

StratPrune: Cascade Stratum Pruning for Efficient Time Series Reasoning

Chenghao Liu¹, Qian Liu¹, Aniket Mahanti¹, Alex Xing Wang², Jingwei Yi³, Wei Fan^{1*}

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

²School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

³Beijing Academy of Artificial Intelligence, China

{chenghao.liu, liu.qian, a.mahanti, wei.fan}@auckland.ac.nz, alex.wang@vuw.ac.nz, jwyi1029@gmail.com

Abstract

Multimodal large language models have recently shown strong potential for time series reasoning, where models jointly process textual context and long numerical sequences. However, dense time series representations introduce a large number of tokens, leading to substantial computational and memory overhead during inference. Existing token pruning methods are mainly designed for visual tokens and rely on global attention or representation geometry, overlooking the fact that token redundancy varies across layers and modalities. This makes pruning-induced perturbations difficult to control, especially when tokens are pruned across multiple transformer layers. In this paper, we propose *StratPrune*, a cascade stratum pruning framework for efficient time series reasoning. *StratPrune* first identifies an early-layer pruning stratum where time series tokens remain less diverse than text tokens, indicating high redundancy suitable for stable pruning. We then derive a hidden-state stability bound that decomposes pruning-induced perturbation into a layer-dependent stability term and a token-level attention saliency term. Guided by this analysis, *StratPrune* allocates layer-wise pruning budgets according to time series diversity and removes low-saliency time series tokens through a context-aware cascade process. Experiments on 7 datasets show that *StratPrune* outperforms strong pruning baselines under matched token budgets, achieving up to 30% relative improvement. Our code is publicly available.¹

1 Introduction

Large Language Models (LLMs) have achieved remarkable success at billion-parameter scales (Team, 2025; DeepSeek-AI, 2026), excelling in challenging tasks (Andriushchenko et al., 2025; Kuntz et al., 2025). Building on this progress, multimodal LLMs, which extend LLMs to process di-

verse modalities, have rapidly scaled up to tackle complex multimodal applications such as visual question answering (Kim et al., 2025), document understanding (Hu et al., 2025), and time series reasoning (Kong et al., 2025). In time series reasoning, models jointly reason over textual context and long numerical sequences, requiring the processing of both semantic and time series information (Guan et al., 2026; Kong et al., 2025).

Although multimodal LLMs generally improve time series reasoning performance, the large number of tokens introduced by dense time series representations substantially increases computational and memory costs (Shi et al., 2025). Existing approaches for token pruning in multimodal LLMs can be broadly categorized into three families. The first is attention-based ranking, which scores each token by the attention it receives and removes the lowest-scoring ones (Bolya et al., 2023; Yang et al., 2025b). The second is geometry-based methods, which operate on hidden-state embeddings to either select a maximally diverse subset under a max-min cosine criterion or merge highly similar tokens via bipartite matching (Alvar et al., 2025; Bolya et al., 2023). The third is adaptive methods that combine attention and geometry signals using input-dependent criteria such as effective rank or attention entropy (Ye et al., 2025; Endo et al., 2025).

While prior methods achieve strong performance in multimodal LLMs, they operate on tokens originating from a single modality and assume relatively uniform redundancy across layers. This becomes particularly severe for long time series inputs containing thousands of numerical data points, where the redundancy of time series tokens varies across layers (Liu et al., 2025b; Shi et al., 2025). Furthermore, existing pruning methods lack a principled characterization of how token removal affects the hidden states of the remaining tokens (Wen et al., 2025). As a result, the representation perturbation induced by pruning remains uncontrolled, partic-

*corresponding author

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

ularly under multi-layer pruning, where perturbations can accumulate across depth.

To address this, we introduce cascade stratum pruning for time series reasoning (StratPrune), an efficient pruning method that cascades token pruning across an early-layer stratum while allocating layer-wise pruning budgets based on modality representation diversity. Specifically, we first introduce a per-modality token diversity and identify an early-layer stratum in which time series tokens remain substantially less diverse than text tokens, indicating a region of high redundancy suitable for stable pruning. We then derive a hidden-state stability bound that decomposes the pruning-induced perturbation into a per-layer stability term and a per-token attention saliency term. Guided by this, we further propose a *stratum-wise pruning policy* that derives per-layer budgets by optimizing a stability-induced surrogate of the perturbation bound. We complement this with a *saliency-guided token selection rule* that ranks time series tokens based on their cross-modal attention scores. In summary, our contributions are as follows:

- We formulate token pruning for time series reasoning from a per-modality, per-layer diversity perspective, which reveals an early-layer stratum where time series tokens exhibit low diversity and can be safely pruned, and enables a centroid-corrected hidden-state stability bound that decomposes pruning-induced perturbations into a per-layer stability term and a per-token attention saliency term.
- We propose *StratPrune*, an efficient pruning framework that instantiates a stability-driven formulation through two tightly coupled modules: a stratum-wise pruning policy that assigns closed-form per-layer budgets derived from the analysis and a saliency-guided token selection rule that prioritizes time series tokens based on cross-modal attention.
- We evaluate StratPrune on 7 tasks in MT-Bench (Chen et al., 2025) that span the weather and finance domains across four backbones. Under the same pruning budget, StratPrune outperforms strong pruning baselines, achieving up to $\sim 30\%$ relative improvement in forecasting tasks and substantial gains in past-trend accuracy.

2 Related Works

LLMs for Time Series Reasoning. Time series reasoning, which aims to model and reason about temporal patterns from sequential numerical data (Chang et al., 2026), has recently emerged as an important capability of Large Language Models (LLMs) (Luo et al., 2025; Chang et al., 2026; Su et al., 2025). Leveraging LLMs, prior work has explored a range of tasks including numerical forecasting (Fan et al., 2022, 2023; Fei et al., 2026), trend classification (Zhou et al., 2025), and analytical question answering (Yugeswardeenoo et al., 2024; Chen et al., 2026) across domains such as weather (Li et al., 2025) and finance (Yu et al., 2024; Liu et al., 2025a). Despite these advances, time series reasoning remains constrained when task-relevant temporal signals are distributed across both numerical dynamics and unstructured contextual narratives. To address this limitation, multimodal time series LLMs have been proposed to extend the LLM-based time series reasoning paradigm by jointly modeling numerical time series and narrative context, such as news articles (Wang et al., 2024), market commentary (Kawarada et al., 2024), and weather reports (Li et al., 2025). These models aim to support more complex temporal reasoning tasks that require integrating signals across both numerical and textual modalities (Guan et al., 2026; Jin et al., 2024; Zhong et al., 2025).

Token Pruning. Token pruning is a compression technique that removes unimportant tokens from the input sequence to reduce compute and KV-cache memory (Yang et al., 2025b). Recently, token pruning has been widely adopted in multimodal LLMs, where visual tokens often constitute a large portion of the input sequence and significantly increase inference cost (Yang et al., 2025a; Zhang et al., 2025). Current methods can be broadly categorized into two families based on how token importance is defined. The first is attention-based ranking, which scores tokens by the attention they receive from other positions and removes low-scoring ones (e.g., FastV (Chen et al., 2024), VisionZip (Yang et al., 2025b)). These methods typically rely on attention signals extracted either from early LLM layers or the vision encoder. The second is geometry-based selection, which operates directly on visual embeddings without using attention, and selects or compresses tokens based on their distribution in representation space. For example, DivPrune (Alvar et al., 2025) preserves

a diverse subset via farthest-point sampling, while Token Merging compresses tokens by merging similar pairs. Beyond the vision modality, compression for time series foundation models has so far focused primarily on weight-level pruning rather than input-token reduction. A third line of work combines attention signals with geometric properties through adaptive, input-dependent criteria. AgilePruner (Baek et al., 2026) dynamically adjusts the pruning threshold based on the effective rank of token embeddings and attention entropy. SCOPE (Deng et al., 2025) instead adopts a greedy token selection strategy guided by both attention importance and coverage gain. Within the time series domain, structured pruning methods rank linear-layer channels in time series foundation models before fine-tuning (Zhao et al., 2025).

3 Method

In this section, we first introduce a per-modality, per-layer diversity analysis that induces a pruning stratum and a corresponding stability bound (Section 3.1). Based on this, we propose a stratum-wise pruning policy and a saliency-guided token selection rule, which determine layer-wise pruning allocation (Section 3.2) and token-level ranking within each layer (Section 3.3), respectively.

3.1 Time Series Pruning Stratum

To identify the layers where time series tokens can be stably pruned in modality-imbalanced inputs, we observe that such layers are characterized by time series (TS) tokens exhibiting relatively homogeneous representations, while text (TEXT) tokens remain comparatively more differentiated. To quantify this phenomenon, we introduce a per-modality *token diversity* statistic $s_X(\ell)$ that measures how hidden representations vary within each modality.

Definition 1 (Token Diversity). Let L denote the number of transformer layers, indexed by $\ell \in \{0, 1, \dots, L\}$ with $\ell = 0$ the input embedding and $\ell \geq 1$ post-attention representations. Let $\mathcal{T}_X$ denote the set of token indices belonging to modality $X \in \{\text{TS}, \text{TEXT}\}$, and let $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$ denote the hidden representation of token i at layer ℓ . We define the token diversity of X at layer ℓ as:

$$s_X(\ell) = \mathbb{E}_{i,j \in \mathcal{T}_X, i \neq j} [1 - \cos(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)})], \quad (1)$$

where the expectation is computed over token pairs. Based on this, we compute the layer-wise token di-

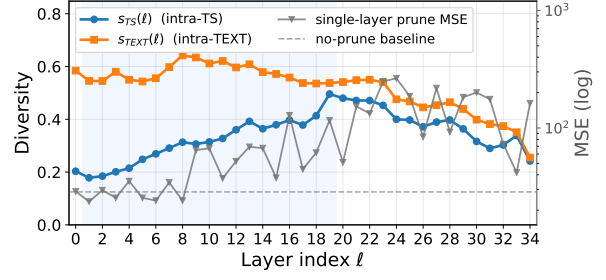


Figure 1: Per-layer within-modality diversity s_{TS} , s_{TEXT} and single-layer pruning MSE for Qwen3-4B. Shaded band: rising segment $\mathcal{L}_{\text{rise}}$; dashed line: no-prune MSE.

versity across all layers of Qwen3-4B (Team, 2025) in Figure 1. We provide observations as follows.

First, text tokens exhibit high diversity in early layers and become progressively more homogeneous in deeper layers. As shown in Fig. 1, $s_{\text{TEXT}}(\ell)$ is high at the input layer and gradually decreases with depth. At early layers, text representations are largely shaped by input token embeddings, leading to more distinct hidden states across tokens and higher within-modality diversity. As depth increases, model layers integrate contextual information across tokens, resulting in more aligned representations and a gradual reduction in diversity. Formally, we define the per-layer rate of change for each modality as $\gamma_X(\ell) := s_X(\ell+1)/s_X(\ell)$. Taking the product across layers yields:

$$\prod_{k=0}^{L-1} \gamma_{\text{TEXT}}(k) = \frac{s_{\text{TEXT}}(L)}{s_{\text{TEXT}}(0)} < 1, \quad (2)$$

which is consistent with a contraction tendency in representation space (Noci et al., 2022).

In contrast, $s_{\text{TS}}(\ell)$ exhibits a non-monotonic trend across depth. At shallow layers, TS representations are driven by low-level numeric structure in the input, resulting in highly similar token embeddings and low within-modality diversity. As depth increases, contextual information from time series values is progressively incorporated, leading to more differentiated representations. This differentiation accumulates until reaching the saturation depth $\ell_{\text{max}} := \arg \max_{\ell} s_{\text{TS}}(\ell)$, after which diversity gradually decreases. Taking the product across layers yields:

$$\prod_{k=0}^{\ell_{\text{max}}-1} \gamma_{\text{TS}}(k) = \frac{s_{\text{TS}}(\ell_{\text{max}})}{s_{\text{TS}}(0)} > 1, \quad (3)$$

indicating that time series diversity accumulates over early layers rather than contracting.

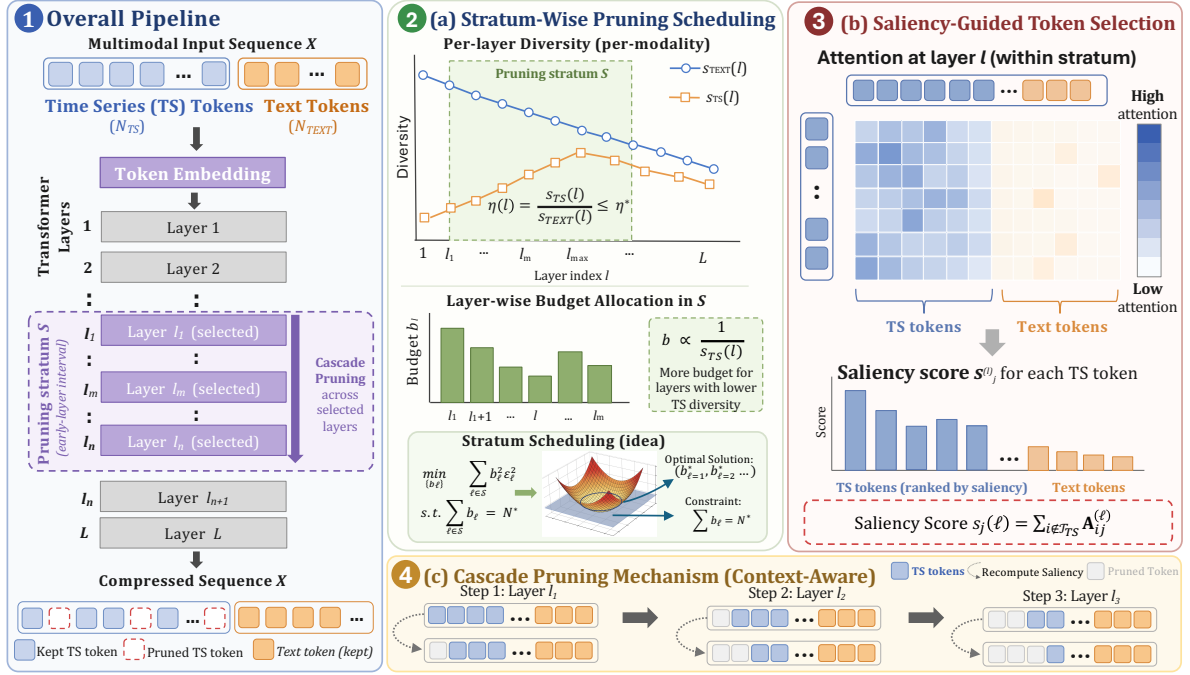


Figure 2: Overview of StratPrune. StratPrune first identifies an early pruning stratum where time series tokens exhibit low within-modality diversity relative to text tokens, then allocates layer-wise pruning budgets according to the stability-induced diversity profile. Within each selected layer, low-saliency time series tokens are removed, and the remaining tokens are progressively forwarded through a context-aware cascade pruning process.

Pruning Stratum. Building on the observations above, text diversity contracts across depth while time series diversity expands over the early layers $[0, \ell_{\max})$. To characterize this cross-modal asymmetry, we define the modality diversity ratio:

$$\eta(\ell) := \frac{s_{\text{TS}}(\ell)}{s_{\text{TEXT}}(\ell)} = \eta(0) \cdot \prod_{k=0}^{\ell-1} \frac{\gamma_{\text{TS}}(k)}{\gamma_{\text{TEXT}}(k)}, \quad (4)$$

where the second equality follows from Eqs. 2 and 3. Motivated by this observation, we define the *pruning stratum* $\mathcal{S}$ as an early-layer interval in which time series tokens can be pruned with minimal impact on downstream representations:

Definition 2 (Pruning Stratum). Given a threshold $\eta^* \in (0, 1)$, the pruning stratum $\mathcal{S}$ is defined as the set of early rising-stage layers where time series representations remain substantially less diverse than text representations:

$$\mathcal{S} := \{\ell \in [1, \ell_{\max}) \mid \eta(\ell) \leq \eta^*\}. \quad (5)$$

The stratum $\mathcal{S}$ serves as a set of candidate layers for time series pruning. However, this candidate set does not always ensure stable downstream representations. As shown in Figure 1, single-layer time series pruning at $\ell \in \mathcal{S}$ induces a non-trivial residual MSE relative to the no-pruning baseline. To

formalize this observation, we introduce a per-layer measure of stability that quantifies the representation displacement induced by time series pruning at $\ell \in \mathcal{S}$. Let $\alpha_{jk}^{(\ell)}$ denote the softmax attention weight from query position j to key position k at layer ℓ , and let $\mathbf{V}^{(\ell)}$ denote the value projection matrix. For a time series pruning subset $\mathcal{P} \subseteq \mathcal{T}_{\text{TS}}$, we define a masked attention operator that excludes pruned tokens from the key-value set. The resulting post-pruning representation at a non-TS position j is given by $\tilde{\mathbf{h}}_j^{(\ell+1)} := \sum_{k \notin \mathcal{P}} \alpha_{jk}^{(\ell)} \mathbf{V}^{(\ell)} \mathbf{h}_k^{(\ell)}$ where attention is restricted to unpruned tokens. The following proposition characterizes the resulting stability of this pruning operation and establishes a uniform bound over layers in $\mathcal{S}$.

Proposition 1 (Pruning Stability). For any $\ell \in \mathcal{S}$ and TS pruning subset $\mathcal{P} \subseteq \mathcal{T}_{\text{TS}}$, let $\bar{\mathbf{h}}_{\text{TS}}^{(\ell)} := \frac{1}{|\mathcal{T}_{\text{TS}}|} \sum_{k \in \mathcal{T}_{\text{TS}}} \mathbf{h}_k^{(\ell)}$ denote the time series centroid at layer ℓ . The centroid-corrected post-attention displacement at any non-TS position j satisfies:

$$\Delta \mathbf{h}_j := \tilde{\mathbf{h}}_j^{(\ell+1)} - \mathbf{h}_j^{(\ell+1)} + \rho_{\mathcal{P}}^{(j)} \mathbf{V}^{(\ell)} \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}, \quad (6)$$

$$\|\Delta \mathbf{h}_j\| \leq C_{\ell} \cdot \rho_{\mathcal{P}}^{(j)} \cdot \sqrt{\eta^* \cdot s_{\text{TEXT}}(\ell)},$$

where $\rho_{\mathcal{P}}^{(j)} := \sum_{k \in \mathcal{P}} \alpha_{jk}^{(\ell)}$ denotes the attention mass that query j assigns to the pruned subset, and $C_{\ell} := \|\mathbf{V}^{(\ell)}\|_{\text{op}}$. The correction term

$+\rho_{\mathcal{P}}^{(j)} \mathbf{V}^{(\ell)} \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}$ removes the shared centroid contribution that remains recoverable from the surviving TS tokens, so that $\Delta \mathbf{h}_j$ isolates the irrecoverable deviation of the pruned tokens from the TS mean. The bound decomposes naturally into two factors. The first, $\rho_{\mathcal{P}}^{(j)}$, captures how strongly query j depends on the pruned tokens through attention. The second, $\sqrt{\eta^* s_{\text{TEXT}}(\ell)}$, captures the intrinsic redundancy of TS representations at layer ℓ , induced by the stratum condition $\eta(\ell) \leq \eta^*$. As $\eta^* \rightarrow 0$, the TS representations become increasingly concentrated around their centroid, causing the perturbation to vanish. This decomposition directly guides the two components: the redundancy term governs layer-wise budget allocation (§3.2), while the attention term drives saliency-guided token selection by minimizing $\rho_{\mathcal{P}}^{(j)}$ (§3.3). Proof in Appendix A.

3.2 Stratum-Wise Pruning Policy

The above section establishes $\mathcal{S}$ as a stable region for time series pruning. Nevertheless, we still consider how to distribute the pruning budget across layers within $\mathcal{S}$, since Equation 6 indicates that the perturbation induced by pruning time series (TS) tokens differs across layers through the layer-dependent scale $\sqrt{s_{\text{TS}}(\ell)}$. This suggests that applying a uniform pruning ratio across $\mathcal{S}$ is suboptimal. Existing attention-based token pruning methods (Alvar et al., 2025; Yang et al., 2025b) typically apply a shared pruning ratio together with a global saliency score across selected layers. However, such strategies ignore the layer-dependent stability structure characterized by Eq. 6. Inspired by stability-aware layer-wise sparsity allocation (Ye et al., 2025; Endo et al., 2025; Tao et al., 2025), we propose a *stratum-wise pruning policy* that adapts pruning decisions to the stability profile of each layer within $\mathcal{S}$. The proposed policy consists of two components: a stability-guided layer selection that determines which layers in $\mathcal{S}$ participate in pruning, and a layer-dependent pruning policy that allocates the global pruning budget N^* according to the per-layer TS diversity.

Stratum Stability for Layer Selection. To establish that every layer in $\mathcal{S}$ admits stable pruning, we derive a per-layer stability margin that bounds the representation displacement uniformly over the choice of TS pruning subset $\mathcal{P}_\ell$. We begin with Proposition 1. Its proof (Appendix A) yields a tighter TS-domain characterization, $\|\Delta \mathbf{h}_j^{(\ell)}\| \leq$

$C_\ell \rho_{\mathcal{P}_\ell}^{(j)} \sqrt{s_{\text{TS}}(\ell)}$, via the identity $\eta(\ell) s_{\text{TEXT}}(\ell) = s_{\text{TS}}(\ell)$, before the relaxation $\eta(\ell) \leq \eta^*$ that gives the looser TEXT-domain form stated in Eq. 6. Combining this tighter form with the normalization property of softmax attention $\rho_{\mathcal{P}_\ell}^{(j)} = \sum_{k \in \mathcal{P}_\ell} \alpha_{jk}^{(\ell)} \leq 1$, we obtain the following bound:

$$\begin{aligned} \|\Delta \mathbf{h}_j^{(\ell)}\| &\leq \rho_{\mathcal{P}_\ell}^{(j)} \varepsilon_\ell \leq \varepsilon_\ell, \\ \varepsilon_\ell &:= C_\ell \sqrt{s_{\text{TS}}(\ell)}, \quad \forall \ell \in \mathcal{S}, \forall \mathcal{P}_\ell \subseteq \mathcal{T}_{\text{TS}}, \end{aligned} \quad (7)$$

where $\mathcal{P}_\ell$ denotes the TS pruning subset at layer ℓ . The first inequality captures the linear sensitivity of the representation displacement to the attention mass assigned to pruned tokens, while the second follows from the upper bound $\rho_{\mathcal{P}_\ell}^{(j)} \leq 1$, yielding a $\mathcal{P}_\ell$ -independent per-layer stability margin ε_ℓ . Here, $\varepsilon_\ell := C_\ell \sqrt{s_{\text{TS}}(\ell)}$ characterizes the intrinsic stability of layer ℓ , where $s_{\text{TS}}(\ell)$ measures the within-modality diversity of TS representations and C_ℓ denotes the operator norm of the value projection. By construction, layers in $\mathcal{S}$ lie in the early stage of the TS diversity profile (Eq. 3), where $s_{\text{TS}}(\ell)$ remains small and TS representations have not yet developed strong heterogeneity. Together with the boundedness of C_ℓ , this implies that ε_ℓ remains uniformly small over $\mathcal{S}$. Consequently, every layer in $\mathcal{S}$ admits a uniformly stable pruning guarantee, and we therefore adopt $\mathcal{S}$ directly as the inference-time pruning layer set of StratPrune.

Dynamic Pruning Scheduling. To leverage the layer stability structure for pruning allocation, we formulate the pruning process as a budget allocation problem over layers. Specifically, we assign a per-layer pruning budget $b_\ell := |\mathcal{P}_\ell|$ for each $\ell \in \mathcal{S}$, where the allocation is dynamically conditioned on the stability margin ε_ℓ defined in Eq. 7. Under this formulation, layers with smaller stability margins admit larger pruning budgets, while more sensitive layers are assigned fewer removals. With the total budget constraint $\sum_{\ell \in \mathcal{S}} b_\ell = N^*$ and the uniform-attention approximation $\rho_{\mathcal{P}_\ell}^{(j)} \propto b_\ell$, the global displacement bound reduces to a quadratic form over layer budgets:

$$\begin{aligned} \min_{\{b_\ell\}} \sum_{\ell \in \mathcal{S}} b_\ell^2 \varepsilon_\ell^2 \quad \text{s.t.} \quad \sum_{\ell \in \mathcal{S}} b_\ell &= N^* \\ \implies b_\ell &= w_\ell N^*, \quad w_\ell := \frac{1/s_{\text{TS}}(\ell)}{\sum_{\ell' \in \mathcal{S}} 1/s_{\text{TS}}(\ell')}. \end{aligned} \quad (8)$$

The resulting allocation $w_\ell \propto 1/s_{\text{TS}}(\ell)$ reveals a coupling between pruning intensity and TS diver-

sity: layers with lower $s_{\text{TS}}(\ell)$ receive larger pruning budgets, while layers approaching the stratum boundary at ℓ_{max} are allocated smaller budgets.

3.3 Saliency-Guided Token Selection

While the stratum-wise policy in Section 3.2 determines both the selected layers $\mathcal{S}$ and their corresponding pruning budgets b_ℓ , it does not specify which individual time series tokens should be removed within each layer. This requires a finer-grained criterion at the token level. To this end, we define token importance by measuring its contribution to non-TS computation. Specifically, Proposition 1 shows that the representation displacement of non-TS queries is governed by their attention to pruned TS tokens, suggesting that TS tokens receiving higher attention from non-TS positions are more critical. Based on this, we aggregate attention from all queries to each TS token, and define the token-level saliency score as:

$$s_j(\ell) = \sum_{i \in \mathcal{T}_{\text{ts}} \cup \mathcal{T}_{\text{text}}} \mathbf{A}_{ij}^{(\ell)}, \quad (9)$$

namely the total attention received by TS token j from all non-TS and TS positions at layer ℓ . This score measures how strongly a token is attended to by the rest of the model, and thus reflects its potential impact on downstream representations if removed. This formulation is consistent with prior attention-based token pruning methods (Chen et al., 2024; Alvar et al., 2025; Yang et al., 2025b), which similarly rank tokens according to their received attention. In contrast to these heuristic designs, our scoring function arises directly from the stability bound in Proposition 1, providing an interpretation of attention as a measure of pruning sensitivity.

4 Experiments

4.1 Experimental Setups

Models and Benchmark. We evaluate pruning on four popular LLM backbones, including Gemma4, the latest open-source model series from Google, and the Qwen3 series (Team, 2025). The Qwen3 models, including Qwen3-4B and Qwen3-8B, share the same architecture but differ in parameter scale, enabling controlled analysis of model size effects. We also include TimeOmni-1-7B (Guan et al., 2026), a model specialized for time series reasoning, to examine whether pruning remains effective on a domain-specialized backbone. We

evaluate on MTBench (Chen et al., 2025), a multi-modal time series benchmark consisting of seven tasks across weather and financial domains. The benchmark covers regression tasks including 7-day and 14-day forecasting and indicator prediction, classification tasks such as trend prediction and news-price correlation, and both short-form and long-form multiple-choice question answering. For evaluation metrics, regression tasks are measured using mean squared error (MSE) and mean absolute error (MAE), while classification and MCQA tasks are evaluated using accuracy.

Baselines. We compare StratPrune with four representative token pruning methods. Although these methods were originally designed primarily for pruning visual tokens in vision-language models, we adapt them to time series inputs by treating each time series token as the basic pruning unit and applying a matched overall pruning ratio for fair comparison. TopV (Yang et al., 2025a) ranks tokens according to the cumulative attention they receive at a fixed early layer and removes the lowest-ranked half. DivPrune (Alvar et al., 2025) formulates token selection as a Max-Min diversity optimization problem, iteratively choosing a subset of tokens that maximizes the minimum pairwise distance in the embedding space. AgilePruner (Baek et al., 2026) dynamically selects between attention-based and diversity-based pruning for each input, guided by common proxies such as effective rank and attention entropy. SCOPE (Deng et al., 2025) performs iterative token selection using a joint criterion that combines attention-based importance with a coverage-gain term measuring each token’s contribution to semantic completeness.

4.2 Results

Table 1 reports the comparison on MTBench under matched $\sim 67\%$ token budgets across four backbones. On the three generalist backbones (Gemma-4-E4B, Qwen3-4B, and Qwen3-8B), StratPrune consistently achieves the lowest forecasting MSE as well as the lowest next-day max, min, and diff MSE across most evaluation settings. It also delivers relative reductions of up to around 30% compared to the strongest pruning baseline on 14-day forecasting and next-day min MSE. In addition, StratPrune attains the highest past-trend accuracy on all generalist backbones. On the specialized TimeOmni-1-7B backbone, the advantage remains clear. StratPrune achieves more than 25% relative

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	Acc.	Acc.
Gemma4-4B	Baseline	100%	26.66	3.94	26.63	3.94	43.5	27.0	32.22	4.42	19.55	3.45	59.61	6.02	48.0	33.5
	TopV	~67%	<u>26.39</u>	4.01	<u>25.55</u>	3.99	41.0	10.5	30.47	4.30	40.94	4.99	56.37	5.85	47.5	33.5
	DivPrune	~67%	33.86	4.54	87.59	7.30	<u>46.5</u>	<u>28.0</u>	53.50	5.71	32.47	4.45	98.98	7.76	<u>48.5</u>	40.5
	SCOPE	~61%	32.32	4.43	25.59	3.95	38.0	23.0	29.94	4.27	52.02	5.63	55.39	5.81	40.5	38.5
	AgilePruner	~73%	26.93	4.05	77.33	6.86	41.5	24.0	<u>25.68</u>	<u>3.95</u>	<u>15.59</u>	<u>3.08</u>	<u>47.51</u>	<u>5.38</u>	42.0	49.0
	StratPrune (Ours)	~67%	25.08	3.28	24.82	3.16	49.0	29.0	23.90	3.52	10.34	2.53	41.62	4.60	51.0	<u>46.5</u>
Qwen3-4B	Baseline	100%	40.89	4.85	61.34	7.83	48.0	36.0	28.90	4.32	17.92	3.17	56.73	6.20	52.0	50.5
	TopV	~67%	49.30	5.39	73.95	8.60	42.5	23.5	27.63	4.24	26.56	3.75	62.58	6.41	47.0	47.0
	DivPrune	~67%	52.02	5.82	78.03	8.83	31.5	<u>52.5</u>	24.82	3.90	15.39	3.92	48.65	6.98	<u>52.5</u>	51.0
	SCOPE	~61%	40.89	4.92	61.34	7.83	38.5	47.0	<u>21.72</u>	<u>3.68</u>	27.39	3.88	37.37	<u>4.72</u>	53.5	<u>52.5</u>
	AgilePruner	~73%	41.31	<u>4.85</u>	61.97	7.87	39.0	55.0	24.20	3.91	<u>15.00</u>	3.87	47.43	6.89	50.0	53.5
	StratPrune (Ours)	~67%	40.74	4.16	56.28	5.77	48.5	55.0	20.35	3.30	11.50	2.62	<u>38.08</u>	4.44	50.0	51.5
Qwen3-8B	Baseline	100%	41.32	4.78	72.40	6.19	36.0	24.5	28.23	4.24	36.35	4.04	62.85	6.81	43.0	39.5
	TopV	~69%	<u>32.68</u>	4.34	73.23	6.32	29.5	<u>25.0</u>	29.42	4.39	58.76	5.67	65.51	8.19	40.5	40.5
	DivPrune	~69%	55.12	5.61	96.46	10.15	27.5	24.5	26.11	4.20	51.52	5.88	58.23	7.42	41.0	38.5
	SCOPE	~54%	35.66	4.54	63.47	5.76	29.5	21.0	26.82	3.91	56.69	5.61	73.68	6.41	40.5	41.0
	AgilePruner	~73%	38.55	4.77	67.08	6.28	32.0	20.5	<u>24.70</u>	<u>3.85</u>	76.60	8.75	<u>55.08</u>	7.22	<u>44.5</u>	<u>42.0</u>
	StratPrune (Ours)	~69%	30.64	4.32	50.00	5.51	38.5	26.5	22.64	3.75	25.19	3.92	45.42	4.78	49.5	48.5
TimeOmni-1-7B	Baseline	100%	53.24	7.30	65.78	8.11	53.0	22.0	57.61	7.59	37.09	5.78	37.53	5.12	25.0	<u>52.0</u>
	TopV	~66%	68.08	8.25	84.42	9.19	45.0	18.0	33.87	5.82	32.28	4.16	30.65	4.90	30.0	37.0
	DivPrune	~66%	<u>52.11</u>	<u>6.59</u>	49.02	<u>6.79</u>	48.0	17.0	66.63	8.16	31.16	4.00	36.38	4.58	24.0	35.0
	SCOPE	~66%	68.92	8.30	85.46	9.24	44.0	21.0	44.72	6.69	49.97	5.50	29.33	4.20	28.0	38.0
	AgilePruner	~72%	73.64	8.58	91.31	9.56	52.0	19.0	<u>29.49</u>	<u>5.43</u>	<u>29.99</u>	4.06	<u>17.53</u>	<u>3.62</u>	<u>34.0</u>	40.0
	StratPrune (Ours)	~66%	37.65	4.79	<u>52.36</u>	5.64	62.0	24.0	28.53	4.42	12.49	2.92	11.28	3.41	45.0	55.0

Table 1: Overall comparison of pruning methods on the MTBench weather domain across four LLM backbones under matched token-budget ratios. The best and second best results are in **bold** and underlined, respectively.

improvement in MSE for 7-day forecasting. It also improves the accuracy of trend prediction by 9% compared to the no-pruning setting. Furthermore, it outperforms all baselines on both MCQA variants, with an 11-point gain on the 7-day MCQA over the strongest pruning baseline.

We further analyze inference efficiency on Qwen3-8B in Table 2. Among the pruning baselines, StratPrune attains the lowest per-token decoding latency and ties SCOPE for the smallest peak memory, while reducing the KV cache to the same size as the other pruners at the matched token budget. The $K=3$ scoring forward passes introduce a modest end-to-end overhead relative to single-pass pruners, but StratPrune remains substantially faster than DivPrune, whose iterative diversity selection dominates its runtime.

5 Further Analysis

Effect of Pruning Ratio. To further evaluate our approach, we analyze the effect of varying the prune ratio r_{ts} on Qwen3-8B, ranging from 0.2 for light pruning to 0.8 for aggressive pruning. As shown in Figure 3a, StratPrune remains competitive

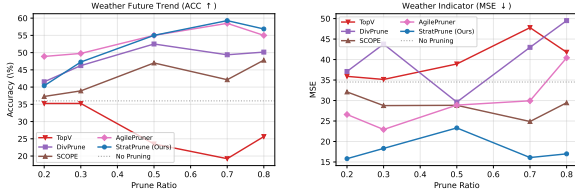
Method	Decode (ms/tok)	E2E (s)	Mem (GB)	KV (MB)	Lat. ($\times$)
No-prune	12.6	0.90	7.90	264.7	1.00
TopV	20.1	1.56	8.44	173.4	0.58
DivPrune	19.6	7.08	8.45	173.2	0.13
SCOPE	19.6	1.54	8.26	173.4	0.58
AgilePruner	19.6	1.53	8.35	173.7	0.59
StratPrune (Ours)	19.3	1.39	8.26	173.4	0.65

Table 2: Inference efficiency on Qwen3-8B at a matched $\sim 67\%$ token budget over a 64-token generation. Decode = ms/token; E2E = end-to-end latency; Mem = peak GPU memory; KV = analytic KV-cache size; Lat. = latency relative to no-prune.

with AgilePruner across the entire ratio range in future trend prediction. Both methods outperform the no-pruning baseline, whereas TopV degrades substantially as the prune ratio increases. Figure 3b presents the results on next-day prediction. StratPrune achieves the lowest error at every prune ratio and maintains stable performance throughout the sweep, keeping the MSE below approximately 16. In contrast, DivPrune and AgilePruner exceed the no-pruning baseline under high pruning ratios.

Ablation Variant	Variant	Next-day MSE ↓	Forecast MSE ↓	Trend Acc ↑	MCQA Acc ↑
(a) K	$K=1$	21.78	39.74	26.9%	43.2%
	$K=2$	22.47	38.80	30.0%	46.8%
	$K=3$	20.86	36.25	32.5%	49.0%
	$K=5$	21.53	37.80	32.4%	48.2%
(b) Re-forward	Single-pass	20.91	40.24	54.2%	44.7%
	Context-aware	20.91	37.81	58.4%	46.9%

Table 3: Ablation study of StratPrune. All cells report the average result. Deployed defaults highlighted.



(a) Future trend prediction.

(b) Indicator prediction.

Figure 3: Pruning performance across token-budget ratios on two weather domain tasks.

Layer-wise Selection. To validate the diversity-based layer selection strategy in §3.1, we evaluate each Qwen3-8B transformer layer and examine the relationship between pruning MSE and the corresponding intra-TS diversity score $s_{TS}(\ell)$. As shown in Figure 4, within the candidate region $\mathcal{L}_{rise}$, pruning MSE increases approximately linearly with $s_{TS}(\ell)$, achieving an R^2 of 0.55 with a bootstrap 95% confidence interval. Moreover, the layer with the lowest diversity score, L_1 , also attains the smallest pruning MSE (23.7, marked by $\star$), suggesting that lower-diversity layers provide more stable pruning behavior. This trend disappears outside the candidate region $[\ell_{max}, L]$, where pruning MSE exhibits larger variance and no consistent ordering across diversity scores is observed. Together, these results empirically support restricting cascade pruning to $\mathcal{L}_{rise}$.

Ablation Study. Table 3 presents an ablation study of three core components of StratPrune. The results show that the proposed stratum provides the most consistent overall performance. Using fewer pruning layers leads to weaker results, while adding more layers produces marginal changes. This suggests that most effective pruning decisions are already captured by a small set of low-diversity layers. Replacing the iterative context-aware cascade with a single forward aggregation step consistently reduces performance, indicating that recomputing token importance after each pruning stage helps later layers adapt to the updated token context.

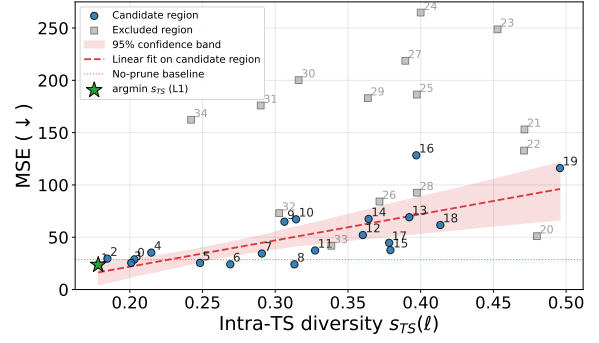


Figure 4: Per-layer pruning MSE versus intra-TS diversity $s_{TS}(\ell)$. Within $\mathcal{L}_{rise}$ (blue), MSE scales approximately linearly with $s_{TS}(\ell)$ ($R^2 = 0.55$), and the minimum-diversity layer matches the empirical minimum-MSE layer ($\star$). This relationship breaks down in the excluded region (gray).

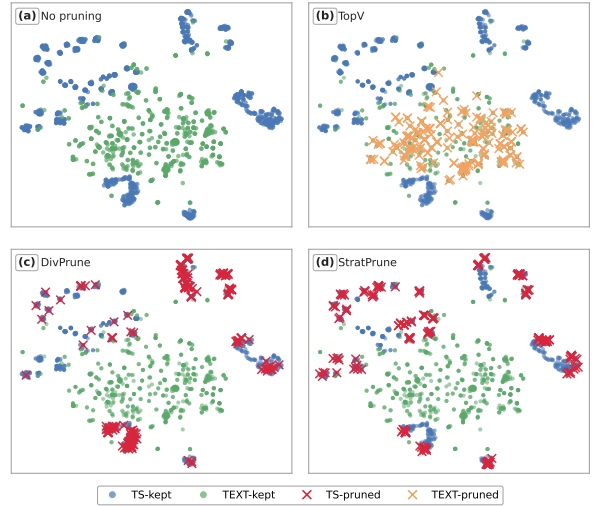


Figure 5: t-SNE visualization of time series and text token hidden states. Pruned tokens are marked by $\times$.

Token Selection Behavior. Figure 5 visualizes hidden-state tokens together with the keep masks produced by different methods. TopV highlights the limitation of modality-unaware pruning, where most removed tokens come from the text region while the time series distribution remains largely unchanged. DivPrune instead removes tokens near the boundary of the time series cluster, preserving redundant interior representations while discarding structurally informative regions. In contrast, StratPrune concentrates pruning within compact low-diversity regions while preserving the overall structure of the time series distribution, consistent with our layer-wise diversity analysis.

6 Conclusion

In this paper, we introduce StratPrune, a cascade stratum pruning framework for time series reasoning in multimodal LLMs. By identifying an early-layer pruning stratum through per-modality diversity analysis and deriving a hidden-state stability formulation, StratPrune enables both layer-wise budget allocation and token selection. Experiments on multiple tasks and LLM backbones demonstrate consistent improvements over strong baselines under matched token budgets across diverse settings. Our results further highlight the importance of modeling layer and modality dependent redundancy for efficient multimodal time series reasoning.

Limitations

While StratPrune demonstrates consistent improvements across multiple tasks and LLM backbones, several limitations remain for future work. First, our diversity analysis focuses on the interaction between time series and text representations. Although this setting captures a common form of multimodal time series reasoning, extending the framework to richer modality combinations such as image, audio, or sensor streams remains an important direction. Second, the proposed pruning stratum is identified from layer-wise representation statistics observed during inference. While the resulting strategy generalizes well across evaluated models and tasks, incorporating instance-adaptive mechanisms that dynamically adjust the stratum for each input may further improve flexibility. Third, our stability formulation characterizes the local effect of pruning on hidden-state representations and provides a principled basis for pruning allocation. Understanding how these local perturbations relate to end-to-end reasoning behavior over long inference trajectories remains an interesting direction for future investigation.

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.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J. Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. 2025. Agentharm: A benchmark for measuring harmfulness of LLM agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Changwoo Baek, Jou Won Song, Sohyeon Kim, and Kyeongbo Kong. 2026. [Agilepruner: An empirical study of attention and diversity for adaptive visual token pruning in large vision-language models](#). *CoRR*, abs/2603.01236.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ching Chang, Yidan Shi, Defu Cao, Wei Yang, Jeehyun Hwang, Haixin Wang, Jiacheng Pang, Wei Wang, Yan Liu, Wen-Chih Peng, and Tien-Fu Chen. 2026. A survey of reasoning and agentic systems in time series with large language models. *Trans. Mach. Learn. Res.*, 2026.

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.

DeepSeek-AI. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). *CoRR*, abs/2606.19348.

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. 2025. SCOPE: saliency-coverage oriented token pruning for efficient multimodal llms. 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.

- Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. 2025. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 22826–22835. IEEE.
- 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 *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7522–7529.
- 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. *arXiv preprint arXiv:2203.07681*.
- 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. *arXiv preprint arXiv:2605.17340*.
- Tong Guan, Zijie Meng, Dianqi Li, Shiyu Wang, Chao-Han Huck Yang, Qingsong Wen, Zuozhu Liu, Sabato Marco Siniscalchi, Ming Jin, and Shirui Pan. 2026. [Timeomni-1: Incentivizing complex reasoning with time series in large language models](#). In *The Fourteenth International Conference on Learning Representations*.
- Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5817–5834. Association for Computational Linguistics.
- 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. 2024. 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.
- Masayuki Kwarada, Tatsuya Ishigaki, and Hiroya Takamura. 2024. Prompting for numerical sequences: A case study on market comment generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 13190–13200. ELRA and ICCL.
- Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Comput. Surv.*, 57(10):249:1–249:35.
- Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. 2025. Time-mqa: Time series multi-task question answering with context enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 29736–29753. Association for Computational Linguistics.
- Thomas Kuntz, Agatha Duzan, Hao Zhao, Francesco Croce, Zico Kolter, Nicolas Flammarion, and Maksym Andriushchenko. 2025. Os-harm: A benchmark for measuring safety of computer use agents. 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*.
- Haobo Li, Zhaowei Wang, Jiachen Wang, Yueya Wang, Alexis Kai-Hon Lau, and Huamin Qu. 2025. Cllmate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 17536–17562. Association for Computational Linguistics.
- 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.
- Yucong Luo, Yitong Zhou, Mingyue Cheng, Jiahao Wang, Daoyu Wang, Tingyue Pan, and Jintao Zhang. 2025. [Time series forecasting as reasoning: A slow-thinking approach with reinforced llms](#). *CoRR*, abs/2506.10630.
- Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurélien Lucchi. 2022. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Feifei Shi, Xueyan Yin, Kang Wang, Wanyu Tu, Qifu Sun, and Huansheng Ning. 2025. [Large language](#)

- models for time series analysis: Techniques, applications, and challenges. *CoRR*, abs/2506.11040.
- 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.
- Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. Dycoke: Dynamic compression of tokens for fast video large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 18992–19001. Computer Vision Foundation / IEEE.
- Qwen Team. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- 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*.
- Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Token pruning in multimodal large language models: Are we solving the right problem? 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 15537–15549. Association for Computational Linguistics.
- Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponuswamy Sadayappan, Xia Hu, and Bo Yuan. 2025a. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19803–19813. Computer Vision Foundation / IEEE.
- Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025b. Visionzip: Longer is better but not necessary in vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19792–19802. Computer Vision Foundation / IEEE.
- Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. 2025. Atp-llava: Adaptive token pruning for large vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24972–24982. Computer Vision Foundation / IEEE.
- 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*.
- Dharunish Yugeswardeenoo, Kevin Zhu, and Sean O’Brien. 2024. Question-analysis prompting improves LLM performance in reasoning tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 543–554. Association for Computational Linguistics.
- Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. 2025. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. 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*.
- 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*.
- 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.
- Jiahui Zhou, Dan Li, Lin Li, Zhuomin Chen, Shunyu Wu, Haozheng Ye, Jian Lou, and Costas J. Spanos. 2025. [Enhancing LLM reasoning for time series classification by tailored thinking and fused decision](#). *CoRR*, abs/2506.00807.

A Proof of Proposition 1 (Pruning Stability on $\mathcal{S}$)

We prove the bound in Eq. 6. Let $\ell \in \mathcal{S}$ be fixed and let $\bar{\mathbf{h}}_{\text{TS}}^{(\ell)} := \frac{1}{|\mathcal{T}_{\text{TS}}|} \sum_{k \in \mathcal{T}_{\text{TS}}} \mathbf{h}_k^{(\ell)}$ denote the TS centroid at layer ℓ .

Step 1: Diversity controls maximum deviation from the centroid. For unit-norm hidden states, $1 - \cos(\mathbf{h}_a, \mathbf{h}_b) = \frac{1}{2} \|\mathbf{h}_a - \mathbf{h}_b\|^2$, so the within-modality cosine diversity coincides with the centered squared spread up to a factor of two. In particular, $s_{\text{TS}}(\ell) = \mathbb{E}_{k \in \mathcal{T}_{\text{TS}}} [\|\mathbf{h}_k^{(\ell)} - \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}\|^2]$. A standard concentration argument (sub-Gaussian or bounded-support tokens) yields a layer-uniform constant $c > 0$ such that:

$$\max_{k \in \mathcal{T}_{\text{TS}}} \|\mathbf{h}_k^{(\ell)} - \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}\| \leq \sqrt{s_{\text{TS}}(\ell) / c}. \quad (10)$$

The factor $1/\sqrt{c}$ is absorbed into the constant C_ℓ in the bound below.

Step 2: Centroid-corrected displacement has a clean form. The pre- and post-pruning attention outputs at any non-TS position j are:

$$\begin{aligned} \mathbf{h}_j^{(\ell+1)} &= \sum_k \alpha_{jk}^{(\ell)} \mathbf{V}^{(\ell)} \mathbf{h}_k^{(\ell)}, \\ \tilde{\mathbf{h}}_j^{(\ell+1)} &= \sum_{k \notin \mathcal{P}} \alpha_{jk}^{(\ell)} \mathbf{V}^{(\ell)} \mathbf{h}_k^{(\ell)}, \end{aligned}$$

whose naive difference is $\tilde{\mathbf{h}}_j^{(\ell+1)} - \mathbf{h}_j^{(\ell+1)} = -\sum_{k \in \mathcal{P}} \alpha_{jk}^{(\ell)} \mathbf{V}^{(\ell)} \mathbf{h}_k^{(\ell)}$. Adding the centroid-correction term $\rho_{\mathcal{P}}^{(j)} \mathbf{V}^{(\ell)} \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}$ (with $\rho_{\mathcal{P}}^{(j)} = \sum_{k \in \mathcal{P}} \alpha_{jk}^{(\ell)}$) and using the identity $\rho_{\mathcal{P}}^{(j)} \bar{\mathbf{h}}_{\text{TS}}^{(\ell)} = \sum_{k \in \mathcal{P}} \alpha_{jk}^{(\ell)} \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}$ yields:

$$\Delta \mathbf{h}_j = \mathbf{V}^{(\ell)} \sum_{k \in \mathcal{P}} \alpha_{jk}^{(\ell)} (\bar{\mathbf{h}}_{\text{TS}}^{(\ell)} - \mathbf{h}_k^{(\ell)}). \quad (11)$$

The centroid correction therefore eliminates the dependence on the absolute hidden-state norm and leaves the per-token *deviation* from the TS centroid.

Step 3: TS-domain bound. Taking norms in Eq. 11, applying triangle inequality, and bounding the convex combination by its maximum element,

$$\begin{aligned} \|\Delta \mathbf{h}_j\| &\leq \|\mathbf{V}^{(\ell)}\|_{\text{op}} \sum_{k \in \mathcal{P}} \alpha_{jk}^{(\ell)} \|\mathbf{h}_k^{(\ell)} - \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}\| \\ &\leq \|\mathbf{V}^{(\ell)}\|_{\text{op}} \rho_{\mathcal{P}}^{(j)} \max_{k \in \mathcal{P}} \|\mathbf{h}_k^{(\ell)} - \bar{\mathbf{h}}_{\text{TS}}^{(\ell)}\|. \end{aligned} \quad (12)$$

Substituting Eq. 10 and absorbing $1/\sqrt{c}$ into C_ℓ gives the tighter *TS-domain* form:

$$\|\Delta \mathbf{h}_j\| \leq C_\ell \cdot \rho_{\mathcal{P}}^{(j)} \cdot \sqrt{s_{\text{TS}}(\ell)}. \quad (13)$$

This is the form used by the per-layer stability margin ε_ℓ in Section 3.2.

Step 4: Stratum relaxation gives the TEXT-domain bound. By the definition of $\mathcal{S}$, every $\ell \in \mathcal{S}$ satisfies $\eta(\ell) \leq \eta^*$. Combined with the identity $s_{\text{TS}}(\ell) = \eta(\ell) s_{\text{TEXT}}(\ell)$,

$$\sqrt{s_{\text{TS}}(\ell)} = \sqrt{\eta(\ell) s_{\text{TEXT}}(\ell)} \leq \sqrt{\eta^* s_{\text{TEXT}}(\ell)}.$$

Substituting into Eq. 13 yields the looser TEXT-domain form stated in the proposition,

$$\|\Delta \mathbf{h}_j\| \leq C_\ell \cdot \rho_{\mathcal{P}}^{(j)} \cdot \sqrt{\eta^* s_{\text{TEXT}}(\ell)}. \quad (14)$$

□

B StratPrune Algorithm

Algorithm 1 summarizes the full StratPrune procedure. The offline phase (lines 1–6) runs once per model on a small calibration set: it estimates the per-modality diversity profile, identifies the pruning stratum $\mathcal{S}$, and pre-computes the layer-wise allocation weights $\{w_\ell\}_{\ell \in \mathcal{S}}$. The inference phase (lines 7–16) is invoked per input prompt: it walks the stratum layers in ascending order, scores the surviving TS tokens by their received attention, and drops the lowest-saliency b_ℓ tokens at each scheduled layer; the resulting cascade keep-mask is then passed to standard generation. The procedure introduces no additional trainable parameters and adds $|\mathcal{S}|$ extra forward passes during scoring (typically $|\mathcal{S}| \leq 3$ in our experiments).

C Baseline Adaptation Details

All four baselines, TopV (Yang et al., 2025a), DivvPrune (Alvar et al., 2025), AgilePruner (Baek et al., 2026), and SCOPE (Deng et al., 2025), were originally developed for vision–language models, where pruning is performed over visual patch tokens produced by a frozen vision encoder. To adapt them to multimodal time series LLMs under a controlled setting, we use each time series token as the basic pruning unit and re-implement all methods within a shared pruning interface. This appendix describes the method-specific adaptation choices used in our experiments, ensuring that the comparisons in Table 1 and Table 4 are fair.

Algorithm 1 StratPrune Inference Procedure

Require: model f , prompt with token sets $\mathcal{T}_{\text{inst}}, \mathcal{T}_{\text{ts}}, \mathcal{T}_{\text{text}}$; TS prune ratio r_{ts} ; threshold η^* ; calibration set $\mathcal{C}$

Ensure: compressed keep-mask $\mathbf{m} \in \{0, 1\}^S$

Offline calibration (run once per model)

1: Compute $s_{\text{TS}}(\ell), s_{\text{TEXT}}(\ell)$ on $\mathcal{C}$ for $\ell = 0, \dots, L$ $\triangleright$ Eq. 1

2: $\ell_{\max} \leftarrow \arg \max_{\ell} s_{\text{TS}}(\ell)$

3: $\eta(\ell) \leftarrow s_{\text{TS}}(\ell) / s_{\text{TEXT}}(\ell)$ for each ℓ $\triangleright$ Eq. 4

4: $\mathcal{S} \leftarrow \{\ell \in [1, \ell_{\max}) : \eta(\ell) \leq \eta^*\}$ $\triangleright$ Def. 2

5: $w_{\ell} \leftarrow \frac{1/s_{\text{TS}}(\ell)}{\sum_{\ell' \in \mathcal{S}} 1/s_{\text{TS}}(\ell')}$ for $\ell \in \mathcal{S}$ $\triangleright$ Eq. 8

6: Sort $\mathcal{S}$ in ascending order of ℓ

Per-inference cascade pruning

7: Initialize active set $\mathcal{A} \leftarrow \mathcal{T}_{\text{ts}}$; $\mathbf{m} \leftarrow \mathbf{1}_{\mathcal{S}}$

8: $N^* \leftarrow \lfloor r_{\text{ts}} |\mathcal{T}_{\text{ts}}| \rfloor$; $n_{\text{done}} \leftarrow 0$

9: **for** $t = 0, \dots, |\mathcal{S}| - 1$ **with** $\ell_t \in \mathcal{S}$ in ascending order **do**

10: Forward f with current $\mathbf{m}$ up to layer ℓ_t ; capture $\mathbf{A}^{(\ell_t)}$

11: $s_j(\ell_t) \leftarrow \sum_{i \in \mathcal{T}_{\text{ts}} \cup \mathcal{T}_{\text{text}}} \mathbf{A}_{ij}^{(\ell_t)}, \forall j \in \mathcal{A}$ $\triangleright$ Eq. 9

12: $b_t \leftarrow \lfloor w_{\ell_t} N^* \rfloor$ if $t < |\mathcal{S}| - 1$ else $N^* - n_{\text{done}}$

13: $\mathcal{P}_t \leftarrow$ bottom-min($b_t, |\mathcal{A}| - 1$) of $\mathcal{A}$ by $s_j(\ell_t)$

14: $\mathbf{m}[j] \leftarrow 0$ for $j \in \mathcal{P}_t$; $\mathcal{A} \leftarrow \mathcal{A} \setminus \mathcal{P}_t$; $n_{\text{done}} \leftarrow n_{\text{done}} + b_t$

15: **end for**

16: **return** $f.\text{GENERATE}(\text{prompt}, \text{attention_mask} = \mathbf{m})$

Shared design choices. Three conventions are applied uniformly across all baselines and to StratPrune itself, so that no method gains an advantage from the boundary conditions:

- *Only time series tokens are prunable.* Instruction tokens $\mathcal{T}_{\text{inst}}$ are always preserved (system prompt, task description, output-format string). Narrative tokens $\mathcal{T}_{\text{text}}$ are also always preserved, since pruning text degrades all methods uniformly; all baselines and our method run in a TS-only pruning regime.
- *Token budget aligned by TS-token retention rate*, i.e., each method targets the same ratio $r_{\text{ts}} = |\mathcal{P}|/|\mathcal{T}_{\text{ts}}|$ of TS tokens to drop, not a ratio of the full sequence. We set the target at $r_{\text{ts}} = 0.33$ (keep $\sim 67\%$ of TS tokens). Two methods (SCOPE, AgilePruner) include in-

ternal adaptive selection rules that cannot be forced to an exact ratio without breaking their design; we report the realized retention in the “Token Budget” column of every results table.

- *Single-layer scoring at layer 4.* For every attention- or hidden-state-based baseline we extract the scoring signal from the same transformer layer ($\ell = 4$) used by FastV in the original paper. This isolates the effect of the scoring rule from the effect of layer choice, and matches the hook layer used by StratPrune in the main comparison.

TopV (FastV-style column-sum attention).

TopV scores each visual token by the column sum of received attention at an early LLM layer (Yang et al., 2025a). We keep the original scoring rule unchanged and only restrict it to the TS slice of the attention matrix:

$$s_j^{\text{TopV}}(\ell=4) = \sum_{i \in \mathcal{T}_{\text{ts}} \cup \mathcal{T}_{\text{text}}} \mathbf{A}_{ij}^{(\ell)}, \quad j \in \mathcal{T}_{\text{ts}},$$

where instruction rows are excluded by construction since most lie before the TS block and are causally blocked. The bottom $r_{\text{ts}} \cdot |\mathcal{T}_{\text{ts}}|$ TS tokens by s_j^{TopV} are masked at generation; text and instruction tokens are untouched. This is the direct LLM-side analog of FastV; the only adaptation is restricting the scored column index set to $\mathcal{T}_{\text{ts}}$.

DivPrune (geometric diversity over TS hidden states).

DivPrune (Alvar et al., 2025) formulates pruning as max–min cosine diversity over visual-encoder embeddings. Since multimodal time series LLMs have no separate visual encoder for TS, we use the LLM’s own layer-4 hidden states $\{\mathbf{h}_j^{(4)}\}_{j \in \mathcal{T}_{\text{ts}}}$ as the geometric substrate, and restrict the selection set to TS token positions:

$$\tilde{\mathbf{E}}_{\text{ts}} = \arg \max_{\substack{S \subset \mathcal{T}_{\text{ts}} \\ |S|=K}} \min_{u,v \in S} [1 - \cos(\mathbf{h}_u^{(4)}, \mathbf{h}_v^{(4)})],$$

solved by the same greedy farthest-point-sampling procedure as the original DivPrune. Text and instruction tokens contribute neither candidates nor distance constraints. The $\sim 860 \times$ runtime penalty reported in our efficiency table reflects FPS over ~ 1000 TS positions per sample at single precision.

AgilePruner (adaptive threshold via effective rank).

AgilePruner (Baek et al., 2026) adaptively widens its similarity threshold using the effective rank (erank) of token embeddings and the attention

entropy of the [CLS] row. Two adaptations are required because multimodal time series LLMs lack a CLS token:

- *Effective rank.* We compute erank on the matrix $\mathbf{H}_{\text{ts}}^{(4)} \in \mathbb{R}^{|\mathcal{T}_{\text{ts}}| \times d}$ of layer-4 TS hidden states, using the original formula $\text{erank}(A) = \exp(-\sum_i q_i \log q_i)$, $q_i = \sigma_i / \sum_j \sigma_j$ with σ_i the singular values of A . This restricts the geometry to TS and removes the dependence on visual-encoder geometry, which is undefined in our setting.
- *Attention entropy.* In place of the CLS row, we use the attention row corresponding to the *last* instruction token, which carries the “return one word” output-format prefix and plays an analogous summarization role at this layer. Entropy is then computed by the original formulation, restricted to TS column indices.

The adaptive threshold rule $\tau_i = \text{order}_i \times (\text{erank}_{\text{input}} / \text{erank}_{\text{avg}} \times 0.01)$ is applied unchanged. Because the threshold is sample-adaptive, the realized TS budget varies per sample.

SCOPE (greedy coverage in TS token space). SCOPE (Deng et al., 2025) greedily picks tokens that maximize additive coverage of high-attention regions. We adapt SCOPE by restricting both the candidate set and the already-selected set to the TS modality: the marginal coverage gain is measured between candidate and already-selected TS tokens using layer-4 hidden states, and text/instruction tokens contribute neither candidates nor coverage targets. Because SCOPE’s selection terminates when marginal gain falls below a threshold rather than at a fixed budget, the realized TS retention varies between $\sim 54\%$ and $\sim 66\%$ across backbones. We report the realized rate in the “Token Budget” column of every results table rather than forcing an exact ratio match, which would require modifying SCOPE’s stopping criterion.

D Extended Layer-wise Selection Results

To further validate the diversity-based layer selection introduced in Section 3.1, we extend the layer-wise selection analysis from Section 5 to two additional backbones, Qwen3-4B and Gemma4-4B. For each backbone, we independently prune each transformer layer under the same pruning ratio and record the resulting task MSE. This analysis examines whether the relationship between intra-TS

diversity $s_{\text{TS}}(\ell)$ and pruning error, originally observed on Qwen3-8B, generalizes across model scale and architecture.

E Extended Results

E.1 Validation of the Per-Layer Stability

Single-layer pruning validates the stratum design. Figure 6 and Figure 7 study the effect of pruning each layer independently under the same pruning ratio. The top panel plots the resulting MSE against the intra-TS diversity $s_{\text{TS}}(\ell)$. Within the candidate region, the MSE increases approximately linearly with $s_{\text{TS}}(\ell)$, while the fitted trend remains stable under the 95% confidence band. This observation is consistent with the stability bound in Eq. 7: since the per-layer perturbation is controlled by $C_\ell \sqrt{s_{\text{TS}}(\ell)}$, the corresponding squared error is expected to scale with $s_{\text{TS}}(\ell)$ up to a layer-dependent constant. Therefore, the empirical sweep supports using $s_{\text{TS}}(\ell)$ as a practical proxy for layer-level pruning sensitivity.

The stratum separates stable and unstable pruning layers. The bottom panel shows the same MSE values ordered by layer index. For the candidate layers before the rising-stage boundary, pruning remains relatively stable: most layers stay close to the no-prune baseline, and the variation is moderate despite using the same pruning ratio. In contrast, once the pruning layer moves outside the stratum, the MSE becomes substantially larger and more volatile, with several late layers producing catastrophic errors. This sharp change indicates that the boundary used to define $\mathcal{S}$ is not merely a heuristic cutoff, but marks a practical stability transition between layers where TS redundancy can be safely removed and layers where pruning disrupts downstream representations.

The minimum-diversity layer explains the allocation rule. The green star marks $\ell^* = \arg \min_\ell s_{\text{TS}}(\ell)$. This layer yields one of the lowest post-pruning MSEs and stays close to the no-prune baseline, matching the prediction that lower intra-TS diversity leads to a smaller pruning-induced perturbation. This result directly motivates our allocation strategy: layers with smaller $s_{\text{TS}}(\ell)$ receive larger pruning budgets through $b_\ell \propto 1/s_{\text{TS}}(\ell)$. Thus, the single-layer sweep provides evidence for both components of our design: restricting pruning to the stable stratum and allocating more pruning to layers with stronger redundancy.

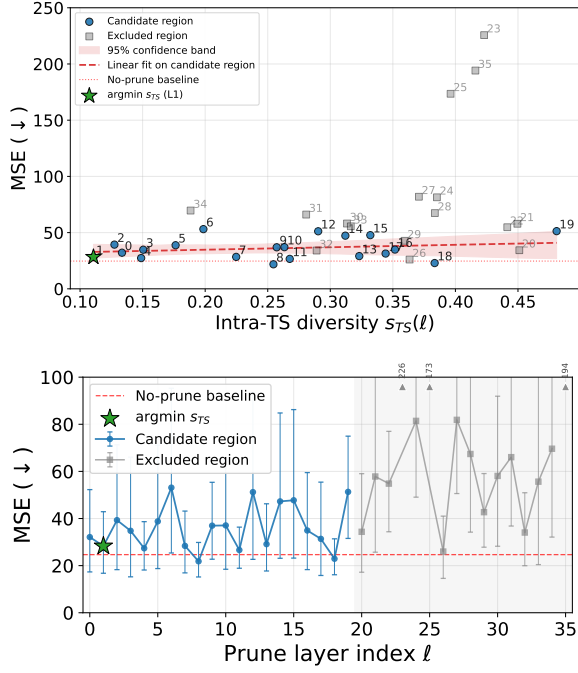


Figure 6: Single-layer pruning sweep on Qwen3-4B. **Top:** MSE versus intra-TS diversity $s_{TS}(\ell)$, with a linear fit over candidate layers. **Bottom:** MSE ordered by layer index. Excluded layers show larger and less stable errors, while the minimum-diversity layer marked by the green star yields low pruning error.

E.2 Evaluation on Finance Domain

Table 4 reports the results on the finance domain using the same experimental protocol, backbones, and matched token-budget setting as Table 1. Across all four backbones, StratPrune consistently achieves the lowest or near-lowest forecasting errors in MAE and MAPE, as well as the highest accuracy on stock trend prediction, news–stock correlation, and multiple-choice question answering. These results show that the improvements observed in weather domain generalize to a different domain.

E.3 Cross-Backbone Layer Diversity

Figure 8 reports the layer-wise diversity profiles of $s_{TS}(\ell)$ and $s_{TEXT}(\ell)$ across all four evaluated backbones. The results show a consistent pattern across models. TS token diversity increases in the early layers and then gradually saturates, whereas TEXT token diversity decreases monotonically with depth. This consistent separation supports the pruning stratum defined in Section 3.1 and confirms that the same layer-selection principle generalizes across different backbones.

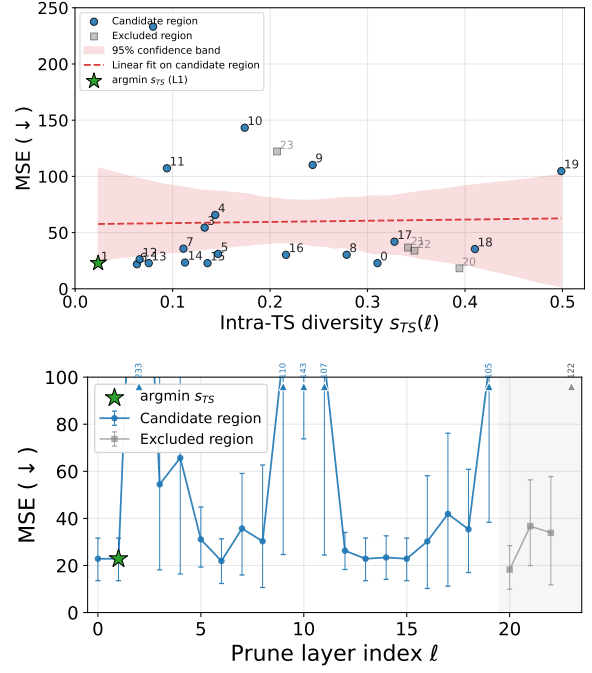


Figure 7: Single-layer pruning sweep on Gemma4-4B. **Top:** MSE versus intra-TS diversity $s_{TS}(\ell)$, with a linear fit over candidate layers. **Bottom:** MSE ordered by layer index. Excluded layers yield larger errors, while the minimum-diversity layer achieves low pruning error.

F Case Study

Figures 9 and 10 present qualitative examples from the weather and finance domains, respectively, illustrating how StratPrune’s stratum selection and budget allocation operate on a single real sample.

Figure 9 presents a real weather forecasting case. Panel (a) shows the input weather report, while Panel (b) shows the 168-hour temperature history together with the 24-hour target window. The ground truth in the target window follows a decreasing trend. Panels (c) and (d) illustrate how StratPrune converts layer-wise diversity statistics into a concrete pruning decision for this sample. Panel (c) reports the per-layer diversity profile $s_{TS}(\ell)$ within the rising segment. Panel (d) shows the corresponding inverse-diversity budget weights, defined by $w_\ell \propto 1/s_{TS}(\ell)$, together with the aggregated keep mask m^{ts} , which retains approximately 67% of the time series tokens.

Figure 10 follows the same overall structure with a representative finance example. Diversity calibration is performed once for each backbone rather than separately for each input. Therefore, Panels (c) and (d) use the same stratum layers and budget weights as Figure 9, while the domain-specific content differs in Panels (a) and (b).

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.
Gemma4-4B	Baseline	100%	4.44	4.62	5.66	5.70	52.0	28.0	44.0	22.0	76.0	42.0	65.0	35.0	60.0	72.0
	TopV	~67%	4.99	4.96	5.51	5.53	76.0	35.0	60.0	28.0	88.0	48.0	70.0	38.0	56.0	60.0
	DivPrune	~67%	8.50	9.20	10.20	11.50	29.0	17.0	25.0	14.0	47.0	25.0	55.0	23.0	47.0	60.0
	SCOPE	~61%	5.10	5.05	6.35	7.20	45.0	24.0	38.0	19.0	55.0	30.0	45.0	22.0	52.0	56.0
	AgilePruner	~73%	5.30	5.25	6.55	7.45	48.0	26.0	42.0	21.0	60.0	32.0	50.0	26.0	53.0	58.0
	StratPrune (Ours)	~67%	4.20	4.40	5.30	5.30	78.0	38.0	65.0	35.0	89.0	50.0	75.0	42.0	64.0	74.0
Qwen3-4B	Baseline	100%	3.77	4.64	5.31	4.45	16.0	10.0	14.0	8.0	24.0	14.0	22.0	12.0	88.0	85.0
	TopV	~67%	3.77	4.52	4.04	3.79	31.8	18.0	28.0	16.0	54.5	30.0	48.0	26.0	86.4	78.0
	DivPrune [†]	~67%	4.82	8.98	7.61	14.15	31.0	17.0	26.0	14.0	60.8	34.0	52.0	28.0	75.0	68.0
	SCOPE	~61%	3.95	4.85	7.85	6.50	25.0	14.0	22.0	12.0	18.0	10.0	16.0	19.0	78.0	70.0
	AgilePruner	~73%	4.05	4.95	7.65	6.30	28.0	15.0	24.0	13.0	16.0	9.0	15.0	18.0	80.0	72.0
	StratPrune (Ours)	~67%	3.34	3.99	3.80	3.50	40.0	22.0	34.0	19.0	63.0	36.0	55.0	30.0	90.0	87.0
Qwen3-8B	Baseline	100%	5.16	5.55	6.52	5.85	33.0	18.0	28.0	15.0	66.0	36.0	56.0	30.0	89.0	79.0
	TopV	~69%	5.79	5.95	6.40	5.80	36.8	20.0	32.0	17.0	68.4	37.0	60.0	32.0	81.1	76.0
	DivPrune [†]	~69%	6.55	6.40	7.20	7.10	32.6	18.0	27.0	14.0	69.5	38.0	61.0	33.0	85.3	87.5
	SCOPE	~54%	5.85	5.70	6.85	6.50	30.0	16.0	25.0	13.0	62.0	33.0	52.0	28.0	83.0	75.0
	AgilePruner	~73%	5.95	5.85	6.95	6.65	32.0	17.0	27.0	14.0	64.0	34.0	54.0	29.0	85.0	76.0
	StratPrune (Ours)	~69%	4.95	5.30	6.10	5.55	38.0	21.0	34.0	19.0	71.0	39.5	62.0	34.0	90.0	89.0
TimeOmni-1-7B	Baseline	100%	3.46	3.30	4.60	4.45	92.0	50.0	78.0	42.0	92.0	52.0	80.0	44.0	48.0	51.0
	TopV	~66%	4.52	4.12	4.56	7.20	73.0	68.0	28.0	15.0	45.0	24.0	38.0	20.0	47.0	38.0
	DivPrune [†]	~66%	4.55	4.13	4.38	8.40	25.0	64.0	62.0	12.0	38.0	20.0	32.0	17.0	38.0	32.0
	SCOPE	~66%	4.50	4.15	4.35	7.80	28.0	75.0	64.0	53.0	40.0	21.0	34.0	18.0	42.0	36.0
	AgilePruner	~72%	4.48	4.11	4.33	7.50	30.0	46.0	36.0	14.0	42.0	22.0	36.0	19.0	44.0	37.0
	StratPrune (Ours)	~66%	3.20	3.10	4.30	4.15	94.0	52.0	80.0	44.0	94.0	54.0	82.0	46.0	52.0	53.0

Table 4: Finance domain results across four backbones.

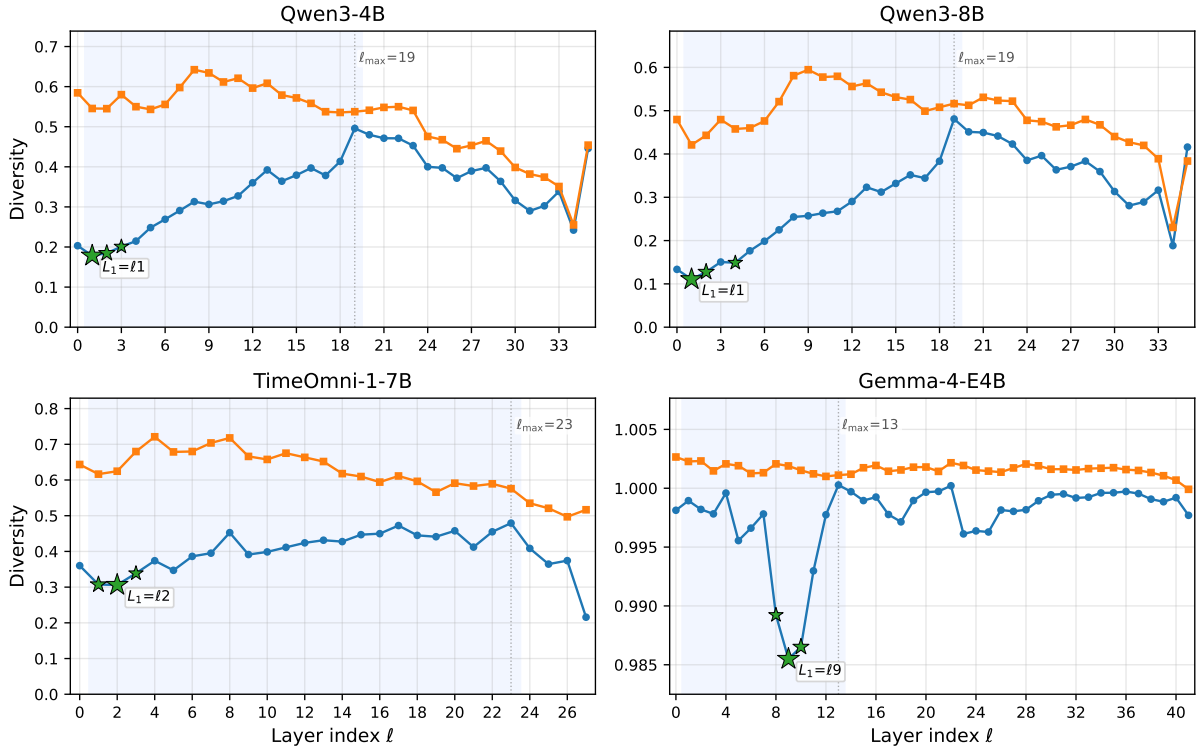


Figure 8: Cross-backbone diversity.

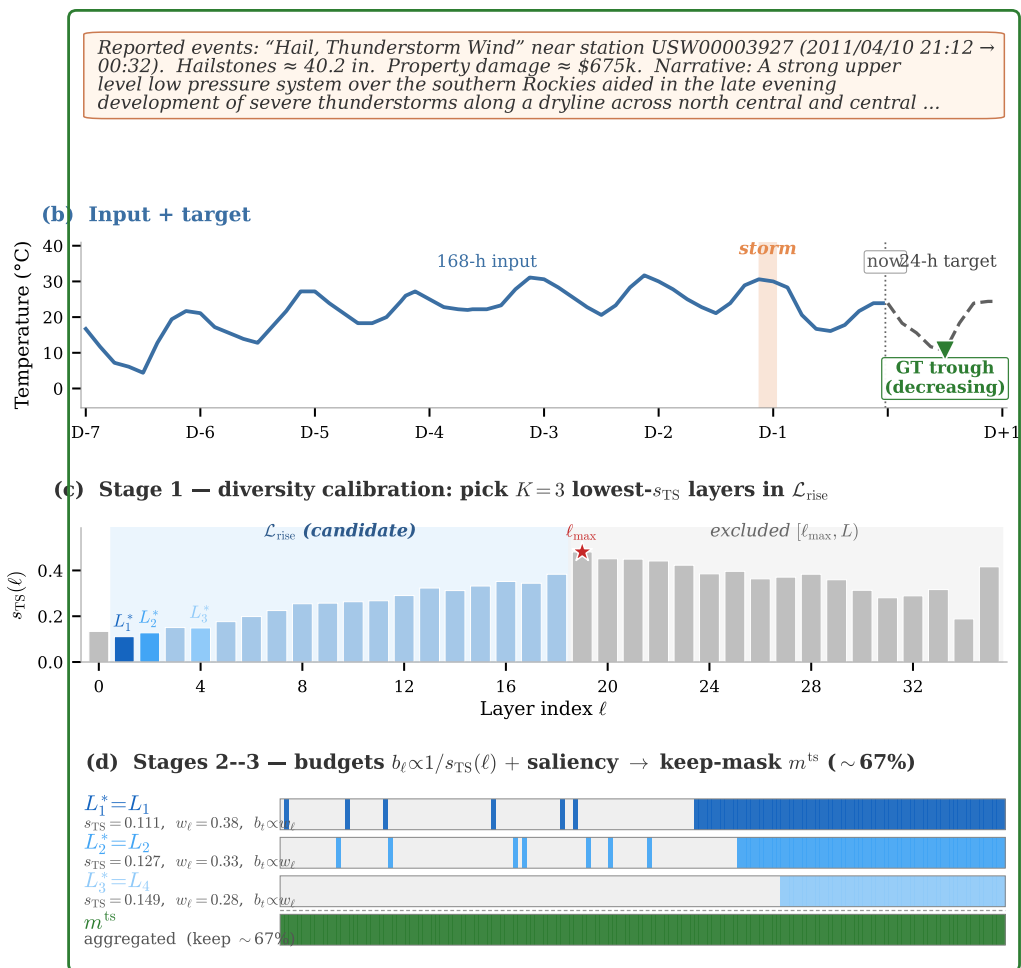


Figure 9: Case Study in Weather Domain

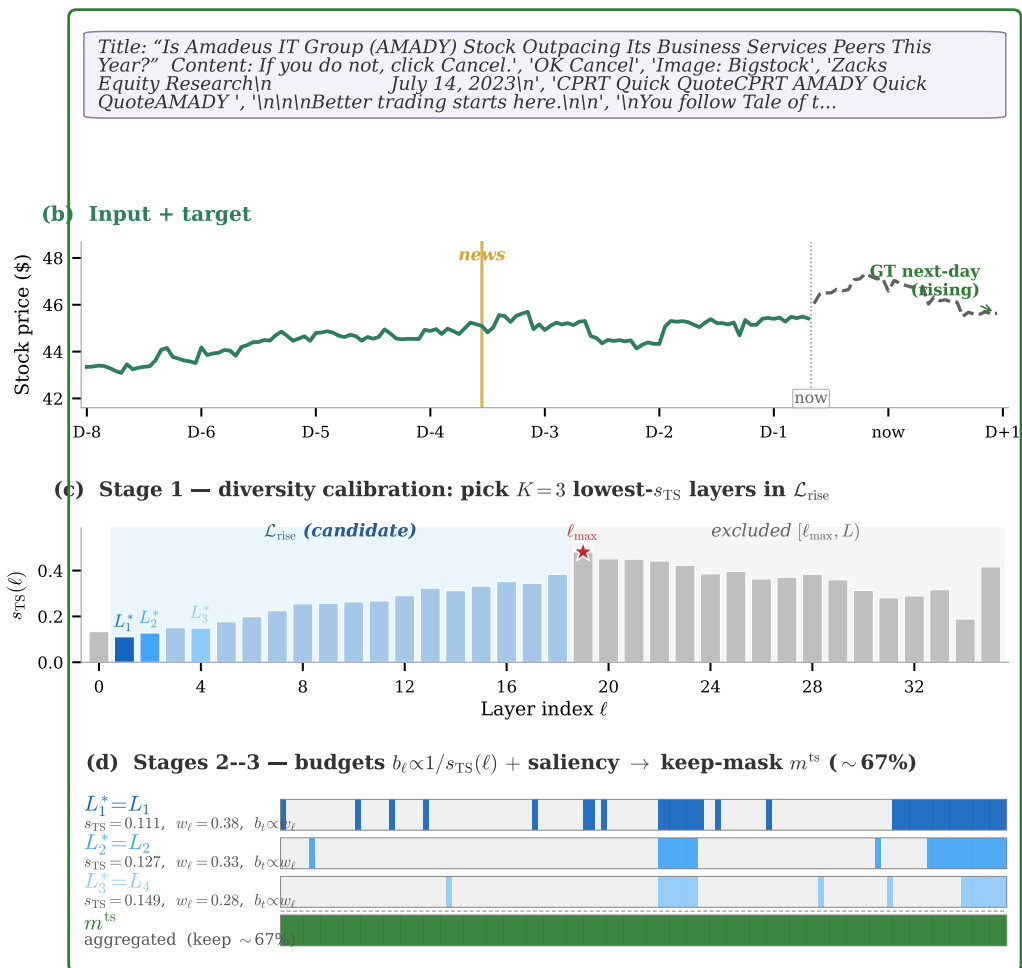


Figure 10: Case Study in Finance Domain