section 4 provides the evaluation protocol, baseline configurations, hyperparameter search, and statistical procedures. Additional details on data alignment, structured feature extraction, implementation, and supplementary experimental results are provided 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, 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. *Transactions on Machine Learning Research*, 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. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, 2020. doi: 10.1007/s10994-020-05910-7.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://mlanthology.org/icml/2024/das2024icml-decoderonly/.

Antonio Gasparrini, Ben Armstrong, and Michael G. Kenward. Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234, 2010. doi: 10.1002/sim.3940.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Yinbo Sun, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, and De-Chuan Zhan. UniCA: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0b5eb45a22ff33956c043dd271f244ea-Abstract-Conference.html.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/680b2a8135b9c71278a09cafb605869e-Abstract-Conference.html.

Litton Jose Kurisinkel, Pruthwik Mishra, and Yue Zhang. Text2TimeSeries: Enhancing financial forecasting through time series prediction updates with event-driven insights from large language models. *arXiv preprint arXiv:2407.03689*, 2024. doi: 10.48550/arXiv.2407.03689. URL https://arxiv.org/abs/2407.03689.

Thomas L. Lee, William Toner, Rajkarn Singh, Artjom Joosen, and Martin Asenov. Lightweight online adaption for time series foundation model forecasts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33736–33764. PMLR, 2025. URL https://proceedings.mlr.press/v267/lee25ag.html.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000. doi: 10.1016/S0169-2070(00)00065-0.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2402.19072.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Zhijian Xu, Yuxuan Bian, Jianyuan Zhong, Xiangyu Wen, and Qiang Xu. Beyond trend and periodicity: Guiding time series forecasting with textual cues. *arXiv preprint arXiv:2405.13522*, 2024a. doi: 10.48550/arXiv.2405.13522. URL https://arxiv.org/abs/2405.13522.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/701251e1db4a2e4dd2ef23f5265d5936-Abstract-Conference.html.

A Data Construction and Temporal Alignment

A.1 Data Sources and Event Records

The forecasting task uses half-hourly Hp30 geomagnetic observations from GFZ and event notifications associated with NASA DONKI WSA–ENLIL simulations. The numerical series provides historical observations and forecasting targets, while published notifications supply structured event information.

Event samples are constructed by linking WSA–ENLIL simulations containing CME inputs to their associated notifications. We retain notifications with valid publication and model completion times, requiring publication to occur after model completion and no later than the estimated shock arrival when an arrival estimate is available. For each simulation, the earliest qualifying notification is selected. The resulting records are organized by unique notification identifiers and retained only when the required numerical history and forecasting targets are available.

Each forecasting sample corresponds to a materialized notification record. Multiple notifications are not jointly encoded by a set-based aggregation module.

A.2 Decision Time and Numerical Alignment

We distinguish the forecasting decision time from the timestamp indexing the last numerical observation. For an event notification published at time τ, define

$$A = \lfloor \tau \rfloor_{30\mathrm{min}}, \qquad t = A - 30\mathrm{min}. \tag{7}$$

Here, t is the start timestamp of the latest completed half-hourly numerical interval. That interval ends at A, which is no later than the notification release time τ.

The numerical input consists of 672 observations indexed up to t, corresponding to 14 days of historical data. To preserve the forecasting alignment, Chronos-2 generates $H+1$ forecast steps. Its first output is discarded, and the subsequent H steps form the evaluated baseline:

$$\hat{\mathbf{y}}^{\,0} = \hat{\mathbf{y}}^{\,\mathrm{Chronos},P50}_{\,2:H+1}. \tag{8}$$

Under the interval-start convention, the first evaluated target interval begins at $t + 60\mathrm{min}$.

This construction separates numerical-history availability from notification-release time and prevents event attributes released after the decision time from entering the prediction.

A.3 Evaluation Cohorts

Model development uses data from 2019–2024 with three expanding-window validation folds: training through 2021 and validating on 2022, training through 2022 and validating on 2023, and training through 2023 and validating on 2024.

The selected configuration is subsequently refitted using the complete 2019–2024 development data. The 2025 cohort serves as a legacy replication, whereas the partial-year 2026 cohort is evaluated with the frozen adapter and evaluation protocol.

The 2026 cohort contains 82 eligible messages with complete numerical history and forecasting targets across all three horizons, corresponding to 20 independent 72-hour event episodes. Ten additional messages are excluded because the downloaded numerical data do not provide the required historical prefix.

The historical numerical files were retrieved retrospectively and represent their available data versions at retrieval time. The evaluation therefore does not constitute a fully archived point-in-time online replay.

B Features and Implementation Details

B.1 Numerical-State Features

The numerical-state representation contains ten features derived from the historical Hp30 series and the frozen Chronos-2 median forecast.

Six features summarize historical observations: the most recent value; the mean and standard deviation over the latest 48 observations; the mean and standard deviation over the latest 336 observations; and the difference between the most recent observation and the observation indexed 48 positions earlier.

Four additional features summarize the H-step baseline forecast: its mean, standard deviation, minimum, and maximum.

These features describe the recent numerical state and the predicted trajectory without updating the forecasting backbone.

B.2 Structured Event Features

For each eligible notification, the structured feature extractor produces eight attributes organized into three groups.

Table 5: Structured event attributes used by the final residual adapter.

Group	Feature
Timing	Published arrival lead time
	Published arrival uncertainty
Strength	Published activity estimate: lower bound
	Published activity estimate: upper bound
	Maximum reported speed
	Mean reported speed
	Reported CME count
Geometry	Maximum reported opening half-angle

Each structured feature is accompanied by a missingness indicator. Missing numerical values in these fields are replaced by zero before feature standardization.

The complete Structured input contains 37 dimensions: ten numerical-state features, one event-presence indicator, ten interactions between numerical-state features and event presence, eight structured event attributes, and eight missingness indicators.

The model operates on extracted structured attributes rather than raw notification text or text embeddings.

B.3 Feature Normalization and Ridge Fitting

For each forecasting horizon, feature normalization parameters are estimated using the corresponding training data. Let $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the feature-wise training means and standard deviations. Standardized features are computed as

$$\mathbf{z}_i = \frac{\mathbf{f}_i - \boldsymbol{\mu}}{\boldsymbol{\sigma}}. \tag{9}$$

Standard deviations smaller than 10^{-6} are replaced by one. The resulting training statistics are saved and reused at inference.

The truncated cosine basis is obtained from

$$C_{h,k} = \cos\left(\frac{\pi(h+1/2)k}{H}\right), \quad k = 0, \ldots, K-1, \tag{10}$$

followed by QR orthonormalization to produce the basis $\mathbf{B}_H$.

For each training sample, we project the residual target onto this basis. A multi-output Ridge regressor then predicts the corresponding coefficient vector using standardized input features.

The training-set mean of the target coefficients is subtracted before fitting and added back at inference. All training samples receive equal weight.

B.4 Residual Constraints and Frozen Configuration

Predicted coefficients are projected back to the forecasting horizon to obtain the raw residual trajectory.

A horizon-specific magnitude threshold c_H is estimated from the selected quantile of absolute training residuals, pooling all training samples and forecast steps within that horizon.

The constrained residual is computed as

$$\boldsymbol{\delta}_\tau = s\, c_H \tanh\left(\frac{\tilde{\mathbf{r}}_\tau}{c_H}\right). \tag{11}$$

The deterministic event mask is subsequently applied before adding the correction to the frozen Chronos-2 median forecast.

The final configuration uses $K = 4$, Ridge regularization $\alpha = 100$, shrinkage $s = 0.75$, and residual-cap quantile $q = 0.99$.

Separate adapters are fitted for $H \in \{96, 144, 192\}$, each with its own basis, Ridge parameters, normalization statistics, and residual cap.

C Supplementary Forecasting Results and Statistical Analysis

C.1 Hierarchical Gain Attribution

Table 6 provides the episode-clustered confidence intervals for the hierarchical comparisons at the primary 72-hour horizon.

The overall reduction from Structured relative to Base is supported by a positive episode-clustered interval. Numerical calibration accounts for an appreciable part of this reduction, whereas the incremental contributions of event presence and structured attributes are less certain.

The exact additive decomposition applies to absolute MAE differences. Relative percentage gains computed against different reference models should not be added directly.

Table 6: Hierarchical forecasting gains during development. Relative improvements use the preceding model as the reference, except for the final comparison.

Comparison	Gain	95% CI
Numeric vs. Base	2.06%	[0.37%, 3.86%]
Presence vs. Numeric	0.68%	[-0.30%, 1.52%]
Structured vs. Presence	0.58%	[-0.37%, 1.49%]
Structured vs. Base	3.30%	[1.41%, 4.92%]

C.2 Additional Feature Ablations

The primary feature-group ablations are reported in Table 3. An additional timing-only control obtains an MAE of 1.2494 at the 72-hour horizon, corresponding to a 2.58% reduction relative to Base.

Removing strength attributes from the full Structured model produces an additional relative gain of 0.70% for the full model, with an episode-clustered 95% confidence interval of [-0.07%, 1.51%].

The corresponding point estimates for timing and geometry do not show additional improvements under the matched ablation protocol.

C.3 Cross-Year Statistical Evidence

At the primary 72-hour horizon, the 2025 legacy replication achieves an 8.97% relative MAE reduction, with an episode-clustered 95% confidence interval of [3.83%, 15.10%].

The partial-year 2026 evaluation achieves a 3.50% relative reduction, with a corresponding interval of [-5.14%, 9.57%].

The 2026 evaluation contains 20 independent event episodes. Its largest episode includes 18 of the 82 eligible messages, and ten episodes exhibit positive average forecasting gains.

Additional hierarchical comparisons on the 2026 cohort were performed after the primary evaluation and are treated as post-hoc diagnostics rather than model-selection evidence.

No ordinary-control windows are available in the final 2026 cohort. Exact fallback is guaranteed by the deterministic event mask and was evaluated on earlier data.

D Supplementary Event-Response Timing Analysis

D.1 Response Labels and Data Integrity

The timing analysis uses an audited response-label dataset constructed from the 2019–2024 development period.

Each primary notification is associated with a deterministically selected candidate arrival. Candidate selection uses notification-linked arrival information and simulation identifiers without accessing future Hp30 outcomes.

Predictor features and retrospectively determined response labels are stored separately. The label records include response-detection status, observed onset and peak times, end times or censoring information, overlapping-event indicators, and episode identifiers.

For right-censored responses, observed end times and realized durations remain missing rather than being replaced with the censoring limit.

The audited dataset contains 319 primary notification messages, including 98 messages with detected responses and 221 without detected responses. The detected responses span 48 event episodes.

Only development-period records are used in these analyses.

D.2 Response Detection

We compare a prevalence baseline, a numerical logistic model, and a structured logistic model for predicting whether a detectable response occurs.

Table 7: Response-detection performance on the 2024 validation data.

Model	AUPRC	AUROC	Brier
Prevalence	0.364	0.500	**0.240**
Numeric logistic	0.404	0.529	0.246
Structured logistic	**0.426**	**0.570**	0.245

The structured model improves the ranking metrics relative to the numerical model, while its Brier score does not improve upon the prevalence baseline.

D.3 Conditional Onset Prediction

Conditional onset prediction is evaluated only for messages with detected responses.

The fixed 2024 validation subset contains 43 detected messages from 18 episodes. A structured ordinal model is compared with a training-set median baseline using episode-equal onset MAE.

The median baseline obtains 11.679 hours, while the selected structured ordinal model obtains 11.347 hours. The absolute improvement is 0.332 hours, with an episode-clustered 95% confidence interval of [0.001, 0.666] hours.

The structured model does not improve accuracy within either six or twelve hours of the observed onset.

D.4 Hindsight Forecasting

We evaluate the potential value of retrospectively known temporal information on 35 uncensored detected messages from 17 event episodes.

Table 8: Hindsight forecasting results. Message-level and episode-equal MAE are reported separately.

Model	Message MAE	Episode-equal MAE
Frozen Chronos-2	1.584	1.309
Direct DCT	1.511	1.291
Oracle	**1.068**	**0.991**

Using retrospectively determined response information, the oracle achieves an episode-equal absolute MAE improvement of 0.317 over frozen Chronos-2, with a 95% confidence interval of [0.111, 0.570].

The contrast between the oracle and predicted-onset results illustrates the difference between the predictive value of temporal structure and its recoverability from available event observations.