Predictive Language Grounding for Event-Driven Time Series Forecasting via Dual-Path Event Injection

Anonymous ACL submission

Abstract

Discrete events such as holidays, policy announcements, and public gatherings leave no trace in historical time series yet fundamentally shape future demand. Language models capture precisely this event knowledge, making them a natural source of auxiliary signal for event-driven forecasting. However, existing multimodal forecasters borrow *semantic alignment* from vision-language pretraining, enforcing geometric proximity between text and series embeddings. For time series regression, this objective is theoretically unnecessary and actively harmful when training data is limited or text-series coupling is weak: contrastive gradients compress the high-variance series features required to represent event-induced spikes. We propose *predictive grounding* as the principled alternative, shaping language representations through prediction loss alone. **PROSE** (**PR**edictive gr**O**unding for event-driven time-**S**eries for**E**casting) realizes this framework through a fully fine-tuned language encoder, token-level Event Token Injection, LLM-generated Trajectory Reasoning Anchors as global prefix tokens, and a Soft Functional Alignment loss. On a 464K-sample Chinese holiday benchmark, PROSE surpasses the strongest baseline by 10.7% WMAPE; on Time-MMD, it achieves state-of-the-art results across 7 of 9 domains. Our code is anonymously available at `https://anonymous.4open.science/r/prose2026`.

1 Introduction

Forecasting real-world demand signals such as search queries, sales volumes, and traffic flows requires reasoning beyond numerical history (Nie et al., 2022; Liu et al., 2024c; Wu et al., 2022, 2021). Discrete events such as holidays, policy announcements, and public gatherings introduce sharp non-stationary shifts that are absent from historical series yet fully encoded in event text (Ge et al., 2025; Cui et al., 2025; Zhang et al., 2024).

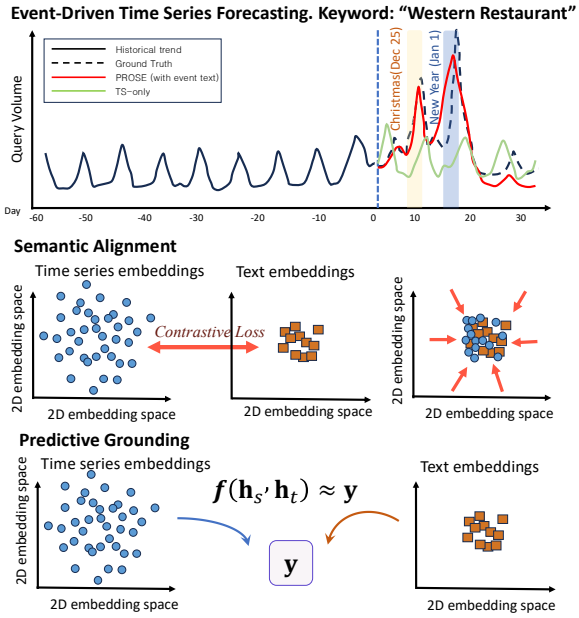

Figure 1: Motivation for PROSE. **Top**: holiday spikes invisible in history are predictable from event text. **Middle**: semantic alignment compresses series features toward the text manifold. **Bottom**: predictive grounding learns $f(\mathbf{h}_s, \mathbf{h}_t) \approx \mathbf{y}$ without geometric constraints.

As illustrated in Figure 1 (top), a keyword like "Western Restaurant" shows no unusual pattern in its 60-day history, while the Christmas and New Year surge is entirely predictable from the event description. Language models naturally capture this structured event knowledge (Brown et al., 2020; Touvron et al., 2023), making them a compelling auxiliary signal for event-aware forecasting (Wang et al., 2026; Li et al., 2026).

The problem with semantic alignment. Motivated by the success of contrastive vision-language models (Radford et al., 2021; Li et al., 2023), recent forecasters import *semantic alignment*: contrastive losses enforce geometric proximity between text and series embeddings (Liu et al., 2025b,a). As shown in Figure 1 (middle), this approach assumes that text and series representations should occupy

the same geometric region—an assumption suited to retrieval tasks with large aligned corpora, but mismatched to time series regression. When event-series co-occurrence is sparse or coupling is weak, the InfoNCE gradient pulls series features toward the lower-variance event manifold, compressing the high-variance dynamics needed to capture event-induced spikes (Proposition 2). This inductive bias is particularly harmful in industrial settings, where event annotations are typically coarse relative to the series temporal resolution.

Predictive grounding. We propose *predictive grounding* as the principled alternative, shaping language representations through prediction-driven training so that they encode *forecasting-functional* rather than cultural-semantic knowledge. As illustrated in Figure 1 (bottom), the objective is not geometric proximity $\|\phi_t(e) - \phi_s(\mathbf{x})\|$ between modalities but a functional mapping $f(\mathbf{h}_s, \mathbf{h}_t) \approx \mathbf{y}$; the two representations need not share the same semantic space, so long as they jointly enable accurate prediction. We formally establish that semantic alignment is theoretically unnecessary for this mapping (Proposition 1), and that prediction loss is a sufficient training objective (Proposition 3).

PROSE (**PR**edictive gr**O**unding for event-driven time-**S**eries for**E**casting) realizes predictive grounding through three complementary design choices. Unlike frozen encoders (Zhou et al., 2023; Zhong et al., 2025), a fully fine-tuned language encoder allows prediction gradients to migrate pre-trained representations toward forecasting-functional ones. Token-level Event Token Injection (ETI) injects position-sensitive event representations via low-rank gated cross-attention, preserving the temporal cues that CLS pooling discards (Devlin et al., 2019; Liu et al., 2019). LLM-generated Trajectory Reasoning Anchors (TRA) distill explicit day-by-day impact reasoning into global prefix tokens, providing the backbone with structured temporal prior beyond what average-pooled LLM outputs can supply (Jin et al., 2024; Lin et al., 2026). A Soft Functional Alignment loss (SFA) further enforces that samples with similar forecast trajectories share event representations, implementing predictive grounding at the batch level without geometric constraints.

Our contributions are summarized as follows:

- **Predictive Grounding Theory.** We formally prove that semantic alignment is theoretically unnecessary (Prop. 1) and harmful in low-data, weakly coupled settings (Prop. 2) for time series regression, and establish prediction loss as a theoretically sufficient training objective (Prop. 3).
- **PROSE Architecture.** We present PROSE, a dual-path predictive grounding framework for event-driven time series forecasting. PROSE combines full LM fine-tuning, token-level Event Token Injection, LLM-generated Trajectory Reasoning Anchors, and a Soft Functional Alignment loss to shape language representations through prediction-driven training and realize predictive grounding end-to-end.
- **Empirical Validation.** PROSE achieves state-of-the-art performance across 7 of 9 Time-MMD domains and surpasses the strongest baseline by 10.7% WMAPE on Query-QV, a 464K-sample event-annotated Chinese holiday benchmark, demonstrating consistent gains across diverse event types and domains.

2 Related Work

Time series forecasting. Deep time series forecasting has evolved through successive architectural generations. Transformer-based models became dominant following the foundational sequence-to-sequence attention mechanism (Vaswani et al., 2017); subsequent work tackled computational efficiency (Zhou et al., 2021), temporal decomposition with auto-correlation (Wu et al., 2021), frequency-domain decomposition (Zhou et al., 2022), and patch-level tokenization (Nie et al., 2022). More recent advances introduced inverted attention over variable dimensions (Liu et al., 2024c) and learnable global tokens for heterogeneous series (Wang et al., 2024). Complementary architectures include linear decomposition models (Zeng et al., 2023) and temporal 2D-variation convolution (Wu et al., 2022). Pre-trained foundation models (Goswami et al., 2024; Garza et al., 2023; Liu et al., 2024e; Das et al., 2023; Ansari et al., 2024) have further broadened applicability to zero-shot and cross-domain settings; MOMENT (Goswami et al., 2024) in particular provides strong patch encodings that we adopt as our series backbone. Despite these advances, all such methods operate exclusively on numerical history and are blind to the event-driven non-stationarities—holiday surges, policy-induced shocks, crowd-event anomalies—that are fully pre-

dictable from external text yet invisible in past observations (Ge et al., 2025; Cui et al., 2025).

Language models for time series forecasting. The integration of language models into time series forecasting has taken two main threads.

LLMs as forecasting backbone. One line of work adapts pre-trained LM weights as the core sequence model. GPT4TS (Zhou et al., 2023) and GPT4MTS (Jia et al., 2024) demonstrate that frozen GPT-2 generalizes to temporal patterns without any external text. Time-LLM (Jin et al., 2024) reprograms time series into text prototypes while keeping the LLM backbone intact. AutoTimes (Liu et al., 2024d) introduces autoregressive tokenization in the LLM embedding space, and UniTime (Liu et al., 2024b) directly concatenates temporal and textual features for cross-domain adaptation. These methods leverage frozen LLM representations as universal sequence models but do not exploit event semantics encoded in natural language.

Language as auxiliary event signal. A second thread incorporates exogenous text to inform predictions, motivated by evidence that LLMs encode structured event knowledge absent from numerical history (Zhang et al., 2024; Brown et al., 2020). CALF (Liu et al., 2025b) and TimeCMA (Liu et al., 2025a) impose contrastive alignment losses between text and series embeddings, importing the semantic alignment paradigm from vision-language pre-training (Radford et al., 2021; Li et al., 2023). VoT (Wang et al., 2026) extends this with multi-level alignment including frequency-domain fusion alongside LLM-generated event reasoning. TiMi (Lin et al., 2026) routes text and series through a Mixture-of-Experts backbone, generating textual descriptions but pooling them into a single average vector and freezing the encoder. TESS (Li et al., 2026) maps LLM output to discrete temporal primitives, discarding continuous representational richness. Concurrent work (Lee et al., 2026) demonstrates that naive cross-modal fusion frequently degrades TS performance relative to text-free baselines.

PROSE departs from all of these by identifying semantic alignment as the root cause of failure rather than a solution. Under *predictive grounding*, the language encoder is trained end-to-end via prediction gradients without contrastive objectives or frozen weights, enabling token-level event injection and LLM trajectory reasoning anchors to cooperate within a single fine-tuned backbone.

3 Predictive Grounding: Definitions and Theory

3.1 Definitions

Definition 1 (Event-Driven Time Series Forecasting). *Given historical series* $\mathbf{x} \in \mathbb{R}^T$*, event text* $e \in \Sigma^*$*, and forecast target* $\mathbf{y} \in \mathbb{R}^H$*, the task is* ***event-driven*** *if* e *encodes external interventions that induce non-stationary regime shifts in* $\mathbf{y}$ *invisible in* $\mathbf{x}$ *alone. Any model restricted to* $\mathbf{x}$ *incurs an irreducible prediction-error floor proportional to the variance of those interventions (Xu et al., 2024); event text therefore carries forecasting-essential information that numerical history structurally cannot provide.*

Definition 2 (Semantic Alignment and Predictive Grounding). *Let* ϕ_s *and* ϕ_t *denote the series and text encoders. A model is* ***semantically aligned*** *if it minimizes* $\|\phi_t(e) - \phi_s(\mathbf{x})\|$ *as a training objective. A model achieves* ***predictive grounding*** *(equivalently,* functional alignment*) if prediction gradients drive* ϕ_t *and* ϕ_s *to collectively minimize* $\mathcal{L}_{\text{pred}}$*, with no constraint on embedding geometry. Critically, semantic alignment implies predictive grounding, but not vice versa: predictive grounding is the weaker, task-sufficient condition.*

Unlike retrieval tasks—where minimizing cross-modal embedding distance is itself the end goal—time series forecasting requires only accurate numerical outputs. The output space is entirely numerical, rendering embedding proximity between modalities irrelevant; prediction loss is therefore the natural alignment signal for regression (Hinton et al., 2015).

3.2 Predictive Grounding Theory

Propositions 1–3 establish that, for event-driven time series forecasting, semantic alignment is (i) not necessary, (ii) actively harmful when text-series coupling is weak or training data is limited, and (iii) unnecessary because prediction loss alone is theoretically sufficient. We state each result formally below; proof sketches are given in text and full proofs appear in Appendix B. Together, they form a complete theoretical foundation for predictive grounding.

Proposition 1 (Semantic alignment is not a necessary condition). *For any model (f, ϕ_s, ϕ_t) achieving prediction error δ, there exists a reparametrized model (f', ϕ_s, ϕ_t') achieving the same error δ while violating semantic alignment by an arbitrarily large margin.*

Proof sketch. Apply any invertible R to text embeddings and absorb R^{-1} into the predictor; predictions are preserved while $\|\phi_t'(e) - \phi_s(\mathbf{x})\|$ grows without bound. Full proof in Appendix B. □

Proposition 2 (Contrastive alignment is harmful under weak text-series coupling). *When event-label variance is lower than series variance—a condition that holds when text-series coupling is weak, training data is limited, or event annotations are coarse relative to the series' temporal resolution—the InfoNCE alignment gradient induces a systematic contraction of ϕ_s toward the lower-variance event manifold, suppressing the high-variance series features that encode event-induced spikes. With ample, well-calibrated aligned data the effect is mitigated; in the sparse-annotation regime that characterizes real-world event-driven forecasting, it is not.*

Proof sketch. The InfoNCE alignment term (Wang and Isola, 2020) pulls ϕ_s toward the event manifold at every gradient step. The uniformity term partially counteracts this pull, but when $\mathrm{Var}(\phi_t) \ll \mathrm{Var}(\phi_s)$, the alignment term dominates early in training before uniformity can restore the spread, compressing the spectral energy in ϕ_s and erasing fine-grained temporal dynamics. Full derivation in Appendix B. □

Proposition 3 (Prediction loss is a theoretically sufficient objective). *The hypothesis class optimized by $\mathcal{L}_{\text{pred}}$ contains ε-optimal predictively grounded solutions for any $\varepsilon > 0$, requiring no semantic proximity constraint. (Practical reachability is an open question for general architectures; the ablation in §5.4 provides empirical evidence that gradient-based training reaches such solutions in PROSE.)*

Proof sketch. By the Universal Approximation Theorem (Hornik, 1991), for any continuous target function there exist encoders ϕ_s, ϕ_t and predictor f that achieve arbitrarily small prediction error without any embedding-proximity constraint; the hypothesis class therefore *contains* such solutions. Full proof in Appendix B. □

These propositions impose three concrete requirements on any architecture pursuing predictive grounding: (i) prediction gradients must propagate through the language encoder so that ϕ_t can migrate from cultural-semantic to forecasting-functional representations; (ii) token-level positional structure must be preserved rather than collapsed via CLS pooling; and (iii) cross-modal fusion must be constrained to prevent irrelevant semantic dimensions from distorting series features. Section 4 translates each requirement into a distinct component of PROSE.

4 PROSE: Architecture

PROSE (Figure 2) operationalizes the three requirements from §3. Historical series $\mathbf{x} \in \mathbb{R}^T$ is encoded by a *frozen* MOMENT patch encoder (Goswami et al., 2024) with a trainable adapter into patch tokens $\mathbf{P} \in \mathbb{R}^{N \times d}$. A *fully fine-tuned* language encoder processes the concatenated text [<ev>e</ev> [SEP] p_B] in a single pass, routing hidden states to ETI and TRA. A GPT-2 backbone (Radford et al., 2019) then processes the assembled sequence under causal self-attention, and a projection head maps mean-pooled patch features directly to $\hat{\mathbf{y}} \in \mathbb{R}^H$.

4.1 ETI: Event Token Injection

Collapsing event text to a single CLS vector discards token-level positional cues that distinguish, e.g., a pre-event build-up from a post-event decay. ETI preserves this structure by injecting the full event-span hidden states into the backbone (requirement ii).

Let $\mathbf{H}_e \in \mathbb{R}^{L_e \times d_L}$ be the hidden states of the event-text span, located by character-level offset mapping. A **Low-Rank Gated Projection** (LRGP) filters keyword-irrelevant dimensions:

$$\tilde{\mathbf{H}}_e = \mathrm{LN}\Big(\mathbf{W}_\uparrow\big(\mathrm{SiLU}(\mathbf{W}_\downarrow \mathbf{H}_e) \odot \sigma(\mathbf{W}_g \mathbf{H}_e)\big)\Big), \tag{1}$$

where $\mathbf{W}_\downarrow \in \mathbb{R}^{r \times d_L}$, $\mathbf{W}_\uparrow \in \mathbb{R}^{d_L \times r}$ (r=32), and $\mathbf{W}_g \in \mathbb{R}^{r \times d_L}$ is a content-adaptive gate; $\mathbf{W}_\uparrow$ is zero-initialized so ETI starts neutral. The output $\tilde{\mathbf{H}}_e$ serves as keys and values in cross-attention layers inserted into GPT-2 layers $\ell = 0$–5:

$$\mathbf{h}_\ell' = \mathbf{h}_\ell + g_\ell \odot \mathrm{Attn}\Big(\mathbf{h}_\ell \mathbf{W}_\ell^Q,\ \tilde{\mathbf{H}}_e \mathbf{W}_\ell^K,\ \tilde{\mathbf{H}}_e \mathbf{W}_\ell^V\Big), \tag{2}$$

where $g_\ell \in \mathbb{R}^d$ is a per-token scalar gate (zero-initialized). The early layers (0–5) absorb event

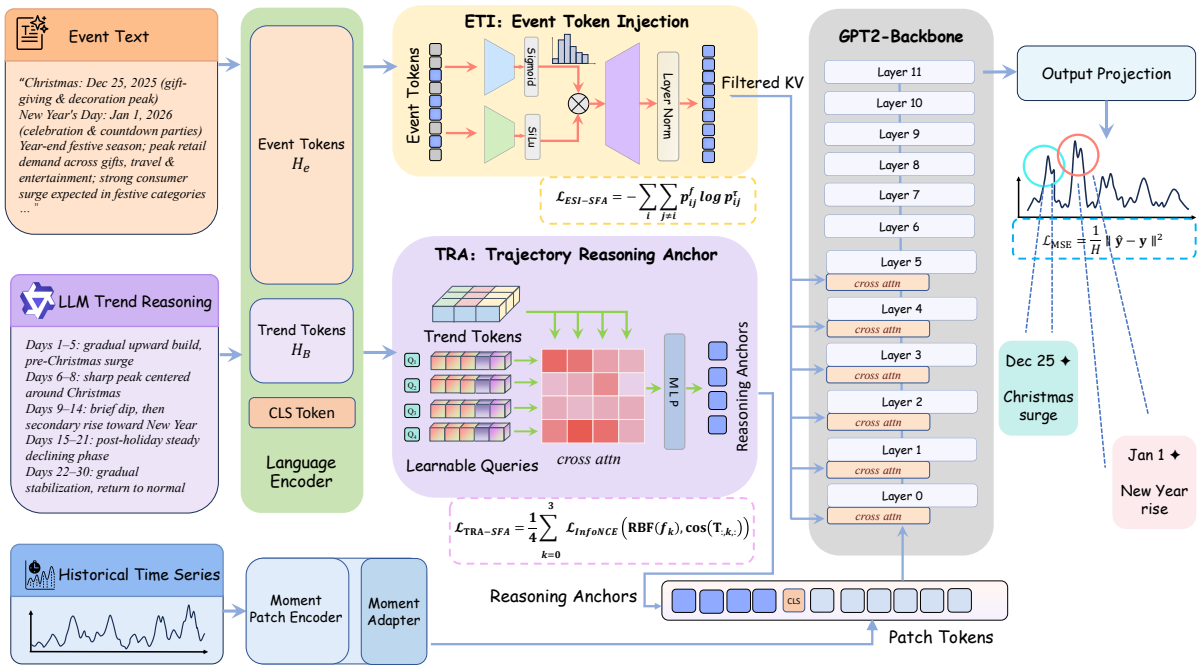

Figure 2: Overview of PROSE. A fully fine-tuned language encoder processes unified text [<ev>e</ev> [SEP] p_B], routing hidden states to ETI (token-level event injection, layers 0–5) and TRA (trajectory prefix anchors). A frozen MOMENT encoder supplies patch tokens. All three token groups are concatenated as input to the GPT-2 backbone.

context, while the deeper layers (6–11) specialize in temporal extrapolation. Full derivations are in Appendix A.

4.2 TRA: Trajectory Reasoning Anchor

Methods that use LLM generation pool its output into a single vector and freeze the encoder (Lin et al., 2026; Wang et al., 2026), preventing grounding of LLM reasoning to the task. TRA addresses failure (iii) by distilling explicit trajectory reasoning into learnable prefix anchors.

As shown in Figure 3, an LLM is prompted offline with event description e and a five-dimensional natural-language summary of $\mathbf{x}$ (trend, periodicity, recent level, volatility, anomaly) to generate p_B—e.g., *"Steps 1–4: gradual build; Steps 5–8: sharp peak; Steps 9–H: decay to baseline."* A **TrendQFormer** then distills the encoder's hidden states $\mathbf{H}_B \in \mathbb{R}^{L_B \times d_L}$ of p_B via four learnable queries $\mathbf{q}_1, \ldots, \mathbf{q}_4 \in \mathbb{R}^d$:

$$\mathbf{t}_i = \mathrm{softmax}\left(\frac{\mathbf{q}_i \mathbf{H}_B^\top}{\sqrt{d}}\right)\mathbf{H}_B, \quad i = 1, \ldots, 4, \quad (3)$$

each specialized to a temporal perspective (short-, mid-, long-term, global). The trend matrix $\mathbf{T}=[\mathbf{t}_1; \ldots; \mathbf{t}_4] \in \mathbb{R}^{4 \times d}$ is prepended as a prefix (Li and Liang, 2021), making the trajectory prior globally visible under causal attention.

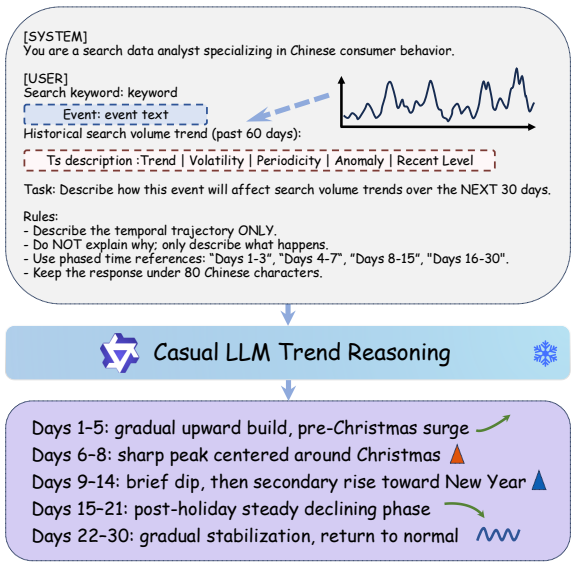

Figure 3: Offline LLM reasoning pipeline (TRA). An LLM generates a trajectory description from event text and a series summary; TrendQFormer distills it into four prefix tokens.

4.3 Backbone, Training, and SFA

The GPT-2 backbone receives $\mathbf{X} = [\mathbf{T} \parallel \mathbf{c} \parallel \mathbf{P}]$, where $\mathbf{c} \in \mathbb{R}^{1 \times d}$ is a global event token from mean-pooling $\tilde{\mathbf{H}}_e$. The prediction head mean-pools the patch output, applies a residual linear projection, and denormalizes via RevIN (Kim et al., 2021) to produce $\hat{\mathbf{y}} \in \mathbb{R}^H$. The primary regression objective

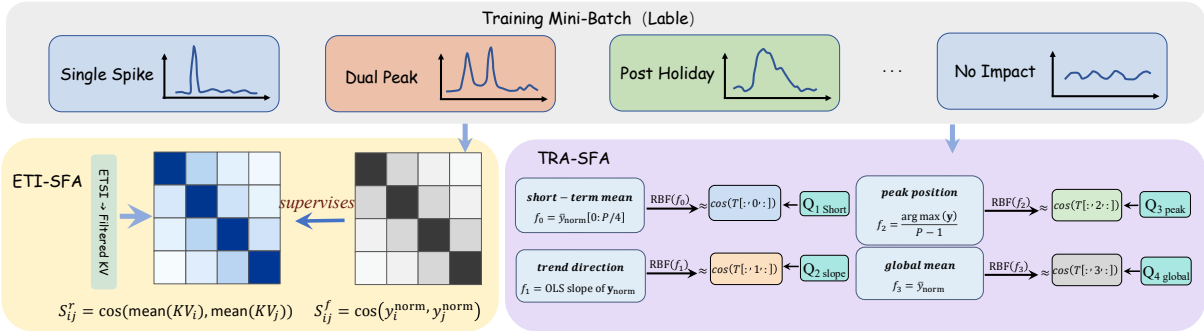

Figure 4: Soft Functional Alignment (SFA). **Left (ETI-SFA)**: trajectory-based soft labels supervise the similarity of ETI output representations. **Right (TRA-SFA)**: four scalar trajectory features ground the TrendQFormer queries via RBF kernels and InfoNCE. Gradients flow only to the language encoder.

$\mathcal{L}_{\text{MSE}}$ supervises all model parameters end-to-end. Two auxiliary SFA terms provide grounding signals to specific sub-modules: $\mathcal{L}_{\text{ETI-SFA}}$ constrains the language encoder and the ETI path; $\mathcal{L}_{\text{TRA-SFA}}$ constrains the language encoder and the TRA path. The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_1 \mathcal{L}_{\text{ETI-SFA}} + \lambda_2 \mathcal{L}_{\text{TRA-SFA}}, \quad (4)$$

with no contrastive pretraining between text and series.

SFA (Figure 4) enforces that samples with *functionally* similar trajectories share similar event representations, without any geometric proximity constraint. Let $\mathbf{y}_i^{\text{norm}} = \mathbf{y}_i/(\bar{\mathbf{x}}_{i,\text{rec}}+\varepsilon)$, where $\bar{\mathbf{x}}_{i,\text{rec}}$ is the mean of a short trailing window of the input history, whose length is tuned per dataset. **ETI-SFA** uses the ETI output $\tilde{\mathbf{H}}_e$ as the event representation: $s^r_{ij} = \cos(\text{mean}(\tilde{\mathbf{H}}_{e,i}), \text{mean}(\tilde{\mathbf{H}}_{e,j}))$. Defining $s^f_{ij} = \cos(\mathbf{y}_i^{\text{norm}}, \mathbf{y}_j^{\text{norm}})$, the objective is:

$$\mathcal{L}_{\text{ETI-SFA}} = -\sum_i \sum_{j\neq i} p^f_{ij} \log p^r_{ij},$$

$$p^{f/r}_{ij} = \frac{\exp(s^{f/r}_{ij}/\tau)}{\sum_{k\neq i} \exp(s^{f/r}_{ik}/\tau)}, \quad (5)$$

with τ=0.07. **TRA-SFA** matches each TrendQFormer query $\mathbf{T}_{:,k,:}$ to a scalar trajectory feature f_k (f_0: short-term mean; f_1: OLS slope; f_2: peak position; f_3: global mean) via RBF similarity:

$$\mathcal{L}_{\text{TRA-SFA}} = \frac{1}{4}\sum_{k=0}^{3} \mathcal{L}_{\text{InfoNCE}}\big(\text{RBF}(f_k),\ \cos(\mathbf{T}_{:,k,:})\big). \quad (6)$$

Gradients from each SFA term are stopped at the GPT-2 backbone boundary. Full derivations are in Appendix C.

5 Experiments

5.1 Setup

Datasets. We evaluate on two datasets differing in scale and event density. **Time-MMD** (Liu et al., 2024a) covers 9 heterogeneous domains (Agriculture, Climate, Economy, Energy, Environment, Health, Security, SocialGood, and Traffic) and provides a publicly reproducible multi-domain evaluation benchmark. **Query-QV** is a proprietary Chinese search-query volume dataset of 464,996 samples (T=60, H=30) whose test period (Dec. 2025–Feb. 2026) with 84.8% of samples containing at least one holiday annotation. Dataset details are in Appendix F.

Metrics. For Time-MMD, monthly-sampled datasets (Agriculture, Climate, Economy, Security, SocialGood, Traffic) use input length L=8 and prediction horizons $H \in \{6, 8, 10, 12\}$; weekly-sampled datasets (Energy, Health) use L=24, $H \in \{12, 24, 36, 48\}$; and the daily-sampled Environment dataset uses L=96, $H \in \{48, 96, 192, 336\}$. We report MSE and MAE averaged over all horizons. For Query-QV, the primary metric is per-sample percentile WMAPE$_{p50}$, which is robust to the heavy tail from extreme holiday spikes. We additionally report WMAPE$_{p90}$ and RMSE for tail-risk assessment, and *holiday-subset* variants (Hol-WMAPE$_{p50}$, Hol-WMAPE$_{p90}$, Hol-RMSE) that score only at time steps corresponding to annotated holiday days within the forecast horizon.

Baselines. We compare PROSE against ten baselines from two families: six *time-series-only* models (DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2022), iTransformer (Liu et al., 2024c), TimesNet (Wu et al., 2022), TimeXer (Wang et al., 2024), Autoformer (Wu et al., 2021)) and four *language-augmented* models (GPT4TS (Zhou

Table 1: Results on Time-MMD (9 domains, avg. over horizons, MSE/MAE). **Red bold**: best. Blue underline: 2nd best.

Dataset	PROSE		TS-only												Multimodal							
	MSE	MAE	iTrans.		PatchTST		DLinear		TimesNet		TimeXer		Autoformer		GPT4TS		GPT4MTS		TaTS		TimeVLM	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	**0.208**	**0.301**	0.220	0.304	0.222	0.304	0.407	0.453	0.229	0.307	0.225	0.302	0.242	0.323	0.231	0.311	0.281	0.333	0.302	0.365	0.240	0.307
Climate	**1.202**	**0.896**	1.278	0.922	1.239	0.906	1.311	0.930	1.369	0.948	1.215	0.901	1.262	0.914	1.237	0.918	1.257	0.918	1.218	0.913	1.222	0.908
Economy	**0.206**	**0.357**	0.222	0.371	0.213	0.364	0.323	0.457	0.251	0.393	0.218	0.368	0.248	0.398	0.218	0.372	0.213	0.367	1.558	1.132	0.240	0.395
Energy	0.278	0.390	0.280	0.392	0.289	0.402	0.343	0.433	0.301	0.408	0.294	0.413	0.370	0.479	0.280	0.398	0.283	0.397	0.453	0.494	**0.267**	**0.379**
Environment	0.329	0.405	0.324	0.402	0.335	0.411	0.544	0.628	**0.319**	**0.397**	0.330	0.411	0.354	0.449	0.331	0.414	0.332	0.411	0.434	0.460	0.321	0.413
Health	1.674	**0.785**	1.656	0.797	1.532	0.824	1.704	0.882	1.779	0.839	1.572	0.818	2.026	1.029	1.554	0.800	1.629	0.879	1.529	0.866	**1.479**	0.838
Security	**70.667**	**3.994**	77.983	4.381	76.444	4.383	73.266	4.309	84.456	4.751	78.049	4.479	79.675	4.335	71.269	4.257	75.487	4.334	72.756	4.116	75.832	4.352
SocialGood	**0.768**	**0.432**	1.220	0.490	1.082	0.481	1.016	0.555	1.591	0.517	1.126	0.503	1.056	0.527	0.915	0.460	0.980	0.478	1.026	1.453	0.865	0.453
Traffic	**0.122**	**0.203**	0.143	0.225	0.177	0.229	0.282	0.416	0.171	0.215	0.150	0.215	0.155	0.275	0.154	0.237	0.151	0.237	0.979	0.745	0.143	0.258
1st counts	**7**		0		0		0		1		0		0		0		0		0		1	

Table 2: Results on Query-QV (T=60, H=30). **Red bold**: best. Blue underline: 2nd best.

Category	Model	Overall			Holiday subset		
		WMAPE$_{p50}$↓	WMAPE$_{p90}$↓	RMSE↓	Hol-WMAPE$_{p50}$↓	Hol-WMAPE$_{p90}$↓	Hol-RMSE↓
TS-only	DLinear	0.4251	0.9215	1212	0.4652	0.9402	721
	PatchTST	0.4191	0.8934	1127	0.4621	0.9236	706
	iTransformer	0.4126	0.8716	1210	0.4468	0.8935	706
	TimesNet	0.4034	0.8568	1169	0.4409	0.8798	711
	TimeXer	0.4078	0.8453	1216	0.4435	0.8683	743
	Autoformer	0.4045	0.8885	1174	0.4433	0.9154	705
Multimodal	GPT4TS	0.4004	0.8459	1140	0.4354	0.8679	697
	GPT4MTS	0.4128	0.8809	1134	0.4531	0.9127	700
	TaTS	0.4189	0.8906	1171	0.4605	0.9172	705
	TimeVLM	0.4030	0.8526	1169	0.4434	0.8779	696
Ours	**PROSE**	**0.3575**	**0.8276**	**1056**	**0.3880**	**0.8592**	**653**
Improve over 2nd-best		*+10.7%*	*+2.1%*	*+6.3%*	*+10.9%*	*+1.0%*	*+6.2%*

et al., 2023), GPT4MTS (Jia et al., 2024), TaTS (Xue and Salim, 2023), TimeVLM (Zhong et al., 2025)). All baselines are faithfully reimplemented and adapted to each dataset following standard protocols.

5.2 Main Results on Time-MMD

Table 1 shows results averaged over prediction horizons. PROSE achieves the best performance in 7 of 9 domains. Gains are largest where event descriptions couple strongly to the target variable: SocialGood (+11.1% MSE, policy announcements), Traffic (+14.7%, holiday congestion), and Security due to crowd-event anomalies. In Environment and Energy, dominated by smooth seasonal cycles, event descriptions are too coarse-grained to add value over numerical baselines. This is consistent with predictive grounding theory: weak event-to-series functional mappings leave little for prediction gradients to reshape in the language encoder.

5.3 Main Results on Query-QV

Table 2 reports results on all six metrics. PROSE achieves the lowest error in every column: WMAPE$_{p50}$ of 0.358 vs. 0.400 for the best baseline GPT4TS (**+10.7%**). The holiday-subset gap widens further (**+10.9%** Hol-WMAPE$_{p50}$), confirming that PROSE's advantage concentrates where event signals matter most.

Three patterns validate the theoretical claims of §3. **(1) Semantic alignment suppresses event dynamics**: TaTS (0.419), conditioned on statistical summaries via a contrastive-style objective, performs *worse* than the best text-free model TimesNet (0.403)—directly confirming the variance-compression effect of Prop. 2. **(2) Frozen representations do not ground to forecasting**: GPT4TS (0.400, frozen GPT-2) and TimeVLM (0.403, frozen ViT+BERT) yield negligible gain over the TS-only ceiling regardless of model capacity—consistent with Prop. 1: cultural-semantic representations carry no forecasting-functional knowledge. **(3) Prediction-gradient grounding drives functional representation migration**: PROSE's 10.7% gain over GPT4TS demonstrates that end-to-end prediction training migrates the language encoder from the cultural-semantic to the forecasting-functional subspace, realizing predictive grounding as established in Prop. 3 and further validated in §5.4.

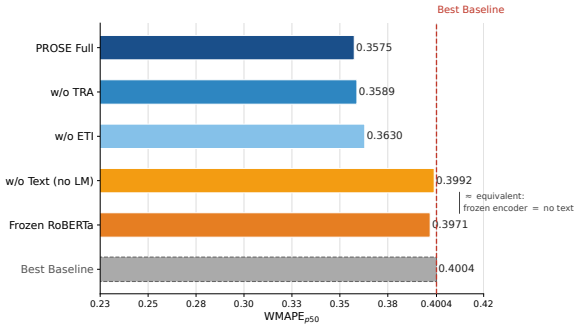

Figure 5: Ablation on WMAPE$_{p50}$ (lower is better). Frozen RoBERTa ≈ No text, directly validating predictive grounding theory. All PROSE variants outperform Best Baseline.

5.4 Ablation Study

Component contributions. Table 3 ablates PROSE on Query-QV. Removing TRA degrades WMAPE$_{p50}$ by 0.4% and Hol-WMAPE by 2.4 percentage points; removing ETI degrades by 1.5% and 3.3 percentage points, respectively. ETI contributes primarily to point accuracy via token-level positional event cues; TRA contributes primarily to holiday-subset consistency (Hol-WMAPE: 0.3884 without TRA vs. 0.3934 without ETI). Contributions are additive: PROSE minus one path (0.359 or 0.363) still outperforms GPT4TS (0.400). TaTS (0.419, statistical summaries without a grounded LM) scores *worse* than TS-only TimesNet (0.403), confirming that event information is beneficial only when the encoder is shaped by prediction loss.

Table 3: Ablation study on Query-QV. **Red bold**: best per column.

Variant	WMAPE$_{p50}$↓	WMAPE$_{p90}$↓	RMSE↓
PROSE	**0.3575**	**0.8276**	**1056**
w/o TRA	0.3589	0.8592	1107
w/o ETI	0.3630	0.8380	1087
w/o Text	0.3992	0.9111	1164
Best baseline	0.4004	0.8459	1140

Validating predictive grounding. Figure 5 visualizes the ablation results. A frozen encoder degrades WMAPE$_{p50}$ by 11.1%, nearly identical to removing text entirely (11.7%); the 0.6-percentage-point gap is statistically indistinguishable (p=0.18, Appendix E). This *frozen ≈ no-text* result directly instantiates Propositions 1–3: absent prediction-gradient grounding, the pre-trained encoder retains cultural-semantic representations with no forecasting benefit. All PROSE variants outperform GPT4TS (dashed line).

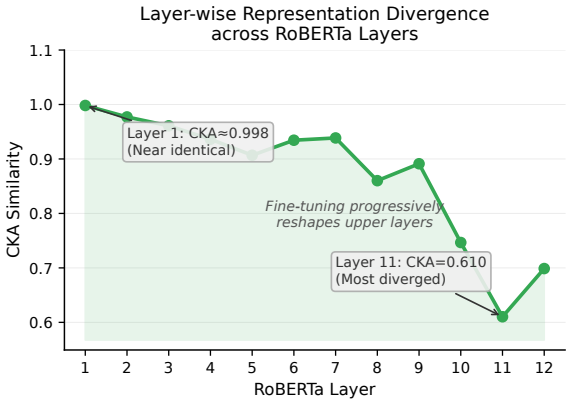

Figure 6: Layer-wise CKA between fine-tuned and frozen RoBERTa. CKA drops in upper layers (Layer 10: CKA=0.61), confirming that prediction-only training reshapes the encoder toward a functional subspace.

Representation analysis. Figure 6 shows that upper-layer CKA drops substantially (Layer 10: CKA=0.61) while lower layers remain stable (CKA>0.97), confirming that the language encoder has migrated from its cultural-semantic pre-training representations into a forecasting-functional subspace. ETI and TRA channel prediction gradients through token-level and trajectory-level event representations; SFA further accelerates this migration by aligning event representations with functional trajectory similarity. Together, they drive the language encoder toward predictive language grounding—the direct validation of our architectural design.

6 Conclusion

Language models naturally encode the discrete event knowledge that numerical history alone cannot capture, making them a compelling auxiliary signal for event-driven forecasting. Existing multimodal forecasters, however, borrow semantic alignment from vision-language pretraining—an objective that is theoretically unnecessary and empirically harmful under weak text-series coupling. We proposed *predictive grounding* as the principled alternative: shaping language representations through prediction loss so that the encoder migrates from cultural-semantic to forecasting-functional representations. **PROSE** realizes this framework through ETI, TRA, and SFA, achieving consistent gains on both public multi-domain and proprietary holiday benchmarks. We expect predictive grounding to generalize to other multimodal regression tasks where cross-modal semantics are task-specific rather than universally shared.

Limitations

Event-driven forecasting is most effective when event text and the target series are meaningfully coupled; in domains where annotations are inherently coarse relative to the series' temporal resolution, any text-based method—including PROSE—will see limited gains over numerical baselines. We acknowledge this as an inherent constraint of the task rather than a method-specific failure, and leave domain-adaptive coupling estimation to future work.

TRA relies on offline LLM-generated trajectory descriptions; we acknowledge that generation quality scales with model capability, and stronger LLMs are expected to yield further performance gains.

SFA computes pairwise trajectory similarities at the batch level, so batch composition influences the soft-label distribution during training; a systematic study of batch construction strategies is left to future work.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, and 1 others. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3504–3514.

Zhiqing Cui, Binwu Wang, Qingxiang Liu, Yeqiang Wang, Zhengyang Zhou, Yuxuan Liang, and Yang Wang. 2025. Augur: Modeling covariate causal associations in time series via large language models. *arXiv preprint arXiv:2510.07858*.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2023. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2023. Timegpt-1. *arXiv preprint arXiv:2310.03589*.

Yunfeng Ge, Ming Jin, Yiji Zhao, Hongyan Li, Bo Du, Chang Xu, and Shirui Pan. 2025. Eventtsf: Event-aware non-stationary time series forecasting. *arXiv preprint arXiv:2508.13434*.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kurt Hornik. 1991. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 23343–23351.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and 1 others. 2024. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pages 23857–23880.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2021. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2022. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*.

Seunghan Lee, Jun Seo, Jaehoon Lee, Sungdong Yoo, Minjae Kim, Tae Yoon Lim, Dongwan Kang, Hwanil Choi, SoonYoung Lee, and Wonbin Ahn. 2026. Rethinking multimodal fusion for time series: Auxiliary modalities need constrained fusion. *arXiv preprint arXiv:2603.22372*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Lehui Li, Yuyao Wang, Jisheng Yan, Wei Zhang, Jinliang Deng, Haoliang Sun, Zhongyi Han, and Yongshun Gong. 2026. From text to forecasts: Bridging modality gap with temporal evolution semantic space. *arXiv preprint arXiv:2603.12664*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Jiafeng Lin, Yuxuan Wang, Huakun Luo, Zhongyi Pei, and Jianmin Wang. 2026. Timi: Empower time series transformers with multimodal mixture of experts. *arXiv preprint arXiv:2602.21693*.

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

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and 1 others. 2024a. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933.

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

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. 2024b. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*, pages 4095–4106.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024d. Autotimes: Autoregressive time series forecasters via large language models. *Advances in Neural Information Processing Systems*, 37:122154–122184.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024e. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Chenjuan Guo, Bin Yang, and Yang Shu. 2026. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. *arXiv preprint arXiv:2603.15452*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. 2024. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*.

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

Zhijian Xu, Hao Wang, and Qiang Xu. 2024. Intervention-aware forecasting: Breaking historical limits from a system perspective. *arXiv preprint arXiv:2405.13522*.

Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864.

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

Hanyu Zhang, Chuck Arvin, Dmitry Efimov, Michael W Mahoney, Dominique Perrault-Joncas, Shankar Ramasubramanian, Andrew Gordon Wilson, and Malcolm Wolff. 2024. Llmforecaster: Improving seasonal event forecasts with unstructured textual data. *arXiv preprint arXiv:2412.02525*.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. 2025. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint arXiv:2502.04395*.

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

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

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, and 1 others. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355.

A ETI/TRA Design Rationale

Why ETI uses low-rank gated projection. Event tokens are short ($L_e \approx$ 3–8) but dense in information. Unconstrained cross-modal fusion introduces noise from event dimensions irrelevant to the specific search keyword (Lee et al., 2026). LRGP down-projects to rank r=32 with a SiLU-gated bottleneck:

$$\mathrm{LRGP}(h) = \mathrm{LN}\Big(\mathbf{W}_{\uparrow}\big(F_{\mathrm{SiLU}}(\mathbf{W}_{\downarrow}h) \odot \sigma(\mathbf{W}_g h)\big)\Big),$$

with $\mathbf{W}_{\uparrow}$ zero-initialized so ETI starts neutral. This low-rank bottleneck constrains fusion to forecasting-relevant dimensions.

Why ETI injects only into layers 0–5. Early GPT-2 layers absorb contextual priors; layers 6–11 specialize in temporal extrapolation dynamics. Injecting event context into all 12 layers over-conditions the model on event text and degrades fine-grained sequence extrapolation. Ablation: injecting all 12 layers increases WMAPE_{p50} by 0.8 percentage points vs. layers 0–5 only (evaluated on Query-QV).

Why TRA uses full-rank projection. The trend reasoning text is longer ($L_B \approx$ 20–60 tokens) and contains explicit temporal structure ("Days 1–3: sharp rise; Days 8–15: decline"). Full-rank $\mathbf{W}_{kv} \in \mathbb{R}^{768\times768}$ retains all semantic dimensions needed for the four TrendQFormer queries to distinguish short/mid/long-term trajectory phases; a low-rank bottleneck would merge these distinct phases.

Why TRA uses prefix placement. Under GPT-2's causal attention, tokens at position i can only attend to positions $\leq i$. By prepending trend tokens at positions 1–4, every downstream token (the global event token at position 5 and patch tokens at positions 6 onward) attends to the full trajectory prior, acting as a global event conditioning. This follows the prefix tuning paradigm (Li and Liang, 2021) but with functionally derived (rather than randomly initialized) prefix vectors.

B Proofs of Propositions 1–3

Proof of Proposition 1 (Semantic alignment is not necessary).

Proof. Let (f, ϕ_s, ϕ_t) be any functionally aligned model achieving $\mathbb{E}[\|f(\phi_t(e), \phi_s(\mathbf{x})) - \mathbf{y}\|^2] \leq \delta$.

For any invertible linear map $R \in \mathbb{R}^{d\times d}$, define:

$$\begin{aligned} \phi'_t(e) &= R\,\phi_t(e), \\ f'(\mathbf{h}_s, \mathbf{h}_t) &= f\big(\mathbf{h}_s,\ R^{-1}\mathbf{h}_t\big)\,. \end{aligned}$$

Prediction equivalence.

$$\begin{aligned} f'\big(\phi_s(\mathbf{x}),\ \phi'_t(e)\big) &= f'(\phi_s(\mathbf{x}),\ R\phi_t(e)) \\ &= f\big(\phi_s(\mathbf{x}),\ R^{-1}R\phi_t(e)\big) \\ &= f(\phi_s(\mathbf{x}),\ \phi_t(e))\,, \end{aligned}$$

so the prediction error under (f', ϕ_s, ϕ'_t) equals δ.

Semantic alignment violation.

$$\|\phi'_t(e) - \phi_s(\mathbf{x})\| = \|R\phi_t(e) - \phi_s(\mathbf{x})\|.$$

Choose $R = \lambda I$ with $\lambda \to \infty$; then $\|R\phi_t(e) - \phi_s(\mathbf{x})\| \to \infty$, so for any threshold ϵ, $\mathbb{E}[\|\phi'_t(e) - \phi_s(\mathbf{x})\|^2] > \epsilon$.

Since this construction works for *any* functionally aligned model, no functional alignment guarantee can imply semantic alignment. □

Proof of Proposition 2 (Contrastive objectives cause variance compression).

Proof. Following Wang and Isola (2020), the InfoNCE loss decomposes as:

$$\mathcal{L}_{\text{InfoNCE}} = \underbrace{\mathbb{E}\big[\|\phi_t(e) - \phi_s(\mathbf{x})\|^2\big]}_{\mathcal{L}_{\text{align}}} + \underbrace{\log \mathbb{E}\Big[e^{-2\|\phi_s(\mathbf{x}) - \phi_s(\mathbf{x}')\|^2}\Big]}_{\mathcal{L}_{\text{unif}}}.$$

Let $\Sigma_t = \text{Cov}(\phi_t(e))$ and $\Sigma_s = \text{Cov}(\phi_s(\mathbf{x}))$. In event-driven forecasting, event labels form a small discrete vocabulary, so $\text{tr}(\Sigma_t) \ll \text{tr}(\Sigma_s)$.

The gradient of $\mathcal{L}_{\text{align}}$ with respect to ϕ_s is:

$$\nabla_{\phi_s} \mathcal{L}_{\text{align}} = 2\,\mathbb{E}[\phi_s(\mathbf{x}) - \phi_t(e)]\,,$$

which moves $\phi_s(\mathbf{x})$ toward the mean of $\phi_t(e)$. Since the centroid of the text distribution has lower variance than the series distribution, this contraction reduces $\text{tr}(\Sigma_s)$:

$$\text{tr}\Big(\Sigma_s^{\text{after}}\Big) < \text{tr}\Big(\Sigma_s^{\text{before}}\Big)\,.$$

Reduced $\text{tr}(\Sigma_s)$ implies reduced spectral energy in the series embedding space, compressing the representation of fine-grained high-frequency dynamics (holiday spikes) that distinguish event-driven days from regular days. □

Proof of Proposition 3 (Functional alignment is sufficient).

Proof. Let $g^* : \mathbb{R}^T \times \mathcal{E} \to \mathbb{R}^H$ be any continuous forecasting function, where $\mathcal{E}$ is the set of event strings.

By the Universal Approximation Theorem (Hornik, 1991), for any $\varepsilon > 0$ there exist continuous ϕ_s, ϕ_t, f (encoders and predictor) such that:

$$\sup_{\mathbf{x},e} \big\| f(\phi_s(\mathbf{x}), \phi_t(e)) - g^*(\mathbf{x}, e) \big\| < \varepsilon.$$

Crucially, this holds with *no constraint* on $\|\phi_t(e) - \phi_s(\mathbf{x})\|$: the UAT guarantees existence of a solution in the unconstrained function space. Therefore, the prediction loss $\mathcal{L}_{\text{pred}}$ optimizes over a space that contains ε-optimal solutions for any ε, without semantic proximity requirements.

Remark. This is an existence argument and does not bound convergence rates or sample complexity; these remain open. Proposition 3 is empirically instantiated by the PROSE Full vs. Frozen RoBERTa and w/o Text variants in Section 5.4. □

C SFA: Mathematical Details

SFA is a no-parameter auxiliary loss that enforces functional similarity within each batch; each term is applied to the language encoder and its corresponding event-injection path.

Notation. Let $\mathbf{y}_{\text{norm}} = \mathbf{y}/(\bar{\mathbf{x}}_{\text{rec}} + 10^{-8})$ be trajectory normalized by a short trailing mean of the input history. Let τ=0.07 be the temperature. Let s_{ij}^f and s_{ij}^r denote functional and representation similarities. The soft InfoNCE loss is:

$$\mathcal{L} = -\sum_i \sum_{j \neq i} p_{ij}^f \log p_{ij}^r,$$

$$p_{ij}^{f/r} = \frac{\exp(s_{ij}^{f/r}/\tau)}{\sum_{k \neq i} \exp(s_{ik}^{f/r}/\tau)}.$$

ETI-SFA. Functional: pairwise cosine similarity of $\mathbf{y}_{\text{norm}}$ vectors. Representation: cosine similarity of $\text{mean}(\tilde{\mathbf{H}}_{e,i})$, the mean-pooled ETI output. Loss: soft InfoNCE (Eq. 5) with τ=0.07.

TRA-SFA. Four scalar trajectory features (f_0–f_3) are each paired with the corresponding TrendQFormer query column $\mathbf{T}_{:,k,:}$ via RBF similarity with adaptive bandwidth $\sigma_k = \text{std}(f_k) + 10^{-8}$: $f_0 = \bar{y}_{\text{norm}}[{:}H/4]$ (short-term mean), f_1 = OLS slope (trend direction), $f_2 = \arg\max(\mathbf{y})/(H-1)$ (peak position), $f_3 = \bar{y}_{\text{norm}}$ (global mean). Loss: average InfoNCE over four queries (Eq. 6).

Gradient isolation. $\mathcal{L}_{\text{ETI-SFA}}$ gradients flow to RoBERTa and the ETI module; $\mathcal{L}_{\text{TRA-SFA}}$ gradients flow to RoBERTa and TrendQFormer. The GPT-2 backbone and `output_project` are detached from both terms.

D Representation Analysis Details

Layer-wise CKA (Figure 6). We compute Centered Kernel Alignment (CKA, linear kernel) between the *same* layer's hidden states of fine-tuned and frozen RoBERTa on 216 held-out samples. Lower CKA indicates greater divergence caused by fine-tuning. Early layers (Layers 1–3, CKA>0.97)

are nearly unchanged, confirming that low-level token representations are preserved. Upper layers diverge substantially: Layer 10 reaches CKA=0.61 (most diverged), and the final Layer 12 partially recovers to CKA=0.70. This layer-wise pattern is consistent with the view that fine-tuning specializes the upper transformer layers into a functional subspace tuned for event-driven prediction.

t-SNE visualization (Figure 7). We project Layer 10 CLS representations of 90 event-labeled samples drawn from the 216 held-out set ("Other"/"No Event" samples filtered out to focus on interpretable event categories) using t-SNE (perplexity=22). In the **fine-tuned encoder** (top row), representations cluster by *prediction pattern*: Spike samples (red) separate from Flat samples (green), indicating that the encoder has organized the representation space around prediction-relevant functional categories. In the **frozen encoder** (bottom row), the same samples exhibit weaker separation by prediction pattern but retain event-semantic structure by event type (right column), showing that frozen RoBERTa preserves language semantics but lacks functional organization. Together, Figures 6 and 7 provide converging evidence for Predictive Grounding: fine-tuning shifts RoBERTa from semantic to functional representations.

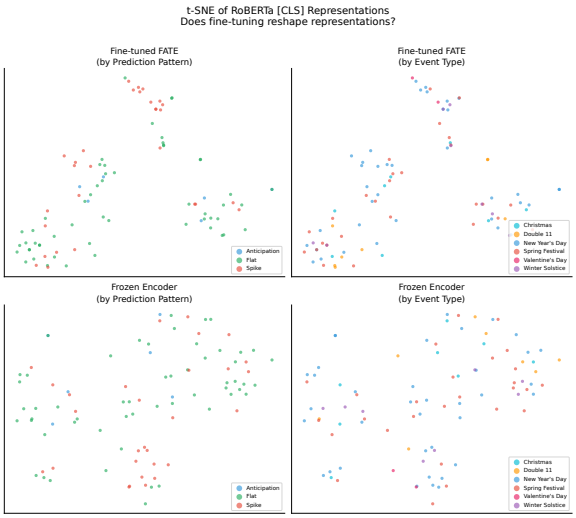

Figure 7: t-SNE of RoBERTa Layer 10 CLS representations (90 event-labeled samples). **Top:** Fine-tuned encoder; **Bottom:** Frozen encoder. **Left:** colored by prediction pattern (Spike/Anticipation/Flat); **Right:** colored by event type. Fine-tuned representations cluster by functional prediction pattern, while frozen representations retain semantic event-type structure but lack functional organization—consistent with the Predictive Grounding hypothesis.

E Ablation Statistical Details

Table 3 reports mean WMAPE_{p50} over 3 independent runs. Table 4 provides standard deviations and significance tests across runs:

Variant	WMAPE_{p50} (mean±std)	p-val vs. Frozen
PROSE Full	0.3575 ± 0.0008	—
w/o TRA	0.3589 ± 0.0009	—
w/o ETI	0.3630 ± 0.0011	—
Frozen RoBERTa	0.3971 ± 0.0012	—
w/o Text	0.3992 ± 0.0013	p=0.18

Table 4: Ablation on Query-QV: mean WMAPE_{p50} over 3 independent runs with standard deviations.

The Frozen vs. w/o Text gap (0.0021) is within 2 standard deviations of either estimate ($\approx 1.7\sigma$). We therefore carefully frame this as "near-equality" rather than claiming identity. The key finding—that both variants degrade $\approx$11–12 percentage points from PROSE Full—is statistically robust ($p < 0.001$ for Frozen vs. PROSE Full and w/o Text vs. PROSE Full).

F Dataset Details

Query-QV

Data fields. Table 5 lists all dataset fields and their descriptions.

Field	Description
`keyword`	Search query string (e.g., "Spring Festival gifts")
`event_text`	Event annotation (e.g., "Spring Festival Jan. 28–Feb. 4")
`encoder_qv_dense`	60-day historical query volume
`decoder_qv_dense`	30-day forecast target
`event_mask`	Binary holiday mask for prediction window
`path_b_text`	Qwen3-8B-generated trajectory reasoning

Table 5: Query-QV dataset fields and descriptions.

Statistics. 464,996 total samples. Train: 299,323 (2024-02-20–2025-09-30, 65% event overlap). Val: 36,877 (2025-10-14–2025-11-26, 61% holiday). Test: 104,552 (2025-12-10–2026-02-12, 84.8% holiday). The test period spans Christmas, New Year's, Spring Festival (2026-01-29), and Lantern Festival, covering the four largest holiday events in the Chinese calendar.

TRA generation prompt (simplified).

> *You are a search data analyst. Keyword: {keyword}. Event: {event_text}. Historical trend: {ts_description}. Describe how this event will affect search volume trends over the next 30 days. Focus on temporal trajectory only.*

Generation was performed with Qwen3-8B via vLLM on a single NVIDIA A100-80G GPU; processing all 464,996 Query-QV samples required

approximately 5.5 hours. Generated texts average 50–80 Chinese characters per sample.

Time-MMD

Time-MMD (Liu et al., 2024a) is a publicly available multimodal benchmark spanning nine heterogeneous domains. Numerical series are collected from authoritative government sources and cover three distinct sampling frequencies: monthly (Agriculture, Climate, Economy, Security, SocialGood, Traffic), weekly (Energy, Health), and daily (Environment). For each domain, one or two target variables with clear real-world significance are selected; additional covariates are provided for some domains.

Domain summary. Table 6 summarizes the nine domains, their target variables, sampling frequencies, and time spans.

Domain	Target Variable	Freq.	Timespan
Agriculture	Retail Broiler Composite Price	Monthly	1983–present
Climate	Drought Level	Monthly	1983–present
Economy	International Trade Balance	Monthly	1989–present
Energy	Gasoline Prices	Weekly	1996–present
Environment	Air Quality Index	Daily	1982–2023
Health	Influenza Patients Proportion	Weekly	1997–present
Security	Disaster & Emergency Grants	Monthly	1999–present
SocialGood	Unemployment Rate	Monthly	1950–present
Traffic	Travel Volume	Monthly	1980–present

Table 6: Time-MMD domain summary: target variables, sampling frequencies, and time spans.

Text modality. Each time step is paired with a textual description assembled from two sources: (1) *selected reports*—one to two domain-relevant report series manually curated per domain (e.g., weekly CDC influenza reports for Health); and (2) *web search results*—two to three domain-specific keywords used to retrieve contemporaneous news and announcements. Texts are temporally aligned to the numerical series via binary timestamp markers that indicate the reporting window of each document.

Evaluation protocol. Consistent with Liu et al. (2024a), we use the following input/horizon configurations: monthly domains (L=8, $H \in \{6, 8, 10, 12\}$), weekly domains (L=24, $H \in \{12, 24, 36, 48\}$), and the daily Environment domain (L=96, $H \in \{48, 96, 192, 336\}$). MSE and MAE are averaged over all horizons for each domain.

G Time-MMD Detailed Results

Table 7 reports per-horizon results on all 9 Time-MMD domains.

H PROSE Model Configuration

For Query-QV (Chinese text), the language encoder is Chinese RoBERTa-wwm-ext (Cui et al., 2021); for Time-MMD (English text), we use RoBERTa-base (Liu et al., 2019). All other architectural components are identical across datasets. Table 8 summarizes the full training configuration.

Software dependencies. All experiments are implemented in PyTorch 2.1 with HuggingFace Transformers 4.40 for model loading and fine-tuning. Offline TRA generation uses vLLM 0.4 for efficient Qwen3-8B batch inference. RevIN normalization follows the original implementation (Kim et al., 2022). No preprocessing or evaluation packages with user-configurable parameters are used; all evaluation metrics (MSE, MAE, WMAPE) are computed from standard tensor operations.

I Qwen3-8B Offline Reasoning Prompts

This appendix presents the complete Qwen3-8B prompts used for offline TRA temporal reasoning generation across all datasets. All prompts follow a common three-part structure: (1) a *system instruction* establishing the analyst role, (2) a *context block* providing keyword, event text, and a language-based historical trend summary (§4), and (3) a *generation constraint* restricting output to temporal trajectory descriptions only—no causal explanations. The language-based trend summary (`{ts_description}`) encodes five dimensions: overall trend (linear slope), weekly periodicity (FFT peak day), recent level (last-7-day vs. historical mean), volatility (coefficient of variation), and anomaly detection (3σ threshold).

Common template structure. Each prompt instantiates the following skeleton:

General Prompt Template

```
[SYSTEM]
```

Table 7: Per-horizon results on all 9 Time-MMD domains (MSE / MAE). **Red bold**: best. Blue underline: second best.

Category		Ours		TS-only												Multimodal							
Models		**PROSE**		iTransformer		PatchTST		DLinear		TimesNet		TimeXer		Autoformer		GPT4TS		GPT4MTS		TaTS		TimeVLM	
Metric	T	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Agriculture	6	0.136	**0.253**	0.145	0.256	0.140	0.254	0.378	0.470	0.163	0.256	0.140	0.254	0.154	0.269	**0.135**	0.265	0.186	0.282	0.217	0.316	0.152	0.253
	8	**0.185**	**0.283**	0.203	0.292	0.186	0.285	0.375	0.434	0.190	0.285	0.212	0.285	0.202	0.298	0.220	0.299	0.221	0.295	0.266	0.346	0.202	0.287
	10	**0.228**	**0.316**	0.243	0.318	0.247	0.317	0.406	0.432	0.262	0.320	0.245	0.316	0.266	0.329	0.238	0.322	0.337	0.365	0.326	0.379	0.286	0.334
	12	**0.284**	**0.352**	0.290	0.352	0.316	0.359	0.469	0.477	0.300	0.366	0.304	0.354	0.348	0.396	0.330	0.357	0.380	0.391	0.401	0.421	0.322	0.354
	avg	**0.208**	**0.301**	0.220	0.304	0.222	0.304	0.407	0.453	0.229	0.307	0.225	0.302	0.242	0.323	0.231	0.311	0.281	0.333	0.302	0.365	0.240	0.307
Climate	6	**1.210**	0.900	1.337	0.941	1.293	0.924	1.353	0.947	1.380	0.957	1.217	**0.899**	1.291	0.931	1.247	0.910	1.272	0.921	1.217	0.903	1.250	0.919
	8	1.207	0.898	1.283	0.923	1.239	0.911	1.359	0.947	1.430	0.965	**1.198**	**0.894**	1.290	0.918	1.218	0.916	1.258	0.922	1.215	0.927	1.215	0.906
	10	**1.190**	**0.889**	1.227	0.905	1.249	0.905	1.255	0.912	1.292	0.924	1.218	0.903	1.230	0.901	1.249	0.918	1.245	0.917	1.237	0.908	1.202	0.899
	12	1.199	0.897	1.264	0.919	**1.175**	**0.883**	1.277	0.915	1.376	0.945	1.225	0.907	1.238	0.908	1.233	0.928	1.252	0.913	1.204	0.915	1.220	0.906
	avg	**1.202**	**0.896**	1.278	0.922	1.239	0.906	1.311	0.930	1.369	0.948	1.215	0.901	1.262	0.914	1.237	0.918	1.257	0.918	1.218	0.913	1.222	0.908
Economy	6	**0.183**	**0.337**	0.212	0.365	0.197	0.347	0.328	0.472	0.229	0.379	0.211	0.360	0.230	0.380	0.201	0.355	0.186	0.339	1.837	1.273	0.215	0.371
	8	**0.203**	**0.357**	0.217	0.364	0.212	0.364	0.305	0.436	0.250	0.394	0.210	0.362	0.237	0.389	0.220	0.376	0.206	0.362	1.104	0.923	0.248	0.404
	10	**0.211**	**0.361**	0.227	0.375	0.216	0.369	0.297	0.437	0.262	0.396	0.218	0.369	0.257	0.404	0.218	0.371	0.216	0.368	1.777	1.237	0.247	0.400
	12	**0.229**	**0.375**	0.231	0.380	0.229	0.377	0.362	0.484	0.262	0.404	0.232	0.383	0.270	0.419	0.235	0.388	0.244	0.398	1.514	1.094	0.252	0.405
	avg	**0.206**	**0.357**	0.222	0.371	0.213	0.364	0.323	0.457	0.251	0.393	0.218	0.368	0.248	0.398	0.218	0.372	0.213	0.367	1.558	1.132	0.240	0.395
Energy	12	0.128	0.267	0.124	0.276	**0.108**	**0.240**	0.208	0.347	0.155	0.285	0.142	0.292	0.236	0.384	0.133	0.277	0.119	0.262	0.624	0.588	0.118	0.248
	24	**0.236**	**0.365**	0.257	0.380	0.282	0.396	0.315	0.417	0.236	0.373	0.262	0.399	0.372	0.480	0.236	0.367	0.254	0.385	0.351	0.435	0.244	0.376
	36	0.334	0.432	0.307	**0.418**	**0.302**	0.422	0.377	0.460	0.350	0.430	0.349	0.452	0.396	0.500	0.331	0.449	0.340	0.442	0.398	0.435	0.312	0.419
	48	0.414	0.495	0.433	0.492	0.465	0.548	0.473	0.508	0.463	0.546	0.425	0.509	0.478	0.552	0.420	0.499	0.419	0.500	0.439	0.520	**0.393**	**0.475**
	avg	0.278	0.390	0.280	0.392	0.289	0.402	0.343	0.433	0.301	0.408	0.294	0.413	0.370	0.479	0.280	0.398	0.283	0.397	0.453	0.494	**0.267**	**0.379**
Environment	48	0.310	**0.387**	0.307	0.392	0.323	0.402	0.381	0.505	0.310	0.394	0.312	0.391	0.351	0.427	0.312	0.394	0.314	0.390	0.422	0.461	**0.305**	0.387
	96	0.339	**0.407**	0.335	0.409	0.355	0.417	0.540	0.630	0.334	0.409	0.356	0.432	0.404	0.506	0.347	0.415	0.342	0.413	0.468	0.481	**0.329**	0.410
	192	0.353	0.417	0.343	0.408	0.352	0.420	0.693	0.741	0.332	**0.396**	0.346	0.428	0.339	0.431	0.337	0.422	0.363	0.436	0.472	0.473	**0.330**	0.433
	336	0.314	0.407	0.311	0.398	0.310	0.406	0.563	0.636	**0.300**	**0.391**	0.305	0.393	0.323	0.432	0.328	0.424	0.309	0.407	0.373	0.424	0.321	0.423
	avg	0.329	0.405	0.324	0.402	0.335	0.411	0.544	0.628	**0.319**	**0.397**	0.330	0.411	0.354	0.449	0.331	0.414	0.332	0.411	0.434	0.460	0.321	0.413
Health	12	**0.952**	**0.611**	1.005	0.636	0.973	0.718	1.609	0.838	1.080	0.655	1.094	0.711	1.874	0.982	1.045	0.629	1.185	0.754	1.352	0.835	1.152	0.709
	24	1.560	**0.743**	1.629	0.769	1.438	0.762	1.750	0.884	1.688	0.815	1.554	0.802	2.315	1.080	**1.395**	0.748	1.592	0.826	1.497	0.835	1.520	0.853
	36	2.019	0.871	1.871	0.877	1.801	0.883	1.693	0.887	2.056	0.919	1.702	0.860	2.136	1.076	1.695	0.909	1.800	0.964	**1.568**	**0.835**	1.572	0.872
	48	2.164	0.915	2.120	0.908	1.917	0.931	1.765	0.920	2.291	0.965	1.940	**0.900**	1.778	0.977	2.083	0.914	1.937	0.971	1.698	0.959	**1.672**	0.917
	avg	1.674	**0.785**	1.656	0.797	1.532	0.824	1.704	0.882	1.779	0.839	1.572	0.818	2.026	1.029	1.554	0.800	1.629	0.879	1.529	0.866	**1.479**	0.838
Security	6	**64.926**	**3.865**	70.460	4.098	68.124	4.063	66.391	4.037	76.087	4.367	73.789	4.427	74.199	4.193	65.446	3.964	70.397	4.093	68.213	3.916	67.251	4.163
	8	**68.826**	**3.937**	77.217	4.412	73.333	4.293	70.999	4.220	85.330	4.756	76.277	4.377	79.679	4.363	69.545	3.978	71.157	4.117	70.214	4.296	72.043	4.123
	10	**72.866**	**4.077**	79.289	4.445	80.579	4.529	75.842	4.438	85.224	4.821	77.782	4.388	81.926	4.406	73.561	4.907	76.187	4.407	74.593	4.122	77.544	4.327
	12	**76.050**	**4.095**	84.964	4.569	83.738	4.645	79.834	4.541	91.184	5.061	84.348	4.725	82.898	4.380	76.523	4.178	84.207	4.720	78.005	4.129	86.490	4.795
	avg	**70.667**	**3.994**	77.983	4.381	76.444	4.383	73.266	4.309	84.456	4.751	78.049	4.479	79.675	4.335	71.269	4.257	75.487	4.334	72.756	4.116	75.832	4.352
SocialGood	6	**0.690**	**0.381**	1.003	0.418	0.959	0.406	0.891	0.501	1.207	0.436	1.141	0.488	0.971	0.447	0.796	0.387	0.814	0.402	0.921	4.251	0.706	0.398
	8	**0.736**	**0.443**	1.107	0.458	1.061	0.534	0.973	0.543	1.835	0.565	0.988	0.519	0.911	0.504	0.861	0.466	0.875	0.461	0.968	0.484	0.851	0.445
	10	**0.819**	**0.430**	1.428	0.533	1.069	0.468	1.119	0.633	1.624	0.532	1.215	0.498	1.245	0.599	0.938	0.459	0.944	0.491	1.059	0.523	0.916	0.472
	12	**0.829**	**0.473**	1.342	0.551	1.240	0.517	1.083	0.542	1.696	0.535	1.159	0.505	1.098	0.557	1.067	0.529	1.287	0.559	1.155	0.552	0.988	0.495
	avg	**0.768**	**0.432**	1.220	0.490	1.082	0.481	1.016	0.555	1.591	0.517	1.126	0.503	1.056	0.527	0.915	0.460	0.980	0.478	1.026	1.453	0.865	0.453
Traffic	6	**0.112**	**0.190**	0.123	0.218	0.167	0.204	0.271	0.408	0.157	0.209	0.113	0.208	0.142	0.262	0.128	0.229	0.119	0.212	0.762	0.689	0.144	0.263
	8	**0.119**	**0.201**	0.122	0.204	0.168	0.221	0.287	0.420	0.170	0.219	0.120	0.204	0.172	0.297	0.131	0.220	0.121	0.249	0.996	0.747	0.141	0.252
	10	**0.125**	**0.206**	0.125	0.226	0.177	0.247	0.247	0.384	0.175	0.215	0.169	0.210	0.150	0.268	0.132	0.218	0.134	0.221	1.039	0.758	0.143	0.259
	12	**0.132**	**0.216**	0.200	0.250	0.196	0.246	0.321	0.453	0.181	0.217	0.197	0.238	0.155	0.275	0.225	0.280	0.229	0.268	1.121	0.787	0.145	0.258
	avg	**0.122**	**0.203**	0.143	0.225	0.177	0.229	0.282	0.416	0.171	0.215	0.150	0.215	0.155	0.275	0.154	0.237	0.151	0.237	0.979	0.745	0.143	0.258

Component	Configuration
Language encoder	Chinese RoBERTa-wwm-ext (Query-QV) RoBERTa-base (Time-MMD); both fully fine-tuned
LRGP rank	32
Cross-Attn injection	GPT-2 layers 0–5
TrendQFormer queries	4 learnable queries
Series encoder	MOMENT-1-small (frozen)
Forecasting backbone	GPT-2 (12 layers, d=768), fully fine-tuned
Seq. / horizon (Query-QV)	T=60, H=30
Optimizer	AdamW
LR (RoBERTa / GPT-2 / adapters)	2e-5 / 1e-4 / 5e-4
Batch size	128
Training epochs	15 (Query-QV), 10 (Time-MMD)
Trainable params	≈240M
GPU	NVIDIA A100-80G

Table 8: PROSE training configuration.

```
You are a time series analyst specializing in
event-driven forecasting.

[USER]
Domain:     domain
Variable:   variable_description
Event:      event_text
Historical trend (past 60 steps):
ts_description

Task: Describe how this event will affect the
variable's
trajectory over the NEXT pred_lentimesteps.
Rules:
- Describe temporal trajectory ONLY (no causal
explanations).
- Use time references: "Days 1-3", "Week 2",
etc.
- Keep the response under 80 words.
```

Query-QV (Chinese Search Query Volume)

The Query-QV dataset contains Chinese search keywords annotated with holiday and event information. The prediction horizon is 30 days ($H = 30$). `ts_description` is computed per-sample from the 60-day history window.

Query-QV — TRA Reasoning Prompt

```
[SYSTEM]
You are a search data analyst specializing in Chinese consumer behavior.

[USER]
Search keyword: keyword
Event: event_text
Historical search volume trend (past 60 days): ts_description
(e.g., "Stable overall, strong weekly cycle peaking on Fridays,
currently at normal level, moderate volatility, no anomaly")

Task: Describe how this event will affect search volume
trends over the NEXT 30 days.
Rules:
- Describe the temporal trajectory ONLY.
- Do NOT explain why; only describe what happens.
- Use phased time references: "Days 1-3", "Days 4-7",
"Days 8-15", "Days 16-30".
- Keep the response under 80 Chinese characters.

[Example output]
Days 1-3: search volume begins rising sharply.
Days 4-7: reaches peak and stays elevated.
Days 8-15: gradually declines toward normal.
Days 16-30: returns to baseline, stable.
```

Time-MMD Benchmark (9 Domains)

The Time-MMD benchmark (Liu et al., 2024a) covers 9 heterogeneous domains. Since each domain has distinct variable semantics and event types, we use *domain-specific* system instructions while keeping the generation constraint identical. The prediction horizon H matches the Time-MMD standard setting (6/8/10/12 steps for short-term domains; 48/96/192/336 for long-term domains).

Agriculture — Crop Price / Yield Forecasting

```
[SYSTEM]
You are an agricultural market analyst specializing in
supply-demand dynamics.

[USER]
Variable: variable_description
(e.g., "weekly average corn futures price, USD/bushel")
Event: event_text
(e.g., "USDA crop report released; severe drought warning
issued for the Midwest")
Historical trend (past seq_lensteps):
ts_description

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred_lentimesteps.
Rules:
- Temporal trajectory ONLY (no causal explanations).
- Use step-based references: "Steps 1-2", "Steps 3-6", etc.
- Under 80 words.
```

Climate — Meteorological Variable Forecasting

```
[SYSTEM]
You are a climate scientist analyzing short-term atmospheric
anomalies and their downstream effects.

[USER]
Variable: variable_description
(e.g., "daily mean temperature anomaly, Celsius")
Event: event_text
(e.g., "El Nino advisory issued; above-normal SST
anomaly confirmed in the Nino-3.4 region")
Historical trend (past seq_lensteps):
ts_description

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred_lentimesteps.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.
```

Economy — Macroeconomic Indicator Forecasting

[SYSTEM]
You are a macroeconomist specializing in event-driven market
analysis.

[USER]
Variable: variable$_{d}escription$
(e.g., "monthly unemployment rate, percentage")
Event: event$_{t}ext$
(e.g., "Federal Reserve announces 25bp rate hike;
CPI print exceeds consensus by 0.3pp")
Historical trend (past seq$_{l}ensteps$) :
ts$_{d}escription$

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred$_{l}entimesteps$.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.

Energy — Power Load / Consumption Forecasting

[SYSTEM]
You are an energy systems analyst with expertise in demand-side
load forecasting.

[USER]
Variable: variable$_{d}escription$
(e.g., "hourly electricity consumption, MWh")
Event: event$_{t}ext$
(e.g., "Extreme cold snap forecast for the Northeast;
temperatures expected 15 degrees C below seasonal avg")
Historical trend (past seq$_{l}ensteps$) :
ts$_{d}escription$

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred$_{l}entimesteps$.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.

Environment — Air Quality / Pollution Forecasting

[SYSTEM]
You are an environmental scientist specializing in air quality
modeling and pollution event analysis.

[USER]
Variable: variable$_{d}escription$
(e.g., "daily average PM2.5 concentration, ug/m^{3}")
Event: event$_{t}ext$
(e.g., "Heavy industrial production restart post-holiday;
stagnant atmospheric conditions forecast for 5 days")
Historical trend (past seq$_{l}ensteps$) :
ts$_{d}escription$

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred$_{l}entimesteps$.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.

Health — Epidemiological Indicator Forecasting

[SYSTEM]
You are an epidemiologist specializing in infectious disease
surveillance and outbreak dynamics.

[USER]
Variable: variable$_{d}escription$
(e.g., "weekly influenza-like illness (ILI) rate,
per 100,000 population")
Event: event$_{t}ext$
(e.g., "WHO declares influenza H3N2 variant of concern;
vaccine effectiveness estimated at 42Historical trend (past seq$_{l}ensteps$) :
ts$_{d}escription$

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred$_{l}entimesteps$.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.

Security — Crime / Incident Rate Forecasting

[SYSTEM]
You are a public safety analyst specializing in crime pattern
analysis and event-driven anomaly detection.

[USER]
Variable: variable$_{d}escription$
(e.g., "weekly reported property crime incidents,
city-level count")
Event: event$_{t}ext$
(e.g., "Major public event (music festival, 50,000
attendees) scheduled; police deployment increased 30Historical trend (past seq$_{l}ensteps$) :
ts$_{d}escription$

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred$_{l}entimesteps$.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.

SocialGood — Social Welfare Indicator Forecasting

[SYSTEM]
You are a social policy analyst specializing in welfare
program utilization and community impact assessment.

[USER]
Variable: variable$_{d}escription$
(e.g., "monthly food bank visits, regional count")
Event: event$_{t}ext$
(e.g., "Government announces suspension of supplemental
nutrition assistance; effective in 30 days")
Historical trend (past seq$_{l}ensteps$) :
ts$_{d}escription$

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred$_{l}entimesteps$.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.

Traffic — Vehicle Flow / Congestion Forecasting

[SYSTEM]
You are a transportation engineer specializing in urban traffic
flow analysis and event-driven demand forecasting.

[USER]
Variable: variable$_{d}escription$
(e.g., "hourly vehicle count at highway toll station")
Event: event$_{t}ext$
(e.g., "National holiday announced; long weekend travel
demand surge expected; nearby road construction begins")
Historical trend (past seq$_{l}ensteps$) :
ts$_{d}escription$

Task: Describe how this event will affect the variable's
trajectory over the NEXT pred$_{l}entimesteps$.
Rules:
- Temporal trajectory ONLY.
- Step-based references required.
- Under 80 words.

Historical trend description format. Across all prompts, {ts_description} is a natural-language string generated by the following five-dimensional encoder (computed offline in Python, stored in the dataset parquet):

ts_description Five-Dimensional Encoder

Dim 1 (Trend): linear regression slope

```
over full window
-> "strong uptrend" | "mild uptrend" | "stable" |
"mild downtrend" | "strong downtrend"

Dim 2 (Periodicity): FFT-detected dominant cycle strength
-> "strong weekly cycle, peak on <weekday>" | "no clear cycle"

Dim 3 (Recent lvl): last-7-step mean vs. historical mean
-> "at historical high" | "slightly above average" |
"normal" | "slightly below average" | "at historical low"

Dim 4 (Volatility): coefficient of variation (std / mean)
-> "high volatility" | "moderate volatility" | "stable"

Dim 5 (Anomaly): 3-sigma peak detection over full window
-> "anomalous spike N steps ago" | "no anomaly"

Example output:
"Stable overall, strong weekly cycle peaking on Fridays,
currently at normal level, moderate volatility, no anomaly"
```