REIN: RESIDUAL MODULATION OF TIME SERIES FORECASTS VIA LLM-REFINED NEWS REGIME SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasts in operational domains are often disrupted by external events such as heatwaves, generator outages, policy interventions, supply shocks, and market incidents. Most deep forecasting models rely on numerical histories, while recent text-augmented methods offer limited control over how far text can move a forecast, which temporal component text should influence, and whether observed gains reflect time-aligned news content rather than news availability or calendar correlations. We introduce REIN (Residual Information from News), a two-stage framework that reformulates news-aware forecasting as bounded residual modulation. Our design leverages an LLM agent to distill raw news records into a leakage-controlled regime bank of compact, time-aligned signals. We further incorporate a PatchTST-based delta module to decompose the residual of a frozen base forecaster into persistent-level, intra-day-shape, and extreme-event-spike channels. Following our design principles, news does not write forecasts directly. It only scales these residual channels through a small set of bounded modulation knobs. This construction enforces a bound on deviation from the base forecast and provides a clean reversion path when the news bank is empty. A counterfactual protocol with news blanking and date permutation, evaluated on disjoint news-active and news-inactive subsets, provides sample-level evidence that improvements are tied to selective use of aligned news. On four heavy-tailed real-world datasets (Electricity Demand, Electricity Price, Traffic, Gas Demand) at horizons up to 720 steps, REIN improves MAE by 2–21% over a strong DLinear base and presents state-of-the-art performance on most settings. Blanking the news increases news-active MAE on all four datasets, by up to 42.6% on the most news-sensitive setting (Traffic, H=96), supporting the view that REIN benefits from structured news signals rather than from adding text indiscriminately.

1 INTRODUCTION

Accurate time series forecasting is crucial to decision-making in energy markets, transportation, finance, and other operational domains, where forecast errors translate directly into additional cost. Modern neural forecasting models generally exploit trend, seasonality, calendar structure, and local temporal dependence effectively, which are determined by numerical history alone (Zeng et al., 2023; Nie et al., 2023; Zhou et al., 2021). This is a fundamental limitation for scenarios, where external events such as heatwaves, outages, policy interventions, supply constraints, and shifts in market sentiment can alter the level, shape, or volatility of a series. These events are by definition poorly represented before their effects are fully reflected in the observed history, yet human analysts routinely consult contemporaneous textual sources such as news feeds, operator notices and weather warnings.

While recent works provide no direct mechanism for incorporating contemporaneous textual evidence, this limitation is especially important under eventful conditions: a purely numerical model can extrapolate regular patterns, but it cannot know that a policy ruling, fuel-supply disruption, weather warning, or market incident has changed the expected regime unless such information has already appeared in the time series. A growing body of work has begun to incorporate text into forecasting through prompt-based LLM forecasters (Gruver et al., 2023; Jin et al., 2024; Wang et al., 2024a), cross-

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.

modal fusion (Zhang and Yan, 2023), and retrieval-augmented approaches (Han et al., 2025). These methods demonstrate the potential of textual context, but there are several methodological gaps for operational use. First, text at large volume is rarely treated as a first-class input for most approaches, leaving potentially informative textual evidence unused or compressed into weak auxiliary features. Second, the correlation between text and forecast is unclear. It is usually impossible to distinguish improvements driven by aligned news content from gains that merely reflect news availability or calendar correlations. Third, the text-to-forecast pathways usually do not expose which temporal component, such as persistent level, intra-day shape, or extreme-event spike, is affected by text. These limitations can make text-augmented forecasting unreliable in operational settings. Ignoring broad news streams may cause models to miss event-driven regime changes, while opaque text-fusion mechanisms can allow noisy or misaligned news to perturb forecasts in uncontrolled ways. Such failure modes are especially consequential in heavy-tailed and regime-switching domains, including electricity-price spikes, gas demand under cold snaps, and sub-daily traffic disruptions, where small changes in event interpretation can lead to large forecasting errors.

In this paper, we focus on news as the most prevalent textual source and raise the research question: can news be incorporated into time-series forecasting so that its impact is bounded by construction, attributable to identifiable forecast components, and provably reverts to a strong numerical base when no such evidence is present? To achieve this goal, we propose **REIN**, which redesigns news augmented forecasting as structured residual modulation, rather than as a free additive feature. The design is based on three principles. First, news should not compete directly with the dominant numerical regularities. A base forecaster is trained on the time series data alone and frozen, and the textual evidence acts only on the residual for correction. Second, the residual itself is interpretably structured rather than a consolidated vector, which will be decomposed into three temporal channels, namely slow level shifts, shape changes, and spike behavior. Third, news should modulate these channels rather than write into them directly. We propose leveraging a small set of bounded modulation knobs that may scale each channel's magnitude, so the news' impact is architecturally constrained and degrades gracefully in the absence of actionable news. Thus, the model yields a provable bound on news-induced deviation from the base forecaster.

Following the principles, it requires turning unstructured raw news records into a signal carrying bounded news-conditioned modulation knobs. Specifically, in REIN, we introduce an LLM agent distills raw news records into structured regime records that encode actionability, topic, intensity, duration, and confidence. These records are aggregated into a leakage-controlled regime bank. On the forecasting side, a PatchTST-based delta module decomposes the frozen base residual into slow, shape, and spike channels whose magnitudes are controlled by the modulation knobs. When no actionable news is present, the modulation collapses toward identity values and the model remains close to the frozen base forecaster, yielding a bounded perturbation guarantee. Our contributions are summarized as follows:

- **Bounded news influence with a formal degradation guarantee.** News enters REIN only through bounded modulation knobs that scale a structured residual on top of a frozen base, yielding an architecturally enforced bound on how far news can move the forecast, and a clean reversion to the base whenever the news bank is empty. This converts "how much can news move the forecast" from an empirical question into a structural guarantee.
- **Channel-decomposed residual that exposes where text acts.** REIN decomposes the residual into three temporally interpretable channels (persistent level, intra-day shape, and extreme-event spike), each scaled by its own bounded news-conditioned knob. This design links each news-conditioned correction to a specific temporal component of the forecast, rather than absorbing textual information into an opaque cross-modal embedding.
- **Counterfactual protocol for sample-level news assessment.** We evaluate every test window under three conditions: true regime pack, blanked regime pack, and date-permuted regime pack. The evaluation is split across news-active and news-inactive subsets. Combined with per-sample help-rate diagnostics, this protocol distinguishes improvements tied to aligned news content from improvements driven by news availability or calendar correlations, and identifies which test samples benefit.
- **Evaluation on heavy-tailed, operationally critical domains.** We validate REIN on four real-world datasets where time series are heavy-tailed and news streams are noisy (Electricity Demand, Electricity Price, Traffic, Gas Demand), with forecast horizons up to 720 steps and

news sources ranging from curated domain feeds to the open GDELT corpus. Across these settings, REIN demonstrates state-of-the-art performance comparing to strong time-series, foundation-model, and LLM-based baselines while preserving controlled news influence.

2 Related Work

Deep time-series forecasting and foundation models. Long-horizon forecasting has been extensively studied with both lightweight decomposition models and Transformer-based architectures. Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), and FEDformer (Zhou et al., 2022) adapt attention and decomposition mechanisms to long sequences, while PatchTST (Nie et al., 2023) improves forecasting by representing each channel as a sequence of patches. In parallel, DLinear (Zeng et al., 2023) shows that simple decomposition-based linear models remain strong baselines, and iTransformer (Liu et al., 2024) models dependencies by treating variates as tokens. More recent foundation models such as TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), and related pretrained forecasters pursue cross-domain transfer through large-scale time-series pretraining. These methods provide strong numerical extrapolators, yet by relying primarily on historical values and calendar covariates, they remain blind to external event information available in textual form.

Text-augmented and LLM-assisted forecasting. Recent work integrates language models and textual information for time-series forecasting along two broad lines. One line of work repurposes LLMs as forecasters or temporal representation models by casting numerical forecasting as text generation, adapting pretrained language backbones, or aligning time-series patches with language-model representations (Gruver et al., 2023; Zhou et al., 2023; Jin et al., 2024; Wang et al., 2024a). Another line incorporates external text such as news, reports, events, or natural-language descriptions through sentiment features, topic signals, retrieval, multimodal encoders, or paired text–time representations (Loughran and McDonald, 2020; Han et al., 2025; Wang et al., 2026). Both lines show the value of textual context, though could not provide explicit information for operational use. These methods often treat text as an unconstrained additive signal or prompt, with no architectural bound on its influence. Thus, it provides limited insights for the localization of any specific temporal component towards the resulting correction. Moreover, the forecast when relevant text is absent is left to the data rather than guaranteed by design.

Residual decomposition, modulation, and counterfactual evaluation. Forecast decomposition and residual modeling are well established tools for robustness and interpretability in time-series forecasting. N-BEATS introduces interpretable basis expansions for trend and seasonality, Autoformer embeds series decomposition as an internal block, and DLinear shows that explicit trend–remainder decomposition can form a strong long-horizon baseline (Oreshkin et al., 2020; Wu et al., 2021; Zeng et al., 2023). Conditional modulation also provides a mechanism for controlling how auxiliary information affects a neural model, as in FiLM-style feature modulation (Perez et al., 2018). These methods inform the decomposition and modulation components in REIN. However, they do not address how to verify whether event-driven textual conditions affect forecasts for the intended reason. Prior counterfactual studies mainly perturb numerical histories, while tests over news conditions remain limited. This gap motivates the three-condition counterfactual protocol we introduce in Section 4, by comparing active, blanked, and date-permuted news regimes.

3 Method

3.1 Overview

Base stage. The structure of REIN is shown in Figure 1. REIN has two stages. In the first stage, called **Base**, a single forecaster f_{base} on numerical histories only is trained and then frozen. We use a frozen DLinear backbone for all four datasets in this paper. DLinear provides a lightweight numerical base that captures dominant trend and seasonal structure with low training cost. Its decomposition is also complementary to the slow / shape / spike residual design: the base absorbs regular calendar-driven patterns, while the delta module focuses on event-driven deviations that news may explain. Finally, DLinear's simple linear signal path makes the base prediction easier to separate from news-driven residual corrections, supporting cleaner attribution and counterfactual analysis.

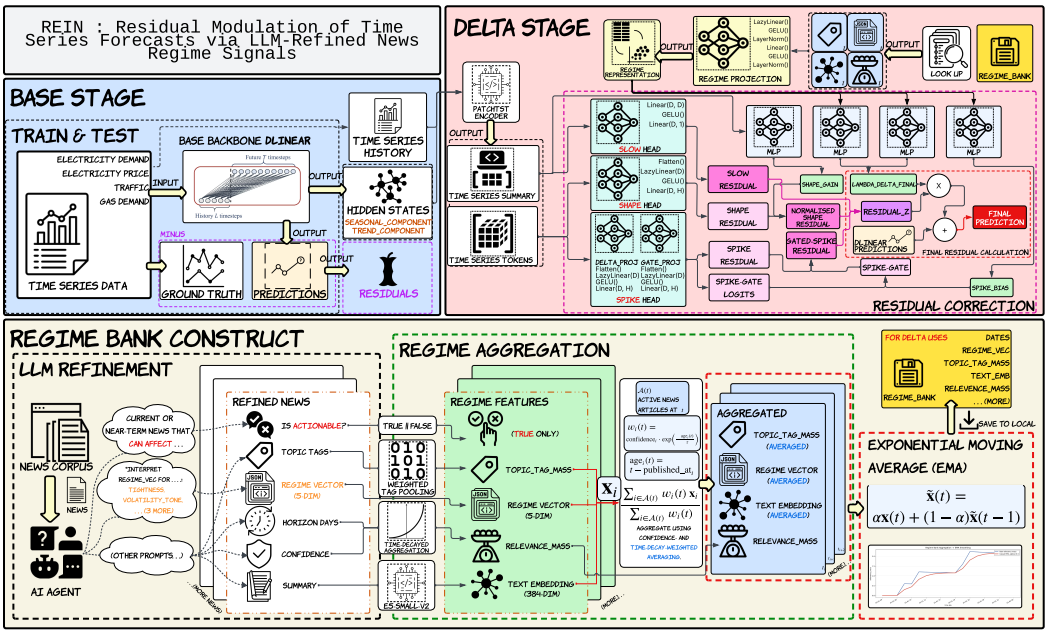


Figure 1: The structure of REIN. First, a frozen DLinear base forecaster produces numerical predictions. REIN then learns residuals between ground truth and base forecasts for bounded news-conditioned modulation in the delta stage. During the regime bank construction, raw news records are refined by an LLM agent into structured regime records, including actionability, topic tags, regime vectors, horizons, confidence scores, summaries, and text embeddings. Active news records are aggregated with confidence and time-decay weights and smoothed into a time-aligned regime bank.

News Regime Bank Construction. After the base stage is trained, REIN constructs the regime bank if no existing bank is available. An LLM agent processes each raw news record using dataset-specific prompts and converts it into a structured regime record. Each record contains an actionability flag, topic tags, a regime vector, an estimated influence horizon, a confidence score, and a concise summary. Non-actionable news records are filtered out, while actionable records are saved locally and later retrieved by the delta stage as time-aligned news signals.

Delta stage. In the **Delta** stage, we train a news-conditioned residual module on top of the frozen base. Given a forecast-origin time t with normalized history $\boldsymbol{h}_z(t) \in \mathbb{R}^L$, calendar features $\boldsymbol{c}_t$, and a regime pack $\mathcal{B}(t)$ retrieved from the news bank (Sec. 3.2), the prediction in z-space is

$$\hat{\boldsymbol{y}}_z(t) \;=\; \hat{\boldsymbol{y}}_z^{\text{base}}(t) \;+\; \lambda_{\text{delta_final}}(t) \cdot \boldsymbol{r}_{\text{delta}}(t), \tag{1}$$

where $\boldsymbol{r}_{\text{delta}}(t) \in \mathbb{R}^H$ is a structured residual correction produced by the delta module (Sec. 3.3) and $\lambda_{\text{delta_final}}(t) \in [0, \lambda_{\max}]$ is a bounded scalar trust coefficient. The subscript z indicates that all training and modulation operate in the z-normalized space; predictions are denormalized only for evaluation.

Multiplicative-only coupling and a degradation bound. A central design choice is that news *never* enters Eq. 1 additively: the regime pack $\mathcal{B}(t)$ controls only $\lambda_{\text{delta_final}}$ and two further bounded knobs (g_{shape}, b_{spike}, defined in Sec. 3.3) that *scale* the time-series-derived residual components. Every news-driven term is multiplied by an activity gate and bounded above by validation-selected caps, so the worst-case perturbation of the final forecast away from the base is bounded:

$$\left\|\hat{\boldsymbol{y}}_z(t) - \hat{\boldsymbol{y}}_z^{\text{base}}(t)\right\|_\infty \;\le\; \lambda_{\max} \cdot \left\|\boldsymbol{r}_{\text{delta}}(t)\right\|_\infty. \tag{2}$$

Moreover, on news-inactive samples ($m(t) \le \theta$, Sec. 3.2) the soft activity gate drives $\lambda_{\text{delta_final}}(t) \to 0$ and the right-hand side approaches zero, so REIN reverts toward the frozen base, a property that purely additive text-to-time-series fusion methods do not provide.

3.2 NEWS REGIME BANK

Each news record is refined by an LLM agent, we use ChatGPT-5.1 under a strict JSON schema, yielding an actionability flag (non-actionable news records will be ignored), a regime vector $\boldsymbol{r}_i$ from a 5-axis template (*tightness* / *demand* / *supply* / *volatility* / *policy*, instantiated per domain), topic tags $\boldsymbol{\tau}_i \in \{0,1\}^{15}$, an in-force horizon h_i in days, a self-assessed confidence $c_i \in [0,1]$, and an E5-small-v2 summary embedding $\boldsymbol{e}_i \in \mathbb{R}^{384}$. Appendix E discusses the rationale for the five-axis regime schema and the safeguards used to reduce black-box LLM news filtering. For each forecast-origin bin t (matching the native series frequency), we retain only news records with $\text{published_at}_i \le t \le \text{published_at}_i + h_i$ and aggregate them with weights $w_i(t) = c_i \exp(-\text{age}_i(t)/\tau)$ at τ=5 days, producing confidence-weighted means $\bar{\boldsymbol{r}}(t), \bar{\boldsymbol{\tau}}(t), \bar{\boldsymbol{e}}(t)$ and the unnormalized relevance mass $m(t) = \sum_i w_i(t)$, all smoothed across t by a causal EMA (α=0.5). A hard activity indicator $A(t) = \mathbb{I}[m(t) > \theta]$ and a soft sigmoid $g_{\text{soft}}(t) = \sigma((m(t)-\theta)/\tau_g)$ (τ_g=0.05) gate the news influence; θ is tuned per dataset on validation MAE ($\theta \in \{0.1, 0.2, 0.7\}$). The retrieved regime pack is $\mathcal{B}(t) = \big(\bar{\boldsymbol{r}}(t), \bar{\boldsymbol{\tau}}(t), \bar{\boldsymbol{e}}(t), m(t)\big)$. Full prompt, per-domain schema, and aggregation pseudocode are in Appendix J.

3.3 DELTA MODULE: DECOMPOSITION AND BOUNDED MODULATION

The delta module ingests $\boldsymbol{h}_z$, a pooled summary $\boldsymbol{f}_{\text{base}}$ from the frozen base, calendar features $\boldsymbol{c}_t$ (DOW one-hot, half-hour-of-day sinusoidal, holiday flag), and $\mathcal{B}(t)$. A PatchTST encoder (2 layers, 4 heads, D=128, patch length 8, stride 4) prepends $\boldsymbol{f}_{\text{base}}$ as an extra token and produces token features $\boldsymbol{T}$ and a pooled summary $\boldsymbol{s}$; the regime path layer-norms each component of $\mathcal{B}(t)$ separately and projects them through a 2-layer GELU MLP to $\boldsymbol{s}_{\text{reg}} \in \mathbb{R}^D$. The structured residual decomposes into a slow level shift, a zero-mean intra-day shape, and a per-step spike correction:

$$\boldsymbol{r}_{\text{delta}}(t) = r_{\text{slow}}(t)\,\mathbf{1}_H + g_{\text{shape}}(t)\,\boldsymbol{r}_{\text{shape}}(t) + \sigma\big(\boldsymbol{\ell}_{\text{spike}}(t) + b_{\text{spike}}(t)\big) \odot \boldsymbol{r}_{\text{spike}}(t), \tag{3}$$

where the three TS-only heads $r_{\text{slow}}{=}f_{\text{slow}}(\boldsymbol{s}) \in \mathbb{R}$, $\boldsymbol{r}_{\text{shape}}{=}f_{\text{shape}}(\boldsymbol{T}) - \overline{f_{\text{shape}}(\boldsymbol{T})} \in \mathbb{R}^H$, and $(\boldsymbol{r}_{\text{spike}}, \boldsymbol{\ell}_{\text{spike}}){=}f_{\text{spike}}(\boldsymbol{T}) \in \mathbb{R}^H \times \mathbb{R}^H$ read only from the time-series tokens. News enters Eq. 3 solely through three bounded modulation knobs:

$$\begin{aligned} \lambda_{\text{delta_final}}(t) &= \text{clip}\big(\lambda_{\text{ts}}(t) + \Delta\lambda_{\text{news}}(t),\, \lambda_{\min},\, \lambda_{\max}\big) \cdot g_{\text{soft}}(t), \\ g_{\text{shape}}(t) &= 1 + \gamma_{\text{shape}} \tanh\big(M_{\text{shape}}(\boldsymbol{s}_{\text{reg}})\big) A(t), \quad b_{\text{spike}}(t) = \gamma_{\text{spike}} \tanh\big(M_{\text{spike}}(\boldsymbol{s}_{\text{reg}})\big) A(t), \end{aligned} \tag{4}$$

with $\lambda_{\text{ts}}(t) \in [\lambda_{\min}, \lambda_{\min}{+}\lambda_{\text{ts}}^{\text{cap}}]$ derived from $\boldsymbol{s}$ and $\Delta\lambda_{\text{news}}(t) \in [-\lambda_{\text{news}}^{\text{cap}}, \lambda_{\text{news}}^{\text{cap}}]{\cdot}A(t)$ derived from $\boldsymbol{s}_{\text{reg}}$ via 2-layer MLPs. On news-inactive samples (A=0, $g_{\text{soft}}{\to}0$) we have $\Delta\lambda_{\text{news}}{=}0$, $g_{\text{shape}}{=}1$, $b_{\text{spike}}{=}0$, $\lambda_{\text{delta_final}}{\to}0$, recovering the bound in Eq. 2; on active samples, news influence is bounded above by the validation-selected caps. The three TS-head targets are derived from the true residual via a 7×48 DOW×half-hour calendar baseline plus a quantile-and-rank intersection spike mask, with full equations in Appendix N. Example diagrams of residual decomposition can be found in Appendix M.

3.4 TRAINING AND COUNTERFACTUAL EVALUATION

The delta-stage loss combines a Smooth-L1 point loss (winsorized z-space quantile loss for Electricity Price), MSE losses on the slow / shape / spike targets, a BCE on the spike-gate logits, and two auxiliary terms:

$$\mathcal{L} = \mathcal{L}_{\text{point}} + w_s\mathcal{L}_{\text{slow}} + w_q\mathcal{L}_{\text{shape}} + w_k(\mathcal{L}_{\text{spike}} + w_g\mathcal{L}_{\text{gate}}) + w_c\mathcal{L}_{\text{consist}} + w_f\mathcal{L}_{\text{counter}}. \tag{5}$$

$\mathcal{L}_{\text{consist}}$ penalizes $\|\hat{\boldsymbol{y}}_z^{\text{normal}} - \hat{\boldsymbol{y}}_z^{\text{blank}}\|^2$ on news-inactive samples to sharpen the inactive-blank gap that the soft gate alone makes small but non-zero; $\mathcal{L}_{\text{counter}}$ is a hinge that requires real-news predictions to beat date-permuted-news predictions on news-active samples, with the permutation seed freshly resampled every batch while the test protocol below uses a held-out fixed seed 2024, preventing overfit-to-permutation leakage. We additionally apply 30% news dropout as a regularizer. At test time each window is evaluated under **Active** (true regime pack), **Blank** (zeroed pack), and **Permuted** (date-shuffled, seed 2024); reporting is split between news-inactive samples (where Active≈Blank should hold by Eq. 2) and news-active samples (where Active should beat both alternatives if news content genuinely helps). Loss-weight values, dropout schedule, and full counterfactual protocol pseudocode are in Appendix I.

4 EXPERIMENTS

4.1 DATASETS

We evaluate REIN on four datasets that span distinct domains, sampling frequencies, and distributional regimes. The first two datasets are 2024 electricity demand and spot electricity price for the New South Wales (NSW) region of the Australian National Electricity Market, both sourced from AEMO (Australian Energy Market Operator). The price series is aggregated from the native 5-minute dispatch interval to a 30-minute resolution. The third dataset consists of hourly road traffic counts from NSW permanent count stations in 2024 (NSW Government), aggregated into a single state-level hourly series. The fourth dataset contains hourly natural gas consumption in the Netherlands in 2025, sourced from ENTSOG (ENTSOG); this series exhibits pronounced seasonal and calendar effects induced by heating demand and industrial offtake. Additional details on the time-series datasets and their paired news corpora are provided in Appendix B. The complete downsampling procedure is described in Appendix C. Forecast horizons are reported in units of each dataset's native sampling frequency; for example, $H = 48$ corresponds to one day for the 30-minute electricity load and price series, and to two days for the hourly gas and traffic series.

4.2 EXPERIMENTAL SETUP

We compare REIN against strong representative baselines from three families. The first family consists of modern deep time-series forecasters, including PatchTST(Nie et al., 2023), iTransformer(Liu et al., 2024), TimesNet(Wu et al., 2023), TimeMixer(Wang et al., 2024b), and TimeXer(Wang et al., 2024c), which represent widely used state-of-the-art architectures for long-horizon numerical forecasting. The second family covers time-series foundation models, including Chronos(Ansari et al., 2024), Chronos-2(Ansari et al., 2025), and Sundial(Liu et al., 2025), evaluated in the zero-shot setting. The third family covers LLM-based and multimodal forecasting methods, including GPT4TS(Zhou et al., 2023), Time-LLM(Jin et al., 2024), ChatTime(Wang et al., 2025), and Time-VLM(Zhong et al., 2025). The latter two provide particularly relevant comparisons as they incorporate textual or multimodal information into time-series forecasting. All methods are evaluated on the same train/validation/test splits, look-back length $L = 48$, and prediction horizons $H \in \{48, 96, 192, 336, 720\}$. We report mean squared error (MSE) and mean absolute error (MAE), where lower values indicate better performance (see Appendix G). In addition to these external baselines, we report the frozen DLinear base used by REIN to isolate the contribution of the proposed residual modulation module.

4.3 MAIN RESULTS

Table 1 reports the main comparison across datasets and horizons. We use this table to answer two questions. First, can bounded residual modulation improve over a frozen numerical base? Second, can it remain competitive with strong time-series, foundation-model, and LLM-based baselines? Across the evaluated settings, REIN is best or second-best on 15 of the 20 dataset–horizon rows. It is strongest on Electricity Demand, Electricity Price, and Gas Demand, while Traffic highlights a more mixed regime in which the proposed delta module improves strongly over the base but does not dominate specialized numerical mixers at all short horizons.

Electricity Price. Electricity Price is the most heavy-tailed target in our evaluation. As shown in Table 1, REIN achieves the best MSE and MAE at four horizons, $H \in \{48, 192, 336, 720\}$, and ranks second at H=96, where TimesNet obtains the lowest error. The advantage is visible at both short and long horizons. For example, at H=48, REIN reduces MAE from the second-best 126.1 to 107.3, while at H=720, it reduces MAE from 128.0 to 112.2. These results suggest that bounded residual modulation is effective for typical price movements, although rare price spikes remain challenging.

Electricity Demand. Electricity Demand provides the most consistent results. REIN achieves the best MSE and MAE at all five horizons in Table 1. The gains are especially clear at shorter and medium horizons, where the proposed model outperforms both time-series baselines and LLM-based forecasting methods. This indicates that the residual correction learned by REIN is well aligned with the regular load structure while still improving over the frozen numerical base.

Table 1: Comparison of forecasting performance on four real-world datasets under a fixed look-back window of L=48 and prediction horizons $H \in \{48, 96, 192, 336, 720\}$. The baseline suite covers strong representative models from multiple categories, including time-series foundation models, LLM-based and multimodal forecasting methods, and state-of-the-art deep time-series architectures. Reported numbers are MSE and MAE; lower is better. To keep the table compact, the MSE/MAE columns of each dataset are scaled by the factor shown in parentheses after the dataset name. Within every (dataset, horizon) row, the **best** result is marked in **red** and the second-best in blue with underline.

Dataset	H	REIN		Sundial†		Chronos†		Chronos-2†		Time-LLM		GPT4TS		ChatTime		Time-VLM		TimeMixer		TimesNet		TimeXer		iTransformer		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity Price (MSE×10^5)	48	**5.617**	107.3	6.594	139.4	13.28	144.0	14.48	139.2	6.231	144.2	6.116	132.4	5.640	**103.4**	5.902	133.1	6.404	167.0	5.834	129.1	6.016	137.6	5.830	126.1	6.065	141.0
	96	5.808	120.4	6.833	146.5	13.86	147.7	10.14	128.9	5.989	131.2	6.049	130.0	8.855	142.8	6.065	134.2	6.055	129.5	**5.785**	**116.6**	5.839	121.6	5.841	124.1	6.200	137.2
	192	**5.878**	**110.6**	7.062	154.3	11.55	147.0	8.665	127.3	6.117	133.9	6.520	136.1	8.046	143.6	6.059	128.7	6.180	141.1	5.968	121.8	6.082	132.0	6.092	136.2	6.124	127.9
	336	**6.222**	**123.5**	7.286	161.0	11.16	149.2	8.256	129.0	6.347	133.8	6.547	138.4	7.330	148.4	6.326	134.4	6.438	135.2	6.283	129.8	6.416	144.1	6.343	136.1	6.281	127.1
	720	**6.976**	**112.2**	8.229	176.9	11.66	155.9	8.226	136.3	7.138	138.2	7.829	153.9	10.569	207.9	7.186	147.7	7.094	134.3	7.163	146.7	7.497	168.8	7.208	150.5	7.034	128.0
Electricity Demand (MSE×10^6)	48	**0.504**	**470**	1.145	798	3.591	1392	2.773	1256	0.606	547	0.658	555	1.500	889	0.563	512	0.551	524	0.631	544	0.544	507	0.521	488	0.577	523
	96	**0.658**	**554**	1.457	909	3.661	1443	2.719	1266	0.892	664	0.899	662	2.020	1080	0.775	613	0.964	718	0.863	654	0.768	613	0.769	604	0.789	618
	192	**0.804**	**621**	1.720	998	3.342	1395	2.588	1252	0.987	717	1.043	732	2.260	1158	0.917	677	1.020	717	1.177	758	0.917	682	0.970	697	0.951	700
	336	**0.969**	**695**	1.841	1043	3.209	1373	2.485	1227	1.091	747	1.182	775	2.370	1191	1.040	724	1.124	757	1.325	796	1.025	716	1.114	747	1.037	720
	720	**1.045**	**723**	1.909	1068	2.901	1317	2.379	1200	1.168	782	1.191	800	2.530	1240	1.150	768	1.498	929	1.367	862	1.160	772	1.316	862	1.137	756
Traffic (MSE×10^8)	48	3.273	12037	5.092	15923	6.385	17392	5.342	15821	5.468	17673	5.760	18305	10.000	22500	4.180	15200	**2.584**	**10503**	2.828	10986	3.101	12763	3.551	13158	3.653	13895
	96	3.651	13098	6.212	17956	7.594	19835	6.281	17381	5.650	18276	6.750	19839	19.400	31600	4.910	16400	**3.193**	**11806**	3.319	11966	3.757	14061	5.019	16031	4.393	15266
	192	3.961	13474	5.613	16648	7.284	18653	7.735	17825	5.683	18113	6.110	18535	28.600	39000	4.340	15300	**3.402**	**12418**	3.608	13036	3.895	14486	4.919	15750	3.920	14483
	336	4.081	14117	5.610	17014	8.529	19898	8.615	18681	5.118	17346	6.200	19007	36.000	45000	4.470	15800	3.602	13356	**3.367**	**12730**	4.497	15921	5.669	17176	3.909	14658
	720	6.739	18402	6.687	18485	9.965	21756	12.36	21631	7.194	20920	8.570	22437	40.000	48700	6.140	18700	5.846	**16941**	7.515	20430	6.408	19281	8.146	20778	**5.311**	17232
Gas Demand (MSE×10^{13}, MAE×10^6)	48	**3.022**	**4.278**	4.108	4.927	3.768	4.575	4.146	4.726	3.845	4.817	3.550	4.993	5.830	5.860	3.560	4.550	3.776	4.750	4.276	5.005	3.694	4.759	3.765	4.661	3.587	4.596
	96	**3.747**	**4.998**	5.193	5.729	4.897	5.460	5.577	5.773	4.909	5.651	4.337	5.610	6.840	6.550	4.570	5.420	5.441	5.882	6.032	6.128	4.839	5.612	5.050	5.705	4.643	5.485
	192	**3.981**	**5.200**	5.194	5.702	5.116	5.611	5.673	5.831	4.762	5.561	4.851	5.933	6.750	6.440	4.740	5.580	5.221	5.753	5.297	5.839	4.824	5.631	5.130	5.742	4.752	5.604
	336	**4.356**	**5.370**	5.489	5.883	5.407	5.825	6.093	6.049	5.196	5.799	4.989	5.885	6.830	6.540	5.030	5.720	5.405	5.876	5.976	6.147	5.150	5.820	5.562	5.989	5.045	5.758
	720	**5.500**	**6.095**	6.665	6.538	6.551	6.479	7.879	6.873	6.379	6.415	7.180	6.920	7.500	6.950	5.990	6.250	6.844	6.654	7.070	6.777	6.212	6.299	6.696	6.580	5.956	6.219

† Zero-shot setting for the pretrained foundation models.

Traffic. Traffic shows a more mixed pattern. REIN improves over the frozen base and remains competitive across horizons, but it does not dominate all numerical baselines. At H=48 and H=96, specialized time-series models such as TimeMixer and TimesNet obtain lower errors. At longer horizons, REIN becomes more competitive. This result highlights an important trade-off: REIN can provide useful news-conditioned residual corrections, but strong numerical architectures may still be preferable when regular temporal patterns dominate the short-horizon task.

Gas Demand. Gas Demand shows stable but less dramatic differences among the strongest models. In Table 1, REIN obtains the best MSE and MAE at all five horizons. The margins over the closest baselines are generally smaller than those observed for Electricity Price and Electricity Demand, which suggests that much of the predictable structure is already captured by numerical histories. Even so, the consistent ranking indicates that the proposed residual modulation provides additional benefit across horizons.

Table 2: Delta-stack ablation against the frozen DLinear base, averaged over five horizons. Positive MSE and MAE gains indicate that the full REIN model improves over the base-only model. Help rate measures the fraction of test samples for which REIN has lower absolute error than the base, while top-10% help rate reports the same quantity on the largest base-error samples.

Dataset	MSE gain	MAE gain	Help rate	Top-10% help rate
Electricity Price	+0.81%	+5.11%	83.1%	84.2%
Electricity Demand	+16.64%	+9.19%	70.5%	77.9%
Traffic	+31.64%	+21.41%	87.3%	80.4%
Gas Demand	+7.88%	+2.30%	65.1%	93.1%

4.4 Residual Channel Validation

To verify that the slow, shape, and spike channels correspond to their intended temporal effects, we conducted a controlled synthetic validation with component-level ground truth. Specifically, at $H = 48$, we inject four types of perturbations into residual windows for each of the four datasets: (i) a constant level shift across the forecast horizon, (ii) a zero-mean intra-horizon shape distortion, (iii) sparse localized spikes, and (iv) a mixed setting containing all three perturbation types. The first three perturbations are designed to correspond to the slow, shape, and spike channels, respectively. Because the injected components are known, we can directly quantify their correspondence with the recovered

Table 3: Counterfactual evaluation of REIN. The full-test block compares the frozen base and REIN. The news-active block evaluates REIN under three regime-pack conditions: Active uses the true aligned regime pack, Blank zeros the regime pack, and Perm shuffles the date-to-regime mapping. Positive Δ values indicate that Active outperforms the corresponding counterfactual condition. Positive and zero Δ values are highlighted in **blue**, while negative values are highlighted in red.

Dataset	H	Full test set						News-active subset									
		Base		REIN		Δ_{Base}		Active		Blank		Δ_{Blank}		Perm		Δ_{Perm}	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity Price (MSE×10^5)	48	5.651	114.18	5.617	107.32	+0.61%	+6.01%	5.990	111.47	6.015	115.49	+0.42%	+3.61%	5.983	111.80	−0.11%	+0.29%
	96	5.849	125.20	5.808	120.36	+0.70%	+3.87%	6.108	124.21	6.137	126.31	+0.47%	+1.68%	6.100	124.21	−0.14%	0.00%
	192	5.939	116.67	5.878	110.56	+1.03%	+5.24%	6.009	112.04	6.032	116.53	+0.38%	+4.01%	5.993	111.32	−0.26%	−0.64%
	336	6.294	129.72	6.222	123.54	+1.15%	+4.76%	6.222	123.54	6.265	129.16	+0.70%	+4.54%	6.226	124.75	+0.07%	+0.98%
	720	7.015	118.94	6.976	112.20	+0.56%	+5.67%	6.976	112.20	7.007	116.72	+0.45%	+4.03%	6.972	112.54	−0.04%	+0.30%
Electricity Demand (MSE×10^6)	48	0.606	514.91	0.504	469.81	+16.92%	+8.76%	0.513	474.28	0.588	507.28	+14.62%	+6.96%	0.509	471.58	−0.88%	−0.57%
	96	0.831	619.23	0.658	554.13	+20.83%	+10.51%	0.666	558.07	0.797	606.54	+19.57%	+8.69%	0.674	562.18	+1.17%	+0.74%
	192	0.992	704.06	0.804	620.65	+18.98%	+11.85%	0.804	620.65	0.947	687.55	+17.82%	+10.78%	0.811	625.76	+0.91%	+0.82%
	336	1.138	753.84	0.969	695.26	+14.85%	+7.77%	0.969	695.26	1.090	738.16	+12.44%	+6.17%	0.980	700.10	+1.08%	+0.70%
	720	1.182	777.78	1.045	722.82	+11.62%	+7.07%	1.045	722.82	1.140	763.40	+9.09%	+5.61%	1.040	721.67	−0.52%	−0.16%
Traffic (MSE×10^8, MAE×10^4)	48	5.508	1.682	3.273	1.204	+40.57%	+28.46%	3.319	1.193	5.242	1.630	+57.93%	+36.65%	3.495	1.235	+5.29%	+3.55%
	96	6.629	1.890	3.651	1.310	+44.93%	+30.71%	3.393	1.258	5.979	1.793	+76.21%	+42.61%	3.579	1.302	+5.49%	+3.55%
	192	5.749	1.694	3.961	1.347	+31.10%	+20.46%	3.127	1.233	4.575	1.539	+46.32%	+24.81%	3.244	1.259	+3.75%	+2.08%
	336	5.816	1.739	4.081	1.412	+29.82%	+18.82%	4.013	1.401	5.297	1.667	+32.00%	+18.95%	4.016	1.407	+0.07%	+0.41%
	720	7.636	2.014	6.739	1.840	+11.75%	+8.62%	6.739	1.840	7.261	1.967	+7.75%	+6.88%	6.924	1.861	+2.74%	+1.14%
Gas Demand (MSE×10^{13}, MAE×10^6)	48	3.212	4.345	3.022	4.278	+5.94%	+1.53%	3.329	4.623	3.529	4.694	+6.00%	+1.54%	3.363	4.636	+1.03%	+0.27%
	96	4.014	5.077	3.747	4.998	+6.65%	+1.56%	4.092	5.373	4.377	5.466	+6.96%	+1.72%	4.145	5.390	+1.29%	+0.31%
	192	4.441	5.353	3.981	5.200	+10.37%	+2.85%	4.204	5.402	4.691	5.557	+11.58%	+2.86%	4.273	5.422	+1.64%	+0.36%
	336	4.748	5.517	4.356	5.370	+8.26%	+2.66%	4.717	5.595	5.165	5.768	+9.51%	+3.10%	4.790	5.624	+1.56%	+0.53%
	720	5.991	6.276	5.500	6.095	+8.21%	+2.90%	5.733	6.245	6.191	6.402	+8.00%	+2.52%	5.834	6.283	+1.76%	+0.61%

residual components. For the slow channel, we report the Pearson correlation with the injected level shift and the fraction of residual energy assigned to the slow component. For the shape channel, we compute the correlation with the injected zero-mean waveform and its corresponding energy ratio. For the spike channel, we compare the recovered spike locations with the injected locations using precision, recall, and F1 score.

As shown in Table 4, persistent offsets are isolated by the slow channel, zero-mean trajectory distortions are primarily captured by the shape channel, and localized impulses are accurately identified by the spike channel. Importantly, the decomposition remains well separated when all three effects coexist: the mixed setting achieves a slow correlation of 0.999, a shape correlation of 0.966, and a spike F1 of 0.933. These results provide controlled component-level evidence that the slow/shape/spike decomposition corresponds to the intended temporal residual patterns rather than acting as three arbitrary residual branches.

Table 4: Controlled validation of the slow, shape, and spike residual channels at $H = 48$. The same pattern is observed across all four datasets.

Synthetic effect	Validation metric	Result
Level shift	Slow correlation / energy ratio	1.000 / 1.000
Shape distortion	Shape correlation / energy ratio	0.985 / 0.965
Localized spike	Precision / recall / F1	1.000 / 1.000 / 1.000
Mixed effects	Slow corr. / shape corr. / spike F1	0.999 / 0.966 / 0.933

5 Ablation Studies

5.1 Effect of the Delta Stack

We first ablate the proposed delta stack by removing the residual correction module and all news-conditioned modulation heads. This comparison isolates whether the learned residual-and-modulation stack provides predictive value beyond the numerical base, while keeping the look-back window, forecast horizons, data split, and normalization unchanged. Table 2 shows that the full REIN model improves over the frozen base on all four datasets. The largest gain appears on Traffic, where the base leaves substantial structured residuals that the delta module can correct. Electricity Demand also shows clear improvement, suggesting that the residual branch complements the regular calendar-driven load structure captured by DLinear. Electricity Price shows a larger MAE gain than MSE

gain, which is consistent with its heavy-tailed distribution: the delta module improves many typical samples, while MSE remains sensitive to rare price spikes. Gas Demand has the smallest average MAE gain, indicating that much of its predictable structure is already captured by numerical histories. Overall, this ablation supports the need for a residual correction module on top of the frozen base, rather than relying on the base forecaster alone.

5.2 Counterfactual News Assessment

We test whether REIN benefits from structured news rather than indiscriminate text input. On the news-active subset, we compare three regime-pack conditions. *Active* uses the true aligned regime pack, *Blank* replaces the regime pack with zeros, and *Perm* globally shuffles the date-to-regime mapping. This protocol tests whether the model benefits from the presence of structured news content and whether the timing of that content contributes to the prediction. As shown in Table 3, blanking the regime pack consistently worsens performance on news-active samples. The effect is largest on Traffic and Electricity Demand, where event-related signals have relatively direct pathways to the target series. The effect is smaller on Electricity Price and Gas Demand, where the relationship between news and future values is weaker, noisier, or already partly absorbed by the numerical base. These results suggest that REIN's gains are not simply caused by the existence of an additional text embedding. The regime bank must provide useful structured information for the active samples. The date-permutation comparison gives a more nuanced picture. Permuting the regime pack generally degrades performance relative to the active setting, but the gap is smaller than the blanking gap. This indicates that temporal alignment matters, although the current regime bank also captures broader event regimes that may persist across nearby forecast origins. This finding supports the use of aligned news while also pointing to a limitation of the present aggregation strategy: confidence-weighted and time-decayed averaging may smooth away fine-grained intra-day event timing.

5.3 Component-wise Ablation

We evaluate the contribution of the main architectural components by individually removing the active threshold and the slow, shape, and spike residual channels while keeping the numerical backbone, training protocol, and news corpus unchanged. The ablations reveal distinct and dataset-dependent roles for these components. The slow and shape channels are generally more influential at medium and long horizons, particularly for Traffic and Gas Demand, whereas the spike channel has a more selective effect in settings with abrupt event-driven deviations. The active threshold primarily improves robustness by suppressing weak or noisy news activations. Overall, the results support the use of structured residual channels rather than treating the delta module as an undifferentiated correction branch. Full results across all datasets and horizons are provided in Appendix F.

6 Conclusion

We introduced REIN, a news-conditioned forecasting framework that incorporates external text through bounded residual modulation over a frozen numerical base. An LLM agent distills raw news offline into a leakage-controlled regime bank, which modulates structured slow, shape, and spike residual channels through a small set of bounded controls. This design localizes the role of textual evidence while preserving fallback to the numerical forecaster when news is weak or absent. Across four operational datasets, REIN improves forecasting accuracy, while counterfactual blanking and date permutation show that the gains depend on aligned news content rather than generic text availability or calendar correlations.

REIN is most effective when news has a relatively direct effect on the target. Its decay-based aggregation can be less suitable for delayed or heterogeneous event impacts, while bounded modulation may under-correct rare extreme spikes. These limitations point to future work on more flexible temporal aggregation and adaptive modulation bounds. More broadly, our results suggest that reliable text-augmented forecasting depends not only on extracting information from text, but also on controlling how that information enters the forecasting process.

REFERENCES

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. doi: 10.1609/aaai.v37i9.26317.

Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, 2023.

M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

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. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024a. Curran Associates Inc. ISBN 9798331314385.

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

S. Han, S. Lee, M. Cha, S. O. Arik, and J. Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 21774–21797, 2025.

H. Wu, J. Xu, J. Wang, and M. Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, pages 27268–27286, 2022.

Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long. iTransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2024.

A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and Y. Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin. One fits all: Power general time series analysis by pretrained lm. In *Advances in Neural Information Processing Systems*, 2023.

T. Loughran and B. McDonald. Textual analysis in finance. *Annual Review of Financial Economics*, 12:357–375, 2020. doi: 10.1146/annurev-financial-012820-032249.

X. Wang, J. Gu, J. Qiu, G. Liu, X. Cai, J. Zhu, Y. Liu, Z. Dong, and J. Zhao. News-driven load forecasting: Generative agents and large language models for unstructured data and event analysis. *Engineering*, 2026. doi: 10.1016/j.eng.2026.02.031.

B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Australian Energy Market Operator. Aggregated price and demand data. [Online]. Available: https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/aggregated-data. Accessed: May 5, 2026.

NSW Government. Road traffic counts hourly sample. [Online]. Available: https://opendata.transport.nsw.gov.au/data/dataset/ef2b0bd2-db1e-48f3-9ea1-2bb9e6bc6504/resource/b8bd742f-af37-4999-887f-6d3657349d13. Accessed: May 5, 2026.

ENTSOG. Entsog. [Online]. Available: https://www.entsog.eu/. Accessed: May 5, 2026.

H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024b.

Y. Wang, H. Wu, J. Dong, Y. Liu, Y. Qiu, H. Zhang, J. Wang, and M. Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 2024c.

A. F. Ansari, O. Shchur, J. Küken, A. Auer, B. Han, P. Mercado, S. S. Rangapuram, H. Shen, L. Stella, X. Zhang, M. Goswami, S. Kapoor, D. C. Maddix, P. Guerron, T. Hu, J. Yin, N. Erickson, P. M. Desai, H. Wang, H. Rangwala, G. Karypis, Y. Wang, and M. Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Y. Liu, G. Qin, Z. Shi, Z. Chen, C. Yang, X. Huang, J. Wang, and M. Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

C. Wang, Q. Qi, J. Wang, H. Sun, Z. Zhuang, J. Wu, L. Zhang, and J. Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12694–12702, 2025. doi: 10.1609/aaai.v39i12.33384.

S. Zhong, W. Ruan, M. Jin, H. Li, Q. Wen, and Y. Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 78478–78497, 2025.

RenewEconomy. News and analysis for the clean energy economy. [Online]. Available: https://reneweconomy.com.au/. Accessed: May 5, 2026.

pv magazine Australia. Photovoltaic markets and technology. [Online]. Available: https://www.pv-magazine-australia.com/. Accessed: May 5, 2026.

WattClarity. Commentary and analysis of australia's nem. [Online]. Available: https://wattclarity.com/. Accessed: May 5, 2026.

Webhose. Free-news-datasets: Weekly free datasets from global news sites. [Online]. Available: https://github.com/Webhose/free-news-datasets. Accessed: May 5, 2026.

The GDELT Project. The gdelt project. [Online]. Available: https://www.gdeltproject.org/. Accessed: May 5, 2026.

A LLM USAGE

Large language models were used in this work for the following purposes:

- **News refinement:** An LLM agent was used to convert raw news records into structured regime records, including actionability flags, topic tags, regime vectors, influence horizons, confidence scores, and summaries.
- **News filtering:** News records judged to be retrospective, routine, or irrelevant to the target series were marked as non-actionable and excluded from the active regime bank.
- **Text embedding:** News record summaries were encoded with E5-small-v2 to provide compact textual representations for the regime bank.
- **Manuscript polishing:** LLM assistance was used to correct grammar and spelling during manuscript preparation.

B DATASET DESCRIPTION

Electricity Demand. 2024 electricity demand from the New South Wales (NSW) region of the Australian National Electricity Market, sourced from AEMO(Australian Energy Market Operator). The raw series is recorded at five-minute resolution and we aggregate it into 30-minute intervals by arithmetic mean within each interval to obtain a consistent forecasting granularity. Demand exhibits strong calendar regularity (weekday/weekend, morning/evening peaks) and is highly weather-sensitive, making it a representative case for short-horizon event-related signals. The matched news corpus contains 4,617 news records, spans late-2023 through early-2025 with >99% dated in 2024, and covers heatwave/cold-snap warnings, generator outages, renewable build-out, and weekday demand commentary that can provide relevant context for next-day load. Sources include RenewEconomy(RenewEconomy) ($\sim$2,040 news records, $\sim$44%), PV-Magazine Australia(pv magazine Australia) ($\sim$1,420, $\sim$31%), WattClarity(WattClarity) ($\sim$740, $\sim$16%), and Webz.io News Dataset RepositoryWebhose. The downsampling algorithm can be found in Appendix C.

Electricity Price. Spot electricity price for the same NSW region and 2024 horizon, sourced from AEMO and aggregated from 5-minute to 30-minute resolution. Unlike demand, the price series is extremely heavy-tailed and can contain short-lived five-minute spikes. We therefore aggregate the raw 5-minute price observations into a robust 30-minute target by taking the median within each interval (see Appendix C). This choice reduces the dominance of isolated dispatch-interval outliers while retaining persistent high-price regimes and broader price movements relevant to short-horizon forecasting. The matched news corpus is the price-side counterpart of the load feed, collected from the same sources as Electricity Demand. It contains 4,645 news records over the same time span and emphasizes price-relevant content (price caps, market notices, fuel costs, transmission constraints, wholesale market commentary) on top of the demand-side context shared with the load corpus.

Traffic. Hourly traffic counts for 2024 from NSW permanent count stations, aggregated into a single state-level hourly series. The series shows strong daily and weekly seasonality with suppressed-volume periods around public holidays and severe-weather events. For news, we pair the series with the open GDELT 2024 global news corpus (The GDELT Project). It contains 11,446 news records published in 2024, with very broad topical coverage, which makes it challenging for REIN: any gain over the base must be extracted by the LLM-refinement stage from a noisy, off-topic stream rather than from a hand-picked feed.

Gas Demand. Hourly Netherlands natural gas consumption for 2025, sourced from ENTSOG (ENTSOG), with strong seasonal and day-of-week patterns driven by heating demand and industrial offtake. The matched news corpus is a GDELT-derived stream which contains 16,166 news records from 2025 and is bilingual (Dutch dominant, English secondary). Relative to other datasets' feeds, this corpus is denser but topically broader, exercising the regime bank's confidence-weighted aggregation under a higher noise floor.

C Downsampling Procedures

The raw NSW electricity price and load series are recorded at five-minute resolution. We downsample both series to 30-minute resolution to match the forecasting granularity used in our experiments. Since price and load have different statistical properties, we use different aggregation rules. Electricity price is aggregated by the within-window median to reduce the effect of isolated dispatch-interval spikes, while electricity load is aggregated by the arithmetic mean to preserve the average system demand over each half-hour interval.

Table 5: Downsampling procedures for converting five-minute NSW electricity price and load records into 30-minute series.

Series	Procedure	Rationale
Electricity Price	For each 30-minute interval, collect the six five-minute price observations within the interval and compute $$y_k^{\text{price}} = \text{median}\{p_t : t \in \mathcal{B}_k\}.$$	The price series is heavy-tailed and can contain isolated dispatch-interval spikes. Median aggregation provides a robust half-hourly target and reduces the dominance of single five-minute outliers.
Electricity Load	For each 30-minute interval, collect the six five-minute load observations within the interval and compute $$y_k^{\text{load}} = \frac{1}{\lvert\mathcal{B}_k\rvert} \sum_{t \in \mathcal{B}_k} d_t.$$	Load is a continuous system-level demand quantity. Mean aggregation preserves the average demand over each half-hour interval and is therefore appropriate for representing system demand at the forecasting granularity.

D Dataset and News Corpus Statistics

Series characteristics and splits. Table 6 summarizes series frequency, units, tail severity, base backbone, and news source. Table 7 reports the chronological train/val/test splits used for every method (no test-set leakage into bank construction). All splits follow a roughly 70/10/20 chronological partition aligned to a common cut at *day 257* (mid-September) and *day 293* (mid-October) of the calendar year.

Table 6: Dataset characteristics. Tail severity is moderate (near-Gaussian) for load/traffic/gas and heavy (extreme right skew with $\sigma \approx 635\,\$/\text{MWh}$) for price.

Dataset	Freq.	H	Unit	Tail	Base	News source
Electricity Demand	30 min	48–720	MW	Moderate	DLinear	RenewEconomy, PV-Mag, WattClarity, Webz.io
Electricity Price	30 min	48–720	$/MWh	Heavy	DLinear	Same as Electricity Demand (price-side feed)
Traffic	1 hour	48–720	vehicles	Moderate	DLinear	GDELT 2024 global news
Gas Demand	1 hour	48–720	m^3	Moderate	DLinear	GDELT 2025

Table 7: Train/validation/test splits. Counts are observations at the native sampling frequency (excluding header row). Forecast windows per split with L=48, stride 1, horizon H are $N_{\text{win}} = \max(0, \text{rows} - L - H + 1)$.

Dataset	Year	Train			Val			Test	
		#obs	start	end	#obs	start	end	#obs	end
Electricity Demand	2024	12,294	Jan 01 00:30	Sep 13 03:00	1,763	Sep 13 03:30	Oct 19 20:30	3,511	Jan 01 2025
Electricity Price	2024	12,294	Jan 01 00:30	Sep 13 03:00	1,763	Sep 13 03:30	Oct 19 20:30	3,511	Jan 01 2025
Traffic	2024	6,147	Jan 01 00:00	Sep 13 02:00	881	Sep 13 03:00	Oct 19 19:00	1,756	Dec 31 23:00
Gas Demand	2025	5,928	Jan 01 06:00	Sep 13 06:00	850	Sep 13 07:00	Oct 19 16:00	1,694	Dec 31 05:00

News corpora and LLM refinement output. Table 8 reports per-corpus refinement statistics: total news, fraction marked actionable by the LLM, mean self-reported confidence and in-force horizon,

and the most frequent topic tags (out of 15). Actionable rates are highest on the curated NSW price feed (19.0%, where the corpus is already topically filtered) and lowest on the Gas Demand stream (1.8%, dominated by long-horizon LNG-infrastructure news that the prompt down-weights). Table 9 reports the resulting time-aligned regime bank: the fraction of forecast-origin bins with non-zero relevance mass, the fraction above the per-dataset activity threshold θ, the mean number of in-force news per active bin, and aggregate $m(t)$ statistics.

Table 8: Per-corpus LLM refinement statistics. Confidence and horizon are over the actionable subset only.

Corpus	news	Actionable	%	$\bar{c}$	$\bar{h}$ (d)	Top topic tags (counts)
Electricity Demand (curated)	4,617	546	11.8%	0.78	7.9	policy_active 176, supply_tight 154, outage 120, renewable_surge 109
Electricity Price (curated)	4,645	881	19.0%	0.71	5.4	supply_tight 310, other 292, renewable_surge 197, outage 191
Traffic (GDELT 2024)	11,446	1,189	10.4%	0.60	6.2	supply_tight 779, outage 470, holiday 423, heatwave 198
Gas Demand (GDELT 2025, NL)	16,166	284	1.8%	0.59	5.3	cold_snap 94, other 75, supply_surplus 61, fuel_shock 36

Table 9: Time-aligned regime-bank coverage. Bank steps are forecast-origin bins at the native frequency; "active" uses the per-dataset θ from Table 11. *docs/active* is the mean number of in-force news within an active bin's h_i-day window.

Dataset	θ	Bank steps	$m>0$	Active	$\bar{m}$	med m	max m	docs/active
Electricity Demand	0.7	17,568	100.0%	96.6%	3.63	2.84	22.6	11.95
Electricity Price	0.7	17,568	99.7%	97.5%	4.16	3.57	22.4	13.18
Traffic	0.2	8,784	100.0%	92.4%	3.18	2.72	13.8	21.00
Gas Demand	0.2	8,736	97.7%	84.0%	1.24	1.02	7.8	4.75

The high active percentage on Electricity Demand and Electricity Price reflects continuous coverage of the curated feeds throughout 2024; the lower coverage on Gas Demand reflects the GDELT keyword filter producing sparser actionable signals around heating-season events. The **news-active subset** reported in the body counterfactual table (Tab. 4) is the test-set restriction of the **Active** column to the test window only and is therefore strictly smaller than the bank-level percentage.

E Rationale and Auditability of the Regime Schema

The regime schema is designed as a compact and auditable interface between unstructured news and the numerical forecasting model. Each refined news record is mapped into a fixed set of fields, including an actionability flag, topic tags, a five-axis regime vector, an influence horizon, a confidence score, a concise summary, and a text embedding. This design avoids using raw LLM outputs as unconstrained forecasting signals.

The five-axis regime vector is chosen to capture broad mechanisms through which external events can affect operational time series. The *tightness* axis represents supply–demand pressure or capacity scarcity. The *demand* axis captures expected changes in usage intensity. The *supply* axis captures resource availability, substitution effects, or supply-side relief. The *volatility* axis captures uncertainty, disruption risk, and spike-prone conditions. The *policy* axis marks active regulatory, market, or operational interventions. Although the domain-specific interpretation of each axis differs across electricity price, electricity demand, traffic, and gas demand, the shared schema provides a consistent representation for cross-domain news refinement.

The schema is intentionally low-dimensional. A larger free-form representation would make the LLM output harder to inspect and easier to overfit, while a smaller representation may fail to separate demand pressure, supply conditions, volatility risk, and policy effects. The five-axis vector therefore serves as a structured bottleneck: it is expressive enough to represent common event mechanisms, but compact enough to support controlled aggregation and bounded residual modulation. More specific record-level information is retained through topic tags, summaries, and embeddings, while the regime vector provides the main structured signal used by the delta module.

Table 10: Full component-wise ablation results across four datasets and five forecasting horizons. Each variant removes one component from the complete REIN architecture while keeping all other settings unchanged. Lower MSE and MAE are better.

To reduce black-box behaviour, the LLM agent is used only as an information extractor. It does not generate forecasts and does not directly control the final prediction. Its output is restricted by a fixed JSON schema with bounded numerical fields and predefined semantic constraints. news that are retrospective, routine, or unrelated to the target are marked as non-actionable and excluded from the active regime bank. For actionable news, the confidence score and influence horizon determine how strongly and how long the news record contributes to the time-aligned regime representation.

This design makes the LLM refinement process inspectable before it affects the forecasting model. Each intermediate record can be logged, checked, and traced back to its source news record. The downstream forecaster only receives aggregated regime signals, whose influence is further limited by activity gates and bounded modulation coefficients. The Active/Blank/Permuted counterfactual protocol then evaluates whether these LLM-derived regime signals improve predictions when aligned with the target time window, rather than merely reflecting the presence of news or calendar-correlated availability.

F Full Component-wise Ablation Results

Table 10 reports the complete component-wise ablation results across all four datasets and five forecasting horizons. We compare the full REIN model with four variants obtained by removing the active threshold, slow channel, shape channel, or spike channel individually. All other components, including the numerical backbone, training protocol, and news corpus, are kept fixed.

The results show that the contribution of each component varies with the temporal characteristics of the dataset and forecasting horizon. The slow and shape channels exhibit their clearest effects at medium and long horizons. This behavior is particularly evident on Traffic, where removing either channel leads to substantial deterioration at longer horizons, indicating that persistent residual shifts and trajectory-level corrections become increasingly important as the forecasting horizon grows.

The spike channel exhibits a more selective pattern. Its removal can substantially degrade performance in event-sensitive settings, while its effect is smaller or occasionally neutral in datasets where abrupt residual events are less informative. This is consistent with the intended role of the spike branch as a specialized mechanism for short-lived and extreme deviations rather than a uniformly active correction component.

The active-threshold ablation produces comparatively smaller but more consistent changes, supporting its role as a robustness mechanism rather than a primary predictive branch. By suppressing weak regime activations, the threshold limits the influence of low-confidence or noisy news signals.

Taken together, the ablations indicate that the residual channels are complementary rather than uniformly beneficial across all datasets and horizons. Their relative importance depends on the temporal structure of the forecasting problem: the slow and shape channels primarily capture persistent and trajectory-wide corrections, while the spike channel provides a specialized pathway for abrupt event-driven deviations.

G Evaluation Metrics

We evaluate forecasting performance using mean squared error (MSE) and mean absolute error (MAE). Both metrics are computed after denormalizing the model outputs back to the original scale of each dataset. Given a test set with N forecasting windows, a prediction horizon H, ground-truth values $y_{i,h}$, and predicted values $\hat{y}_{i,h}$, MSE and MAE are defined as

$$\mathrm{MSE} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} \left(y_{i,h} - \hat{y}_{i,h} \right)^2 ,$$

$$\mathrm{MAE} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{h=1}^{H} |y_{i,h} - \hat{y}_{i,h}| \,.$$

MSE penalizes large errors more strongly because the error term is squared, making it sensitive to rare but severe forecasting mistakes such as electricity-price spikes. MAE measures the average absolute deviation and is therefore more robust to extreme values. Reporting both metrics allows us to evaluate not only typical forecasting accuracy, but also the extent to which each model controls large errors in heavy-tailed operational time series.

H Hyperparameter Details

This section reports the implementation-level hyperparameters used for REIN across all datasets. We include both the base-stage settings and the delta-stage settings, together with the news-modulation coefficients, activity thresholds, spike-target construction parameters, and regime-bank configuration. These values are fixed for each dataset and are used for all reported horizons unless otherwise stated.

Table 11 lists the complete hyperparameter settings used to produce every REIN result in the body.

Table 11: Complete hyperparameter settings across datasets. Backbone choice is per dataset; all delta modules share the same PatchTST encoder.

Parameter	**Demand**	**Traffic**	**Gas Demand**	**Price**
Base backbone	DLinear	DLinear	DLinear	DLinear
Base epochs / lr	40 / 10^{-3}	100 / 10^{-3}	40 / 10^{-3}	40 / 10^{-3}
Early-stop patience	5	5	5	5
Delta hidden / layers / heads	128/2/4	128/2/4	128/2/4	128/2/4
Patch len / stride	8/4	8/4	8/4	8/4
Delta lr / epochs	10^{-4}/30	10^{-4}/50	5×10^{-6}/30	5×10^{-6}/30
Pretrain epochs / lr	12/10^{-3}	12/10^{-3}	12/10^{-3}	12/10^{-3}
Warmup / Grad clip	5%/1.0	5%/1.0	5%/1.0	5%/1.0
$\lambda_{\min}/\lambda_{\text{ts}}^{\text{cap}}/\lambda_{\text{news}}^{\text{cap}}/\lambda_{\max}$	.05/.30/.12/.45	.05/.30/.12/.60	.05/.30/.12/.45	.05/.30/.12/.45
γ_{shape} / γ_{spike}	0.20/0.75	0.70/0.75	0.20/0.75	0.20/0.75
θ (active mass)	0.7	0.2	0.2	0.7
Spike k / target %	3.0/10%	3.0/10%	3.0/10%	4.0/6%
News blank prob	30%	30%	30%	30%
Text encoder	E5-sm-v2	E5-sm-v2	E5-sm-v2	E5-sm-v2
Regime bank freq.	30 min	1 hour	1 hour	30 min
News corpus	Curated NSW	GDELT 2024	GDELT 2025 (NL gas)	Curated NSW

I Training Details and Counterfactual Protocol

This appendix specifies the loss decomposition referenced by Eq. 5, the news-dropout regularizer, the optimizer and scheduler configuration, and the per-window counterfactual evaluation protocol.

Notation. Within a mini-batch, let $\boldsymbol{e}^{\star} \in \mathbb{R}^{B\times H}$ be the true z-space residual of the frozen base, decomposed into $(\boldsymbol{e}^{\star}_{\text{slow}}, \boldsymbol{e}^{\star}_{\text{shape}}, \boldsymbol{e}^{\star}_{\text{spike}})$ via Eq. 10 (Appendix N), and let $\mathbb{S} \in \{0,1\}^{B\times H}$ be the corresponding spike mask. The model produces final z-space prediction $\hat{\boldsymbol{y}}_z$, three TS-head outputs $(\hat{r}_{\text{slow}}, \hat{\boldsymbol{r}}_{\text{shape}}, \hat{\boldsymbol{r}}_{\text{spike}})$, spike-gate logits $\hat{\boldsymbol{\ell}}_{\text{spike}}$ with sigmoid gate $\hat{g}_{\text{spike}}{=}\sigma(\hat{\boldsymbol{\ell}}_{\text{spike}})$, and the gated spike contribution $\hat{\boldsymbol{r}}_{\text{spike}} \odot \hat{g}_{\text{spike}}$. The active mask $A(t){=}\mathbb{I}[m(t){>}\theta]$ partitions the batch into news-active and news-inactive samples.

Loss decomposition. Here, the overline denotes averaging over all batch and horizon positions, i.e., $\overline{\hat{g}}_{\text{spike}} = \frac{1}{BH}\sum_{b=1}^{B}\sum_{h=1}^{H} \hat{g}_{\text{spike},b,h}$ and $\overline{\mathbb{S}} = \frac{1}{BH}\sum_{b=1}^{B}\sum_{h=1}^{H} \mathbb{S}_{b,h}$. The total delta-stage loss

expands Eq. 5 as

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{point}} &+ w_s \cdot \mathrm{MSE}(\hat{r}_{\text{slow}},\, e^{\star}_{\text{slow}}) + w_q \cdot \mathrm{SL1}(\hat{\boldsymbol{r}}_{\text{shape}},\, \boldsymbol{e}^{\star}_{\text{shape}}) \\ &+ w_k \cdot \Big(\tfrac{\sum \mathbb{S} \odot \mathrm{SL1}(\hat{g}_{\text{spike}} \hat{\boldsymbol{r}}_{\text{spike}}, \boldsymbol{e}^{\star}_{\text{spike}})}{\max(1, \sum \mathbb{S})} + w_g \cdot \big[\mathrm{BCE}(\hat{\boldsymbol{\ell}}_{\text{spike}}, \mathbb{S}) + \mathrm{SL1}(\overline{\hat{g}_{\text{spike}}},\, \overline{\mathbb{S}}) \big] \Big) \\ &+ w_c \cdot \mathbb{E}_{A=0}\big[\|\hat{\boldsymbol{y}}_z^{\text{normal}} - \hat{\boldsymbol{y}}_z^{\text{blank-all}}\|_2^2 \big] + w_f \cdot \mathcal{L}_{\text{counter}} \\ &+ w_i \cdot \mathbb{E}_{A=0}\big[\|\hat{\boldsymbol{y}}_z - \hat{\boldsymbol{y}}_z^{\text{base}}\|_2^2 \big] + w_d \cdot \mathbb{E}_{A=1}\big[1 - \cos\big(\hat{\boldsymbol{y}}_z - \hat{\boldsymbol{y}}_z^{\text{base}},\, \boldsymbol{e}^{\star}\big) \big], \end{aligned} \tag{6}$$

where SL1 is Huber/Smooth-L1 with δ=1 and BCE is binary cross-entropy with logits. $\mathcal{L}_{\text{point}}$ is Smooth-L1 on $(\hat{\boldsymbol{y}}_z, \boldsymbol{y}_z)$ for non-price datasets; for Electricity Price it is a winsorized Smooth-L1 plus a 0.1-weighted three-point pinball loss at quantiles $\{0.1, 0.5, 0.9\}$ in raw \$/MWh, rescaled by the global z-score σ to keep the auxiliary z-space terms commensurable. $\mathcal{L}_{\text{counter}}$ is a two-margin hinge,

$$\mathcal{L}_{\text{counter}} = \mathbb{E}_{A=1 \,\wedge\, \text{not blanked}} \Big[\big(\|\hat{\boldsymbol{y}}_z - \boldsymbol{y}_z\|_1 - \|\hat{\boldsymbol{y}}_z^{\text{blank}} - \boldsymbol{y}_z\|_1 + \delta \big)_+ + \big(\|\hat{\boldsymbol{y}}_z - \boldsymbol{y}_z\|_1 - \|\hat{\boldsymbol{y}}_z^{\text{perm}} - \boldsymbol{y}_z\|_1 + \delta \big)_+ \Big], \tag{7}$$

with margin δ=0.02 in z-space and the permutation seed resampled at every batch (training only; test uses a held-out fixed seed, see below).

Loss weights. Default weights, used for every dataset unless noted, are

$$w_s{=}1.0, \quad w_q{=}1.0, \quad w_k{=}1.0, \quad w_g{=}0.25, \quad w_c{=}0.05, \quad w_f{=}0.10, \quad w_i{=}0.10, \quad w_d{=}0.05.$$

News dropout, optimizer, and schedule. At each training step, with probability `news_blank_prob`=0.30 each window's regime pack is replaced by zeros; this dropout is the regularizer that lets the consistency, counterfactual, and inactive-residual losses sample inactive-like behavior on news-rich batches. Training uses AdamW with β_1=0.9, β_2=0.999, weight decay 10^{-5}, gradient clipping 1.0, batch size 16. The learning-rate schedule is linear warmup over 5% of total steps to the peak rate (Tab. 11), then cosine decay to 5% of the peak. The base stage trains identically with cosine decay and early stopping patience 5 on validation MAE.

Pre-training of news-conditioned modulation heads. Before joint delta training, the three news-conditioned modulation heads ($\Delta\lambda_{\text{news}}$, g_{shape}, b_{spike}) are pretrained for 12 epochs at LR 10^{-3} on a self-supervised target derived from $\bar{\boldsymbol{r}}(t)$, $\bar{\boldsymbol{\tau}}(t)$, and $m(t)$. This pretraining warm-starts the modulation knobs to non-trivial values so that the joint training does not get stuck at the identity initialization $\Delta\lambda_{\text{news}}{\equiv}0$, $g_{\text{shape}}{\equiv}1$, $b_{\text{spike}}{\equiv}0$.

Counterfactual test protocol. Algorithm 1 formalises the per-window protocol behind Tables 3 and 4 of the body. The permutation bank is built once at test time by globally shuffling the (date $\to$ regime) mapping with seed 2024 (held out from the training-time seed); this is the sole source of randomness in the test evaluation.

Algorithm 1 Counterfactual evaluation of REIN on the test set.

Require: frozen base f_{base}, trained delta module, regime bank $\mathcal{B}$, test windows $\{(\boldsymbol{h}_z(t), \boldsymbol{y}_z(t))\}_{t \in \mathcal{T}}$, threshold θ, permutation seed s=2024.

1: $\mathcal{B}_{\text{perm}} \leftarrow$ shuffle the date-axis of $\mathcal{B}$ with $\mathrm{rng}(s)$ (regime_vec, topic_tag_mass, text_emb, relevance_mass, in_force_doc_count are all permuted by the same index).
2: **for** each test window t **do**
3: $\quad \mathcal{B}(t) \leftarrow$ time-aligned regime pack at t; $\quad \mathcal{B}_{\text{perm}}(t) \leftarrow$ pack at t from $\mathcal{B}_{\text{perm}}$.
4: $\quad \hat{\boldsymbol{y}}^{\text{Active}} \leftarrow \text{REIN}\big(\boldsymbol{h}_z(t), \mathcal{B}(t)\big)$; $\hat{\boldsymbol{y}}^{\text{Blank}} \leftarrow \text{REIN}\big(\boldsymbol{h}_z(t), \mathbf{0}\big)$; $\hat{\boldsymbol{y}}^{\text{Perm}} \leftarrow \text{REIN}\big(\boldsymbol{h}_z(t), \mathcal{B}_{\text{perm}}(t)\big)$.
5: $\quad A(t) \leftarrow \mathbb{I}[m(t) > \theta]$.
6: **end for**
7: Report MAE/MSE on **Full test** and on the *news-active* subset $\{t : A(t){=}1\}$ and the *news-inactive* subset $\{t : A(t){=}0\}$.
8: $\Delta_{\text{Blank}}{=}\mathrm{MAE}(\hat{\boldsymbol{y}}^{\text{Blank}}) - \mathrm{MAE}(\hat{\boldsymbol{y}}^{\text{Active}})$ on the active subset (news-blanking gain).
9: $\Delta_{\text{Perm}}{=}\mathrm{MAE}(\hat{\boldsymbol{y}}^{\text{Perm}}) - \mathrm{MAE}(\hat{\boldsymbol{y}}^{\text{Active}})$ on the active subset (date-permutation gain).

Validation-time selection. We select the best epoch on the validation MAE constrained by two news-attribution checks: real-news active MAE must beat (a) blanked-active MAE or (b) permuted-active MAE by at least ϵ_2=0.01 in raw MAE units; epochs that fail both checks are not eligible to be the saved checkpoint. This prevents selecting an epoch whose validation gain is driven by reductions on news-inactive windows alone.

J LLM REFINEMENT PROMPT AND PER-DOMAIN SCHEMA

To construct the news regime bank, we use a dataset-specific LLM refinement prompt for each target domain. All variants share the same output schema and validity constraints, while the definition of "actionable" news and the semantic interpretation of the regime dimensions are adapted to the target series.

Shared Output Schema. For every news record, the LLM is instructed to return strict JSON with the following fields:

```
{is_actionable: boolean, topic_tags: string[],
regime_vec: {tightness: float, demand_outlook: float,
renewable_surplus: float, volatility_tone: float,
policy_in_effect: float}, horizon_days: int, confidence:
float, summary: string}.
```

The fields `tightness`, `demand_outlook`, and `renewable_surplus` are constrained to $[-1, 1]$, `volatility_tone` to $[0, 1]$, `policy_in_effect` to $\{0, 1\}$, and `horizon_days` to $[0, 14]$. The summary is required to be factual and under 60 words. news tagged as `retrospective` or `routine` are forced to be non-actionable.

Prompt Design Principle. The prompts favor recall over precision for borderline cases: if a news record has a plausible near-term pathway to the target, it is kept actionable with lower confidence rather than discarded. This design allows the downstream regime bank to aggregate weak signals while using confidence weighting, temporal decay, and active-mass gating to suppress noise.

Schema validation and decoding policy. For every news record we send a system prompt + user prompt pair (full prompts below) to GPT-5.1 with `temperature`=0 and `max_tokens`=260. The response is parsed as JSON; any of the following force a non-actionable record without a retry: missing keys, type mismatches, range violations on the regime axes, `topic_tags` containing an out-of-vocabulary tag, or `topic_tags` that contain `retrospective` or `routine` (which the prompt mandates be non-actionable). Refined records are cached per (corpus, schema variant) at `_shared_refine_cache/v4/refined_*.jsonl` keyed by SHA-1 of `(title,date,url)`, so identical news are not re-queried across runs. The empirical actionable rates and confidence statistics on each corpus are reported in Table 8.

Time-aligned regime bank construction. Algorithm 2 formalises the aggregation that turns a refined-news-record stream into the time-aligned bank used for retrieval at training and test time. The two design choices that prevent same-day textual leakage on sub-daily forecasts are: (i) a date-only news record (no published time) is shifted to the next day's first bin under the sub-daily frequency (line 2); (ii) every aggregation step uses only news with $\text{published_at} \leq t$ (line 8). These two together ensure the bank is causal at the native series frequency.

Table 12: Dataset-specific prompt semantics used for LLM-based news refinement. Each prompt uses the same JSON schema, but adapts the actionable-event definition and regime-vector interpretation to the forecasting target.

Dataset	Actionable News Definition	Regime-Vector Interpretation
Electricity Price	Current or near-term news that can plausibly affect NSW wholesale price level or volatility within 14 days, including generator outages, deratings, maintenance, rebidding, interconnector constraints, fuel shocks, heat-driven demand peaks, renewable droughts or surges, reserve scarcity, market notices, and market interventions.	`tightness`: near-term NSW supply-demand tightness and scarcity pressure on price. `demand_outlook`: load and cooling-demand pressure. `renewable_surplus`: renewable abundance or import relief that suppresses price; negative values indicate renewable scarcity. `volatility_tone`: spike, rebidding, outage, constraint, or scarcity-driven volatility risk. `policy_in_effect`: active policy or market intervention.
Gas Demand	Current or near-term news affecting Dutch or Northwest European gas demand, gas-market tightness, storage stress, balancing conditions, or gas-for-power burn, including cold snaps, warm spells, storage depletion or refill stress, LNG or pipeline disruptions, import bottlenecks, supply-security alerts, and power-system conditions that alter gas burn.	`tightness`: Dutch/NW European gas-market tightness and scarcity pressure. `demand_outlook`: weather-sensitive heating demand and gas-for-power demand. `renewable_surplus`: conditions reducing gas demand or gas burn; negative values indicate weak renewables or substitution constraints. `volatility_tone`: uncertainty from storage stress, outages, balancing tension, or supply shocks. `policy_in_effect`: active storage mandate, emergency action, or market intervention.
Traffic	Current or near-term news that can affect road traffic volume, congestion, route availability, commuter behavior, freight flows, or hourly traffic volatility, including holidays, major events, protests, road closures, crashes with lingering disruption, transit outages, severe weather, evacuations, fuel shocks, freight bottlenecks, congestion charging, and traffic-control interventions.	`tightness`: road-network tightness, congestion pressure, and capacity scarcity. `demand_outlook`: travel demand from commuting, freight, holidays, or major events. `renewable_surplus`: mobility relief or alternate-capacity availability reducing road usage; negative values indicate weak relief and higher road usage. `volatility_tone`: disruption risk from closures, weather, protests, strikes, or events. `policy_in_effect`: active restriction, emergency measure, or traffic intervention.
Electricity Demand	Current or near-term energy news that can affect NSW load within 14 days. The prompt emphasizes currently in-force conditions and filters out retrospective reports, routine commentary, and news without a plausible near-term demand signal.	`tightness`: near-term system tightness relevant to demand-side pressure. `demand_outlook`: expected load pressure, especially weather-sensitive demand. `renewable_surplus`: conditions that may relieve net demand pressure. `volatility_tone`: uncertainty or disruption risk affecting load shape. `policy_in_effect`: active policy or operational intervention.

Algorithm 2 Build time-aligned regime bank from refined news.

Require: refined records $\{r_i\}_{i=1}^{N}$ with fields $(\text{published_at}, h_i, c_i, \boldsymbol{r}_i, \boldsymbol{\tau}_i, \text{summary}_i)$, native frequency f (minutes), bank window $[t_0, t_1]$, decay τ=5 d, EMA α=0.5, encoder ϕ.
Drop r_i with $h_i \le 0$ or $c_i \le 0$ or is_actionable=false.
if $f < 1440$ and r_i has date-only published_at **then**
  $\text{published_at}_i \leftarrow \text{floor}(\cdot, \text{day}) + 1\text{d}$ ▷ date-only news record available next day for sub-daily
end if
$\boldsymbol{e}_i \leftarrow \phi(\text{summary}_i)$ via E5-small-v2 (mean-pool, prefix "passage:").
Build $T \leftarrow \text{date_range}(t_0, t_1, \text{step}=f)$.
for $t \in T$ **do**
  $W_t \leftarrow \{(i, w_i) : \text{published_at}_i \le t \le \text{published_at}_i + h_i\,\text{d},\ w_i = c_i \cdot \exp(-\text{age}_i(t)/\tau)\}$
  if W_t empty **then continue**
  end if
  $m(t) \leftarrow \sum_{(i,w)} w$; $Z \leftarrow \max(10^{-6}, m(t))$.
  $\bar{\boldsymbol{r}}(t) \leftarrow \sum (w/Z)\boldsymbol{r}_i$; $\bar{\boldsymbol{\tau}}(t) \leftarrow \sum (w/Z)\boldsymbol{\tau}_i$; $\bar{\boldsymbol{e}}(t) \leftarrow \sum (w/Z)\boldsymbol{e}_i$.
end for
Apply causal EMA with α to $\bar{\boldsymbol{r}}, \bar{\boldsymbol{\tau}}, \bar{\boldsymbol{e}}, m$ along T.
return bank $\{(t, \bar{\boldsymbol{r}}(t), \bar{\boldsymbol{\tau}}(t), \bar{\boldsymbol{e}}(t), m(t))\}_{t \in T}$.

Full per-domain prompts. Each prompt is the system message paired with a user message of the form `published_at: <date>\n title: <title>\n content: <body>\n Return one JSON object.` The news article body is truncated to 220 words before being injected.

The four domain-specific system prompts are reproduced verbatim below; the closed topic vocabulary (15 tags) and the JSON schema are shared and listed in Table 12.

[Electricity Demand / Load]

```
You are refining generic energy news for NSW electricity load forecasting
into strict JSON. Return JSON only, no markdown.  The key question is
whether the article describes a currently in-force market condition
that can still matter within the next 14 days.  If the article is
retrospective, routine commentary, historical recap, or not currently
actionable, set is_actionable=false.  Use only the closed topic
vocabulary:  <TAGS>.  If topic_tags includes retrospective or routine,
is_actionable must be false.  Return fields exactly:  {is_actionable:boolean,
topic_tags:string[], regime_vec:{tightness:float, demand_outlook:float,
renewable_surplus:float, volatility_tone:float, policy_in_effect:float},
horizon_days:int, confidence:float, summary:string}.  tightness,
demand_outlook, renewable_surplus are in [-1,1].  volatility_tone is in
[0,1].  policy_in_effect is 0 or 1.  horizon_days is an integer in [0,14].
summary must stay under 60 words and be factual.
```

[Electricity Price]

```
You are refining generic energy news for NSW electricity price forecasting
into JSON. Return JSON only, no markdown.  Favor recall over precision
for is_actionable.  If an article has any plausible near-term pathway to
move NSW wholesale price level or volatility, keep it actionable even if
the impact is indirect, probabilistic, or only moderate.  For price, keep
near-term operational and market-microstructure drivers, not just broad
macro supply-demand themes.  Treat the following as potentially actionable if
they are current, imminent, recently announced and still relevant, or likely
to shape trader expectations within the next 14 days:  generator outages,
deratings, planned maintenance, forced outages, unit returns, rebidding
or bidding behavior, interconnector flows or constraints, transmission
limits, fuel supply or spot-price shocks, heatwaves and cooling-demand peaks,
renewable drought or renewable surges, reserve scarcity, market notices,
operational advisories, and market interventions.  Do not filter out an
article just because it is operational, plant-specific, scheduled rather than
forced, framed as a market notice or advisory, or about another connected NEM
region that can move NSW prices through interconnectors, shared weather, or
fuel markets.  Routine maintenance notices, forecast updates, renewable
output outlooks, and interconnector advisories can still be actionable
if they plausibly affect price or volatility.  Set is_actionable=false
only if the article is clearly retrospective, purely corporate or project
news without a near-term operational cue, generic commentary with no
concrete signal, or not plausibly relevant to NSW wholesale price within
14 days.  For borderline cases, prefer is_actionable=true with lower
confidence rather than defaulting to false.  Use only the closed topic
vocabulary:  <TAGS>.  If an article is clearly price-relevant but no topic
fits well, keep is_actionable=true and include other instead of routine.
Reserve routine for news with no clear operational or market signal.  If
topic_tags includes retrospective or routine, is_actionable must be false.
Map price drivers into the existing tags and regime fields as follows:
outage, derating, maintenance, unit trip, tight reserve margin -> outage
or supply_tight; fuel shortage or fuel-cost shock -> fuel_shock; interstate
flow changes, interconnector outages, binding transmission constraints ->
interconnector_limit; heat-driven demand peaks -> heatwave; cold-driven
demand peaks -> cold_snap; high wind, solar, hydro, or excess imports
suppressing price -> renewable_surge or supply_surplus; low wind, low
solar, drought, or import scarcity lifting price -> renewable_drought or
supply_tight; rebidding, scarcity pricing, spike risk, or abrupt operational
uncertainty should raise volatility_tone; administered pricing, market
suspension, or intervention already in force -> market_intervention or
policy_active and policy_in_effect=1.  Interpret regime_vec for price
as:  tightness = near-term NSW supply-demand tightness and scarcity
pressure on price; demand_outlook = near-term load and cooling-demand
pressure; renewable_surplus = expected renewable abundance or import
relief that suppresses price, negative when renewable scarcity lifts price;
volatility_tone = spike, scarcity, rebidding, outage, or constraint-driven
price volatility risk; policy_in_effect = 1 only when a policy or market
intervention is already active.  For indirect but still relevant signals, use
smaller-magnitude regime values instead of zeroing them out.  Return fields
exactly:  <SCHEMA>.  Range constraints as in [Load].
```

[Traffic]

```
You are refining generic news for hourly road traffic count forecasting into
```

strict JSON. Return JSON only, no markdown. The goal is to keep news that describe a current or near-term condition that can plausibly change road traffic volume, congestion, route availability, commuter behavior, freight flow, or hourly traffic volatility within the next 14 days. Treat the following as potentially actionable if they are current, imminent, recently announced and still relevant, or likely to shape travel behavior within 14 days: public holidays, school holidays, long weekends, major sports or entertainment events, festivals, protest activity, road closures, lane reductions, tunnel or bridge restrictions, major crashes with lingering disruption, rail strikes or transit outages that spill travelers onto roads, airport disruptions, port disruptions, severe weather, heatwaves, cold snaps, flooding, evacuations, fuel shortages, fuel-price shocks, major logistics or freight bottlenecks, congestion charging, and traffic-control interventions already in force. Do not filter out an article just because it is operational, local, event-specific, or framed as a traffic notice, transport advisory, or disruption update. Set is_actionable=false only if the article is clearly retrospective, generic commentary with no live signal, corporate or infrastructure planning news without near-term travel impact, or not plausibly relevant to road traffic counts within 14 days. Use only the closed topic vocabulary: <TAGS>. If an article is clearly traffic-relevant but the topic vocabulary is awkward, keep is_actionable=true and include other instead of routine. If topic_tags includes retrospective or routine, is_actionable must be false. Map traffic drivers into the existing tags and regime fields as follows: public holidays, school holidays, long weekends, vacation travel, or event-driven travel surges -> holiday; road closures, crashes with ongoing disruption, rail or transit outages, strikes, or infrastructure failures -> outage; bridge, tunnel, motorway, border, corridor, or lane-capacity restrictions -> interconnector_limit; heat-driven travel changes, heatwave warnings, or fire-weather disruptions -> heatwave; snow, ice, or cold-weather commuting disruption -> cold_snap; fuel shortages, strong fuel-cost shocks, or freight cost shocks -> fuel_shock; traffic control, emergency restrictions, congestion charging, or government intervention already active -> policy_active or market_intervention and policy_in_effect=1; route reopening, transit relief, remote-work or modal-shift relief, or unusually weak travel demand -> supply_surplus or renewable_surge; network scarcity, strong commuter demand, freight bottlenecks, or event-driven congestion pressure -> supply_tight or renewable_drought. Interpret regime_vec for traffic as: tightness = near-term road-network tightness, congestion pressure, and capacity scarcity; demand_outlook = near-term traffic demand pressure from commuting, freight, holidays, and major events; renewable_surplus = use this slot as mobility relief and alternate-capacity availability that reduces road counts or congestion, negative when relief is weak and road usage is pushed higher; volatility_tone = disruption and uncertainty risk for hourly traffic from outages, closures, weather, protests, strikes, or major events; policy_in_effect = 1 only when a policy, restriction, emergency measure, or traffic intervention is already active. Return fields exactly: <SCHEMA>. Range constraints as in [Load].

[Gas Demand]
You are refining generic energy news for Netherlands gas demand forecasting into strict JSON. Return JSON only, no markdown. Keep news that describe a current or near-term condition that can change Dutch or Northwest European gas demand, gas supply tightness, storage stress, balancing conditions, or gas-for-power burn within the next 14 days. Treat the following as potentially actionable if they are current, imminent, or still in force: cold snaps, warm spells, heating demand shifts, storage depletion or refill stress, LNG terminal disruptions, pipeline outages or constraints, import or export bottlenecks, supply security alerts, balancing-market stress, fuel shortages, market intervention already in force, and power-system conditions that materially alter gas burn. Do not filter out an article just because it is framed as security of supply, storage, infrastructure maintenance, terminal availability, or system balancing, as long as it can plausibly affect gas demand or tightness within 14 days. Set is_actionable=false only if the article is retrospective, routine commentary, historical recap, corporate or board news, long-horizon hydrogen/CCS infrastructure news without near-term gas-market impact, or not plausibly relevant to Dutch gas demand or Dutch/NW European gas tightness within 14 days. Use only the closed topic vocabulary: <TAGS>. If an article is clearly gas-demand relevant but no topic fits well, keep is_actionable=true and include other instead of routine. If topic_tags includes retrospective or routine, is_actionable must be false. Map gas-demand drivers into the existing tags and regime fields as follows: cold-driven heating demand or cold weather risk -> cold_snap; mild weather, weak heating demand, or demand relief -> supply_surplus or other; storage depletion, low inventories, import risk, or supply scarcity -> supply_tight; LNG, pipeline, terminal, compressor, or field disruptions -> outage, interconnector_limit, or fuel_shock depending on context; high wind/solar or other substitution that reduces gas burn -> renewable_surge or supply_surplus; low wind/solar or power-market stress that raises gas burn -> renewable_drought or supply_tight; government

```
emergency actions, storage mandates, or market intervention already active
-> policy_active or market_intervention and policy_in_effect=1.  Interpret
regime_vec for gas demand as:  tightness = near-term Dutch/NW European gas
market tightness and scarcity pressure; demand_outlook = near-term Dutch
gas demand pressure, especially weather-sensitive heating demand and gas
burn; renewable_surplus = conditions that reduce gas demand or gas-for-power
burn, negative when weak renewables or substitution constraints increase gas
usage; volatility_tone = uncertainty and volatility risk from storage stress,
outages, balancing tension, or supply shocks; policy_in_effect = 1 only when
a policy or intervention is already active.  Return fields exactly:  <SCHEMA>.
Range constraints as in [Load].
```

The 15-tag closed vocabulary is <TAGS> = {supply_tight, supply_surplus, heatwave, cold_snap, holiday, outage, fuel_shock, renewable_surge, renewable_drought, interconnector_limit, policy_active, market_intervention, retrospective, routine, other}. <SCHEMA> repeats the JSON object listed at the top of this appendix. The Load prompt is the shortest because the curated NSW demand feed is already topically coherent; the Price, Traffic, and Gas prompts apply progressively broader recall instructions as the underlying corpora become noisier.

K Compute and Reproducibility

Hardware and software. All experiments run on a single workstation with 2×NVIDIA GeForce RTX 5090 (32 GB each), 48 logical CPUs, 125 GB system RAM, NVIDIA driver 570.211.01, CUDA 12.8. Python 3.10.19, PyTorch 2.9.1+cu128, transformers 4.57.1, numpy 1.26.4, pandas 2.3.3. Each individual training run uses one GPU; the second GPU is used only to parallelize independent (dataset, horizon) configurations.

Parameter counts and FLOPs scale. The frozen DLinear base has $2(LH+H)$ parameters. With L=48 this ranges from 4,704 at H=48 to 70,560 at H=720. The delta module is dominated by the regime projector and PatchTST encoder; full-model parameter counts are $\{1.97\text{M}, 1.99\text{M}, 2.02\text{M}, 2.08\text{M}, 2.23\text{M}\}$ at $H \in \{48, 96, 192, 336, 720\}$ respectively. The text encoder used during bank build is E5-small-v2 (∼33 M parameters) and is frozen and used only at refinement time, not at training or test time.

Wall-clock training cost. Per (dataset, horizon) configuration on one RTX 5090, with batch size 16 and the schedule in Table 11, base-stage training completes in ∼1–3 min and delta-stage training (12 pretrain + 30–50 main epochs) in ∼10–25 min depending on horizon and training-set size. The complete grid for the four datasets times five horizons fits in approximately 8 GPU-hours. The one-time LLM refinement of all four corpora (37,874 news total) costs ∼45 minutes of API time at GPT-5.1, `temperature`=0, `max_tokens`=260; refined records are cached on disk and re-used across all training runs and ablations.

Determinism and seeds. All Python, NumPy, and PyTorch random states are seeded with the global seed 2024; CUDA non-determinism (atomicAdd, matmul tiling) is not eliminated, so re-running a configuration may produce numerical differences below 0.1% on MAE. The counterfactual permutation seed is held out at 2024 at test time and freshly resampled at every batch during training (Sec. I).

L Sensitivity and Cap Calibration

This appendix documents how the bounded modulation caps in Eq. 4 are selected per dataset and how the knobs settle in practice on the test set. The caps are chosen on the validation MAE under the eligibility constraint of Sec. I (real-news active MAE must beat blanked or permuted active MAE by at least ϵ_2=0.01); they are not tuned on the test set.

Per-dataset cap calibration. The four caps that govern news influence are $\lambda^{\text{cap}}_{\text{news}}$ (the additive news contribution to the trust coefficient), $\lambda_{\max}$ (the hard ceiling on $\lambda_{\text{delta_final}}$), γ_{shape} (the bound on the shape-channel news scaling around 1), and γ_{spike} (the bound on the news-conditioned spike-gate logit bias). Together with the activity threshold θ, they fully determine the worst-case deviation in Eq. 2.

Table 13 translates the chosen caps into the worst-case bound on $\hat{\boldsymbol{y}}_z - \hat{\boldsymbol{y}}_z^{\text{base}}$ in z-space units. Traffic uses the highest γ_{shape} (0.70) because intra-day shape distortions during holidays and incidents are the dominant news signal in the GDELT 2024 stream; the curated NSW feeds use a smaller γ_{shape}=0.20 because most actionable Demand/Price news affect the level rather than the shape.

Table 13: Per-dataset bounded-modulation caps and the implied worst-case bound. Columns 2–5 are inputs; column 6 reports the architecturally enforced cap on $\|\hat{\boldsymbol{y}}_z - \hat{\boldsymbol{y}}_z^{\text{base}}\|_\infty$ when the residual heads saturate at a unit z-space norm.

Dataset	$\lambda_{\text{news}}^{\text{cap}}$	$\lambda_{\max}$	γ_{shape}	γ_{spike}	θ	$\|\hat{\boldsymbol{y}}_z - \hat{\boldsymbol{y}}_z^{\text{base}}\|_\infty \leq$
Electricity Demand	0.12	0.45	0.20	0.75	0.7	0.45
Electricity Price	0.12	0.45	0.20	0.75	0.7	0.45
Traffic	0.12	0.60	0.70	0.75	0.2	0.60
Gas Demand	0.12	0.45	0.20	0.75	0.2	0.45

Realised knob values on the test set. Table 14 reports test-time means of the three news-conditioned knobs at every horizon, computed from the same checkpoints that produced Table 1. Two consistent patterns appear: (i) $\bar{\lambda}_{\text{delta_final}}$ on news-rich datasets (Demand, Traffic) settles in 0.33–0.44, well below the cap, indicating that the bound is non-trivially active rather than slack; (ii) $\bar{g}_{\text{shape}}$ rises with horizon on Traffic (1.42 $\rightarrow$ 1.40 around the cap of 1.70) but stays close to identity on Demand/Price/Gas, consistent with the per-domain prompt design (Traffic emphasizes intra-day disruption events, while Demand/Price emphasize level shifts).

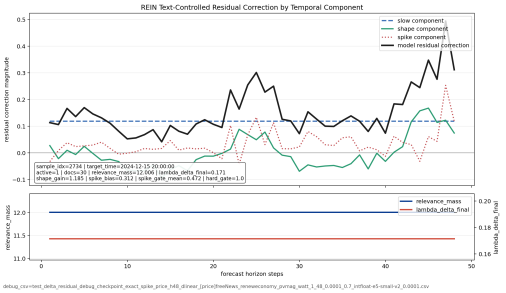
Table 14: Test-set means of the three news-conditioned modulation knobs. $\bar{\lambda}_{\textit{delta_final}}$ averages over the full test set (clipped to $[\lambda_{\min}, \lambda_{\max}]$). $\bar{g}_{\textit{shape}}$ is the multiplicative scaling of the shape head (target identity is 1). $\bar{b}_{\textit{spike}}$ is the additive spike-gate logit bias (target identity is 0). *Spike hit* is the fraction of horizon steps for which the spike gate fires above 0.8.

Dataset	H=48	H=96	H=192	H=336	H=720	**Cap**
$\bar{\lambda}_{\text{delta_final}}$ Demand	0.399	0.335	0.428	0.351	0.440	0.45
$\bar{\lambda}_{\text{delta_final}}$ Price	0.194	0.193	0.208	0.199	0.213	0.45
$\bar{\lambda}_{\text{delta_final}}$ Traffic	0.434	0.430	0.405	0.404	0.396	0.60
$\bar{\lambda}_{\text{delta_final}}$ Gas	0.197	0.229	0.267	0.198	0.222	0.45
$\bar{g}_{\text{shape}}$ Demand	1.192	1.193	1.200	1.200	1.200	[0.80,1.20]
$\bar{g}_{\text{shape}}$ Price	1.160	1.176	1.184	1.200	1.200	[0.80,1.20]
$\bar{g}_{\text{shape}}$ Traffic	1.421	1.476	1.439	1.420	1.396	[0.30,1.70]
$\bar{g}_{\text{shape}}$ Gas	1.047	1.034	1.089	1.066	1.317	[0.80,1.20]
$\bar{b}_{\text{spike}}$ Demand	−0.378	−0.470	−0.227	−0.205	+0.148	[-0.75,0.75]
$\bar{b}_{\text{spike}}$ Price	−0.167	−0.140	−0.601	−0.624	−0.265	[-0.75,0.75]
$\bar{b}_{\text{spike}}$ Traffic	−0.113	−0.206	−0.101	+0.031	+0.170	[-0.75,0.75]
$\bar{b}_{\text{spike}}$ Gas	−0.023	−0.054	−0.135	+0.015	−0.240	[-0.75,0.75]
Spike-hit % Demand	3.8	1.9	5.5	4.4	8.9	—
Spike-hit % Price	2.7	2.0	6.4	5.2	6.8	—
Spike-hit % Traffic	5.5	6.3	7.3	7.3	6.2	—
Spike-hit % Gas	0.2	1.2	2.1	0.0	1.0	—

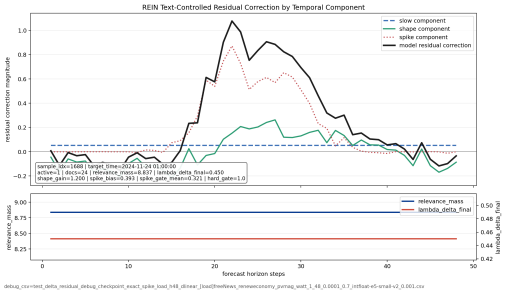
Discussion. The realised $\bar{g}_{\text{shape}}$ values often saturate at the upper end of the allowed range on Demand and Price, suggesting that a small fraction of news-active samples request more shape modulation than the cap permits; tightening these caps further would lose the gains in Table 3, while widening them risks dominating the bound in Eq. 2 on a small number of high-mass windows. The current per-dataset values were selected such that the validation eligibility constraint of Sec. I is satisfied with a non-trivial margin, while keeping the worst-case bound on $\hat{\boldsymbol{y}}_z - \hat{\boldsymbol{y}}_z^{\text{base}}$ at most 0.60 in z-space units (i.e., at most 0.60σ in the original scale of the series). On Traffic, $\bar{b}_{\text{spike}}$ flips sign with horizon, consistent with the model encoding "mute the spike head at short horizons (incidents already absorbed by the base) but encourage spikes at long horizons (events not yet visible)." The spike-hit percentages remain well below the target rate (10% for non-price, 6% for price), indicating the spike gate is conservative in deployment, which is the desired behavior under heavy-tailed targets.

M DIAGRAMS OF RESIDUAL DECOMPOSITION

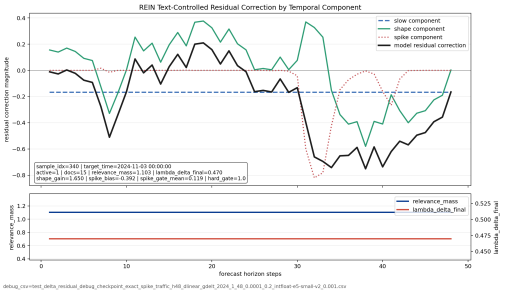
This appendix provides additional visualizations of the residual decomposition learned by REIN across different forecasting horizons (Figures 2–6). Each figure shows representative news-active examples from the four evaluated domains. The upper plot in each panel decomposes the final residual correction into slow, shape, and spike components, while the lower plot reports the news relevance mass and the bounded base modulation coefficient λ_{base}.



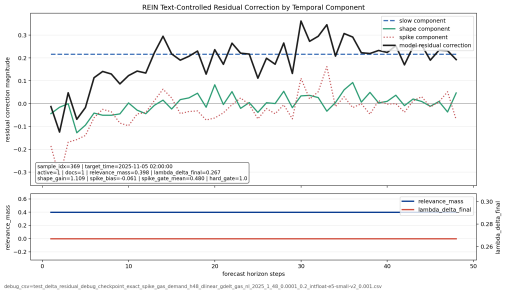
(a) Electricity Price.



(b) Electricity Load.



(c) Traffic.



(d) Gas Demand.

Figure 2: Examples of REIN's text-controlled residual decomposition at forecasting horizon $H = 48$ across four domains.

N DECOMPOSED RESIDUAL TARGETS

This appendix gives the full construction of the slow / shape / spike training targets used by Eq. 3. We pre-compute on the training set a calendar baseline $\mathbf{C} \in \mathbb{R}^{7\times 48}$ defined cell-wise as the mean residual at each (day-of-week, half-hour-of-day) bin, and the standard deviation σ_e of the $\mathbf{C}$-detrended residual. For a sample at t with horizon target times $\{u_h\}_{h=1}^{H}$ and true residual $\boldsymbol{e}^\star(t) = \boldsymbol{y}_z(t) - \hat{\boldsymbol{y}}_z^{\text{base}}(t)$, let

$$\mu(t) = \tfrac{1}{H}\textstyle\sum_h e_h^\star(t), \qquad \tilde{e}_h^\star(t) = e_h^\star(t) - \mu(t) - \mathbf{C}\big[\text{dow}(u_h),\, \text{hod}(u_h)\big]. \tag{8}$$

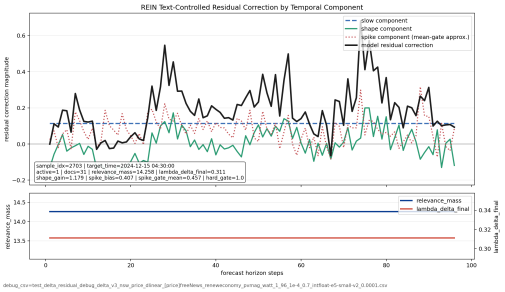
The spike mask is the intersection of an absolute and a rank criterion,

$$\mathbb{S}_h(t) = \mathbb{I}\big[|\tilde{e}_h^\star(t)| > \max(k\sigma_e,\, q_{1-p})\big] \,\wedge\, \mathbb{I}\big[h \in \text{top-}p \text{ of } |\tilde{\boldsymbol{e}}^\star(t)|\big], \tag{9}$$

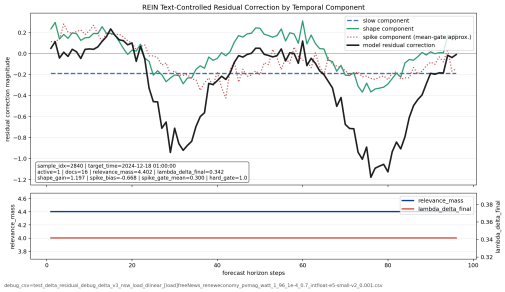
with k=3, p=0.10, and q_{1-p} the empirical $(1-p)$-quantile of $|\tilde{e}^\star|$ on the training set; the rank criterion is essential on the heavy-tailed price residual where σ_e alone would saturate the spike target. The spike, slow, and shape targets are

$$e^\star_{\text{spike},h} = \tilde{e}^\star_h \cdot \mathbb{S}_h, \qquad e^\star_{\text{slow}} = \mu + \tfrac{1}{H}\textstyle\sum_h \big(e^\star_h - \mu - e^\star_{\text{spike},h}\big), \qquad e^\star_{\text{shape},h} = e^\star_h - e^\star_{\text{slow}} - e^\star_{\text{spike},h}, \tag{10}$$

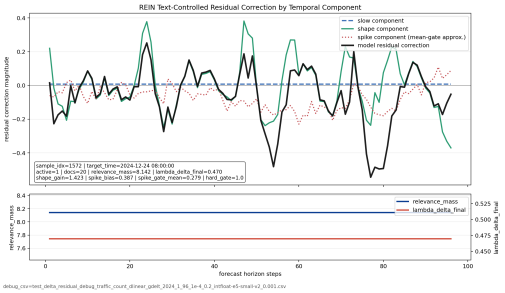
which by construction satisfy $e^\star_h = e^\star_{\text{slow}} + e^\star_{\text{shape},h} + e^\star_{\text{spike},h}$ for every h with $\sum_h e^\star_{\text{shape},h} = 0$. The three TS heads in Eq. 3 are trained against $e^\star_{\text{slow}}$, $\boldsymbol{e}^\star_{\text{shape}}$, and $\boldsymbol{e}^\star_{\text{spike}}$ respectively, and the spike-gate logits $\boldsymbol{\ell}_{\text{spike}}$ are trained against the binary mask $\mathbb{S}$.



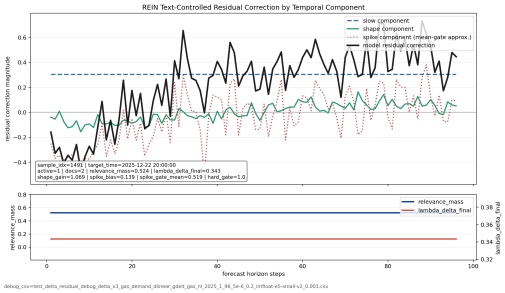
(a) Electricity Price.



(b) Electricity Load.

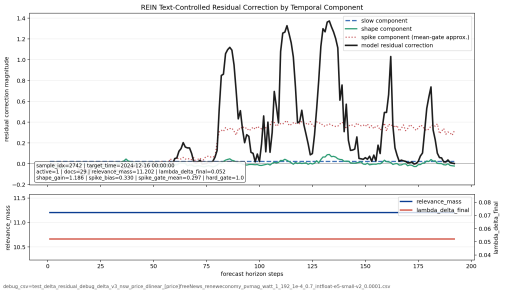


(c) Traffic.

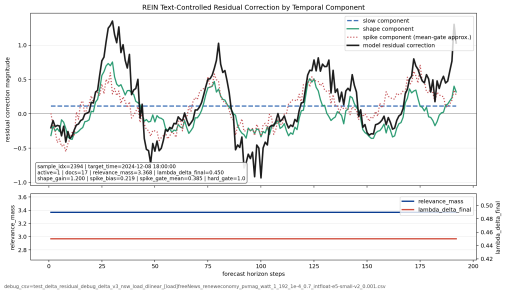


(d) Gas Demand.

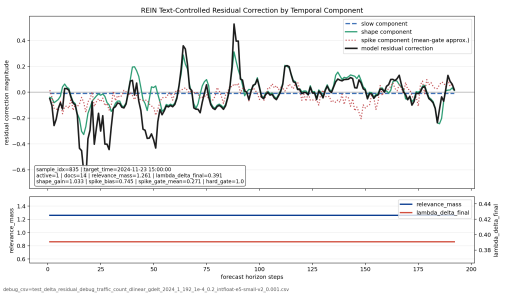
Figure 3: Examples of REIN's text-controlled residual decomposition at forecasting horizon $H = 96$ across four domains.



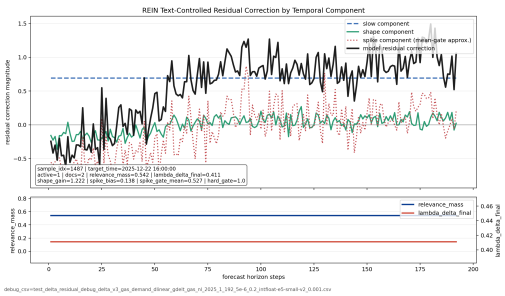
(a) Electricity Price.



(b) Electricity Load.

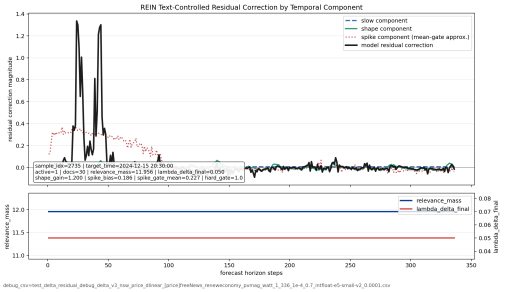


(c) Traffic.

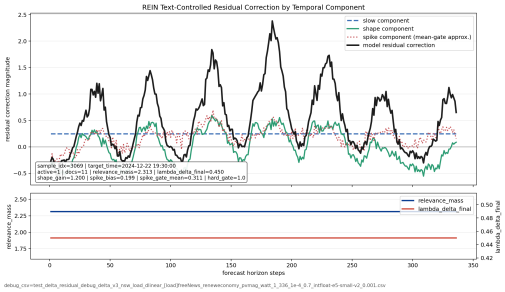


(d) Gas Demand.

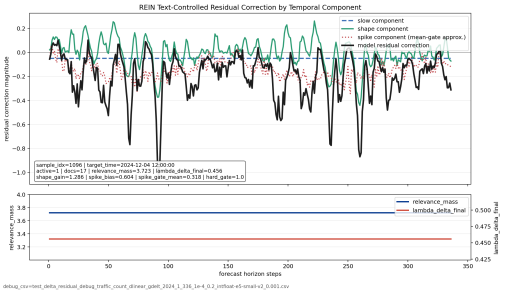
Figure 4: Examples of REIN's text-controlled residual decomposition at forecasting horizon $H = 192$ across four domains.



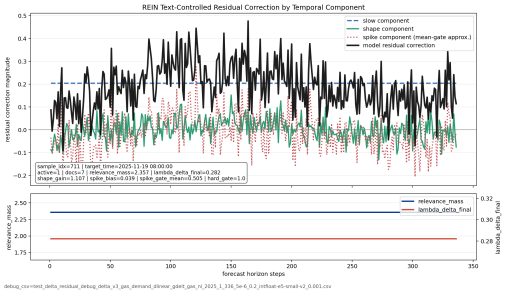
(a) Electricity Price.



(b) Electricity Load.

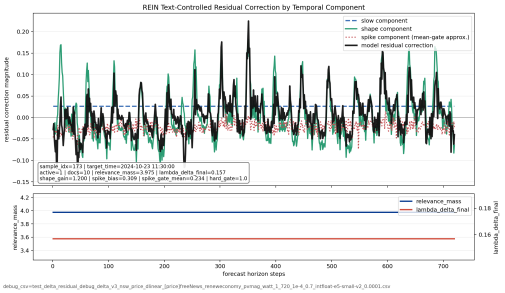


(c) Traffic.

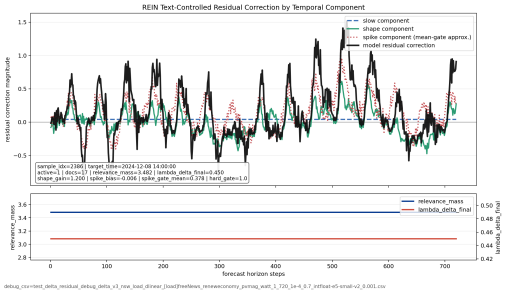


(d) Gas Demand.

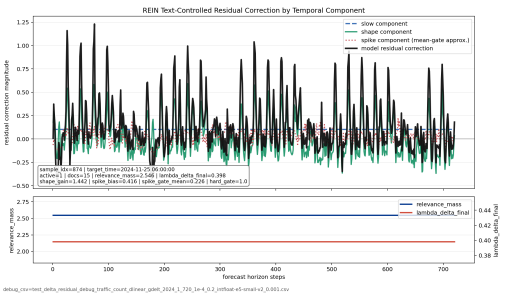
Figure 5: Examples of REIN's text-controlled residual decomposition at forecasting horizon $H = 336$ across four domains.



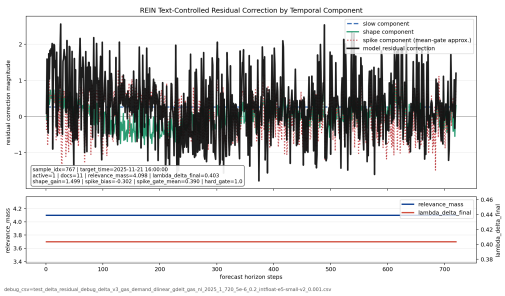
(a) Electricity Price.



(b) Electricity Load.



(c) Traffic.



(d) Gas Demand.

Figure 6: Examples of REIN's text-controlled residual decomposition at forecasting horizon $H = 720$ across four domains.