READING THE DATE, NOT THE EVENT: THE CALENDAR SHORTCUT IN TEXT-ASSISTED TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-assisted forecasting pairs numerical windows with contemporaneous documents to capture event semantics unavailable from the series alone. Evidence for this premise rests largely on modality ablation (adding text helps, removing it hurts), which establishes predictive signal but does not identify its source. We design controlled interventions on trained forecasters to localize that signal: reversing event semantics alters MAE by less than 1%, whereas masking only dates and timestamp expressions reproduces at least 74% of the full text removal penalty in five of the six settings. We term this mode *the calendar shortcut*: forecasters recover periodic phase from textual calendar cues while largely bypassing event content. To overcome it, we propose State Conditioned Paired Residual Adaptation (SPRA), a lightweight, backbone agnostic framework for state-conditioned residual learning. Shortcut controlled document pairs preserve event content while suppressing calendar cues, and a paired consistency objective reduces the adapter's reliance on such cues. A low-rank adapter then estimates what the text contributes beyond the observed trajectory and maps it to a target specific multi-horizon residual correction of a frozen backbone. Across nine domains and five backbones, SPRA ranks first in 79 of 90 comparisons, improves upon its unimodal backbone in 44 of 45 settings, and retains its gains under the same shortcut interventions, supporting attribution to forecast-relevant event content rather than incidental calendar identity.

1 INTRODUCTION

Time series forecasts guide decision making in energy, climate, and other event sensitive domains (Xu et al., 2026; Zhang et al., 2026; Li et al., 2026b). However, a numerical history records only the traces that past events have left on measured variables. Outbreak reports and market announcements influence future dynamics before their effects become visible within the lookback window. Text-assisted forecasting pairs each numerical window with contemporaneous documents intended to provide event related evidence that the time series alone cannot capture (Hu et al., 2026).

Existing approaches differ primarily in where this evidence is incorporated into the forecasting pipeline. Feature engineering methods encode documents as auxiliary variables and integrate them through concatenation, prompting, or learned covariates (Jia et al., 2024; Li et al., 2026c). Inference level methods instead rely on a language model to interpret the documents and combine its assessment with a numerical prediction (Liu et al., 2024a; Wang et al., 2026). Despite their architectural diversity, these methods aim to use contemporaneous documents as an additional source of forecasting evidence. This target is plausible and has motivated the development of increasingly elaborate fusion and reasoning mechanisms.

However, it remains unclear which information in the paired documents drives the improvement. Calendar position is a legitimate forecasting signal. A document can encode both this information and event specific evidence, so improved accuracy after adding text does not identify which source the model uses. If the gain is supplied mainly by calendar identity, it may be obtainable without processing document content. The additional promise of text is its ability to inform predictions about event dependent changes that calendar cues alone do not specify. Motivated by the broader

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

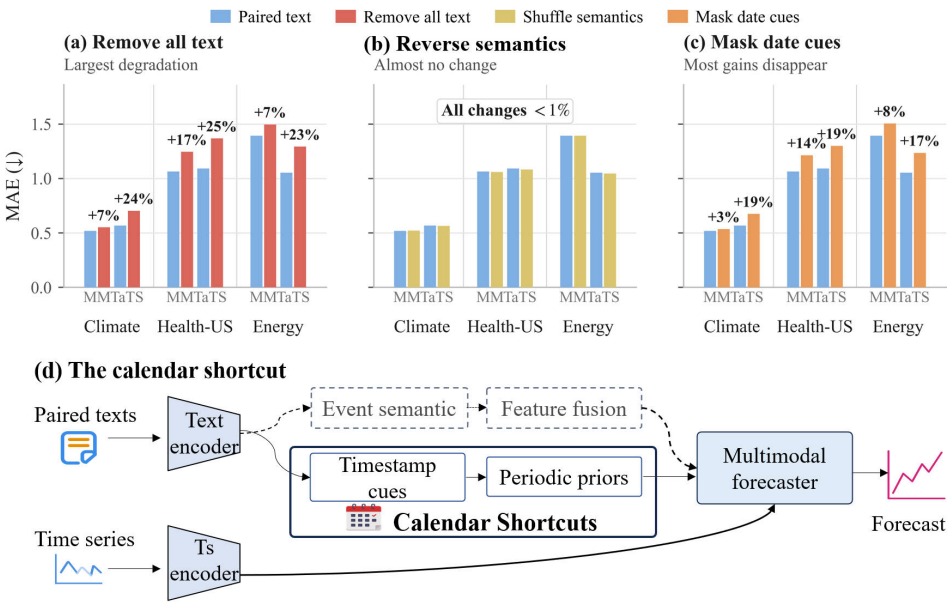


Figure 1: Three controlled interventions, applied to the same trained forecasters, localize the predictive signal carried by paired text.

problem of shortcut learning (Geirhos et al., 2020), we investigate this distinction through controlled interventions on event content and calendar expressions.

Figure 1 separates these sources through three interventions across six types of settings. Removing all paired text increases MAE by 7%–25%, confirming substantial predictive information. Reversing event semantics, such as rise to decline and stability to volatility, changes MAE by less than 1%. Most observed gains therefore survive semantic reversal. Masking only explicit dates and timestamp expressions, with the rest of each document unchanged, increases MAE by 3%–19% and reproduces at least 74% of the full text-removal penalty in five of the six settings. These interventions reveal a calendar shortcut: forecasters extract calendar identity from paired text to recover periodic priors while largely bypassing event content (Figure 1d). They do not establish that all textual semantics are ignored; they show that calendar information accounts for most of the text-removal penalty.

The preceding analysis shows that controlling calendar shortcuts is necessary before gains from paired text can be attributed to event information. We further observe that the usefulness of text information depends on the current numerical state. For example, a report of the "supply disruption" may provide substantial predictive value while the recent trajectory remains stable, but offer little additional information once the disruption is already reflected in the time series. Motivated by these observations, we propose State Conditioned Paired Residual Adaptation (SPRA) framework for numerical forecasting backbones. SPRA first constructs paired text views that preserve event content while varying calendar and template cues, and enforces consistency between the resulting corrections to reduce reliance on these shortcuts. It conditions the contribution of textual information on a compact representation of the observed numerical trajectory, allowing the current state to regulate when and how strongly the text should affect the forecast. Finally, rather than replacing the numerical prediction, SPRA uses text to estimate the residual error of the frozen backbone through a shared low-rank decoder, with the correction calibrated on a later chronological block.

Our contributions are threefold. (1) We identify a calendar shortcut in text-assisted forecasting, showing that much of the textual gain comes from calendar cues rather than event semantics. (2) We propose State Conditioned Paired Residual Adaptation (SPRA), a lightweight, backbone agnostic framework that suppresses this shortcut while incorporating text according to the current numerical state through residual adaptation. (3) Across nine domains, SPRA outperforms mainstream forecasting methods, ranking first in 79 of 90 domain$\times$backbone$\times$metric comparisons. Further analysis demonstrates adaptability across five backbones with low training time and memory overhead.

2 Related Work

2.1 Time Series Forecasting

Numerical forecasting provides the foundation on which text-assisted methods build. Transformer variants incorporate temporal structure through frequency decomposition, distribution shift normalization, and patch or variate tokenization (Zhou et al., 2022; Liu et al., 2022; Nie et al., 2023; Liu et al., 2024b); linear, convolutional, and multi-scale architectures remain competitive (Zeng et al., 2023; Wu et al., 2023; Wang et al., 2024; Luo & Wang, 2024); and pretrained foundation models replace dataset-specific training with reusable models adapted downstream (Woo et al., 2024; Liu et al., 2024c; Shi et al., 2025; Liu et al., 2025b;c; Gao et al., 2026; Li et al., 2026a). All three lines infer the future from traces that past events have already imprinted on the series and therefore cannot anticipate an event whose effect has not yet entered the lookback window, which is precisely the gap that text-assisted forecasting addresses.

2.2 Multimodal Time Series Forecasting with Text Incorporation

One family repurposes pretrained language or vision backbones as numerical encoders without consuming documents (Jin et al., 2024; Pan et al., 2024; Cao et al., 2024; Chen et al., 2025; Zhong et al., 2025). When temporally aligned text is available, existing approaches fall into two groups. Feature fusion methods transform text into representations that are aligned with numerical features. GPT4MTS encodes news through LLM prompts, TaTS compresses text embeddings into auxiliary covariates, and MM-TSFLib combines separately modeled numerical and textual representations (Jia et al., 2024; Li et al., 2026c; Liu et al., 2024a). ChatTime further integrates multiple modalities within unified representations or generative architectures (Wang et al., 2025). In contrast, reasoning-based methods use language models to interpret textual evidence. VoT combines an LLM derived event forecast with a numerical forecast in the frequency domain (Wang et al., 2026). Aurora injects text and image domain knowledge into temporal representations for zero-shot inference (Wu et al., 2026). Despite increasingly elaborate mechanisms for incorporating text, it remains unclear whether gains reflect event understanding or temporal shortcuts. Concurrent work identifies the text branch contributes little content dependent signal, and supervises the text branch on the residuals of a jointly trained backbone (Nguyen et al., 2026). SPRA learns from residuals, but corrects a frozen backbone with chronologically constructed targets and addresses a different mode, namely the reliance on calendar cues.

3 Methodology

Problem setup and leakage-free residual targets. Consider a time series observed up to a forecast origin t, where $\boldsymbol{X}_t \in \mathbb{R}^{L\times d}$ denotes the preceding L time steps and $\mathcal{T}_t$ denotes the set of documents available by time t. The H-step forecasting target is $\boldsymbol{y}_t = [y_{t+1}, \ldots, y_{t+H}]$. We use PatchTST (Nie et al., 2023) as the running example of a numerical backbone f_θ. The backbone produces an anchor forecast $\boldsymbol{b}_t = f_\theta(\boldsymbol{X}_t)$ and leaves a residual $\boldsymbol{r}_t = \boldsymbol{y}_t - \boldsymbol{b}_t$. For the text branch, $\boldsymbol{r}_t$ is the prediction target and $\boldsymbol{c}_t$ the corresponding output, since $\|\boldsymbol{r}_t - \boldsymbol{c}_t\|^2 = \|\boldsymbol{y}_t - (\boldsymbol{b}_t + \boldsymbol{c}_t)\|^2$. To prevent future information leakage, we construct the residuals sequentially: for each adapter training interval, the forecasting backbone is fitted using observations preceding that interval. This procedure preserves the causal forecasting setting and ensures that no future information is used.

Let $\boldsymbol{z}_t^c$ encode the document content available at t and $\boldsymbol{u}_t$ summarize the numerical state. The adapter g_ϕ is trained toward the conditional residual, $\boldsymbol{c}_t^c = g_\phi(\boldsymbol{z}_t^c, \boldsymbol{u}_t) \approx \mathbb{E}[\boldsymbol{r}_t \mid \boldsymbol{z}_t^c, \boldsymbol{u}_t]$, with $g_\phi(\boldsymbol{0}, \boldsymbol{u}_t) = \boldsymbol{0}$. The term $g_\phi(\boldsymbol{z}^c, \boldsymbol{u})$ specifies the role of text: to explain the anchor forecast error under the current trajectory regime. The constraint $g_\phi(\boldsymbol{0}, \boldsymbol{u}_t) = \boldsymbol{0}$ enforces text necessity, ensuring that zero textual evidence yields zero correction for any state.

Overview. During training, SPRA proceeds through three stages. Paired text views expose the calendar-and-template surface during training so that it can be neutralized (§3.1). A compact trajectory state, aligned with the text along a few shared directions, gates how much of the available evidence acts (§3.2–3.4). A low-rank decoder turns the gated evidence into one coherent multi-horizon correction (§3.5). The sentence encoder stays frozen.

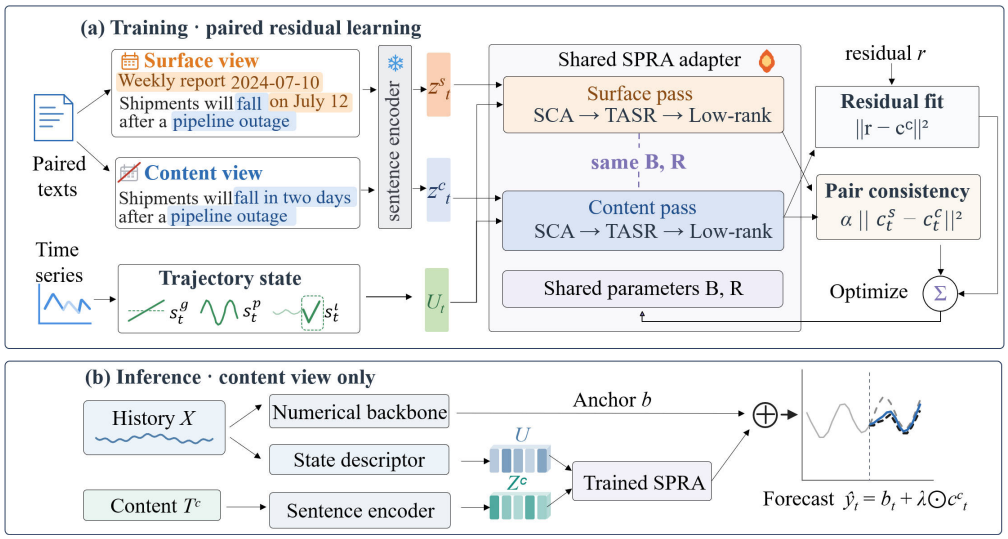


Figure 2: Overview of SPRA. Paired text views preserve event semantics while controlling calendar cues. At inference, the content view correction refines the forecast produced by the backbone.

3.1 Paired Text Views

For each document, SPRA constructs two views of the same event. The content view converts event dates into offsets relative to the forecast origin t, removes absolute timestamps and recurring report templates, and preserves the event entities, quantities, direction, intensity, and lead time. For example, at a July 10 origin, "shipments will fall on July 12" becomes "shipments will fall in two days." The paired surface view retains the original dates and report template while preserving identical event content. This construction separates calendar identity from event timing: event lead times are retained, while absolute calendar expressions and recurring report templates are controlled in the textual views. The paired consistency objective targets the adapter's dependence on these textual cues. A frozen sentence encoder maps the two views to z_t^c and z_t^s.

3.2 Compact Trajectory State

The state descriptor answers one focused question: what regime is present when the documents are available? It summarizes the observed target history at three scales:

$$\boldsymbol{u}_t = R_s\,[\boldsymbol{s}_t^{\mathrm{g}}; \boldsymbol{s}_t^{\mathrm{p}}; \boldsymbol{s}_t^{\ell}] \in \mathbb{R}^{d_s}. \tag{1}$$

The global vector $\boldsymbol{s}_t^{\mathrm{g}}$ records level, trend, endpoint, and coarse shape; the periodic vector $\boldsymbol{s}_t^{\mathrm{p}}$ records reliable recurrence and low-frequency energy; and the local vector $\boldsymbol{s}_t^{\ell}$ describes the variation, slope, and latest change within an adaptive recent segment. The matrix R_s is a low-capacity projection fitted on training histories. The resulting $\boldsymbol{u}_t$ is a regime descriptor that is inexpensive to compute.

3.3 Shared Cross-Alignment

The state $\boldsymbol{u}_t$ describes the observed trajectory, whereas the text embedding $\boldsymbol{z}_t^v$ describes an external event. Directly concatenating them leaves the residual decoder to discover all cross-modal interactions in the original feature spaces. We instead use **Shared Cross-Alignment** (SCA) to construct a compact routing space before residual learning. SCA identifies paired directions along which numerical regimes and content-text patterns repeatedly co-vary in the training data.

Let $\widetilde{\boldsymbol{u}}_t$ and $\widetilde{\boldsymbol{z}}_t^v$ denote features standardized with training-set statistics, where $v \in \{\mathrm{c}, \mathrm{s}\}$ selects the content or surface view. SCA estimates cross-covariance from the content view and retains its leading r_{ca} singular directions:

$$C_{uz} = \frac{1}{N}\sum_{t=1}^{N} \widetilde{\boldsymbol{u}}_t \widetilde{\boldsymbol{z}}_t^{\mathrm{c}\top} \approx P_u \Sigma_r P_z^{\top}, \qquad \boldsymbol{a}_t = P_u^{\top}\widetilde{\boldsymbol{u}}_t, \qquad \boldsymbol{q}_t^v = P_z^{\top}\widetilde{\boldsymbol{z}}_t^v. \tag{2}$$

Here N is the number of training examples, $P_u \in \mathbb{R}^{d_s \times r_{\text{ca}}}$ and $P_z \in \mathbb{R}^{d_z \times r_{\text{ca}}}$ contain the retained state and text directions, and Σ_r contains their singular values. The outputs $\boldsymbol{a}_t$ and $\boldsymbol{q}_t^v$ are paired routing-channel activations. Only the content view defines the basis, preventing the controlled calendar surface from determining the alignment. The same P_z then projects both views, so their downstream corrections are compared in identical coordinates. Residual supervision in §3.5, rather than SCA itself, determines which routed patterns are forecast-relevant.

3.4 Text Anchored State Routing

SCA places the numerical state and text in paired coordinates $\boldsymbol{a}_t$ and $\boldsymbol{q}_t^v$. The remaining question is how the state should change the text before it is converted into a forecast correction. We do not feed $\boldsymbol{a}_t$ directly to the decoder. Instead, the **Text Anchored State Router** (TASR) forms two adjusted versions of the same text activation:

$$\begin{aligned} \boldsymbol{m}_{t,1}^v = \boldsymbol{q}_t^v \odot \tanh(\boldsymbol{a}_t), \qquad \boldsymbol{m}_{t,2}^v = \boldsymbol{q}_t^v \oslash \big(\mathbf{1} + |\boldsymbol{a}_t - \boldsymbol{q}_t^v|\big), \\ \boldsymbol{h}_t^v = \big[\widetilde{\boldsymbol{z}}_t^v; \boldsymbol{m}_{t,1}^v; \boldsymbol{m}_{t,2}^v\big]. \end{aligned} \tag{3}$$

Here $\odot$ and $\oslash$ denote element-wise multiplication and division. Both terms start from $\boldsymbol{q}_t^v$: $\boldsymbol{m}_{t,1}^v$ applies the signed state activation, whereas $\boldsymbol{m}_{t,2}^v$ attenuates a channel as its distance from the state channel grows. The original text is retained alongside these adjustments in $\boldsymbol{h}_t^v$. Thus the same report can be represented differently when its effect is absent versus already visible in the series. Missing text sets all text terms to zero, so the state alone cannot produce a correction.

3.5 Low-Rank Trajectory Decoding and Learning

The decoder produces k window-specific weights. Inspired by reduced rank regression and neural basis-expansion forecasting (Oreshkin et al., 2020; Ni et al., 2023), we represent the residual trajectory in a learned low-dimensional output subspace. Each basis vector specifies an H step correction pattern shared across windows, while state-conditioned text determines its window-specific weight. Let $R \in \mathbb{R}^{d_h \times k}$ produce the weights and write $B = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_k] \in \mathbb{R}^{H \times k}$ for the learned shapes:

$$\boldsymbol{\beta}_t^v = R^\top \boldsymbol{h}_t^v, \qquad \boldsymbol{c}_t^v = B\boldsymbol{\beta}_t^v = \sum_{j=1}^{k} \beta_{t,j}^v \boldsymbol{\phi}_j. \tag{4}$$

Each $\boldsymbol{\phi}_j \in \mathbb{R}^H$ is a learned correction shape, which may emphasize near or later horizons, and $\boldsymbol{\beta}_t^v \in \mathbb{R}^k$ specifies how strongly the shapes are added or subtracted for the current origin.

Learning objective. The content view learns the realized anchor error, while the paired view tests whether controlled calendar and template changes alter the resulting correction:

$$\mathcal{L}_{\text{SPRA}} = \underbrace{\frac{1}{N}\sum_{t=1}^{N} \|\boldsymbol{r}_t - \boldsymbol{c}_t^{\text{c}}\|_2^2}_{\text{residual fitting}} + \alpha_{\text{pair}} \underbrace{\frac{1}{N}\sum_{t=1}^{N} \|\boldsymbol{c}_t^{\text{s}} - \boldsymbol{c}_t^{\text{c}}\|_2^2}_{\text{paired consistency}} + \eta \|BR^\top\|_F^2. \tag{5}$$

The first term learns the state-conditioned correction from chronological anchor errors. The second acts at the prediction level: it reduces sensitivity to the specified surface intervention without forcing the two text embeddings to coincide (Kurtuluş et al., 2023); $\alpha_{\text{pair}} \geq 0$ controls this constraint. In the final term, $BR^\top$ is the matrix that converts $\boldsymbol{h}_t^v$ into $\boldsymbol{c}_t^v$. Penalizing the sum of its squared entries discourages unnecessarily large corrections that fit noise; $\eta \geq 0$ controls this shrinkage, while k controls how many correction shapes are used.

Training and inference. We first train and freeze the numerical backbone, construct chronological anchor errors, and then fit the SPRA adapter while keeping the sentence encoder frozen. A later validation block estimates a shrinkage vector $\boldsymbol{\lambda} \in [0,1]^H$ to attenuate unsupported horizon-wise corrections. The deployed forecast is

$$\widehat{\boldsymbol{y}}_t = \boldsymbol{b}_t + \boldsymbol{\lambda} \odot \boldsymbol{c}_t^{\text{c}}. \tag{6}$$

At inference, only $\boldsymbol{X}_t$ and the content text view are required. The realized error $\boldsymbol{r}_t$ is neither observed nor needed; SPRA predicts $\boldsymbol{c}_t^{\text{c}}$ directly from the available text and numerical state.

Table 1: Time-MMD multimodal forecasting (MSE/MAE, averaged over prediction lengths). For every backbone and metric the best result is marked **red** and the second best blue.

Domain	Method	iTransformer		PatchTST		Crossformer		DLinear		FEDformer	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	Uni-modal	0.163	0.270	0.165	0.277	1.345	0.807	0.506	0.512	0.305	0.426
	MM-TSFLib	0.213	0.344	0.211	0.342	1.393	0.829	0.551	0.524	0.327	0.472
	TaTS	0.125	0.227	0.185	0.284	0.697	**0.574**	0.414	0.442	**0.279**	0.395
	SPRA (ours)	**0.111**	**0.201**	**0.151**	**0.254**	**0.678**	0.586	**0.342**	**0.429**	0.283	**0.383**
Climate	Uni-modal	0.305	0.423	0.306	0.413	0.291	0.403	0.303	0.414	0.297	0.416
	MM-TSFLib	0.345	0.452	0.338	0.439	0.312	0.425	0.311	0.419	0.295	0.421
	TaTS	0.265	0.387	0.267	0.392	0.257	0.381	0.245	0.369	0.254	0.378
	SPRA (ours)	**0.256**	**0.377**	**0.252**	**0.376**	**0.245**	**0.365**	**0.240**	**0.363**	**0.249**	**0.375**
Economy	Uni-modal	0.281	0.421	0.294	0.436	0.402	0.512	0.318	0.449	0.307	0.443
	MM-TSFLib	0.298	0.438	0.286	0.429	0.437	0.536	0.324	0.454	0.279	0.420
	TaTS	0.239	0.381	0.247	0.389	0.301	0.437	0.259	0.398	0.252	0.394
	SPRA (ours)	**0.221**	**0.362**	**0.224**	**0.366**	**0.263**	**0.404**	**0.241**	**0.379**	**0.233**	**0.373**
Energy	Uni-modal	0.248	0.385	0.287	0.419	0.286	0.424	0.332	0.449	0.252	0.367
	MM-TSFLib	0.313	0.432	0.310	0.421	0.439	0.522	0.339	0.455	0.309	0.431
	TaTS	0.297	0.423	0.243	0.360	0.318	0.454	0.393	0.482	0.367	0.484
	SPRA (ours)	**0.246**	**0.367**	**0.241**	**0.354**	**0.278**	**0.398**	**0.293**	**0.418**	**0.242**	**0.365**
Environment	Uni-modal	0.352	0.416	0.350	0.409	0.367	0.451	0.375	0.474	0.354	0.427
	MM-TSFLib	0.355	0.410	0.360	0.411	0.390	0.442	0.352	0.427	0.388	0.432
	TaTS	0.315	0.407	0.313	0.406	0.327	0.437	0.331	0.443	0.314	0.414
	SPRA (ours)	**0.309**	**0.398**	**0.312**	**0.401**	**0.317**	**0.406**	**0.312**	**0.410**	**0.311**	**0.396**
Health-US	Uni-modal	1.865	0.850	1.876	0.862	1.721	0.842	2.570	1.089	1.795	0.907
	MM-TSFLib	2.005	0.905	2.058	0.918	1.696	0.883	2.729	1.133	1.924	1.015
	TaTS	1.418	**0.750**	1.462	**0.766**	**1.364**	0.743	1.807	0.904	1.476	0.837
	SPRA (ours)	**1.374**	0.759	**1.384**	0.769	1.372	**0.723**	**1.773**	**0.881**	**1.434**	**0.761**
Security	Uni-modal	81.081	4.764	80.432	4.796	86.164	5.423	69.699	3.823	72.216	4.032
	MM-TSFLib	105.152	6.212	95.254	5.466	97.889	6.222	69.593	3.748	**70.539**	4.419
	TaTS	79.358	4.328	78.013	**4.189**	85.154	5.291	70.142	**3.639**	71.517	4.313
	SPRA (ours)	**73.954**	**4.239**	**73.822**	4.215	**83.266**	**5.103**	**68.018**	3.641	70.774	**3.989**
Social Good	Uni-modal	1.305	0.478	1.246	0.486	0.826	**0.404**	1.359	0.768	1.208	0.638
	MM-TSFLib	2.935	0.864	2.496	0.777	1.258	0.686	1.326	0.738	2.141	0.860
	TaTS	1.333	0.530	1.283	0.525	0.848	0.462	1.601	0.934	1.086	0.532
	SPRA (ours)	**1.058**	**0.451**	**1.022**	**0.454**	**0.811**	0.416	**1.291**	**0.719**	**1.009**	**0.464**
Traffic	Uni-modal	0.164	0.230	0.122	0.205	0.177	0.373	0.211	0.390	0.158	0.252
	MM-TSFLib	0.251	0.365	0.167	0.281	0.150	0.326	0.220	0.395	0.191	0.302
	TaTS	0.136	0.205	0.122	0.196	0.126	0.287	0.140	0.289	**0.131**	**0.212**
	SPRA (ours)	**0.126**	**0.199**	**0.116**	**0.192**	**0.121**	**0.266**	**0.128**	**0.261**	0.132	0.214

4 EXPERIMENTS

We empirically validate whether SPRA improves diverse backbones and competes with multimodal methods (§4.2); whether its gains come from event content or from calendar identity (§4.3); which components are necessary (§4.4); and what the learned correction encodes and costs (§4.5–§4.6).

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on Time-MMD (Liu et al., 2024a), a multimodal benchmark that pairs numerical windows with retrieved reports across nine domains: Agriculture, Climate, Economy, Energy, Environment, Health-US, Security, Social Good, and Traffic (App. C). Per-domain lookback and horizon follow native reporting frequencies: $L{=}8$, $\mathcal{H}{=}\{6, 8, 10, 12\}$ for Agriculture, Climate, Economy, Security, Social Good, and Traffic; $L{=}36$, $\mathcal{H}{=}\{12, 24, 36, 48\}$ for Health-US and Energy; $L{=}96$, $\mathcal{H}{=}\{48, 96, 192, 336\}$ for Environment.

Baselines and Metrics. SPRA is evaluated as a plug-in to five numerical backbones: iTransformer, PatchTST, Crossformer, DLinear, and FEDformer. Text-assisted baselines include MM-TSFLib (output level fusion (Liu et al., 2024a)) and TaTS (auxiliary covariates (Li et al., 2026c)). The multimodal comparison additionally includes Aurora, VoT, CALF, Time-VLM, ChatTime, and

Table 2: Comparison with multimodal and reasoning-based methods. SPRA is instantiated on PatchTST. The best result is marked **red** and the second best blue.

Methods	SPRA		Aurora		VoT		CALF		Time-VLM		ChatTime		Time-LLM	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	**0.151**	**0.254**	0.156	0.260	0.159	0.271	0.178	0.292	0.163	0.265	0.172	0.284	0.168	0.277
Climate	**0.252**	**0.376**	0.273	0.400	0.280	0.410	0.302	0.399	0.287	0.421	0.266	0.395	0.337	0.432
Economy	**0.224**	**0.366**	0.264	0.421	0.249	0.410	0.236	0.390	0.242	0.382	0.255	0.399	0.231	0.375
Energy	0.241	**0.354**	**0.237**	0.377	0.255	0.386	0.251	0.363	0.249	0.407	0.256	0.396	0.267	0.370
Environment	**0.312**	**0.401**	0.368	0.461	0.337	0.449	0.346	0.427	0.329	0.437	0.356	0.419	0.321	0.411
Health-US	1.384	**0.769**	1.453	0.777	**1.342**	0.772	1.536	0.794	1.495	0.785	1.592	0.804	1.419	0.792
Security	**73.822**	**4.215**	77.882	4.594	79.728	4.405	87.110	4.847	84.157	4.721	76.037	4.320	81.942	4.489
Social Good	**1.022**	0.454	1.078	0.472	1.165	0.465	1.053	**0.445**	1.104	0.483	1.206	0.491	1.134	0.451
Traffic	**0.116**	**0.192**	0.125	0.204	0.119	0.201	0.122	0.197	0.129	0.209	0.137	0.221	0.132	0.215

Table 3: Counterfactual attribution of textual gains. $\bar{\rho}$, $\bar{\sigma}$ average Eq. (7) over domains. The backbone model is PatchTST.

	Climate			Health-US			Energy			Summary	
Method	Δ_{rem}	Δ_{sem}	Δ_{date}	Δ_{rem}	Δ_{sem}	Δ_{date}	Δ_{rem}	Δ_{sem}	Δ_{date}	$\bar{\rho}\downarrow$	$\bar{\sigma}\uparrow$
MM-TSFLib (fusion)	7.2	0.4	3.1	16.8	0.6	13.9	7.1	−0.3	8.2	0.80	0.02
TaTS (fusion)	23.8	0.5	19.2	25.1	0.7	19.2	22.8	0.2	16.9	0.77	0.02
VoT (reasoning)	0.9	0.0	0.0	1.6	0.1	0.0	1.1	−0.1	0.1	0.03	−0.01
SPRA, α_{pair}=0	9.9	8.7	2.4	12.2	10.6	3.0	18.6	16.9	3.5	0.23	0.89
SPRA	9.8	10.9	0.5	12.1	12.6	0.7	18.4	21.3	0.8	**0.05**	**1.10**

Time-LLM (Wu et al., 2026; Wang et al., 2026; Liu et al., 2025a; Zhong et al., 2025; Wang et al., 2025; Jin et al., 2024). We report MSE and MAE averaged over all forecast horizons per domain.

Protocol. All datasets follow a chronological 70/10/20 split into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{cal}}$, and $\mathcal{D}_{\text{test}}$, with no temporal overlap. The numerical backbone is trained first and then frozen, after which residual targets are constructed from blocked out-of-fold prediction errors (App. B). The sentence encoder remains frozen throughout adapter training. For fair comparison, all methods use the same numerical inputs, retrieved text, and chronological splits. We generate no text: every document is a Time-MMD retrieval whose publication timestamp is no later than the forecast origin t.

4.2 Overall Forecasting Performance

Across the 90 dataset–backbone–metric cells in Table 1, SPRA ranks first in 79 and second in 11. Relative to the numerical backbone, it improves MSE in all 45 configurations and MAE in 44. The exception is Social Good/Crossformer, where MAE increases from 0.404 to 0.416 despite a lower MSE. These results demonstrate broad improvements across the evaluated backbones, while leaving the information source of these gains to the controlled intervention analysis.

Comparison with Multimodal and Reasoning Methods. Table 2 compares SPRA with six multimodal or reasoning methods across nine domains. SPRA achieves the lowest MSE in seven domains and the lowest MAE in eight domains. The largest relative MSE improvement over the next best method occurs on Climate (5.3% over ChatTime) and Agriculture (3.2% over Aurora). SPRA also leads both metrics on Security, indicating robustness under noisy textual context. The exceptions reveal metric specific trade-offs: on Energy, SPRA provides the best MAE but its MSE is 1.7% above Aurora; on Health-US, VoT achieves the best MSE; and on Social Good, SPRA leads MSE while CALF gives the lowest MAE.

4.3 Counterfactual Attribution

We apply the same diagnostic to SPRA and baselines (MM-TSFLib, TaTS and VoT). At test time, *text removal*, *semantic perturbation* (event content replaced, calendar expressions kept), and *date*

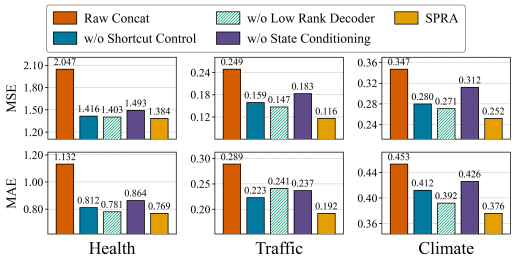


Figure 3: Architectural ablation on PatchTST.

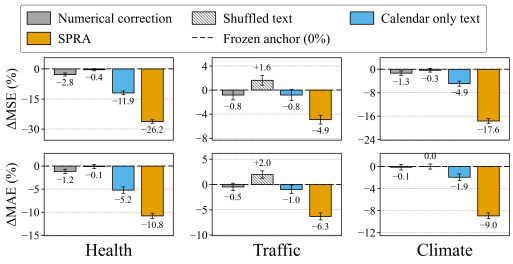


Figure 4: Text necessity controls.

masking (calendar expressions removed, content kept) act on the raw documents while parameters, numerical inputs, and calibration stay fixed. For each intervention $x \in \{\mathrm{rem}, \mathrm{sem}, \mathrm{date}\}$ we report the relative increase in test error,

$$\Delta_x = \frac{\mathrm{MAE}_x - \mathrm{MAE}}{\mathrm{MAE}} \times 100\%, \qquad \rho = \frac{\Delta_{\mathrm{date}}}{\Delta_{\mathrm{rem}}}, \qquad \sigma = \frac{\Delta_{\mathrm{sem}}}{\Delta_{\mathrm{rem}}}, \tag{7}$$

where MAE and MAE_x are the errors of the same model with the original and the intervened documents. Δ_{rem} measures sensitivity to text removal. For $\Delta_{\mathrm{rem}} > 0$, ρ and σ normalize the effects of date masking and semantic perturbation by this sensitivity. They characterize responses to the specified interventions, rather than additive fractions of predictive information. We examine these responses jointly with forecasting error and the retraining controls below to assess whether document content contributes predictive information.

The fusion methods lose most of their textual gain under date masking ($\bar{\rho} = 0.80$, 0.77) and almost none under semantic perturbation ($\bar{\sigma} = 0.02$): they reproduce the calendar shortcut in every domain. VoT is equally insensitive to semantics but carries little calendar information either; deleting its documents changes MAE by 0.9–1.6%, so in our reimplementation the reasoning pathway adds little rather than adding the wrong signal. The mechanisms differ, but none of the baselines derives its textual gain from the events the documents describe.

SPRA shows the opposite profile. The gains from text assistance remain nearly unchanged under date masking ($\bar{\rho} = 0.05$) but vanish under semantic perturbation ($\bar{\sigma} = 1.10$); $\sigma > 1$ indicates that perturbed reports misdirect the correction, which presupposes that the correction depends on event content. The low ρ is partly a consequence of construction: canonicalization alone (α_{pair}=0) leaves $\bar{\rho} = 0.23$ from underspecified cues such as named months, which paired consistency removes. A high σ, by contrast, is not implied by construction, as VoT relies little on dates yet ignores event content. The gains of SPRA thus stem from event content rather than calendar identity.

4.4 Ablation Studies

Architectural components. Figure 3 removes one stage of SPRA at a time. *Raw Concat* is worst everywhere, so gains do not follow from adding text alone. Removing *state conditioning* causes the largest MSE increase among the components in all three domains (1.384→1.493, 0.116→0.183, 0.252→0.312): the same report should not induce a fixed correction across numerical regimes. An unrestricted decoder in place of the *low-rank decoder* also degrades accuracy, most visibly in Traffic MAE. Removing *shortcut control* (surface view, no paired consistency) raises MSE by 2.3–37%, consistent with absolute calendar identity extrapolating poorly across a chronological split.

Text necessity controls. We retrain SPRA with three replacements for its text input (Figure 4): *numerical correction*, which applies the same correction stage to numerical history; *shuffled text*, which breaks document window alignment; and *calendar-only text*, which retains only calendar expressions. Numerical correction and shuffled text change MSE by less than 3% relative to the anchor, with shuffled text increasing Traffic MSE by 1.6%, whereas SPRA reduces MSE by 4.9%–26.2%. Calendar-only text recovers less than half of SPRA's MSE reduction in each domain, including 11.9% versus 26.2% on Health-US. None of these replacements reproduces SPRA's gains. Together with the intervention results in §4.3, these controls support the predictive contribution of aligned event content when reliance on textual calendar cues is suppressed.

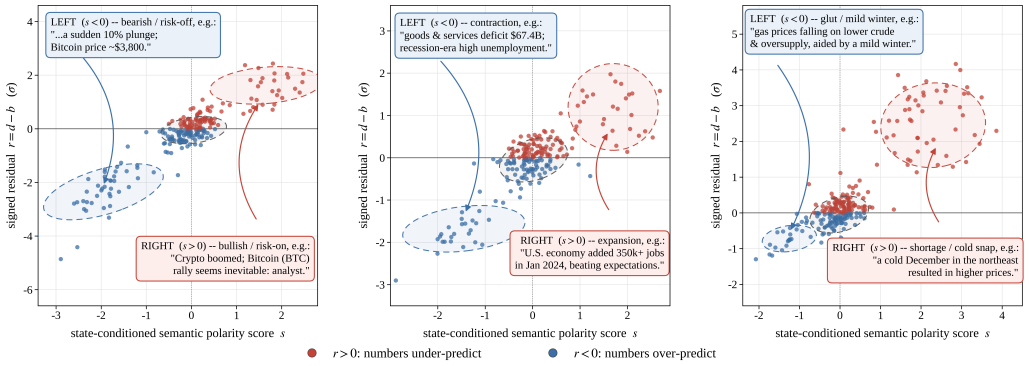


Figure 5: Alignment between state-conditioned semantic polarity and backbone forecast residuals.

4.5 Case Study

Figure 5 tests whether SPRA orders reports by the numerical anchor's errors. For each held-out window, we compute a *state-conditioned semantic polarity score* s: the standardized coefficient of the learned low-rank residual map's leading direction $\boldsymbol{\omega}_1$ (App. G). Its sign indicates an upward or downward anchor adjustment; $|s|$ measures evidence strength under the current numerical trajectory. Because s uses only documents and observed history, its relation to the realized residual $r = y - b$ is evaluated out of sample.

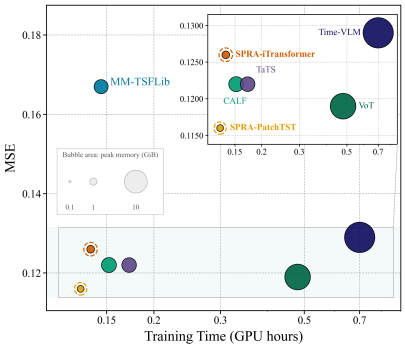


Figure 6: Comparison of efficiency on the Traffic dataset.

Across all three sectors, economy (left), social good (center), and energy (right), s tracks both the sign and the magnitude of the anchor residual. The positive end is populated by reports of cryptocurrency rallies, labor market expansion, and cold driven demand, for which the anchor, extrapolating only the observed trajectory, underpredicts the realized values. SPRA raises the forecast along the leading correction shape in proportion to s. Windows near the origin receive little correction, consistent with state conditioning attenuating reports whose implications are weak or already reflected in the series.

4.6 Efficiency

Figure 6 reports the Traffic domain trade-off among test MSE, total training time (GPU hours), and peak GPU memory (bubble area). SPRA-PatchTST and SPRA-iTransformer both reside in the optimal region, achieving the lowest MSE with training cost under 0.15 GPU hours and peak memory below 0.11 GiB. Relative to VoT, SPRA-PatchTST achieves comparable MSE while requiring nearly $4\times$ less training time and over two orders of magnitude less memory. CALF and Time-VLM incur substantially higher compute and memory budgets for similar or inferior accuracy. The same holds with iTransformer, so the advantage comes from the residual design rather than the backbone.

5 Conclusion

We identify a calendar shortcut in text-assisted forecasting: the evaluated fusion models show substantial error increases under date masking but little change under the tested semantic perturbations. SPRA combines shortcut controlled text views with state-conditioned, low-rank residual learning to correct a frozen numerical backbone. Experiments across nine domains and five backbones show broad forecasting gains, with intervention and retraining controls supporting the contribution of aligned event content. Future work will examine transfer across document sources and temporal regimes, and calibrate correction uncertainty to determine when unreliable or conflicting reports should leave the numerical forecast unchanged.

AI USE STATEMENT

We used GPT-5.6 for language polishing of the manuscript, including improvements to grammar, wording, and readability. AI assistance was also used for LaTeX typesetting and for searching relevant research literature to help ensure that the literature review was sufficiently comprehensive. No AI tools were used for research ideation, hypothesis formulation, or methodological or experimental design. The authors take full responsibility for the final manuscript, including all AI-assisted edits, and for the accuracy and integrity of its scientific content.

REPRODUCIBILITY STATEMENT

The source code has been uploaded as supplementary material. App. A gives the complete SPRA specification, including the paired-view construction rules (App. A.1), the state features (App. A.2), and the closed-form estimator of the residual map (App. A.5). App. B states the chronological fitting, selection, and calibration protocol.

REFERENCES

Defu Cao, Furong Jia, Sercan O. Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YH5w12OUuU.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8979–9007. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25be.html.

Wentao Gao, Xiaojing Du, Xiongren Chen, Yifan Guo, Andres Mauricio Cifuentes-Bernal, Renqiang Luo, and Ziqi Xu. Energy-efficient training-free zero-inflation correction for rainfall forecasting with time-series foundation models. In *Proceedings of the ACM Web Conference 2026*, pp. 8840–8850, 2026.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Shiyan Hu, Jianxin Jin, Yang Shu, Peng Chen, Bin Yang, and Guo Guo. Towards multimodal time series anomaly detection with semantic alignment and condensed interaction. In *International Conference on Learning Representations*, pp. 92903–92933, 2026.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. GPT4MTS: Prompt-based large language model for multimodal time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23343–23351, 2024. doi: 10.1609/aaai.v38i21.30383. URL https://ojs.aaai.org/index.php/AAAI/article/view/30383.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Unb5CVPtae.

Emirhan Kurtuluş, Zichao Li, Yann Dauphin, and Ekin Dogus Cubuk. Tied-augment: controlling representation similarity improves data augmentation. In *International Conference on Machine Learning*, pp. 17994–18007. PMLR, 2023.

Xingwang Li, Fei Teng, Tianrui Li, and Qiang Duan. Preference guided meta-learning for cross domain time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2026a.

Xingwang Li, Fei Teng, Xin Wu, and Qiang Duan. Uncovering hidden degeneration: A physics-guided bidirectional inference framework for industrial time series prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23204–23211, 2026b.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=a1zBg9cBvt.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-2476. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8e7768122f3eeec6d77cd2b424b72413-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

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

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=JePfAI8fah.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399. PMLR, 2024c. URL https://proceedings.mlr.press/v235/liu24cb.html.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-context transformers for unified time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=KMCJXjlDDr.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 2025c. URL https://proceedings.mlr.press/v267/liu25be.html.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Huu Hiep Nguyen, Minh Hoang Nguyen, Dung Nguyen, and Hung Le. Does text actually help? Uncovering and resolving text collapse in multimodal time series forecasting. *arXiv preprint arXiv:2606.19413*, 2026.

Zelin Ni, Hang Yu, Shizhan Liu, Jianguo Li, and Weiyao Lin. BasisFormer: Attention-based time series forecasting with learnable and interpretable basis. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

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

Zijie Pan, Yushan Jiang, Sahil Garg, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. S^2IP-LLM: Semantic space informed prompt learning with LLM for time series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39135–39153. PMLR, 2024. URL https://proceedings.mlr.press/v235/pan24c.html.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=e1wDDFmlVu.

Jannik Strötgen and Michael Gertz. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 321–324, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1071/.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. ChatTime: A unified multimodal time series foundation model bridging numerical and textual data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):12694–12702, 2025. doi: 10.1609/aaai.v39i12.33384. URL https://ojs.aaai.org/index.php/AAAI/article/view/33384.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

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

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

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VVJ6Ck9JBl.

Ronghui Xu, Jihao Chen, Jindong Tian, Chenjuan Guo, and Bin Yang. Most: A foundation model for multi-modality spatio-temporal traffic prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1716–1727, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

Tengxue Zhang, Biao Ouyang, Yang Shu, Xinyang Chen, Guo Guo, and Bin Yang. Swiftts: A swift selection framework for time series pre-trained models via multi-task meta-learning. In *International Conference on Learning Representations*, volume 2026, pp. 102214–102234, 2026.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78478–78497. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhong25a.html.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

Appendix Overview

The appendix completes the specification of SPRA (App. A), states the chronological fitting, selection, and calibration protocol (App. B), and describes the data and the document-availability rule (App. C). It then details the experiments of §4: the calendar-shortcut diagnostic and counterfactual attribution (App. D), the architectural ablation (App. E), the text-necessity controls (App. F), and the case-study score (App. G). Implementation, baseline, and efficiency details follow in App. H, and limitations in App. I.

A Complete SPRA Specification

A.1 Paired Text Construction

Two roles of temporal expressions. Absolute document dates and recurring report boilerplate can reveal calendar phase even when event content is ignored, whereas deleting every temporal expression would also destroy information about *when* an event occurs. The paired views separate these two cases with a deterministic, rule-based transformation that follows standard temporal-expression normalization practice (Strötgen & Gertz, 2010). No generative model is queried at any stage, so the transformation cannot introduce information that is absent from the original document. Each resolvable temporal expression receives one of two roles:

- a *document-time* mention stamps the document itself: a dateline, a release date, or a reporting-period expression such as "as of", "week ending", or "during the week of";
- an *event-time* mention situates an event, typically as the temporal argument of an event predicate ("will fall on July 12", "expected in March 2024").

An expression is assigned document time if it coincides with the document timestamp at the domain's reporting granularity or occurs in one of the reporting-period patterns; every other resolvable expression is treated as event time.

Construction. For a document $D_{t,i}$ available at forecast origin t, the two views are obtained as follows.

1. *Detection and normalization.* Fully specified calendar expressions (day–month–year dates, ISO timestamps, month–year expressions, and explicit date ranges) are detected and normalized to calendar positions.
2. *Role assignment.* Each normalized expression is labelled document time or event time by the rule above.
3. *Content view* $D^{c}_{t,i}$. Document-time mentions are replaced by the generic token `[DOC_TIME]`. Event-time mentions are converted to a signed offset from t, bucketed at the domain's native frequency (e.g., `[REL_FUTURE_2D]`, `[REL_PAST_1W]`), and verbalized

Table 4: Paired-view construction. The surface view is the original document; the content view withholds absolute calendar identity while preserving event content and relative timing. Example (c) is the illustration of Figure 2 at a July 10 origin; (a) and (b) are Time-MMD records (App. D.3).

	Surface view D^{s} (original)	Content view D^{c}
(a)	The national average retail gasoline price increased to \$3.258 per gallon *on December 26, 2011*, \$0.029 more than last week.	The national average retail gasoline price increased to \$3.258 per gallon `[DOC_TIME]`, \$0.029 more than last week.
(b)	The proportion of flu patients was 2.5% during week 44 *(October 28–November 3, 2007)*.	The proportion of flu patients was 2.5% during week 44 (`[DOC_TIME]`).
(c)	*Weekly report 2024-07-10.* Shipments will fall *on July 12* after a pipeline outage.	Shipments will fall in two days after a pipeline outage.

In (b), the underspecified identifier "week 44" is retained; it is a residual calendar cue of the kind discussed above and is removed by the date-masking intervention of App. D.1.

before encoding ("in two days", "one week earlier"), so that the frozen encoder receives natural language. Recurring report shells are removed (see below). Entities, quantities, directional predicates, intensities, and relative lead times are kept verbatim.

4. *Surface view* $D^{\mathrm{s}}_{t,i}$. The original document after the same deduplication: it carries identical event content, but retains every absolute calendar expression and the report shell.

Table 4 shows the transformation on three documents.

Underspecified expressions. Expressions that cannot be resolved to a unique calendar position without further context—bare month names in running prose, week numbers without a year, seasons, and holidays—are left unchanged, because an incorrect resolution would corrupt event timing. They therefore remain in the content view and are the main source of the residual calendar reliance measured at $\alpha_{\mathrm{pair}}{=}0$ in Table 3. Because sentence embeddings of explicit dates and of such expressions share directions, penalizing the response along the explicit surface displacement also attenuates the response to them (App. A.5). Table 3 measures this effect rather than assuming it: $\bar{\rho}$ falls from 0.23 at $\alpha_{\mathrm{pair}}{=}0$ to 0.05 with paired consistency.

Report shells. Shell handling is deliberately conservative and is not part of the calendar-shortcut definition; we posit no semantic taxonomy of templates. A span is removed only if it is near-identical boilerplate that (i) recurs in at least eight training windows, (ii) occurs in at least 30% of those windows, and (iii) appears in a stable document position. The detector is fitted on the chronologically earliest training documents only. If no stable shell is detected, the two views differ only in absolute temporal surface forms. The interventions in §4.3 and App. D concern explicit calendar cues and do not depend on treating arbitrary prose as a template.

Encoding and aggregation. At most 24 documents available at t are used. After duplicate removal, each view is encoded by the frozen sentence encoder $e(\cdot)$ and aggregated as

$$\boldsymbol{z}^{v}_{t} = \frac{\sum_{i\in\mathcal{I}_t} e(D^{v}_{t,i})}{\left\|\sum_{i\in\mathcal{I}_t} e(D^{v}_{t,i})\right\|_2 + \epsilon}, \qquad v \in \{\mathrm{c},\mathrm{s}\}, \tag{8}$$

where $\mathcal{I}_t$ indexes documents available no later than t and ϵ is a small constant. If $\mathcal{I}_t = \emptyset$, both text vectors are set to zero and the window is flagged as text-free (App. A.4). The surface displacement $\Delta\boldsymbol{z}_t = \boldsymbol{z}^{\mathrm{s}}_t - \boldsymbol{z}^{\mathrm{c}}_t$ isolates the calendar and shell information that the content view withholds. Only $\boldsymbol{z}^{\mathrm{c}}_t$ receives residual supervision and is used at test time; the surface view enters solely through the paired-consistency term of Eq. (5). Because e is frozen, that term cannot be satisfied by collapsing the encoder so that the two views coincide.

A.2 Numerical State Features

The state descriptor uses only the target history in $\boldsymbol{X}_t$ and is inexpensive to compute. Table 5 makes the three groups of §3.2 explicit. Every quantity is computed independently for each window before the training-only projection R_s is fitted.

The recent segment adapts to the lookback and contains 4, 8, and 16 observations for $L = 8$, 36, and 96, respectively. Standard deviations below 10^{-6} are replaced by one before normalization, and

Table 5: Components of the compact trajectory state.

Group	Recorded quantities	Role
Global $\boldsymbol{s}_t^{\mathrm{g}}$	endpoint; window mean and scale; least-squares slope; normalized minimum and maximum; coarse down-sampled shape	overall level, direction, and broad geometry of the observed trajectory
Periodic $\boldsymbol{s}_t^{\mathrm{p}}$	normalized low-frequency Fourier energy; three coarse frequency-band energies; autocorrelation at valid training-derived seasonal lags	recurrence, recorded only when the look-back contains enough observations to estimate it
Local $\boldsymbol{s}_t^{\ell}$	mean, variation, slope, root-mean-square change, endpoint, and latest difference in the most recent segment	distinguishes a stable regime from a change already visible near the origin

candidate autocorrelation lags outside $[1, L-1]$ contribute zeros. The concatenated descriptor is reduced to $d_s = 16$ dimensions by a projection fitted only on training histories, so $\boldsymbol{u}_t \in \mathbb{R}^{d_s}$ in §3.2. The construction avoids long-window statistics, whose estimates would be unreliable for the short $L = 8$ tasks.

A.3 Standardization and Shared Cross-Alignment

Let $\boldsymbol{\mu}_u, \boldsymbol{\sigma}_u$ and $\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z$ denote element-wise means and standard deviations computed on adapter-training examples. The standardized features of §3.3 are

$$\widetilde{\boldsymbol{u}}_t = (\boldsymbol{u}_t - \boldsymbol{\mu}_u) \oslash \max(\boldsymbol{\sigma}_u, 10^{-6}), \qquad \widetilde{\boldsymbol{z}}_t^v = (\boldsymbol{z}_t^v - \boldsymbol{\mu}_z) \oslash \max(\boldsymbol{\sigma}_z, 10^{-6}), \tag{9}$$

with the maximum taken element-wise. SCA computes the singular value decomposition of the content-view cross-covariance C_{uz} of §3.3 and keeps its r_{ca} leading singular triplets, which give P_u, Σ_r, and P_z; this truncation is the approximation $C_{uz} \approx P_u \Sigma_r P_z^\top$ of the main text. Only the content view defines these axes; the same fitted P_z then projects both views. Paired consistency therefore compares predictions in one coordinate system, and the surface view cannot rotate the alignment toward calendar identity.

A.4 Text Necessity

The routed feature of Eq. (3) has dimension $d_h = d_z + 2r_{\mathrm{ca}}$. The following property makes precise the statement of §3.4 that the state alone cannot produce a correction.

Proposition 1 (Text necessity)**.** *For every view v, origin t, and routing channel j, $|m_{t,1,j}^v| \le |q_{t,j}^v|$ and $|m_{t,2,j}^v| \le |q_{t,j}^v|$. If no document is available at t, then $\boldsymbol{c}_t^v = \mathbf{0}$ and the deployed forecast of Eq. (6) equals the anchor $\boldsymbol{b}_t$, for every state $\boldsymbol{u}_t$.*

Proof. Bound. By Eq. (3), $m_{t,1,j}^v = q_{t,j}^v \tanh(a_{t,j})$ and $m_{t,2,j}^v = q_{t,j}^v/(1 + |a_{t,j} - q_{t,j}^v|)$. Since $|\tanh(\cdot)| \le 1$ and $1 + |a - q| \ge 1$, both $|m_{t,1,j}^v|$ and $|m_{t,2,j}^v|$ are at most $|q_{t,j}^v|$.

Necessity. Standardizing a zero vector would give $-\boldsymbol{\mu}_z \oslash \boldsymbol{\sigma}_z \neq \mathbf{0}$. Text-free windows are therefore flagged *before* standardization, and $\widetilde{\boldsymbol{z}}_t^v$ and $\boldsymbol{q}_t^v = P_z^\top \widetilde{\boldsymbol{z}}_t^v$ are reset to zero *after* it. By the bound, $\boldsymbol{m}_{t,1}^v = \boldsymbol{m}_{t,2}^v = \mathbf{0}$, hence $\boldsymbol{h}_t^v = \mathbf{0}$. The decoder $\boldsymbol{c}_t^v = BR^\top \boldsymbol{h}_t^v$ of Eq. (4) is linear and has no additive bias, so $\boldsymbol{c}_t^v = \mathbf{0}$ and, by Eq. (6), $\widehat{\boldsymbol{y}}_t = \boldsymbol{b}_t + \boldsymbol{\lambda} \odot \mathbf{0} = \boldsymbol{b}_t$. Neither step depends on $\boldsymbol{u}_t$. □

The first part implies that the numerical state can attenuate a text channel or reverse its sign, but cannot enlarge it beyond the text activation. Consequently, the removal intervention of §4.3 returns the frozen anchor exactly, and Δ_{rem} of SPRA in Table 3 is fixed by the uni-modal and SPRA entries of Table 1: for PatchTST, $\Delta_{\mathrm{rem}} = 9.8$, 12.1, and 18.4% on Climate, Health-US, and Energy are the relative MAE gains $0.413/0.376$, $0.862/0.769$, and $0.419/0.354$ minus one.

A.5 Analytical Estimation of the Low-Rank Residual Map

The adapter parameters of §3.5 are estimated in closed form from the chronological training residuals rather than by gradient descent. Stack the routed content features into $H_{\mathrm{c}} \in \mathbb{R}^{N \times d_h}$, the

surface features into H_s, and the residual targets into $Y \in \mathbb{R}^{N\times H}$. With $\Delta = H_s - H_c$ and $A = RB^\top \in \mathbb{R}^{d_h\times H}$, Eq. (5) becomes

$$\min_{\mathrm{rank}(A)\le k} \frac{1}{N}\|Y - H_c A\|_F^2 + \frac{\alpha_{\mathrm{pair}}}{N}\|\Delta A\|_F^2 + \eta\|A\|_F^2. \tag{10}$$

Define

$$Q = \frac{1}{N}H_c^\top H_c + \frac{\alpha_{\mathrm{pair}}}{N}\Delta^\top\Delta + \eta I, \qquad C = \frac{1}{N}H_c^\top Y, \tag{11}$$

and let $Q = LL^\top$ be a Cholesky factorization, which exists because Q is positive definite for $\eta > 0$. Substituting $G = L^\top A$ turns the objective into $\|G - L^{-1}C\|_F^2$ up to a constant, and $\mathrm{rank}(G) = \mathrm{rank}(A)$. By the Eckart–Young theorem, if $L^{-1}C = U\Sigma V^\top$, the globally optimal solution of Eq. (10) is

$$\widehat{A}_k = L^{-\top}U_{1:k}\Sigma_{1:k}V_{1:k}^\top. \tag{12}$$

A compact SVD $\widehat{A}_k = \bar{U}\bar{\Sigma}\bar{V}^\top$ yields one deployable factorization, $R = \bar{U}\bar{\Sigma}^{1/2}$ and $B = \bar{V}\bar{\Sigma}^{1/2}$, so that $A = RB^\top$ and the column-vector map is $BR^\top$. Only the product is identifiable; joint sign or rotation changes of R and B leave predictions unchanged.

Role of the consistency term. Eq. (11) shows that α_{pair} enters only through $\Delta^\top\Delta$: paired consistency is a generalized ridge penalty that shrinks the residual map along the directions in which the two views differ. It leaves directions carrying event content, which the views share, unpenalized. This is why it can lower calendar reliance with almost no change in clean accuracy: in Table 3, moving from α_{pair}=0 to the default lowers $\bar{\rho}$ from 0.23 to 0.05, while Δ_{rem}, which by Proposition 1 is each variant's relative MAE gain over the anchor, changes by at most 0.2 percentage points in every domain.

B Chronological Fitting, Selection, and Calibration

B.1 Chronological Residual Targets

The outer split is chronological: the first 70% of observations form $\mathcal{D}_{\mathrm{train}}$, the next 10% form $\mathcal{D}_{\mathrm{cal}}$, and the last 20% form $\mathcal{D}_{\mathrm{test}}$. Forecast windows that would cross a boundary are discarded. Within $\mathcal{D}_{\mathrm{train}}$, anchor residuals are produced by fitting the numerical backbone on an earlier prefix and predicting a later block; if an origin predicts indices $t{+}1, \ldots, t{+}H$, the block boundary is purged so that none of those labels overlaps the next block. These are the blocked out-of-fold errors of §4.1: every residual target is an error on observations unseen by the backbone fit that produced it. The complete sequence is:

1. fit each numerical backbone on an earlier prefix and collect later-block residuals inside $\mathcal{D}_{\mathrm{train}}$;
2. fit SCA and the analytical residual map for each candidate specification on an inner fitting block;
3. rank candidates by corrected MSE on later, non-overlapping development blocks and freeze the selected specification;
4. re-estimate SCA, B, and R on all chronological training residuals;
5. estimate the shrinkage $\boldsymbol{\lambda}$ on $\mathcal{D}_{\mathrm{cal}}$ without changing the representation or the residual map;
6. evaluate $\mathcal{D}_{\mathrm{test}}$ once, with every fitted quantity frozen.

Future targets serve as residual labels only in Steps 1–4 and as validation responses in Step 5. At deployment the model receives only the observed history and the documents available at the forecast origin.

B.2 Hyperparameter Selection

All SPRA hyperparameters form one shared configuration (Table 9 in App. H.1); no domain, backbone, or test split receives separately tuned values. The alignment width r_{ca} and the ridge coefficient η are selected in Step 3 by corrected MSE macro-averaged over the development blocks of all domains.

B.3 Chronological Calibration

The later validation block of §3.5 is $\mathcal{D}_{\mathrm{cal}}$. For forecast step h, the least-squares scale on $\mathcal{D}_{\mathrm{cal}}$ and its clipped version are

$$\tilde{\lambda}_h = \frac{\sum_{t \in \mathcal{D}_{\mathrm{cal}}} c^{\mathrm{c}}_{t,h}\, r_{t,h}}{\sum_{t \in \mathcal{D}_{\mathrm{cal}}} (c^{\mathrm{c}}_{t,h})^2 + \varepsilon}, \qquad \lambda_h = \mathrm{clip}(\tilde{\lambda}_h, 0, 1), \tag{13}$$

with $\varepsilon = 10^{-5}$. The total correction energy on $\mathcal{D}_{\mathrm{cal}}$ is further capped at 5% of the validation residual energy. Both operations can only attenuate the learned correction, and the same $\boldsymbol{\lambda}$ is used for every test window; no test response affects it. The numerical-correction control of App. F uses the calibration block in exactly the same way.

C Datasets, Document Availability, and Preprocessing

Table 6 summarizes the nine Time-MMD domains used in all experiments. Text records count report and search files together.

Table 6: Dataset statistics after timestamp deduplication. d is the number of retained numerical variables.

Domain	Rows	d	Date range	L	Forecast lengths	Text records
Agriculture	532	1	1980-01-01 – 2024-04-01	8	6/8/10/12	2543
Climate	1272	6	2000-01-04 – 2024-05-14	8	6/8/10/12	2851
Economy	447	3	1987-01-01 – 2024-03-01	8	6/8/10/12	2749
Energy	1622	9	1993-04-05 – 2024-04-29	36	12/24/36/48	2668
Environment	15979	2	1980-01-01 – 2023-09-30	96	48/96/192/336	2468
Health-US	1384	10	1997-09-29 – 2024-05-06	36	12/24/36/48	2951
Security	309	1	1998-09-01 – 2024-05-01	8	6/8/10/12	21880
Social Good	916	1	1948-01-01 – 2024-04-01	8	6/8/10/12	4961
Traffic	651	1	1970-01-01 – 2024-03-01	8	6/8/10/12	2676

Document availability. A document is admitted at forecast origin t only if its publication timestamp—for Time-MMD reports, the end of the reporting period—is no later than t. Documents dated after t are never visible, in training or in testing. No text is generated: every document is a Time-MMD retrieval, and the content views and interventions only rewrite retrieved documents. The same rule and the same document sets are used for every method, including all baselines (App. H.2).

Numerical preprocessing. Rows are sorted by timestamp, and duplicate timestamps are averaged for numeric columns. A column is retained only if its coverage on the training prefix is at least 70% and it takes more than one distinct value. Infinite values are treated as missing; missing entries are forward-filled, then filled with the training median, and finally with zero only if no training statistic exists. Centering and scaling use the first 70% of observations exclusively, with a 10^{-6} scale floor. No observation is removed on the basis of validation or test behavior. Windows use stride one and preserve chronological order.

D Calendar-Shortcut Diagnostic and Counterfactual Attribution

D.1 Interventions

All interventions act on the raw documents of a trained model at test time; parameters, numerical inputs, and calibration are held fixed, and the same intervened documents are supplied to every method before any model-specific processing (for SPRA, before paired-view construction; for VoT, before event extraction). None of the interventions is used for training or model selection.

- **Text removal** replaces the document set by the empty set.

- **Semantic perturbation** replaces the event content while keeping the calendar surface. Each window receives the reports of another window of the same domain under a fixed derangement drawn before any test target is evaluated, and forecast-relevant directional predicates are reversed with a fixed antonym lexicon (e.g., increase/decrease, rise/fall, above/below, surplus/deficit, stable/volatile). The receiving window's own calendar expressions are then re-inserted in place of those of the donor report.
- **Date masking** replaces absolute dates, named months, week identifiers, and timestamp-like expressions by a neutral `[DATE]` token, while keeping quantities, entities, directional predicates, and relative expressions such as "last week".

For each model, intervention, and domain, Δ is the relative change of MAE, averaged over the four forecast lengths, with respect to the same model's unmodified predictions; ρ and σ follow Eq. (7). The ratios are descriptive. When Δ_{rem} is small they are dominated by noise, so such rows are best read through the raw changes.

D.2 Full Attribution Results

Table 7 reports the per-domain changes behind Table 3 with standard errors over 20 seeds, together with the per-domain ratios. MM-TSFLib, TaTS, and SPRA use the PatchTST backbone, as in Table 3. The MM-TSFLib and TaTS rows are the settings of Figure 1, where values are rounded to integers. Across these six settings, date masking reproduces a median 79% of the removal penalty (range 43–115%). A ratio above one (Energy, MM-TSFLib) means that masking dates harms the model more than deleting the text: the masked documents are not only uninformative but also out of distribution for a model that learned to read their timestamps.

The calendar-shortcut claim therefore rests on the two fusion forecasters, for which it holds in all three domains, and App. D.4 reproduces the same profile for five further multimodal methods. Removing VoT's documents raises its MAE by at most 1.6%, so its text pathway carries little information of either kind; the broader observation that the textual gain does not come from event content applies to every baseline we examine.

Table 7: Per-domain counterfactual attribution (relative MAE change in %, mean $\pm$ s.e. over 20 seeds; the domain averages of ρ and σ are $\bar{\rho}$ and $\bar{\sigma}$ in Table 3). For VoT, Δ_{rem} is below 2% in every domain, so its ratios are dominated by noise.

Domain	Method	Δ_{rem}	Δ_{sem}	Δ_{date}	ρ	σ
Climate	MM-TSFLib	7.2 ± 0.4	0.4 ± 0.2	3.1 ± 0.3	0.43	0.06
	TaTS	23.8 ± 0.6	0.5 ± 0.3	19.2 ± 0.5	0.81	0.02
	VoT	0.9 ± 0.2	0.0 ± 0.1	0.0 ± 0.1	0.00	0.00
	SPRA, α_{pair}=0	9.9 ± 0.3	8.7 ± 0.4	2.4 ± 0.2	0.24	0.88
	SPRA	9.8 ± 0.3	10.9 ± 0.4	0.5 ± 0.1	0.05	1.11
Health-US	MM-TSFLib	16.8 ± 0.6	0.6 ± 0.3	13.9 ± 0.5	0.83	0.04
	TaTS	25.1 ± 0.7	0.7 ± 0.3	19.2 ± 0.6	0.76	0.03
	VoT	1.6 ± 0.3	0.1 ± 0.1	0.0 ± 0.1	0.00	0.06
	SPRA, α_{pair}=0	12.2 ± 0.4	10.6 ± 0.5	3.0 ± 0.3	0.25	0.87
	SPRA	12.1 ± 0.4	12.6 ± 0.5	0.7 ± 0.2	0.06	1.04
Energy	MM-TSFLib	7.1 ± 0.4	-0.3 ± 0.2	8.2 ± 0.4	1.15	-0.04
	TaTS	22.8 ± 0.6	0.2 ± 0.2	16.9 ± 0.5	0.74	0.01
	VoT	1.1 ± 0.2	-0.1 ± 0.1	0.1 ± 0.1	0.09	-0.09
	SPRA, α_{pair}=0	18.6 ± 0.5	16.9 ± 0.6	3.5 ± 0.3	0.19	0.91
	SPRA	18.4 ± 0.5	21.3 ± 0.7	0.8 ± 0.2	0.04	1.16

D.3 Auditable Examples

The following Time-MMD records are quoted from the released data together with the observed target values surrounding each report. They show that date masking changes calendar identity without deleting the forecast-relevant predicate.

Table 8: Calendar-shortcut diagnostic for five further multimodal methods: relative MAE change (%) with respect to each model's own unmodified paired-text predictions, averaged over four forecast lengths in each of Climate, Health-US, and Energy.

Method	Text removal Δ_{rem}	Semantic perturbation Δ_{sem}	Date masking Δ_{date}
CALF	$+8.0$	-0.1	$+5.6$
ChatTime	$+10.3$	-0.1	$+7.9$
Aurora	$+4.3$	0.0	$+3.2$
Time-VLM	$+1.1$	0.0	$+0.8$
Time-LLM	$+3.3$	0.0	$+2.7$

Climate. Original: "As of November 2000, drought affected approximately 7% of the contiguous United States, with abundant rains bringing drought relief to the southern Plains and Deep South." Date-masked: "As of `[DATE]`, drought affected approximately 7% of the contiguous United States, with abundant rains bringing drought relief ..." The antonym step of the semantic perturbation replaces "bringing drought relief" with "intensifying drought conditions" and keeps the calendar phrase. The surrounding target values rise from 49.98, 62.07, 69.11, 70.05 to 70.22, 71.61, 72.21, 75.39.

Health-US. Original: "The proportion of flu patients was 2.5% during week 44 (October 28–November 3, 2007)." The date-masked view keeps the measured rate but replaces the week identifier and the date range. The target tail is 1.36, 1.45, 1.53, 1.67, 1.97, illustrating why event level and calendar phase are both potentially predictive unless they are explicitly separated.

Energy. Original: "The national average retail gasoline price increased to $3.258 per gallon on December 26, 2011, $0.029 more than last week." The date-masked view keeps the increase, the level, and the weekly comparison and replaces the absolute date; the antonym step of the semantic perturbation changes increased/more to decreased/less without altering the calendar expression. The surrounding sequence is 3.368, 3.350, 3.346, 3.290, 3.317, 3.358, 3.441.

Residual calendar cues. The canonicalizer rewrites fully specified dates but keeps underspecified expressions (App. A.1), such as "week 44" above or "a cold December in the northeast" and "aided by a mild winter" in the Energy reports of Figure 5. Such expressions carry the calendar reliance that remains at $\alpha_{\mathrm{pair}}{=}0$ and are removed by date masking.

D.4 Extension to Further Multimodal Methods

The diagnostic is not specific to SPRA or to the two fusion baselines of Table 3. We therefore applied the same three interventions of App. D.1 to the trained CALF, ChatTime, Aurora, Time-VLM, and Time-LLM models of App. H.2, holding parameters, numerical inputs, and calibration fixed. Table 8 averages the four forecast lengths in each of Climate, Health-US, and Energy, that is, 12 tasks per model; VoT is reported in Table 3.

The clearest contrast appears in ChatTime. Removing its paired text increases MAE by 10.3%, which would suggest a substantial benefit from the text input. Yet reversing the reported event directions changes MAE by only -0.1%, while masking calendar expressions increases it by 7.9%. A large text-removal penalty therefore does not, by itself, establish that a model uses the events that the documents describe. The same ordering holds for CALF, Aurora, Time-VLM, and Time-LLM: date masking has a much larger effect than semantic perturbation. For the four methods whose removal penalty exceeds 3%, the ratio of the date-masking penalty to the text-removal penalty ranges from 0.70 to 0.82, close to the $\bar{\rho}$ of the two fusion baselines in Table 3. Time-VLM is left out of that range: its Δ_{rem} of 1.1% is small enough for the ratio to be dominated by noise, the same reservation that App. D.1 records for VoT. These interventions separate two properties that a modality ablation alone cannot: whether paired text improves accuracy, and whether that improvement depends on the tested event content.

A plausible explanation for this shared pattern is that temporally aligned documents expose calendar expressions from which a forecaster can recover periodic phase. Using those expressions may be easier than identifying which heterogeneous event reports are relevant to the target and how their implications depend on the current numerical state. Because the pattern appears across architec-

turally different methods, greater model complexity alone does not ensure that forecasting gains come from event information. The table records input–output responses and does not identify the internal features that each model uses.

E Architectural Ablation Details

The ablations in Figure 3 change one mechanism at a time. *Raw Concat* replaces SCA and TASR with direct concatenation of $\widetilde{\boldsymbol{u}}_t$ and $\widetilde{\boldsymbol{z}}_t^{\mathrm{c}}$. *Without shortcut control* trains and tests on the surface view without paired consistency; it differs from $\alpha_{\mathrm{pair}} = 0$ in Table 3, which keeps the content view and removes only the consistency term. *Without state conditioning* maps the content text directly to the residual. *Without the low-rank decoder* replaces $BR^\top$ by an unrestricted linear residual map. The frozen anchor, the text encoder, and the splits are shared by all variants.

F Text-Necessity Controls

Each control in Figure 4 is retrained with its textual input replaced; the frozen PatchTST anchor, the splits, the chronological residual targets, and the calibration of App. B.3 are shared with SPRA. The controls are ordered by the aligned information they retain.

- **Numerical correction (no text).** The documents are replaced by the numerical history of the window: the routed feature becomes $\boldsymbol{h}_t = [\widetilde{\boldsymbol{u}}_t; \widetilde{\boldsymbol{x}}_t]$, the standardized state descriptor (App. A.2) concatenated with the standardized lookback values of the target. Because every state-dependent term of the router is anchored on text (Proposition 1), these features enter the rank-k residual map directly (Eq. (10) with $\alpha_{\mathrm{pair}} = 0$). The control therefore has the same correction stage, training residuals, and calibration block as SPRA and differs only in the absence of text: any gain attributable to the additional stage or to the calibration data would appear here.
- **Shuffled text.** The document sets of training windows are permuted by one fixed random permutation and those of test windows by an independent one, which breaks the alignment between documents and windows; content and surface views are permuted jointly. The pipeline is otherwise identical to SPRA.
- **Calendar-only text.** Each document is replaced by the concatenation of its calendar expressions, both fully specified and underspecified (App. A.1), encoded by the same frozen encoder, and passed through the SPRA pipeline with $\alpha_{\mathrm{pair}} = 0$, since no content view remains to be paired.

The calibration of App. B.3 is estimated on $\mathcal{D}_{\mathrm{cal}}$, where it can only attenuate a correction; it does not force a control without usable information to reproduce the anchor on the test split. Numerical correction and shuffled text stay within 3% of the anchor in MSE, and shuffled text raises Traffic MSE by 1.6% (Figure 4), consistent with a small correction that does not generalize. Calendar-only text improves on the anchor but recovers less than half of the MSE reduction of SPRA: 11.9% versus 26.2% on Health-US, 0.8% versus 4.9% on Traffic, and 4.9% versus 17.6% on Climate. Figure 4 reports means over 20 seeds with standard-error bars.

G Case-Study Score

Definition. The factors B and R of §3.5 are identifiable only up to a joint invertible transformation (App. A.5), so for $k > 1$ the coefficient $\beta_{t,1}^{\mathrm{c}}$ of Eq. (4) is not uniquely defined. We therefore read the correction through the singular value decomposition of the product, which is unique up to sign when the singular values are distinct:

$$BR^\top = \sum_{j=1}^{k} \gamma_j \boldsymbol{\psi}_j \boldsymbol{\omega}_j^\top, \qquad \boldsymbol{c}_t^{\mathrm{c}} = \sum_{j=1}^{k} s_{t,j}\, \boldsymbol{\psi}_j, \qquad s_{t,j} = \gamma_j \boldsymbol{\omega}_j^\top \boldsymbol{h}_t^{\mathrm{c}}, \tag{14}$$

with singular values $\gamma_1 \geq \cdots \geq \gamma_k$, orthonormal correction shapes $\boldsymbol{\psi}_j \in \mathbb{R}^H$ spanning the column space of B, and orthonormal read-out directions $\boldsymbol{\omega}_j \in \mathbb{R}^{d_h}$. Figure 5 displays the leading coefficient

$s_t = s_{t,1}$, standardized over the displayed windows: this is the state-conditioned semantic polarity score s of §4.5, and $\boldsymbol{\omega}_1$ is the leading direction referred to there. The definition does not require the adapter to have rank one; the figure shows the dominant direction of the rank-k map. The sign of $(\boldsymbol{\psi}_1, \boldsymbol{\omega}_1)$ is fixed so that $s_t > 0$ raises the forecast on average over the horizon.

Why the score is state-conditioned. Partitioning $\boldsymbol{\omega}_1 = [\boldsymbol{\omega}_1^z; \boldsymbol{\omega}_1^{(1)}; \boldsymbol{\omega}_1^{(2)}]$ conformably with $\boldsymbol{h}_t^{\mathrm{c}} = [\widetilde{\boldsymbol{z}}_t^{\mathrm{c}}; \boldsymbol{m}_{t,1}^{\mathrm{c}}; \boldsymbol{m}_{t,2}^{\mathrm{c}}]$ in Eq. (3) gives

$$s_t = \gamma_1 \Big(\boldsymbol{\omega}_1^{z\top} \widetilde{\boldsymbol{z}}_t^{\mathrm{c}} + \boldsymbol{\omega}_1^{(1)\top} \big(\boldsymbol{q}_t^{\mathrm{c}} \odot \tanh(\boldsymbol{a}_t) \big) + \boldsymbol{\omega}_1^{(2)\top} \big(\boldsymbol{q}_t^{\mathrm{c}} \oslash (\mathbf{1} + |\boldsymbol{a}_t - \boldsymbol{q}_t^{\mathrm{c}}|) \big) \Big), \tag{15}$$

before standardization. The first term is a state-independent reading of the content view; the other two read the same text activations after state routing and are bounded coordinate-wise by $|\boldsymbol{q}_t^{\mathrm{c}}|$ (Proposition 1). The text therefore supplies the evidence, the state modulates how strongly and with which sign it enters s_t, and $s_t = 0$ when no document is available.

Axes and protocol. All points are held-out test windows. Test targets enter only the vertical axis, the realized anchor residual $r_t = y_t - b_t$ averaged over the horizon and standardized, and never the score. Annotated snippets are drawn from the two score tails and the central decile before forecast errors are inspected, which limits outcome-based selection. Every annotated document is dated no later than its forecast origin (App. C). Finally, the deployed correction is further scaled by $\boldsymbol{\lambda}$ (Eq. (6)) and includes the remaining $k - 1$ directions, so the figure characterizes the direction and relative strength of the correction, not its absolute size.

H Implementation, Baselines, and Efficiency

H.1 Configuration

Table 9: Shared configuration of all main experiments.

Item	Value	Item	Value
Split	70/10/20 chronological	Batch size	32
Optimizer (backbones)	AdamW	Weight decay	10^{-4}
Learning rate	10^{-3}	Maximum epochs	45
Early stopping	patience 6 on $\mathcal{D}_{\mathrm{cal}}$	Encoder layers	2
Backbone width	$d_{\mathrm{model}} = 64$	Attention heads	4
Feed-forward width	128	Numerical dropout	0.15
Text encoder	BGE-small-en-v1.5 (frozen)	Maximum tokens	256
Text embedding width	$d_z = 384$	Maximum documents	24
State width	$d_s = 16$	SCA width	$r_{\mathrm{ca}} = 8$
Residual rank	$k = 4$	Pair weight	$\alpha_{\mathrm{pair}} = 1$
Ridge coefficient	$\eta = 0.01$	Gradient clipping	1.0

Only the numerical backbones are optimized with AdamW; the sentence encoder is frozen, and the residual map is obtained in closed form (App. A.5), so the routed feature has $d_h = d_z + 2r_{\mathrm{ca}} = 400$ dimensions. Main results use 20 numerical-backbone seeds. Runs used Python 3.13.5, PyTorch 2.12.0, and CUDA 13.0 on a single NVIDIA GeForce RTX 5090 GPU (32 GiB).

H.2 Baselines

Text-assisted baselines (Table 1). MM-TSFLib and TaTS are trained as plug-ins to the same five backbones, with the numerical inputs, retrieved documents, chronological splits, and seeds used for SPRA.

Multimodal and reasoning-based methods (Table 2). Aurora, VoT, CALF, Time-VLM, Chat-Time, and Time-LLM are rerun under the same experimental setting as SPRA: the chronological splits, numerical inputs, retrieved documents, and document-availability rule of §4.1 and App. C. Each method is run with its standard implementation as described in its original paper, without reducing its model components.

VoT in the attribution diagnostic (Table 3). The diagnostic feeds intervened documents to a trained model. Because every baseline is trained under our protocol, the interventions of App. D.1 are applied directly to the trained VoT model, which §4.3 refers to as our reimplementation.

H.3 Efficiency Measurement

Training time is the complete optimization stage in GPU hours, not one epoch; peak memory is the maximum GPU allocation. Both are recorded for every method during its training run of App. H.2, on the hardware of App. H.1. Frozen embedding caches are built once and excluded from training time for all methods. Every reported time and memory value is a measurement taken during that run; none is carried over from a published paper. Figure 6 reports the Traffic comparison and App. H.4 the remaining domains.

H.4 Efficiency Across Domains

Figure 7 extends the Traffic comparison of Figure 6 to all nine domains with the same encoding: horizontal position is total training time, vertical position is test MSE, and bubble area is monotone in peak GPU memory. The conventions of App. H.3 apply throughout. SPRA-PatchTST stays in the low-cost region in every domain, and the second numerical backbone occupies a nearby region, so the resource profile is not tied to one backbone architecture. Accuracy differences remain domain-dependent: VoT attains a lower MSE on Health-US (Table 2) at a markedly higher training cost, while in the other domains SPRA combines competitive or lower MSE with shorter training times and smaller memory footprints, most visibly against Time-VLM and VoT.

I Limitations and Scope

We treat calendar phase as a predictor that belongs in the numerical backbone as an explicit covariate; combining such covariates with SPRA is complementary and left to future work, together with implicit temporal-cue diagnostics, multilingual event normalization, and online calibration under distribution shift.

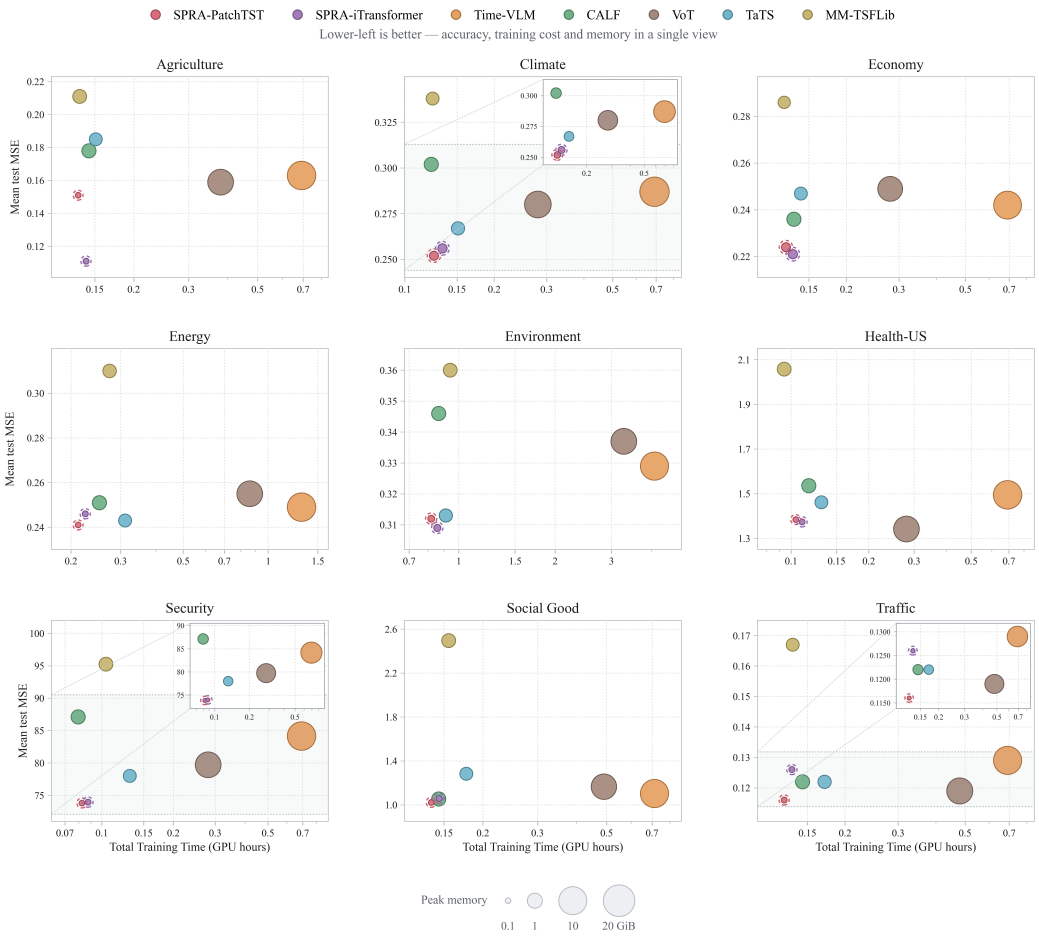


Figure 7: Accuracy, training time, and peak GPU memory across the nine domains; lower left is better. Dashed rings mark the SPRA variants. All time and memory values are measured on the hardware of App. H.1 during the runs of App. H.2 (App. H.3); the plotted values are released with the figure script.