Semantics-Enhanced Retrieval-Augmented Time Series Forecasting

Anonymous Authors[1]

Abstract

Time series forecasting models often benefit from historical patterns. Inspired by Retrieval-Augmented Generation (RAG), recent research explored retrieving relevant historical time series segments to enhance forecasting. However, relying solely on time series similarity is often insufficient for retrieval under non-stationarity. To address this, we propose a multimodal approach: a **S**emantics-**E**nhanced **R**etrieval-**A**ugmented Time Series **F**orecasting framework, SERAF. Unlike mainstream approaches that depend only on time series similarity, SERAF conducts dual retrieval over the time series and their self-generated textual descriptions. It retrieves two complementary sets of historical patterns and corresponding futures, which are selectively and jointly used to guide future predictions. Experiments across seven real-world datasets demonstrate the effectiveness of SERAF in bridging numerical and semantic views of time series compared with state-of-the-art baselines.

1. Introduction

Multivariate time series forecasting predicts future trajectories from historical observations and is central to traffic (Lippi et al., 2013), energy (Daut et al., 2017), finance (Poon & Granger, 2003), and climate (Price et al., 2025). Methods have progressed from classical statistical models such as ARIMA (Box et al., 2015) to deep forecasters such as DLinear, a robust MLP-based linear model (Zeng et al., 2022), and PatchTST, which uses patch-level time series representations (Nie et al., 2025). Recently, LLM-based methods such as Time-LLM (Jin et al., 2023) and GPT4TS (Zhou et al., 2023) have used textual context to inject background knowledge into forecasting.

The most widely adopted approach for supporting context-aware generation is retrieval-augmented generation (RAG), which retrieves relevant documents from large external databases and has become a key component of modern LLM pipelines (Lewis et al., 2020). Motivated by RAG, recent studies on time series forecasting have explored retrieval-based approaches that construct historical databases to retrieve similar patterns, thereby explicitly leveraging the entire history to guide future prediction. Representative efforts along this direction include RAFT, introducing multi-periodicity for retrieval-based forecasting (Han et al., 2025); TimeRAG, integrating retrieved sequences into LLM-based forecasters (Yang et al., 2025); TRACE, aligning external text with time series for multimodal retrieval (Chen et al., 2025); TS-RAG, augmenting time series foundation models (TSFMs) via adaptive retrieval (Ning et al., 2025); and TimeRAF, incorporating channel prompting into retrieval-augmented TSFMs (Zhang et al., 2025).

However, for challenging non-stationary time series, most retrieval methods remain confined to time series similarity (Han et al., 2025; Yang et al., 2025; Ning et al., 2025; Zhang et al., 2025), while multimodal approaches often rely on large external text corpora or LLM-generated descriptions, leading to inefficiency and limited scalability (Chen et al., 2025). To address this gap, we propose SERAF, a **S**emantics-**E**nhanced **R**etrieval-**A**ugmented Time Series **F**orecasting framework. Beyond retrieving relevant patterns via TS similarity, SERAF introduces a semantic retrieval module based on textual descriptions generated directly from time series segments. SERAF performs retrieval from both temporal and semantic perspectives, and adaptively fuses the retrieved results to enhance forecasting, all within a lightweight pipeline that requires no external annotations or domain-specific texts.

This is useful because two time series segments may differ in raw scale or local shape while sharing high-level attributes such as season, trend, and volatility. By indexing these attributes as text, SERAF can retrieve semantically relevant historical futures that are not necessarily nearest under TS similarity, complementing numerical retrieval without relying on external text resources.

Our main contributions are summarized as follows:

- We propose a novel semantic retrieval strategy based

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

on textual descriptions, enriching the retrieval process beyond purely numerical time series similarity.

- The textual descriptions are automatically generated from time series segments, requiring no external texts, thus ensuring efficiency and scalability.

- Extensive experiments on seven real-world datasets demonstrate SERAF's forecasting improvements.

2. Method

2.1. Problem Formulation

Given a historical input sequence $\mathbf{X}_{t-L+1:t} = \{X_{t-L+1}, X_{t-L+2}, \ldots, X_t\} \in \mathbb{R}^{L\times C}$, where L is the look-back window and C is the number of channels, the goal is to predict $\mathbf{Y}_{t+1:t+H} = \{Y_{t+1}, Y_{t+2}, \ldots, Y_{t+H}\}$ over the next H time steps.

2.2. Overview

As shown in Figure 1, SERAF performs retrieval-augmented forecasting from both temporal and semantic information. Given an input time series, a trainable encoder produces a naive prediction, while TS-similar historical segments are retrieved from a time series database. In parallel, a textual description of the input is embedded by a frozen text model to retrieve semantically similar descriptions from an aligned description database. The retrieved futures from both modalities are Gaussian-weighted, fused, gated with the naive prediction, and projected to produce the final prediction. This semantic retrieval dimension enriches the search space beyond temporal patterns and improves forecasting.

2.3. TS Database and TS Description Database

We construct the Time Series (TS) database by sliding a length-L window over the training set with stride 1 for dense historical coverage. Each segment is paired with its future sequence of length H. The TS database is $D_T = \{(\mathbf{P}_T^i, \mathbf{F}_T^i)\}_{i=1}^N$, where each pair contains a historical segment $\mathbf{P}_T^i$ and its future $\mathbf{F}_T^i$.

In parallel, we build a TS Description Database by generating a natural language description for each pair in D_T with a predefined template. As illustrated in Figure 2, each description includes time period, season, main trend, and main volatility. Time period and season come from timestamps, while trend and volatility use the most frequent channel-level pattern. Trend is categorized as upward, downward, or stable, and volatility as high, medium, or low. We denote the database as $D_S = \{\mathbf{Q}_S^i\}_{i=1}^N$, where each description Q_S^i aligns with $\mathbf{P}_T^i$ in D_T.

2.4. Retrieval from Time Series Similarity

For TS retrieval, given the j-th input sequence $\mathbf{X}^j$, we compute the similarity score ρ_{ij} between $\mathbf{X}^j$ and each historical segment $\mathbf{P}_T^i$ in D_T using a similarity function sim:

$$\rho_{ij} = sim(\mathbf{X}^j, \mathbf{P}_T^i), \quad i \in [1, N]. \tag{1}$$

We use Pearson's correlation because it reduces the effects of scale variation and value shifts while emphasizing monotonic trends (Han et al., 2025). To avoid leakage, historical segments in D_T that overlap with the input are excluded during training. The valid index set is:

$$\mathcal{I}_{\text{valid}} = \begin{cases} \{\, i \in [1, N] \mid i \notin [\max(1, j-(L+H-1)), \\ \quad \min(N, j+(L+H-1))] \,\}, \text{ if training}, \\ \{\, i \in [1, N] \,\}, \text{ otherwise}. \end{cases} \tag{2}$$

The Top-K index set $\mathcal{K}_T^j$ and retrieval set $\mathcal{R}_T^j$ are:

$$\mathcal{K}_T^j = \text{Top-}K\{\rho_{ij} \mid i \in \mathcal{I}_{\text{valid}}\}, \ \ |\mathcal{K}_T^j| = K, \tag{3}$$

$$\mathcal{R}_T^j = \{\, (\mathbf{P}_T^k, \mathbf{F}_T^k) \mid k \in \mathcal{K}_T^j \,\}. \tag{4}$$

The K retrieved segments are weighted by a Gaussian kernel, assigning larger weights to higher similarities:

$$\alpha_T^{kj} = \frac{\exp\left(-\frac{(1-\rho_{kj})^2}{2\tau^2}\right)}{\sum\limits_{i\in\mathcal{K}^j} \exp\left(-\frac{(1-\rho_{ij})^2}{2\tau^2}\right)}, \tag{5}$$

where τ is the Gaussian bandwidth.

The TS-similarity retrieved future is:

$$\hat{\mathbf{F}}_T^j = \sum_{k\in\mathcal{K}_T^j} \alpha_T^{kj}\, \mathbf{F}_T^k. \tag{6}$$

2.5. Retrieval from Semantic Similarity

For semantic retrieval, we generate a textual description $\mathbf{Q}^j$ for input $\mathbf{X}^j$ using the same template as in Figure 2. This extracts attributes from the time series itself without external text. A frozen text embedding model encodes $\mathbf{Q}^j$ into $\mathbf{E}^j$ and each historical description $\mathbf{Q}_S^i$ into $\mathbf{E}_S^i$. Semantic similarity is computed as:

$$s_{ij} = sim(\mathbf{E}^j, \mathbf{E}_S^i), \quad i \in [1, N], \tag{7}$$

where sim is cosine similarity. Analogous to Equation 4, we retrieve Top-K TS pairs and define:

$$\mathcal{R}_S^j = \{\, (\mathbf{P}_T^k, \mathbf{F}_T^k) \mid k \in \mathcal{K}_S^j \,\}, \tag{8}$$

where $\mathcal{K}_S^j$ is the Top-K index set. The retrieved segments are weighted as in Eq. 5, giving:

$$\hat{\mathbf{F}}_S^j = \sum_{k\in\mathcal{K}_S^j} \alpha_S^{kj}\, \mathbf{F}_T^k. \tag{9}$$

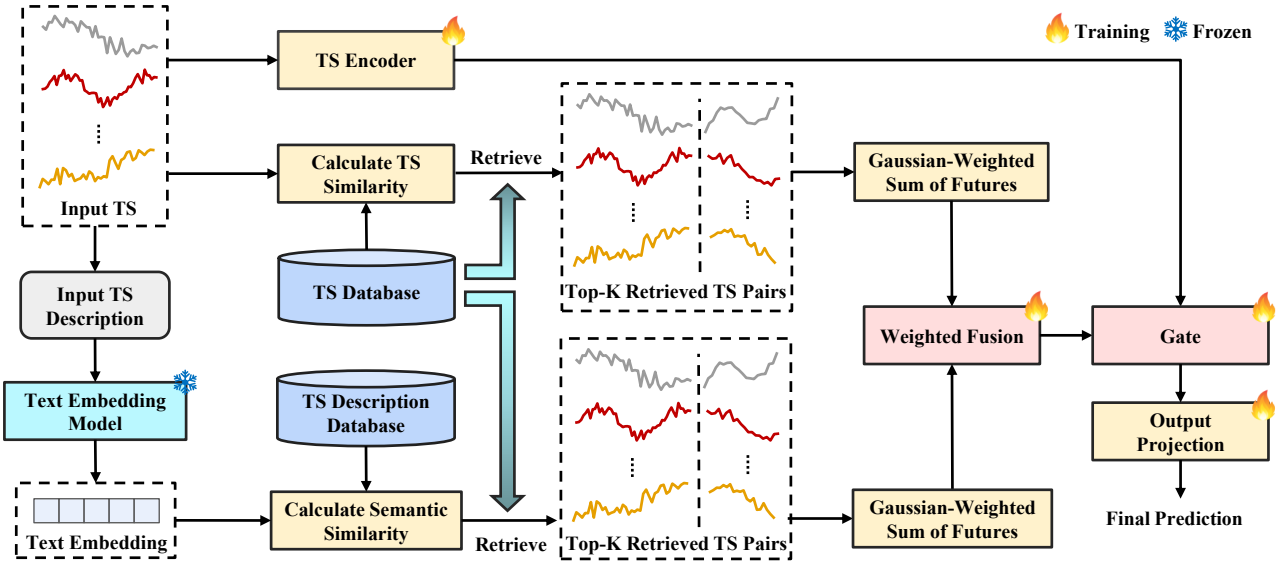

Figure 1. Overview of SERAF. Each retrieved TS pair contains a historical time series segment and its corresponding future.

Time period: from Thursday, January 4, 00:00 to Friday, February 2, 23:00;
Season: Winter;
Main trend: upward;
Main volatility: high.

Figure 2. An example of textual description of time series.

2.6. Fusion and Final Prediction

To adaptively balance the contributions of semantic and temporal retrieval, their aggregated futures $\hat{\mathbf{F}}_S^j$ and $\hat{\mathbf{F}}_T^j$ are first fused with a learnable weight:

$$\hat{\mathbf{F}}^j = w\,\hat{\mathbf{F}}_S^j + (1-w)\,\hat{\mathbf{F}}_T^j, \tag{10}$$

where $w \in (0,1)$ is a trainable parameter.

In addition to the two retrieval modules, the input X^j is passed through a linear time series (TS) encoder to produce a naive prediction $\hat{\mathbf{X}}^j$. This output is then adaptively integrated with $\hat{\mathbf{F}}^j$ through a gating mechanism, which dynamically balances their relative contributions:

$$\mathbf{G}^j = \beta\,\hat{\mathbf{X}}^j + (1-\beta)\hat{\mathbf{F}}^j, \quad \beta = \sigma\left(W[\hat{\mathbf{X}}^j;\hat{\mathbf{F}}^j]\right). \tag{11}$$

where $[\cdot;\cdot]$ denotes concatenation, W is a learnable projection, and $\sigma(\cdot)$ is the sigmoid function.

Finally, the representation $\mathbf{G}^j$ is mapped by an output projection to the final prediction $\hat{\mathbf{Y}}^j$, and the model is trained by minimizing the mean squared error (MSE) loss.

3. Experiment

3.1. Experimental Settings

Datasets. We train and evaluate SERAF on seven widely used multivariate time series datasets: ETTh1, ETTh2, ETTm1, ETTm2, Exchange, Weather, and Electricity (Wu et al., 2021).

Baselines. We compare SERAF against seven state-of-the-art time series forecasting models. Autoformer (Wu et al., 2021) and PatchTST (Nie et al., 2025) are Transformer-based forecasters. TimeMixer (Wang et al., 2024) is an MLP-based multiscale mixing model. DLinear (Zeng et al., 2022) employs a lightweight yet robust linear model. RAFT (Han et al., 2025) enhances linear models with multi-period retrieval modules based on TS similarity. CycleNet (Lin et al., 2024) explicitly captures periodic patterns on linear backbones. TimesNet (Wu et al., 2023) detects dominant periods via Fourier analysis.

Implementation details. We follow the experimental settings of RAFT (Han et al., 2025). The batch size is set to 32, and the Adam optimizer is employed. The input length is set as 720 and τ is set as 0.1. We use all-MiniLM-L6-v2 as the text embedding model. All experiments are conducted on one NVIDIA A100 GPU. Each result reported in the tables is reproduced and averaged over three independent runs.

3.2. Main Results

The forecasting results are shown in Table 1. SERAF achieves the best MSE on all four ETT datasets and the best or tied-best MAE on three of them, with second-best MAE on ETTh2. Compared with the retrieval-based base-

Table 1. Comparison of SERAF and baselines over seven datasets. All the results are with the same input time series length = 720 and averaged across four horizons (96, 192, 336, 720). Best results are shown in **bold** and second-best results are underlined.

	SERAF		RAFT		CycleNet		PatchTST		DLinear		TimeMixer		TimesNet		Autoformer	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.417**	**0.432**	0.418	0.434	0.437	0.450	0.696	0.595	0.521	0.508	0.447	0.440	0.909	0.679	1.090	0.785
ETTh2	**0.348**	0.401	0.358	0.408	0.368	0.406	0.472	0.479	0.445	0.472	0.365	**0.395**	0.486	0.491	0.494	0.516
ETTm1	**0.346**	**0.377**	0.347	**0.377**	0.365	0.390	0.551	0.489	0.400	0.422	0.381	0.396	0.926	0.653	4.233	1.653
ETTm2	**0.252**	**0.317**	0.257	0.321	0.285	0.338	0.354	0.383	0.290	0.346	0.275	0.323	0.418	0.426	1.363	0.830
Weather	0.235	0.280	0.241	0.287	**0.224**	**0.266**	0.232	0.274	0.238	0.289	0.240	0.272	0.475	0.421	0.424	0.449
Exchange	0.419	0.430	0.449	0.444	**0.403**	**0.426**	0.436	0.455	0.465	0.463	0.504	0.468	0.681	0.534	2.053	1.064
Electricity	**0.156**	0.257	0.160	0.259	0.157	**0.252**	0.216	0.318	0.225	0.319	0.169	0.268	0.193	0.304	0.227	0.364

Table 2. Ablation study of SERAF on ETTh2 and ETTm2 datasets, with MSE and MAE averaged over four forecasting horizons (96, 192, 336, 720). Best results are shown in **bold**.

	Full (SERAF)		w/o text		w/o gate		w/o weight	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	**0.348**	**0.401**	0.354	0.405	0.351	0.404	0.350	0.403
ETTm2	**0.252**	**0.317**	0.254	0.319	0.253	0.318	0.252	0.318

Table 3. Inference-stage efficiency profiling on ETTh1 with horizon 96 and batch size 32.

Method	Parameters (M)	Model GPU Memory (MiB)	Inference Time (s/iter)
DLinear	0.138	0.53	0.0203
RAFT	0.097	0.37	0.0230
SERAF	**0.088**	**0.34**	**0.0075**

line RAFT, SERAF matches or improves both metrics across all seven datasets, reducing the averaged MSE and MAE by 2.56% and 1.42%, respectively. Compared with CycleNet, SERAF reduces the averaged MSE and MAE by 2.95% and 1.34%, respectively. These results suggest that semantic retrieval consistently improves over TS-similarity retrieval, while its gains may be less pronounced on datasets whose future dynamics are harder to capture through coarse semantic descriptions alone.

3.3. Ablation Study

We conduct an ablation study on ETTh2 and ETTm2 to evaluate the contributions of key components in SERAF, as shown in Table 2. The results are averaged across four forecasting horizons (96, 192, 336, 720), with Full (SERAF) using the averaged results from Table 1. Removing the semantic retrieval module (w/o text) leads to the most significant average drop, confirming the complementary benefits of semantic similarity beyond pure time series similarity. Excluding the gating mechanism (w/o gate) and replacing it with simple concatenation, as in RAFT (Han et al., 2025), results in a general performance drop, highlighting the role of gating in adaptively balancing the encoder's prediction with retrieval-enhanced signals. Finally, replacing the learnable weighted fusion with uniform averaging (w/o weight) also degrades performance, demonstrating the effectiveness of relevance-aware weighting in fusing the two retrieval results.

3.4. Inference Efficiency Analysis

We profile inference-stage efficiency on ETTh1 with forecasting horizon 96 and batch size 32. All compared backbones are lightweight MLP-based forecasters. As shown in Table 3, SERAF uses the smallest forecasting backbone, with 0.088M parameters and 0.34 MiB model GPU memory, compared with 0.138M parameters and 0.53 MiB for DLinear and 0.097M parameters and 0.37 MiB for RAFT. SERAF achieves the fastest inference speed, requiring 0.0075 seconds per iteration, compared with 0.0203 and 0.0230 seconds per iteration for DLinear and RAFT, respectively.

4. Conclusion

In this paper, we proposed SERAF, which enhances retrieval-based forecasting by incorporating semantic retrieval with self-generated textual descriptions of time series. By constructing parallel time series and text databases, SERAF retrieves complementary historical patterns and adaptively fuses them through weighted fusion, while a gating module balances retrieval signals with the TS encoder's initial prediction. Experiments on seven real-world datasets and ablation studies highlight SERAF's design and demonstrate accuracy improvements. For future work, we will explore richer text templates and refined Top-K retrieval to reduce redundancy and improve robustness, toward more interpretable and generalizable retrieval-augmented forecasting.

References

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Chen, J., Zhao, Z., Nurbek, G., Feng, A., Maatouk, A., Tassiulas, L., Gao, Y., and Ying, R. Trace: Grounding time series in context for multimodal embedding and retrieval. *arXiv preprint arXiv:2506.09114*, 2025.

Daut, M. A. M., Hassan, M. Y., Abdullah, H., Rahman, H. A., Abdullah, M. P., and Hussin, F. Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review. *Renewable and Sustainable Energy Reviews*, 70:1108–1118, 2017.

Han, S., Lee, S., Cha, M., Arik, S. O., and Yoon, J. Retrieval augmented time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Lin, S., Lin, W., Hu, X., Wu, W., Mo, R., and Zhong, H. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024.

Lippi, M., Bertini, M., and Frasconi, P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882, 2013.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2025.

Ning, K., Pan, Z., Liu, Y., Jiang, Y., Zhang, J. Y., Rasul, K., Schneider, A., Ma, L., Nevmyvaka, Y., and Song, D. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. *arXiv preprint arXiv:2503.07649*, 2025.

Poon, S.-H. and Granger, C. W. J. Forecasting volatility in financial markets: A review. *Journal of economic literature*, 41(2):478–539, 2003.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J. Y., and Zhou, J. Timemixer: Decomposable multi-scale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023.

Yang, S., Wang, D., Zheng, H., and Jin, R. Timerag: Boosting llm time series forecasting via retrieval-augmented generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? *arXiv preprint arXiv:2205.13504*, 2022.

Zhang, H., Xu, C., Zhang, Y.-F., Zhang, Z., Wang, L., and Bian, J. Timeraf: Retrieval-augmented foundation model for zero-shot time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Zhou, T., Niu, P., Sun, L., Jin, R., et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.