HIP: FORECAST-CALIBRATED EVENT REPRESENTATIONS FOR MULTI-HORIZON TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecast-relevant events do not by themselves specify how their influence should be represented across a future trajectory. We introduce **Hierarchical Impact Pathways (HIP)**, a structured intermediate representation that maps semantic event hypotheses and temporal attributes into horizon-wise conditioning signals for multi-horizon forecasting. HIP explicitly separates what an event is hypothesized to affect from how its influence is represented across future prediction steps. Because semantically plausible event descriptions are not necessarily predictively useful, we further introduce **RL-HIP**, which calibrates event retention and pathway attributes using downstream forecasting feedback while keeping the forecasting model and pathway policy separately trained. We evaluate the framework on electricity-demand and Bitcoin forecasting with multiple numerical backbones. RL-HIP provides strong and consistent improvements for electricity demand, while the gains on Bitcoin are smaller and more architecture-dependent. These results suggest that the utility of external events in multi-step forecasting depends not only on their relevance, but also on how their predicted influence is represented and calibrated across future horizons.

1 INTRODUCTION

External events can change a future trajectory before their effects are visible in the observed series. Modern numerical forecasters are increasingly effective at extrapolating temporal patterns (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Zeng et al., 2023; Nie et al., 2022; Liu et al., 2024b), while recent multimodal and event-aware methods incorporate textual context through prompting, retrieval, alignment, multimodal fusion, and event reasoning (Liu et al., 2024a; Williams et al., 2025; Jia et al., 2024; Zhong et al., 2025; Zhang et al., 2025a; Wang et al., 2024; 2026; Ge et al., 2026; Zhang et al., 2025b). These developments establish that external context can be useful, but they leave a distinct multi-horizon question unresolved: *once an event is judged relevant, how should its influence be distributed across future prediction steps?* Relevance identifies what may matter; it does not specify when an effect should begin, how long it should persist, or how strongly it should condition different parts of the future trajectory.

We view this as a representation problem rather than only a retrieval problem. Two relevant events can imply different temporal profiles. A heatwave warning may raise residential cooling demand after a short delay and remain influential for several hours, whereas a transport disruption may affect a different demand component over a shorter interval. Representing both events only as retrieved text, a static embedding, or free-form reasoning leaves the event-to-trajectory mapping implicit. Our goal is therefore not to estimate real-world causal effects, but to construct an explicit, forecast-oriented intermediate representation that can express how an event *may* influence multiple future horizons.

In this paper, we introduce **Hierarchical Impact Pathways (HIP)**. HIP maps a retained event to a structured hypothesis containing semantic attributes—event category, hypothesized mechanism, affected component, and direction—together with temporal and calibration attributes—lag, duration, persistence, intensity, and uncertainty. These attributes are not endpoints by themselves: they are

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.

converted into a horizon-wise conditioning profile, yielding the sequence

$$\text{event} \rightarrow \text{structured impact hypothesis} \rightarrow \text{horizon-wise conditioning signal} \rightarrow \text{forecast}.$$

The central role of HIP is therefore to transform static event information into a representation that can vary across future prediction steps.

A second gap arises because semantic plausibility is not the same as predictive usefulness. An LLM can produce a reasonable pathway, yet assign an event excessive intensity, an unsuitable onset, or a persistence pattern that is poorly aligned with the forecasting task. We therefore introduce **RL-HIP**, which uses downstream forecasting feedback to calibrate event retention and pathway attributes. The LLM-generated pathway serves as a structured prior, while reinforcement learning adjusts the pathway policy within the constrained HIP space. This differs from optimizing free-form textual reasoning: the object being calibrated is an explicit forecast-oriented representation whose attributes are subsequently mapped to different future horizons.

We evaluate the framework on two multi-horizon domains with different temporal resolutions: Australian regional electricity demand, where the previous 24 hours are used to forecast the next 12 hours at 30-minute resolution ($L = 48, H = 24$), and daily Bitcoin prices, where the previous seven days are used to forecast the next seven days ($L = 7, H = 7$). Across four numerical backbones, RL-HIP achieves the lowest mean MAE on all four AULF models and on three of four numerical Bitcoin backbones; on AULF, the reduction relative to No events is 5.4–10.9%. The representative LLM-forecasting interface is less consistent, which we treat as a scope result rather than evidence of universal event-conditioning gains. Supplementary experiments in the appendix examine forecast calibration, Static versus Temporal conditioning, and alternative forecast geometries; these analyses are used to characterize when the representation is useful rather than as additional headline claims.

The contribution of this work is threefold in substance, though tightly connected in one framework. We formulate event-conditioned forecasting as a horizon-resolved representation problem; we introduce HIP as a structured interface that maps semantic and temporal event hypotheses into future-step-specific conditioning signals; and we develop forecast-feedback calibration to align those hypotheses with downstream predictive utility. The resulting framework separates event relevance, structured representation, and forecasting calibration while keeping causal interpretation outside the scope of the method.

2 Related Work

Numerical and Language-Model Time-Series Forecasting. Modern forecasting spans Transformer, linear, patch-based, inverted-Transformer, convolutional, and multi-scale architectures, including Informer, Autoformer, FEDformer, DLinear, PatchTST, iTransformer, ModernTCN, and TimeMixer++ (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Zeng et al., 2023; Nie et al., 2022; Liu et al., 2024b; Luo & Wang, 2024; Wang et al., 2025). Language and foundation models have also been adapted to temporal data, with methods such as TimeCMA and CALF aligning time-series and language-model representations (Liu et al., 2025b;c). RL-HIP is complementary to these backbone developments: it changes how external events are represented before conditioning the forecaster rather than redesigning the forecasting architecture.

Text-Enhanced and Event-Aware Forecasting. Multimodal benchmarks including Time-MMD, MoTime, and the TimeText Corpus support evaluation of aligned textual and numerical forecasting (Liu et al., 2024a; Zhou et al., 2025; Kim et al., 2024), and systematic evidence shows that multimodal gains depend strongly on whether text contributes predictive information beyond the numerical history (Zhang et al., 2025a). Existing methods inject text through prompting (Jia et al., 2024; Williams et al., 2025; Liu et al., 2025a), frequency or representation alignment (Lo et al., 2024; Seo et al., 2024; Zhou et al., 2024; Yao et al., 2025), vision-language or multimodal fusion (Zhong et al., 2025; Jiang et al., 2025; Chen et al., 2026), or trajectory ranking with textual context (Nguyen et al., 2026). These approaches demonstrate the utility of external context, but they typically represent the event-to-trajectory relation through prompts, latent alignment, or generated reasoning rather than through an explicit intermediate profile whose values are resolved over forecast horizons.

Event-aware forecasting has a longer domain-specific history in finance and demand prediction (Schumaker & Chen, 2009; Rodrigues et al., 2019). Recent methods use LLM-derived event insights (Kurisinkel et al., 2024), news retrieval and filtering with reflection (Wang et al., 2024),

event-driven reasoning and alignment (Wang et al., 2026), and event-aware multimodal routing (Du et al., 2025). CAMEF introduces causal learning and counterfactual event augmentation (Zhang et al., 2025b), EventTSF models fine-grained event–series synchronization (Ge et al., 2026), and hierarchical or structured event representations have also been explored through semantic graphs and explicit event dimensions (Cen et al., 2026; Zhu et al., 2026). HIP differs by representing each retained event as a forecast-oriented state with an affected component, hypothesized mechanism, direction, lag, persistence, intensity, and uncertainty, which is then converted into a horizon-wise conditioning profile.

Forecast-Feedback Calibration. Downstream forecast feedback has been used to revise textual reasoning or event information, including reflection in From News to Forecast (Wang et al., 2024), error-driven critique and refinement in TimeXL (Jiang et al., 2025), and reinforcement learning over textual content in Text Reinforcement (Su et al., 2025). RL-HIP instead applies forecast feedback to a constrained pathway space, calibrating event retention and explicit semantic and temporal attributes that are subsequently converted into horizon-wise conditioning signals. Standard policy-gradient methods such as REINFORCE and PPO provide the optimization machinery (Williams, 1992; Schulman et al., 2017); our contribution is the structured impact representation and its coupling to multi-horizon forecasting.

Table 1: Comparison with representative event-aware and text-enhanced forecasting methods. "Forecast feedback" denotes any explicit use of forecasting error or outcomes to revise event/text processing; "direct structured calibration" is reserved for methods that directly optimize a constrained structured state such as HIP. The table is intended to clarify the interface, not to imply that a checkmark alone establishes empirical superiority.

Method	External events	Event selection	Structured impact state	Mechanism / component	Explicit temporal attributes	Horizon-wise impact profile	Forecast feedback	Direct structured calibration
From News to Forecast (Wang et al., 2024)	✓	✓	–	–	–	–	✓	–
VoT (Wang et al., 2026)	✓	✓	–	–	–	–	✓	–
CAMEF (Zhang et al., 2025b)	✓	–	✓	✓	–	–	–	–
EventTSF (Ge et al., 2026)	✓	–	✓	–	✓	–	–	–
Text Reinforcement (Su et al., 2025)	✓	–	–	–	–	–	✓	–
HIP / RL-HIP (ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1 summarizes the distinction more narrowly than a binary "uses feedback / does not use feedback" comparison. From News to Forecast, VoT, and Text Reinforcement all exploit forecasting outcomes or errors in their reasoning or text-processing loops (Wang et al., 2024; 2026; Su et al., 2025). HIP instead makes the event-to-forecast interface explicit through a constrained semantic–temporal state and a horizon-wise numerical profile; RL-HIP directly calibrates decisions in that structured space. The novelty claim is therefore not the generic use of forecast feedback, but the representation being calibrated and the way it is converted into horizon-aligned conditioning.

3 METHODOLOGY

Figure 1 summarizes the framework. At each forecast origin, the pathway policy converts available events and forecasting context into a structured HIP state; the temporal adapter maps that state to a horizon-wise conditioning tensor for the final forecaster. During PPO calibration, a separate frozen reward forecaster supplies forecast feedback. After calibration, the pathway policy is frozen and the evaluation forecaster is trained independently. The policy therefore makes context-conditioned structured decisions at each origin without assuming action-dependent transitions between forecast origins.

At forecast origin t, let $\mathbf{x}_{t-L+1:t}$ denote the historical target series, $\mathbf{s}_{t-L+1:t+H}$ the structured covariates used by the forecasting protocol, and $\mathcal{E}_t = \{e_i\}_{i=1}^{N_t}$ the textual events available before the origin. The target is $\mathbf{y}_{t+1:t+H}$. Rather than passing text directly to the forecaster, HIP introduces a pathway policy

$$\mathcal{Z}_t = g_\theta(\mathcal{E}_t, \mathbf{x}_{t-L+1:t}, \mathbf{s}_{t-L+1:t+H}). \tag{1}$$

The retained pathway set is transformed by an event-to-forecast adapter $A(\cdot)$ into a horizon-aligned tensor. Once calibration is complete, the final evaluation forecaster f_{ϕ_F} predicts

$$\hat{\mathbf{y}}_{t+1:t+H} = f_{\phi_F}\left(\mathbf{x}_{t-L+1:t}, \mathbf{s}_{t-L+1:t+H}, A(\mathcal{Z}_t)\right). \tag{2}$$

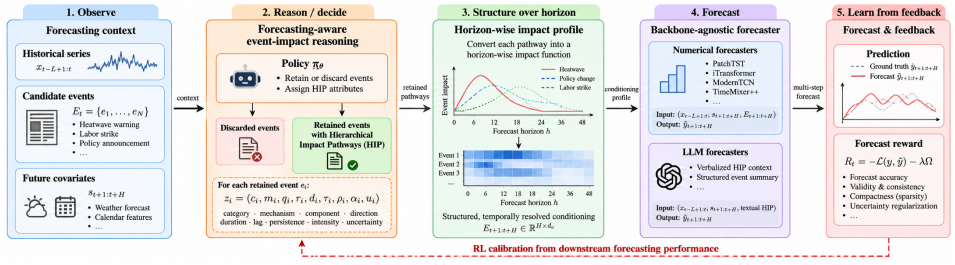

Figure 1: Overview of HIP and RL-HIP. Candidate events are transformed into structured semantic and temporal hypotheses, mapped to a horizon-wise conditioning profile, and supplied to a numerical forecasting backbone. The dashed path is used only during pathway calibration: forecasting feedback updates the pathway policy rather than the forecasting backbone. HIP attributes are forecast-oriented hypotheses, not identified causal effects.

This decomposition separates candidate external information, the structured hypothesis describing how it may matter, and the forecasting model that maps the resulting conditioning state to the future trajectory. The pathway vocabulary and implementation details are given in Appendix B.

3.1 Hierarchical Impact Pathway Representation

For each candidate event, the policy first decides whether it should be retained. A retained event is represented as

$$z_i = (c_i, m_i, q_i, r_i, d_i, \tau_i, \rho_i, \alpha_i, u_i), \tag{3}$$

where c_i is category, m_i the hypothesized mechanism, q_i the affected target component, r_i an input-side direction label, d_i duration class, τ_i lag, ρ_i persistence, α_i intensity, and u_i a policy-derived uncertainty discount. The first four attributes describe *what form of effect is hypothesized*; the remaining attributes describe *how that effect is represented over the forecast window*. This separation matters because events with similar semantic relevance can imply different onset, persistence, and conditioning strength across future steps.

The representation is hierarchical rather than a set of independent labels. The policy first predicts a coarse event category, conditions the hypothesized mechanism on that category, and then predicts the affected component and subsequent effect attributes in the context of the preceding decisions. This factorization restricts the policy to a structured state space and makes the resulting conditioning representation directly auditable. LLM-generated pseudo-pathways provide the supervised warm start, but they are treated as structured priors rather than ground truth; forecast feedback may subsequently alter both retention and attributes. Accordingly, the mechanism field is a forecasting hypothesis, not an identified causal mechanism. Post-calibration labels may also behave as semantically initialized predictive codes, so we do not claim that every rewritten label remains a faithful explanation of its source article. The full factorization and compatibility constraints are given in Appendix A.

3.2 From Pathways to Horizon-Wise Forecast Conditioning

HIP differs from a static event embedding by assigning each retained pathway a horizon-dependent profile. Let $\psi(c_i, m_i, q_i, r_i) \in \mathbb{R}^{d_e}$ denote the learned semantic embedding, $\eta(d_i) \in \mathbb{R}^{d_e}$ the learned duration-conditioned modulation, and

$$\kappa(h; \tau_i, \rho_i) = \mathbb{I}(h \geq \tau_i) \exp[-\rho_i(h - \tau_i)]$$

the onset-and-decay kernel. The event-conditioning vector at horizon h is

$$\mathbf{e}_{t+h} = \sum_{z_i \in \mathcal{Z}_t} \alpha_i \kappa(h; \tau_i, \rho_i) \sigma(r_i)(1 - u_i) \big[\eta(d_i) \odot \psi(c_i, m_i, q_i, r_i)\big], \tag{4}$$

and the complete conditioning tensor is

$$\mathbf{E}_{t+1:t+H} = [\mathbf{e}_{t+1}, \ldots, \mathbf{e}_{t+H}]^\top \in \mathbb{R}^{H \times d_e}. \tag{5}$$

Lag controls when the pathway kernel becomes active and persistence controls its exponential decay. Duration enters through the learned modulation $\eta(d_i)$ rather than imposing a hard cutoff; intensity scales the pathway magnitude, direction supplies an input-side sign, and uncertainty attenuates pathways for which the policy distribution is more diffuse. These quantities describe the conditioning signal presented to the forecaster, not guaranteed properties of the final prediction. In particular, a positive direction label does not force the final forecast upward, and a lag value does not prevent the forecasting model from mixing information across horizons.

The purpose of the adapter is therefore representational: it makes the future horizon an explicit axis of event conditioning instead of reusing one static event vector at every step. A pathway can have little influence immediately after the forecast origin, become active after a delay, and then persist or decay over later horizons. This is the main structural distinction between HIP and event interfaces that provide only a single event embedding or free-form textual context.

For the numerical forecasters, HIP is injected through a shared horizon-wise decoding interface rather than by modifying the backbone's historical input. Let $\mathbf{z}_t$ denote the representation produced by a forecasting backbone from the observed target history. At each future step h, the decoder concatenates $\mathbf{z}_t$, the future-indexed structured covariates $\mathbf{s}_{t+h}$, and the HIP signal $\mathbf{e}_{t+h}$ before mapping the resulting vector to $\hat{y}_{t+h}$. Thus, the numerical backbones differ in how they encode historical dynamics while sharing the same event-conditioning interface. Exact dimensions and decoder implementations are given in Appendix G.

3.3 Forecast-Aware Pathway Calibration

Semantically plausible pathways need not be predictively useful, so RL-HIP calibrates the discrete pathway policy with downstream forecasting feedback. During calibration, a frozen reward forecaster f_{ϕ_R}—distinct from the final evaluation forecaster f_{ϕ_F}—scores the sampled HIP state against a no-event reference using the same checkpoint, numerical history, and structured covariates. Specifically,

$$\hat{\mathbf{y}}^{\mathrm{ref}}_{t+1:t+H} = f_{\phi_R}(\mathbf{x}_{t-L+1:t}, \mathbf{s}_{t-L+1:t+H}, \mathbf{0}), \qquad \hat{\mathbf{y}}^{\mathrm{HIP}}_{t+1:t+H} = f_{\phi_R}\big(\mathbf{x}_{t-L+1:t}, \mathbf{s}_{t-L+1:t+H}, \mathbf{E}^{\mathrm{HIP}}_t\big), \tag{6}$$

where $\mathbf{E}^{\mathrm{HIP}}_t = \mathcal{A}(\mathcal{Z}_t)$ is the tensor induced by the sampled pathway set. For each forecast origin we use horizon-averaged MAE,

$$\ell_t(\hat{\mathbf{y}}) = \frac{1}{H}\sum_{h=1}^{H} |y_{t+h} - \hat{y}_{t+h}|, \tag{7}$$

and define the forecasting reward as relative loss reduction over the no-event reference,

$$R^{\mathrm{forecast}}_t = \frac{\ell_t\big(\hat{\mathbf{y}}^{\mathrm{ref}}\big) - \ell_t\big(\hat{\mathbf{y}}^{\mathrm{HIP}}\big)}{\ell_t(\hat{\mathbf{y}}^{\mathrm{ref}}) + 10^{-6}}. \tag{8}$$

Positive reward therefore means that the sampled HIP state improves the frozen reward forecaster relative to the same model with no event tensor. The implemented regularized reward is

$$R_t = R^{\mathrm{forecast}}_t + 0.30R^{\mathrm{val}}_t - 0.05R^{\mathrm{comp}}_t - 0.05R^{\mathrm{unc}}_t, \tag{9}$$

where validity measures schema compliance among retained pathways, compactness penalizes dense/high-intensity representations, and the uncertainty term penalizes large effective weight on diffuse policy decisions. The final calibration reward does not include a consistency term. Rewards are not batch-clipped or normalized; PPO advantages are standardized within the rollout batch.

Because retention and pathway attributes are discrete, we optimize the policy with PPO. PPO is used here as an optimizer for context-conditioned structured actions, not as a claim of long-horizon environment control. The learning stages are deliberately separated: the LLM supplies pseudo-pathways, supervised imitation initializes the policy, PPO calibrates that policy against the frozen reward forecaster, and the resulting policy is frozen before the final forecasting model is trained. This separation prevents the reported evaluation gains from arising through simultaneous updates of the pathway policy and final forecaster. Appendix A gives the policy factorization, uncertainty definition, reward terms, PPO objective, and full training procedure.

4 EXPERIMENTS

We first evaluate the complete RL-HIP pipeline across backbones on AULF electricity demand and daily Bitcoin. We then use a separate simplified AULF controlled configuration, together with alternative forecast geometries, to characterize forecast calibration, Static versus Temporal conditioning, and temporal-scope sensitivity. The controlled configuration shares the same AULF task geometry, chronological forecasting split, and test support as the main benchmark, but uses a separately trained calibration setup described in Appendix D.1. Full controlled tables and diagnostics are reported in Appendices D and C.

The numerical evaluation uses PatchTST (Nie et al., 2022), iTransformer (Liu et al., 2024b), ModernTCN (Luo & Wang, 2024), and TimeMixer++ (Wang et al., 2025). For each backbone we compare No events, Raw-event, LLM-HIP, and RL-HIP. No events uses only the numerical forecasting inputs; Raw-event adds retrieved article text through a static event representation; LLM-HIP uses structured pathways produced directly from the LLM annotations; and RL-HIP uses the forecast-calibrated pathway policy. These arms progressively introduce external text, structured event representation, and forecast-feedback calibration. Numerical results use five forecaster seeds; Tables 2 and 3 report the mean across seeds. We also report one representative LLM-forecasting interface per domain: Llama-2-7B-Chat for AULF and Qwen2.5-7B for Bitcoin, each under the same full-test, five-seed protocol within its domain.

Data protocol. For AULF, forecast origins are split chronologically into 3,285 training instances through October 2020, 365 validation instances from late 2020, and the full 2021 test period with 1,825 region–day origins. For Bitcoin, the numerical forecasting splits use 2018–2019 for training, January–March 2020 for validation, and April 2020–May 2021 for testing. Model checkpoints and the Shrinkage-HIP coefficient are selected using validation performance, while the frozen test sets are used for reporting. Candidate news is retrieved from the preceding seven-day window at each forecast origin. Exact split boundaries and stage-wise data access are documented in Appendix B.2.

Training/evaluation separation. LLM pseudo-pathways initialize the policy; PPO calibrates it against the frozen reward forecaster f_{ϕ_R}; the calibrated policy is then frozen before the final forecaster f_{ϕ_F} is trained and evaluated. Thus the reported seed variation reflects final-forecaster initialization under a fixed pathway policy rather than policy-training uncertainty.

4.1 MULTI-HORIZON FORECASTING PERFORMANCE

Electricity demand. AULF contains five Australian regions (NSW, QLD, VIC, SA, and TAS) and is sampled every 30 minutes. We use one forecast origin per day and region, a 48-step lookback, and a 24-step forecast horizon. Thus the main task predicts the next 12 hours rather than a full day. Table 2 reports the full test-set results. The numerical result is consistent across architectures. RL-HIP obtains the lowest mean MAE, MSE, RMSE, and MAPE for all four backbones. Relative to No events, the MAE reduction is 5.90% for PatchTST, 5.43% for iTransformer, 10.94% for ModernTCN, and 7.86% for TimeMixer++. The uncalibrated LLM-HIP representation does not improve the numerical baselines at this geometry, whereas the complete RL-HIP pipeline reduces MAE by 6.08–10.74% relative to LLM-HIP. This comparison shows that the calibrated pipeline is more effective than the direct LLM pathway interface in this setting, but it does not by itself isolate the incremental effect of PPO from imitation, retention changes, or other coupled changes in the policy. Likewise, the four-way ladder should not be read as evidence that any individual HIP attribute is independently necessary. The representative LLM forecaster shows a different behavior: No events is best, so the direct verbalized interface does not reproduce the consistency observed for numerical conditioning.

Bitcoin. Bitcoin provides a daily cross-domain evaluation with a seven-day lookback and seven-day forecast horizon. The $H = 7$ setting matches the temporal vocabulary used for the daily Bitcoin pathways, allowing lag and persistence choices to remain active within the forecast window. Table 3 reports the results. The Bitcoin result is favorable in mean ranking but substantially weaker than AULF. RL-HIP has the lowest mean MAE for iTransformer, ModernTCN, and TimeMixer++, while PatchTST remains best without events. Relative to LLM-HIP, RL-HIP lowers mean MAE for all four numerical backbones, but the absolute differences are modest and no paired uncertainty analysis is reported for these Table 3 contrasts. We therefore treat Bitcoin primarily as a cross-domain scope

Table 2: AULF electricity-demand forecasting. Numerical results are means over five forecaster seeds on the full test set with a fixed calibrated policy. The representative LLM forecaster follows the same full-test, five-seed evaluation protocol. Lower is better.

Forecaster	Method	MAE	MSE $\times 10^{-3}$	RMSE	MAPE (%)
PatchTST	No events	181.15	82.57	287.35	5.75
	Raw-event	181.29	83.51	288.97	5.78
	LLM-HIP	182.10	86.77	294.52	5.96
	RL-HIP	**170.46**	**76.65**	**276.82**	**5.57**
iTransformer	No events	179.61	81.26	285.05	5.82
	Raw-event	180.54	81.94	286.25	5.87
	LLM-HIP	180.86	84.16	290.07	5.96
	RL-HIP	**169.86**	**75.90**	**275.45**	**5.58**
ModernTCN	No events	187.92	92.96	304.86	6.28
	Raw-event	185.35	89.93	299.82	6.17
	LLM-HIP	187.50	92.41	303.79	6.43
	RL-HIP	**167.36**	**74.33**	**272.61**	**5.59**
TimeMixer++	No events	186.32	88.94	298.18	6.17
	Raw-event	186.37	88.86	297.92	6.19
	LLM-HIP	189.19	92.71	304.43	6.41
	RL-HIP	**171.68**	**77.18**	**277.70**	**5.71**
LLM forecaster	**No events**	**176.58**	**95.35**	**308.22**	**5.70**
	Raw-event	196.07	115.52	339.88	6.04
	LLM-HIP	194.16	112.30	335.11	6.43
	RL-HIP	194.30	112.94	336.05	6.19

result rather than as evidence of a uniform gain. The representative LLM forecaster also attains its lowest mean MAE under RL-HIP, subject to the same scope interpretation.

4.2 Controlled Analysis of Calibration and Temporal Structure

As a supplement to the main benchmark, we run matched controls under a separate simplified AULF calibration configuration. The control study retains the same forecasting task geometry, chronological forecasting split, and full $1{,}825$-origin test set, but uses separately trained reward-forecaster and policy checkpoints with the controlled calibration settings specified in Appendix D.1. All variants within a given controlled comparison use the same controlled configuration and paired final-forecaster seeds. We therefore interpret within-study differences rather than compare absolute MAE values directly with Table 2. Remaining comparison-specific confounds are stated explicitly in the corresponding appendix sections.

The supplementary controls show that post-imitation calibration can improve the Temporal interface on several backbones, while the value of horizon-varying conditioning itself is architecture-dependent. Static conditioning is stronger before calibration, whereas a post-calibration Temporal advantage appears only for a subset of backbones. Because the current Static construction also changes realized tensor magnitude, these comparisons characterize the tested interfaces rather than isolate temporal shape alone. Additional attenuation, retention, and forecast-geometry controls are reported in Appendices D and C.

4.3 From Structured Events to Forecast Adjustment: NSW Case Study

We use the NSW forecast for 2021-11-20 as an illustrative example of the event-to-trajectory interface. Within this visualization, No events obtains MAE 321.37 and RL-HIP obtains 155.15, a reduction of 166.22 MW (51.7%). The calibrated policy retains 14 pathways. The retained representation can be inspected at both the pathway and horizon levels. Four of the 14 retained pathways originated from source annotations with `keep=0` and were retained or relabelled after forecast cal-

Table 3: Bitcoin forecasting results. All rows report means over five seeds on the full held-out test set.

Forecaster	Method	MAE ($\times 10^3$)	MSE ($\times 10^6$)	RMSE ($\times 10^3$)	MAPE (%)
PatchTST	**No events**	**1.467**	**6.720**	**2.592**	**5.05**
	Raw-event	1.547	7.453	2.728	5.28
	LLM-HIP	1.493	6.860	2.619	5.15
	RL-HIP	1.486	6.801	2.608	5.13
iTransformer	No events	1.579	7.607	2.757	5.33
	Raw-event	1.576	7.632	2.762	5.31
	LLM-HIP	1.550	7.422	2.722	5.25
	RL-HIP	**1.539**	**7.292**	**2.699**	**5.19**
ModernTCN	No events	1.547	7.281	2.698	5.33
	Raw-event	1.549	7.339	2.709	5.35
	LLM-HIP	1.574	7.398	2.719	5.49
	RL-HIP	**1.515**	**7.054**	**2.655**	**5.21**
TimeMixer++	No events	1.641	**8.315**	**2.883**	5.52
	Raw-event	1.663	8.800	2.965	5.51
	LLM-HIP	1.658	8.496	2.910	5.61
	RL-HIP	**1.629**	8.330	2.884	**5.46**
LLM forecaster	No events	1.367	5.774	1.580	4.80
	Raw-event	1.414	6.087	1.635	5.03
	LLM-HIP	1.373	5.762	1.587	4.90
	RL-HIP	**1.353**	**5.667**	**1.567**	**4.75**

ibration, illustrating that the calibrated policy is not simply reproducing the initial LLM decisions. Appendix F reports the ranked pathways, unsigned pathway-pressure curves, aggregate pressure, and same-checkpoint pathway-deletion sensitivity. These quantities expose the conditioning state supplied to the forecaster; they should not be interpreted as signed causal contributions or independent semantic-validity scores.

4.4 Discussion and Limitations

The evidence supports a specific rather than universal scope for event-conditioned forecasting. RL-HIP is strongest on the AULF numerical forecasting task, where it achieves the lowest mean MAE across all four evaluated backbones. On Bitcoin, the gains are smaller and architecture-dependent, while direct LLM forecasting interfaces are less consistent. The controlled AULF study further shows that post-imitation calibration improves MAE on three of four backbones, whereas a clear advantage of the Temporal interface over its Static counterpart appears only for PatchTST and ModernTCN. Static conditioning is better at the imitation stage on all four backbones. Together with the shorter-horizon diagnostics, these results suggest that the utility of structured event conditioning depends on the interaction among forecast geometry, calibration, and downstream architecture rather than on temporal structure being uniformly beneficial.

Several limitations qualify the interpretation of these results. First, the supervised imitation warm start is pretrained on a global text-only pseudo-label corpus that is not chronologically restricted to the forecasting development period: it includes articles overlapping the 2021 AULF test period and later article records. This stage does not use forecasting targets, test errors, or target-bearing test examples, but the article corpus is not temporally disjoint from the forecasting test period. The reported protocol should therefore be interpreted as cross-temporal text-only pretraining rather than prospective zero-shot generalization to unseen future article distributions. Second, HIP attributes are forecast-oriented hypotheses rather than identified causal mechanisms, and we do not independently validate the semantic fidelity of pathway labels after calibration. The calibrated pathways should therefore be interpreted as auditable predictive states rather than faithful explanations of the source articles. Third, the current Static–Temporal comparison preserves the horizon-mean pathway profile but not the realized conditioning-tensor norm, so it does not cleanly isolate temporal variation from conditioning magnitude. Fourth, although Full RL improves over retention-only calibration,

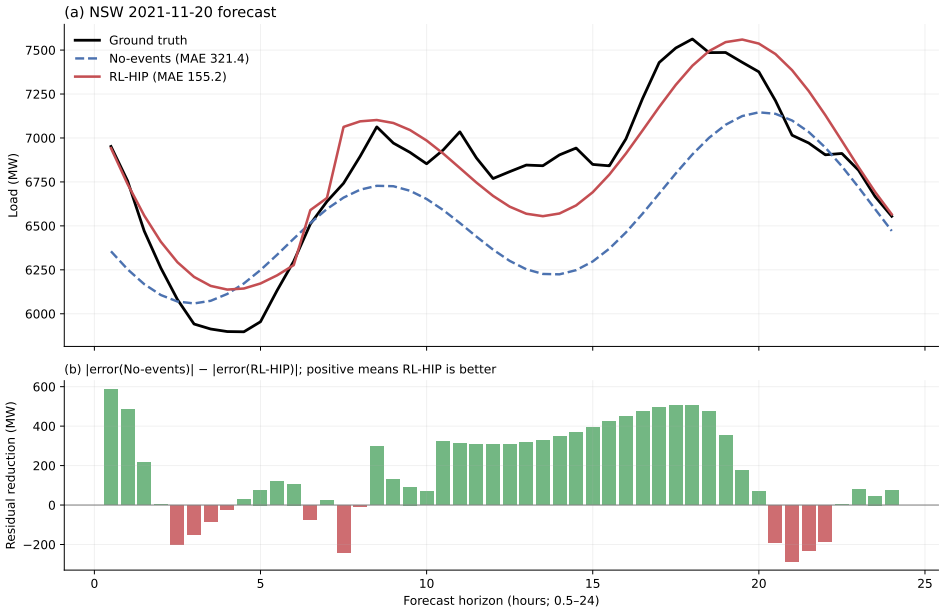

Figure 2: NSW 2021-11-20 illustrative iTransformer case. The figure compares the No-events and RL-HIP forecast trajectories with the ground truth and shows horizon-wise residual reduction. RL-HIP reduces case-level MAE from 321.37 to 155.15. The figure is used to visualize the event-conditioned representation rather than to establish an aggregate causal effect.

a validation-selected scalar shrinkage baseline is statistically indistinguishable from Full RL in the PatchTST control. Simple attenuation therefore remains a plausible explanation for a substantial portion of the calibration benefit, and the current experiments do not establish that any individual semantic or temporal rewrite is necessary. More generally, calibration jointly changes event retention, pathway attributes, and signal magnitude, making their individual contributions difficult to identify from the present ablations. The Context-only comparison likewise establishes a difference between the tested pipelines but is not perfectly sample-matched to Real-News. Finally, the reported confidence intervals quantify final-forecaster initialization variability conditional on one fixed policy realization per configuration and do not capture PPO policy-training variability. Future work should address these limitations through magnitude-matched temporal controls, broader shrinkage comparisons, repeated policy training, and independent validation of post-calibration semantic fidelity.

5 Conclusion

We introduced HIP as a structured event representation for multi-horizon time-series forecasting and RL-HIP as a forecast-feedback mechanism for calibrating that representation. The central distinction is between knowing that an event is relevant and specifying how its influence should be distributed across future prediction steps. HIP makes this intermediate representation explicit through semantic and temporal attributes, while RL-HIP aligns the resulting pathway decisions with downstream forecasting utility.

Across four numerical backbones, RL-HIP is consistently strongest on the AULF next-12-hour benchmark and favorable on three of four backbones in seven-day Bitcoin forecasting. Supplementary controlled and forecast-geometry analyses show that the benefit of calibration and horizon-varying conditioning depends on the downstream architecture and temporal scope. They also indicate that simple attenuation can explain a substantial portion of the PatchTST calibration gain. Taken together, the results support HIP as a forecast-calibrated, horizon-resolved representation layer while cautioning against universal claims about event augmentation, treating calibrated labels as identified causal effects, or assuming that complex attribute rewriting is always necessary.

AI USE STATEMENT

Generative AI tools were used in this work to support methodological discussion, provide feedback on experimental design and result interpretation, and assist with drafting and editing portions of the manuscript. AI-generated suggestions and text were reviewed, revised, and verified by the authors before inclusion. Generative AI was not used to create synthetic datasets or to generate reported experimental measurements. Any mathematical, methodological, or scientific content suggested with AI assistance was independently checked by the authors. The authors were responsible for the implementation, execution, and verification of the experiments, as well as for checking citations and scientific claims. The authors take full responsibility for the final content of the paper, including all text, claims, results, and artifacts produced with the assistance of generative AI.

REFERENCES

Caichun Cen, Hongyan Liu, Beigu Hu, Shang Cao, Ziqi Qu, Yanyan Liang, Zeling Xu, and Yuntao Zou. Hierarchical semantic extraction and heterogeneous graph neural networks for event-driven time series forecasting. *Scientific Reports*, 2026. doi: 10.1038/s41598-026-56483-9.

Peng Chen, Siyuan Wang, Shiyan Hu, Xingjian Wu, Yang Shu, Zhongwen Rao, Meng Wang, Yijie Li, Bin Yang, and Chenjuan Guo. Empowering time series analysis with large-scale multimodal pretraining. *arXiv preprint arXiv:2602.05646*, 2026.

Xinke Du, Jinfei Cao, Xiyuan Jiang, Qin Wang, Boyao Xu, Ziyang Liu, Yikun Chen, and ChunHong Yuan. Integrating event information and multi dimensional relationships for improved financial time series forecasting. *Scientific Reports*, 15:38099, 2025. doi: 10.1038/s41598-025-22926-y.

Yunfeng Ge, Ming Jin, Yiji Zhao, Hongyan Li, Bo Du, Chang Xu, and Shirui Pan. Eventtsf: Event-aware non-stationary time series forecasting. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, 2026. Accepted paper; preprint arXiv:2508.13434.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23343–23351, 2024. doi: 10.1609/aaai.v38i21.30383.

Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. Timexl: Explainable multi-modal time series prediction with llm-in-the-loop. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4520.

Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. Multi-modal forecaster: Jointly predicting time series and textual data. *arXiv preprint arXiv:2411.06735*, 2024. doi: 10.48550/arXiv.2411.06735.

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.

Chanjuan Liu, Shengzhi Wang, and Enqiang Zhu. Dp-gpt4mts: Dual-prompt large language model for textual-numerical time series forecasting. *arXiv preprint arXiv:2508.04239*, 2025a. doi: 10.48550/arXiv.2508.04239.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18780–18788, 2025b. doi: 10.1609/aaai.v39i18.34067.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-mmd: Multi-domain multimodal dataset for time series analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-2476. Datasets and Benchmarks Track.

Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18915–18923, 2025c. doi: 10.1609/aaai.v39i18.34082.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024b.

Ming-Chih Lo, Ching Chang, and Wen-Chih Peng. Text2freq: Learning series patterns from text via frequency domain. In *NeurIPS 2024 Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=Pi6sA1MSSr.

Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024.

Huu Hiep Nguyen, Dung Nguyen, Minh Hoang Nguyen, Dai Do, and Hung Le. Llm as forecasting planner: Training-free text conditioning for time-series foundation models. *arXiv preprint arXiv:2607.24892*, 2026.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Filipe Rodrigues, Ioulia Markou, and Francisco C Pereira. Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. *Information Fusion*, 49: 120–129, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2): 1–19, 2009.

Jun Seo, Hyeokjun Choe, Seohui Bae, Soyeon Park, Jinseok Yang, Dongwan Kang, and Woohyung Lim. Adaptive information routing for multi modal time series forecasting. In *NeurIPS 2024 Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=3Im0luRZHS.

Chen Su, Yuanhe Tian, Yan Song, and Yongdong Zhang. Text reinforcement for multimodal time series forecasting. *arXiv preprint arXiv:2509.00687*, 2025. doi: 10.48550/arXiv.2509.00687.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Chenjuan Guo, Bin Yang, and Yang Shu. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0TAFiyHgEl. ICLR 2026 Poster.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153, 2024.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66887–66944. PMLR, 2025.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Yueyang Yao, Jiajun Li, Xingyuan Dai, MengMeng Zhang, Xiaoyan Gong, Fei-Yue Wang, and Yisheng Lv. Context-aware probabilistic modeling with llm for multimodal time series forecasting. *arXiv preprint arXiv:2505.10774*, 2025.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Xiyuan Zhang, Boran Han, Haoyang Fang, Abdul Fatir Ansari, Shuai Zhang, Danielle C. Maddix, Cuixiong Hu, Andrew Gordon Wilson, Michael W. Mahoney, Hao Wang, Yan Liu, Huzefa Rangwala, George Karypis, and Bernie Wang. When does multimodality lead to better time series forecasting? *arXiv preprint arXiv:2506.21611*, 2025a.

Yang Zhang, Wenbo Yang, Jun Wang, Qiang Ma, and Jie Xiong. Camef: Causal-augmented multimodality event-driven financial forecasting by integrating time series patterns and salient macroeconomic announcements. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3867–3878. Association for Computing Machinery, 2025b. doi: 10.1145/3711896.3736872.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78478–78497. PMLR, 2025.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

Xin Zhou, Weiqing Wang, Shilin Qu, Zhiqiang Zhang, and Christoph Bergmeir. Unveiling the potential of text in high-dimensional time series forecasting. In *NeurIPS 2024 Workshop on Time Series in the Age of Large Models*, 2024. Also available as arXiv:2501.07048.

Xin Zhou, Weiqing Wang, Francisco J. Baldán, Wray L. Buntine, and Christoph Bergmeir. Motime: A dataset suite for multimodal time series forecasting. *arXiv preprint arXiv:2505.15072*, 2025. doi: 10.48550/arXiv.2505.15072.

Daohan Zhu, Sitong Ge, Ruofei Wang, Honggu Chen, Yubo Hou, Tao Wan, and Zengchang Qin. Beyond sentiment: Structured information extraction from financial news. *arXiv preprint arXiv:2607.28496*, 2026.

A Detailed Methodology

This appendix provides the mathematical and implementation details omitted from Section 3. The main text introduces the complete event-to-forecast pipeline, the HIP state in Eq. (3), the horizon-wise conditioning tensor in Eq. (5), and the regularized forecasting reward in Eq. (9). Here we expand the policy factorization, uncertainty calculation, auxiliary reward terms, PPO optimization, and temporal adapter without redefining the main formulation.

A.1 Hierarchical Policy Factorization

For each candidate event e_i, the policy encoder combines a frozen 768-dimensional DistilBERT title embedding, a one-layer GRU encoding of the numerical target history, and a projected mean of the structured covariates to produce a fused state $\mathbf{h}_i \in \mathbb{R}^{256}$. Candidate events are conditionally independent given their own text and the shared numerical/covariate context. The implemented event-action factorization is

$$\begin{aligned}
&\pi_\theta(b_i, c_i, m_i, q_i, r_i, d_i, \tau_i, \rho_i, \alpha_i \mid \mathbf{h}_i) \\
&= \pi^b(b_i \mid \mathbf{h}_i)\pi^c(c_i \mid \mathbf{h}_i)\pi^m(m_i \mid \mathbf{h}_i, c_i)\pi^q(q_i \mid \mathbf{h}_i, c_i, m_i) \\
&\quad \times \pi^r(r_i \mid \mathbf{h}_i, c_i, m_i, q_i)\pi^d(d_i \mid \mathbf{h}_i, c_i, m_i, q_i, r_i) \\
&\quad \times \pi^\tau(\tau_i \mid \mathbf{h}_i, c_i, m_i, q_i, r_i, d_i)\pi^\rho(\rho_i \mid \mathbf{h}_i, c_i, m_i, q_i, r_i, d_i, \tau_i) \\
&\quad \times \pi^\alpha(\alpha_i \mid \mathbf{h}_i, c_i, m_i, q_i, r_i, d_i, \tau_i, \rho_i).
\end{aligned} \tag{10}$$

The attribute heads use a 32-dimensional running context formed by summing embeddings of previously sampled attributes; retention is not inserted into this autoregressive attribute context. Actions are sampled event by event and their log probabilities are summed to form an origin-level joint policy probability. The final policy does not mask incompatible labels at the logits. Instead, sampled retained pathways are scored by the validity component of the reward during PPO; deployed greedy outputs are subsequently tensorized according to the schema.

A.2 Policy-Derived Uncertainty

Uncertainty is not an additional free categorical action in the default implementation. Instead, it is derived from the entropy of the policy distributions. Let $\mathcal{A}_z = \{c, m, q, r, d, \tau, \rho, \alpha\}$ denote the sampled pathway attributes. We compute

$$u_i = \frac{1}{|\mathcal{A}_z|} \sum_{a \in \mathcal{A}_z} \frac{\mathcal{H}(\pi_\theta^a(\cdot \mid \mathbf{h}_i))}{\log |\mathcal{A}^a|}, \tag{11}$$

where $\mathcal{A}^a$ is the discrete vocabulary for attribute a. Hence $u_i \in [0, 1]$ is high when the policy is diffuse across plausible labels. This quantity is a policy-entropy measure, not a calibrated probability that the pathway hypothesis is incorrect. In Eq. (4), the factor $(1 - u_i)$ discounts uncertain pathways before they are aggregated into the event-impact tensor.

A.3 Auxiliary Reward Components

The implemented reward in Eq. (9) contains validity, compactness, and uncertainty terms. For one origin, validity is the fraction of retained pathways satisfying the vocabulary and category–mechanism–component / category–duration compatibility rules; an empty retained set receives validity zero. Compactness is

$$R_t^{\text{comp}} = \frac{|\mathcal{Z}_t|}{N_t + 10^{-6}} + \frac{1}{|\mathcal{Z}_t| + 10^{-6}} \sum_{z_i \in \mathcal{Z}_t} \bar{\alpha}_i, \tag{12}$$

where N_t is the number of candidates and $\bar{\alpha}_i$ is the numeric intensity. The uncertainty term is

$$R_t^{\text{unc}} = \frac{1}{|\mathcal{Z}_t| + 10^{-6}} \sum_{z_i \in \mathcal{Z}_t} u_i \bar{\alpha}_i. \tag{13}$$

The final reward weights are $+0.30$ for validity and -0.05 each for compactness and uncertainty. A consistency/path-distance term is not active in the final calibration configuration. These terms regularize the structured conditioning state; schema validity is not evidence of semantic correctness.

A.4 PPO Optimization

Pathway retention and attribute assignments are discrete decisions, so the policy is calibrated with PPO. Let $r_t(\theta)$ denote the ratio between the updated and behavior-policy probabilities of the sampled structured action, and let A_t denote the estimated advantage under the reward in Eq. (9). The clipped objective is

$$J^{\mathrm{PPO}}(\theta) = \mathbb{E}\left[\min\left(r_t(\theta)A_t, \mathrm{clip}\left(r_t(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}\right)A_t\right)\right]. \tag{14}$$

During policy calibration, the reward forecaster f_{ϕ_R} is kept fixed so that reward changes reflect pathway decisions rather than simultaneous changes in the numerical predictor. One PPO optimization sample is one forecast origin containing the joint action over its candidate events. The value baseline uses summary statistics of the target history and mean structured covariates; the one-step advantage is $A_t = R_t - V(s_t)$ and is standardized within each rollout batch. PPO uses clipping $\epsilon_{\mathrm{clip}} = 0.2$, two epochs per update, policy learning rate 3×10^{-5}, AdamW weight decay 10^{-5}, and 50 updates. There are no action-dependent state transitions across forecast origins and no GAE. The calibrated policy is then frozen before constructing the pathway representations used to train and evaluate the final forecaster f_{ϕ_F}. We keep these two forecaster roles separate in notation because the first supplies the policy-learning signal, whereas the second measures the performance of the frozen representation.

A.5 Temporal Adapter

The main text defines the horizon-wise conditioning signal in Eq. (4). The delayed-decay kernel used by the current adapter is

$$\kappa(h; \tau_i, \rho_i) = \mathbb{I}(h \geq \tau_i)\exp[-\rho_i(h-\tau_i)], \tag{15}$$

with h and τ_i expressed in forecast-step units. The semantic vector is $\psi(c, m, q, r) = \mathrm{Emb}_c(c) + \mathrm{Emb}_m(m) + \mathrm{Emb}_q(q) + \mathrm{Emb}_r(r)$ with $d_e = 16$. Duration selects a trainable 16-dimensional modulation vector $\eta(d)$, initialized to ones, which is multiplied elementwise with ψ before aggregation; it does not impose a hard temporal cutoff. Direction maps to $+1$ for positive, -1 for negative, and 0 for mixed/neutral; intensity maps low/medium/high to $0.34/0.67/1.00$. For AULF, one hour corresponds to two forecast steps and persistence classes are converted to decay rates by $\rho = \log(2)/\text{half-life}_{\text{steps}}$. Thus, for the main $H = 24$ task, lags from 0h through 11h activate within the forecast window, while lags whose converted onset is at least 24 steps contribute zero throughout indices $0, \ldots, 23$. Unlike Raw-event conditioning, which broadcasts one pooled event vector to every horizon position, HIP can generate a different conditioning vector for each future step.

A.6 Training and Inference Procedure

The complete pipeline contains three stages. First, a local Llama-2-7B-Chat annotator produces text-only pseudo-pathways that provide weak structured supervision for the policy; numerical forecasting targets are not provided to the annotator. The main-policy warm start uses a global article-level pseudo-label corpus whose article timestamps are not restricted to the forecasting development period and include records overlapping the 2021 AULF test period as well as later records. We therefore treat supervised imitation as cross-temporal text-only pretraining rather than as a strictly chronological forecasting split. Second, the pathway policy is warm-started from these pseudo-labels and calibrated with forecasting feedback using PPO while a reward forecaster f_{ϕ_R} is fixed. PPO uses AULF training origins only and saves the final update-50 policy. Third, the calibrated policy is frozen, pathway tensors are generated for the forecasting data, and the final numerical forecasters f_{ϕ_F} are trained on the training origins with validation-MAE checkpoint selection before evaluation on the held-out test origins. Each numerical backbone uses a separately trained reward forecaster and calibrated policy because the forecasting reward depends on the downstream architecture and forecast geometry.

At forecast construction time, event candidates are retrieved from the preceding seven calendar days and are restricted by publication timestamp to occur no later than the forecast origin. Forecasting targets used to fit f_{ϕ_R} or construct PPO rewards come only from the development pipeline; the final test targets are used only for reporting. The chronological restriction applies to target-bearing forecasting stages and per-origin event retrieval; it does not make the preceding text-only imitation corpus temporally disjoint, as described above. Exact split dates and the stage-by-stage data flow are given in Appendix B.2.

B Pathway Construction, Vocabulary, and Data Statistics

This appendix describes the structured annotation space used by HIP and summarizes the pathway data used in the electricity experiments. These statistics characterize the representation and its calibration behavior; they are not semantic-accuracy labels for individual events.

B.1 Pseudo-Pathway Construction

Ground-truth event-impact pathways are not available in the electricity data. We therefore use an LLM annotator to convert each candidate article into an initial retain/discard decision and, for retained events, a structured pathway compatible with Eq. (3). The pseudo-label pipeline can be summarized as

$$e_i \longrightarrow \text{LLM pseudo-pathway} \longrightarrow \text{schema validation} \longrightarrow \text{policy warm start} \longrightarrow \text{RL calibration}. \tag{16}$$

The LLM output is therefore a weakly supervised hypothesis about forecast relevance and event impact. Schema validation checks whether the produced labels belong to the allowed vocabulary and whether the combination satisfies the implemented compatibility rules. RL subsequently changes both retention decisions and pathway attributes according to forecasting reward. We do not treat either the initial or calibrated pathway as an independently verified causal annotation.

B.2 Stage-by-Stage Data Flow

The forecasting splits are chronological and fixed. For AULF, training origins span 2019-01-02–2020-10-19, validation origins span 2020-10-20–2020-12-31, and the test set is the full 2021 calendar year. With five regions and one origin per region per day, this yields 3,285/365/1,825 train/validation/test origin-region instances. The main geometry is $L = 48, H = 24$. Bitcoin uses daily $L = 7, H = 7$ forecasting with target periods 2018-01-01–2019-12-31 for training, 2020-01-01–2020-03-31 for validation, and 2020-04-01–2021-05-20 for testing, yielding 724/85/409 origins after requiring each complete seven-day target window to remain inside its split.

Table 4: Auditable stage-by-stage data flow for the reported HIP/RL-HIP pipeline.

Stage	Data / split	Forecast-target access	Selection rule	Output
LLM pseudo-label generation	Global article corpus, including 2021 test-period and later article records; title, publication time, optional region, and summary/full text	None	Schema parsing/validation; retry on invalid output	Pseudo-label JSONL
Supervised imitation	Grouped-title 70/15/15 split of the global text/pseudo-label corpus; not chronologically tied to forecasting splits	None	Best epoch by keep-F1 + mean attribute accuracy + sampled validity	Imitation-HIP warm policy
Reward forecaster	AULF train origins; AULF validation for checkpoint selection	Training targets for fitting	Minimum validation MAE	Frozen f_{ϕ_R}
PPO calibration	AULF training origins only; frozen f_{ϕ_R}	Training targets enter through R_t^{forecast}	No validation checkpoint search; final update 50	Frozen RL-HIP policy
Shrinkage selection	AULF validation origins; $\lambda \in \{0, 0.05, \ldots, 1\}$	Validation targets	Minimum mean validation MAE; smaller λ breaks ties	Selected λ
Final forecaster	AULF/Bitcoin train origins with validation checkpointing	Training targets for fitting; validation targets for selection	Minimum validation MAE across 50 epochs	Final f_{ϕ_F} per arm/seed
Final test scoring	Frozen policy and trained f_{ϕ_F} on held-out test origins	Test targets only for metric computation	None	Reported MAE/MSE/RMSE/MAPE

The first two rows describe text-only weak-supervision stages rather than target-bearing forecasting stages. Because their article corpus extends across and beyond the 2021 test period, the imitation warm start is cross-temporal; forecasting targets remain unavailable to both stages. For AULF, the first lookback window of a split may draw historical values from the preceding period, but the forecast origin and target horizon belong to the assigned split. For Bitcoin, an origin is retained only when its complete seven-day target window remains within the same split; adjacent origins therefore have overlapping target windows only inside a split. The reward forecaster and final forecaster are separately trained checkpoints, and the policy is frozen before final-forecaster fitting.

Table 5: Electricity-domain HIP vocabulary, separated into semantic attributes and temporal/control attributes.

(a) Semantic attributes

Attribute	Definition	Vocabulary / value range
Category c_i	Coarse semantic type of the event	`infrastructure_failure`, `economic_activity`, `public_health`, `extreme_weather`, `social_activity`, `policy_holiday`, `sports_culture`
Mechanism m_i	Hypothesized pathway through which the event affects demand	`residential_cooling_demand_increase`, `residential_heating_demand_increase`, `commercial_activity_reduction`, `commercial_activity_increase`, `industrial_activity_change`, `grid_disruption`, `emergency_response_demand_increase`, `behavioral_shift_indoor`, `evacuation_demand_drop`
Component q_i	Affected part of the electricity-demand system	`residential_load`, `commercial_load`, `industrial_load`, `public_service_load`, `total_system_load`
Direction r_i	Expected input-side sign used by the conditioning representation; it does not guarantee the sign of the final forecast change	`positive`, `negative`, `mixed`, `neutral`

(b) Temporal and control attributes

Attribute	Definition	Vocabulary / value range
Duration d_i	Coarse duration class used by the learned duration modulation; it does not by itself impose a hard cut-off	`real_time`, `short_term`, `medium_term`, `long_term`, `recurring`
Lag τ_i	Delay between forecast origin and onset of the pathway conditioning kernel	$\{0\text{h}, 1\text{h}, \ldots, 24\text{h}, 1\text{–}7\text{d}, > 7\text{d}\}$
Persistence ρ_i	Persistence class mapped to the decay behavior of the conditioning kernel	$\{1\text{h}, 2\text{h}, \ldots, 24\text{h}, 1\text{–}3\text{d}, 3\text{–}7\text{d}, > 7\text{d}\}$
Intensity α_i	Magnitude assigned before tensor aggregation	`low`, `medium`, `high`
Uncertainty u_i	Policy-entropy discount applied to the pathway contribution; not a calibrated probability of semantic correctness	$u_i \in [0, 1]$, derived from normalized policy entropy in the default implementation

B.3 Electricity-Domain Pathway Vocabulary

Table 5 lists the pathway attributes and the electricity-domain vocabulary used in the current manuscript. The schema is shared across forecast instances, whereas the semantic vocabulary is domain-specific. The fine lag and persistence bins correspond to the final experiments.

The hierarchy should not be interpreted as an identified causal graph. It is a constrained forecasting representation: category narrows plausible mechanisms, mechanism narrows plausible affected components, and the remaining attributes specify signed temporal behavior. The mechanism field is therefore a model hypothesis about how an event may affect demand, not a verified causal mechanism.

B.4 Pseudo-Pathway Dataset Distribution

The canonical electricity annotation set contains 222,365 news rows. The LLM-based pseudo-pathway construction retains 27,569 rows, corresponding to a 12.4% keep rate (Table 6).

The pseudo-pathway distribution is strongly imbalanced toward public-health events and positive short- or medium-term effects. This partly reflects the 2020–2021 news environment. These counts describe the weak-label corpus rather than semantic annotation accuracy and should not be assumed to generalize unchanged to other domains or time periods.

Table 6: Pseudo-pathway corpus statistics. Panel (a) summarizes retention, panel (b) reports the category distribution, and panel (c) reports direction and duration distributions among retained pseudo-pathways.

(a) Retention statistics

Quantity	Count	Share
Total news rows	222,365	100.0%
Retained pseudo-pathways	27,569	12.4%
Discarded rows	194,796	87.6%

(b) Category distribution

Category	Count	Share	Category	Count	Share
public_health	11,915	43.2%	policy_holiday	3,265	11.8%
extreme_weather	4,932	17.9%	economic_activity	184	0.7%
infrastructure_failure	3,842	13.9%	sports_culture	70	0.3%
social_activity	3,361	12.2%			

(c) Direction and duration

Attribute	Value	Count	Share	Attribute	Value	Count	Share
Direction	positive	20,697	75.1%	Duration	short_term	15,606	56.6%
Direction	negative	6,871	24.9%	Duration	medium_term	6,711	24.3%
Direction	mixed	1	0.0%	Duration	real_time	5,251	19.0%
				Duration	long_term	1	0.0%

B.5 Auxiliary Representation Statistics

Table 17 in Appendix E reports representation statistics from the additional $48 \to 48$ full-day robustness diagnostic. Within that diagnostic, most of the reduction in retained count and pathway-distance statistics relative to direct LLM-HIP already appears after supervised imitation: average retained pathways change from 16.00 (LLM-HIP) to 9.98 (imitation) and 10.16 (RL-HIP), while D_{path} changes from 1.1382 to 0.6649 and 0.6512. RL-HIP further lowers mean intensity from 0.8111 after imitation to 0.6636. These quantities are used only to characterize the full-day robustness diagnostic and are not combined numerically with the simplified $H = 24$ controlled study.

B.6 Forecast-Aware Calibration Examples

We inspect two event-heavy instances to determine whether RL merely filters LLM pathways or also rewrites the structured state. Table 7 combines the retention and attribute-rewrite diagnostics for the same two forecast instances. The reported calibration artifacts contain 32 candidate events in each case.

These cases show that forecast-aware calibration acts on both event retention and the structured pathway state. They should not be interpreted as semantic validation of the rewritten labels. Some changes may reduce forecasting loss while remaining difficult to justify from the article alone; stronger expert or rule-based semantic constraints are therefore a natural extension of the current framework.

C Sensitivity to Forecast Geometry

The main paper uses AULF $48 \to 24$ and Bitcoin $7 \to 7$ as representative multi-horizon settings. The experiments below examine how the event-conditioning result changes under alternative temporal geometries. They are not intended to identify an optimal horizon. In particular, some comparisons change both lookback and horizon, and the temporal vocabularies are domain-specific; we therefore interpret the results as evidence of forecast-geometry sensitivity rather than a causal effect of horizon length alone.

Table 7: Forecast-aware calibration behavior for two event-heavy electricity forecasts. Panel (a) summarizes retention; panel (b) reports rewrites among events retained by both LLM-HIP and RL-HIP. Percentages in panel (b) use the jointly retained set as the denominator.

(a) Retention behavior

Statistic	VIC 2021-02-13	NSW 2021-03-22
Candidate events	32	32
LLM-kept pathways	19	10
RL-kept pathways	17	12
Keep agreement	30/32 (93.8%)	28/32 (87.5%)
Kept by both	17	9
RL-only	0	3
LLM-only	2	1

(b) Attribute rewrites among jointly retained events

Attribute	VIC: changed / 17	NSW: changed / 9
Category	8/17 (47.1%)	2/9 (22.2%)
Mechanism	5/17 (29.4%)	2/9 (22.2%)
Component	7/17 (41.2%)	4/9 (44.4%)
Direction	9/17 (52.9%)	2/9 (22.2%)
Duration	11/17 (64.7%)	2/9 (22.2%)
Lag	13/17 (76.5%)	2/9 (22.2%)
Persistence	4/17 (23.5%)	6/9 (66.7%)
Intensity	11/17 (64.7%)	4/9 (44.4%)

Table 8: AULF MAE under two fixed-origin multi-horizon settings. The $48 \to 48$ task forecasts the next 24 hours, whereas $48 \to 24$ forecasts the next 12 hours. Lower is better within each geometry; absolute MAE values should not be compared as if the two tasks had identical targets.

	$48 \to 48$				$48 \to 24$			
Backbone	No events	Raw-event	LLM-HIP	RL-HIP	No events	Raw-event	LLM-HIP	RL-HIP
PatchTST	243.28	238.35	230.10	**220.80**	181.15	181.29	182.10	**170.46**
iTransformer	244.00	240.32	224.73	**217.47**	179.61	180.54	180.86	**169.86**
ModernTCN	238.61	237.34	**220.31**	220.64	187.92	185.35	187.50	**167.36**
TimeMixer++	239.98	236.76	219.75	**214.83**	186.32	186.37	189.19	**171.68**

C.1 AULF Fixed-Origin Horizon Comparison

A HIP arm is best for all four numerical backbones at both horizons. RL-HIP is best for three of four backbones at $H = 48$ and all four at $H = 24$. The mean relative MAE reduction of RL-HIP over No events is 9.53% at $H = 48$ and 7.53% at $H = 24$. The difference in effect size indicates sensitivity to geometry, but the two settings do not support a simple monotonic relationship between horizon length and RL benefit.

C.2 Immediate One-Step Forecasting

We next predict only the next half-hour observation using the previous 48 observations. This setting uses an $H = 1$-specific calibration and timestamp-aware event retrieval. No events has the lowest MAE for PatchTST and iTransformer, whereas Raw-event has the lowest MAE for ModernTCN and TimeMixer++. Structured HIP/RL-HIP therefore does not provide a consistent one-step advantage.

Table 9: Immediate one-step AULF forecasting ($48 \to 1$) across four numerical backbones. Displayed MSE is raw MSE divided by 10^3. Values are mean $\pm$ sample standard deviation; lower is better.

Backbone	Method	MAE	MSE $\times 10^{-3}$	RMSE	MAPE (%)
PatchTST	**No events**	$\mathbf{53.15 \pm 0.41}$	$\mathbf{6.870 \pm 0.058}$	$\mathbf{82.88 \pm 0.35}$	$\mathbf{1.86 \pm 0.02}$
	Raw-event	53.91 ± 1.43	7.019 ± 0.298	83.77 ± 1.77	1.92 ± 0.05
	LLM-HIP	53.71 ± 0.29	7.057 ± 0.097	84.00 ± 0.58	1.90 ± 0.03
	RL-HIP	53.87 ± 0.50	7.031 ± 0.099	83.85 ± 0.59	1.88 ± 0.01
iTransformer	**No events**	$\mathbf{50.01 \pm 0.09}$	$\mathbf{6.226 \pm 0.007}$	$\mathbf{78.90 \pm 0.04}$	1.86 ± 0.02
	Raw-event	50.22 ± 0.19	6.288 ± 0.083	79.29 ± 0.52	1.85 ± 0.01
	LLM-HIP	50.81 ± 0.58	6.338 ± 0.192	79.61 ± 1.20	1.83 ± 0.04
	RL-HIP	50.10 ± 0.15	6.260 ± 0.077	79.12 ± 0.48	$\mathbf{1.82 \pm 0.02}$
ModernTCN	No events	51.50 ± 0.57	6.366 ± 0.087	79.79 ± 0.55	1.87 ± 0.05
	Raw-event	$\mathbf{51.05 \pm 0.47}$	$\mathbf{6.290 \pm 0.112}$	$\mathbf{79.31 \pm 0.71}$	$\mathbf{1.83 \pm 0.05}$
	LLM-HIP	51.41 ± 0.31	6.349 ± 0.097	79.68 ± 0.61	1.84 ± 0.02
	RL-HIP	51.32 ± 0.43	6.330 ± 0.140	79.56 ± 0.88	1.86 ± 0.03
TimeMixer++	No events	62.46 ± 8.59	9.138 ± 2.497	95.03 ± 12.68	2.49 ± 0.08
	Raw-event	$\mathbf{58.15 \pm 3.80}$	8.131 ± 1.075	90.04 ± 5.89	$\mathbf{2.16 \pm 0.21}$
	LLM-HIP	64.92 ± 8.97	9.783 ± 2.505	98.33 ± 13.04	2.57 ± 0.85
	RL-HIP	58.63 ± 1.41	$\mathbf{7.971 \pm 0.295}$	$\mathbf{89.27 \pm 1.66}$	2.23 ± 0.14

C.3 AULF Lookback–Horizon Grid

To separate the primary half-day result from a single choice of lookback, we also evaluate six PatchTST geometries on a common ordered set of forecast origins. One day corresponds to 48 half-hour steps.

Table 10: PatchTST lookback–horizon analysis on AULF. Values are mean $\pm$ sample standard deviation over seeds 0–4. Lower MAE is better.

Lookback $\to$ horizon	No events	Raw-event	LLM-HIP	RL-HIP
$7 \to 1$ days	233.86 ± 1.98	238.38 ± 4.85	222.14 ± 2.52	$\mathbf{214.29 \pm 3.27}$
$5 \to 1$ days	236.52 ± 1.94	239.44 ± 4.52	225.07 ± 3.83	$\mathbf{214.70 \pm 2.35}$
$3 \to 1$ days	238.05 ± 4.76	236.40 ± 3.37	219.81 ± 1.56	$\mathbf{218.75 \pm 3.40}$
$2 \to 1$ days	235.96 ± 2.18	234.23 ± 3.35	223.23 ± 1.61	$\mathbf{222.65 \pm 3.44}$
$1 \to 1$ days	244.15 ± 1.59	238.97 ± 1.78	228.94 ± 4.61	$\mathbf{223.30 \pm 0.71}$
$1 \to 0.5$ days	181.51 ± 1.11	180.32 ± 1.03	183.62 ± 3.15	$\mathbf{171.13 \pm 1.44}$

RL-HIP has the lowest mean MAE in all six geometries of this controlled grid. The grid shows that the method is not tied to one lookback length, but it is not an optimization study over H: most rows

share the same one-day horizon, and the half-day row changes the target support. The main-paper $48 \to 24$ values differ slightly because the common-origin grid removes early origins that cannot support the longest lookback.

C.4 Bitcoin Daily Geometry

Table 11: Bitcoin MAE under $7 \to 7$ and $5 \to 3$ daily forecasting geometries. Numerical rows are shown to isolate forecast-geometry sensitivity across the four backbone architectures. RL gain denotes the relative MAE reduction of RL-HIP over No events within the same geometry.

	$7 \to 7$					$5 \to 3$				
Backbone	No	Raw	LLM-HIP	RL-HIP	RL gain	No	Raw	LLM-HIP	RL-HIP	RL gain
PatchTST	**1.467**	1.547	1.493	1.486	-1.30%	**0.982**	1.001	0.994	0.991	-0.92%
iTransformer	1.579	1.576	1.550	**1.539**	+2.53%	1.007	1.006	**1.001**	1.036	-2.88%
ModernTCN	1.547	1.549	1.574	**1.515**	+2.07%	**0.992**	0.997	0.993	1.003	-1.11%
TimeMixer++	1.641	1.663	1.658	**1.629**	+0.73%	1.068	1.105	1.068	**1.065**	+0.28%

The average RL gain across the four numerical backbones changes from approximately +1.01% at $7 \to 7$ to -1.16% at $5 \to 3$. The daily Bitcoin pathway vocabulary was designed around a seven-day horizon and includes lag actions extending to seven days and beyond. Under an origin-relative three-day adapter, only the earliest lag choices can activate within the forecast window; later choices are effectively inactive. This mismatch is a plausible explanation for the reduced leverage of calibration at $H = 3$, but it does not isolate horizon length as the cause because the lookback changes from seven to five days at the same time.

C.5 AUD/USD as an Additional Scope Check

Table 12: AUD/USD MAE under two daily forecasting geometries. MAE is shown in units of 10^{-3}. The absence of a consistent RL-HIP advantage under either geometry provides a negative cross-domain scope check.

	$7 \to 7$				$5 \to 3$			
Backbone	No events	Raw-event	LLM-HIP	RL-HIP	No events	Raw-event	LLM-HIP	RL-HIP
PatchTST	**5.639**	5.641	6.089	5.980	**3.893**	3.921	3.995	4.048
iTransformer	5.664	5.662	**5.594**	5.694	3.978	**3.906**	4.066	3.998
ModernTCN	5.615	**5.463**	5.540	5.522	3.960	**3.903**	3.933	3.932
TimeMixer++	6.687	7.360	**6.449**	6.883	4.259	**4.098**	4.505	4.382

AUD/USD does not show a consistent RL-HIP advantage under either daily geometry. We therefore use it as a scope check rather than primary evidence. Together with the Bitcoin comparison, this result cautions against interpreting temporal span alone as sufficient for event conditioning to help: the exogenous information must also contain incremental predictive signal for the target.

D Additional Controlled Ablations

This appendix reports matched controls that test whether the AULF gain can be explained by the controller ignoring news content, by event selection alone, or by simple global attenuation. The component-isolation experiments use a separate simplified controlled AULF calibration configuration. Unless explicitly stated otherwise, they retain the same $L = 48, H = 24$ forecasting geometry, chronological forecasting split, and full $1{,}825$-origin test set as the main benchmark, while using separately trained reward-forecaster and policy checkpoints for systematic matched comparisons. All variants within each controlled experiment use the same controlled configuration and paired final-forecaster seeds. We therefore interpret within-study differences rather than compare their absolute error values directly with Table 2. Remaining comparison-specific confounds are stated explicitly below. Five paired final-forecaster seeds are used where paired uncertainty is reported.

Table 13: Controlled AULF analysis of supervised imitation versus RL calibration and Static versus Temporal conditioning under the simplified $L = 48, H = 24$ controlled configuration. Static and Temporal variants use the same policy-generated pathways and separately trained matched forecasters. $\Delta_{\mathrm{S-T}} = \mathrm{MAE}(\mathrm{Static}) - \mathrm{MAE}(\mathrm{Temporal})$, so positive values favor Temporal. Results use five paired forecaster seeds with a fixed policy; intervals are paired 95% CIs. Absolute MAE values are intended for within-study comparison and are not directly comparable with Table 2.

Backbone	Policy stage	Static MAE	Temporal MAE	$\Delta_{\mathrm{S-T}}$	Paired 95% CI
PatchTST	Imitation	$\mathbf{181.391 \pm 2.525}$	184.960 ± 3.408	-3.569	$[-5.422, -1.715]$
	RL	182.136 ± 1.461	$\mathbf{175.719 \pm 1.212}$	$+6.418$	$[+3.897, +8.938]$
iTransformer	Imitation	$\mathbf{176.695 \pm 3.067}$	181.217 ± 1.484	-4.522	$[-8.519, -0.525]$
	RL	179.362 ± 1.080	$\mathbf{177.588 \pm 2.577}$	$+1.773$	$[-1.350, +4.896]$
ModernTCN	Imitation	$\mathbf{183.593 \pm 3.632}$	190.875 ± 3.781	-7.282	$[-14.167, -0.396]$
	RL	186.261 ± 2.282	$\mathbf{179.960 \pm 2.737}$	$+6.300$	$[+2.246, +10.354]$
TimeMixer++	Imitation	$\mathbf{180.540 \pm 5.257}$	188.384 ± 2.152	-7.844	$[-12.519, -3.170]$
	RL	$\mathbf{188.376 \pm 6.204}$	188.623 ± 5.007	-0.247	$[-2.244, +1.751]$

D.1 Four-Backbone Calibration and Temporal-Profile Details

This study is designed as a matched component analysis rather than a reproduction of the final benchmark configuration. For each backbone, we vary the policy stage (Imitation-HIP versus RL-HIP) and the conditioning interface (Static versus Temporal), while keeping the remaining controlled setup fixed and training five paired final-forecaster seeds for every arm.

The main benchmark and the controlled study share the same $L = 48, H = 24$ geometry, chronological split, AULF data, pseudo-label corpus, warm imitation checkpoint, backbone families, final-forecaster training procedure, and test support. They differ in the calibration configuration used to obtain the frozen pathway policy. The final benchmark trains its reward forecaster from direct LLM-HIP-aligned pathways with a pathway cap of 16 and uses PPO rollout batches of 512, whereas the simplified controlled configuration uses Imitation-HIP-generated alignments with a pathway cap of 32 and PPO rollout batches of 64. The reward forecasters and calibrated policies are trained separately for the two configurations. Consequently, absolute RL-HIP MAE values from this controlled study are not expected to reproduce those in Table 2; the relevant quantities are the matched differences among variants within the controlled study.

The control table uses the same fixed-policy, five-forecaster-seed protocol for all four numerical backbones. Using the Temporal interface in both stages, the paired Imitation$-$RL MAE differences are $+9.241$ MW for PatchTST (95% CI $[+4.080, +14.402]$), $+3.628$ for iTransformer ($[+0.499, +6.757]$), $+10.915$ for ModernTCN ($[+6.071, +15.758]$), and -0.239 for TimeMixer++ ($[-4.511, +4.033]$). The first three therefore support an incremental post-imitation calibration benefit; TimeMixer++ does not. Its RL policy also retains only 0.937 pathways per origin and covers 35.12% of origins, compared with 9.965 pathways and 99.95% coverage under imitation. This co-occurs with the null MAE result but is not treated as a causal explanation.

For Static versus Temporal conditioning, Static is better at the imitation stage on all four backbones. After RL, Static$-$Temporal MAE is $+6.418$ MW for PatchTST ($[+3.897, +8.938]$), $+1.773$ for iTransformer ($[-1.350, +4.896]$), $+6.300$ for ModernTCN ($[+2.246, +10.354]$), and -0.247 for TimeMixer++ ($[-2.244, +1.751]$). Thus the post-RL Temporal advantage is clear for PatchTST and ModernTCN, unresolved for iTransformer, and absent for TimeMixer++. Mean-preserving Static profiles are mechanically horizon-constant, but their realized tensor norms are substantially smaller than Temporal: under RL they are $7.068/20.686$ (Static/Temporal) for PatchTST, $8.872/24.808$ for iTransformer, $10.243/30.495$ for ModernTCN, and $1.659/5.167$ for TimeMixer++. The experiment therefore establishes the behavior of the horizon-mean control but does not fully separate temporal shape from realized conditioning magnitude.

D.2 Context-Only Controller versus Real-News RL-HIP

The Context-only controller keeps the same candidate slots and masks but replaces each article representation with a fixed zero vector during imitation, PPO, and deployment; numerical history and structured covariates remain available. Table 14 compares it with the Real-News RL-HIP arm under the same simplified controlled configuration.

Table 14: PatchTST Context-only control under the simplified AULF $48 \rightarrow 24$ controlled configuration. Positive paired differences correspond to lower error for Real-News.

Variant	MAE	MSE $\times 10^{-3}$	RMSE	MAPE (%)
Context-only, count-preserving	183.732 ± 1.281	86.126 ± 1.978	293.456 ± 3.378	$\mathbf{5.812 \pm 0.074}$
Real-News RL-HIP	$\mathbf{175.719 \pm 1.212}$	$\mathbf{79.864 \pm 1.047}$	$\mathbf{282.597 \pm 1.855}$	5.855 ± 0.053

For MAE, Context-only minus Real-News is $+8.014$ MW with paired 95% CI $[+5.592, +10.435]$; MSE and RMSE also favor Real-News, while the MAPE interval includes zero. The learned behaviors differ substantially: Context-only retains on average 0.298 pathways per origin and covers only 0.93% of origins, compared with 4.045 retained pathways and 88.22% coverage for Real-News RL-HIP. This provides evidence that removing article content materially changes the learned forecasting pipeline.

The comparison is not a perfectly isolated text-content effect. The Real-News arm preserves its canonical article-level, text-only imitation warmup, whereas Context-only must construct forecast-origin/candidate pairs to expose numerical history and covariates after text removal. This changes supervised sample weighting and split construction in addition to text availability. A stronger future control would train Real-News on the same origin/candidate imitation dataset.

D.3 Global Attenuation and Retention-Only Calibration

We next test two simpler explanations for the RL-HIP gain. *Shrinkage-HIP* multiplies the Imitation-HIP conditioning tensor by one global scalar λ selected using validation MAE only from the grid $\{0, 0.05, \ldots, 1.0\}$. Validation selects $\lambda = 0.15$, after which a new final forecaster is trained and evaluated. *Retention-only RL* starts from the same imitation checkpoint as Full RL but updates only the keep/discard head during PPO; all semantic and temporal attribute heads remain frozen.

Table 15: PatchTST controls for signal attenuation and event retention under the same simplified AULF $48 \rightarrow 24$ controlled configuration. The shrinkage coefficient is selected on validation data only.

Variant	MAE	MSE $\times 10^{-3}$	RMSE	MAPE (%)	Mean tensor norm
Imitation-HIP	184.960 ± 3.408	89.339 ± 2.624	298.870 ± 4.410	6.176 ± 0.149	78.684
Shrinkage-HIP ($\lambda = 0.15$)	177.100 ± 2.087	82.317 ± 2.382	286.885 ± 4.173	$\mathbf{5.725 \pm 0.073}$	15.671
Retention-only RL	180.596 ± 1.584	83.600 ± 1.341	289.129 ± 2.327	5.820 ± 0.062	19.941
Full RL-HIP	$\mathbf{175.719 \pm 1.212}$	$\mathbf{79.864 \pm 1.047}$	$\mathbf{282.597 \pm 1.855}$	5.855 ± 0.053	20.686

Relative to Retention-only RL, Full RL lowers MAE by 4.877 MW with paired 95% CI $[+1.553, +8.202]$ for Retention-only$-$Full RL; the MSE and RMSE intervals also exclude zero, while MAPE does not. Selection alone therefore does not reproduce the Full-RL forecasting result. In contrast, Shrinkage-HIP is only 1.382 MW worse than Full RL, with paired 95% CI $[-1.663, +4.427]$ for Shrinkage$-$Full RL; the corresponding MSE, RMSE, and MAPE intervals also include zero. Hence the current experiment does not establish an advantage of Full RL over validation-selected global attenuation.

The controls together suggest that forecast calibration changes more than event retention, but a substantial portion of the improvement can be explained by reducing the effective strength of the conditioning signal. This motivates interpreting RL-HIP as forecast-driven calibration of a structured conditioning state rather than as evidence that every semantic or temporal rewrite is independently necessary.

The reported tensor norms are measured from each condition's own trained adapter on the test set. Shrinkage is applied after semantic/temporal aggregation as $\mathbf{E}_{\text{shrink}} = \lambda \mathbf{E}$, but each λ cell trains a separate final forecaster and trainable pathway adapter. Therefore the selected $\lambda = 0.15$ condition

does not share the same semantic-embedding and duration-modulation parameters as the separately trained Imitation-HIP condition, and its mean tensor norm need not equal 0.15×78.684. For each seed, we average the Frobenius norm $\|\mathbf{E}_i\|_F$ over test origins and then average that statistic over the five forecaster seeds.

E Representation Diagnostics

We complement the forecasting results with representation-level diagnostics. *Supervised imitation* denotes the pathway policy after warm-starting on LLM-generated pseudo-pathways but before forecast-feedback RL calibration. It is included here to separate the effect of imitating the structured prior from the additional effect of forecast-based calibration; it is not a separate forecasting architecture.

The simplified controlled-protocol forecasting comparison between Imitation-HIP and RL-HIP is reported in Section 4.2; the diagnostics below focus on how the internal pathway representation changes across learning stages.

E.1 Forecasting Diagnostic Across LLM, Imitation, and RL Stages

As an additional forecast-geometry robustness diagnostic, Table 16 reports a PatchTST stage comparison under $L = 48, H = 48$. The comparison includes direct LLM-generated pathways, the supervised-imitation checkpoint before PPO, and the forecast-calibrated RL-HIP policy. This full-day diagnostic is reported separately from both the main $L = 48, H = 24$ benchmark in Table 2 and the simplified $H = 24$ controlled study in Appendix D.1.

Table 16: PatchTST forecasting diagnostic across representation-learning stages under the AULF $48 \to 48$ full-day robustness setting. This diagnostic is reported separately from the main $48 \to 24$ benchmark and the simplified $H = 24$ controlled study. Lower is better.

Method	MAE (MW)	MAPE (%)
No events	246.63	7.98
Raw-event	241.50	7.75
LLM-HIP	243.92	8.01
Supervised imitation	237.76	7.93
RL-HIP	**230.93**	**7.59**

Within this diagnostic, supervised imitation reduces MAE from 243.92 to 237.76 relative to direct LLM-HIP, while forecast-feedback calibration further reduces MAE from 237.76 to 230.93. The latter corresponds to a 6.83-MW, or approximately 2.87%, reduction relative to the imitation checkpoint. This result provides a robustness check on the stage-wise trend under a full-day horizon.

The representation statistics below belong to the same full-day robustness diagnostic and are interpreted separately from the simplified $H = 24$ controlled study in Appendix D.1.

Table 17: Representation-level pathway statistics. Lower D_{path} indicates more consistent pathway assignments. These quantities characterize the internal representation rather than causal or semantic correctness of individual pathways.

Metric	LLM-HIP	Supervised imitation	RL-HIP
Validity	1.00	1.00	1.00
Avg. retained	16.00	9.98	10.16
Mean intensity	0.8505	0.8111	0.6636
$D_{\mathrm{path}} \downarrow$	1.1382	0.6649	0.6512
Intensity–residual Spearman	0.1968	0.5074	0.3798

The three-stage comparison shows that representation changes should not be attributed to RL by comparing only LLM-HIP with RL-HIP. Supervised imitation already accounts for most of the reduction in retained count and D_{path}, while RL-HIP primarily produces a further reduction in mean

intensity and a modest additional reduction in D_{path}. Supervised imitation also has the strongest intensity–residual Spearman correlation. These diagnostics therefore characterize how the representation evolves across LLM initialization, imitation, and forecast-feedback calibration; they do not establish that PPO uniquely produces compactness, semantic validity, or every favorable representation statistic.

F Extended NSW Case Study

The main text uses NSW 2021-11-20 only as an illustrative example. This appendix exposes the corresponding pathway-level artifacts for auditability. The final RL policy retains 14 pathways for the instance; four originate from source annotations with `keep=0` and are retained or relabelled after calibration. This behavior demonstrates that forecast calibration can modify the initial LLM proposal, but it does not establish that every rewritten label is semantically correct.

Top-five retained pathways ranked by unsigned representation-pressure sum

rank	event_id / title	category / mechanism	direction	lag	persistence	pressure sum
1	event_159131 Rain threatening to disrupt search for William...	extreme_weather residential_cooling_demand_increase	positive	6h	3-7d	24.42
2	event_159086 Scary footage as waterfall cascades over road in...	infrastructure_failure grid_disruption	negative	6h	3-7d	15.44
3	event_159120 New find at William Tyrrell search site as...	public_health behavioral_shift_indoor	positive	6h	3-7d	7.22
4	event_158994 Oysters from Coffin Bay recalled after 56 Vibrio...	public_health behavioral_shift_indoor	positive	7h	3-7d	7.02
5	event_159125 'Dangerous views': Scott Morrison's vaccine...	public_health behavioral_shift_indoor	positive	7h	3-7d	6.86

Figure 3: Ranked retained pathways for the NSW 2021-11-20 iTransformer case. Ranking and pathway fields are representation-level model outputs and are not causal-effect or semantic-validity scores.

The following diagnostics show when retained pathways are active inside the conditioning representation. The pathway-pressure quantities are unsigned: they characterize internal conditioning strength and should not be read as signed forecast contributions in MW.

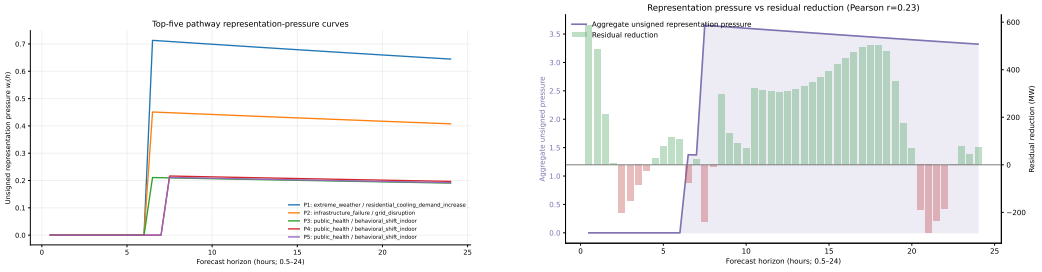

Figure 4: Representation dynamics for the NSW 2021-11-20 case. Left: top-five unsigned pathway-pressure curves. Right: aggregate unsigned pathway pressure and residual reduction across the forecast horizon. Neither panel is a signed causal contribution decomposition.

Finally, we disable selected pathways inside the same trained RL-HIP checkpoint to measure sensitivity of the prediction to the retained representation. This is a *pathway-deletion sensitivity analysis*: the all-off condition is not equivalent to an independently trained No-events model and is not a real-world counterfactual.

G Implementation and Reproducibility Details

G.1 Policy Encoder and Structured Action

Event titles are encoded once with frozen `distilbert-base-uncased` embeddings (768 dimensions; maximum token length 128; mean pooling). The policy projects this vector through

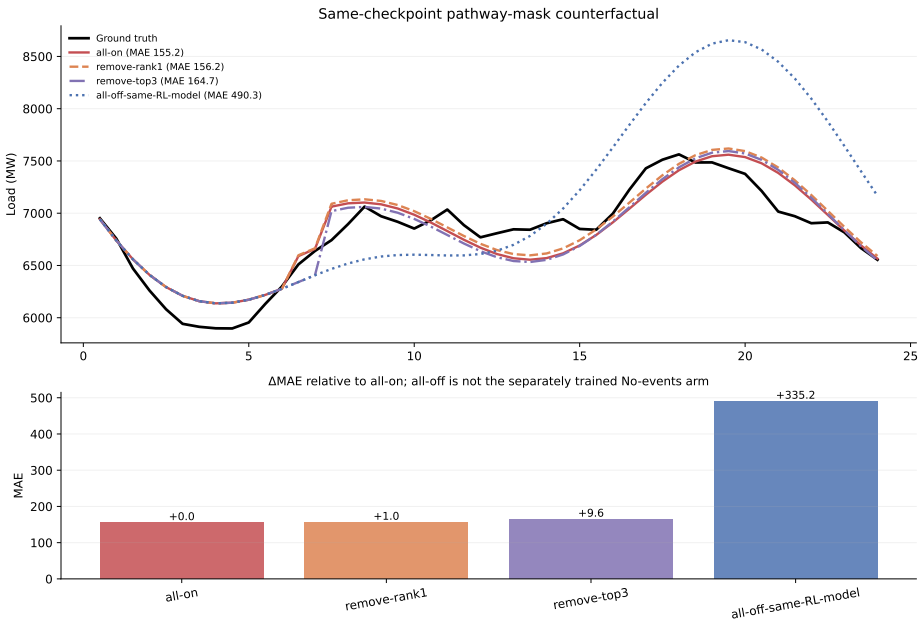

Figure 5: Same-checkpoint pathway-deletion sensitivity for the NSW 2021-11-20 iTransformer case. The diagnostic tests dependence on the retained pathway representation inside RL-HIP; "all off" is not the independently trained No-events forecast.

$\mathrm{Linear}(768, 256)$+GELU. The scalar target history is encoded by a one-layer GRU with hidden size 128. Structured covariates are averaged over the concatenated lookback/future-covariate sequence and projected through $\mathrm{Linear}(d_s, 64)$+GELU. Concatenating the three blocks gives 448 dimensions, followed by LayerNorm and $\mathrm{Linear}(448, 256)$+GELU to form $\mathbf{h}_i$. Previously sampled pathway attributes are represented by 32-dimensional embeddings whose sum forms the autoregressive attribute context. Retention uses a separate $\mathrm{Linear}(256, 2)$ head. Candidate actions are sampled independently conditional on their event/context states and are combined through an origin-level summed log probability for PPO.

G.2 Semantic Adapter, Direction, Intensity, and Temporal Mapping

The forecaster-side semantic embedding is

$$\psi(c, m, q, r) = \mathrm{Emb}_c(c) + \mathrm{Emb}_m(m) + \mathrm{Emb}_q(q) + \mathrm{Emb}_r(r), \qquad d_e = 16.$$

These four embedding tables belong to the forecasting adapter, not the policy; they are trained with the reward forecaster and, independently, with each final forecaster. Direction uses $\sigma(\texttt{positive}) = +1$, $\sigma(\texttt{negative}) = -1$, and $\sigma(\texttt{mixed}) = \sigma(\texttt{neutral}) = 0$. Intensity uses $\alpha \in \{0.34, 0.67, 1.00\}$ for low/medium/high. The duration label selects a learned 16-dimensional vector $\eta(d)$ initialized to ones.

For AULF, lag labels expressed in hours are converted using two 30-minute steps per hour and clamped to the forecast horizon. In the main $H = 24$ setting, 0h–11h map to onsets 0–22 steps, while labels whose converted onset is at least 24 are inactive over horizon indices 0–23. Persistence labels are interpreted as half-lives and converted through $\rho = \log(2)/\text{half-life}_{\text{steps}}$. For Bitcoin, one forecast step is one day and the same half-life conversion is applied in daily units.

G.3 Conditioning Injection

All four numerical backbones preserve their native historical encoder and use the same future-conditioned decoder interface. Let $\mathbf{z}_{\text{rep}} \in \mathbb{R}^{128}$ be the backbone history representation. At each future step h, the decoder concatenates

$$[\mathbf{z}_{\text{rep}},\ \mathbf{s}_{t+h},\ \mathbf{e}_{t+h}],$$

which has dimension $128 + 19 + 16 = 163$ for AULF, and maps it through $\mathrm{Linear}(163, 256) \to \mathrm{GELU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(256, 1)$. The 19-dimensional AULF structured-covariate vector follows the inherited data construction of Wang et al. (2024) and comprises 8 weather variables, 6 calendar/time variables, and 5 regional one-hot indicators. HIP therefore conditions the future decoder rather than changing the backbone's historical input dimension. The Raw-event baseline mean-pools frozen DistilBERT title embeddings over retrieved articles, projects the 768-dimensional vector to 16 dimensions, broadcasts the same vector to every horizon, and injects it at the same decoder location.

For the LLM forecasting interface, the inherited forecasting prompt is kept fixed and the event slot is replaced by either "No relevant news available" or a list of serialized pathways. Each pathway verbalizes its semantic attributes, lag, persistence, intensity, and uncertainty; the full numerical tensor is not serialized step by step.

G.4 Annotation and Imitation

The AULF pseudo-label annotator is a local Llama-2-7B-Chat checkpoint run through vLLM. The first decoding attempt is greedy (temperature 0); parsing/validation failures allow up to two retries with temperature 0.7 and top-$p = 0.9$. The annotator receives the article title, publication time, optional region hint, and summary/full article text, and returns a JSON retain/discard decision plus pathway fields for retained articles. It is instructed to default to discard unless the event plausibly shifts electricity demand at population scale, to use only the predefined vocabulary, and to provide an evidence span and one-sentence pathway explanation. Forecasting targets are never supplied to the annotator. For the main policy, the pseudo-label corpus used for imitation is global rather than forecasting-chronological: it contains article records overlapping the 2021 test period and later records. Accordingly, imitation is a cross-temporal text-only warm start; it does not consume forecasting targets or test metrics. Imitation learning uses the resulting pseudo-labels as weak supervision and selects its checkpoint by keep-F1 plus mean attribute accuracy plus sampled schema validity.

G.5 Reward-Forecaster and PPO Configuration

For each numerical backbone, the reward forecaster uses the same backbone family as the corresponding final forecaster but is trained as a separate checkpoint and frozen during PPO. It is optimized with MAE and selected by minimum validation MAE. Pathway dropout 0.5 is used during reward-forecaster training. The final AULF benchmark uses direct LLM-HIP-aligned reward-forecaster pathways with a pathway cap of 16. The simplified controlled configuration in Appendix D.1 instead uses Imitation-HIP-generated alignments with a pathway cap of 32; the reward forecasters and calibrated policies are trained separately for these two configurations.

The policy optimizer is AdamW with learning rate 3×10^{-5} and weight decay 10^{-5}. PPO uses clip $\epsilon = 0.2$, two optimization epochs per update, policy-gradient norm clipping at 0.5, no entropy bonus, and 50 updates. The final AULF benchmark uses rollout batches of 512, whereas the simplified controlled configuration uses rollout batches of 64. Other PPO settings listed here are shared unless an explicitly labeled forecast-geometry robustness experiment states otherwise. A separately optimized value network maps $[\mathrm{mean}(x), \mathrm{std}(x), \mathrm{mean}(s)]$ through $\mathrm{Linear}(2 + d_s, 128) \to \mathrm{GELU} \to \mathrm{Linear}(128, 1)$; its AdamW learning rate is 10^{-3}. Advantages use the one-step estimate $R_t - V(s_t)$ and are standardized within the rollout batch. There are no action-dependent transitions across forecast origins.

G.6 Final Forecasters and Metric Aggregation

Each final numerical forecaster is trained independently for 50 epochs with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, batch size 128) and the minimum-validation-MAE checkpoint is retained. Target normalization uses per-instance RevIN computed from the lookback series. For each seed, metrics are computed over all origin–horizon elements:

$$\mathrm{MAE} = \mathrm{mean}\,|y - \hat{y}|, \quad \mathrm{MSE} = \mathrm{mean}(y - \hat{y})^2, \quad \mathrm{RMSE} = \sqrt{\mathrm{MSE}}, \quad \mathrm{MAPE} = 100\ \mathrm{mean}\,\frac{|y - \hat{y}|}{|y|}.$$

Unless otherwise specified, experiments use five final-forecaster seeds. Tables 2 and 3 report the arithmetic mean only; appendix controls may additionally report sample standard deviations and paired confidence intervals. For paired five-seed controls, confidence intervals are computed from seed-matched differences as

$$\bar{d} \pm t_{0.975,4} \frac{s_d}{\sqrt{5}}, \qquad t_{0.975,4} = 2.7764.$$

These intervals quantify final-forecaster initialization variability conditional on the frozen control policy.