Rationale-Grounded In-Context Learning for Time Series Reasoning with Multimodal Large Language Models

Anonymous ACL submission

Abstract

The underperformance of existing multimodal large language models for time series reasoning lies in the absence of rationale priors that connect temporal observations to their downstream outcomes, which leads models to rely on superficial pattern matching rather than principled reasoning. We therefore propose the rationale-grounded in-context learning for time series reasoning, where rationales work as guiding reasoning units rather than post-hoc explanations, and develop the `RationaleTS` method. Specifically, we firstly induce label-conditioned rationales, composed of reasoning paths from observable evidence to the potential outcomes. Then, we design the hybrid retrieval by balancing temporal patterns and semantic contexts to retrieve correlated rationale priors for the final in-context inference on new samples. We conduct extensive experiments to demonstrate the effectiveness and efficiency of our proposed `RationaleTS` on three-domain time series reasoning tasks. We will release our code for reproduction.

1 Introduction

Time series reasoning is fundamental to decision making in ubiquitous real-world domains, such as air pollution warning (Cui et al., 2025), transportation management (Yu et al., 2024), and healthcare monitoring (Liu et al., 2023). The reasoning performance hinges on not only extrapolating historical trends, but also modeling the interaction among multiple variables over time and analyzing how temporal contexts correspond to future outcomes (Jiang et al., 2025). Thus, time series reasoning actually covers diverse tasks of prediction, classification, and anomaly detection, where the generated results must be supported by the evidence from the historical horizons and temporal contexts (Kong et al., 2025; Ni et al., 2025).

Recent advances in Multimodal Large Language Models (MLLMs) have motivated their use for time series reasoning by converting numerical sequences into visual charts, promising jointly perceiving temporal patterns and generating natural language explanations in a unified framework (Liu et al., 2025; Zhong et al., 2025; Wang et al., 2025). However, despite the improved modality alignment compared with converting numerical data into textual tokens in LLMs (Jin et al., 2023; Liu et al., 2024b; Chang et al., 2025), existing approaches tend to yield results and explanations solely on *superficial temporal similarity or local pattern matching*, hardly generating reliable and evidence-based reasoning (as shown in Figure 1).

The underlying reason of the above problem is not the insufficient accessed data or model capacity, but **the lack of explicit rationale priors empowering in-context learning for time series reasoning**. Thus, we propose *rationale-grounded in-context learning for time series reasoning*. Each rationale consists of structured reasoning paths, connecting the observable cross-variable coordination with specific downstream implications, which work as reasoning guides for in-context learning rather than post-hoc explanations of given outcomes. By grounding in-context reasoning on these rationales, MLLMs can deduce why particular temporal contexts lead to some outcomes, making MLLMs less prone to hallucinated or unjustified conclusions.

Given this insight, we introduce `RationaleTS`, a novel method that enhances time series reasoning ability of MLLMs with rationale-grounded in-context learning (Figure 1). We first induce ground truth-conditioned rationales to build reasoning paths between cross-variable observations and implications on specific outcomes. We then design a hybrid retrieval mechanism to retrieve guiding rationales for a given sample, balancing both temporal patterns and semantic contexts. Finally, we complement the in-context inference of MLLMs with these rationale priors for evidence-grounded outcome predictions and interpretations. Note that

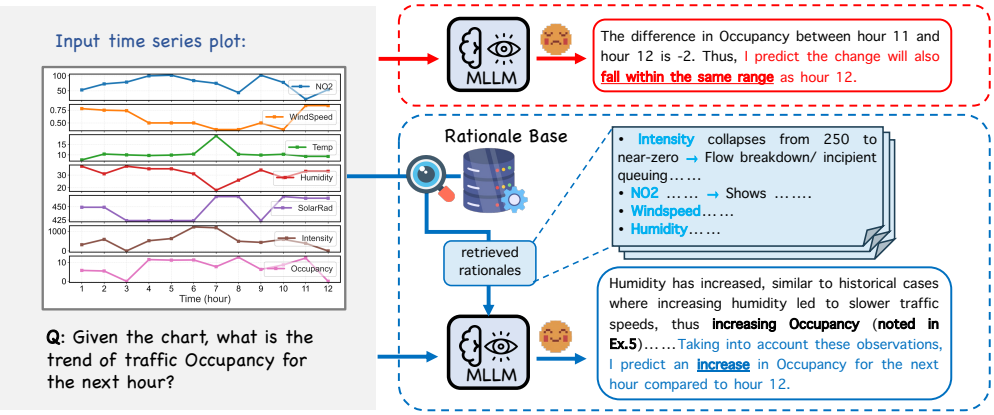

Figure 1: Comparison in time series reasoning paradigms with MLLMs (red part) and rationale-grounded in-context learning in RationaleTS (blue part). In MLLMs the prediction outcome is generated by pattern extrapolation, while in RationaleTS, rationales provide reasoning priors connecting observations and implications, for the in-context learning on new samples.

our method differs fundamentally from existing Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Jiang et al., 2023; Zheng et al., 2025) or exemplar-based In-Context-Learning (ICL) (Wei et al., 2022) prompting methods. RationaleTS retrieves guiding rationales, instead of referenced samples, labels, or factual knowledge, which provides prior reasoning paths for in-domain tasks. Our contributions are summarized as follows:

- We identify the key limitation of existing MLLMs for time series reasoning as the absence of rationale priors that connect temporal observations to their downstream outcomes.
- We introduce the rationale-grounded in-context learning for time series reasoning, which treats rationales as guiding reasoning units rather than post-hoc explanations.
- We propose RationaleTS, which induces label-consistent rationales and retrieves temporal-and-semantic similar rationale priors for in-context reasoning of MLLMs on new samples.
- The extensive experiments across three real-world time series reasoning datasets demonstrate that grounding reasoning on rationales promises improved effectiveness and efficiency.

2 Related Works

2.1 LLMs and MLLMs for Time Series

Given the reasoning and interpretability of LLMs, existing works have attempted to bring such ability to the time series community (Jin et al., 2023; Zhou et al., 2023; Gruver et al., 2023; Liu et al., 2024b; Xie et al., 2024; Wang et al., 2025; Zhong et al., 2025). Most of these works treat time series data as numerical sequences and try to tackle the problem of modality alignment by tokenization, reprogramming and prompt engineering (Rasul et al., 2023; Cheng et al., 2025; Ni et al., 2025). However, LLMs struggle with capturing series-level contexts due to limited horizon windows. MLLMs provide a promising way to visualize the numerical data, which can align the time steps of different variables (OpenAI et al., 2024; Comanici et al., 2025a; Liu et al., 2025). Recent works on MLLMs mainly concentrate on chart comprehension and value perceptions, instead of complicated temporal reasoning (Zhou and Yu, 2024; Zhang et al., 2025).

2.2 Rationale Generation

The existing works have suggested that explicit rationales can enhance the reasoning ability of LLMs, compared with just true answers (Wei et al., 2022; Zhang et al., 2024). Prior works propose to treat the generated rationales as supervision signals for training small models (Hsieh et al., 2023; Wang et al., 2023). For the label-conditioned rationale generation, we can track back to (Camburu et al., 2018), where human explanations are collected conditioned on gold labels. While in STaR (Zelikman et al., 2022), a bootstrapping approach is introduced, where the LLM generates rationales conditioned on the correct answer to enlarge the fine-tuning datasets for iteratively training itself. (Chen et al., 2023) automatically aligns the generated rationales with the correct answers and thus constructs the self-training datasets for small language models. These methods primarily aim to gen-

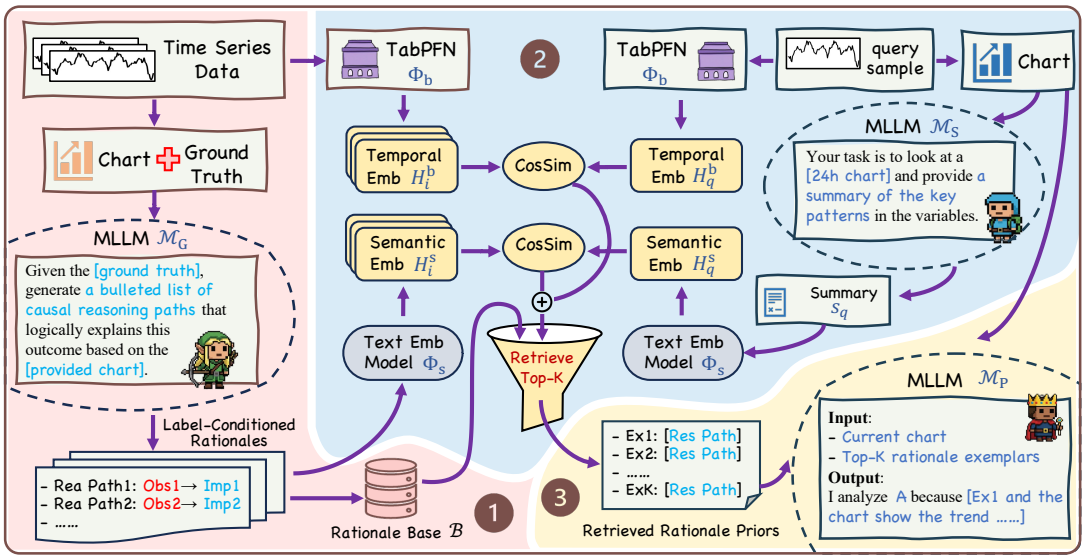

Figure 2: The workflow of RationaleTS, which includes ❶ Abductive Rationale Generation (§3.2.1), ❷ Hybrid Retrieval (§3.2.2), and ❸ Rationale-Grounded In-Context Inference (§3.2.3).

erate rationales for the downstream fine-tuning the model itself or student models (Shinn et al., 2023; Madaan et al., 2023; Liu et al., 2024a). While in our work, label-conditioned rationales include reasoning paths from the temporal variable changes to the implications. The rationales can constitute high-quality knowledge base for in-context learning rather than fine-tuning models.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{D}$ denote the time series dataset, where each sample $X_i \in \mathbb{R}^{T \times N}$ has T historical time steps and N variables. y_i may represent continuous values in future steps or the discrete future trend. Following (Jiang et al., 2025; Lee et al., 2025), we focus on the latter one and $y_i \in \mathbb{Z}_{\geq 0}$ denotes the discrete classification of a variable's future trend (e.g., 0: "increase"; 1: "decrease"; 2: "stable"). It is more beneficial to decision-critical applications, where early warnings rely on the evaluation of future trend direction, compared with accurate but uncertain continuous predictions. Instead of simple classification, we aim to excavate the synergistic effects of different variables from historical contexts and yield results and explanations via rationale-grounded in-context learning.

3.2 RationaleTS

The workflow of our proposed RationaleTS is shown in Figure 2. In the process of **abductive rationale generation**, we concatenate each time series chart and corresponding ground-truth labels to encourage the MLLM to generate hindsight reason paths, thus providing the guiding rationale base for downstream in-domain reasoning tasks. We then propose a **hybrid retrieval** mechanism to retrieve Top-K label-free rationale priors, which balances temporal patterns and semantic contexts, for the final rationale-grounded **in-context inference**.

3.2.1 Abductive Rationale Generation

The pretrained MLLMs have naive understanding of domain-specific knowledge and may result in hallucination issues, hardly learning the synergistic effects of variables and the contextual information for complicated reasoning. Inspired by (Zelikman et al., 2022), we propose an abductive rationale generation mechanism, where the MLLM is tasked with justifying the ground-truth results by constructing evidence-grounded reasoning paths.

For each time series sample X_i, we firstly obtain the corresponding visual chart, denoted as X_i^c. We concatenate X_i^c and y_i to encourage the pretrained MLLM $\mathcal{M}_G$ to generate label-conditioned rationales r_i, i.e., $r_i \leftarrow \mathcal{M}_G(X_i^c, y_i)$. The process can be seen as a conditional text generation task. The rationale r_i (supposing including $\mathcal{T}_i$ tokens) is generated autoregressively as:

$$P(r_i \mid X_i^c, y_i) = \prod_{t=1}^{\mathcal{T}_i} P_{\mathcal{M}_G}(r_{i,t} \mid X_i^c, y_i, r_{i,<t}).$$

The label-conditioned rationales are organized as a bulleted list of reasoning paths following the for-

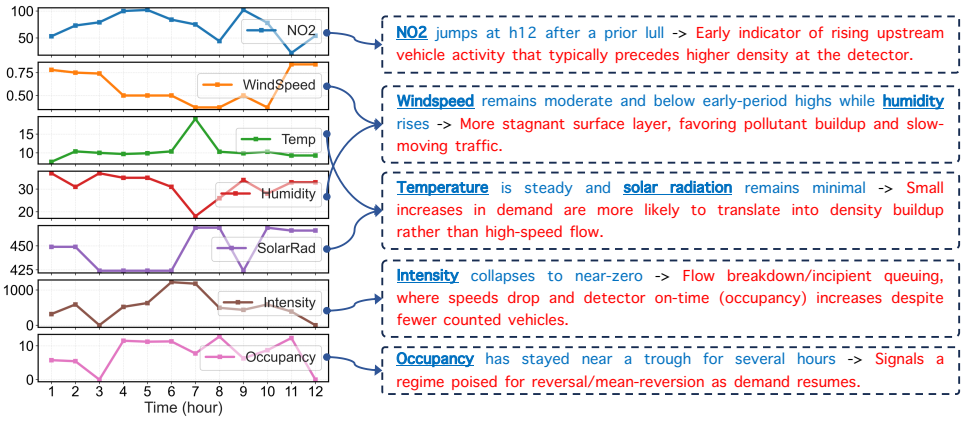

Figure 3: **Left**: Time series chart of a sample from Traffic dataset. **Right**: The generated rationales include 5 reasoning paths. Blue: Observations. Red: Implications. Each reasoning path provides the evidence-grounded analysis on implications to the final outcome.

mat of "`Observation` $\rightarrow$ `Implication`", which describe how different variables synergize towards future results, as shown in Figure 3. It is beneficial to construct a high-quality rationale database, i.e., $\mathcal{B} = \{r_i\}_{i=1}^{D}$ for in-context inference. Note that we avoid revealing the true labels in rationales to encourage principled reasoning rather than simple imitation (detailed results are provided in §4.3.1).

3.2.2 Hybrid Retrieval

The effectiveness of in-context inference largely hinges on the retrieved exemplars (rationales in our paper) (Brown et al., 2020; Alayrac et al., 2022; Wang et al., 2024a). In time series reasoning tasks, a natural way is to retrieve the rationales of time series with similar temporal patterns, which however lacks semantic contexts of observations. For example, the decrease in the density of PM 2.5 in some cases may result from either higher wind speed or seasonality. Besides, text summary provides coarse-grained abstraction but not quantitative details, which may aggravate hallucination issues. Thus, we propose a hybrid retrieval approach to unify the statistical priors and semantic contexts.

Data-Centric Similarity. The accuracy of time series analysis depends on how to model the synergies in different variables, while the existing time series foundation models mainly leverage the channel-independent strategy (Nie et al., 2022; Woo et al., 2024; Ansari et al., 2024). Given the effectiveness of TabPFN in time series forecasting (Hoo et al., 2024), we leverage the frozen representation power of TabPFN as a universal time-series encoder to obtain temporal embedding.

Specifically, we reorganize raw time series X_i into tabular data X_i^{b}, with each column and row corresponding to a variable and a time step respectively. Similarly, for a *query* [1] sample X_q, we have X_q^{b}. Let $\Phi_{\mathrm{b}}(\cdot)$ denote the TabPFN encoder. We can obtain the temporal embeddings as:

$$H_q^{\mathrm{b}} = \Phi_{\mathrm{b}}(X_q^{\mathrm{b}}), H_i^{\mathrm{b}} = \Phi_{\mathrm{b}}(X_i^{\mathrm{b}}). \quad (1)$$

H_q^{b} and H_i^{b} provide data priors on the coordination of temporal variables. We then compute the data-centric similarity as:

$$Sim_i^{\mathrm{b}} = \mathbf{cos}(H_q^{\mathrm{b}}, H_i^{\mathrm{b}}) = \frac{H_q^{\mathrm{b}} \cdot H_i^{\mathrm{b}}}{\left\|H_q^{\mathrm{b}}\right\| \left\|H_i^{\mathrm{b}}\right\|}. \quad (2)$$

Semantic-Centric Similarity. The generated factual rationales in §3.2.1 promise high-quality semantic contexts, including observations and potential effects. However, the proposed abductive rationale generation will not work for query samples without ground-truth labels. Therefore, we adopt an intermediate MLLM $\mathcal{M}_{\mathrm{S}}$ to generate the text summary of a query chart, which aligns with rationales in both modality and semantic space. Note that $\mathcal{M}_{\mathrm{S}}$ merely abstracts temporal changes, i.e., providing the "`Observations`" in reasoning paths. Compared with complicated reasoning, this is a simple task $\mathcal{M}_{\mathrm{S}}$ can handle.

Specifically, given a query chart X_q^{c}, the text summary is generated as $s_q \leftarrow \mathcal{M}_{\mathrm{S}}(X_q^{\mathrm{c}})$. Let $\Phi_{\mathrm{e}}(\cdot)$ denote the text embedding model. We can obtain the semantic embeddings and evaluate the semantic-centric similarity as:

$$H_q^{\mathrm{s}} = \Phi_{\mathrm{s}}(s_q), H_i^{\mathrm{b}} = \Phi_{\mathrm{b}}(r_i). \quad (3)$$

$$Sim_i^{\mathrm{s}} = \mathbf{cos}(H_q^{\mathrm{s}}, H_i^{\mathrm{s}}). \quad (4)$$

[1] We name a *test* sample a *query* here (An et al., 2025).

Table 1: Comparison of time series reasoning paradigms.

Method	Reasoning Unit	Reasoning Prior	In-Context Usage	Limitation
ICL	Exemplars	Implicit	Retrieved exemplars	Focusing on sample similarity
RAG	Retrieved knowledge	Implicit	Retrieved documents	Not for Time series reasoning
Rationale-supervised	Rationale-as samples	Fixed	/	Hardly connecting reasoning to rationales
`RationaleTS`	Rationales	**Explicit**, **Transferrable**	Retrieved rationales	—

Hybrid Fusion. To unify the statistical priors and semantic contexts, we combine the above similarity scores with a balancing factor λ:

$$Sim_i^{\text{final}} = \lambda \cdot Sim_i^{\text{b}} + (1-\lambda) \cdot Sim_i^{\text{s}}. \quad (5)$$

We construct $\mathcal{R}$ by retrieving rationales from $\mathcal{B}$ with top K highest hybrid similarity scores:

$$\mathcal{R} = \{r_i \mid i \in \arg\text{top-K}_{i\in[1,\dots,D]}\, Sim_i^{\text{final}}\}. \quad (6)$$

3.2.3 In-Context Inference

The retrieved explicit rationales empower the MLLM $\mathcal{M}_{\text{P}}$ to perform in-context inference, by transferring the logical deduction patterns from these reasoning priors to the new query sample. Specifically, we concatenate the query chart X_q^{c} and rationale set $\mathcal{R}$ as augmented input and the inference process of $\mathcal{M}_{\text{P}}$ is formulated as:

$$(\hat{r}_q, \hat{y}_q) = \arg\max_{r,y} P_{\mathcal{M}_{\text{P}}}(r, y \mid \mathcal{R}; X_q^{\text{c}}). \quad (7)$$

Typically, this is decomposed into a two-step generation process:

$$P_{\mathcal{M}_{\text{P}}}(r, y \mid \mathcal{R}; X_q^{\text{c}}) = \underbrace{P_{\mathcal{M}_p}(r \mid \mathcal{P})}_{\text{Reasoning Generation}} \underbrace{P_{\mathcal{M}_p}(y \mid \mathcal{P}, r)}_{\text{Final Result}}. \quad (8)$$

We bootstrap $\mathcal{M}_{\text{P}}$ to generate reasoning first and then results, which ensures the final inference logically consistent with the visual evidence and contextual knowledge from rationale priors. The detailed process of `RationaleTS` is provided in Appendix B.

3.3 Method Analysis

We compare different time series reasoning paradigms in Table 1. In ICL (Wang et al., 2024a) or RAG (Jiang et al., 2023; Zheng et al., 2025) paradigms, the models are augmented by retrieved units, which provides similar samples but implicit reasoning priors. While in rationale-supervised methods (Shinn et al., 2023; Madaan et al., 2023; Liu et al., 2024a), rationales work as supervised signals for fine-tuning models, where the reasoning process is not grounded on rationales. Our proposed `RationaleTS` goes beyond the above limitations, by grounding in-context reasoning on explicit rationales which contain reasoning paths from observations to implications.

4 Experiments

4.1 Experimental Settings

Datasets and Tasks. We evaluate the performance of `RationaleTS` on datasets of three domains: finance, transportation, and energy. These datasets all include multiple variables and pose complicated time series reasoning tasks. The datasets and the corresponding tasks are described as follows. More details are presented in Appendix C.

Finance includes the daily records of 9 financial indicators from January 2019 to December 2023 (Lee et al., 2025). Our task is to reason the S&P 500 in the next day will *increase by over 1%, decrease by over 1%, or remain stable* w.r.t the last day of a given 20-day period. **Traffic** includes the hourly records of 7 weather and transportation indicators from January 2019 to June 2019 (Iskandaryan et al., 2022). The task is to reason the occupancy of the next hour, w.r.t the last hour of a 12-hour period, will *increase by 2, decrease by 2, or remain stable*. **Power** includes 10-min records of 9 variables from meteorologic system and wind turbine SCADA in 2021 (Zhou et al., 2024). We aim to infer whether the average active power in the next 6 hours *will surpass* that of the past 24 hours.

Baselines. We conduct the comparison experiments with three types of baselines. **(1) Time series reasoning methods with LLMs**: Moirai (Woo et al., 2024), ChatTS (Xie et al., 2025), ChatTime (Wang et al., 2025), and TimeXL (Jiang et al., 2025); **(2) VL-Time** (Liu et al., 2025)

Table 2: Time series reasoning performance comparison. **Bold**: the best. Underline: the second best.

Dataset	Finance		Power		Traffic	
Metric	F1	AUC	F1	AUC	F1	AUC
Moirai	36.57	53.44	50.35	54.23	62.23	71.81
ChatTS	30.30	50.54	43.26	52.63	18.27	44.31
ChatTime	6.74	50.00	28.33	50.49	10.79	50.00
TimeXL	19.61	50.73	67.39	67.50	20.07	41.36
GPT-4o	55.65	66.73	68.02	67.83	37.93	53.45
GPT-5	66.53	74.90	69.50	69.32	61.38	71.03
gemini-2.0-flash	22.58	41.94	58.50	61.67	47.59	60.69
gemini-2.5-flash	28.40	50.00	41.00	50.00	48.97	61.72
qwen-vl-max	44.35	58.27	69.61	69.41	9.84	33.11
qwen3-vl-plus	43.55	57.66	70.60	70.15	13.79	35.34
textual	24.60	43.45	52.50	55.84	32.41	49.31
textual+CoT	30.59	50.27	66.50	64.17	37.93	53.45
textual+ICL	29.84	47.38	64.00	61.12	36.55	52.41
visual	45.16	58.87	63.00	55.62	40.69	55.52
visual+CoT	51.61	63.71	67.00	61.24	42.07	56.55
visual+ICL	62.50	71.88	65.50	62.02	44.14	58.10
RationaleTS	**69.76**	**77.32**	**71.50**	**72.87**	**66.21**	**74.66**

with different base MLLMs: GPT-4o (Hurst et al., 2024), GPT-5[2], gemini-2.0-flash (Team et al., 2023), gemini-2.5-flash (Comanici et al., 2025b), qwen-vl-max (Bai et al., 2025), and qwen3-vl-plus (Yang et al., 2025). **(3) We evaluate the performance of a same MLLM (`GPT-4o-mini`) in the textual and visual modality**. We also augment the MLLM with CoT (Wei et al., 2022) and In-Context Learning (ICL) (Wang et al., 2024a). The detailed prompts are provided in Appendix D.

Implementations. We employ `GPT-5` to generate rationales and `GPT-4o-mini` to generate text summary and perform the final prediction. We adopt `text-embedding-3-large` as the text embedding model. The datasets are divided with the ratio of 8:2. We construct rationale base on the 80% samples and perform in-context inference on the 20%. For fair comparison, we report the performance of `RationaleTS` and the zero-shot baselines on the 20% samples. We evaluate the performance of all methods with the widely-used F1 score and AUC. The parameters of K and λ are set as 5 and 0.8.

4.2 Main Results

Table 2 shows the numerical results of the proposed `RationaleTS` and different baselines. The baselines in Type (1) perform poorly in time series reasoning, where the tokenization of numerical data can hardly reserve the intrinsic temporal patterns, thus affecting the learning of coordination of different variables. In contrast, MLLMs have general better understanding of multi-variable time series. By comparing models of the same series, higher accuracy is obtained with higher edition, which indicates the scaling laws retain in MLLMs for time series reasoning (Kaplan et al., 2020). In baselines of Type (3), we respectively input the numerical data and visual plots to the same MLLM, i.e., `GPT-4o-mini`. The key motivation is that the visualization can augment the time series reasoning capability of MLLM, with the F1 score increasing by 13.11% on average. Moreover, on two modalities, both CoT and ICL can improve the reasoning performance. Our proposed `RationaleTS` outperforms on all datasets, indicating the the effectiveness of in-context inference, grounded on the high-quality rationale base and hybrid retrieval.

Table 3: Ablation results on Finance and Power datasets. **Bold**: the best. Underline: the second best.

Datasets	Finance		Power	
Variants	F1	AUC	F1	AUC
A.1 w/ chart	64.11	73.08	62.50	63.38
A.2 w/ label	64.92	73.69	68.00	67.49
A.3 w/ both	62.50	71.88	64.50	65.08
B.1 w/o data	64.11	70.38	52.50	58.82
B.2 w/o semantic	66.13	74.60	66.50	64.73
B.3 random	57.08	60.92	63.50	59.21
RationaleTS	**69.76**	**77.32**	**71.50**	**72.87**

4.3 Analysis

4.3.1 Ablation Study

The ablation results are reported in Table 3. In **A.1**-**A.3**, we integrate the visual charts, ground-truth labels, or both into rationales for in-context inference. Neither of the two can improve the reasoning performance. The disclosure of labels may induce the MLLM to directly copy the results, instead of decision after reasoning. On the other hand, the visual charts make the MLLM trapped into local pattern matching, which may provide the opposite evidence against true outcomes.

In **B.1** and **B.2**, we ablate the data-centric and semantic-centric similarity respectively. A key observation is that ablating either would decrease the effectiveness of retrieval and ultimately impact the reasoning performance. In **B.3**, we randomly select 5 most similar rationales, which decreases the F1 score by 12.68% and 8% respectively on

[2]https://cdn.openai.com/gpt-5-system-card.pdf

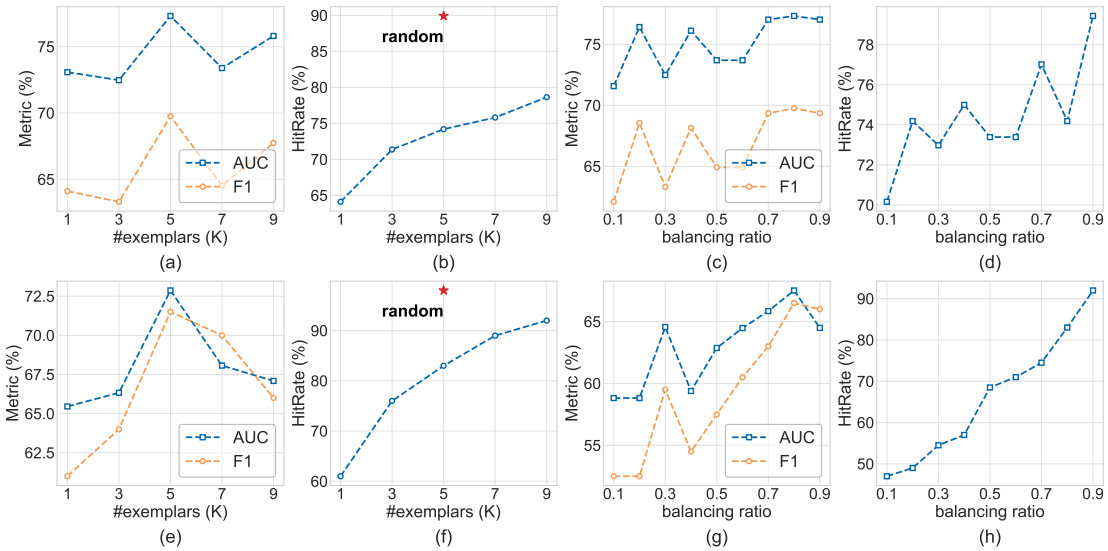

Figure 4: Hyperparamter analysis of K and λ on Finance ((a)-(d)) and Power ((e)-(h)) datasets.

two datasets. The outperformance of RationaleTS w.r.t **B.1-B.3** demonstrates the proposed hybrid retrieval mechanism can unify statistical priors and semantic contexts and retrieve high-quality rationales for in-context inference.

4.3.2 Sensitivity Investigation

We conduct the sensitivity investigation of the number of rationales K and balancing ratio λ on Finance and Power datasets, as shown in Figure 4. Besides AUC and F1 score, we adopt the $HitRate$ metric to evaluate the retrieval accuracy, which is computed as:

$$HitRate = \frac{\sum_q \mathbb{1}(\exists y_i = y_q, \forall r_i \in \mathcal{R})}{D_q}. \quad (9)$$

$\mathbb{1}(\exists y_i = y_q, \forall r_i \in \mathcal{R})$ denotes the indicator function, which represents whether at least one of the retrieved rationales has the same label with the query. D_q denotes the number of query samples.

As shown in Figure 4 (a) and (e), more rationales do not guarantee improved performance. Less rationales may not provide enough referenced knowledge to benefit reasoning, while more rationales mean more noise knowledge misleading the reasoning process. The performance is optimal when $K = 5$. As shown in Figure 4 (b) and (f), the HitRate increases with more rationales. The HitRate is much higher in the setting of random selection. However, as shown in Table 3, **B.3** underperforms RationaleTS a lot, which indicates that higher HitRate may not correspond to better performance.

Figure 4 (c), (d), (g), and (h) show that the three metrics have a general increasing trend with the balancing ratio λ, which indicates that data-centric similarity may play a more significant role. RationaleTS achieves the best on both datasets in the setting of $\lambda = 0.8$, which does not correspond to the highest HitRate, as shown in Figure 4 (d) and (h). Thus in hybrid retrieval, the semantic-centric similarity, on the other hand, can compensate for the loss of semantic contexts.

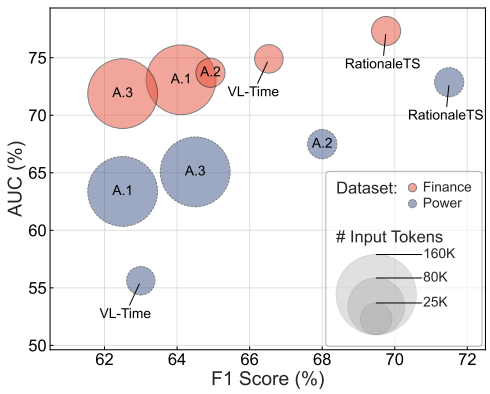

Figure 5: Efficiency analysis in terms of AUC, F1 score, and # Input Tokens on Finance and Power datasets.

4.3.3 Efficiency Analysis

We conduct efficiency analysis by comparing the F1 score, AUC and averaged number of input tokens of $\mathcal{M}_{\mathrm{P}}$. Figure 5 shows the comparison among RationaleTS, VL-Time and three variants (in Table 3). More tokens correspond to larger bubble size. In **A.1** and **A.3**, the visual charts are incorporated into the prompts of $\mathcal{M}_{\mathrm{P}}$, which increases the tokens by 5.7× compared with RationaleTS.

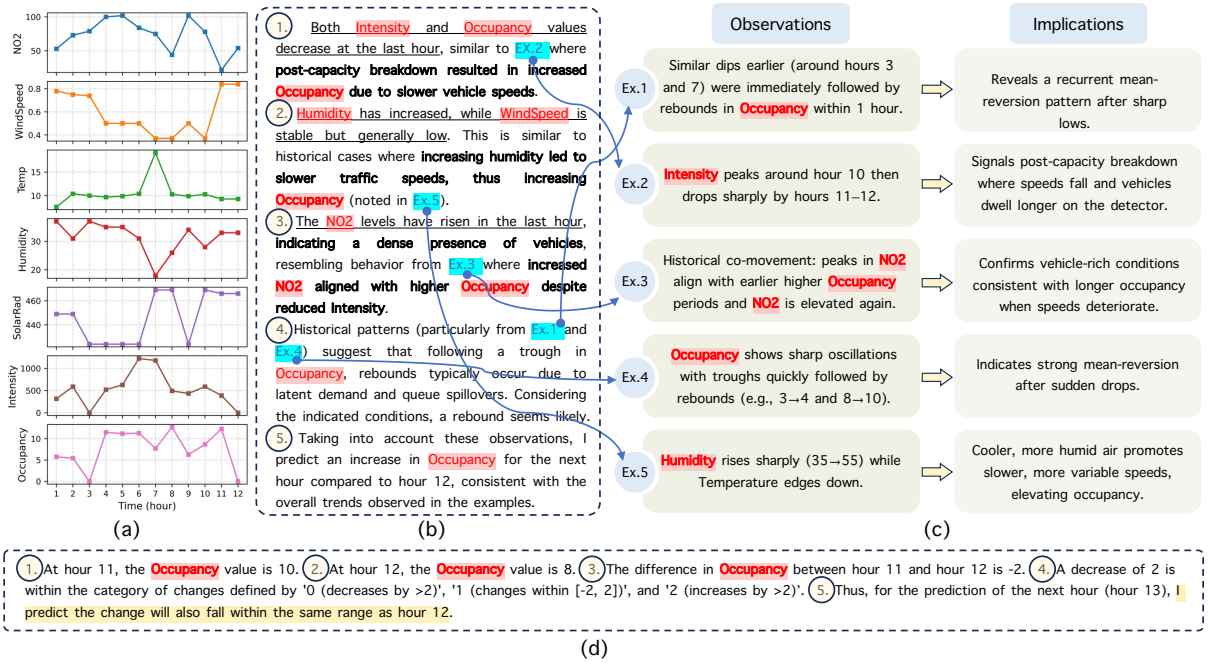

Figure 6: Case study on Traffic dataset. (a) Time series plot with 7 variables. (b) Time series reasoning of $\mathcal{M}_{\mathrm{P}}$. Underline: Observation form the chart. **Bold**: Potential conclusion after referencing rationales. (c) The guiding reasoning paths in retrieved rationales. (d) Reasoning process of VL-Time.

However, the averaged F1 score decreases by 7.23% on two datasets. In **A.2**, the ground-truth labels are included, which results in an averaged 4.17% F1 decrease. In VL-Time, $\mathcal{M}_{\mathrm{P}}$ perform zero-shot inference of each query, which has fewer tokens but 5.86% averaged F1 score drop on two datsests. Hence, compared with the four methods, the proposed `RationaleTS` is effective and efficient, with good balance between token usage and performance improvement.

4.4 Case Study

We provide the case analysis in Figure 6. We have a key observation from Figure 6 (b) that each reasoning step follows the process of ❶ summarizing the observations from the chart; ❷ seeking to specific reasoning paths from rationales for similar reasoning patterns; ❸ generating the potential conclusion on the future trend of `Occupancy`. This process benefits $\mathcal{M}_{\mathrm{P}}$ to resort to specific evidence and then produce the results, avoiding the arbitrary guess and direct imitation.

Beyond co-variables, the temporal patterns of `Occupancy` in the historical horizons are also analyzed in reasoning step 4. The underlying reasons for the fluctuation in `Occupancy` are also analyzed. Moreover, in reasoning step 5, the final conclusion is generated by considering the coordination of different variables, instead of merely depending on `Intensity`, which is most correlated with the targeted variable `Occupancy`. While in Figure 6(d), VL-Time merely focuses on the changes of `Occupancy`, instead of the coordination of different variables, and produce the results following the former time stamp. To conclude, the reasoning MLLM $\mathcal{M}_{\mathrm{P}}$ can perform effective rationale-grounded in-context inference by generating step-by-step reasoning with the process of observation, reference, and conclusion.

5 Conclusion and Future Work

In this paper, we identify the key limitation of MLLMs for time series reasoning tasks, namely the absence of rationale priors that connect temporal observations to their downstream implications. We thus introduce the rationale-grounded in-context learning and propose the `RationaleTS` method, which induces label-consistent rationales and retrieves temporal-and-semantic similar rationale priors for in-context reasoning on new query samples. Extensive experiments demonstrate the outperformance of `RationaleTS` on three reasoning tasks. In the future work, we will explore how to construct cross-domain rationale priors and improve the rationales' structures of reasoning paths to enable more accurate retrieval.

Limitations

The proposed method heavily depends on the generated reasoning paths in abductive rationale generation process, which are not further evaluated. Following existing works, this paper explores the problems of future trend prediction. More time series reasoning tasks should be considered in the future work.

Ethical Statement

This work employs publicly available multi-modal large language models as foundational components of the proposed framework. These models are used without additional training on private or proprietary data. All datasets involved in our experiments are obtained from public sources and do not contain personally identifiable information. Potential risks of this work include the misuse of forecasting results in automated decision-making systems. Our method is designed as a decision-support tool rather than a fully autonomous system. We adopt AI Assistants for polishing the original content, rather than for suggesting new content.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Kaikai An, Fangkai Yang, Liqun Li, Junting Lu, Sitao Cheng, Shuzheng Si, Lu Wang, Pu Zhao, Lele Cao, Qingwei Lin, and 1 others. 2025. Thread: A logic-based data organization paradigm for how-to question answering with retrieval augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18300–18319.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. 2024. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*. Expert Certification.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. 2025. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–20.

Wei-Lin Chen, An-Zi Yen, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. ZARA: Improving few-shot self-rationalization for small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4682–4693, Singapore. Association for Computational Linguistics.

Mingyue Cheng, Yiheng Chen, Qi Liu, Zhiding Liu, Yucong Luo, and Enhong Chen. 2025. Instructime: Advancing time series classification with multimodal language modeling. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 792–800.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025b. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. 2024. The tabular foundation model tabPFN

outperforms specialized time series forecasting models based on simple features. In *NeurIPS Workshop on Time Series in the Age of Large Models*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ditsuhi Iskandaryan, Francisco Ramos, and Sergio Trilles. 2022. Bidirectional convolutional lstm for the prediction of nitrogen dioxide in the city of madrid. *PloS one*, 17(6):e0269295.

Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. 2025. Timexl: Explainable multi-modal time series prediction with llm-in-the-loop. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and 1 others. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Yaxuan Kong, Yiyuan Yang, Shiyu Wang, Chenghao Liu, Yuxuan Liang, Ming Jin, Stefan Zohren, Dan Pei, Yan Liu, and Qingsong Wen. 2025. Position: Empowering time series reasoning with multimodal llms. *arXiv preprint arXiv:2502.01477*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18082–18090.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2024a. Chain of hindsight aligns language models with feedback. In *The Twelfth International Conference on Learning Representations*.

Haoxin Liu, Chenghao Liu, and B Aditya Prakash. 2025. A picture is worth a thousand numbers: Enabling llms reason about time series via visualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7486–7518.

Qingxiang Liu, Xu Liu, Chenghao Liu, Qingsong Wen, and Yuxuan Liang. 2024b. Time-ffm: Towards lm-empowered federated foundation model for time series forecasting. *Advances in Neural Information Processing Systems*, 37:94512–94538.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Jingchao Ni, Ziming Zhao, ChengAo Shen, Hanghang Tong, Dongjin Song, Wei Cheng, Dongsheng Luo, and Haifeng Chen. 2025. Harnessing vision models for time series analysis: A survey. *arXiv preprint arXiv:2502.08869*.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv e-prints*, pages arXiv–2211.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, and 1 others. 2023. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2025. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12694–12702.

Liang Wang, Nan Yang, and Furu Wei. 2024a. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, and 1 others. 2024b. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, pages 53140–53164. PMLR.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2024. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2025. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. In *Proceedings of the VLDB Endowment, 2025*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qi Yu, Weilong Ding, Hao Zhang, Yang Yang, and Tianpu Zhang. 2024. Rethinking attention mechanism for spatio-temporal modeling: A decoupling perspective in traffic flow prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3032–3041.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Yunkai Zhang, Yawen Zhang, Ming Zheng, Kezhen Chen, Chongyang Gao, Ruian Ge, Siyuan Teng, Amine Jelloul, Jinmeng Rao, Xiaoyuan Guo, and 1 others. 2025. Insight miner: A time series analysis dataset for cross-domain alignment with natural language. *arXiv preprint arXiv:2512.11251*.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. 2025. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*.

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

Jingbo Zhou, Xinjiang Lu, Yixiong Xiao, Jian Tang, Jiantao Su, Yu Li, Ji Liu, Junfu Lyu, Yanjun Ma, and Dejing Dou. 2024. Sdwpf: a dataset for spatial dynamic wind power forecasting over a large turbine array. *Scientific Data*, 11(1):649.

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

Zihao Zhou and Rose Yu. 2024. Can llms understand time series anomalies? *arXiv preprint arXiv:2410.05440*.

A More Related Works

A.1 Augmented Language Models

The augmented language models aim to tackle hallucination issues of LLMs or MLLMs by complementing them with external knowledge for improved reasoning ability. (Wei et al., 2022) proposes CoT prompting, encouraging models to generate intermediate reasoning steps. (Zhang et al., 2024) introduces Multimodal-CoT to vision-language domain. RAG typically retrieves documents or simple Question-Answer pairs rather than complicated logic flows (Lewis et al., 2020; Jiang et al., 2023; Salemi and Zamani, 2024; Wang et al., 2024b). Furthermore, standard retrieval struggles to align the statistical properties of time series with semantic reasoning. Our `RationaleTS` bridges the gaps by proposing a hybrid retrieval (integrating data priors and semantic contexts) to retrieve rationale priors, enabling in-context inference.

B Algorithm

Algorithm 1 present the process of `RationaleTS`, with the blue parts represent the construction of rationale base based on training datasets, which includes generating rationales (Line 2) and obtaining temporal and semantic embeddings (Line 3). In the inference phase, given a query sample, the temporal and semantic embeddings are firstly obtained (Line 5-6). We then incorporate statistical priors and semantic contexts to retrieve K rationales with top similarity scores (Line 7-11), based on which the in-context inference is conducted to generate the underlying reasons and results (Line 12-13).

Algorithm 1: `RationaleTS`

Input: Dataset $\{X_i, X_i^{\mathrm{c}}, X_i^{\mathrm{b}}, y_i\}_{i=1}^{D}$; Pretrained MLLMs $\mathcal{M}_{\mathrm{G}}, \mathcal{M}_{\mathrm{S}}$, and $\mathcal{M}_{\mathrm{P}}$; λ; Query: $(X_q, X_q^{\mathrm{c}}, X_q^{\mathrm{b}})$

Output: $\hat{y}_q$ and $\hat{r}_q$

1 **for** $i \in [1, D]$ **do**
2 $\quad r_i \leftarrow \mathcal{M}_{\mathrm{G}}(X_i^{\mathrm{c}}, y_i)$
3 $\quad H_i^{\mathrm{b}} = \Phi_{\mathrm{b}}(X_i^{\mathrm{b}}),\ H_i^{\mathrm{s}} = \Phi_{\mathrm{s}}(r_i)$
4 *// In Inference Phase*
5 $s_q \leftarrow \mathcal{M}_{\mathrm{S}}(X_q^{\mathrm{c}})$
6 $H_q^{\mathrm{b}} = \Phi_{\mathrm{b}}(X_q^{\mathrm{b}}),\ H_q^{\mathrm{s}} = \Phi_{\mathrm{s}}(s_q)$
7 **for** $i \in [1, D]$ **do**
8 $\quad Sim_i^{\mathrm{b}} = \mathbf{cos}(H_q^{\mathrm{b}}, H_i^{\mathrm{b}})$
9 $\quad Sim_i^{\mathrm{s}} = \mathbf{cos}(H_q^{\mathrm{s}}, H_i^{\mathrm{s}})$
10 $\quad Sim_i^{\mathrm{final}} = \lambda \cdot Sim_i^{\mathrm{b}} + (1-\lambda) \cdot Sim_i^{\mathrm{s}}$
11 $\mathcal{R} = \{r_i \mid i \in \arg\text{top-K}_{i\in[1,\ldots,D]}\, Sim_i^{\mathrm{final}}\}$
12 $\hat{r}_q, \hat{y}_q \leftarrow \mathcal{M}_{\mathrm{P}}(\mathcal{R}, X_q^{\mathrm{c}})$
13 **return** $\hat{r}_q$ and $\hat{y}_q$

C Details of Datasets

We evaluate on three benchmarks covering the domains of finance, transportation, and energy. The details of the datasets are presented in Table 4. We provide the details in the following parts.

Finance (Lee et al., 2025): This dataset includes 9 indicators: **`S&P 500`, `VIX`, `Nikkei 225`, `FTSE 100`, `Gold Futures`, `Crude Oil Futures`, `EUR/USD`, `USD/JPY`, and `USD/CNY`**. As shown in Figure 7, `S&P 500` correlates with the other variables, especially `Nikkei 225`, `Gold Futures`, and `Crude Oil Futures`.S Therefore, we analyze the future trend of `S&P 500` based on the historical contexts of all nine variables. Specifically, we analyze whether the indicator of `S&P 500` in the next one day will *decrease by over 1 % (labeled as 0), remain stable (labeled as 1), or increase by over 1 % (labeled as 2)* w.r.t the last day of a given 20-day period.

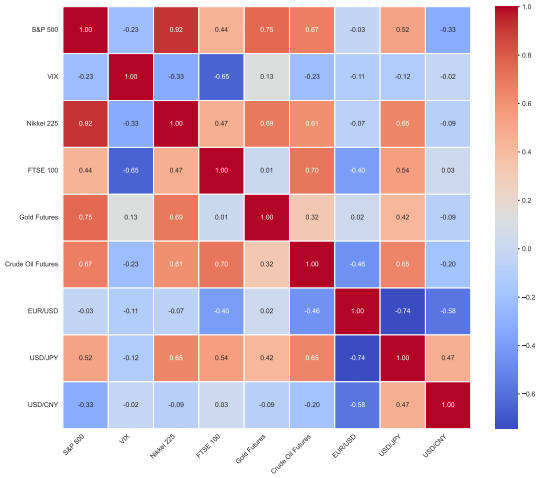

Figure 7: Correlation Matrix of 9 variables in Finance dataset.

Traffic (Iskandaryan et al., 2022): This dataset includes 5 weather indicators: **`NO2`, `WindSpeed`, `Temperature`, `Humidity`, and `SolarRad`**; and 2 transportation indicators: **`Intensity` and `Occupancy`**. The traffic `Intensity` evaluates the number of vehicles per hour, while the `Occupancy` indicates the proportion of time that road detectors are occupied by vehicles in an hour, which reflects the traffic jam level. We illustrate the correlation matrix of these 7 variables in Figure 8. The 5

Table 4: Details of three benchmarking datasets.

Dataset	Frequency	#Variables	#Time Stamps	Duration	#Samples	Label Distribution
Finance	1-D	9	1258	2019/1–2023/12	1238	13.78% / 17.04% / 69.18%
Traffic	1-H	7	4344	2019/1–2019/6	722	14.95% / 52.22% / 32.83%
Power	10-Min	9	49760	2021/1–2021/12	997	42.05% / 57.95%

weather indicators may affect traffic needs and thus correlated with the `Intensity` and `Occupancy`. In this paper, we analyze the `Occupancy` of the next hour, w.r.t the last hour of a 12-hour period, will *decrease by 2 (labeled as 0), remain stable (labeled as 1), or increase by 2 (labeled 2).*

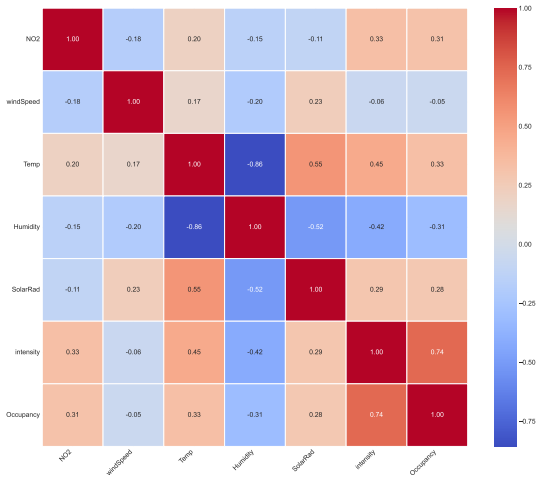

Figure 8: Correlation Matrix of 7 variables in Traffic dataset.

Power (Zhou et al., 2024): This dataset includes wind speed measurements (**`Wspd`, `Wspd_w`**), environmental and internal temperatures (**`Etmp`, `Itmp`**), blade pitch angles (**`Pab1`, `Pab2`, and `Pab3`**), rotor speed (**`Sp`**), and active power output (**`Patv`**), collectively characterizing the aerodynamic input, control actions, mechanical state, and power generation of wind turbines. Figure 9 shows the correlation matrix of these 9 variables. In this paper, we aim to analyze whether the average active power `Patv` in the next 6 hours will *surpass that of the past 24 hours (labeled as 1) or not (labeled 0).*

D Details of Prompts

We provide the prompts of the adopted three MLLMs $\mathcal{M}_{\mathrm{G}}$, $\mathcal{M}_{\mathrm{S}}$, and $\mathcal{M}_{\mathrm{P}}$ as follows. Moreover, we also present the prompts for baselines in Type (3).

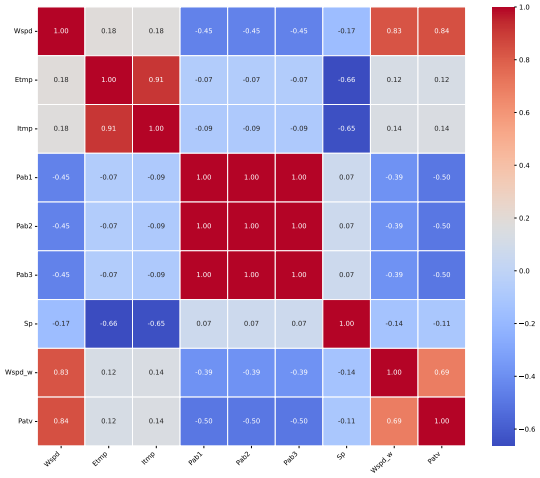

Figure 9: Correlation Matrix of 9 variables in Power dataset.

1. Prompt in $\mathcal{M}_{\mathrm{G}}$ for abductive rationale generation

System Prompt:
You are a senior [`specific domain`, *e.g., traffic and urban*] analyst. Given the actual outcome, your task is to generate a concise, 'gold-standard' causal reasoning path that logically explains this outcome based on the provided [`specific domain`, *e.g., traffic*] chart. This path will be used for a retrieval system. **Do not mention the actual outcome or the final prediction in your reasoning text.**
User Prompt:
The actual outcome for the next [future windows] was: **{true_label_meaning}**.
Please provide the ideal reasoning path that explains this outcome based on the attached [historical windows] data chart.
Your Task
Provide a bulleted list of key causal factors. Each bullet point must follow the format: 'Observation -> Implication'. Focus on describing the *dynamics* and *patterns*.

2. Prompt in $\mathcal{M}_{\mathrm{S}}$ for generating text summary

System Prompt:
You are a concise [`specific domain`, *e.g., traffic*] data analyst. Your task is to look at a [`historical windows`, *e.g., 12-hour*] [`specific domain`, *traffic*] chart and provide a brief, factual summary of the most prominent patterns.
User Prompt:
Analyze the attached [`historical windows`, *e.g., 12-hour*] [`specific domain`, *e.g., traffic*] data chart. Provide a one-paragraph summary describing the key trends you observe in variables. Be factual and objective.

3. Prompt in $\mathcal{M}_{\mathrm{P}}$ for in-context inference

System Prompt:
You are a world-class [`specific domain`, *e.g., wind power generation*] expert.
You will be given a new [`historical windows`, *e.g., 24-hour*] data chart and several relevant historical reasoning paths.
Your task is to first study the historical examples, then analyze the new chart, and finally analyze the [`targeted variable`, *e.g., power output*] trend for the next [`future windows`, *e.g., 6 hours*].
User Prompt:
Here are some relevant historical reasoning paths:
`{examples}`
Your Task
Now, analyze the **`new attached chart`**. Based on your analysis of this new chart AND the patterns learned from the historical examples, predict whether the [`targeted variable`, *e.g., average active power ('Patv')*] in the next [`future windows`, *6 hours*] will [`specific reasoning task`, *be higher than the average of the past 24 hours*]. Categorize your prediction as [`discrete labels and meanings`, *e.g., 0 (decrease by more than 1%), 1 (remain neutral (i.e., between -1% and 1%)), or 2 (increase by more than 1%)*].
Provide your answer in a valid JSON format with 'reasoning' and 'prediction' keys. Your 'reasoning' should be a step-by-step analysis that explicitly references both the new chart's data and the logic from the provided examples.

4. Prompt for zero-shot inference in textual modality

System Prompt:
You are a world-class [`specific domain`, *e.g., traffic*] expert.
You will be given [`historical windows`, *e.g., 12-hour*] [`specific domain`, *e.g., traffic and environment*] data.
Your task is to analyze the data and predict the [`targeted variable`, *e.g., Occupancy*] trend for the next [`future windows`, *e.g., hour*].
User Prompt:
Time-Series Data
Here is the [`historical windows`, *e.g., 12-hour*] data for a specific location:
[time series data]
Your Task
Analyze the provided data. Predict the change in [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*] compared to the last hour in the data. Categorize your prediction as [`discrete labels and meanings`, *e.g., 0 (decreases by >2), 1 (changes within [-2, 2]), or 2 (increases by >2)*] .
Provide your answer in a valid JSON format with 'reasoning' and 'prediction' keys. Your 'reasoning' should be a step-by-step analysis of the data.

5. Prompt for ICL inference in textual modality

System Prompt:
You are a world-class [`specific domain`, *e.g., traffic*] expert.
You will be shown several examples of [`historical windows`, *e.g., 12-hour*] data, each paired with its correct label indicating the [`targeted variable`, *e.g., Occupancy*] change for the next [`future windows`, *e.g., hour*]. Your task is to learn the patterns from these examples and then predict the change for new, unseen data.
User Prompt:
Analyze the following examples. Each example consists of time-series data and its corresponding label for the [`targeted variable`, *e.g., Occupancy*] change.
[Example i: time series data; label meanings]
Your Task
Now, analyze the **new data** below. Based on the patterns you observed in the examples, predict the change in [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*]. Categorize your prediction as [`discrete labels and meanings`, *e.g., 0 (decreases by >2), 1 (changes within [-2, 2]), or 2 (increases by >2)*].
New Data
[time series data]
Provide your answer in a valid JSON format with 'reasoning' and 'prediction' keys. Your reasoning should be a step-by-step analysis of the new data, drawing parallels to the provided examples where applicable.

6. Prompt for CoT inference in textual modality

System Prompt:
You are a world-class [`specific domain`, *e.g., traffic*] expert.
You will be given [`historical windows`, *e.g., 12-hour*] [`specific domain`, *e.g., traffic*] data. Your task is to analyze the data and predict the [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*].
User Prompt:
Time-Series Data
Here is the [`historical windows`, *e.g., 12-hour*] data for a specific location:
[time series data]
Your Task
Analyze the provided data. Predict the change in [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*] compared to the last hour in the data. Categorize your prediction as [`discrete labels and meanings`, *e.g., 0 (decreases by >2), 1 (changes within [-2, 2]), or 2 (increases by >2)*].
Please provide the ideal reasoning path that explains your prediction based on the provided data, following the format: 'Observation -> Implication'.
Provide your answer in a valid JSON format with 'reasoning' and 'prediction' keys. Your 'reasoning' should be a step-by-step analysis of the data.

7. Prompt for zero-shot inference in visual modality

System Prompt:
You are a world-class [`specific domain`, *e.g., traffic*] expert.
You will be given [`historical windows`, *e.g., 12-hour*] [`specific domain`, *e.g., traffic and environment*] data chart.
Your task is to analyze the chart and predict the [`targeted variable`, *e.g., Occupancy*] trend for the next [`future windows`, *e.g., hour*].
User Prompt:
Your Task
Analyze the **Attached Chart**. Predict the change in [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*] compared to the last hour in the data. Categorize your prediction as [`discrete labels and meanings`, *e.g., 0 (decreases by >2), 1 (changes within [-2, 2]), or 2 (increases by >2)*] .
Provide your answer in a valid JSON format with 'reasoning' and 'prediction' keys. Your 'reasoning' should be a step-by-step analysis of the chart.

8. Prompt for ICL inference in visual modality

System Prompt:
You are a world-class [`specific domain`, *e.g., traffic*] expert.
You will be shown several examples of [`historical windows`, *e.g., 12-hour*] data chart, each paired with its correct label indicating the [`targeted variable`, *e.g., Occupancy*] change for the next [`future windows`, *e.g., hour*]. Your task is to learn the patterns from these examples and then predict the change for new, unseen chart.
User Prompt:
Analyze the following examples. Each example consists of a chart and its corresponding label for the [`targeted variable`, *e.g., Occupancy*] change.
[Example i: time series chart; label meanings]
Your Task
Now, analyze the **Attached Chart**. Based on the patterns you observed in the examples, predict the change in [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*]. Categorize your prediction as [`discrete labels and meanings`, *e.g., 0 (decreases by >2), 1 (changes within [-2, 2]), or 2 (increases by >2)*].
Provide your answer in a valid JSON format with 'reasoning' and 'prediction' keys. Your reasoning should be a step-by-step analysis of the new chart, drawing parallels to the provided examples where applicable.

9. Prompt for CoT inference in visual modality

System Prompt:
You are a world-class [`specific domain`, *e.g., traffic*] expert.
You will be given [`historical windows`, *e.g., 12-hour*] [`specific domain`, *e.g., traffic*] data chart. Your task is to analyze the chart and predict the [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*].
User Prompt:
Your Task
Analyze the **Attached Chart**. Predict the change in [`targeted variable`, *e.g., Occupancy*] for the next [`future windows`, *e.g., hour*] compared to the last hour in the chart. Categorize your prediction as [`discrete labels and meanings`, *e.g., 0 (decreases by >2), 1 (changes within [-2, 2]), or 2 (increases by >2)*].
Please provide the ideal reasoning path that explains your prediction based on the attached chart, following the format: 'Observation -> Implication'.
Provide your answer in a valid JSON format with 'reasoning' and 'prediction' keys. Your 'reasoning' should be a step-by-step analysis of the chart.