MISSTIME: STATE-AWARE RETRIEVAL FOR GENERAL TIME SERIES ANALYSIS WITH MISSING VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series analysis is crucial in a wide range of real-world applications. However, multivariate time series often contain missing variables in practice due to sensor or communication failures, disrupting the cross-variable correlations and degrading performance across diverse downstream tasks. Existing approaches use imputation models or retrieval of historical instances as additional context to reconstruct the missing variables. However, we identify that they still face two key challenges: (i) *reliance on the imputation model* and (ii) *retrieval ambiguity* when incorporating historical references. Motivated by this, we propose **MissTime**, a direct retrieval framework that identifies similar historical instances and reconstructs the missing variables with the retrieved full observations. Specifically, MissTime introduces two key components: (i) Full Historical Bank Collection, which curates a historical time series bank with fully observed variables, providing the basis for data reconstruction; and (ii) State-Aware Retrieval, which extracts the underlying state representations from both time series with missing variables and fully observed time series. It then aligns samples with similar states through a contrastive disambiguation objective, mitigating retrieval ambiguity caused by partially observed variables. We conduct extensive experiments on time series benchmark datasets, demonstrating the consistent and significant efficacy of MissTime under different missing ratios across diverse downstream tasks. Our code, datasets and implementation details are available at: https://anonymous.4open.science/r/MissTime-F956.

1 INTRODUCTION

Time series analysis is crucial in extensive real-world applications, including weather prediction Karevan & Suykens (2020), medical diagnosis Kiyasseh et al. (2021), industrial maintenance Rabatel et al. (2011), and financial assistance Doran (1974). Modern time series architectures capture both the temporal dependencies and inter-variable correlations to learn versatile representations. They have achieved strong performance across diverse downstream tasks, including forecasting, classification, imputation, and anomaly detection Wu et al. (2023); Wang et al. (2025); Woo et al. (2024).

However, in practice, multivariate time series often contain missing variables. For example, in weather forecasting, entire variables are frequently missing due to sensor or communication failures. The variable-missing challenge disrupts the feature space and inter-channel correlations, leading to significant performance degradation across diverse tasks Yu et al. (2016); Du et al. (2023); Hao et al. (2025), which is shown in Fig. 4. Although controlled data curation can ensure a basic level of completeness during data collection Ding et al. (2024), variable missing during inference is difficult to tackle since the model must produce timely predictions from the missing inputs Liang et al. (2025).

To make robust predictions from time series with missing variables, existing methods can be broadly classified into two categories: (i) Task-oriented Representation Learning, which design specified architectures to extract robust representations, alongside task-tailored loss functions to make reliable predictions directly from missing data, such as GinAR Yu et al. (2024), Merlin Yu et al. (2025b) and HCIB Zhang et al. (2025). However, this group of methods is often tailored to a specific task, limiting their transferability across different tasks Liu et al. (2025a); Yu et al. (2025a); (ii) Data Reconstruction methods, which first recover the missing variables from observed ones and then use the reconstructed data as input to the downstream model for prediction. For this recovery step, a common approach is to train an imputation model. For instance, SAITS Du et al. (2023) uses

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

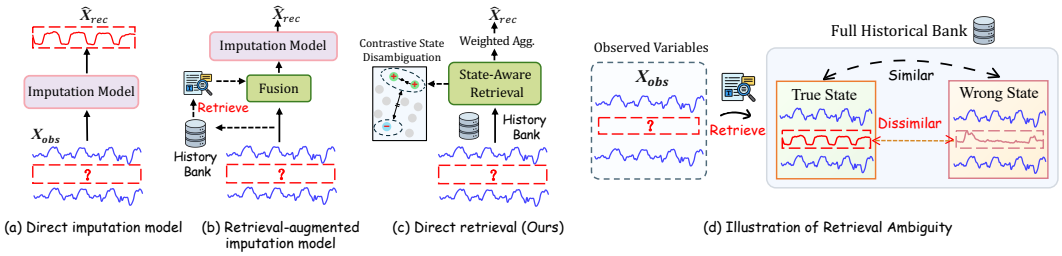


Figure 1: (a)-(c) Existing methods rely on direct imputation models or retrieval-augmented imputation models to reconstruct missing variables. In contrast, state-aware retrieval differentiates ambiguous states and reconstructs the missing variables by directly using retrieved historical sequences; (d) Historical instances that resemble observed channels may differ substantially in missing channels, causing retrieval ambiguity under state mismatch.

an attention-based imputation model to capture variable-wise correlations of observed time series and generate missing variables from the observed variables. MTSCI Zhou et al. (2024) and Glocal-IB Yang et al. (2026) employ generative diffusion and information bottleneck-based imputation models, respectively, to improve the recovery quality. However, reconstruction from the imputation model becomes highly uncertain at high missing ratios Jing et al. (2022). To improve the recovery reliability, recent work has retrieved similar historical full observations as additional context to assist the imputation model for data recovery. For instance, ReTime Jing et al. (2022) retrieves related sequences and fuses their representations with those of current observations before feeding into the imputation model. ALER-TI Truong et al. (2026) retrieves similar historical sequences through latent embedding alignment and integrates them into the imputation model through an adaptation module.

Despite the advances, existing retrieval-augmented reconstruction approaches still face two key challenges: ❶ **Reliance on the imputation model**. As shown in Fig. 1, although existing methods have retrieved fully observed historical instances as references, the missing variables are still generated by an imputation model whose training objective may have misalignments with the downstream task. Consequently, even a low imputation error can still yield suboptimal downstream performance Zhang et al. (2025). ❷ **Retrieval ambiguity.** Historical instances that resemble the current sequences in observed channels may differ substantially in missing channels, making it difficult to determine which retrieved instances are suitable for imputation. For example, a rainy summer night and a sunny winter morning may exhibit similar temperature patterns but markedly different precipitation levels. Such ambiguity arises when similar partial observations correspond to different system states.

In this paper, we introduce **MissTime**, a framework that reconstructs missing variables in time series by directly using retrieved historical sequences. MissTime contains two key components: **(i) Full Historical Bank** which curates historical time series with fully observed variables, providing the basis for data reconstruction; **(ii) State-Aware Retrieval**, which extracts the underlying state representations from both time series with missing variables and fully observed time series. It then aligns samples with similar state representations while separating those with different states through a contrastive disambiguation objective, thereby reducing retrieval ambiguity caused by state mismatch. With State-Aware Retrieval, MissTime achieves reliable data reconstruction by directly using retrieved historical sequences as imputation. Moreover, its plug-and-play design is naturally adaptable across diverse downstream tasks and different backbones. Our contributions can be summarized as follows:

- We identify two key challenges in existing retrieval-augmented reconstruction methods for missing time series: *reliance on imputation model* and *retrieval ambiguity*. This motivates the design of MissTime, which reconstructs missing data by directly retrieving from full time series.
- MissTime collects a full historical bank to curate historical time series with fully observed variables. State-aware retrieval further extracts the underlying system states from both partial and full observations and selects samples with similar states, mitigating the retrieval ambiguity.
- We conduct extensive experiments across five time series analysis tasks, including long-term forecasting, short-term forecasting, classification, imputation, and anomaly detection, demonstrating MissTime's effectiveness across varying missing ratios and generality on different backbones.

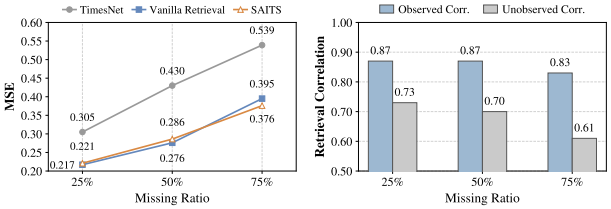
2 PRELIMINARIES

2.1 PROBLEM STATEMENT

Let a multivariate time series $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\} \in \mathbb{R}^{L \times C}$, where L denotes the sequence length, and C be the number of variables. During inference, a proportion $M\%$ of the C variables of the test sample $\mathbf{X}_{test}$ are randomly missing, and the values of the corresponding variables across the entire sequence are set to zero Yu et al. (2024). Our goal is to recover the missing variables based on the test instance with partially observed variables $\mathbf{X}_{obs}$. Then, the recovered time series $\hat{\mathbf{X}}$ should apply to diverse downstream tasks, including classification, forecasting, anomaly detection, and imputation.

2.2 PRELIMINARY STUDY OF VARIABLE RECONSTRUCTION VIA DIRECT RETRIEVAL

Before diving into technical details, we conduct preliminary studies to investigate: (i) the feasibility of using direct retrieval to recover the missing variables; (ii) the limitations of using vanilla retrieval. Specifically, we retrieve fully observed historical samples based on *feature similarity* between the input's observed channels and the corresponding channels of historical samples. The weighted sum of the retrieved instances is used to reconstruct the missing variables in $\mathbf{X}_{test}$, and the recovered input is then fed into the model for prediction. We provide the experimental details in App. B.4. We conduct experiments on Weather Wu et al. (2023) dataset and conclude the following observations:



(a) Results of direct retrieval. (b) Limits of vanilla retrieval.

Figure 2: (a) Performance of direct retrieval is comparable to imputation models. (b) The correlation between ground-truth values in unobserved channels decreases at high missing ratios.

Obs. 1. Direct retrieval-based reconstruction achieves performance comparable to that of imputation models. Fig. 2a demonstrates the forecasting performance of direct retrieval and a representative imputation model, SAITS Du et al. (2023). From the figure, it is observed that vanilla retrieval provides significant benefits: the simple feature similarity matching strategy substantially improves the robustness of TimesNet under all missing ratios and achieves performance comparable to SAITS. However, at higher missing ratios, direct retrieval underperforms SAITS in forecasting.

Obs. 2. The retrieved values of vanilla retrieval deviate substantially from the ground truth at high missing ratios. Fig. 2b describes the correlations between the retrieved values and the ground truth in observed and missing channels, respectively. From the figure, we observe that retrieved values in observed channels remain highly correlated with ground-truth values, whereas this correlation declines markedly in missing channels as the missing ratio increases. Thus, vanilla feature similarity matching on observed variables does not guarantee the similarity on all variables.

Discussion. Our observations from Fig. 2a and Fig. 2b suggest that direct retrieval is a promising and feasible approach to reconstruct the missing variables in multivariate time series. However, retrieval based solely on feature similarity in observed channels remains challenging: similarity in these channels does not guarantee similarity across all variables. Consequently, retrieved values may differ substantially from the true missing values, limiting reconstruction quality and downstream performance. This motivates us to develop a more advanced retrieval method that identifies historical samples with overall similar patterns across all variables based only on the observed variables.

3 THE FRAMEWORK OF MISSTIME

Our preliminary analysis in Fig. 2a has verified the feasibility of using direct retrieval to reconstruct the missing variables. Additionally, the results in Fig. 2b also indicate that similarity in the observed variables does not guarantee similarity across all variables, named *retrieval ambiguity*. Moreover, the historical data instances may also contain missing channels in reality. To address these challenges, we propose MissTime, which contains three key components: (i) **Full Historical Bank Collection** curates a reliable bank comprising fully observed history instances; (ii) **State-Aware Retrieval**

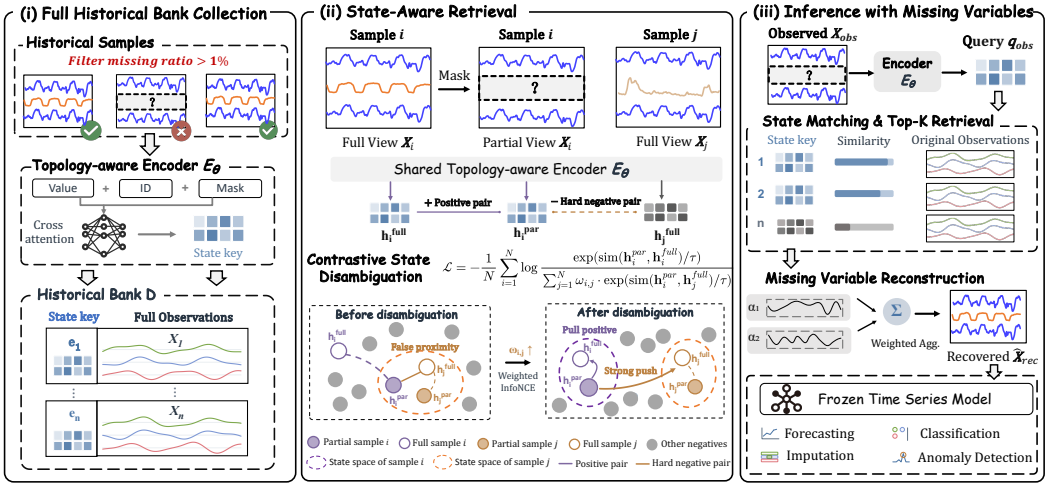


Figure 3: Framework of MissTime. It contains three key modules: (i) **Full Historical Bank Collection** curates fully observed historical instances; (ii) **State-Aware Retrieval** extracts underlying state representations from both time series with missing variables and fully observed time series. It does state alignment through contrastive state disambiguation; (iii) **Inference with Missing Variables** performs inference from the recovered data.

extracts underlying state representations from both time series with missing variables and fully observed time series. It aligns samples with similar states through contrastive state disambiguation; (iii) **Inference with Missing Variables** integrates the downstream time series model and performs inference with the reconstructed data. The details of each module are explained in Fig. 3.

3.1 Full Historical Bank Collection

Filtering Series with High Missing Rates. A full historical bank comprises time series instances with fully observed variables, providing the basis for retrieval. However, historical samples may also contain missingness, which introduces noise and degrades the retrieval quality. We therefore filter out samples with missing ratios $> 1\%$. Specifically, given the historical collection $\mathcal{X}_{\text{hist}}$, we retain:

$$m_i = \frac{\text{number of missing entries in } \mathbf{X}_i}{L \times C}, \qquad \mathcal{X}_{\text{filtered}} = \{\mathbf{X}_i \in \mathcal{X}_{\text{hist}} \mid m_i < 0.01\}. \tag{1}$$

Topology-Aware Encoding. Each filtered instance is then projected into a latent space for retrieval. To extract underlying state representations from $\mathbf{X}_i$ with either missing or fully observed variables, we must model both temporal and topological dependencies among variables. Hence, we first construct a composite embedding for each variable by adding its value, identity, and observability embeddings:

$$\mathbf{Z}_c^{\text{com}} = \mathbf{E}_{\text{value}}(\mathbf{X}_c) + \mathbf{E}_{\text{identity}}(\text{ID}_c) + \mathbf{E}_{\text{obs}}(\text{mask}_c), \tag{2}$$

where $\mathbf{X}_c$ denotes the input sequence of variable c, ID_c is its channel identifier used to index a learnable identity embedding, and $\text{mask}_c \in \{0, 1\}$ indicates whether the channel is missing (1) or observed (0). Subsequently, we apply self-attention to $\mathbf{Z}^{\text{com}} = [\mathbf{Z}_1^{\text{com}}, \mathbf{Z}_2^{\text{com}}, \ldots, \mathbf{Z}_C^{\text{com}}] \in \mathbb{R}^{C \times D}$ along the variate dimension to propagate topological information and get a state representation $\mathbf{h}$:

$$\mathbf{h} = \text{Channel-Attn}(\mathbf{Q}_Z, \mathbf{K}_Z, \mathbf{V}_Z), \tag{3}$$

where Channel-Attn denotes self-attention across channels, and $\mathbf{Q}_Z, \mathbf{K}_Z, \mathbf{V}_Z \in \mathbb{R}^{C \times D}$ are the query, key, and value representations derived from $\mathbf{Z}^{\text{com}}$. The resulting state representation $\mathbf{h} \in \mathbb{R}^{C \times D}$ captures inter-variable dependencies by aggregating information across all channels.

Full Historical Bank Collection. Finally, the full historical bank is defined as a standardized pair consisting of the original time series and its corresponding state representations, formally:

$$\mathcal{B} = \{(\mathbf{X}_i, \mathbf{e}_i), |i = 1, 2, \ldots, n\}, \tag{4}$$

where $\mathcal{B}$ denotes the collected full historical bank, $\mathbf{X}_i$ represents the i-th historical time series, and $\mathbf{e}_i$ represents the state representation computed by the topology-aware encoder.

3.2 State-Aware Retrieval

Even with a historical bank of fully observed samples, retrieving based on partially observed variables still faces the challenge of *Retrieval Ambiguity*. To address this, we further propose **State-Aware Retrieval**, which extracts the underlying state representations from both time series with missing variables and fully observed time series. Through a contrastive disambiguation objective, it aligns representations of samples with similar states while separating those with different states, thereby reducing retrieval ambiguity caused by state mismatch. We elaborate on the specifics below.

Contrastive State Disambiguation. Our objective is to learn state representations that remain consistent under variable missingness while distinguishing samples with different underlying states. For each training instance, we treat its partially observed and fully observed views as a positive pair, since both correspond to the same sample and underlying state. Let $\mathbf{h}_i^{\text{par}}$ and $\mathbf{h}_i^{\text{full}}$ denote their respective state representations. Other fully observed instances in the batch serve as negatives, because they correspond to different states. We optimize a weighted InfoNCE objective:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{h}_i^{par}, \mathbf{h}_i^{full})/\tau)}{\sum_{j=1}^{N} \omega_{i,j} \cdot \exp(\text{sim}(\mathbf{h}_i^{par}, \mathbf{h}_j^{full})/\tau)}, \tag{5}$$

where N is the batch size, τ is the contrastive temperature, and $\text{sim}(\cdot, \cdot)$ denotes the similarity function. The weight $\omega_{i,j}$ controls the contribution of negative sample j for query i. We next define these weights to emphasize negatives associated with retrieval ambiguity.

Mining Hard Negative States. Standard InfoNCE does not explicitly distinguish negatives according to the mismatch between their partial and complete observations. For retrieval, particularly challenging negatives resemble the query under partial observations but differ in their complete representations. Such samples can be mistaken for suitable references when retrieval relies on observed-channel similarity alone. We therefore assign larger weights to negatives exhibiting both high similarity between partially observed representations and low similarity between fully observed representations:

$$\omega_{i,j} = \exp\big(\frac{\text{sim}(\mathbf{h}_i^{par}, \mathbf{h}_j^{par})}{\tau_w}\big) \cdot (1 - \text{sim}(\mathbf{h}_i^{full}, \mathbf{h}_j^{full}), \tag{6}$$

where τ_w controls the sensitivity of the weights to partial-representation similarity. The first factor emphasizes negatives that appear similar under partial observations, while the second uses differences between complete representations as a signal of state mismatch. Together, they focus training on distinguishing samples that are easily confused from partially observed variables. As illustrated in Fig. 3, this objective encourages the encoder to capture state-discriminative information from the available observations, supporting the retrieval of historical samples with overall similar patterns.

3.3 Inference with Missing Variables

Given the observed input $\mathbf{X}_{\text{obs}}$, we compute its state representation $\mathbf{h}^{\text{par}}$ and retrieve top-k instances $\mathcal{C}$ from $\mathcal{B}$ that exhibit the most similar states, and then get the recovered $\hat{\mathbf{X}}$ via a weighted sum of $\mathcal{C}$.

$$\mathcal{C} = \text{TopK}\left(\left\{(\mathbf{X}_i, \text{sim}(\mathbf{h}^{\text{par}}, \mathbf{h}_i^{\text{full}}))\right\}_{i=1}^{n}\right), \quad \widehat{\mathbf{X}} = \mathbf{X}_{\text{obs}} + \mathbf{M} \odot \text{WeightedSum}(\mathcal{C}), \tag{7}$$

where $\mathbf{M}$ is a binary mask indicating missing variables, $\odot$ denotes element-wise multiplication, and $\text{WeightedSum}(\cdot)$ denotes the weighted aggregation of retrieved samples. The recovered $\hat{\mathbf{X}}$ is then fed into a downstream time series model to produce the final task output $\mathbf{y} = f_{\text{Time-Series-Model}}(\hat{\mathbf{X}})$. Here, the output $\mathbf{y}$ denotes task-specific predictions, such as classification logits or forecast values.

4 EXPERIMENTS

In this section, we conduct extensive experiments, aiming to answer the following research questions: **RQ1 (Main Performance):** How does MissTime perform under variable missingness across diverse downstream tasks? **RQ2 (Generality Analysis):** Can the efficacy of MissTime be generalized to different time series backbones? **RQ3 (Ablation Study):** How does each component contribute to the overall performance of MissTime? **RQ4 (Case Study):** How does state-aware retrieval identify historical samples with similar patterns across all variables using only the observed ones? **RQ5 (Parameter Sensitivity):** How do key hyperparameters affect the performance of MissTime?

4.1 EXPERIMENTAL SETUP

Datasets To verify the effectiveness of MissTime, we conduct experiments on benchmark multivariate time series datasets across five mainstream tasks, including short- and long-term forecasting, imputation, classification, and anomaly detection Wang et al. (2024). Please refer to Appendix B.1 for details about the benchmarks and datasets.

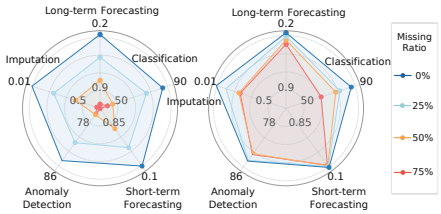


Figure 4: Overall improvement of MissTime across five downstream tasks and four missing ratios based on TimesNet.

Baselines. We compare MissTime with four categories of SOTA representative methods, including: (1) Representative time series forecasters: TimesNet Wu et al. (2023), TimeMixer Wang et al. (2024) and iTransformer Liu et al. (2024b); (2) Data reconstruction methods: SAITS Du et al. (2023), GinAR (forecasting) Yu et al. (2024) and HCIB (classification) Zhang et al. (2025); (3) Model with specialized designs for mitigating channel noise: TimeFilter Hu et al. (2025); (4) Retrieval-augmented time series models: We integrate two existing retrieval mechanisms of RAFT Han et al. (2025) and TS-RAG Ning et al. (2026). Please refer to Appendix B.2 for details about the compared methods.

Implementation Details. We use `TimesNet` Wu et al. (2023) as the backbone for its general-purpose design and good performance across tasks. MissTime operates under the Missing Completely At Random (MCAR) assumption Liang et al. (2025) and with different missing ratios across 25%, 50%, and 75%. Refer to App. B.3 for additional details on hyperparameters and settings across tasks.

4.2 MAIN RESULTS (RQ1)

Obs. 0. MissTime exhibits superior performance across all tasks. Fig. 4 reports that MissTime's retrieval mechanism brings substantial improvements across five diverse downstream tasks under different variable-missing conditions. Notably, compared with the setting without MissTime, the improvement becomes more pronounced as the missing ratio increases, demonstrating the strong effectiveness of MissTime's retrieval-based reconstruction for missing variables in the time series.

4.2.1 SHORT-TERM FORECASTING

Obs.1. MissTime demonstrates excellent performance on the short-term forecasting task. Tab. 1 reports the average performance across different missing ratios. We report the results of all baseline methods on the PeMS benchmark. MissTime substantially improves the robustness of TimesNet, achieving over 50% relative gains in both MSE and MAE. Moreover, MissTime outperforms SAITS, a representative imputation model-based method, highlighting the benefit of retrieval-based imputation for downstream tasks. Complete results under each missing ratio are detailed in App. A.1.

4.2.2 LONG-TERM FORECASTING

Obs. 2. MissTime achieves the best performance among the compared methods for the long-term forecasting tasks. Tab. 1 compares MissTime with comprehensive baseline methods across different missing ratios and reports the MSE and MAE metrics. MissTime achieves the best or second-best performance among all baselines and improves by over 40% on datasets with strong cross-channel correlations such as ECL and Traffic, enhancing the robustness of TimesNet. This

Table 1: Results of forecasting and imputation. We present values averaged over four missing ratios. Lower MSE and MAE are better. We highlight the **best** and second-best results. AVE. averages across datasets. - indicates out of memory. Blue shading indicates relative improvements ($\Delta\%$) over TimesNet, with darker shades denoting larger gains; Full results with different missing ratios and horizons are in App. A.1, A.2, A.3, A.4, and A.5.

(a) Short-term Forecasting

Method	PEMS03		PEMS04		PEMS07		PEMS08		AVE.		Δ (%) ↑
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
TimesNet (2023)	0.444	0.462	0.464	0.480	0.440	0.441	0.544	0.502	0.473	0.471	0.00
TimeMixer (2024)	0.344	0.403	0.380	0.428	0.358	0.397	0.421	0.424	0.376	0.413	+16.46
iTransformer (2024b)	0.421	0.434	0.450	0.458	0.425	0.429	0.504	0.460	0.450	0.445	+5.19
GinAR (2025a)	0.444	0.468	0.473	0.492	–	–	0.574	0.509	0.497	0.490	-2.21
TimeFilter (2025)	1.780	0.862	0.468	0.475	0.450	0.448	0.505	0.468	0.801	0.563	-44.41
RAFT (2025)	0.253	0.348	0.287	0.363	0.283	0.347	0.225	0.312	0.262	0.343	+35.96
TS-RAG (2026)	0.453	0.429	0.375	0.392	0.375	0.380	0.471	0.438	0.419	0.410	+12.29
SAITS (2023)	0.159	0.271	0.133	0.252	0.122	0.207	0.184	0.264	0.150	0.249	+57.83
MissTime (Ours)	**0.141**	**0.250**	**0.109**	**0.222**	**0.115**	**0.207**	**0.184**	**0.257**	**0.139**	**0.234**	+60.63

(b) Long-term Forecasting

	ETT		ECL		Weather		Traffic		AVE.		Δ (%) ↑
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
TimesNet (2023)	0.751	0.565	0.513	0.506	0.361	0.377	1.019	0.551	0.661	0.500	0.00
TimeMixer (2024)	**0.626**	0.536	0.477	0.466	0.345	0.361	0.844	0.483	0.573	0.462	+10.48
iTransformer (2024b)	0.733	0.554	0.484	0.472	0.356	0.369	0.812	0.471	0.596	0.467	+8.22
GinAR (2025a)	0.746	0.569	0.538	0.527	0.342	0.359	–	–	0.542	0.485	-0.27
TimeFilter (2025)	0.667	0.539	0.495	0.483	0.332	0.355	0.923	0.552	0.604	0.482	+6.04
RAFT (2025)	0.734	0.568	0.255	0.337	0.274	0.312	0.748	0.435	0.503	0.413	+20.65
TS-RAG (2026)	0.857	0.572	0.305	0.372	0.357	0.340	0.723	0.399	0.561	0.421	+15.51
SAITS (2023)	0.756	0.601	0.233	0.324	0.263	0.292	0.639	0.357	0.473	0.394	+24.87
MissTime (Ours)	0.631	**0.536**	**0.225**	**0.317**	**0.254**	**0.276**	**0.615**	**0.329**	**0.431**	**0.365**	+30.91

(c) Imputation

	ETT		ECL		Weather		AVE.		Δ (%) ↑
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
TimesNet (2023)	0.558	0.424	0.538	0.507	0.320	0.318	0.472	0.416	0.00
TimeMixer (2024)	0.572	0.421	0.512	0.460	0.316	0.308	0.467	0.396	+2.97
iTransformer (2024b)	0.554	0.469	0.504	0.428	0.319	0.313	0.459	0.403	+2.94
GinAR (2025a)	0.597	0.502	0.639	0.592	0.334	0.338	0.523	0.477	-12.76
TimeFilter (2025)	0.653	0.541	0.619	0.570	0.408	0.425	0.560	0.512	-20.81
RAFT (2025)	0.598	0.497	0.733	0.567	0.475	0.390	0.602	0.485	-21.98
TS-RAG (2026)	0.624	0.512	0.538	0.507	0.315	0.314	0.492	0.444	-5.52
SAITS (2023)	0.575	0.396	0.207	0.264	0.139	0.149	0.307	0.270	+35.09
MissTime (Ours)	**0.490**	**0.395**	**0.183**	**0.263**	**0.137**	**0.149**	**0.270**	**0.269**	+39.09

demonstrates that state-aware retrieval can effectively retrieve historical moments using only observed variables. Full results under different missing ratios and forecasting horizons are reported in App. A.2.

4.2.3 IMPUTATION

Obs. 3. MissTime exhibits superior performance on imputation tasks, even when compared with classic imputation model. Tab. 1 shows the model's ability to recover complete channels for missing variables using the observed channels. Notably, MissTime even surpasses SAITS, which is specifically designed for multivariate time series imputation, demonstrating the effectiveness of retrieving highly similar historical instances in reconstruction. Moreover, the results in Tab. 7 further show the efficacy under high missing ratios. Plearse refer to App. A.3 for full results.

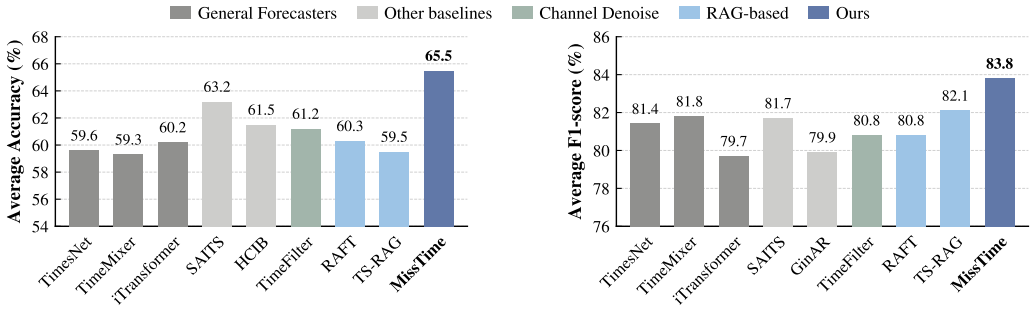


Figure 5: Results of classification and anomaly detection. The results are averaged over all datasets. Higher accuracy and F1-score indicate better performance. Full results are provided in Tab. 8 and 9.

4.2.4 CLASSIFICATION

Obs. 4. MissTime exhibits superior performance on classification tasks. Fig. 5a reports the accuracy averaged across all datasets and missing ratios. MissTime not only achieves a substantial 5.9% improvement in accuracy over TimesNet, but also consistently outperforms all baselines, including HCIB, which is specifically designed for the classification tasks. Furthermore, RAFT and TS-RAG yield inferior performance due to retrieval ambiguity. Refer to App. A.4 for full results.

4.2.5 ANOMALY DETECTION

Obs. 5. MissTime demonstrates great results on the anomaly detection task. Fig. 5b reports the F1-score averaged across all datasets and missing ratios under unsupervised time series anomaly detection. As shown in the figure, MissTime consistently outperforms all baselines in multivariate time series anomaly detection. It also surpasses SAITS by 1.4%, further validating the benefit of directly leveraging real-world historical data for downstream tasks. We further observe that TS-RAG, which employs a pre-trained time series foundation model as the retrieval encoder, also has good performance on this task. Please refer to App. A.5 for full results under each missing ratio.

4.3 GENERALITY ANALYSIS (RQ2)

Obs. 6. MissTime generalizes well across time series backbones. Fig 6 reports forecasting MSE on Electricity at 25%, 50%, and 75% missing ratios. To investigate the generality of MissTime across time series backbones, we compare TimesNet Wu et al. (2023), TimeMixer Wang et al. (2024), and Moirai Woo et al. (2024) with their MissTime variants. The relative reduction of the MissTime variants ranges from 35.5% to 69.0%, and the reduction increases significantly as the missing ratio increases. Full results on different tasks and the related analysis are reported in App. A.6.

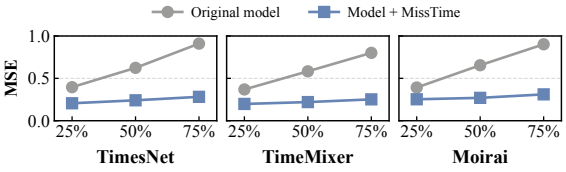


Figure 6: Generality analysis on Electricity dataset at 25%, 50%, and 75% missing ratios. Lower MSE means better.

4.4 ABLATION STUDY (RQ3)

Obs. 7. Each module of MissTime is critical for the performance under variable missingness. Fig 7 reports the contribution of each component in MissTime. We compare the full MissTime model with variants that remove the topology-aware component, hard-negative learning, or both. Removing hard-negative learning produces the largest single-component degradation on PEMS04, UGesture, and SMD, to go a step further removing both components gives the lowest accuracy and F1 on the latter two datasets. Please refer to App. A.7 for more ablation results and analysis.

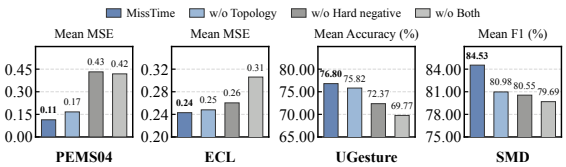


Figure 7: Ablation results averaged over 25%, 50%, and 75% missing ratios on different benchmarks. Lower MSE and higher accuracy/F1 mean better performance.

4.5 Case Study (RQ4)

Obs. 8. MissTime benefits from state-aware retrieval, which successfully recovers time series on both observed and missing channels. Fig 8 compares MissTime and Vanilla reconstructions on Electricity over a 96-hour window, showing two observed variables above and two missing variables below. Both methods capture the main periodic patterns in observed variables, achieving correlations above 0.97. The differences are clearer for the selected missing variables: Vanilla follows the recurring cycles but overestimates elevated segments, whereas MissTime more closely preserves their levels and amplitudes. Since MissTime retrieves historical contexts through similar system states, rather than observed patterns. In the two displayed missing-variable examples, MissTime reduces RMSE from 0.385 to 0.115 (70.1%) and from 0.641 to 0.207 (67.7%), respectively. Despite its larger RMSE, Vanilla has slightly higher correlation in the first example, illustrating that matching temporal patterns does not necessarily yield accurate reconstruction magnitudes. Taken together, these examples suggest that MissTime can recover time series on both observed and missing channels. Please refer to App. A.8 for more cases.

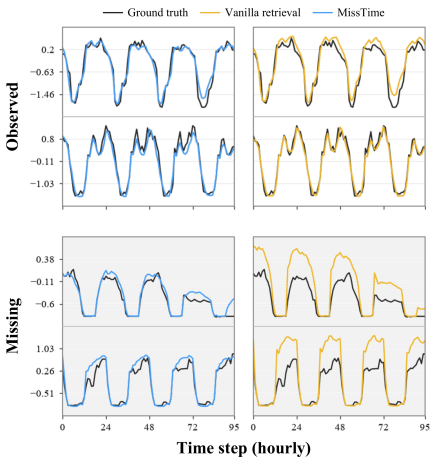


Figure 8: Case study: reconstructed data of MissTime and vanilla retrieval on observed and missing channels, respectively.

4.6 Parameter Analysis (RQ5)

Obs. 9. MissTime can maintain high performance under different hyperparameter settings. Fig 9 reports the sensitivity of MissTime to key retrieval hyperparameters on three representative datasets across different downstream tasks. Overall, the results reveal a consistent high-performing region around moderate hyperparameter settings rather than an isolated optimum. On SMD, increasing k from 3 to 5 consistently improves the F1-score, while further increasing it to 7 brings negligible gains, suggesting that a moderate number of retrieved instances is sufficient to provide reliable historical evidence. On UGesture, $\tau_w = 0.10$ consistently yields the strongest accuracy across different retrieval sizes. A similar pattern is observed on Electricity, while the performance varies only marginally on retrieval size. These observations indicate that moderate retrieval size and hard-negative mining temperature provide a balance between retrieving sufficient historical evidence and emphasizing informative hard negative states, and that MissTime remains effective under different hyperparameter settings. Please refer to A.9 for more results.

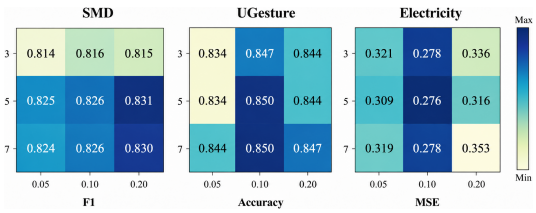


Figure 9: Hyperparameter sensitivity of MissTime on three datasets. Columns vary the hard-negative mining temperature $\tau_w \in \{0.05, 0.10, 0.20\}$, and rows vary the retrieval size $k \in \{3, 5, 7\}$.

5 Conclusion

We presented MissTime, a general framework that reconstructs missing variables by directly retrieving from historical observations. It combines a Full Historical Bank with State-Aware Retrieval to select historical instances with similar underlying states. A contrastive disambiguation objective aligns partial and full observations while distinguishing ambiguous matches, mitigating retrieval ambiguity under state mismatch. Experiments across various tasks and multiple backbones demonstrate consistent performance improvement under varying missing ratios. These results support state-aware historical retrieval as an effective and adaptable approach to missing-variable reconstruction for general time series analysis. By using actual historical observations for reconstruction, MissTime handles missing variables by directly retrieving and achieves strong performance on diverse tasks.

AI USE STATEMENT

Generative AI tools were used to assist with the manuscript's wording and presentation. Any use of large language models as part of the research methodology is described in the relevant technical sections. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

Our research focuses on multivariate time series analysis with missing variables and involves no studies with human participants, animal experiments, or environmentally sensitive materials. All datasets used in our experiments, including ETT, Electricity, Traffic, Weather, PeMS, selected datasets from the UEA archive, and SMD, MSL, SMAP, and PSM, are publicly available benchmarks widely used in time series research. We strictly adhere to the ICLR Code of Ethics, particularly regarding data privacy, transparency, and responsible computing practices. We anticipate no ethical risks, conflicts of interest, or concerns related to bias, fairness, or security arising from this research.

REPRODUCIBILITY STATEMENT

We provide comprehensive materials to support the reproducibility of our work. The MissTime framework, training objective, and inference procedure are described in Section 3, with algorithms provided in Appendix E. Dataset specifications, baseline methods, and experimental configurations are detailed in Section 4.1 and Appendices B.1–B.3, including preprocessing, missing-variable settings, hyperparameters, evaluation metrics, and computational resources. Detailed results across all five downstream tasks, together with generality, ablation, and sensitivity analyses, are reported in Appendix A. Our code, datasets, and implementation details are available at https://anonymous.4open.science/r/MissTime-F956.

REFERENCES

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

Jialin Chen, Ziyu Zhao, Gaukhar Nurbek, Aosong Feng, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. Trace: Grounding time series in context for multimodal embedding and retrieval. *arXiv preprint arXiv:2506.09114*, 2025.

Xiaoou Ding, Yichen Song, Hongzhi Wang, Donghua Yang, Chen Wang, and Jianmin Wang. Clean4tsdb: A data cleaning tool for time series databases. *Proceedings of the VLDB Endowment*, 17(12):4377–4380, 2024.

HE Doran. Prediction of missing observations in the time series of an economic variable. *Journal of the American Statistical Association*, 69(346):546–554, 1974.

Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pp. 1651–1661. PMLR, 2020.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *Forty-first International Conference on Machine Learning*, 2024.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

Qi Hao, Runchang Liang, Yue Gao, Hao Dong, Wei Fan, Lu Jiang, and Pengyang Wang. Is precise recovery necessary? a task-oriented imputation approach for time series forecasting on variable subset. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

Baoyu Jing, Si Zhang, Yada Zhu, Bin Peng, Kaiyu Guan, Andrew Margenot, and Hanghang Tong. Retrieval based time series forecasting. *arXiv preprint arXiv:2209.13525*, 2022.

Nur Dalila KA, Nooritawati Md Tahir, Zatul Iffah Abd Latiff, Mohamad Huzaimy Jusoh, Yoshikawa Akimasa, et al. Missing data imputation of magdas-9's ground electromagnetism with supervised machine learning and conventional statistical analysis models. *Alexandria Engineering Journal*, 61 (1):937–947, 2022.

Zahra Karevan and Johan AK Suykens. Transductive lstm for time-series prediction: An application to weather forecasting. *Neural Networks*, 125:1–9, 2020.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021.

Runchang Liang, Qi Hao, Yue Gao, Kunpeng Liu, Lu Jiang, Pengyang Wang, and Minghao Yin. Imputation via domain adaptation: Rethinking variable subset forecasting from knowledge transfer. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 1683–1694, 2025.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. *Advances in Neural Information Processing Systems*, 37:2766–2786, 2024a.

Mengna Liu, Pengshuai Yao, Xu Cheng, and Shengyong Chen. An end-to-end model for time series classification in the presence of missing values. *Expert Systems with Applications*, pp. 127954, 2025a.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024b.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, pp. 32369–32399. PMLR, 2024c.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Forty-second International Conference on Machine Learning*, 2025c.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Yiming Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster, 2025. URL https://arxiv.org/abs/2503.07649.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. *Advances in Neural Information Processing Systems*, 38:163170–163199, 2026.

Isaac Kofi Nti, Moses Teimeh, Owusu Nyarko-Boateng, and Adebayo Felix Adekoya. Electricity load forecasting: a systematic review. *Journal of Electrical Systems and Information Technology*, 7(1):13, 2020.

Julien Rabatel, Sandra Bringay, and Pascal Poncelet. Anomaly detection in monitoring sensor data for preventive maintenance. *Expert Systems with Applications*, 38(6):7003–7015, 2011.

Jussi Ronkainen and Antti Iivari. Designing a data management pipeline for pervasive sensor communication systems. *Procedia Computer Science*, 56:183–188, 2015.

Kutay Tire, Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. Retrieval augmented time series forecasting. *CoRR*, 2024.

Xuan-Thong Truong, Trung-Kien Le, Tung Kieu, Thi-Thu Nguyen, and Nhat-Hai Nguyen. Aler-ti: Aligned latent embedding retrieval for time series imputation. *arXiv preprint arXiv:2607.07640*, 2026.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis, 2025. URL `https://arxiv.org/abs/2410.16032`.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53140–53164, 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023.

Jie Yang, Kexin Zhang, Guibin Zhang, Philip S Yu, and Kaize Ding. Glocal information bottleneck for time series imputation. *Advances in Neural Information Processing Systems*, 38:104452–104484, 2026.

Silin Yang, Dong Wang, Haoqi Zheng, and Ruochun Jin. 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.

Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, and Yongjun Xu. Ginar: An end-to-end multivariate time series forecasting model suitable for variable missing. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3989–4000, 2024.

Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, Zhulin An, Qi Wang, and Yongjun Xu. Ginar+: A robust end-to-end framework for multivariate time series forecasting with missing values. *IEEE Transactions on Knowledge and Data Engineering*, 2025a.

Chengqing Yu, Fei Wang, Chuanguang Yang, Zezhi Shao, Tao Sun, Tangwen Qian, Wei Wei, Zhulin An, and Yongjun Xu. Merlin: Multi-view representation learning for robust multivariate time series forecasting with unfixed missing rates. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3633–3644, 2025b.

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. *Advances in neural information processing systems*, 29, 2016.

Shuo Zhang, Jing Wang, Shiqin Nie, Jinghang Yue, Weikang Zhu, and Youfang Lin. Loss or gain: Hierarchical conditional information bottleneck approach for incomplete time series classification. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3796–3807, 2025.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022.

Jianping Zhou, Junhao Li, Guanjie Zheng, Xinbing Wang, and Chenghu Zhou. Mtsci: A conditional diffusion model for multivariate time series consistent imputation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 3474–3483. ACM, October 2024. doi: 10.1145/3627673.3679532. URL http://dx.doi.org/10.1145/3627673.3679532.

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

APPENDIX CONTENTS

A Full Results **15**

A.1 Additional Results on Short-term Forecasting 15

A.2 Additional Results on Long-term Forecasting 16

A.3 Additional Results on Imputation 18

A.4 Additional Results on Classification 19

A.5 Additional Results on Anomaly Detection 20

A.6 Additional Results on Generality Analysis 21

A.7 Additional Results on Ablation Study 22

A.8 Additional Results on Case Study 22

A.9 Additional Results on Parameter Analysis 23

B Details of Experimental Setup **25**

B.1 Dataset Specifications 25

B.2 Baseline Methods 25

B.3 Experimental Configurations 26

B.4 Implementation Details of the Preliminary Retrieval Study 28

C FAQs **29**

C.1 Is the assumption of a fully observed historical database practical? 29

C.2 How does the contrastive state disambiguation handle hard negatives? 29

C.3 What is the computational and memory overhead of the retrieval pipeline? 30

C.4 How does MissTime perform under harder missingness scenarios and distribution shifts? 30

C.5 Does the Top-k weighted sum cause oversmoothing? 31

C.6 Why report results under the 0% missingness setting? 31

D Related Work **31**

D.1 Modern Time Series Analysis 31

D.2 Handling Missing Values in Time Series 31

D.3 Retrieval-Augmented Techniques in Time Series 32

E ALGORITHMS OF MISSTIME **32**

F Limitations **33**

A FULL RESULTS

This section provides a detailed breakdown of results under each missing ratio and dataset, complementing the averaged results in the main paper. We report all task-level tables and summarize key empirical observations below.

A.1 ADDITIONAL RESULTS ON SHORT-TERM FORECASTING

Table 2: Full short-term forecasting results on PEMS03, PEMS04, PEMS07, and PEMS08 with the forecasting horizon $T = 12$ under different missing ratios. The experiments evaluate missing ratios of 25%, 50%, and 75%, the input sequence length L is set to 96 for all baselines, and the best and second-best results are marked in **bold** and underlined, respectively.

Models		MissTime		TimesNet		TimeMixer		iTransformer		SAITS		GinAR		TimeFilter		RAFT		TS-RAG	
Missing(%)		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	0	0.106	0.217	0.106	0.217	**0.063**	**0.168**	0.071	0.177	0.135	0.249	0.108	0.226	0.150	0.259	0.104	0.219	0.098	0.207
	25	**0.122**	**0.235**	0.331	0.380	0.215	0.304	0.305	0.349	0.131	0.239	0.333	0.390	2.099	0.967	0.138	0.250	0.150	0.251
	50	**0.144**	**0.254**	0.553	0.542	0.410	0.470	0.538	0.521	0.174	0.290	0.555	0.549	2.966	1.200	0.357	0.443	0.350	0.400
	75	**0.191**	**0.294**	0.784	0.709	0.689	0.670	0.770	0.690	0.194	0.305	0.781	0.707	1.904	1.023	0.412	0.481	1.215	0.859
	Avg.	**0.141**	**0.250**	0.444	0.462	0.344	0.403	0.421	0.434	0.159	0.271	0.444	0.468	1.780	0.862	0.253	0.348	0.453	0.429
PEMS04	0	0.094	0.207	0.094	0.207	**0.069**	**0.173**	0.094	0.202	0.104	0.221	0.116	0.240	0.074	0.178	0.114	0.227	0.095	0.208
	25	**0.105**	**0.219**	0.336	0.385	0.236	0.324	0.332	0.373	0.125	0.243	0.360	0.414	0.327	0.373	0.130	0.244	0.136	0.246
	50	**0.115**	**0.228**	0.582	0.568	0.459	0.506	0.568	0.543	0.134	0.252	0.594	0.578	0.604	0.583	0.393	0.458	0.321	0.381
	75	**0.122**	**0.235**	0.846	0.759	0.754	0.709	0.808	0.716	0.169	0.290	0.821	0.737	0.866	0.765	0.512	0.522	0.948	0.734
	Avg.	**0.109**	**0.222**	0.464	0.480	0.380	0.428	0.450	0.458	0.133	0.252	0.473	0.492	0.468	0.475	0.287	0.363	0.375	0.392
PEMS07	0	0.086	0.190	0.086	0.190	**0.057**	**0.152**	0.069	0.170	0.085	0.187	-	-	0.058	0.156	0.083	0.184	0.084	0.185
	25	**0.105**	**0.200**	0.319	0.357	0.221	0.295	0.307	0.341	0.108	0.201	-	-	0.309	0.344	0.190	0.268	0.140	0.231
	50	**0.122**	**0.212**	0.562	0.525	0.429	0.469	0.547	0.512	0.129	0.213	-	-	0.561	0.529	0.377	0.437	0.409	0.430
	75	**0.147**	**0.225**	0.794	0.690	0.725	0.673	0.787	0.683	0.157	0.228	-	-	0.870	0.764	0.480	0.498	0.865	0.675
	Avg.	**0.115**	**0.207**	0.440	0.441	0.358	0.397	0.425	0.429	0.122	0.207	-	-	0.450	0.448	0.283	0.347	0.375	0.380
PEMS08	0	0.121	0.224	0.121	0.224	0.074	**0.173**	0.088	0.193	0.120	0.224	0.141	0.255	**0.072**	0.173	0.153	0.276	0.121	0.225
	25	0.164	**0.243**	0.399	0.405	0.286	0.327	0.354	0.373	**0.145**	0.244	0.462	0.439	0.370	0.386	0.191	0.293	0.213	0.286
	50	**0.200**	**0.267**	0.679	0.595	0.503	0.501	0.653	0.551	0.200	0.271	0.731	0.595	0.645	0.569	0.237	0.318	0.410	0.438
	75	**0.255**	**0.294**	0.979	0.785	0.821	0.697	0.923	0.722	0.270	0.318	0.961	0.748	0.933	0.743	0.321	0.361	1.140	0.802
	Avg.	**0.184**	**0.257**	0.544	0.502	0.421	0.424	0.504	0.460	0.184	0.264	0.574	0.509	0.505	0.468	0.225	0.312	0.471	0.438

Table 2 reports short-term forecasting results on PEMS03, PEMS04, PEMS07, and PEMS08, with an input length of $L = 96$ and a prediction horizon of $T = 12$. The experiments evaluate missing ratios of 25%, 50%, and 75%, together with a fully observed reference at 0%. MSE and MAE measure forecasting error, with lower values indicating better performance. The Avg. rows summarize all four displayed missing ratios, including 0%, while the following comparisons focus on performance under missing observations.

MissTime achieves the lowest MSE and MAE on PEMS03, PEMS04, and PEMS07 at every nonzero missing ratio. Its advantage over TimesNet becomes particularly pronounced under severe missingness. On PEMS04, for example, the MSE of MissTime increases only from 0.105 at 25% missingness to 0.122 at 75%, whereas the corresponding TimesNet error increases from 0.336 to 0.846. At 75% missingness, this represents an 85.6% reduction in MSE. MissTime also improves upon the competitive SAITS baseline in this setting, reducing MSE/MAE from 0.169/0.290 to 0.122/0.235. These results show that MissTime can maintain substantially more accurate forecasts as fewer variables remain observable.

The results on PEMS08 reveal a more nuanced comparison. At 25% missingness, SAITS achieves a lower MSE than MissTime, with 0.145 versus 0.164, while MissTime obtains a slightly lower MAE of 0.243 versus 0.244. At 50% and 75% missingness, the two methods tie in MSE at 0.200 and 0.270, respectively, but MissTime retains the lower MAE. In particular, its MAE at 75% missingness is 0.294, compared with 0.318 for SAITS. Thus, the advantage on PEMS08 is more consistent in MAE than in MSE, illustrating that the magnitude of improvement depends on both the dataset and evaluation metric.

Overall, MissTime achieves the best or tied-best reported result in 23 of the 24 dataset–missing-ratio–metric combinations under nonzero missingness. At 0%, it matches TimesNet on all four datasets, providing a common fully observed reference. Together, these results support the effectiveness

of retrieval-enhanced reconstruction for preserving short-term forecasting accuracy under variable missingness, with particularly substantial gains over TimesNet when the missing ratio is high.

A.2 Additional Results on Long-term Forecasting

Table 3: Full long-term forecasting results with the forecasting horizon $T = 96$ under different missing ratios. The input sequence length L is set to 96 for all baselines. The best and second-best results are in **bold** and underlined, respectively.

Models		MissTime		TimesNet		TimeMixer		iTransformer		SAITS		GinAR		TimeFilter		RAFT		TS-RAG	
Missing(%)		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT	0	0.405	0.424	0.406	0.424	**0.374**	0.398	0.393	0.409	0.412	0.428	0.396	0.419	0.387	**0.396**	0.396	0.410	0.406	0.425
	25	0.544	0.500	0.611	0.527	**0.494**	**0.466**	0.564	0.503	0.651	0.546	0.562	0.510	0.563	0.493	0.600	0.509	0.580	0.519
	50	**0.724**	**0.587**	0.776	0.620	0.735	0.596	0.765	0.612	0.905	0.669	0.777	0.627	0.730	0.588	0.855	0.613	0.779	0.613
	75	**0.850**	**0.633**	0.910	0.690	0.900	0.682	0.910	0.690	1.056	0.752	0.950	0.720	0.888	0.676	1.181	0.749	1.063	0.725
	Avg.	0.631	**0.536**	0.751	0.565	**0.626**	0.536	0.733	0.554	0.756	0.601	0.746	0.569	0.667	0.539	0.734	0.568	0.857	0.572
ECL	0	0.171	0.276	0.171	0.276	0.155	0.247	0.165	0.256	0.174	0.279	0.232	0.331	**0.138**	**0.235**	0.210	0.302	0.171	0.276
	25	0.206	**0.301**	0.396	0.429	0.369	0.393	0.378	0.400	**0.203**	0.302	0.451	0.472	0.372	0.401	0.239	0.324	0.215	0.312
	50	0.241	0.331	0.624	0.583	0.583	0.538	0.589	0.543	**0.236**	**0.330**	0.642	0.594	0.625	0.577	0.270	0.348	0.333	0.399
	75	**0.282**	**0.361**	0.860	0.737	0.801	0.685	0.802	0.688	0.319	0.386	0.825	0.711	0.847	0.720	0.300	0.372	0.500	0.502
	Avg.	**0.225**	**0.317**	0.513	0.506	0.477	0.466	0.484	0.472	0.233	0.324	0.538	0.527	0.495	0.483	0.255	0.337	0.305	0.372
Weather	0	0.171	0.220	0.171	0.220	0.164	0.211	0.181	0.223	0.168	0.217	0.163	0.211	**0.156**	**0.203**	0.196	0.255	0.172	0.221
	25	**0.219**	0.255	0.305	0.330	0.285	0.312	0.296	0.320	0.221	**0.249**	0.285	0.309	0.276	0.306	0.224	0.279	0.296	0.299
	50	**0.275**	**0.287**	0.430	0.433	0.414	0.415	0.417	0.419	0.286	0.301	0.401	0.407	0.388	0.406	0.277	0.318	0.445	0.389
	75	**0.353**	**0.340**	0.539	0.525	0.518	0.505	0.529	0.513	0.376	0.402	0.517	0.507	0.509	0.504	0.398	0.396	0.516	0.452
	Avg.	**0.254**	**0.276**	0.361	0.377	0.345	0.361	0.356	0.369	0.263	0.292	0.342	0.359	0.332	0.355	0.274	0.312	0.357	0.340
Traffic	0	0.595	0.320	0.595	0.320	0.470	0.293	0.413	0.287	0.610	0.327	-	-	**0.391**	**0.260**	0.696	0.429	0.594	0.321
	25	**0.609**	**0.327**	0.828	0.455	0.719	0.420	0.685	0.401	0.634	0.354	-	-	0.822	0.513	0.730	0.433	0.612	0.331
	50	**0.617**	**0.331**	1.138	0.626	0.968	0.546	0.948	0.534	0.647	0.364	-	-	1.148	0.680	0.764	0.437	0.686	0.369
	75	**0.639**	**0.340**	1.513	0.805	1.219	0.673	1.210	0.662	0.667	0.388	-	-	1.332	0.753	0.800	0.442	0.900	0.477
	Avg.	**0.615**	**0.329**	1.019	0.551	0.844	0.483	0.812	0.471	0.639	0.357	-	-	0.923	0.552	0.748	0.435	0.723	0.399

Table 4: Full long-term forecasting results with the forecasting horizon $T = 192$ under different missing ratios. The input sequence length L is set to 96 for all baselines. The best and second-best results are in **bold** and underlined, respectively.

Models		MissTime		TimesNet		TimeMixer		iTransformer		SAITS		GinAR		TimeFilter		RAFT		TS-RAG	
Missing(%)		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT	0	0.461	0.457	0.461	0.457	**0.432**	0.437	0.935	0.697	0.470	0.461	0.458	0.459	0.438	**0.424**	0.445	0.440	0.469	0.463
	25	0.617	0.545	0.652	0.559	**0.558**	**0.506**	0.999	0.731	0.685	0.559	0.620	0.538	0.623	0.534	0.728	0.560	0.601	0.540
	50	0.777	0.622	0.821	0.648	**0.747**	**0.605**	1.070	0.769	0.881	0.657	0.767	0.615	0.807	0.636	1.037	0.688	0.807	0.627
	75	1.011	0.724	0.950	0.711	**0.917**	**0.691**	1.127	0.797	1.059	0.748	0.936	0.704	0.969	0.723	1.214	0.771	1.076	0.722
	Avg.	0.717	0.587	0.721	0.594	**0.664**	**0.560**	1.033	0.749	0.774	0.606	0.695	0.579	0.709	0.579	0.856	0.615	0.738	0.588
Weather	0	0.226	0.265	0.226	0.265	0.209	0.252	0.228	0.261	0.223	0.263	0.213	0.257	**0.204**	**0.248**	0.236	0.293	0.225	0.264
	25	0.270	**0.293**	0.343	0.363	0.315	0.341	0.337	0.352	0.265	0.305	0.321	0.345	0.320	0.341	**0.263**	0.316	0.334	0.328
	50	0.328	**0.327**	0.452	0.453	0.420	0.429	0.434	0.436	**0.291**	0.330	0.423	0.429	0.422	0.426	0.305	0.346	0.458	0.404
	75	**0.390**	**0.369**	0.553	0.537	0.525	0.518	0.532	0.519	0.423	0.454	0.527	0.516	0.534	0.522	0.404	0.408	0.542	0.462
	Avg.	0.304	**0.314**	0.394	0.405	0.367	0.385	0.383	0.392	**0.301**	0.338	0.371	0.387	0.370	0.384	0.302	0.341	0.390	0.365
ECL	0	0.188	0.290	0.188	0.290	0.169	0.260	0.171	0.260	0.189	0.291	0.261	0.352	**0.156**	**0.250**	0.210	0.305	0.189	0.292
	25	0.237	0.329	0.415	0.447	0.382	0.404	0.380	0.403	**0.215**	**0.312**	0.464	0.481	0.375	0.402	0.237	0.326	0.228	0.326
	50	**0.253**	**0.342**	0.656	0.605	0.593	0.547	0.591	0.546	0.295	0.368	0.648	0.599	0.599	0.555	0.266	0.348	0.340	0.407
	75	**0.288**	**0.370**	0.865	0.739	0.800	0.687	0.804	0.689	0.322	0.390	0.833	0.717	0.816	0.699	0.295	0.371	0.505	0.510
	Avg.	**0.242**	**0.333**	0.531	0.520	0.486	0.475	0.487	0.475	0.255	0.340	0.552	0.537	0.487	0.477	0.252	0.338	0.316	0.384
Traffic	0	0.619	0.330	0.619	0.330	0.497	0.304	0.433	0.289	0.624	0.334	-	-	**0.413**	**0.269**	0.647	0.407	0.620	0.331
	25	0.630	**0.335**	0.876	0.473	0.739	0.428	0.694	0.416	**0.628**	0.340	-	-	0.800	0.490	0.679	0.412	0.637	0.338
	50	0.639	**0.340**	1.141	0.620	0.984	0.552	0.951	0.544	**0.634**	0.347	-	-	1.091	0.642	0.718	0.417	0.708	0.375
	75	**0.682**	**0.366**	1.571	0.827	1.231	0.677	1.216	0.673	0.702	0.398	-	-	1.306	0.734	0.756	0.422	0.899	0.470
	Avg.	**0.643**	**0.343**	1.052	0.563	0.863	0.490	0.824	0.481	0.647	0.355	-	-	0.903	0.534	0.700	0.415	0.716	0.379

Tables 3, 4, 5, and 6 report long-term forecasting results on ETT, Weather, ECL, and Traffic with a fixed input length of $L = 96$ and prediction horizons of $T \in \{96, 192, 336, 720\}$. Each table evaluates missing ratios of 25%, 50%, and 75%, alongside a fully observed reference at 0%. MSE

Table 5: Full long-term forecasting results with the forecasting horizon $T = 336$ under different missing ratios. The input sequence length L is set to 96 for all baselines. The best and second-best results are in **bold** and underlined, respectively.

Models		MissTime		TimesNet		TimeMixer		iTransformer		SAITS		GinAR		TimeFilter		RAFT		TS-RAG	
	Missing(%)	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT	0	0.487	0.471	0.487	0.471	0.491	0.467	0.970	0.715	0.512	0.483	0.481	0.471	0.491	**0.453**	0.487	0.465	**0.486**	0.471
	25	0.629	0.552	0.655	0.555	0.628	0.549	1.037	0.751	0.686	0.556	0.662	0.562	0.660	0.547	0.693	0.554	**0.624**	**0.546**
	50	0.768	**0.618**	0.816	0.646	**0.757**	0.621	1.098	0.784	0.875	0.651	0.822	0.649	0.822	0.645	0.890	0.640	0.847	0.642
	75	0.967	**0.704**	0.938	0.714	**0.915**	0.711	1.134	0.800	1.046	0.750	0.942	0.712	0.958	0.722	1.164	0.747	1.061	0.714
	Avg.	0.713	**0.586**	0.724	0.597	**0.698**	0.587	1.060	0.763	0.780	0.610	0.727	0.599	0.733	0.592	0.809	0.602	0.755	0.593
Weather	0	0.299	0.314	0.299	0.314	0.268	0.294	0.283	0.301	0.280	0.305	0.284	0.304	**0.262**	**0.291**	0.282	0.331	0.301	0.315
	25	0.350	0.342	0.406	0.403	0.373	0.381	0.380	0.381	0.332	0.351	0.375	0.382	0.356	0.369	**0.308**	**0.348**	0.417	0.380
	50	0.400	**0.369**	0.492	0.479	0.472	0.461	0.466	0.456	0.430	0.430	0.463	0.458	0.450	0.450	**0.352**	0.379	0.522	0.444
	75	**0.439**	**0.397**	0.570	0.548	0.551	0.531	0.546	0.529	0.551	0.523	0.552	0.535	0.547	0.531	0.443	0.433	0.573	0.477
	Avg.	**0.345**	**0.356**	0.442	0.436	0.416	0.417	0.419	0.417	0.398	0.402	0.419	0.420	0.404	0.410	0.346	0.373	0.453	0.404
ECL	0	0.206	0.309	0.206	0.309	0.186	0.276	0.187	0.276	0.208	0.308	0.281	0.369	**0.169**	**0.266**	0.223	0.319	0.204	0.307
	25	0.247	0.339	0.426	0.457	0.392	0.414	0.392	0.415	**0.235**	**0.330**	0.470	0.487	0.391	0.416	**0.235**	**0.330**	0.240	0.336
	50	**0.265**	**0.352**	0.648	0.603	0.594	0.551	0.601	0.555	0.301	0.375	0.652	0.602	0.614	0.566	0.274	0.357	0.337	0.407
	75	0.304	0.384	0.881	0.747	0.802	0.690	0.804	0.692	0.328	0.396	0.830	0.716	0.824	0.704	**0.303**	**0.379**	0.484	0.499
	Avg.	**0.256**	**0.346**	0.540	0.529	0.494	0.483	0.496	0.485	0.268	0.352	0.558	0.544	0.500	0.488	0.262	0.348	0.316	0.387
Traffic	0	0.627	0.334	0.627	0.334	0.507	0.302	0.444	0.292	0.641	0.342	-	-	**0.429**	**0.277**	0.653	0.410	0.631	0.337
	25	0.647	**0.343**	0.862	0.468	0.753	0.427	0.705	0.419	0.644	0.346	-	-	0.805	0.489	0.690	0.414	**0.643**	**0.343**
	50	**0.652**	**0.346**	1.114	0.606	0.999	0.553	0.965	0.547	0.655	0.356	-	-	1.087	0.630	0.731	0.418	0.693	0.372
	75	0.690	0.369	1.355	0.735	1.242	0.678	1.234	0.675	**0.657**	**0.356**	-	-	1.309	0.727	0.766	0.422	0.887	0.467
	Avg.	0.654	**0.348**	0.990	0.536	0.875	0.490	0.837	0.483	**0.649**	0.350	-	-	0.908	0.531	0.710	0.416	0.714	0.380

Table 6: Full long-term forecasting results with the forecasting horizon $T = 720$ under different missing ratios. The input sequence length L is set to 96 for all baselines. The best and second-best results are in **bold** and underlined, respectively.

Models		MissTime		TimesNet		TimeMixer		iTransformer		SAITS		GinAR		TimeFilter		RAFT		TS-RAG	
	Missing(%)	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT	0	**0.509**	0.492	**0.509**	0.492	0.541	0.508	1.103	0.783	0.553	0.521	0.551	0.526	0.510	**0.484**	0.510	0.508	0.533	0.510
	25	0.691	0.601	0.652	0.568	0.667	0.577	1.130	0.796	0.710	0.579	0.637	0.561	**0.628**	**0.547**	0.740	0.601	0.661	0.579
	50	0.790	0.643	0.819	0.655	0.826	0.657	1.160	0.813	0.889	0.670	**0.779**	**0.635**	0.784	0.639	0.908	0.680	0.875	0.649
	75	0.991	0.722	0.938	0.719	0.928	0.712	1.167	0.817	1.067	0.755	**0.898**	**0.697**	0.943	0.725	1.124	0.760	1.272	0.764
	Avg.	0.745	0.615	0.730	0.609	0.741	0.614	1.140	0.802	0.805	0.631	**0.716**	0.605	**0.716**	**0.599**	0.821	0.637	0.835	0.626
Weather	0	0.372	0.363	0.372	0.363	0.351	0.349	0.358	0.350	0.355	0.352	0.383	0.365	**0.340**	**0.342**	0.345	0.381	0.377	0.365
	25	**0.270**	**0.293**	0.463	0.443	0.436	0.424	0.434	0.418	0.361	0.371	0.421	0.412	0.414	0.408	0.360	0.390	0.474	0.417
	50	0.413	**0.385**	0.538	0.511	0.495	0.482	0.504	0.483	**0.387**	0.400	0.490	0.476	0.490	0.475	0.391	0.411	0.559	0.466
	75	**0.434**	**0.399**	0.600	0.568	0.564	0.544	0.565	0.542	0.461	0.476	0.565	0.544	0.567	0.545	0.470	0.451	0.583	0.485
	Avg.	**0.372**	**0.360**	0.493	0.471	0.462	0.450	0.465	0.448	0.391	0.400	0.465	0.449	0.453	0.443	0.392	0.408	0.498	0.433
ECL	0	0.489	0.431	0.489	0.431	0.226	0.311	0.223	0.307	0.242	0.335	0.321	0.398	**0.197**	**0.292**	0.258	0.350	0.256	0.349
	25	**0.277**	**0.360**	0.450	0.474	0.417	0.438	0.422	0.439	0.297	0.369	0.486	0.502	0.416	0.442	0.278	0.364	0.290	0.377
	50	0.301	**0.377**	0.652	0.607	0.616	0.570	0.616	0.570	0.318	0.388	0.661	0.611	0.637	0.591	**0.299**	0.379	0.395	0.450
	75	0.330	0.401	0.854	0.735	0.813	0.701	0.812	0.701	0.348	0.410	0.835	0.722	0.838	0.723	**0.321**	**0.395**	0.509	0.520
	Avg.	0.349	0.392	0.599	0.537	0.518	0.505	0.518	0.504	0.301	0.376	0.576	0.558	0.522	0.512	**0.289**	**0.372**	0.363	0.424
Traffic	0	0.668	0.354	0.668	0.354	0.530	0.317	0.477	0.311	0.680	0.363	-	-	**0.462**	**0.295**	0.695	0.431	0.667	0.352
	25	0.676	**0.355**	0.911	0.485	0.777	0.439	0.725	0.434	0.683	0.368	-	-	0.834	0.493	0.732	0.433	**0.675**	0.358
	50	0.679	**0.357**	1.143	0.614	1.017	0.560	0.975	0.556	**0.675**	0.364	-	-	1.129	0.636	0.765	0.435	0.716	0.381
	75	0.712	**0.378**	1.414	0.755	1.261	0.683	1.235	0.681	**0.700**	0.388	-	-	1.351	0.738	0.801	0.438	0.896	0.469
	Avg.	**0.684**	**0.361**	1.034	0.552	0.896	0.500	0.853	0.496	0.685	0.371	-	-	0.944	0.541	0.748	0.434	0.739	0.390

and MAE measure forecasting error, with lower values indicating better performance. These results allow us to examine how the effectiveness of MissTime varies with the prediction horizon, missing ratio, and dataset.

Across all four horizons, MissTime achieves lower MSE and MAE than TimesNet on Weather, ECL, and Traffic at every nonzero missing ratio. At $T = 96$, MissTime obtains the best results in both metrics on all four datasets under 75% missingness. On Traffic, for example, it achieves an MSE/MAE of 0.639/0.340, compared with 1.513/0.805 for TimesNet, reducing both errors by approximately 57.8%. The improvements also persist at $T = 192$: on ECL with 75% missingness, MissTime reduces MSE/MAE from 0.865/0.739 to 0.288/0.370, corresponding to reductions of 66.7% and 49.9%, respectively. These comparisons show that MissTime can substantially improve forecasting accuracy when a large proportion of variables is unavailable.

The benefits remain evident at longer prediction horizons. On Weather, MissTime achieves the lowest MSE and MAE among the compared methods at 75% missingness for all four horizons. Its MSE/MAE is 0.439/0.397 at $T = 336$ and 0.434/0.399 at $T = 720$, compared with 0.570/0.548 and 0.600/0.568 for TimesNet. On Traffic at $T = 720$ and 75% missingness, MissTime obtains 0.712/0.378, whereas TimesNet reaches 1.414/0.755. These results support the usefulness of retrieval-enhanced reconstruction for forecasting over extended horizons under severe variable missingness.

The improvements nevertheless depend on the dataset and competing method. On ETT, the advantage becomes less consistent at longer horizons: at $T = 720$ and 75% missingness, MissTime obtains 0.991/0.722, while GinAR achieves 0.898/0.697 and TimesNet achieves 0.938/0.719. On ECL under the same setting, RAFT also performs slightly better than MissTime, with 0.321/0.395 versus 0.330/0.401. Overall, the four tables demonstrate consistent improvements over TimesNet on Weather, ECL, and Traffic across the evaluated horizons, while showing that MissTime does not uniformly outperform every alternative. The results support its effectiveness for long-term forecasting under partial observation, with the strength of the benefit varying across datasets and error metrics.

A.3 Additional Results on Imputation

Table 7: Full missing-variable imputation results on ETT, Weather, and ECL under different missing ratios. The input sequence length L is set to 96 for all baselines. The best and second-best results are in **bold** and underlined, respectively.

Models		MissTime		TimesNet		TimeMixer		iTransformer		SAITS		GinAR		TimeFilter		RAFT		TS-RAG	
Missing(%)		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT	25	**0.239**	0.242	0.326	0.264	0.272	0.215	0.289	0.308	0.243	**0.189**	0.359	0.361	0.452	0.427	0.376	0.367	0.379	0.360
	50	**0.515**	0.407	0.570	0.432	0.568	0.418	0.562	0.477	0.577	**0.396**	0.589	0.496	0.658	0.544	0.568	0.492	0.660	0.548
	75	**0.715**	**0.538**	0.778	0.576	0.876	0.631	0.812	0.623	0.904	0.604	0.843	0.648	0.850	0.652	0.849	0.633	0.832	0.629
	Avg.	**0.490**	**0.395**	0.558	0.424	0.572	0.421	0.554	0.469	0.575	0.396	0.597	0.502	0.653	0.541	0.598	0.497	0.624	0.512
Weather	25	0.073	0.100	0.163	0.170	0.154	0.162	0.161	0.163	**0.046**	**0.074**	0.175	0.194	0.306	0.341	0.340	0.261	0.143	0.162
	50	**0.165**	**0.165**	0.321	0.315	0.309	0.303	0.319	0.314	0.168	0.169	0.337	0.346	0.403	0.423	0.555	0.424	0.314	0.312
	75	**0.174**	**0.182**	0.476	0.468	0.485	0.460	0.478	0.461	0.204	0.203	0.490	0.475	0.515	0.511	0.531	0.485	0.487	0.469
	Avg.	**0.137**	**0.149**	0.320	0.318	0.316	0.308	0.319	0.313	0.139	0.149	0.334	0.338	0.408	0.425	0.475	0.390	0.315	0.314
ECL	25	**0.130**	**0.177**	0.303	0.347	0.256	0.254	0.254	0.228	0.158	0.181	0.451	0.472	0.375	0.404	0.480	0.401	0.303	0.346
	50	**0.175**	0.277	0.537	0.507	0.524	0.490	0.503	0.424	0.178	**0.262**	0.642	0.594	0.632	0.584	0.775	0.579	0.539	0.509
	75	**0.245**	**0.336**	0.773	0.668	0.757	0.637	0.755	0.633	0.285	0.348	0.825	0.711	0.850	0.723	0.945	0.722	0.772	0.667
	Avg.	**0.183**	**0.263**	0.538	0.507	0.512	0.460	0.504	0.428	0.207	0.264	0.639	0.592	0.619	0.570	0.733	0.567	0.538	0.507

Table 7 reports missing-variable imputation results on ETT, Weather, and ECL with an input length of $L = 96$ and missing ratios of 25%, 50%, and 75%. MSE and MAE measure reconstruction error, while the Avg. rows summarize performance across the three missing ratios.

MissTime achieves the lowest MSE in eight of the nine dataset–missing-ratio combinations and the lowest average MSE on all three datasets. However, its advantage in MAE is less consistent at moderate missingness. SAITS obtains lower MAE on all three datasets at 25% missingness and on ETT and ECL at 50%. On Weather at 25%, SAITS also achieves lower MSE, indicating that MissTime does not uniformly outperform specialized imputation methods across all settings.

At 75% missingness, MissTime achieves the lowest MSE and MAE on every dataset. On ETT, for example, it obtains 0.715/0.538, compared with 0.904/0.604 for SAITS. Averaged equally across the three datasets, its MSE and MAE are 18.6% and 8.6% lower than those of SAITS in this setting.

These results support the effectiveness of MissTime for missing-variable reconstruction, with the most consistent benefits appearing under severe missingness.

A.4 Additional Results on Classification

Table 8: Full classification results on ten datasets from the UEA archive under different missing ratios. We report the classification accuracy (%) under missing ratios of 0%, 25%, 50%, and 75%, and the final row averages each method over the ten datasets.

Models		MissTime	TimesNet	TimeMixer	iTransformer	SAITS	GinAR	TimeFilter	RAFT	TS-RAG
Datasets	Missing(%)	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC	ACC
EthanolConcentration	0	28.1	28.1	25.9	26.1	27.0	24.3	25.8	24.7	31.2
	25	24.7	17.1	29.7	25.3	24.7	20.2	26.9	24.0	29.5
	50	27.7	35.4	25.5	24.6	31.6	25.1	11.0	26.5	28.4
	75	32.2	32.7	19.8	20.8	29.7	21.3	26.2	27.0	28.1
FaceDetection	0	64.8	64.8	65.6	63.3	64.4	66.5	64.8	63.6	63.1
	25	64.4	62.6	61.7	61.7	64.4	63.0	61.5	61.5	61.0
	50	61.3	61.2	62.6	60.2	62.9	61.3	62.0	59.3	58.7
	75	60.1	59.9	58.7	57.8	63.0	59.4	62.0	58.3	57.3
Handwring	0	33.6	33.6	26.1	27.2	34.5	33.3	34.8	21.2	32.0
	25	29.4	24.0	18.7	18.9	26.0	20.4	24.6	20.0	25.1
	50	20.4	18.6	12.1	17.3	20.9	20.0	20.8	18.8	14.4
	75	18.7	14.2	12.6	10.7	18.2	17.1	19.2	16.6	13.6
Heartbeat	0	74.1	74.1	66.8	68.8	70.7	69.3	71.2	70.7	70.7
	25	71.7	73.2	66.8	72.7	68.3	63.4	70.2	72.7	71.7
	50	73.2	56.1	71.2	74.1	67.8	62.9	71.7	71.6	70.7
	75	70.7	50.2	73.2	72.7	69.3	65.4	73.2	71.7	74.1
Japanese	0	95.1	95.1	95.6	93.1	96.2	94.7	90.1	95.6	95.0
	25	89.6	83.2	78.0	86.0	89.1	86.9	77.7	88.2	81.2
	50	83.8	68.1	74.6	65.6	75.3	75.3	56.1	75.2	78.9
	75	72.3	41.1	57.2	53.4	66.1	63.6	50.2	66.0	48.9
PEMS-SF	0	82.1	82.1	75.7	82.7	79.8	84.4	77.5	82.1	79.8
	25	80.3	80.3	71.1	82.7	83.2	82.1	82.7	76.1	76.3
	50	75.1	75.1	73.4	85.5	83.2	78.6	80.9	68.3	43.4
	75	74.9	54.9	63.0	31.8	75.1	71.1	71.1	63.6	59.8
SelfRegulationSCP1	0	87.0	87.0	86.3	88.1	86.0	85.0	85.7	81.9	83.3
	25	84.3	80.9	78.2	88.7	87.0	77.5	88.7	86.7	83.6
	50	83.3	79.9	71.7	73.4	84.0	78.2	89.1	84.2	75.1
	75	82.3	65.5	73.0	68.9	89.4	81.9	88.4	82.9	65.9
SelfRegulationSCP2	0	50.0	50.0	51.1	51.1	46.1	50.0	48.9	42.8	50.0
	25	51.7	44.4	55.0	52.8	48.3	55.6	48.9	43.9	50.0
	50	51.1	47.8	41.7	46.7	41.7	46.7	47.2	44.6	47.8
	75	51.7	47.8	50.6	52.8	47.8	52.8	52.2	48.9	46.1
SpokenArabicDigits	0	98.7	98.7	99.8	99.0	96.0	98.8	96.3	98.4	97.0
	25	95.0	93.6	88.2	85.3	85.9	86.6	88.3	85.6	85.6
	50	88.2	80.8	75.3	73.7	81.4	73.1	82.5	72.5	71.5
	75	70.5	49.7	49.5	59.5	67.5	51.0	57.0	50.8	59.5
UWaveGestureLibrary	0	87.8	87.8	85.6	86.3	85.9	87.2	85.9	78.4	86.9
	25	81.2	77.5	67.8	75.6	75.9	76.6	70.9	72.2	70.3
	50	78.8	57.5	61.6	66.3	62.5	67.2	53.4	60.3	53.8
	75	70.3	54.4	51.3	57.2	53.1	64.1	54.3	53.4	60.9
Average	-	**65.51**	59.73	59.31	60.21	63.25	61.55	61.25	60.27	59.51

Table 8 reports classification accuracy on ten datasets under missing ratios of 0%, 25%, 50%, and 75%. MissTime achieves the highest overall average accuracy of 65.51%, compared with 59.62% for TimesNet and 63.25% for SAITS, corresponding to improvements of 5.78 and 2.26 percentage points, respectively. These results demonstrate improved average classification performance across the evaluated datasets and missingness conditions.

The advantage over TimesNet becomes larger as missingness increases. Averaged across the ten datasets, the accuracy improvements are 3.55, 6.24, and 13.33 percentage points at 25%, 50%, and 75% missingness, respectively. At 75%, MissTime retains an average accuracy of 60.37%, whereas TimesNet drops to 47.04%. Particularly large gains occur on Japanese and SpokenArabicDigits,

where MissTime achieves 72.3% and 70.5%, compared with 41.1% and 49.7% for TimesNet. These comparisons indicate that MissTime can preserve useful information for classification when many variables are unavailable.

The improvements are not uniform across individual datasets. At 75% missingness, SAITS outperforms MissTime on FaceDetection and SelfRegulationSCP1, achieving 63.0% and 89.4% compared with 60.1% and 82.3%, respectively. Nevertheless, MissTime achieves the highest mean accuracy at each evaluated missing ratio. Overall, the results support its effectiveness in improving classification robustness under variable missingness, with particularly substantial average gains over TimesNet under severe missingness.

A.5 Additional Results on Anomaly Detection

Table 9: Full anomaly detection results on MSL, PSM, SMAP, and SMD under different missing ratios, where each method is evaluated with precision (P), recall (R), and F1-score (F1) in percent.

Models	Datasets	MSL			PSM			SMAP			SMD			Average
	Missing (%)	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
MissTime (**Ours**)	0%	89.6	74.6	81.5	98.4	94.9	96.6	90.0	56.6	69.5	88.5	82.2	85.2	83.2
	25%	89.3	74.5	81.2	98.1	96.0	97.0	88.7	55.5	68.3	88.1	82.4	85.1	**82.9**
	50%	89.7	72.7	80.3	97.9	96.0	97.0	88.5	55.9	68.5	88.3	76.9	84.2	**82.5**
	75%	90.6	82.3	86.3	97.3	95.3	96.3	91.4	70.7	79.7	85.6	75.7	84.3	**86.7**
TimesNet	0%	89.7	74.6	81.5	98.8	90.2	94.3	90.7	61.8	73.5	88.5	82.4	85.3	83.7
	25%	89.4	72.9	80.3	98.7	92.2	95.3	90.2	55.9	69.0	87.5	78.5	82.8	81.9
	50%	90.2	73.0	80.7	98.7	88.0	93.1	89.6	53.7	67.2	86.6	72.4	78.9	81.2
	75%	90.4	73.0	80.7	98.9	91.3	94.9	89.2	52.1	65.8	87.8	71.4	78.8	80.8
TimeMixer	0%	89.2	74.4	81.1	98.6	92.7	95.6	90.8	61.4	73.3	87.0	79.8	83.3	83.3
	25%	88.9	70.0	78.3	98.6	92.8	95.6	90.5	59.5	71.8	86.2	77.9	81.8	81.9
	50%	89.4	71.8	79.7	98.7	92.1	95.3	90.5	59.3	71.7	83.7	72.8	77.9	81.2
	75%	89.5	72.0	79.8	98.7	93.6	96.1	90.2	58.0	70.6	85.4	69.4	76.6	80.8
iTransformer	0%	87.7	70.2	78.0	97.9	92.8	95.3	90.7	52.9	66.8	87.3	79.0	82.9	80.8
	25%	88.0	70.8	78.5	97.9	91.8	94.8	90.5	53.0	66.8	85.6	74.7	79.8	80.0
	50%	88.1	71.4	78.9	97.8	87.7	92.5	90.3	53.0	66.8	85.7	77.0	81.2	79.9
	75%	88.5	66.4	75.9	98.5	92.4	95.4	89.8	52.1	66.0	86.3	68.4	76.3	78.4
SAITS	0%	89.2	73.9	80.8	98.4	94.7	96.5	90.0	56.6	69.5	88.6	82.8	85.6	83.1
	25%	89.2	72.4	79.9	98.5	93.9	96.2	89.8	55.8	68.8	86.1	78.6	82.2	81.8
	50%	89.8	72.6	79.2	98.6	95.5	96.0	89.0	55.7	67.8	89.4	79.1	80.9	81.0
	75%	89.1	70.8	78.9	99.0	95.0	97.0	89.0	55.4	68.6	86.8	72.7	79.1	80.9
GinAR	0%	88.9	74.4	81.0	98.3	89.8	93.9	90.1	55.3	68.5	86.9	80.2	83.4	81.7
	25%	88.7	72.2	79.6	98.5	88.9	93.4	89.8	53.1	66.7	86.5	77.2	81.6	80.3
	50%	89.0	73.5	80.5	98.8	90.0	94.2	89.6	53.6	67.1	85.9	73.2	79.1	80.2
	75%	89.3	74.7	81.3	98.9	90.3	94.4	89.6	53.6	67.1	80.4	56.9	66.6	77.4
TimeFilter	0%	89.3	74.5	81.3	97.9	92.2	95.0	89.4	53.9	67.2	87.2	81.1	84.0	81.9
	25%	89.4	74.6	81.3	98.1	92.9	95.4	89.3	53.6	67.0	86.3	76.2	81.0	81.2
	50%	89.3	74.1	81.0	98.3	92.4	95.2	89.1	53.6	66.9	85.3	74.7	79.7	80.7
	75%	89.0	71.2	79.1	98.4	91.6	94.9	88.9	53.4	66.7	84.2	70.1	76.5	79.3
RAFT	0%	89.9	75.1	81.8	98.7	94.7	96.6	89.9	53.8	67.3	87.3	81.0	84.0	82.4
	25%	90.0	72.7	80.4	98.7	94.3	96.4	90.0	53.8	67.4	87.3	79.9	83.4	81.9
	50%	89.2	70.8	78.9	98.9	93.4	96.1	90.0	53.4	67.0	87.6	77.1	82.0	81.0
	75%	88.9	70.0	78.4	98.9	90.8	94.7	89.6	53.1	66.7	85.2	61.4	71.4	77.9
TS-RAG	0%	89.6	74.7	81.5	98.5	95.0	96.7	90.7	61.8	73.5	88.5	82.1	85.2	**84.2**
	25%	89.5	76.7	82.6	97.7	94.4	96.0	89.6	55.7	68.7	87.7	81.1	84.3	82.9
	50%	89.9	77.0	82.9	97.9	94.2	96.0	89.5	53.6	67.1	86.2	74.0	79.6	81.4
	75%	90.0	79.7	84.5	97.5	93.7	95.6	89.2	52.1	65.7	85.6	63.7	73.0	79.7

Table 9 reports anomaly detection results on MSL, PSM, SMAP, and SMD under four missing ratios, using precision, recall, and F1-score. The fully observed setting provides a reference, while the remaining settings evaluate robustness when increasing proportions of input variables are unavailable.

MissTime achieves the highest reported average F1 at all three nonzero missing ratios, tying TS-RAG at 25%. On PSM, its F1 remains relatively stable, reaching 97.0% at both 25% and 50% missingness and 96.3% at 75%. At 75%, MissTime achieves 79.7% F1 on SMAP, exceeding TimeMixer by 9.1 percentage points, and 86.3% on MSL, outperforming TS-RAG by 1.8 points. The SMAP

improvement is accompanied by substantially higher recall (70.7% versus 58.0%) and slightly higher precision (91.4% versus 90.2%), showing that increased anomaly coverage does not come at the expense of precision.

At 75% missingness, MissTime exceeds the strongest baseline in average F1 by 4.5 percentage points. On SMD, its reported F1 decreases by only 0.9 percentage points relative to the fully observed setting, indicating resilience to severe variable missingness.

The improvements are not uniform across all settings. TS-RAG performs better on MSL and TimeMixer on SMAP at 25% and 50%, while SAITS achieves higher F1 on PSM at 75%. Overall, the results demonstrate competitive detection performance across datasets, with particularly strong gains under severe variable missingness.

A.6 Additional Results on Generality Analysis

Figure 10 examines the applicability of MissTime across two backbones and three downstream tasks through paired comparisons of each original model and its MissTime variant. The rows correspond to TimesNet and TimeMixer, while the columns show short-term forecasting on PEMS03, classification on SCP, and anomaly detection on SMD. Each pair is evaluated at missing ratios of 25%, 50%, and 75%, allowing us to assess whether the benefit persists across backbones, tasks, and levels of missingness.

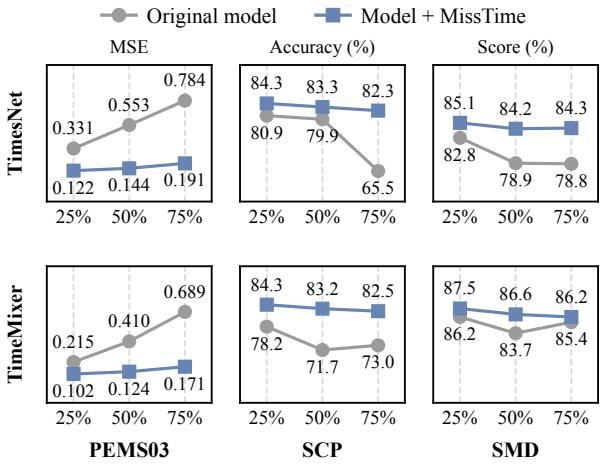


Figure 10: Generality analysis on PEMS03, SCP, and SMD at 25%, 50%, and 75% missing ratios. Gray circles and blue squares denote the original backbones and their MissTime variants, respectively. We report forecasting MSE, classification accuracy, and anomaly detection score.

On PEMS03, MissTime reduces forecasting MSE for both backbones at every missing ratio. For TimesNet, the relative MSE reductions are 63.1%, 74.0%, and 75.6% at 25%, 50%, and 75% missingness, respectively; the corresponding reductions for TimeMixer are 52.6%, 69.8%, and 75.2%. At 75%, TimesNet's MSE decreases from 0.784 to 0.191, while TimeMixer's decreases from 0.689 to 0.171. Thus, although forecasting error still increases as more variables become unavailable, MissTime substantially limits this increase for both backbones. The increasing relative reductions indicate that its forecasting benefit is particularly pronounced under severe variable missingness.

On SCP, MissTime improves classification accuracy across all three missing ratios. The gains for TimesNet are 3.4, 3.4, and 16.8 percentage points, while those for TimeMixer are 6.1, 11.5, and 9.5 points. With MissTime, accuracy decreases only from 84.3% to 82.3% for TimesNet and from 84.3% to 82.5% for TimeMixer as missingness increases from 25% to 75%. In comparison, the original TimesNet drops from 80.9% to 65.5%. These results show that the benefit extends to classification, where maintaining useful information under missing observations is reflected in more stable prediction accuracy.

On SMD, the reported detection scores also improve for both backbones at every missing ratio. TimesNet gains 2.3, 5.3, and 5.5 percentage points, reaching 85.1%, 84.2%, and 84.3%, respectively. TimeMixer gains 1.3, 2.9, and 0.8 points, with scores of 87.5%, 86.6%, and 86.2%. The magnitude of improvement depends on the backbone and missing ratio: TimesNet receives its largest gain at 75%, whereas TimeMixer benefits most at 50%. This variation complements the forecasting and classification results by showing consistent improvements without requiring the gains to increase monotonically in every task.

Taken together, the improvements across all six backbone–task combinations provide empirical support for the applicability of MissTime beyond a single forecasting architecture or prediction

objective. The results are consistent with retrieval-based reconstruction being useful for both continuous forecasting and discriminative tasks under variable-missing conditions. The size of the benefit remains task- and backbone-dependent, so these findings establish generality across the evaluated settings while motivating further evaluation on additional models and datasets.

A.7 Additional Results on Ablation Study

Table 10 reports ablation results on four datasets: PEMS04 for short-term forecasting, ECL for long-term forecasting, UGesture for classification, and SMD for anomaly detection. We compare the full MissTime model with variants that remove the topology-aware component, hard negative learning, or both. The results cover 25%, 50%, and 75% missing ratios, and Avg summarizes their mean.

For forecasting, the full model achieves the lowest MSE on both PEMS04 and ECL at all three missing ratios. On PEMS04, its average MSE is 0.114, compared with 0.166 without topology, 0.432 without hard negatives, and 0.419 without both components. The error increase from removing hard negatives becomes particularly pronounced at higher missing ratios. On ECL, the full model also achieves the lowest average MSE of 0.243, with the largest degradation observed when both components are removed.

For classification on UGesture, the full model consistently achieves higher accuracy than the ablated variants across the displayed missing ratios. Its average accuracy is 76.80%, compared with 75.82% without topology, 72.37% without hard negatives, and 69.77% without both components. These results suggest that both components contribute to classification performance, with a larger average decrease after removing hard negative learning.

For anomaly detection on SMD, the full model achieves the highest average F1 of 84.53% and remains comparatively stable as missingness increases. At 25% missingness, the variant without hard negatives scores slightly higher than the full model (85.14% versus 85.10%), but the full model performs best at 50% and 75%. Thus, the benefits are clearer under more severe missingness. Overall, the full model achieves the best average performance on all four datasets, supporting the contributions of both components while showing that their effects vary across tasks and missing ratios.

Table 10: **Ablation results under different missing ratios.** Avg reports the mean over 25%, 50%, and 75%. A leading "-" denotes component removal. The metrics are MSE on PEMS04 and ECL, accuracy on UGesture, and F1-score on SMD. **Bold** and underlined mark the best and second-best results.

Method	PEMS04	ECL	UGesture	SMD
	MSE ↓	MSE ↓	Acc.(%)↑	F1(%)↑
(a) 25% Missing Ratio				
MissTime (Full)	**0.105**	**0.206**	**81.30**	85.10
- Topology	0.128	0.207	81.25	84.21
- Hard negative	0.139	0.221	72.80	**85.14**
- Both	0.139	0.303	71.80	82.36
(b) 50% Missing Ratio				
MissTime (Full)	**0.115**	**0.241**	**78.80**	**84.20**
- Topology	0.137	0.248	77.80	80.92
- Hard negative	0.399	0.252	76.50	82.19
- Both	0.380	0.307	70.00	79.80
(c) 75% Missing Ratio				
MissTime (Full)	**0.122**	**0.282**	**70.30**	**84.30**
- Topology	0.234	0.289	68.40	77.80
- Hard negative	0.758	0.308	67.80	74.32
- Both	0.737	0.308	67.50	76.90
(d) Avg				
MissTime (Full)	**0.114**	**0.243**	**76.80**	**84.53**
- Topology	0.166	0.248	75.82	80.98
- Hard negative	0.432	0.260	72.37	80.55
- Both	0.419	0.306	69.77	79.69

A.8 Additional Results on Case Study

Figure 11 compares the retrieval reconstructions of MissTime and Vanilla on the Electricity dataset. The upper group presents six observed variables, whose values are available during retrieval, while the lower group presents six variables masked at test time. Each row compares both methods against the same ground truth over a 96-hour input window, using standardized values and identical vertical scales. The observed variables share the first test window, whereas the missing-variable windows are selected to illustrate cases where MissTime achieves lower reconstruction error.

For the observed variables, both methods capture the dominant periodic patterns, including the timing of peaks, troughs, and transitions between high- and low-value intervals. Their correlations with the ground truth exceed 0.97 in all six displayed examples. MissTime achieves lower RMSE in five of these examples, although Vanilla performs better in the fifth row. These results indicate that

both methods can closely reproduce the temporal structure of the displayed observed variables, with differences primarily in local amplitudes and signal levels.

Larger discrepancies appear in the selected missing-variable examples. In the first missing row, Vanilla exhibits a persistent upward offset, whereas in the second it substantially underestimates the signal level. In subsequent rows, Vanilla often retains the recurring temporal pattern but misestimates the magnitude of the reconstructed values. MissTime generally follows the ground-truth levels more closely, reducing RMSE by 67.7%–85.3% across the six selected windows.

These cases show that temporal similarity alone does not ensure accurate reconstruction. MissTime better recovers signal levels and amplitudes while preserving temporal structure. This evidence is limited to the selected windows and should be considered alongside aggregate test-set results.

A.9 Additional Results on Parameter Analysis

Figure 12 presents six heatmaps that complement the parameter analysis in Fig. 7 by reporting additional evaluation metrics on SMD, UGesture, and Electricity. Each heatmap evaluates the same nine combinations of retrieval size $k \in \{3, 5, 7\}$ and hard-negative mining temperature $\tau_w \in \{0.05, 0.10, 0.20\}$. Rows correspond to retrieval sizes, and columns correspond to temperature values. The reported metrics are Accuracy, Precision, and Recall for SMD, AUC and F1-score for UGesture, and MAE for Electricity. Higher values indicate better performance for the classification metrics, whereas lower MAE indicates better forecasting performance.

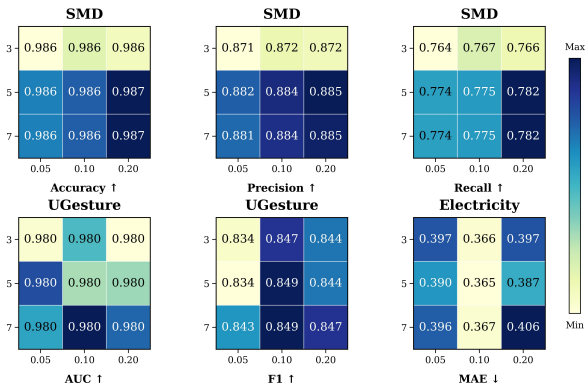


Figure 12: Additional hyperparameter sensitivity results of MissTime. Rows indicate $k \in \{3, 5, 7\}$, and columns indicate $\tau_w \in \{0.05, 0.10, 0.20\}$. Higher is better except for MAE. The panels report accuracy, precision, and recall on SMD, AUC and F1-score on UGesture, and MAE on Electricity.

These heatmaps allow us to examine both the effect of each hyperparameter and whether its effect changes with the other. Reading down a column reveals the influence of retrieval size at a fixed temperature, while reading across a row shows temperature sensitivity for a fixed retrieval size.

We compare the numerical values within each panel, since the metrics have different scales and optimization directions. In particular, the additional metrics help distinguish settings that appear similar under an aggregate measure but differ in precision, recall, or F1-score. The following analysis therefore focuses on the observed changes within each dataset and on whether a single configuration consistently performs best across metrics.

SMD. Accuracy remains between 0.986 and 0.987 across all evaluated configurations, whereas Precision and Recall show clearer differences. Increasing k from 3 to 5 improves both metrics at every temperature. At $\tau_w = 0.20$, for example, Precision increases from 0.872 to 0.885, and Recall increases from 0.766 to 0.782. Increasing k further to 7 provides no additional improvement in Recall and changes Precision only slightly. These results indicate diminishing gains from increasing retrieval size within the tested range: five retrieved instances already attain the highest reported Precision and Recall when paired with $\tau_w = 0.20$.

UGesture. AUC remains at 0.980 to the reported precision for all nine configurations, but F1-score varies from 0.819 to 0.849. The highest F1-score is obtained at $k = 7$ and $\tau_w = 0.10$. However, the same temperature yields the lowest F1-score when $k = 5$, showing that its effect depends on retrieval size. In comparison, $\tau_w = 0.20$ produces F1-scores of 0.844, 0.844, and 0.847 for $k = 3$, 5, and 7, respectively. Thus, the unchanged AUC does not imply that all configurations are equivalent: F1-score reveals sensitivity that is not visible in AUC.

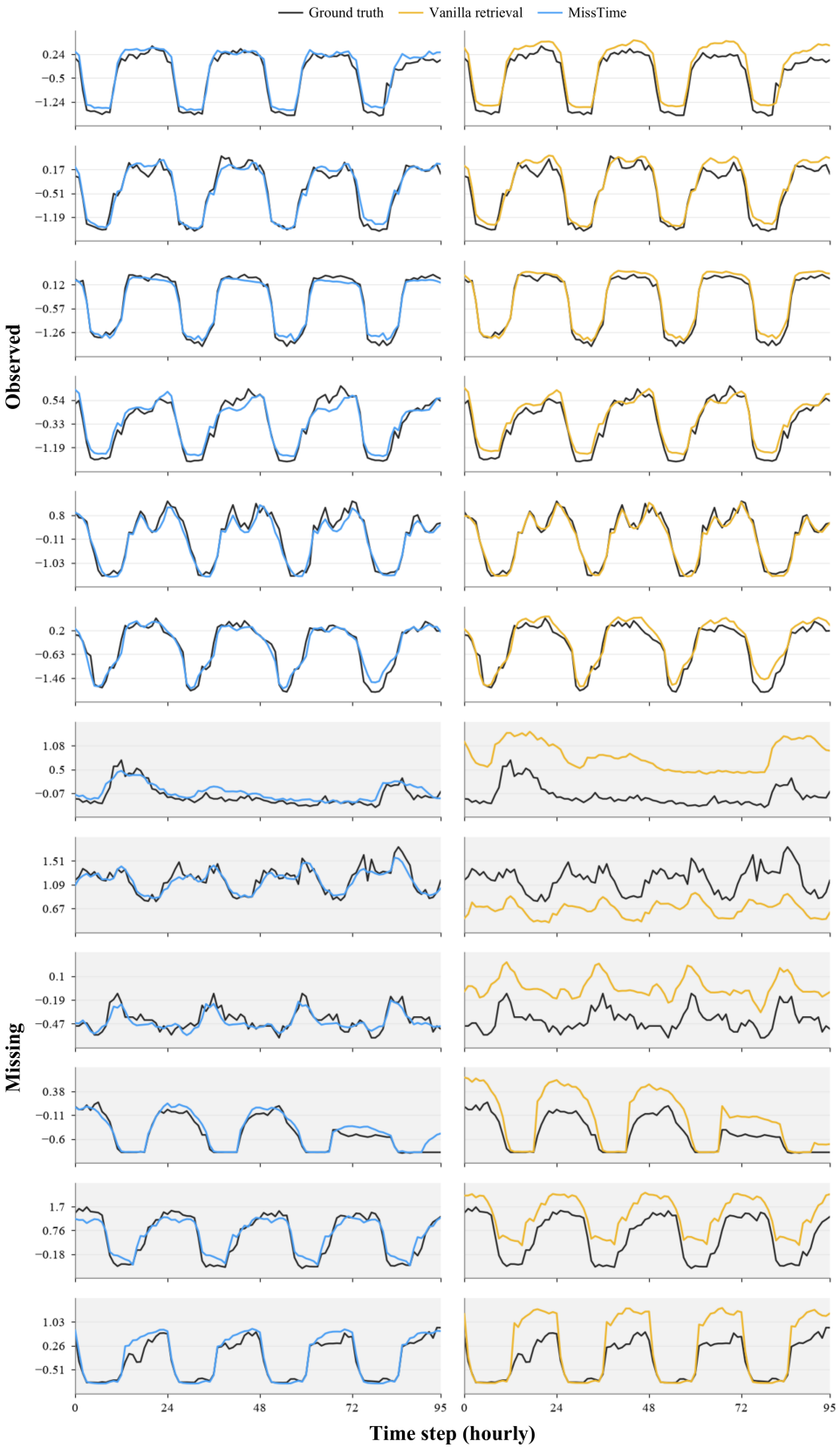


Figure 11: Reconstruction of six observed (top) and six selected missing (bottom) variables on Electricity. MissTime (left, blue) and Vanilla (right, gold) are compared with ground truth (black).

Electricity. The temperature exhibits a consistent pattern across retrieval sizes. Setting $\tau_w = 0.10$ gives the lowest MAE in every row, with values of 0.366, 0.365, and 0.371 for $k = 3$, 5, and 7, respectively. The best evaluated configuration is therefore $k = 5, \tau_w = 0.10$. At this retrieval size, changing the temperature to 0.05 or 0.20 increases MAE to 0.390 or 0.387. Moreover, increasing k from 5 to 7 raises MAE at every tested temperature, indicating that retrieving more instances does not necessarily improve forecasting performance.

Overall conclusion. The additional results show that hyperparameter sensitivity depends on both the dataset and the evaluation metric. SMD Accuracy and UGesture AUC remain nearly unchanged, while Precision, Recall, F1-score, and MAE reveal more differentiated responses. A retrieval size of $k = 5$ reaches the best observed results on SMD and Electricity with suitable temperatures, but UGesture achieves its highest F1-score at $k = 7$. These observations support a modest retrieval size as a practical starting point, while favoring validation-based selection of the temperature and retrieval size rather than a universally optimal configuration.

B Details of Experimental Setup

B.1 Dataset Specifications

We evaluate MissTime on five time series tasks, with detailed dataset statistics summarized in Table 11. The table reports the number of variables, sequence configurations, application domains, and dataset sizes in the order of training, validation, and test sets. For long-term forecasting, we use ETT, Electricity, Traffic, and Weather with prediction horizons of $\{96, 192, 336, 720\}$. Short-term forecasting is evaluated on PEMS03, PEMS04, PEMS07, and PEMS08, which contain traffic measurements sampled every five minutes, using a prediction horizon of 12 steps. For imputation, we use ETT, Electricity, and Weather with a sequence length of 96. Classification experiments cover ten multivariate datasets from the UEA archive, with dimensions ranging from 3 to 963 and sequence lengths from 29 to 1751. For anomaly detection, we use SMD, MSL, SMAP, and PSM with a window length of 100. These datasets span diverse domains and vary substantially in dimensionality, temporal resolution, and data volume, enabling evaluation across different time series characteristics and task requirements.

B.2 Baseline Methods

We describe the compared methods and/or how we modify them to our variable missing settings in detail below. Firstly, we compare MissTime with SOTA and representative time series forecasters without RAG, including:

- **TimesNet** Wu et al. (2023) is a representative task-general backbone for time series analysis by transforming the 1D time series into a set of 2D tensors based on multiple periods. It achieves consistent performance in five mainstream time series analysis tasks, including short- and long-term forecasting, imputation, classification, and anomaly detection.
- **TimeMixer** Wang et al. (2024) is a fully MLP-based architecture that takes full advantage of disentangled multiscale series in both past extraction and future prediction phases.
- **iTransformer** Liu et al. (2024b) is a Transformer-based forecaster that embeds each channel of the time series sequence as a separate token and leverages self-attention to model the global dependencies of multivariate time series. It is also experimentally verified to perform well on various tasks.

Secondly, we compare with existing data imputation methods, including the two-stage and end-to-end methods:

- **SAITS** Du et al. (2023) is a representative imputation method for multivariate time series based on the self-attention mechanism. It captures both the temporal dependencies and channel correlations in order to fully utilize the local time series to fill the missingness.
- **GinAR** Yu et al. (2024) is an end-to-end network that combines interpolation attention across channels and a graph-enhanced recursive network to address the missing variable challenge.

- **HCIB** Zhang et al. (2025) focuses on classification-oriented representation learning with GRU-based imputation simultaneously.

Then, we include one recent study with specialized designs for mitigating channel noise into comparision:

- **TimeFilter** Hu et al. (2025) is a GNN-based framework that captures the patch-level spatial-temporal dependency with a graph. It then applies a Mixture-of-Experts (MoE)-style dynamic router to filter out noisy or redundant channels.

Lastly, we include two RAG-enhanced time series forecasting models as retrieval-based baselines:

- **TS-RAG** Ning et al. (2026) is a retrieval-augmented framework designed to enhance the zero-shot forecasting performance of time series foundation models. We adopt its retriever module, which uses a pre-trained time series encoder to retrieve semantically historical segments, and leverage the retrieved segments to impute the missing channels.
- **RAFT** Han et al. (2025) is a retrieval-augmented time series forecasting method that exploits historical patterns to improve linear forecasting models. We use its multi-period Pearson correlation-based retriever to retrieve similar historical sequences and impute the missing channels.

Table 11: Dataset detailed descriptions. The dataset size is organized in (Train, Validation, Test).

Tasks	Dataset	Dim	Series Length	Dataset Size	Information (Frequency)
Long-term Forecasting	ETT	7	{96, 192, 336, 720}	(8545, 2881, 2881)	Electricity (15 min)
	Electricity	321	{96, 192, 336, 720}	(18317, 2633, 5261)	Electricity (Hourly)
	Traffic	862	{96, 192, 336, 720}	(12185, 1757, 3509)	Transportation (Hourly)
	Weather	21	{96, 192, 336, 720}	(36792, 5271, 10540)	Weather (10 min)
Short-term Forecasting	PEMS03	358	12	(15617,5135,5135)	Transportation (5min)
	PEMS04	307	12	(10172,3375,3375)	Transportation (5min)
	PEMS07	883	12	(16911,5622,5622)	Transportation (5min)
	PEMS08	170	12	(10690,3548,265)	Transportation (5min)
Imputation	ETT	7	96	(8545, 2881, 2881)	Electricity (15 mins)
	Electricity	321	96	(18317, 2633, 5261)	Electricity (15 mins)
	Weather	21	96	(36792, 5271, 10540)	Weather (10 mins)
Classification (UEA)	EthanolConcentration	3	1751	(261, 0, 263)	Alcohol Industry
	FaceDetection	144	62	(5890, 0, 3524)	Face (250Hz)
	Handwriting	3	152	(150, 0, 850)	Handwriting
	Heartbeat	61	405	(204, 0, 205)	Heart Beat
	JapaneseVowels	12	29	(270, 0, 370)	Voice
	PEMS-SF	963	144	(267, 0, 173)	Transportation (Daily)
	SelfRegulationSCP1	6	896	(268, 0, 293)	Health (256Hz)
	SelfRegulationSCP2	7	1152	(200, 0, 180)	Health (256Hz)
	SpokenArabicDigits	13	93	(6599, 0, 2199)	Voice (11025Hz)
	UWaveGestureLibrary	3	315	(120, 0, 320)	Gesture
Anomaly Detection	SMD	38	100	(566724, 141681, 708420)	Server Machine
	MSL	55	100	(44653, 11664, 73729)	Spacecraft
	SMAP	25	100	(108146, 27037, 427617)	Spacecraft
	PSM	25	100	(105984, 26497, 87841)	Server Machine

B.3 Experimental Configurations

Device Details. All experiments are implemented in PyTorch and conducted on one NVIDIA A6000 48GB GPU. We use the Adam optimizer with an initial learning rate of $1e^{-4}$.

Metric Details. To provide a thorough and multi-faceted assessment of the model capabilities across diverse analytic scenarios, we employ a comprehensive set of standard evaluation metrics that are widely established within the time series analysis community.

For the tasks of forecasting (covering both short-term and long-term horizons) and imputation, our primary objective is to minimize the deviation between the generated sequence and the actual data. To quantify this, we utilize the Mean Squared Error (MSE) and Mean Absolute Error (MAE). The Mean Squared Error is particularly sensitive to large deviations due to the quadratic penalty term, thereby highlighting instances where the model makes significant prediction errors. Conversely, the Mean Absolute Error provides a linear assessment of the average error magnitude, offering a perspective that is more robust to outliers. In the context of multivariate time series, these metrics are computed by averaging the errors across all predicted time steps and all variable dimensions. Lower values for both metrics indicate a higher degree of predictive precision. These metrics are formally defined as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2, \quad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\mathbf{y}_i - \hat{\mathbf{y}}_i|, \tag{8}$$

where $\mathbf{y}_i$ denotes the ground truth vector and $\hat{\mathbf{y}}_i$ represents the predicted value for the i-th sample, while N signifies the total number of samples in the evaluation set.

For the classification task, we employ Accuracy as the primary performance indicator to evaluate the effectiveness of the model in correctly mapping complex temporal dynamics to discrete class labels. This metric serves as a direct measure of generalization capability. It is calculated as the ratio of the number of samples correctly assigned to their respective ground truth classes to the total number of samples in the dataset:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}}. \tag{9}$$

In the domain of anomaly detection, data distributions are often highly imbalanced because anomalous events occur far less frequently than normal patterns. Consequently, simple accuracy can be misleading. To address this, we follow standard protocols by reporting Precision (P), Recall (R), and the F1-score ($F1$). Precision measures the reliability of the alarms raised by the model by quantifying the proportion of true anomalies among all detected anomalies. Recall assesses the coverage of the model by measuring the proportion of actual anomalies that were successfully identified. A higher F1-score indicates that the model achieves a robust trade-off between minimizing false alarms and avoiding missed detections.

Experiment Details. All experimental evaluations were implemented using the PyTorch deep learning framework and executed on a computational platform equipped with a single NVIDIA A6000 48G GPU. To guarantee the reproducibility and statistical reliability of our results, we fixed the random seeds for the CPU, GPU, and all numpy operations at the beginning of each experimental run. The model parameters were optimized using the Adam algorithm, which was selected for its efficiency in handling non-stationary objectives and sparse gradients. We determined the optimal initial learning rate through a grid search within the range of 10^{-4} to 10^{-3}. To further stabilize the training process and facilitate convergence to optimal minima, we employed a dynamic learning rate adjustment strategy. Specifically, we utilized a cosine annealing scheduler or a step decay policy to gradually reduce the learning rate as the training epochs progressed. We also incorporated an early stopping mechanism that monitors the loss on the validation set. The training process was automatically terminated if the validation performance did not improve for a predefined number of consecutive epochs, which effectively mitigates the risk of overfitting. Prior to feeding the data into the network, all input time series were standardized using Z-score normalization to ensure zero mean and unit variance across variables. The batch size was rigorously tuned to either 16 or 32 to balance memory utilization and gradient estimation accuracy, contingent upon the specific scale of the dataset and the complexity of the model architecture.

- **Missing Variable Settings:** To evaluate the robustness of *MissTime* and baselines under variable missingness, we randomly mask variables (channels) from the input multivariate time series. We experiment with missing ratios of $\{0\%, 25\%, 50\%, 75\%\}$ across all tasks.

- **Forecasting:** For long-term forecasting on ETT, Electricity, Traffic, and Weather datasets, we set the input sequence length $L = 96$ and evaluate on prediction horizons $T \in \{96, 192, 336, 720\}$. For short-term forecasting on PEMS datasets, the input length is $L = 96$ and the prediction horizon is $T = 12$.
- **Imputation:** We perform imputation experiments on ETT, Electricity, and Weather datasets. The input length is fixed at 96. Models are tasked with reconstructing the missing values in the masked channels.
- **Classification:** We utilize the UEA multivariate time series classification archive. The data splitting follows the official training/testing sets provided by the archive.
- **Anomaly Detection:** Experiments are conducted on SMD, MSL, SMAP, and PSM datasets using a sliding window approach with a window size of 100. We employ the reconstruction error as the anomaly score criterion.

B.4 Implementation Details of the Preliminary Retrieval Study

We conduct the preliminary study on the Weather dataset using TimesNet as the forecasting backbone. We compare vanilla direct retrieval with the backbone baseline and SAITS-based imputation under different missing ratios. This section details the retrieval procedure, reconstruction process, and correlation metrics.

Historical bank and query construction. We construct a historical bank $\mathcal{D} = \{\mathbf{X}_i\}_{i=1}^{n}$, where each $\mathbf{X}_i \in \mathbb{R}^{L \times C}$ is a fully observed historical sample with L time steps and C variables. For an incomplete test instance $\mathbf{X}_{\text{test}} \in \mathbb{R}^{L \times C}$, let $\mathcal{O}$ and $\mathcal{U}$ denote the sets of observed and missing channels, respectively. We concatenate the values in the observed channels to form the retrieval query and extract the corresponding values from each historical sample:

$$\mathbf{q}_{\text{obs}} = \text{vec}\left(\mathbf{X}_{\text{test}}[:, \mathcal{O}]\right), \qquad \mathbf{k}_{i,\text{obs}} = \text{vec}\left(\mathbf{X}_i[:, \mathcal{O}]\right), \tag{10}$$

where $\text{vec}(\cdot)$ denotes vectorization. The same channel set and vectorization order are used for the query and all candidates.

Similarity-based retrieval. We compute the feature similarity between the query and each candidate using only the observed channels:

$$s_i = \text{Sim}\left(\mathbf{q}_{\text{obs}}, \mathbf{k}_{i,\text{obs}}\right). \tag{11}$$

Let $\mathcal{I}_K$ denote the indices of the K candidates with the highest similarity scores:

$$\mathcal{I}_K = \text{TopK}\left(\{s_i\}_{i=1}^{n}\right). \tag{12}$$

Here, $\text{TopK}(\cdot)$ returns candidate indices. The selected historical samples provide values for all channels, including those missing from the test instance.

Retrieval-based reconstruction. We compute a weighted aggregate of the retrieved samples:

$$\widetilde{\mathbf{X}} = \sum_{i \in \mathcal{I}_K} w_i \mathbf{X}_i, \qquad w_i \geq 0, \qquad \sum_{i \in \mathcal{I}_K} w_i = 1, \tag{13}$$

where w_i denotes the aggregation weight assigned to candidate i. We use this aggregate to fill the missing channels while preserving the observed values:

$$X_{t,c}^{\text{rec}} = \begin{cases} X_{\text{test},t,c}, & c \in \mathcal{O}, \\ \widetilde{X}_{t,c}, & c \in \mathcal{U}. \end{cases} \tag{14}$$

The completed input $\mathbf{X}^{\text{rec}}$ is then fed into TimesNet for forecasting. No additional imputation model is used to generate the missing values.

Retrieval correlation. We evaluate whether the retrieved values reflect the ground truth separately in observed and missing channels. For test instance j, let $\mathbf{X}_j^*$ denote its complete ground-truth sequence and $\widetilde{\mathbf{X}}_j$ denote the weighted retrieval aggregate before replacement of observed values. For either channel set $\mathcal{S} \in \{\mathcal{O}_j, \mathcal{U}_j\}$, we define

$$\mathbf{x}_{j,\mathcal{S}} = \mathrm{vec}\left(\mathbf{X}_j^*[:,\mathcal{S}]\right), \qquad \widetilde{\mathbf{x}}_{j,\mathcal{S}} = \mathrm{vec}\left(\widetilde{\mathbf{X}}_j[:,\mathcal{S}]\right). \tag{15}$$

We compute Pearson correlation for each instance and average the results over N evaluated test instances:

$$\mathrm{Corr}_{\mathrm{obs}} = \frac{1}{N}\sum_{j=1}^{N} \rho\left(\mathbf{x}_{j,\mathcal{O}_j}, \widetilde{\mathbf{x}}_{j,\mathcal{O}_j}\right), \tag{16}$$

$$\mathrm{Corr}_{\mathrm{miss}} = \frac{1}{N}\sum_{j=1}^{N} \rho\left(\mathbf{x}_{j,\mathcal{U}_j}, \widetilde{\mathbf{x}}_{j,\mathcal{U}_j}\right), \tag{17}$$

where $\rho(\cdot,\cdot)$ denotes Pearson correlation. These metrics are reported as Observed Retrieval Corr. and Unobserved Retrieval Corr., respectively. Ground-truth values in missing channels are used only for evaluation and are unavailable to the retrieval procedure.

C FAQs

C.1 Is the assumption of a fully observed historical database practical?

Firstly, although historical data may also contain a certain level of missingness, MissTime's Full Historical Bank Collection module addresses this challenge by curating historical data with fully observed variables, providing the basis for data reconstruction. Secondly, the assumption of a complete historical dataset represents a standard and practical setting in many real-world "offline-to-online" deployments Yu et al. (2024; 2025a); Du et al. (2023). While streaming data during online inference often suffers from missing variables due to sensor or communication failures, controlled data curation can ensure a basic level of data integrity during the data collection process Ronkainen & Iivari (2015); Ding et al. (2024). For instance, during collection phases, redundant sensors can be deployed and closely monitored to ensure data integrity. Furthermore, modern data cleaning techniques provide reliable solutions for effective offline quality control and data repair.

Beyond this assumption, MissTime is also highly robust even when the historical training data is partially observed. To validate this, we conduct experiments under two training missing scenarios: (1) *missing variables*, where a portion p of variables are randomly missing; and (2) *missing time steps*, where all variables of a certain length l of time steps are missing. The missing ratio of the test data is fixed at 25%. As shown in Table 12, MissTime consistently achieves better performance than existing imputation (SAITS) and end-to-end models (GinAR) under both settings of training data missingness.

Table 12: Forecasting performance (MSE) on PEMS03 with partially observed training data.

Models	Missing Variables Ratio (p)				Missing Time Steps Length (l)			
	0%	10%	20%	30%	$l=3$	$l=6$	$l=9$	$l=12$
MissTime	**0.122**	**0.128**	**0.149**	**0.199**	**0.122**	**0.124**	**0.133**	**0.139**
SAITS	0.131	0.149	0.177	0.208	0.131	0.139	0.147	0.158
GinAR	0.303	0.343	0.349	0.358	0.303	0.313	0.318	0.322

C.2 How does the contrastive state disambiguation handle hard negatives?

The goal of the contrastive state disambiguation objective is to capture subtle but critical differences about the underlying system states from partial observations. Assigning equal weights to all negative samples is ineffective because most partial-complete sample pairs are highly dissimilar across all channels. This leads to large distances in the latent space, making them too easy to separate.

Consequently, the model tends to stay in a comfort zone and becomes less effective in learning discriminative representations.

To this end, our ambiguity-aware weighting dynamically assigns larger weights to "hard negative samples." These samples satisfy two criteria: they are highly similar under the observed (partial) channels but distinctly different under the unobserved channels. By explicitly pulling away these deceptive look-alike samples, the objective facilitates the identification of subtle differences and effectively eliminates the retrieval ambiguity caused by partial observations.

C.3 What is the computational and memory overhead of the retrieval pipeline?

To assess the computational and memory overhead, we provide a detailed breakdown of the retrieval pipeline. As shown in Table 13, the time cost of constructing the knowledge base (KB) and performing retrieval is extremely light and comparable to standard model training and inference.

In terms of memory, the knowledge base requires only about 100 MB of CPU memory (for storing embeddings and raw data), which is negligible compared to modern TB-scale memory capacities. The additional GPU memory required for the retrieval operation is merely around 40 MB. The overall space complexity grows linearly with the dataset size $\mathcal{O}(N \cdot C \cdot D)$, ensuring that MissTime is highly scalable and practical for real-world deployments.

Table 13: Time and memory breakdown analysis of the retrieval pipeline.

Time Breakdown (Seconds)					
Dataset	Encoder Train	KB Construct	Model Train	Retrieval	Inference
ETT	70.67	1.23	150.13	2.01	0.52
Traffic	1198.30	38.30	1200.59	30.80	4.50
PEMS07	2185.87	55.32	940.29	25.60	6.30
Memory Breakdown (MB)					
Item	Encoder	KB (Embedding)	KB (Data)	Model	Retrieving
Cost	10.5	115.51	21.66	2.5	40.2

C.4 How does MissTime perform under harder missingness scenarios and distribution shifts?

Out-of-distribution (OOD) scenarios and non-random continuous missingness are critical challenges in time series analysis. Under such harder missingness scenarios, MissTime is naturally more robust than existing imputation approaches. Existing models rely on a "guessing" mechanism, which may introduce substantial errors when encountering unseen patterns or contiguous block missingness. In contrast, MissTime captures underlying system states and retrieves authentic historical patterns with similar temporal dynamics, anchoring the imputations to real-world physics.

To validate this property, we conduct experiments on the Electricity dataset under two harder scenarios: (1) *OOD data*, simulated by injecting varying levels of Gaussian noise ($X' = X + \epsilon$) into the test data; and (2) *Block Missingness*, where contiguous segments of length l are completely masked. Table 14 shows that MissTime significantly outperforms baselines, confirming that retrieving realistic historical references effectively reduces the discrepancy introduced by unseen distributions and continuous data loss.

Table 14: Forecasting performance (MSE) on Electricity under OOD and block missingness scenarios.

Models	OOD Noise Level (ϵ)				Continuous Block Length (l)			
	0%	10%	20%	30%	$l=3$	$l=6$	$l=9$	$l=12$
MissTime	**0.225**	**0.228**	**0.233**	**0.256**	**0.171**	**0.185**	**0.193**	**0.208**
SAITS	0.233	0.245	0.268	0.289	0.174	0.188	0.196	0.216
GinAR	0.360	0.412	0.473	0.594	0.222	0.235	0.246	0.277

C.5 Does the Top-k weighted sum cause oversmoothing?

A common concern with weighted aggregation strategies is the potential oversmoothing of sharp temporal patterns or subtle anomaly signals. However, time series data is inherently driven by underlying physical or systemic states, leading to recurring temporal patterns.

Our state-aware retrieval mechanism explicitly captures this whole-system state, ensuring that the retrieved instances are not just superficially similar, but represent the true, recurring physics of the system. To validate that aggregation does not wash out critical signals, we conduct a sensitivity analysis on the Top-k parameter for anomaly detection on the SMD dataset (Table 15). The results indicate that rather than causing oversmoothing, an appropriate aggregation ($k = 10$) of faithful historical segments actually enhances the robustness of anomaly detection under missing variables.

Table 15: Anomaly detection F1-Score (%) on SMD with different Top-k retrieval sizes.

Retrieval Size	$k = 1$	$k = 3$	$k = 5$	$k = 10$
F1-Score (%)	85.2	85.1	85.6	**86.2**

C.6 Why report results under the 0% missingness setting?

Under a 0% missingness ratio, no variables are masked in the input data. Consequently, MissTime does not trigger the retrieval and imputation pipeline. In this scenario, the data flows directly into the downstream model, meaning the performance of MissTime is theoretically and mathematically identical to the vanilla downstream backbone.

We deliberately include the 0% setting in our comprehensive results tables to establish an absolute performance anchor. This serves two purposes: (1) It prevents situations where certain baseline architectures inherently perform better under fully observed conditions, which could obscure the actual robustness improvements when missingness occurs; (2) It ensures a fair, baseline-normalized comparison of performance degradation across all evaluated methods. *(Note: A minor typographical discrepancy for the ETTh1 0% result in the initial submission has been corrected in the final tables to strictly match the vanilla backbone).*

D Related Work

D.1 Modern Time Series Analysis

Time series analysis plays a crucial role in identifying and predicting temporal patterns across a wide range of domains and tasks, including forecasting Nti et al. (2020), classification Kiyasseh et al. (2021), anomaly detection Rabatel et al. (2011), and imputation Doran (1974). The evolution of modern general time series models (GTSMs) in recent years has reshaped the paradigm of time series analysis by introducing general-purpose architectures Zhou et al. (2023); Wu et al. (2023); Wang et al. (2025) and time series foundation models (TSFMs) Goswami et al. (2024); Liu et al. (2024c); Wang et al. (2025); Liu et al. (2025c). Among them, TimesNet Wu et al. (2023) and TimeMixer++ Wang et al. (2025) aim to create a unified model architecture capable of handling a broad range of time series tasks across domains. GPT4TS Zhou et al. (2023) explores this objective through a pre-trained GPT2 language model. Concurrently, Timer Liu et al. (2024c) and Moment Goswami et al. (2024) unify the multiple tasks via large-scale pre-training. Moirai Woo et al. (2024), Chronos Ansari et al. (2024) and TimerXL Liu et al. (2025b) have further advanced the development of zero-shot forecasting.

D.2 Handling Missing Values in Time Series

Missing data is common in time series and can severely impact downstream tasks like forecasting and classification. The mainstream imputation-based approaches first impute missing values with an imputation model and then applies the downstream tasks. Imputation methods range from simple mean substitution KA et al. (2022) to advanced deep learning methods via Transformer Du et al. (2023), VAEs Fortuin et al. (2020) and Diffusion Models Zhou et al. (2024). Yet, they suffer from decoupled design and imputation bias. To address this, representation learning-based models have been proposed involving specialized network designs for joint representation recovery and

downstream task optimization. RainDrop Zhang et al. (2022) uses GNN to embed and classify irregularly sampled time series. G-MS Liu et al. (2025a) and HCIB Zhang et al. (2025) focus on classification-oriented representation learning with GRU-based imputation and an information bottleneck approach, respectively. GinAR Yu et al. (2024; 2025a) proposes a graph-enhanced recursive network to handle the variable missing challenge in time series forecasting, while Merlin Yu et al. (2025b) utilizes teacher-student model to align the incomplete embedding with complete ones. Existing approaches either impute missing values explicitly via imputation models or implicitly via end-to-end architectures, both of which rely on 'guessing' missing data and fail to generalize well across diverse tasks. In contrast, we aim to address this issue through a task-agnostic and plug-in framework, greatly facilitated by the general-purpose design and successful pre-training of MTSMs.

D.3 Retrieval-Augmented Techniques in Time Series

Recent works have explored the integration of retrieval mechanisms into time series. Among them, several approaches integrate the retrieval technique with non-pretraining time series forecasting models including ReTime Jing et al. (2022), RATD Liu et al. (2024a) and RAFT Han et al. (2025). Several other appraoches integrate the retrieval with pre-trained TSFMs for zero-shot forecasting improvement,including TimeRAG Yang et al. (2025), RAF Tire et al. (2024) and TS-RAG Ning et al. (2025). TRACE Chen et al. (2025) is a multimodal time-series retriever with aligned context embeddings, supporting TSFM with text-aware retrieval.

E Algorithms of MissTime

Algorithm 1 Overall Pipeline of MissTime

Require: Full training set $\mathcal{D}$; historical collection $\mathcal{X}_{\text{hist}}$; zero-masked test inputs $\mathcal{Q} = \{(\mathbf{X}_{\text{obs}}^{(q)}, \mathbf{m}^{(q)})\}$; initial encoder parameters θ; training settings (N, τ, τ_w); retrieval size k; frozen downstream model f

Ensure: Reconstructed inputs and task predictions for all test instances

Offline preparation (once)

$\mathcal{X}_{\text{filtered}} \leftarrow \{\mathbf{X} \in \mathcal{X}_{\text{hist}} \mid m(\mathbf{X}) < 0.01\}$ ▷ Filter historical windows

$\theta \leftarrow \textsc{TrainRetriever}(\mathcal{D}, \theta, N, \tau, \tau_w)$ ▷ Algorithm 2

$\mathcal{B} \leftarrow \{(\mathbf{X}_i, \textsc{StateEncode}(\mathbf{X}_i, \mathbf{0}; \theta)) \mid \mathbf{X}_i \in \mathcal{X}_{\text{filtered}}\}$

Online inference (reuse the encoder and bank)

for all $(\mathbf{X}_{\text{obs}}, \mathbf{m}) \in \mathcal{Q}$ **do**

  Broadcast $\mathbf{m}$ over time to obtain $\mathbf{M}$

  if $\mathbf{M} = \mathbf{0}$ **then**

    $\widehat{\mathbf{X}} \leftarrow \mathbf{X}_{\text{obs}}$ ▷ Bypass retrieval at 0% missingness

  else

    $\mathbf{h}^{\text{par}} \leftarrow \textsc{StateEncode}(\mathbf{X}_{\text{obs}}, \mathbf{m}; \theta)$

    $s_i \leftarrow \text{sim}(\mathbf{h}^{\text{par}}, \mathbf{e}_i)$ for all $(\mathbf{X}_i, \mathbf{e}_i) \in \mathcal{B}$

    $\mathcal{C} \leftarrow \text{TopK}_k\left(\{(\mathbf{X}_i, s_i)\}_{i=1}^{|\mathcal{B}|}\right)$

    $\widehat{\mathbf{X}} \leftarrow \mathbf{X}_{\text{obs}} + \mathbf{M} \odot \text{WeightedSum}(\mathcal{C})$ ▷ Fill missing variables

  end if

  $\mathbf{y} \leftarrow f(\widehat{\mathbf{X}})$ ▷ Downstream prediction

  output $(\widehat{\mathbf{X}}, \mathbf{y})$

end for

Algorithm 2 State-Aware Retrieval: Retriever Training

Require: Full training set $\mathcal{D}$; initial encoder parameters θ; batch size N; contrastive temperature τ; hard-negative weighting temperature τ_w
Ensure: Trained parameters θ of the topology-aware retriever

function TRAINRETRIEVER($\mathcal{D}, \theta, N, \tau, \tau_w$)
  while training has not converged **do**
    Sample a minibatch $\{\mathbf{X}_i\}_{i=1}^{N}$ from $\mathcal{D}$
    Sample channel masks $\mathbf{m}_i$ and their temporal broadcasts $\mathbf{M}_i$, for all i
    $\mathbf{X}_i^{\mathrm{par}} \leftarrow (\mathbf{1} - \mathbf{M}_i) \odot \mathbf{X}_i$ for all i ▷ Construct partial views
    $\mathbf{h}_i^{\mathrm{par}} \leftarrow \textsc{StateEncode}(\mathbf{X}_i^{\mathrm{par}}, \mathbf{m}_i; \theta)$ for all i
    $\mathbf{h}_i^{\mathrm{full}} \leftarrow \textsc{StateEncode}(\mathbf{X}_i, \mathbf{0}; \theta)$ for all i ▷ Same-sample positives
    $\omega_{i,i} \leftarrow 1$ for all i
    $\omega_{i,j} \leftarrow \exp\left(\dfrac{\mathrm{sim}(\mathbf{h}_i^{\mathrm{par}}, \mathbf{h}_j^{\mathrm{par}})}{\tau_w}\right)\left[1 - \mathrm{sim}(\mathbf{h}_i^{\mathrm{full}}, \mathbf{h}_j^{\mathrm{full}})\right]$ for all $i \neq j$
    $a_{i,j} \leftarrow \exp\big(\mathrm{sim}(\mathbf{h}_i^{\mathrm{par}}, \mathbf{h}_j^{\mathrm{full}})/\tau\big)$ for all i, j
    $\mathcal{L} \leftarrow -\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \log \dfrac{a_{i,i}}{\sum_{j=1}^{N} \omega_{i,j} a_{i,j}}$ ▷ Weighted InfoNCE
    Update θ to minimize $\mathcal{L}$
  end while
  return θ
end function

function STATEENCODE($\mathbf{X}, \mathbf{m}; \theta$)
  $\mathbf{Z}_c^{\mathrm{com}} \leftarrow \mathbf{E}_{\mathrm{value}}(\mathbf{X}_c) + \mathbf{E}_{\mathrm{identity}}(\mathrm{ID}_c) + \mathbf{E}_{\mathrm{obs}}(m_c)$ for $c = 1, \ldots, C$
  $\mathbf{Z}^{\mathrm{com}} \leftarrow [\mathbf{Z}_1^{\mathrm{com}}, \ldots, \mathbf{Z}_C^{\mathrm{com}}]$
  Derive $\mathbf{Q}_Z, \mathbf{K}_Z, \mathbf{V}_Z$ from $\mathbf{Z}^{\mathrm{com}}$
  return $\mathrm{Channel\text{-}Attn}(\mathbf{Q}_Z, \mathbf{K}_Z, \mathbf{V}_Z)$ ▷ Attend across variables
end function

F LIMITATIONS

Due to space constraints, we only present quantitative results; more qualitative visualizations can be provided in the future work. We adopt a shared set of hyperparameters across datasets for fair comparison; performance might be further improved with dataset-specific tuning.