

TimePrune: Budgeted Temporal Pruning for Efficient Multimodal Time Series Analysis

Chenghao Liu¹, Qian Liu¹, Aniket Mahanti¹, Jindong Han², Min Li³, Wei Fan^{1*}

¹School of Computer Science, University of Auckland, New Zealand

²School of Artificial Intelligence, Shandong University, China

³School of Computer Science and Engineering, Central South University, China

{chenghao.liu, liu.qian, a.mahanti, wei.fan}@auckland.ac.nz, jindong.han@sdu.edu.cn, limin@mail.csu.edu.cn

Abstract

The recent advancement of time series has witnessed Vision Language Models (VLMs) as a powerful tool for a variety of multimodal time series analysis tasks, including forecasting, classification, question answering, etc. However, with time series as an explicit modality, inputs with long time series horizons yield excessive numeric tokens for VLMs, resulting in high inference latency. Moreover, such long contexts often contain redundant or noisy segments that might distract the model and thus degrade downstream performance. Although previous token pruning methods can reduce visual or textual tokens, they are mostly unimodal and do not explicitly explore multimodal semantics for time series. To address this, we propose *TimePrune*, a plug-and-play temporal pruning framework for VLMs in multimodal time series analysis, which can accelerate VLM inference without fine-tuning the VLM backbone. Specifically, TimePrune formulates time series pruning as a budgeted temporal selection problem and introduces a Trimodal Budgeted Network to score temporal patches via jointly conditioning on linguistic and visual semantics. Then, under a fixed budget, this framework is optimized via the proposed Narrative-Calibrated Binned Grounding constraints, to further facilitate context-aware and temporally-coherent pruning. Extensive experiments on four time series tasks across application domains on multiple VLMs show that TimePrune could consistently achieve better performances. For example, on Qwen3VL-8B, TimePrune improves average forecasting performance by 33.91%. Our code is publicly available.¹

1 Introduction

Vision Language Models (VLMs) have emerged as a natural extension of Large Language Models

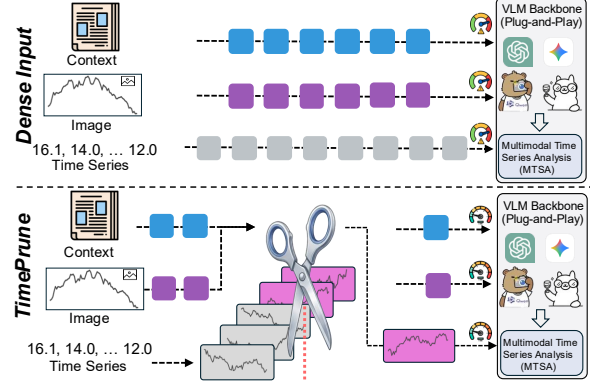


Figure 1: **Top:** Existing multimodal time series pipelines feed dense inputs into a VLM, resulting in high computational cost and latency. **Bottom:** The proposed TimePrune accelerates inference by selectively removing redundant tokens.

(LLMs), transferring their strong language understanding capabilities to multimodal contexts such as images and video (Liu et al., 2023; Lin et al., 2024; Team, 2025a). Thus, VLMs have demonstrated strong generalization across a wide range of real-world applications, including document understanding (Nacson et al., 2025; Liu et al., 2025a), chart and diagram reasoning (Mukhopadhyay et al., 2024), and video analysis (Arefeen et al., 2024). More recently, with time series being an important modality in scenarios across domains such as finance, weather, etc, VLMs have been increasingly explored in multimodal time series analysis, where they show powerful performance in interpreting temporal signals, narrative descriptions, and auxiliary modalities (Jiang et al., 2025).

Multimodal time series analysis (MTSA) covers a broad spectrum of tasks, ranging from forecasting (Fan et al., 2022, 2023; Fei et al., 2026), classification (Li et al., 2023) to anomaly detection (Zhou and Yu, 2025), as well as causal discovery (Shen et al., 2025). To address these MTSA tasks, recent VLM-based approaches typically incorporate time

*corresponding author

¹<https://github.com/coenliu/timeprune>

series alongside textual and tabular signals (Feng et al., 2025), by encoding them into unified prompts for instruction fine-tuning (Wang et al., 2024), aligning multimodal representations (Lee et al., 2025), or leveraging meta-information (e.g., domain, data statistics and granularity, and auxiliary covariates) to provide real-world context (Liu et al., 2024). However, directly adapting VLMs to the time series domain introduces a distinct computational bottleneck: as models often need to process long, high-frequency numerical time series sequences, *this substantially increases inference cost and limits scalability of VLMs*. Moreover, lengthy time horizons are frequently noisy and contain irrelevant temporal segments. When too many redundant numeric tokens are retained, *these distractions could largely hinder the performance of downstream time series tasks*.

To address the above challenges, we propose **TimePrune**, a novel model-agnostic framework for efficient and effective multimodal time series analysis. The core idea of TimePrune is to perform context-aware selective pruning on the time series in order to make VLMs retain semantically informative tokens and prune redundant noise. Specifically, we first formalize the token selection for time series as a *Budgeted Temporal Pruning* problem, which aims to maximize downstream utility under a strict budget constraint by selecting at most context-conditioned temporal patches via conditioned on the multimodal context to form a pruned context. Next, since existing VLM pruning methods usually focus on spatial redundancy in images, we perform trimodal-conditioned budgeted pruning via a *Trimodal Budgeted Network*, which could learn the interactions among text, image, and time to dynamically filter out irrelevant temporal segments based on textual and visual signals. Finally, to mitigate the risk of breaking temporal continuity during pruning, we propose *Narrative Calibrated Binned Grounding*, an optimization objective that leverages textual narratives as supervision. It employs a joint regularization loss to ensure that the selected tokens remain narrative-grounded and coherent. In summary, our contributions are as follows:

- We propose **TimePrune**, a plug-and-play framework for frozen VLMs that employs a *Trimodal Budgeted Network* to dynamically select informative time series tokens based on multimodal semantics.
- We introduce **Narrative Calibrated Binned**

Grounding, which leverages textual narratives as supervisory signals. This mechanism optimizes a compound loss to strictly enforce temporal coherence and cross-modal grounding under a constrained token budget.

- We conduct extensive experiments in finance and weather domains. The results demonstrate that TimePrune achieves a superior latency-accuracy trade-off compared to baselines, delivering speedups while maintaining or even improving predictive performance.

2 Related Works

LLMs for Time Series. Recent studies have extensively explored the adaptation of LLMs to time series analysis (Jiang et al., 2025; Liu et al., 2025c). Some approaches incorporate textual information into time series tasks using LLMs (Wang et al., 2025a; Jin et al., 2024b; Su et al., 2025), while others leverage cross-modal interactions to achieve deeper insights and robust predictive and analytical performance (Zhong et al., 2025). Beyond these directions, recent research combines LLMs with time series modeling to advance a variety of downstream applications, including forecasting (Gruver et al., 2023), anomaly detection (Liu et al., 2025d), and financial prediction (Yu et al., 2024; Liu et al., 2025a). However, few studies explicitly incorporate task-aware temporal segment selection when applying LLMs to long-horizon inputs.

Multimodal Time Series Analysis. Multimodal approaches have shown that integrating diverse data modalities enhances contextual representation, leading to better performance of MLLMs in forecasting (Liu et al., 2025b), classification (Lee et al., 2025), and causal discovery (Zheng et al., 2024). Beyond time series signals, recent research also explores the incorporation of textual and visual modalities, where text provides semantic context and images capture spatial cues, together enhancing the forecasting ability of multimodal LLMs (Jin et al., 2024a; Zhong et al., 2025; Chen et al., 2026). Few of them incorporate selection mechanisms to filter out irrelevant temporal segments, often leading to redundancy and increased computational costs when processing long-horizon data.

MLLMs Token Pruning. Computational efficiency is a significant challenge in MLLM-based time series analysis, especially when dealing with long input horizons and multivariate streams (Liu et al., 2025c). Recent work has explored strategies

for efficient time series modeling, including storing the last token in forecasting (Liu et al., 2025b), a pruning and fine-tuning approach that structurally adapts time series foundation models to specific forecasting tasks (Zhao et al., 2025), and SPAT, which applies sensitivity-based structured pruning to remove low importance multi-head attention modules in Transformer forecasters (Guo et al., 2025). However, these methods predominantly rely on unimodal signals, overlooking the potential of utilizing a cross-modal semantic context to guide the pruning of temporal patches.

3 TimePrune

In this section, we propose *TimePrune*, a model-agnostic budgeted temporal pruning framework for multimodal time series analysis. As illustrated in Figure 2, TimePrune first formulates budgeted temporal pruning under a fixed token budget as described in Sec. 3.1. It then introduces a lightweight Trimodal Budgeted Network that scores temporal patches conditioned on textual and visual context as described in Sec. 3.2. Finally, it trains the selector with Narrative Calibrated Binned Grounding, using pseudo labels and coherence regularization as described in Sec. 3.3.

Task Formulation. we consider the task of *Multimodal Time Series Analysis*, where the goal is to generate a target response $\mathbf{Y}$ (e.g., forecast values or analytical answers) from a multimodal input. Each instance is represented as $\mathcal{D} = (\mathbf{P}, \mathcal{C}, \mathcal{V})$, where $\mathbf{P} = \{p_1, \dots, p_N\}$ is the sequence of patched temporal tokens from a multivariate time series ($p_i \in \mathbb{R}^d$), $\mathcal{C}$ is a sequence of context tokens (e.g., news, queries), and $\mathcal{V}$ is the visual input. The frozen VLM backbone with parameters Φ induces a conditional distribution:

$$P_{\Phi}(\mathbf{Y} \mid \mathbf{P}, \mathcal{C}, \mathcal{V}), \quad (1)$$

3.1 Budgeted Temporal Pruning

Multimodal time series analysis typically feeds the full patched sequence $\mathbf{P} = \{p_i\}_{i=1}^N$, together with textual context $\mathcal{C}$ and a plot token $\mathcal{V}$, into a frozen VLM backbone $P_{\Phi}(\mathbf{Y} \mid \mathbf{P}, \mathcal{C}, \mathcal{V})$. However, when N is large, directly processing $\mathbf{P}$ incurs high computation and latency. We therefore study temporal token pruning for VLM time series analysis: given a token budget $K \ll N$, we seek a **context-conditioned** subset $\tilde{\mathbf{P}}_{\theta} \subseteq \mathbf{P}$ that preserves the most informative windows *conditioned on* $(\mathcal{C}, \mathcal{V})$, and is injected into VLM before inference.

Formally, we define a pruning policy f_{θ} that selects $\tilde{\mathbf{P}}_{\theta} = f_{\theta}(\mathbf{P}, \mathcal{C}, \mathcal{V})$ subject to $|\tilde{\mathbf{P}}_{\theta}| \leq K$. Under the frozen backbone Φ , the ideal objective:

$$\max_{\theta} \mathbb{E}_{\mathcal{D}} \left[\log P_{\Phi}(\mathbf{Y} \mid \tilde{\mathbf{P}}_{\theta}, \mathcal{C}, \mathcal{V}) \right], \quad (2)$$

which characterizes the best reasoning performance achievable under a fixed budget. In practice, direct optimization is intractable due to the non-differentiable top- k selection (Jang et al., 2017), motivating a lightweight trainable Trimodal Budgeted Network and a narrative calibrated binned grounding surrogate objective for alignment.

3.2 Trimodal Budgeted Network

The core of TimePrune is a **Trimodal Budgeted Network** g_{θ} that assigns temporal patches by jointly conditioning on linguistic and visual semantics in the multimodal context $(\mathcal{C}, \mathcal{V})$. Concretely, we encode the patched time series into temporal embeddings and construct a context memory from frozen text and vision encoders; the temporal embeddings then attend to this memory via multi-head cross-attention, followed by a lightweight MLP head to produce logits $\mathbf{s}_{\theta}$ and probabilities $\mathbf{p}_{\theta} = \sigma(\mathbf{s}_{\theta})$. At inference time, we apply top- k selection on $\mathbf{p}_{\theta}$ to obtain a sparse mask and pruned temporal patches $\tilde{\mathbf{P}}_{\theta}$ under the token budget.

Temporal Patching and Instance Normalization. Given an input series $\mathbf{x} \in \mathbb{R}^T$, we first apply instance normalization to reduce scale variation across instances:

$$\hat{\mathbf{x}} = \frac{\mathbf{x} - \mu(\mathbf{x})}{\sigma(\mathbf{x})}, \quad (3)$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are computed over the time dimension. We then discretize $\hat{\mathbf{x}}$ into N temporal patches and compute bin-level features (e.g., summary statistics and relative position) to form $X \in \mathbb{R}^{N \times F}$, which is fed into the Time Encoder.

Multimodal Context Encoding. From patched temporal tokens $\mathbf{P} = \{p_1, \dots, p_N\}$ we derive bin-level features $X \in \mathbb{R}^{N \times F}$ (statistics and relative position) and encode them with a Time Encoder:

$$\mathbf{H}_t = \text{TimeEnc}(X). \quad (4)$$

A frozen text encoder produces text embeddings:

$$\mathbf{H}_{\text{txt}} = \text{Enc}_{\text{text}}(\mathcal{C}), \quad (5)$$

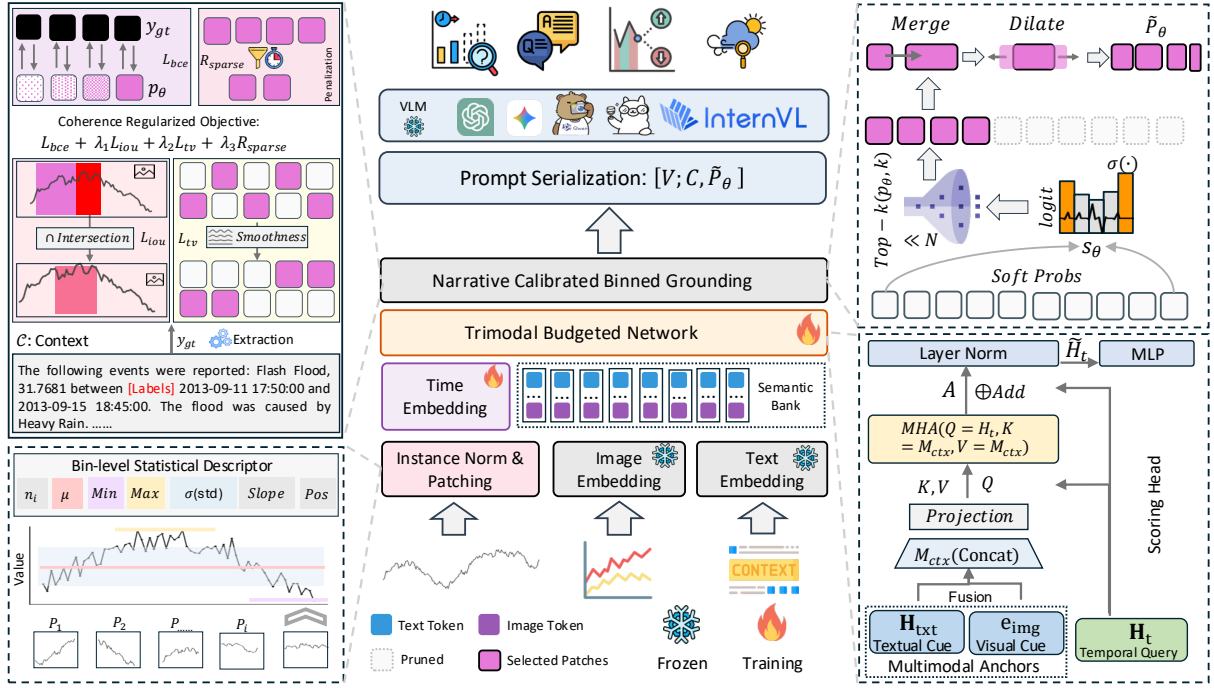


Figure 2: An overview of TimePrune. We normalize and patch the time series into N windows, then use a lightweight trimodal budgeted network conditioned on $(C, \mathcal{V})$ to score windows and select top- k patches $\tilde{\mathbf{P}}_\theta$ for a frozen VLM prompt $[V; C; \tilde{\mathbf{P}}_\theta]$. The gate is trained with narrative derived pseudo labels $\mathbf{y}_{gt}$ and a compound loss.

and a frozen vision encoder yields a visual token:

$$\mathbf{e}_{img} = \text{Enc}_{img}(\mathcal{V}). \quad (6)$$

Both are projected to the D -dimensional space and concatenated:

$$\mathbf{M}_{ctx} = [\text{Proj}_{txt}(\mathbf{H}_{txt}); \text{Proj}_{img}(\mathbf{e}_{img})]. \quad (7)$$

Budgeted Patch Selection via Trimodal Gating. We align patch relevance via cross-attention to the multimodal anchors and then convert the scores into a budget constrained top- k mask.

$$\mathbf{A} = \text{MHA}(\mathbf{Q} = \mathbf{H}_t, \mathbf{K} = \mathbf{M}_{ctx}, \mathbf{V} = \mathbf{M}_{ctx}), \quad (8)$$

followed by a residual connection and layer normalization:

$$\tilde{\mathbf{H}}_t = \text{LayerNorm}(\mathbf{H}_t + \mathbf{A}). \quad (9)$$

A lightweight MLP head produces patch-level logits and probabilities:

$$\mathbf{s}_\theta = h_\theta(\tilde{\mathbf{H}}_t) \in \mathbb{R}^N, \quad \mathbf{p}_\theta = \sigma(\mathbf{s}_\theta). \quad (10)$$

At inference time, $\mathbf{p}_\theta$ is converted into a sparse mask $\mathbf{m}_\theta \in \{0, 1\}^N$ via top- k selection and the pruned set is:

$$\tilde{\mathbf{P}}_\theta = \{p_i \in \mathbf{P} \mid m_{\theta,i} = 1\}, \quad (11)$$

realizing $g_\theta : (\mathbf{P}, \mathcal{C}, \mathcal{V}) \mapsto \mathbf{s}_\theta$ as used in the window pruning formulation.

VLM Inference on Pruned Context. Finally, the selected temporal patches $\tilde{\mathbf{P}}_\theta$ are serialized together with the original textual context $\mathcal{C}$ and visual plot token $\mathcal{V}$ into a multimodal prompt, which is fed into the frozen VLM backbone to generate the final response. By exposing only a small number of context-conditioned windows to the VLM, TimePrune reduces the effective context length while maintaining task performance.

3.3 Narrative Calibrated Binned Grounding

To avoid fragmented selections that break temporal continuity during pruning, we train the network with narrative-derived pseudo labels using a differentiable surrogate objective that enforces the token budget and temporal coherence.

Surrogate Optimization. Because the pruning decision involves a discrete top- k and the backbone Φ is kept frozen, we optimize a weakly supervised grounding objective rather than Eq. (2). We construct patch-level pseudo labels $\mathbf{y}_{gt} \in \{0, 1\}^N$ from narrative timestamps, and define:

$$\mathbf{s}_\theta = g_\theta(\mathbf{P}, \mathcal{C}, \mathcal{V}) \in \mathbb{R}^N \quad (12)$$

as patch-level logits, with selection probabilities.

$$\mathbf{p}_\theta = \sigma(\mathbf{s}_\theta) \in (0, 1)^N \quad (13)$$

At inference time, we prune $\mathbf{P}$ by selecting the top- k patches according to $\mathbf{p}_\theta$. During training, we optimize continuous $\mathbf{p}_\theta$ and never apply hard top- k ; top- k is used at inference, while sparsity is encouraged through $\mathcal{R}_{\text{sparse}}$ as a differentiable proxy to the budget.

$$\min_{\theta} \mathcal{L}_{\text{total}}(\mathbf{p}_\theta, \mathbf{y}_{\text{gt}}). \quad (14)$$

To encourage the gating probabilities to respect the token budget, we enforce an expected-budget constraint on $\mathbf{p}_\theta$. Let $\rho = K/N$ denote the target keep ratio. We define:

$$\mathcal{R}_{\text{sparse}}(\mathbf{p}_\theta) = \left(\frac{1}{N} \sum_{i=1}^N p_{\theta,i} - \rho \right)^2 \quad (15)$$

We obtain $\mathbf{y}_{\text{gt}} \in \{0, 1\}^N$ from the narrative by extracting temporal expressions in $\mathcal{C}$ (e.g., timestamps or time ranges) and mapping them to discretized patches, setting $y_{\text{gt},i} = 1$ if the i -th patch overlaps a mentioned interval.

3.3.1 Coherence Regularized Objective.

Let $\mathbf{s}_\theta = g_\theta(\mathbf{P}, \mathcal{C}, \mathcal{V})$ denote the patch-level logits and $\mathbf{p}_\theta = \sigma(\mathbf{s}_\theta)$ the corresponding probabilities. To obtain temporally coherent and sparse selections, we minimize a compound loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bce}} + \lambda_1 \mathcal{L}_{\text{iou}} + \lambda_2 \mathcal{L}_{\text{tv}} + \lambda_3 \mathcal{R}_{\text{sparse}}, \quad (16)$$

Coherence Regularization. We use a binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ to encourages patch-level agreement with pseudo event interval. This follows the weakly supervised alignment established in temporal localization (Gao et al., 2017). Although BCE is a point-wise objective, it provides weak supervision for trimodal grounding by encouraging the selector to activate temporal patches that are consistent with high-level linguistic cues (e.g., “flash flood”), aligning with weakly supervised temporal localization where supervision comes from narrative context rather than dense annotations.

$\mathcal{L}_{\text{iou}}$ is the Soft-IoU loss (Rahman and Wang, 2016), which promotes span-level completeness by maximizing the overlap between the predicted soft mask and the pseudo event interval. We represent the pseudo interval as a binary mask aligned with temporal patches. In contrast to point-wise supervision, Soft-IoU optimizes global coverage, encouraging the selector to preserve contiguous event windows rather than isolated high-score patches.

Budget-Constrained Grounding. To further mitigate fragmented selections, $\mathcal{L}_{\text{tv}}$ applies Total Variation (TV) regularization (Mahendran and Vedaldi, 2015) to penalize rapid score changes between neighboring temporal patches. While Soft-IoU enforces global span coverage, TV imposes local smoothness, and together they reduce spikes and produce stable, contiguous pruning masks.

Finally, $\mathcal{R}_{\text{sparse}}$ penalizes overly dense mask activations to explicitly enforce the target sparsity budget in Eq. (2), pushing the selector to allocate its limited budget to the most informative temporal patches. During training, we freeze VLM backbone and update the trimodal budgeted network.

4 Experiments

In this section, we verify our proposed method on different datasets and VLMs. We first introduce our experimental setup in subsection 4.1 and then report the main results in subsection 4.2. Analysis experiments are presented in subsection 4.3.

4.1 Experimental Setup

Models and Evaluation Benchmark. We employ the Qwen3-VL series (Team, 2025b) and InternVL3.5 series (Wang et al., 2025b) as the backbone. We conduct our experiments on MT-Bench (Chen et al., 2025), a multimodal time series benchmark designed to assess cross-modal reasoning in LLMs. MTBench couples high frequency numerical series with aligned textual narratives in both *Finance* and *Weather* domains. The details of dataset can be found in Appendix C.

Baselines. To verify the effectiveness of our trimodal-guided approach, we compare TimePrune against three pruning strategies. All baselines operate under the same sparsity budget. **FastV** is a vision-based method and we adapt it by treating each temporal patch as a token and pruning it (Chen et al., 2024). **DivPrune** is a diversity-based pruning method that reformulates token selection as a Max-Min Diversity Problem and we apply this diversity maximization rule to temporal patch tokens (Alvar et al., 2025). Finally, **Random** samples windows stochastically from the full sequence, serving as a lower bound.

Metrics. For the financial domain, we use Mean Absolute Error (**MAE**) and Mean Absolute Percentage Error (**MAPE**) to evaluate regression performance. The news-driven tasks include correlation prediction and multiple-choice QA, both evaluated

Method			Token Budget	Forecasting (↓)				Trend Pred. (↑)		Next-day Prediction (↓)						MCQA (↑)	
				7-Day		14-Day		Past	Future	Max		Min		Diff		7-Day	14-Day
				MSE	MAE	MSE	MAE			MSE	MAE	MSE	MAE	MSE	MAE		
Qwen3VL-8B	Baseline	100%	29.04	4.29	25.50	3.93	<u>46.30</u>	25.00	29.53	4.11	23.41	3.61	50.01	5.72	<u>58.33</u>	<u>60.00</u>	
	Random	~30%	25.52	3.91	22.00	3.56	42.85	39.05	23.29	3.57	29.33	3.70	<u>24.36</u>	<u>3.86</u>	57.14	56.66	
	FastV	~30%	24.84	3.91	<u>19.36</u>	3.38	41.17	39.86	24.68	3.54	21.90	3.30	25.14	3.84	57.85	58.33	
	DivPrune	~30%	<u>23.32</u>	<u>3.63</u>	19.54	<u>3.35</u>	44.53	39.56	<u>21.82</u>	<u>3.51</u>	<u>21.88</u>	<u>3.33</u>	24.77	3.74	55.00	57.50	
	TimePrune	~30%	16.64	3.08	19.09	3.32	47.05	40.17	16.52	3.05	11.99	2.66	20.94	3.32	60.00	62.50	
InternVL3.5-8B	Baseline	100%	<u>25.28</u>	<u>4.04</u>	24.62	3.88	64.06	35.15	30.42	4.15	17.39	3.17	35.06	4.59	48.34	43.54	
	Random	~30%	32.53	4.47	25.58	3.90	57.50	34.16	27.11	4.04	15.04	<u>3.00</u>	<u>22.76</u>	3.50	64.28	64.28	
	FastV	~30%	37.03	4.64	22.97	3.68	56.66	35.83	<u>23.13</u>	<u>3.64</u>	19.05	3.10	22.82	<u>3.42</u>	62.85	<u>66.40</u>	
	DivPrune	~30%	31.53	4.22	<u>21.32</u>	<u>3.53</u>	59.16	<u>35.85</u>	27.28	4.00	<u>14.94</u>	3.01	25.18	3.66	<u>65.00</u>	62.25	
	TimePrune	~30%	22.81	3.57	20.82	3.52	<u>60.00</u>	37.50	21.24	3.19	13.23	2.55	19.90	2.78	67.14	66.42	
Qwen3VL-4B	Baseline	100%	23.98	3.77	26.19	3.94	52.37	26.66	36.81	4.66	18.86	3.23	57.66	6.08	<u>54.57</u>	<u>56.59</u>	
	Random	~30%	25.74	3.92	22.04	3.57	<u>55.69</u>	45.83	24.51	3.68	<u>16.24</u>	3.05	35.30	4.50	52.09	53.52	
	FastV	~30%	28.02	3.87	<u>19.41</u>	<u>3.36</u>	54.36	54.67	<u>19.98</u>	<u>3.26</u>	19.39	<u>3.03</u>	<u>20.51</u>	<u>3.39</u>	50.69	52.64	
	DivPrune	~30%	<u>24.16</u>	<u>3.68</u>	19.64	3.37	55.43	<u>55.94</u>	28.50	3.91	17.10	3.11	41.92	4.94	50.54	54.41	
	TimePrune	~30%	15.91	2.98	19.19	3.34	55.74	56.20	18.71	3.28	15.08	2.80	17.14	3.20	55.65	59.11	
InternVL3.5-4B	Baseline	100%	<u>23.42</u>	<u>3.76</u>	<u>22.67</u>	3.73	46.03	24.21	26.47	3.68	19.70	3.24	<u>21.08</u>	<u>3.33</u>	<u>53.49</u>	<u>55.00</u>	
	Random	~30%	32.98	4.51	27.18	4.00	42.21	48.33	34.79	4.56	21.96	3.55	36.29	4.17	22.14	54.28	
	FastV	~30%	24.16	3.86	25.57	3.89	43.33	<u>50.35</u>	<u>23.92</u>	<u>3.62</u>	20.64	<u>3.20</u>	25.80	3.54	52.85	52.14	
	DivPrune	~30%	31.63	4.23	25.13	<u>3.62</u>	44.16	49.16	29.66	4.10	<u>17.78</u>	3.21	35.47	4.23	51.42	52.85	
	TimePrune	~30%	21.99	3.72	21.32	3.53	<u>45.83</u>	52.50	22.71	3.51	14.37	3.14	20.73	2.67	55.71	58.57	

Table 1: Overall performance of multimodal time series analysis in the **Weather** domain. $\downarrow$ indicates that lower is better, while $\uparrow$ indicates that higher is better. The best result in each block is **bolded**, and the second-best is underlined. In Appendix, results of Finance domain in Table 7; results of closed-source models are in Tables 9–10.

Method	Prefill	TPS.	Spd.	Performance (avg.)			
	s	token/s	$\times$	Forecasting	Trend	Next-day	MCQA
Full Context	12.71	107.81	1.0 $\times$	27.27	35.65	34.31	59.16
DivPrune	4.76	121.18	1.88 $\times$	21.43	42.04	22.84	56.25
FastV	4.65	130.91	2.12 $\times$	19.13	40.51	21.42	58.09
TimePrune	4.12	136.55	2.59$\times$	17.86	43.61	16.48	61.25
vs. FastV	-11.39%	+4.30%	+22.16%	-6.58%	+7.65%	-23.06%	+6.00%

Table 2: Overall efficiency comparison. **Prefill** (s) is time to first token, **TPS** is throughput (tokens/s), and **Spd.** is speedup vs. Full Context. **Performance (avg.)** reports average scores on four representative tasks. In vs. *FastV*, **green** marks increases in higher-is-better metrics, while **blue** marks decreases in lower-is-better metrics.

with accuracy. For the weather data, we adopt Mean Squared Error (MSE) and MAE, along with next-day maximum temperature, minimum temperature, and their difference as regression targets. Weather trend prediction and multiple-choice QA are also evaluated using accuracy.

Implementation Details. We process each multimodal instance by discretizing the time series window into $N = 24$ non-overlapping bins. For each bin, we extract a 7-dimensional feature vector containing statistical metrics (mean, std, min, max, slope, count) and a relative positional feature (i/N).

Further details of the training implementation and hyperparameters can be found in Appendix B.

4.2 Main Results

4.2.1 Weather Domain

Experimental results in Table 1 show that TimePrune consistently enhances the performance across various tasks under the same token budget. Across all models, TimePrune achieves the best or second-best results in most metrics, with particularly strong gains on forecasting and fine-grained next-day prediction. For Qwen3VL-8B, TimePrune improves the 7-day MSE by 28.64% relative to DivPrune, and further achieves a 24.28% relative reduction in the next-day MAX MSE. On InternVL3.5-8B, TimePrune remains competitive across all tasks and achieves the highest overall MCQA accuracy in the short- and long-term, scoring 67.14 and 66.42, respectively. Notably, TimePrune substantially improves Qwen3VL-4B’s performance. Relative to DivPrune, it yields a 34.1% improvement in 7-day forecasting MSE, an 11.8% improvement in next-day minimum-value MSE, and a 10.1% increase in 7-day MCQA accuracy. TimePrune also generalizes well to In-

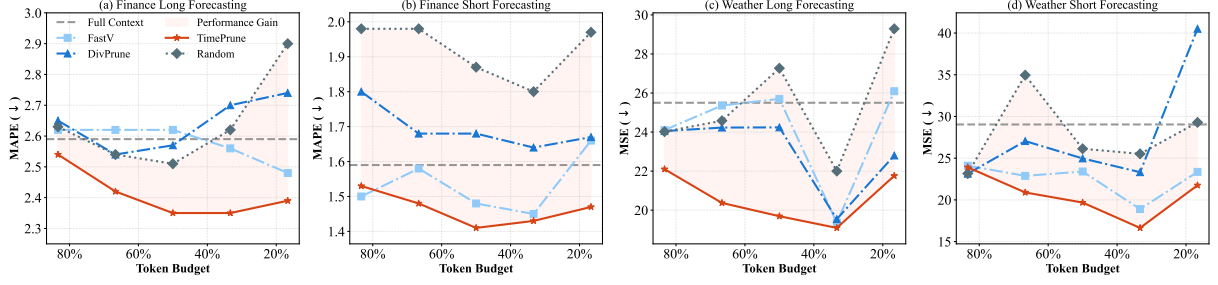


Figure 3: Performance sensitivity under token-budget pruning on finance and weather. Token budget decreases from left to right; (a–b) report finance long/short forecasting (MAPE↓) and (c–d) report weather long/short forecasting (MSE↓). Shaded regions indicate the performance gain of TimePrune over Random under the same budget.

ternVL3.5 family, delivering strong performance on both 8B and 4B models. In contrast, the gains on trend prediction are relatively small than those on regression tasks, likely because trend classification depends more on a few observations and is therefore less sensitive to moderate token budget reductions. Despite these differences, TimePrune demonstrates consistent benefits across model scales and architectures. Further experimental results in the finance domain can be found in Appendix Table 7. **Efficiency Analysis on Weather Tasks.** In addition to task performance, we follow (Dong et al., 2025) to evaluate the overall inference efficiency. As shown in Table 2, we report the Time-to-First-Token (Prefill), Throughput (TPS), and end-to-end speedup under the same fixed token budget with Qwen3VL-8B model and evaluation settings as our main results. Under a fixed budget, TimePrune achieves the superior efficiency profile, recording a 4.12s prefill time and a $2.59\times$ speedup over the Full Context baseline. It also outperforms FastV, reducing prefill latency by 11.39% and increasing throughput by 4.30%. Crucially, these gains enhance rather than compromise quality. TimePrune achieves the lowest error rates on Next-day and Forecasting tasks (16.48 and 17.86, respectively).

4.3 Model Analysis

4.3.1 Ablation Study

We evaluate each modality in the trimodal gating module using Qwen3VL-8B in Table 3. Accuracy is averaged over short and long QA, and MAE is averaged over short and long forecasting. Each variant retrain the gating network under the same objective while removing a specific conditioning signal: *Time only* discards both text and visual inputs; *w/o Text Guidance* conditions the gate only on the visual trend plot; and *w/o Visual Guidance* keeps text guidance but drops the visual token. All

Modality Variant	Finance		Weather	
	Acc. (↑)	MAE (↓)	Acc. (↑)	MSE (↓)
Time-only	83.07 (-8.91)	2.12 (-0.21)	60.83 (-0.42)	22.24 (-4.38)
w/o Text Guidance	86.92 (-5.06)	2.23 (-0.32)	56.68 (-4.57)	22.00 (-4.14)
w/o Visual Guidance	87.61 (-4.37)	2.13 (-0.22)	59.16 (-2.09)	21.52 (-3.66)
TimePrune (Ours)	91.98	1.91	61.25	17.86

Table 3: Ablation on Modality. We report the **Accuracy** (%, ↑) for MCQA tasks and **MSE** (↓) and **MAE** (↓) for forecasting tasks. Values in parentheses denote the **performance drop** compared to our TimePrune method. “w/o” denotes the removal of a specific component. The best results are highlighted in **bold**.

pruning variants use the same token budget $\sim 30\%$.

Compared with all ablated variants, TimePrune achieves the best overall results, reaching the highest QA accuracy and the lowest forecasting error. Relative to Time-only, conditioning the gate on a single modality yields task-dependent effects: it improves Finance QA but consistently increases forecasting error, and for Weather QA, using only text or only visual guidance reduces accuracy rather than improving it. Among dual-modality variants, text guidance is more informative than visual guidance for QA and forecasting, yet removing either modality still degrades performance, suggesting that text and vision provide complementary cues.

4.3.2 Sensitivity to Pruning Ratio

In this section, we show the performance of baselines and our method across various token budgets for the Qwen3VL-8B model. The results are shown in Figure 3 where the y-axis represents MAPE and MSE. We compare TimePrune against FastV, DivPrune, and Random, and include a dashed gray line as the full context baseline. Across all token budgets, TimePrune consistently yields the lowest error rates, demonstrating strong robustness to aggressive context pruning. Performance notably peaks at around the 33% budget level. In

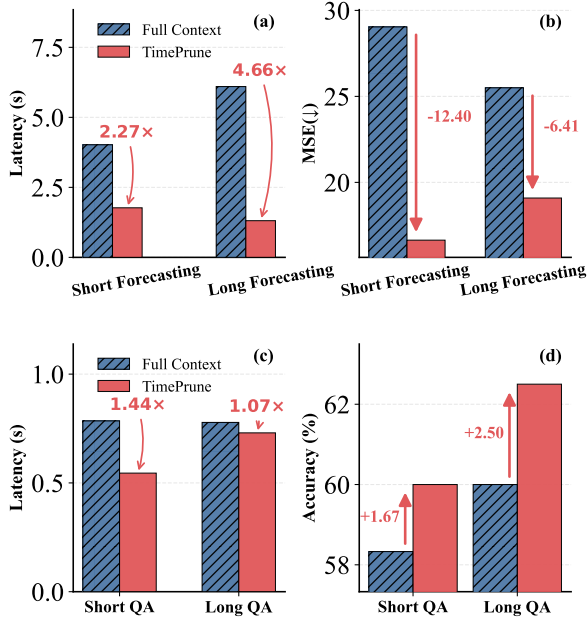


Figure 4: Efficiency and effectiveness analysis on Weather using Qwen3VL-8B. Panels (a) and (c) display the inference latency with **relative speedup**, while panels (b) and (d) present the task performance.

the Finance domain, TimePrune reaches a minimum MAPE of 2.35 for Long forecasting and a near-optimal 1.43 for Short forecasting. The effectiveness is even more pronounced in the Weather domain, where the model attains its lowest MSE values of around 19 and 16 for Long and Short tasks, respectively. Furthermore, the shaded regions indicate that the performance gap over the Random baseline widens as the budget tightens. Although errors increase slightly at the extreme 17% budget, TimePrune maintains a consistent lead over competing pruning strategies.

4.3.3 Efficiency and Effectiveness Analysis

The experiments discussed above demonstrate that our method can substantially reduce the number of time tokens. Here, we examine the inference efficiency in real-world scenarios. We report the total running time on the Weather dataset in Figure 4. In the Weather Forecasting task, TimePrune yields substantial time savings without sacrificing quality. For instance, it reduces inference latency by 2.27 $\times$ for short forecasting and 4.66 $\times$ for long forecasting relative to the Full Context baseline, while simultaneously lowering the MSE by 12.40 and 6.41, respectively. Furthermore, in the short QA and long QA tasks, our method accelerates inference by 1.44 $\times$ and 1.07 $\times$, respectively. Together, these results indicate that pruning redundant

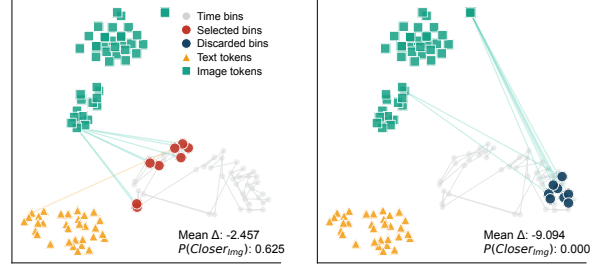


Figure 5: t-SNE visualization of latent representations for time bins and context tokens. **Left**: highlighted selected time bins; **Right**: highlighted discarded time bins. For each highlighted bin, a line connects it to the most-attended token. Insets report the mean distance gap $\Delta = d_{\text{text}} - d_{\text{img}}$ and proportion $P(d_{\text{img}} < d_{\text{text}})$, where d_{text} and d_{img} are nearest-neighbor distances to text and image tokens.

time tokens can concurrently accelerate inference and enhance forecasting quality, particularly as the forecasting horizon extends. Additional results in the Finance domain can be found in Appendix D.2.

4.4 Visualizing Trimodal Alignment

The t-SNE method (Maaten and Hinton, 2008) projects the learned embeddings of time bins (gray) and context tokens into 2D (Figure 5). The left panel highlights *selected* time bins (red), while the right panel highlights *discarded* bins (blue); each highlighted bin is connected to its most-attended context token. Selected bins exhibit stronger visual association, with a higher geometry-based image proximity $P(\text{Closer}_{img}) = 0.625$ and a less negative mean gap $\Delta = d_{\text{text}} - d_{\text{img}} = -2.457$. In contrast, discarded bins show $P(\text{Closer}_{img}) = 0.000$ and a much more negative $\Delta = -9.094$, indicating they are consistently closer to text than to image tokens in the embedding geometry.

5 Conclusion

In this work, we introduce TimePrune, a lightweight time token pruning framework that accelerates VLM inference on multimodal time series tasks by selecting task-relevant temporal segments. Across multiple benchmarks, subtasks, and VLM backbones, TimePrune delivers a stronger efficiency-quality than prior pruning methods and unimodal baselines, substantially reducing end-to-end latency for QA and forecasting while maintaining or improving accuracy and remaining robust under varying token budgets. Moreover, TimePrune is a plug-in pruning module that can be attached to VLMs without modifying the backbone.

Limitations

Although TimePrune keeps the VLM backbone frozen, it still requires training a lightweight selector using narrative-derived pseudo labels. The current supervision process depends on temporal expressions or event descriptions extracted from narratives. Therefore, the framework does not directly support fully narrative-free settings where raw time series data are available. In addition, when narratives are ambiguous, incomplete, or noisy, the extracted event intervals may be inaccurate. These intervals may also fail to perfectly correspond to the temporal segments that are most useful for downstream prediction or reasoning, which can reduce the quality of the selected patches.

The current evaluation is limited to the finance and weather domains. Although these domains contain different temporal patterns and reasoning tasks, broader validation is still needed before claiming general applicability across time series domains. In particular, the effectiveness of TimePrune on medical signals, industrial sensors, traffic data, and other application settings remains unexplored. These domains may exhibit different temporal structures, event frequencies, and narrative styles, which could require changes to pseudo-label construction.

The transferability of a learned selector also requires further investigation. A selector trained in one domain may not directly generalize to unseen domains, and transfer across different VLM backbones may require additional adaptation. Potential failure cases may occur when relevant evidence is weakly expressed in the narrative or located outside the extracted event interval.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China (No. 62402531, No. 62225209).

References

- Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9392–9401. Computer Vision Foundation / IEEE.
- Md. Adnan Arefeen, Biplob Debnath, Md. Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. Vita: An efficient video-to-text algorithm using VLM for rag-based video analysis system. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 2266–2274. IEEE.
- Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassioulas, Yifeng Gao, and Rex Ying. 2025. [Mtbench: A multimodal time series benchmark for temporal reasoning and question answering](#). *CoRR*, abs/2503.16858.
- Junyi Chen, Qian Liu, Jing Sun, and Yi Zhang. 2026. Truelens: video fake news detection with dual level evidence gathering and consolidation. In *Proceedings of the ACM Web Conference 2026*, pages 7178–7188.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pages 19–35. Springer.
- Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. 2025. [Mmtok: Multimodal coverage maximization for efficient inference of vlms](#). *CoRR*, abs/2508.18264.
- Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. 2023. Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 7522–7529. AAAI Press.
- Wei Fan, Shun Zheng, Xiaohan Yi, Wei Cao, Yanjie Fu, Jiang Bian, and Tie-Yan Liu. 2022. DEPTS: deep expansion learning for periodic time series forecasting. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Jingru Fei, Kun Yi, Alex Xing Wang, Qingsong Wen, Xiangxiang Zhu, and Wei Fan. 2026. Olivia: Harmonizing time series foundation models with power spectral density. *CoRR*, abs/2605.17340.
- Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025. Citygpt: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 591–602. ACM.

- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.
- Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Suhan Guo, Jiahong Deng, Mengjun Yi, Furao Shen, and Jian Zhao. 2025. SPAT: sensitivity-based multihead-attention pruning on time series forecasting models. *CoRR*, abs/2505.08768.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. 2025. Multi-modal time series analysis: A tutorial and survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 6043–6053. ACM.
- 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. 2024a. Time-llm: Time series forecasting by re-programming large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024b. Position: What can large language models tell us about time series analysis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 22260–22276. PMLR / OpenReview.net.
- Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. 2025. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 18082–18090. AAAI Press.
- Jun Li, Che Liu, Sibao Cheng, Rossella Arcucci, and Shenda Hong. 2023. Frozen language model helps ECG zero-shot learning. In *Medical Imaging with Deep Learning, MIDL 2023, 10-12 July 2023, Nashville, TN, USA*, volume 227 of *Proceedings of Machine Learning Research*, pages 402–415. PMLR.
- Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE.
- Chenghao Liu, Qian Liu, Ziqin Zhu, Hao Fei, and Aniket Mahanti. 2025a. David vs. goliath: Cost-efficient financial QA via cascaded multi-agent reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.
- Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. 2025b. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 18780–18788. AAAI Press.
- Chenxi Liu, Shaowen Zhou, Qianxiong Xu, Hao Miao, Cheng Long, Ziyue Li, and Rui Zhao. 2025c. Towards cross-modality modeling for time series analytics: A survey in the LLM era. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 10564–10572. ijcai.org.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2025d. Large language models can deliver accurate and interpretable time series anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 4623–4634. ACM.
- Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. 2024. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 4095–4106. ACM.

- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5188–5196. IEEE Computer Society.
- Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024. Unraveling the truth: Do vlms really understand charts? A deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 16696–16717. Association for Computational Linguistics.
- Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben-Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. 2025. Docvlm: Make your VLM an efficient reader. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 29005–29015. Computer Vision Foundation / IEEE.
- Md. Atiqur Rahman and Yang Wang. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In *Advances in Visual Computing - 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings, Part I*, volume 10072 of *Lecture Notes in Computer Science*, pages 234–244. Springer.
- ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. 2025. Exploring multi-modal data with tool-augmented LLM agents for precise causal discovery. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 636–660. Association for Computational Linguistics.
- Yuchen Su, Yonghua Zhu, Ruofan Wang, Zijian Huang, Diana Benavides-Prado, and Michael J Witbrock. 2025. A survey of pun generation: Datasets, evaluations and methodologies. *EMNLP (Findings)*, pages 7375–7395.
- Gemma Team. 2025a. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Qwen Team. 2025b. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2025a. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 12694–12702. AAAI Press.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internv13.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Su-chow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. Fincon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Lifan Zhao, Yanyan Shen, Zhaoyang Liu, Xue Wang, and Jiaji Deng. 2025. Less is more: Unlocking specialization of time series foundation models via structured pruning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.
- Lecheng Zheng, Zhengzhang Chen, Jingrui He, and Haifeng Chen. 2024. MULAN: multi-modal causal structure learning and root cause analysis for microservice systems. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 4107–4116. ACM.
- 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. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.
- Zihao Zhou and Rose Yu. 2025. Can llms understand time series anomalies? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

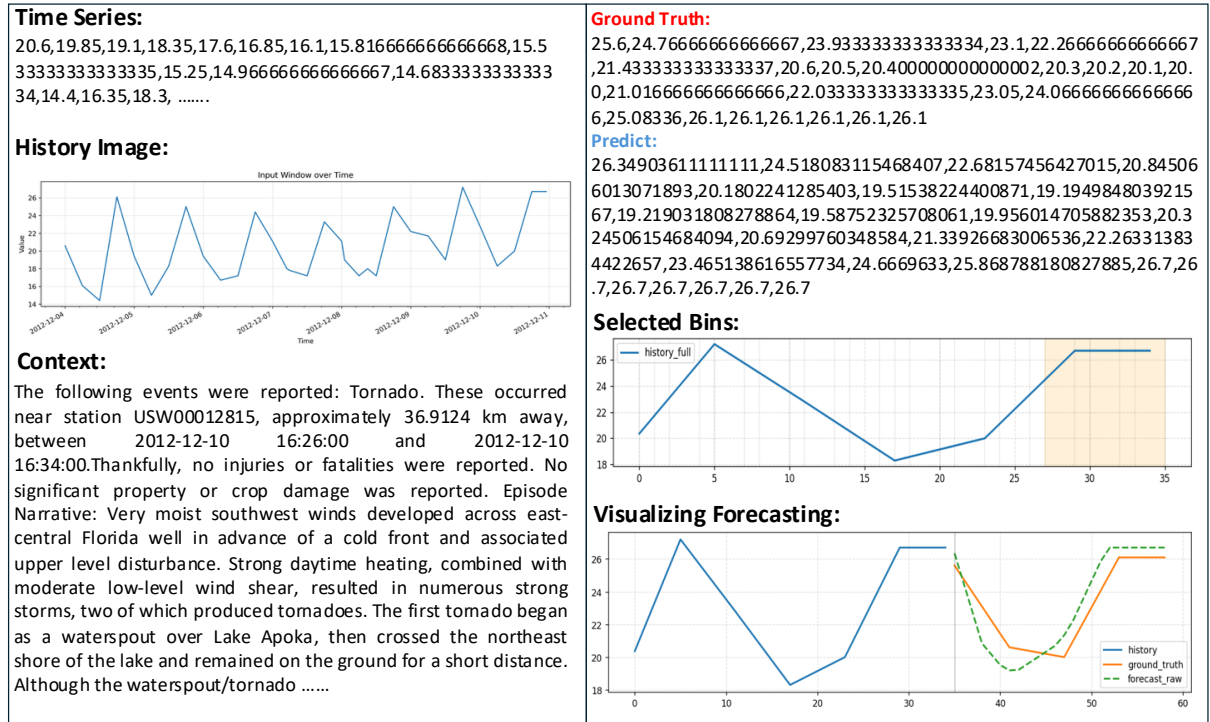


Figure 6: Weather Domain Forecasting Case Study

A Case Study

Figure 6 presents a forecasting case study with both the model inputs and outputs. On the **left**, we show the raw time series input window together with its plotted trajectory, as well as the associated event context text that provides external information potentially relevant to the target period. On the **right**, we report the ground-truth future values and our model’s predicted sequence, showing that the forecast closely follows the target trend. We also visualize the model’s selected time bins, where the shaded region indicates the subset of historical segments retained by our gating mechanism to form a compact yet informative representation. Finally, the forecasting visualization overlays the historical window with both the ground truth continuation and our prediction.

B Training Details

Table 6 summarizes the main hyperparameters used in our experiments, including the architecture configuration, training setup, and the weights assigned to each loss component. These settings are chosen to provide stable optimization while encouraging smooth and sparse bin selection. Moreover, the Time Encoder maps the 7-dimensional bin features to a latent space with $d = 128$ and applies a 2-layer Transformer encoder with 4 at-

tention heads and learnable positional embeddings. For context, a frozen all-MiniLM-L6-v2 encoder produces sentence-level text embeddings and a frozen clip-vit-base-patch16 encoder extracts a global visual token from the time series plot. All experiments are run on a single RTX 5090 GPU with 32GB of memory.

C Dataset Specifications

Table 5 reports dataset specifications for Finance and Weather across four subtasks each, including prediction targets, sample counts, horizons, time resolution, and input modalities. Both domains use hourly multimodal inputs combining time series, text, and plot images, while the subtasks span forecasting, classification-based trend and correlation prediction, and MCQA. The table highlights how dataset sizes and horizons vary across subtasks and domains under a consistent input format.

Finance Domain. Each finance sample consists of an equity price time series, where every hourly observation is paired with its timestamp, together with associated news articles.

Forecasting. The short forecasting setting takes the past 7 days of hourly prices paired with 168 timestamped steps as input and predicts the next 1 day of hourly prices while long forecasting instead

Aspect	Weather	Finance
Train / Val / Test	1567 / 196 / 196	600 / 75 / 75
Selector training	One selector trained independently for this domain	One selector trained independently for this domain
Shared across tasks	Yes, within the same domain	Yes, within the same domain
Shared across VLM backbones	Yes, within the same domain	Yes, within the same domain
Trainable components	Lightweight selector only	Lightweight selector only
Frozen components	VLM backbone and pretrained text/vision encoders	VLM backbone and pretrained text/vision encoders
Epochs	5	5
Batch size	8	8
Optimizer	AdamW, learning rate 5×10^{-4}	AdamW, learning rate 5×10^{-4}
Model selection	Best checkpoint based on validation loss	Best checkpoint based on validation loss

Table 4: Selector training configuration.

uses the past 30 days of hourly prices to forecast the next 7 days of hourly prices.

Stock Trend Prediction. Using the same input–output windows as forecasting, MTBench maps the realized return over the prediction horizon to discrete trend labels. The scheme is based on percentage change: a three–way setting with *negative*, *neutral*, and *positive* trends, and a five–way setting with *strong decrease*, *mild decrease*, *near–flat*, *mild increase*, and *strong increase*.

News–stock Correlation. Given the multimodal context, the model is asked to predict the trend label. In the short forecasting task, the label is computed from the cumulative return over the next 1 day; in the long forecasting task, from the cumulative return over the next 7 days. Performance is reported as classification accuracy under both the three–way and five–way label schemes.

News–driven Question Answering. News–driven QA evaluates cross–modal reasoning between the textual narrative and the time series. In the multiple–choice QA task, the model selects the correct statement about price dynamics and events given the same 7–day or 30–day history and its corresponding news context.

Weather Domain. In the weather domain, the *Forecasting* and *MCQA* tasks follow the same short– and long–horizon settings as in finance, but operate on hourly temperature series paired with narrative

descriptions of meteorological events. *Trend Prediction* focuses on the directional behavior of future temperature: given the same input–output windows as in forecasting, MTBench converts the future temperature profile into discrete trend labels that indicate whether the overall evolution is cooling, roughly stable, or warming, with both coarse three–way and finer five–way categorizations. *Next–day Prediction* involves regressing the maximum temperature, minimum temperature, and temperature difference for the next day, given past temperature observations and their textual and visual contexts.

D Experiment

D.1 Additional Main Result

D.1.1 Finance Domain

As shown in Table 7, TimePrune consistently delivers superior performance across all subtasks under a constrained token budget. In particular, building upon the strong full–context performance of Qwen3VL–8B, TimePrune achieves the best results for all metrics within the block, while using substantially fewer tokens. For forecasting, TimePrune achieves an additional 12.8% MAE relative reduction in short forecasting and 13.6% MAPE relative reduction in long forecasting compared to DivPrune, while also improving trend prediction and news–stock correlation accuracies. A similar trend is observed on InternVL3.5–8B as well as smaller backbones (Qwen3VL–4B and InternVL3.5–4B), demonstrating that our trimodal pruning method can robustly enhance multi–task finance performance under strict token budgets.

Table 9 and Table 10 summarize additional evaluations under token–reduction across two domains. They compare proprietary LLMs (GPT–4o, Gemini, Claude) with an open–source VLM backbone (Qwen3VL–2B) equipped with different pruning strategies at comparable token budgets. Table 9 reports Finance results on multi–horizon forecasting, stock trend prediction, news–stock correlation, and MCQA, while Table 10 reports Weather results on four distinct tasks: forecasting, trend prediction, next–day prediction, and MCQA.

Across both domains, the main trend is consistent: pruning baselines (FastV, DivPrune, Random) often degrade performance, whereas TimePrune achieves the best overall results among open–source pruning methods, remaining competitive with or improving upon the baseline on several metrics. Overall, these tables support the conclusion that

Domain	Subtask	Prediction Target	Num.	Horizon	Time Res.	Modalities
Finance	Forecasting	MAE, MAPE	750, 525	7-day, 30-day	hourly	time series + text + plot image
	Stock Trend Prediction	3-way, 5-way classification	750, 525	7-day, 30-day	hourly	time series + text + plot image
	News-Stock Correlation	3-way, 5-way classification	750, 525	7-day, 30-day	hourly	time series + news text + plot image
	MCQA	Accuracy	491, 516	7-day, 30-day	hourly	time series + text + plot image
Weather	Forecasting	MSE, MAE	1959, 1959	7-day, 14-day	hourly	time series + event text + plot image
	Trend Prediction	Past, Future classification	1959	14-day	hourly	time series + text + plot image
	Next-day Prediction	Max, Min, Diff regression	1959	7-day	hourly	time series + text + plot image
	MCQA	Accuracy	645, 728	7-day, 14-day	hourly	time series + text + plot image

Table 5: Dataset specifications by domain and subtask. Each row corresponds to a subtask evaluated under the same multimodal input format.

Hyperparameters
Architecture
Input bins N : 24
Input feature dim d_{in} : 7
Model hidden dim d_{model} : 128
Transformer heads n_{heads} : 4
Transformer layers n_{layers} : 2
Image size: 224
Training
Epochs: 10
Batch size: 8
Optimizer: AdamW
Learning rate: 5×10^{-4}
Weight decay: 1×10^{-4}
Loss
BCE: BCELoss
IoU weight: 0.5
TV weight: 0.01
Sparsity weight: 0.001

Table 6: Parameters used in our experiments.

time-aware pruning preserves task-relevant temporal evidence more effectively.

D.2 Latency Test on Finance Domain

In Figure 7, it reduces end to end latency by 1.75 times for short-horizon forecasting and 1.34 times for long-horizon forecasting, compared to Full Context. Meanwhile, it consistently lowers forecasting error, with absolute MAE reductions of 0.09 for short-horizon and 0.59 for long-horizon settings. For finance QA, TimePrune yields an even larger latency gain: it speeds up short QA by 8.62 $\times$ and long QA by 1.57 $\times$. Despite the substantial acceleration, accuracy is not compromised and instead improves by +0.94 and +1.60 points for short and long QA, respectively.

D.3 Comparison with the Training-Free Narrative-Span Baseline

To examine whether the Trimodal Budgeted Network is necessary, we construct a simple training-free alternative based directly on the pseudo la-

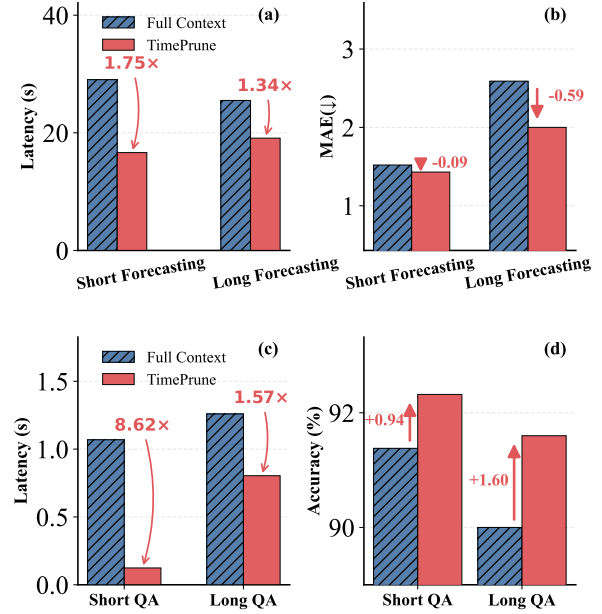


Figure 7: Efficiency and effectiveness analysis on Finance using Qwen3VL-8B.

bels y_{gt} introduced in Sec. 3.3. These labels already identify narrative-aligned temporal regions in the time series, raising the question of whether a learned selection mechanism is required. We therefore define a **Narrative-Span** baseline that bypasses both the Trimodal Budgeted Network and the coherence-regularized training objective. It directly retains time-series bins whose timestamps overlap with the extracted narrative intervals, without any additional learning or optimization. Unlike our method, Narrative-Span does not enforce a fixed retention budget. Instead, the number of retained bins is fully determined by the coverage of the narrative spans in each sample, which can vary substantially across inputs and therefore does not correspond to a constant ratio such as $\sim 30\%$.

As shown in Table 8, Narrative-Span exhibits inconsistent performance across tasks and model backbones. In several settings, it even underper-

Method			Token Budget		Forecasting (↓)				Stock Trend Prediction (↑)				News-stock Correlation (↑)				MCQA (↑)	
					7-Day		30-Day		7-Day		30-Day		7-Day		30-Day		7-Day	30-Day
			MAE	MAPE	MAE	MAPE	3-way	5-way	3-way	5-way	3-way	5-way	3-way	5-way	Acc.	Acc.		
Qwen3VL-8B	Baseline	100%	1.52	2.06	2.59	4.64	28.34	23.42	26.45	20.83	60.00	34.78	46.67	25.00	91.38	90.00		
	Random	~30%	1.80	2.52	2.67	5.30	46.73	39.63	52.19	36.57	62.98	34.47	55.00	32.80	92.25	89.80		
	FastV	~30%	1.45	2.09	2.51	4.95	48.91	41.45	51.80	37.50	55.60	33.20	55.40	32.60	91.89	90.80		
	DivPrune	~30%	1.64	2.31	2.58	4.55	48.13	40.67	52.38	37.33	61.91	34.26	57.60	33.60	92.04	90.20		
	TimePrune	~30%	1.43	2.00	2.39	3.93	50.21	42.64	53.89	38.09	63.34	35.32	63.19	34.68	92.32	91.60		
InternVL3.5-8B	Baseline	100%	2.60	5.96	4.66	9.46	23.43	17.18	41.40	20.31	56.35	29.27	45.24	26.19	56.45	54.03		
	Random	~30%	1.47	2.81	2.65	5.77	33.43	25.81	50.00	28.78	57.72	26.83	46.77	26.61	88.65	89.53		
	FastV	~30%	1.40	2.27	2.31	4.28	24.09	19.63	55.71	25.71	58.06	29.84	48.80	27.20	88.04	87.05		
	DivPrune	~30%	1.35	2.33	2.48	4.71	30.12	22.04	48.61	23.61	57.26	29.03	48.39	29.84	87.30	88.21		
	TimePrune	~30%	1.26	2.16	2.03	3.71	35.30	28.91	56.26	29.40	58.54	30.92	61.54	31.73	89.58	91.40		
Qwen3VL-4B	Baseline	100%	2.09	2.84	2.81	5.37	26.66	20.83	45.83	24.16	55.00	33.33	45.00	22.50	81.66	88.00		
	Random	~30%	1.35	2.79	2.69	5.08	47.56	38.37	50.95	31.42	54.37	33.36	52.91	31.98	80.04	84.49		
	FastV	~30%	1.56	3.34	2.55	4.65	46.48	36.75	51.61	32.85	56.69	34.22	53.10	31.59	85.94	88.56		
	DivPrune	~30%	1.36	2.38	2.63	4.26	45.94	36.75	51.52	31.86	57.23	35.25	53.16	32.75	81.69	87.98		
	TimePrune	~30%	1.30	2.19	2.40	3.62	50.37	40.58	52.00	33.90	59.47	36.46	59.81	35.94	88.18	89.14		
InternVL3.5-4B	Baseline	100%	1.82	3.51	3.54	6.69	19.29	15.78	38.28	18.12	42.22	27.78	28.57	14.29	93.54	89.51		
	Random	~30%	1.46	2.87	2.46	5.54	23.75	11.87	37.50	15.62	37.69	23.08	42.31	26.15	91.93	87.13		
	FastV	~30%	1.55	3.19	2.39	5.27	22.50	18.75	35.00	16.25	40.95	25.71	39.23	23.85	92.66	89.53		
	DivPrune	~30%	1.39	2.39	2.50	4.62	23.12	19.30	38.12	16.22	36.92	23.85	41.54	26.02	91.20	88.52		
	TimePrune	~30%	1.35	2.24	2.23	4.00	25.00	21.07	38.75	19.53	48.41	31.11	43.08	27.38	94.90	90.69		

Table 7: Overall performance of time series analysis on the **Finance** domain. $\downarrow$ indicates that lower is better, while $\uparrow$ indicates that higher is better. The best result in each block is **bolded**, and the second-best is underlined.

forms simple baselines such as Random pruning, suggesting that restricting selection to narrative-aligned timestamps can omit temporally relevant segments that are not explicitly grounded in the text. Although Narrative-Span can achieve competitive results in a few short-horizon or low-complexity settings, these gains are not stable and do not generalize across tasks. In particular, its performance degrades substantially on long-horizon forecasting and multimodal reasoning tasks, where temporal dependencies extend beyond explicitly mentioned narrative intervals.

In contrast, TimePrune consistently delivers stronger and more stable results across all evaluated backbones and metrics. The improvements are especially pronounced in long-range prediction scenarios, highlighting its ability to better capture task-relevant temporal structure. These results suggest that narrative timestamps alone are insufficient as a selection principle. Instead, effective temporal pruning requires integrating multimodal signals beyond text, including visual and contextual cues, and identifying semantically informative regions that are not explicitly referenced in the narrative.

Method			Token Budget	Forecasting (↓)				Trend Pred. (↑)		Next-day Prediction (↓)						MCQA (↑)	
				7-Day		14-Day		Past	Future	Max		Min		Diff		7-Day	14-Day
				MSE	MAE	MSE	MAE			MSE	MAE	MSE	MAE	MSE	MAE		
Qwen3VL-8B	Baseline	100%	29.04	4.29	25.50	3.93	<u>46.30</u>	25.00	29.53	4.11	23.41	3.61	50.01	5.72	<u>58.33</u>	<u>60.00</u>	
	Random	~30%	25.52	3.91	22.00	3.56	42.85	39.05	23.29	3.57	29.33	3.70	<u>24.36</u>	3.86	57.14	56.66	
	FastV	~30%	24.84	3.91	<u>19.36</u>	3.38	41.17	<u>39.86</u>	24.68	3.54	21.90	<u>3.30</u>	25.14	3.84	57.85	58.33	
	DivPrune	~30%	<u>23.32</u>	<u>3.63</u>	19.54	<u>3.35</u>	44.53	39.56	<u>21.82</u>	<u>3.51</u>	21.88	3.33	24.77	<u>3.74</u>	55.00	57.50	
	Narrative-Span	—	27.77	4.19	31.00	4.32	45.45	35.45	31.30	4.24	<u>19.61</u>	<u>3.30</u>	52.70	5.79	56.02	56.17	
	TimePrune	~30%	16.64	3.08	19.09	3.32	47.05	40.17	16.52	3.05	11.99	2.66	20.94	3.32	60.00	62.50	
InternVL3.5-8B	Baseline	100%	25.28	4.04	24.62	3.88	64.06	35.15	30.42	4.15	17.39	3.17	35.06	4.59	48.34	43.54	
	Random	~30%	32.53	4.47	25.58	3.90	57.50	34.16	27.11	4.04	15.04	3.00	<u>22.76</u>	3.50	64.28	64.28	
	FastV	~30%	37.03	4.64	22.97	3.68	56.66	35.83	<u>23.13</u>	<u>3.64</u>	19.05	3.10	22.82	<u>3.42</u>	62.85	<u>66.40</u>	
	DivPrune	~30%	31.53	4.22	<u>21.32</u>	<u>3.53</u>	59.16	35.85	27.28	4.00	14.94	3.01	25.18	3.66	<u>65.00</u>	62.25	
	Narrative-Span	—	<u>24.08</u>	<u>3.83</u>	23.75	3.79	58.18	34.46	30.13	4.11	<u>14.60</u>	<u>2.87</u>	34.18	4.55	54.31	58.23	
	TimePrune	~30%	22.81	3.57	20.82	3.52	<u>60.00</u>	37.50	21.24	3.19	13.23	2.55	19.90	2.78	67.14	66.42	
Qwen3VL-4B	Baseline	100%	<u>23.98</u>	3.77	26.19	3.94	52.37	26.66	36.81	4.66	18.86	3.23	57.66	6.08	<u>54.57</u>	<u>56.59</u>	
	Random	~30%	25.74	3.92	22.04	3.57	<u>55.69</u>	45.83	24.51	3.68	<u>16.24</u>	3.05	35.30	4.50	52.09	53.52	
	FastV	~30%	28.02	3.87	<u>19.41</u>	<u>3.36</u>	54.36	54.67	<u>19.98</u>	3.26	19.39	<u>3.03</u>	<u>20.51</u>	<u>3.39</u>	50.69	52.64	
	DivPrune	~30%	24.16	<u>3.68</u>	19.64	3.37	55.43	<u>55.94</u>	28.50	3.91	17.10	3.11	41.92	4.94	50.54	54.41	
	Narrative-Span	—	24.92	3.90	21.85	3.64	51.36	27.27	34.16	4.52	18.11	3.15	50.85	5.70	53.83	55.58	
	TimePrune	~30%	15.91	2.98	19.19	3.34	55.74	56.20	18.71	<u>3.28</u>	15.08	2.80	17.14	3.20	55.65	59.11	
InternVL3.5-4B	Baseline	100%	23.42	3.76	<u>22.67</u>	3.73	46.03	24.21	26.47	3.68	19.70	3.24	<u>21.08</u>	<u>3.33</u>	<u>53.49</u>	<u>55.00</u>	
	Random	~30%	32.98	4.51	27.18	4.00	42.21	48.33	34.79	4.56	21.96	3.55	36.29	4.17	22.14	54.28	
	FastV	~30%	24.16	3.86	25.57	3.89	43.33	<u>50.35</u>	<u>23.92</u>	<u>3.62</u>	20.64	3.20	25.80	3.54	52.85	52.14	
	DivPrune	~30%	31.63	4.23	25.13	<u>3.62</u>	44.16	49.16	29.66	4.10	17.78	3.21	35.47	4.23	51.42	52.85	
	Narrative-Span	—	<u>22.19</u>	3.67	23.68	3.82	45.20	28.18	25.59	3.71	<u>16.38</u>	2.93	21.60	3.34	51.97	53.84	
	TimePrune	~30%	21.99	<u>3.72</u>	21.32	3.53	<u>45.83</u>	52.50	22.71	3.51	14.37	<u>3.14</u>	20.73	2.67	55.71	58.57	

Table 8: Comparison with the training-free **Narrative-Span** baseline on the **Weather** domain. Narrative-Span retains only the time-series bins overlapping with extracted narrative time spans, without training a gating network; its token budget is not strictly fixed because retention depends on narrative coverage. The best result in each block is **bolded**, and the second-best is underlined.

Method	Token Budget	Forecasting (↓)				Stock Trend Prediction (↑)				News-stock Correlation (↑)				MCQA (↑)		
		7-Day		30-Day		7-Day		30-Day		7-Day		30-Day		7-Day	30-Day	
		MAE	MAPE	MAE	MAPE	3-way	5-way	3-way	5-way	3-way	5-way	3-way	5-way	Acc.	Acc.	
Proprietary LLMs																
GPT-4o	N/A	1.59	2.54	2.33	3.52	42.81	36.45	47.35	30.58	53.60	31.00	57.60	34.60	65.10	52.80	
Gemini	N/A	1.62	3.51	2.43	3.26	47.30	41.50	44.90	29.70	51.80	26.40	59.60	34.80	63.60	50.30	
Claude	N/A	1.42	2.09	2.06	2.84	44.90	33.40	52.05	31.70	50.40	29.00	57.90	34.30	75.60	61.10	
Open-source VLMs																
Qwen3VL-2B	Baseline	100%	1.97	2.78	2.40	5.01	20.00	17.50	38.31	18.35	48.33	16.67	43.14	19.17	71.69	60.34
	FastV	~30%	1.32	2.26	2.48	5.57	28.72	19.32	49.33	31.42	50.00	22.50	43.33	20.09	72.50	60.83
	DivPrune	~30%	1.52	3.25	2.54	4.70	28.53	19.11	49.14	29.52	51.67	24.17	43.33	20.83	74.16	58.33
	Random	~30%	1.38	2.92	2.68	6.21	28.89	18.81	48.00	29.14	50.83	21.67	41.67	20.00	71.66	63.73
	TimePrune	~30%	1.29	2.37	2.33	4.17	29.89	19.81	49.52	32.23	52.50	22.50	45.00	21.67	75.00	65.85

Table 9: **Additional Results on the Finance Domain**. ↓ indicates that lower is better, while ↑ indicates that higher is better. The best result is **bolded**. Results for proprietary LLMs are reproduced from [Chen et al. \(2025\)](#).

Method	Token Budget	Forecasting (↓)				Trend Pred. (↑)		Next-day Prediction (↓)						MCQA (↑)		
		7-Day		14-Day		Past	Future	Max		Min		Diff		7-Day Acc.	30-Day Acc.	
		MSE	MAE	MSE	MAE			MSE	MAE	MSE	MAE	MSE	MAE			
Proprietary LLMs																
GPT-4o	N/A	17.55	3.11	40.43	4.49	66.36	43.54	19.58	3.02	15.39	2.76	18.84	3.20	41.7	44.8	
Gemini	N/A	24.31	3.67	29.47	4.03	56.96	51.76	16.39	2.96	16.27	2.93	23.21	3.63	43.4	54.0	
Claude	N/A	22.48	3.50	25.08	3.75	59.78	56.87	18.69	3.21	13.42	2.63	19.10	3.26	51.8	51.2	
Open-source VLMs																
Qwen3VL-2B	Baseline	100%	31.54	4.49	30.36	4.44	44.53	21.66	40.97	5.01	33.53	4.38	56.50	6.12	51.66	54.16
	FastV	~30%	44.05	5.35	22.77	3.67	44.16	29.16	17.19	3.16	35.13	4.38	34.06	4.83	50.00	52.85
	DivPrune	~30%	35.91	4.76	23.16	3.70	45.00	35.00	19.73	3.35	41.95	5.07	34.65	4.81	52.85	52.14
	Random	~30%	34.01	4.68	25.61	3.92	42.50	36.66	27.60	3.85	42.16	5.02	37.16	4.80	52.14	49.28
	TimePrune	~30%	29.72	4.24	20.82	3.51	45.33	38.53	15.56	3.05	17.59	3.10	33.94	4.81	55.71	55.00

Table 10: **Additional Results on the Weather Domain.** ↓ indicates that lower is better, while ↑ indicates that higher is better. The best result is **bolded**. Results for proprietary LLMs are reproduced from [Chen et al. \(2025\)](#).

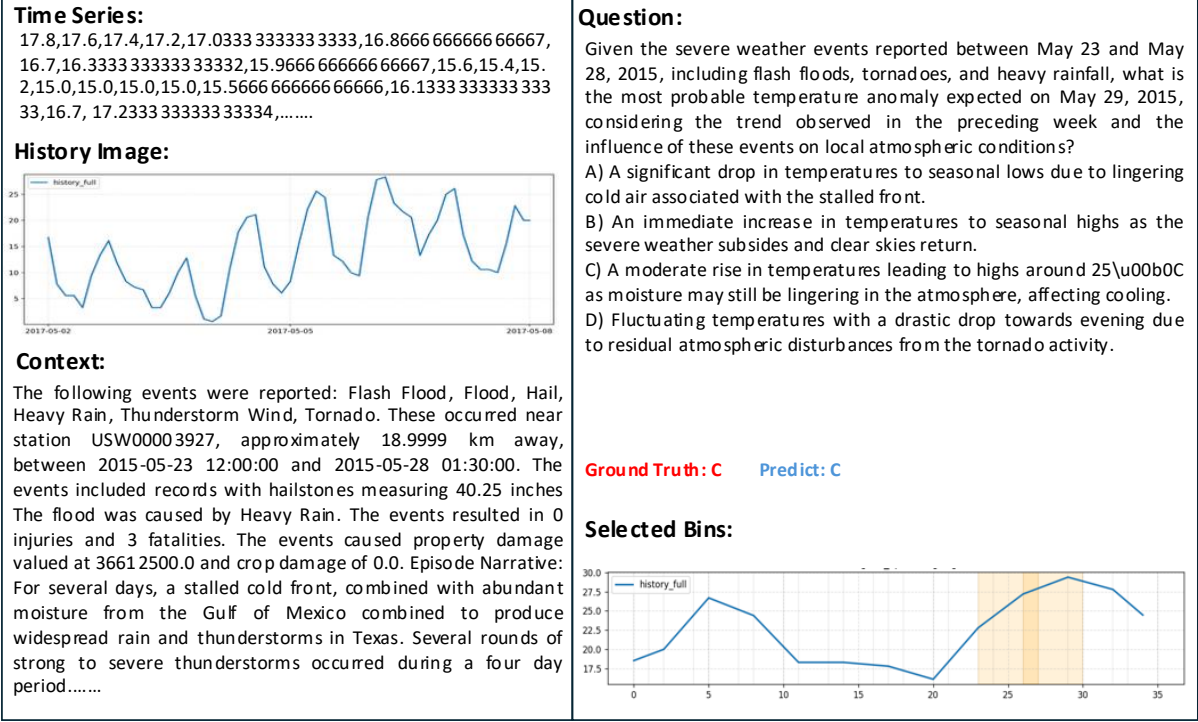
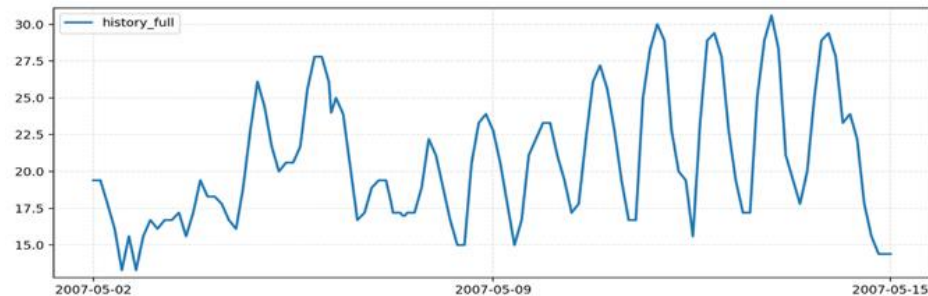


Figure 8: Weather Domain MCQA Case Study

Time Series:

19.4, 19.4, 19.4, 19.4, 18.866666666666667, 18.333333333333332, 17.8,
17.233333333333334, 16.666666666666668, 16.1, 15.166666666666668,
14.233333333333334, 13.3, 14.066666666666666, 14.833333333333334, 15.6,
14.833333333333334, 14.066666666666666, 13.3, 14.066666666666666,
14.833333333333334, 15.6, 15.966666666666667, 16.333333333333332, 16.7, 16.5,
16.3, 16.1, 16.3, 16.5, 16.7, 16.7, 16.7, 16.7, ...

History Image:



Context:

The following events were reported: Flash Flood, Flood, Hail, Heavy Rain. These occurred near station USW00003947, approximately 35.4383 km away, between 2007-05-06 06:00:00 and 2007-05-15 08:00:00. The events included records with hailstones measuring 0.8333333333333334 inches. The flood was caused by Heavy Rain. Thankfully, no injuries or fatalities were reported. The events caused property damage valued at 1467000.0 and crop damage of 2050000.0. Episode Narrative: A slow moving cold front and upper level storm system, combined

Ground Truth: Stable **Predict: Stable**

Selected Bins:

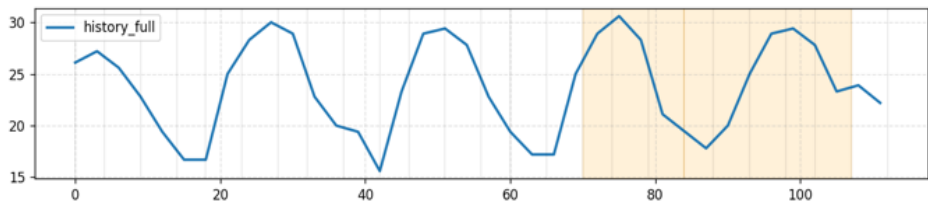


Figure 9: Weather Domain Trend Prediction Case Study

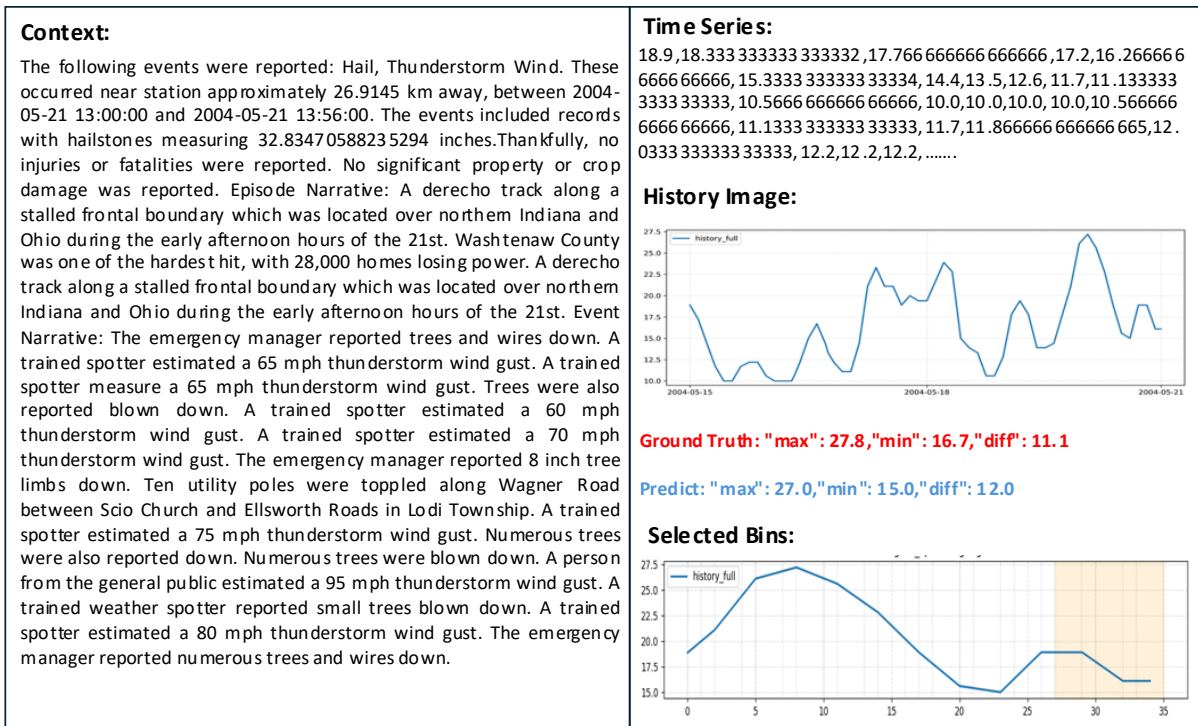


Figure 10: Weather Domain Indicator Prediction Case Study

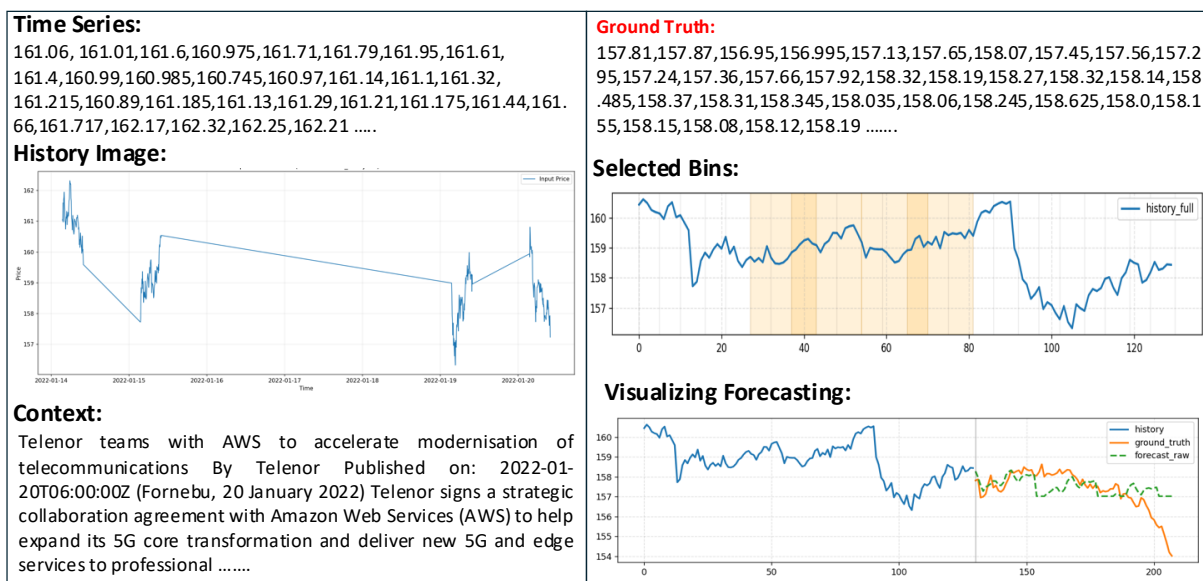


Figure 11: Finance Domain Short Forecasting Case Study

Time Series:

2.21,2.24,2.24,2.25,2.2539,2.265,2.25,2.255,2.23,2.235,2.23,2.2399,2.23,2.235,2.2373,2.2399,2.2319,2.2641,2.27,2.2701,2.275,2.26,2.25,2.255,2.255,2.2501,2.255,2.25,2.2576,2.25,2.25,2.2512,2.25,2.2599, ...

History Image:



Context:

Title: Nordic American Tankers (NAT) Reports Q2 Loss, Misses Revenue Estimates \n Content: ['We use cookies to understand how you use our site and to improve your experience. This includes personalizing content and advertising. To learn more, click here. By continuing to use our site, you accept our use of cookies, revised Privacy Policy and Terms of Service.', \"Don't Know Your Password?\", \"', 'New to Zacks? Get started here.', \"Don't Know Your Password?\", \"\", 'You are being directed to ZacksTrade, a division of LBMZ.....

Question:

"Which of the following statements about NAT's stock price and the given financial analysis is incorrect?\n

- A. The earnings surprise of -12.50% indicates that the stock is undervalued and offers a strong buying opportunity for investors looking for bargains following the news release.\n
- B. The company's Zacks Rank of #4 (Sell) following the news suggests that analysts expect continued underperformance relative to the market, which may lead to further declines in the stock price.\n
- C. Following the publication of Nordic American Tankers' disappointing Q2 earnings report, the stock price declined from an initial value of \$2.35 to around \$2.24 within the same trading day.\n
- D. The earnings report indicated a substantial miss in revenue estimates, which is likely to exert downward pressure on the stock price in the near term due to investor sentiment."

Ground Truth: A

Predict: A

Selected Bins:

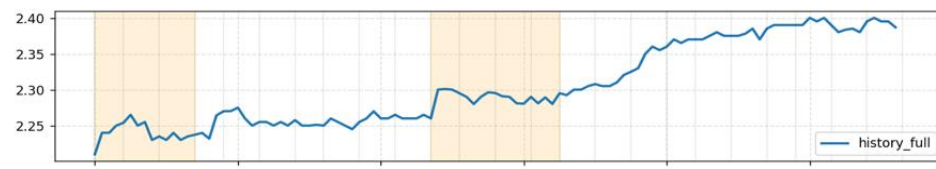
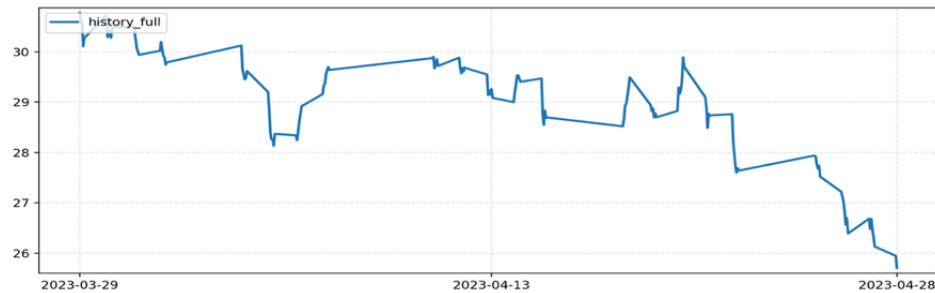


Figure 12: Finance Domain MCQA Case Study

Time Series:

30.775,30.61,30.55,30.11,30.275,30.71,30.525,30.29,30.42, 30.32,30.28,
30.49,30.5,30.435,30.26,30.08,30.01,29.94,30.02,30.195,30.035,29.92,29.88,
29.745,29.79,30.125,29.65,29.56,29.45,29.48,29.62,29.2,28.85,28.4,28.26,
28.26,28.135,28.3705,28.34,28.245,28.43,28.63,28.76,28.92, ...

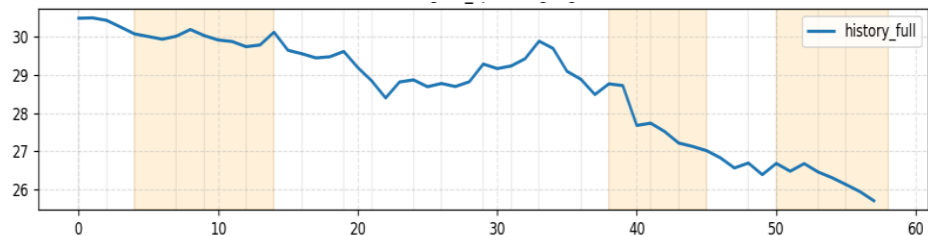
History Image:



Context:

Analyst Ratings for First Interstate BancSys\nBy Benzinga Insights\nPublished on: 2023-04-27T15:02:01Z\n\nAnalysts have provided the following ratings for First Interstate BancSys (NASDAQ:FIBK) within the last quarter: the last 3 months, 6 analysts have offered 12-month price targets for First Interstate BancSys. The company has an average price target of \$39.0 with a high of \$43.00 and a low of \$33.00. \nBelow is a summary of how these 6 analysts rated First Interstate BancSys over the past 3 months. The greater the number of bullish ratings, the ...Full story available on Benzinga.com\n\nArticle Content:\n\nAnalysts have provided the following ratings for First

Selected Bins:



Visualizing Forecasting:

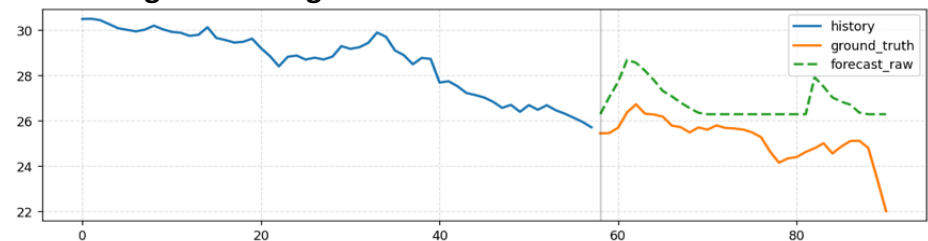


Figure 13: Finance Domain Long Forecasting Case Study

Time Series:

88.96,89.64,90.14,90.42,90.795,90.85,90.975,90.884,91.0375,90.87,91.39,91.49,91.65,91.89,91.69,91.62,91.54,91.64,91.61,91.72,91.5132,91.28,90.965,91.37,91.52,91.3,91.33,91.4,91.19,90.97,90.89,90.81,...

History Image:



Context:

Analyst Ratings for Owens-Corning
Content: ['\xa0', '"Over the past 3 months, 9 analysts have published their opinion on Owens-Corning OC stock. These analysts are typically employed by large Wall Street banks and tasked with understanding a company's business to predict how a stock will trade over the upcoming year.', 'In the last 3 months, 9 analysts have offered 12-month price targets for Owens-Corning. The company has an average price target of \$91.22 with a high of \$128.00 and a low of \$77.00.', 'Below is a summary of how these 9 analysts rated Owens-Corning over the past 3 months. The greater the number of bullish ratings, the more positive analysts are on the stock and the greater the number of bearish ratings, the more negative analysts are on the stock', 'This current average represents a 9.46% decrease from the previous average price target of \$100.75.', 'Analysts are

Ground Truth: Moderate Negative Correlation

Predict: Moderate Negative Correlation. \n\nThe analysis text indicates a decrease in the average price target, suggesting a negative sentiment. The price series shows a general downward trend in the last few days, which aligns with the negative sentiment. Therefore, there is a moderate negative correlation between the analysis sentiment and the stock price fluctuations.

Selected Bins:

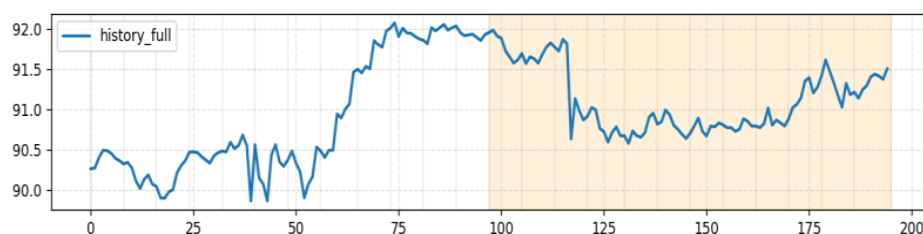


Figure 14: Finance Domain Trend Prediction Case Study