ARISE: RETRIEVAL-AUGMENTED FORECASTING FOR IRREGULAR TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) supports time series forecasting by reusing historical windows and their subsequent observations. However, most existing methods are designed for regularly sampled data and face two challenges when extended to irregular multivariate time series (IMTS). Uneven observation density makes fixed windows prone to insufficient information or redundancy, while irregular sampling and asynchronous observations complicate the retrieval of relevant historical cases. To address these challenges, we propose ARISE, a retrieval-augmented forecasting framework for IMTS. First, ARISE introduces a dynamic window construction mechanism that learns a sample-specific left boundary for each historical window under multiple training objectives while keeping its right boundary fixed, yielding compact history–future pairs with predictive value. Second, a multi-evidence knowledge retrieval mechanism assesses historical relevance from three perspectives: observation patterns, temporal dynamics, and value information. By combining multi-view candidate retrieval with maximal marginal relevance (MMR) selection, ARISE balances relevance and diversity to retrieve complementary historical evidence. A retrieval-based memory fusion module then uses cross-attention to incorporate the retrieved histories and their observed continuations into the predictor's hidden representations. Finally, a forecasting decoder produces the final prediction. Experimental results on multiple real-world datasets show that ARISE outperforms existing state-of-the-art methods in forecasting accuracy.

1 INTRODUCTION

Forecasting irregular multivariate time series (IMTS) has become an important research topic in recent years. It has many applications in healthcare, biomechanics, climate science, and astronomy (Zhang et al., 2024; Yalavarthi et al., 2024). Unlike regularly sampled time series, IMTS record observations at unequal time intervals, as shown in Figure 1(a). Different variables are also often observed at different times (Kidger et al., 2020). Unequal intervals lead to uneven information density. Differences in observation times also make it hard to model relationships between variables. These features make accurate forecasting more difficult (Liu et al., 2026a;b; Qiu et al., 2026b).

Retrieval-augmented generation (RAG) has also been used for time series forecasting in recent years (Han et al., 2025; Ning et al., 2025). RAG combines a model with an external memory instead of relying only on model parameters to learn time series patterns (Lewis et al., 2020). This allows the model to directly use historical cases when making predictions. For regularly sampled time series, a knowledge base usually stores historical windows together with their later observations, as shown in Figure 1(b). Given the current input, the model retrieves related historical cases and uses their later observations to support forecasting (Han et al., 2025; Ning et al., 2025). This approach directly reuses useful historical experience. It provides extra evidence for forecasting by showing how the system developed after similar historical conditions.

However, most existing time series RAG methods are designed for regularly sampled data. They usually retrieve and use knowledge from historical windows with equal time intervals and aligned variables (Han et al., 2025; Ning et al., 2025). In IMTS, unequal sampling intervals and different observation times across variables change how historical cases are represented and matched. This makes existing RAG methods hard to apply directly (Liu et al., 2026b; Qiu et al., 2026b). We

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.

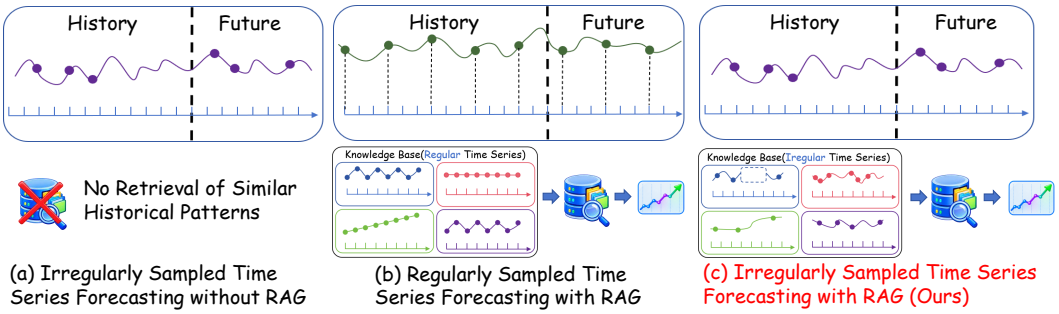

Figure 1: Extending retrieval-augmented forecasting from regular to irregular time series.

therefore propose a retrieval-augmented forecasting framework designed for IMTS, as shown in Figure 1(c). It builds a knowledge base from irregular historical observations and retrieves related historical evidence to support the current forecast. However, two key challenges must be addressed.

Challenge 1: How can we build a meaningful knowledge base for IMTS? RAG methods for regular time series usually build a knowledge base by dividing training data into fixed windows. With regular sampling, the same time span contains the same number of observations. Different windows therefore tend to contain similar amounts of information. This does not hold for IMTS. With irregular sampling, windows that cover the same time span may contain very different numbers of observations. In sparse regions, a fixed window may contain too little history to show how the system changed before the forecast. In dense regions, the same window may contain many observations that are close together in time. Some of these observations may repeat similar information and add unnecessary redundancy. Different observation times across variables further increase the differences in useful information between windows. Therefore, the primary challenge in extending RAG to IMTS is how to adaptively determine the window span of each historical case to retain sufficient prediction-relevant information while reducing redundancy.

Challenge 2: How can we retrieve relevant knowledge for IMTS? Existing time series RAG methods often retrieve historical candidate windows based on value correlation or similarity between learned sequence representations (Han et al., 2025; Ning et al., 2025). They usually assume that the time structure of different windows is similar enough for comparison. Similar values or representations can then reflect related patterns of change. This assumption does not hold for IMTS. Windows may differ greatly in observation times, sampling density, and missing data patterns. They may also differ in how much observation times vary across variables. Similar sequences of observed values do not always mean that the system changes in a similar way. For example, two windows may show similar value changes over different time spans. Their rates of change can then be very different. In contrast, windows with similar patterns of change may have very different observed values because of sparse sampling or different value scales (Yalavarthi et al., 2024). Directly reusing value or representation similarity from regular time series can therefore lead to misleading matches. IMTS need a retrieval strategy that accounts for their irregular observation patterns.

To address these two challenges, we propose ARISE, a retrieval-augmented forecasting framework for IMTS. First, we design a **dynamic window construction mechanism** to create knowledge entries that are useful for forecasting. It keeps the right boundary of each historical window fixed and learns the left boundary through several training objectives. The selected windows retain complete events, have limited lengths, and contain useful information for forecasting. They form reusable knowledge entries. Second, we propose a **multi-evidence knowledge retrieval mechanism** for irregular observations. It measures how the current input relates to historical cases through observation patterns, dynamics, and value features. We combine multi-view candidate retrieval with maximal marginal relevance (MMR) selection. This keeps the results relevant and diverse while reducing the effect of repeated knowledge. Meanwhile, we design a **retrieval-based memory fusion module**. It uses the current sequence representation as the query, representations of retrieved historical windows as keys, and representations of their corresponding ground-truth future observations as values. The module incorporates historical knowledge into the predictor's hidden representation. Finally, a forecasting decoder produces the final prediction.

Our contributions are summarized as follows:

- We propose ARISE, a retrieval-augmented forecasting framework for irregular multivariate time series. It extends retrieval-augmented forecasting to irregularly sampled data and enables the model to use relevant historical cases to improve future prediction.
- We design a dynamic window construction mechanism for knowledge base construction. Guided by multiple training objectives, it learns a sample-specific left boundary while keeping the right boundary fixed, yielding compact history–future pairs with predictive value.
- We introduce a multi-evidence knowledge retrieval mechanism based on observation patterns, temporal dynamics, and value information. By combining multi-view retrieval with MMR selection, ARISE balances relevance and diversity to retrieve complementary historical evidence under irregular sampling.
- Extensive experiments on 11 real-world irregular time series datasets with three forecasting horizons demonstrate that ARISE achieves leading overall forecasting performance compared with strong baselines. Code and data are available at `https://anonymous.4open.science/r/ARISE-E162/`.

2 Related Work

2.1 Irregular Multivariate Time Series Forecasting

Models for irregular series must handle unequal sampling intervals, asynchronous variables, and missing values. GRU-D uses masks and elapsed times to model informative missingness (Che et al., 2018). Neural ODEs evolve hidden states between observations (Chen et al., 2018; Rubanova et al., 2019); GRU-ODE-Bayes adds Bayesian observation updates (De Brouwer et al., 2019). Neural Flows instead parameterizes solution curves without numerical ODE solvers (Biloš et al., 2021). SeFT represents timestamped observations as sets (Horn et al., 2020), while mTAN uses time embeddings and attention to obtain fixed-length representations (Shukla & Marlin, 2021). Some of these methods were originally evaluated on classification or interpolation rather than forecasting.

Graph and patch-based forecasters model dependencies across observations and variables. GraFITi represents measurements as edges in a sparse bipartite graph (Yalavarthi et al., 2024), and ASTGI builds adaptive neighborhoods among spatio-temporal observation points (Liu et al., 2026a). tPatchGNN groups observations into time-aligned patches (Zhang et al., 2024); Hi-Patch uses hierarchical graphs for densely and sparsely sampled variables (Luo et al., 2025). APN learns patch boundaries and time-aware aggregation (Liu et al., 2026b), while TFMixer combines temporal patch mixing with a non-uniform Fourier branch (Qiu et al., 2026b). For regular forecasting, SEER enhances and replaces patch representations (Qiu et al., 2026a). ARISE constructs an external memory by pairing dynamically selected training histories with their observed futures, then retrieves relevant cases by comparing sampling patterns, temporal dynamics, and value features, providing complementary historical evidence beyond the current input to support irregular time series forecasting.

2.2 Retrieval-Augmented Time Series Forecasting

Retrieval-augmented forecasting reuses historical cases and their observed futures. RAFT retrieves windows by Pearson correlation and combines similarity-weighted futures with the query (Han et al., 2025). TS-RAG retrieves pretrained sequence embeddings and fuses them through an Adaptive Retrieval Mixer (Ning et al., 2025). Cross-RAG applies input-conditioned attention with historical keys and future values (Lee et al., 2026); RATL routes retrieved residuals of a frozen forecaster (He et al., 2026). SARAF adjusts retrieval diversity and aggregation to stationarity (Zhou et al., 2026a), while SERAF combines numerical matching with textual descriptions (Zhou et al., 2026b).

ARISE uses three views of one history–future memory and retains their separate matching scores. Maximal marginal relevance (Carbonell & Goldstein, 1998) reduces candidate redundancy. Before reading, each candidate future is aligned to the query's target times and value scale, with unsupported positions masked. The method combines prediction-oriented memory construction with sampling-aware retrieval and target-specific reading; it does not treat the history-key/future-value assignment alone as a new mechanism.

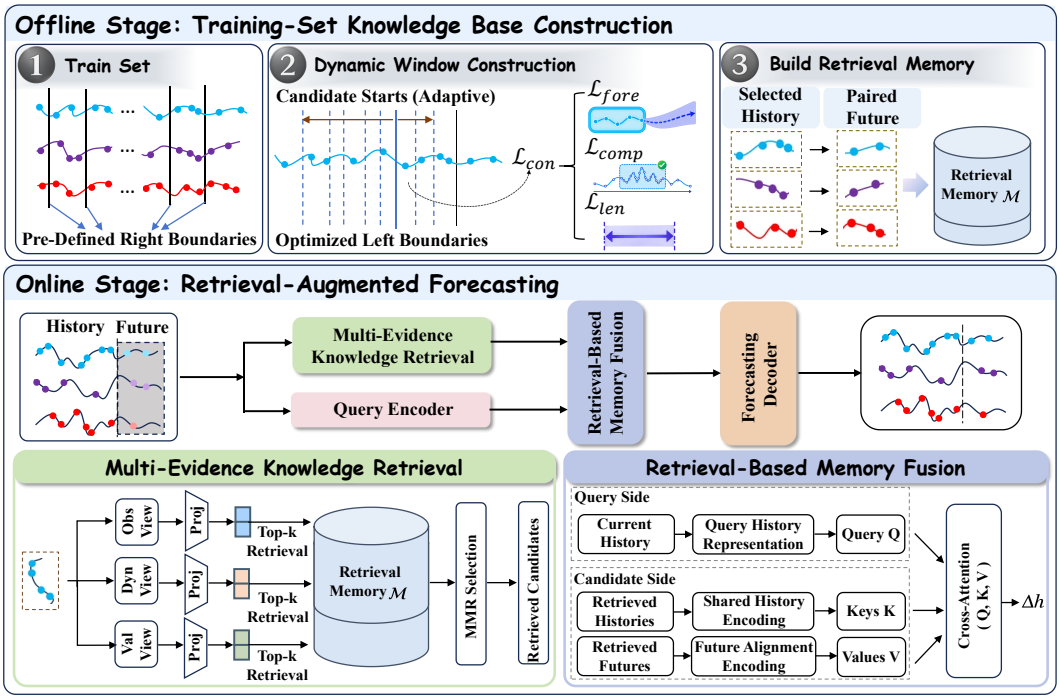

Figure 2: Overview of ARISE. Offline construction stores selected training histories, paired futures, and retrieval embeddings. Online three-view retrieval and MMR select candidates. The Query Side supplies queries; the Candidate Side encodes retrieved histories and aligned futures as keys and values. The correction Δh augments the base hidden state before decoding. The shaded query future contains unknown targets, not inputs.

3 Methodology

3.1 Problem Definition

An irregular multivariate time series (IMTS) consists of N observation sequences $\mathcal{O} = \{\mathcal{O}^v\}_{v=1}^{N}$, where $\mathcal{O}^v = \{(t_k^v, x_k^v)\}_{k=1}^{L_v}$. Variable v has L_v observations; t_k^v and x_k^v are the time and value of observation k. Sampling intervals may vary within and across variables. Given future queries $\mathcal{Q} = \{\{q_h^v\}_{h=1}^{H_v}\}_{v=1}^{N}$, with H_v target times for variable v, ARISE predicts values $\widehat{\mathcal{Y}} = \{\{\widehat{y}_h^v\}_{h=1}^{H_v}\}_{v=1}^{N}$ using a training-set memory $\mathcal{M}$:

$$\widehat{\mathcal{Y}} = f_{\text{ARISE}}(\mathcal{O}, \mathcal{Q}; \mathcal{M}). \tag{1}$$

3.2 Framework Overview

ARISE has two stages—see Figure 2. Offline, Dynamic Window Construction selects training histories and stores them with paired futures and retrieval embeddings. Online, Multi-Evidence Knowledge Retrieval matches observation, dynamics, and value patterns, then removes redundant candidates. In Retrieval-Based Memory Fusion, the Query Side uses the Query Encoder's history representation; the Candidate Side encodes retrieved histories and aligned futures as attention keys and values. The reader updates the backbone's hidden state before decoding. Dynamic selection changes the retrieval context, not the backbone's historical input.

3.3 Offline Stage: Training-Set Knowledge Base Construction

3.3.1 Train Set

For subsequent operations, represent each history by values $X \in \mathbb{R}^{L\times N}$, ordered timestamps $T \in \mathbb{R}^{L}$, and mask $M \in \{0,1\}^{L\times N}$; $M_{lv} = 1$ marks an observation at (l, v). Future values $Y \in \mathbb{R}^{H\times N}$,

times $T^* \in \mathbb{R}^H$, and mask $M^Y \in \{0,1\}^{H\times N}$ use H target positions. Both L and H count positions, not elapsed durations. Indices i, q denote samples, j memory entries, and l, h, u history, query-future, and candidate-future positions. Each training sample has a fixed cutoff and paired future. The cutoff is shared across variables, but different samples need not have equally spaced cutoffs. Its visible history is $\mathcal{H}_i = (X_i, T_i, M_i^{\text{base}})$ under the backbone's context policy.

3.3.2 Dynamic Window Construction

Equal observation counts can span different durations under irregular sampling. We therefore fix the right endpoint and adapt the historical start candidates to each sample's observed time span. Let a_i, b_i be the first and last observed times in $\mathcal{H}_i$ (zero if empty). The last observation b_i need not equal the forecasting cutoff. With $d_i = \max(b_i - a_i, 10^{-6})$, we define $C \geq 2$ sample-adaptive candidate starts using shared relative positions $\rho_m = (m-1)/(C-1)$, $m = 1, \ldots, C$:

$$\ell_{im} = a_i + \rho_m d_i, \qquad M_{ilv}^{(m)} = M_{ilv}^{\text{base}} \mathbf{1}\{\ell_{im} \leq T_{il} \leq b_i\}. \tag{2}$$

Here ℓ_{im} is candidate m's left boundary and $\mathbf{1}\{\cdot\}$ is the indicator. Variables share the boundary but retain their observation masks; no resampling is applied.

A shared scorer f_{cut} maps features ψ_{im}—three-view descriptors, boundary differences, and coverage statistics—to a scalar. The binary eligibility mask $\chi_{im}^{\text{sel}} \in \{0,1\}$ checks observation counts, coverage, duration, and full-history safeguards. With temperature $\tau_{\text{cut}} > 0$, candidate probabilities and the selected mask are

$$\begin{aligned} p_{im} &= \text{MaskedSoftmax}_m\big(f_{\text{cut}}(\psi_{im})/\tau_{\text{cut}}; \chi_{im}^{\text{sel}}\big), \\ m_i^* &= \arg\max_m p_{im}, \qquad M_i^{\text{sel}} = M_i^{(m_i^*)}. \end{aligned} \tag{3}$$

p_{im} is normalized over eligible candidates, and m_i^* is the highest-scoring index. The relative fractions ρ_m are shared across samples, while the resulting candidate starts adapt to each sample's observed time span; the scorer then learns which candidate becomes the left boundary. If no candidate passes the checks, the full-history mask provides a fallback; an empty history remains empty.

For view $r \in \mathcal{V} = \{\text{obs}, \text{dyn}, \text{val}\}$, candidate embedding $e_{im}^r \in \mathbb{R}^{d_e}$ is computed as in Section 3.4.1. The scorer and view encoders are trained jointly with a projection $P_{\text{cut}} : \mathbb{R}^{3d_e} \to \mathbb{R}^{d_{\text{cut}}}$ using the soft history representation

$$z_i = \text{Normalize}\left(\sum_{m=1}^{C} p_{im} P_{\text{cut}}[e_{im}^{\text{obs}}; e_{im}^{\text{dyn}}; e_{im}^{\text{val}}]\right). \tag{4}$$

Semicolons denote concatenation, d_e, d_{cut} are embedding widths, and $\text{Normalize}(z) = z/\max(\|z\|_2, 10^{-8})$. The construction loss is

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{forecast}} + \lambda_b \mathcal{L}_{\text{completeness}} + \lambda_l \mathcal{L}_{\text{length}}, \tag{5}$$

where $\lambda_b, \lambda_l \geq 0$ weight the two regularizers. $\mathcal{L}_{\text{forecast}}$ makes window selection prediction-oriented by using time-aware future similarity to form mini-batch triplets and supervise the historical representations. $\mathcal{L}_{\text{completeness}}$ encourages the selected boundary to preserve a complete local event by favoring a larger representation difference across the boundary than between adjacent regions within the retained window. $\mathcal{L}_{\text{length}}$ controls the temporal extent of the selected history according to its effective observational coverage, preventing unnecessarily broad or poorly supported windows while retaining sufficient historical evidence. Together, the three objectives encourage the selector to construct windows that preserve complete prediction-relevant events, maintain appropriate lengths, and provide useful evidence for forecasting. Gradients are propagated through the soft representation and probability-weighted objectives rather than the hard $\arg\max$, while future observations are used only as training supervision.

3.3.3 Build Retrieval Memory

We restore the validation-selected construction checkpoint and freeze the selector, view encoders, and fitted statistics. Admission and deduplication leave J cases, at most one per training sample:

$$\mathcal{M} = \big\{\big(\mathcal{H}_j^{\text{sel}}, \mathcal{F}_j, \{e_j^r\}_{r\in\mathcal{V}}, \mu_j^{\text{sel}}, \sigma_j^{\text{sel}}, \pi_j\big)\big\}_{j=1}^{J}. \tag{6}$$

Each entry pairs $\mathcal{H}_j^{\mathrm{sel}} = (X_j, T_j, M_j^{\mathrm{sel}})$ with $\mathcal{F}_j = (Y_j, \tau_j, M_j^Y)$, where $\tau_{ju} = T_{ju}^* - b_j$. Selected-history means and stabilized standard deviations $\mu_j^{\mathrm{sel}}, \sigma_j^{\mathrm{sel}} \in \mathbb{R}^N$ support value alignment; π_j records entry, sample, and source identities. The three embeddings index the same case, not separate memories. Empty histories and exact-identity duplicates are excluded. Missing futures may be stored but provide no supported evidence. The stored futures retain their observation masks and time axes. The selected-history scales are used for future transfer, not as replacements for the backbone normalization statistics.

3.4 Online Stage: Retrieval-Augmented Forecasting

3.4.1 Multi-Evidence Knowledge Retrieval

The *Observation* view describes sampling frequency through per-variable gap statistics, a relative-time observation histogram, coverage, and availability. It distinguishes dense and sparse sampling, not signal oscillation frequency. *Dynamics* describes normalized values and changes per unit time between consecutive observations; *Value* retains levels, scales, and ranges. For selected query history $\mathcal{H}_q^{\mathrm{sel}}$, descriptor map ϕ^r and frozen view encoder E_r give

$$e_q^r = \mathrm{Normalize}\big(E_r(\phi^r(\mathcal{H}_q^{\mathrm{sel}}))\big), \qquad s_{qj}^r = \cos(e_q^r, e_j^r). \tag{7}$$

The cosine score s_{qj}^r compares embeddings rather than raw sequences. Memory and query histories use the same frozen view encoders. Rate features connect consecutive observed values, not padded neighbors. After excluding self/same-source cases, each view recalls up to K entries. We deduplicate their union by entry ID and retain all three scores $\mathbf{s}_{qj} = [s_{qj}^{\mathrm{obs}}, s_{qj}^{\mathrm{dyn}}, s_{qj}^{\mathrm{val}}]$, with mean $\bar{s}_{qj} = \frac{1}{3}\sum_r s_{qj}^r$. Removing cases without support at any target position gives $\mathcal{U}_q^+$; see Section 3.4.3. Starting with $\mathcal{R} = \emptyset$, maximal marginal relevance (MMR) (Carbonell & Goldstein, 1998) selects

$$j^* = \arg\max_{j \in \mathcal{U}_q^+ \setminus \mathcal{R}} \left[\bar{s}_{qj} - \lambda_{\mathrm{MMR}} \max_{j' \in \mathcal{R}} \left(\tfrac{1}{3} \sum_{r \in \mathcal{V}} \cos(e_j^r, e_{j'}^r) \right) \right]. \tag{8}$$

Here $\lambda_{\mathrm{MMR}} \geq 0$ penalizes similarity to selected cases j', with zero penalty for an empty set. Selection stops at budget R or exhaustion, yielding $\mathcal{R}_q$. Each selected candidate retains its own future and all three view scores.

3.4.2 Query Encoder

The backbone's frozen history encoder E_B, including its configured normalization, produces

$$\mathbf{h}_{\mathrm{base}} = E_B(X_q, T_q, M_q^{\mathrm{base}}), \qquad \mathbf{h}_{\mathrm{sel}} = E_B(X_q, T_q, M_q^{\mathrm{sel}}). \tag{9}$$

Both states lie in $\mathbb{R}^{N \times D}$, where D is the hidden width. The Query Side uses $\mathbf{h}_{\mathrm{sel}}$ for reading, while $\mathbf{h}_{\mathrm{base}}$ retains the full available context. Shared History Encoding on the Candidate Side computes $\mathbf{h}_j = E_B(X_j, T_j, M_j^{\mathrm{sel}})$. These states are distinct from the retrieval embeddings used for case matching.

3.4.3 Retrieval-Based Memory Fusion

Future Alignment Encoding on the Candidate Side uses query offsets $\tau_{qh} = T_{qh}^* - b_q$ and stored candidate offsets τ_{ju}. Gaussian weights $\kappa_{qjhu}^t = \exp[-(\tau_{ju} - \tau_{qh})^2/(2\sigma_t^2)]$, with bandwidth $\sigma_t > 0$, give kernel mass Z, support mask S, and aligned values $\widetilde{Y}$:

$$\begin{aligned} Z_{qjhv} &= \sum_u \kappa_{qjhu}^t M_{juv}^Y, \qquad S_{qjhv} = \mathbf{1}\{Z_{qjhv} > \epsilon_t\}, \\ \widetilde{Y}_{qjhv} &= \frac{\sum_u \kappa_{qjhu}^t M_{juv}^Y Y_{juv}}{\max(Z_{qjhv}, \epsilon_t)}, \qquad \epsilon_t = 10^{-6}. \end{aligned} \tag{10}$$

Selected-history statistics transfer $\widetilde{Y}$ to the query's value scale and then its backbone normalization domain. No query future values or masks are used. A content encoder maps the transferred value,

support, and scaled target offset to $\mathbf{u}_{qjhv}$; score encoder E_s gives $e^s_{qj} = E_s(\mathbf{s}_{qj})$. The reader forms

$$\begin{aligned} \mathbf{Q}_q(h,v) &= W_Q[\mathbf{h}_{\text{sel}}(v); \text{TE}(T^*_{qh})], \\ \mathbf{K}_{qj}(v) = W_K\mathbf{h}_j(v) + U_K e^s_{qj}, &\qquad \mathbf{V}_{qj}(h,v) = W_V\mathbf{u}_{qjhv}. \end{aligned} \tag{11}$$

Here TE encodes target times, while $\mathbf{h}(v)$ denotes the hidden vector of variable v; W_Q, W_K, U_K, W_V are learned projections to width D_A (biases implicit). Each of n_{head} heads has width $d_{\text{head}} = D_A/n_{\text{head}}$; superscript a denotes a head slice. At target (h,v), supported cases $\mathcal{J}_{qhv} = \{j \in \mathcal{R}_q : S_{qjhv} = 1\}$ compete with a learned null key $\mathbf{K}^a_\emptyset$ and zero value. Attention (Vaswani et al., 2017) assigns

$$\alpha^a_{qjhv} = \text{Softmax}_{j\in\mathcal{J}_{qhv}\cup\{\emptyset\}}\left(\frac{(\mathbf{Q}^a_q(h,v))^\top \mathbf{K}^a_{qj}(v)}{\sqrt{d_{\text{head}}}}\right), \tag{12}$$

The weights α^a_{qjhv} include null ($j = \emptyset$). Each head forms a context and contributes to the correction:

$$\begin{aligned} c^a_q(h,v) &= \sum_{j\in\mathcal{J}_{qhv}} \alpha^a_{qjhv}\mathbf{V}^a_{qj}(h,v), \\ \Delta h(h,v) &= g_{\text{ret}} W_O[c^1_q(h,v); \ldots; c^{n_{\text{head}}}_q(h,v)]. \end{aligned} \tag{13}$$

The bias-free W_O projects back to width D; g_{ret} is fixed to one in the standard configuration. Null allows reduced reliance on retrieval; if no case is supported, $\Delta h = 0$.

3.4.4 Forecasting Decoder

The backbone decoder maps the augmented representation to predictions:

$$\mathbf{h}_{\text{fused}}(h,v) = \mathbf{h}_{\text{base}}(v) + \Delta h(h,v), \qquad \widehat{Y} = \mathcal{D}_B(\mathbf{h}_{\text{fused}}, T^*; \mathcal{C}_q). \tag{14}$$

The decoder $\mathcal{D}_B$ includes output postprocessing; $\mathcal{C}_q$ holds query normalization statistics. The pretrained backbone remains fixed. Zero-initialized W_O starts from its forecast, while content, score, and attention parameters learn from observed-target MAE $\mathcal{L}_{\text{pred}}$ through the frozen decoder.

4 Experiments

4.1 Experimental Settings

Datasets: To evaluate ARISE across different irregular sampling regimes, we conduct experiments on eleven real-world IMTS benchmarks: PhysioNet (Silva et al., 2012), HumanActivity (Vidulin et al., 2010), USHCN (Easterling, 2002), CESNET (Koumar et al., 2025), Pamap2 (Reiss, 2012), EPA-Air (U.S. Environmental Protection Agency, 2026), ClusterTrace (Weng et al., 2022), APTC (Souza, 2018), FNSPID (Dong et al., 2024), GDELT (Leetaru & Schrodt, 2013), and Seabirds (Browning et al., 2018). These datasets cover clinical monitoring, human activity, environmental sensing, network and system traces, financial/event streams, and ecological observations, and exhibit different levels of sparsity, sampling-frequency variation, and cross-variable asynchrony. Their observation densities and sampling-gap distributions differ substantially, providing a broad test of whether retrieval remains reliable across heterogeneous irregular patterns. We keep the benchmark-provided timestamps and observation masks rather than interpolating them onto a regular grid. The original 60%/20%/20% training/validation/test partition is retained before window construction, using record-wise splits for PhysioNet, USHCN, and APTC and chronological splits for the remaining datasets.

Baselines: We comprehensively evaluate ARISE against nine IMTS forecasting baselines: TFMixer (Qiu et al., 2026b), ASTGI (Liu et al., 2026a), APN (Liu et al., 2026b), KAFNet (Zhou et al., 2026c), Hi-Patch (Luo et al., 2025), HyperIMTS (Li et al., 2025), GNeuralFlow (Mercatali et al., 2024), GraFITi (Yalavarthi et al., 2024), and Warpformer (Zhang et al., 2023).

Implementation Details: ARISE is implemented in PyTorch 2.6.0 (Paszke et al., 2019) and trained on NVIDIA GeForce RTX 3090 GPUs with 24 GB memory. Within ARISE, construction checkpoints are selected by the validation construction objective and fusion checkpoints by validation per-variable MAE. Retrieval statistics and the knowledge memory are fitted from the training split only, and the selector and view encoders are frozen before fusion training. "Drop Last" is disabled at test time for all three settings to retain the final incomplete batch (Qiu et al., 2024).

Table 1: Forecasting performance on eleven IMTS datasets. Overall performance is evaluated by mean MSE and MAE over three forecasting horizons ($\times 10^{-2}$). The best and second-best results are highlighted in red bold and blue with an underline, respectively. Full per-horizon results are reported in Table 3.

Models	ARISE (ours)		TFMixer (2026)		ASTGI (2026)		APN (2026)		KAFNet (2026)		Hi-Patch (2025)		HyperIMTS (2025)		GNeuralFlow (2024)		GraFITi (2024)		Warpformer (2023)	
Metrics	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
PhysioNet	**0.449**	**3.070**	0.558	3.568	0.518	3.793	0.471	3.655	0.567	4.284	0.551	3.942	0.588	4.273	1.040	5.826	0.567	4.038	1.125	5.783
HumanActivity	**0.362**	**3.586**	0.363	3.604	0.390	4.004	0.366	3.732	0.383	3.917	0.395	4.127	0.391	4.004	1.841	10.504	0.378	3.902	1.393	8.703
USHCN	52.00	**25.10**	**50.48**	25.78	52.36	34.46	54.86	35.90	64.47	42.13	58.32	40.36	68.96	47.08	59.13	38.76	59.58	39.63	53.19	37.95
CESNET	2.169	**8.953**	2.327	10.006	**2.160**	9.340	2.319	10.159	2.351	10.382	2.435	10.325	2.657	11.890	2.339	10.337	2.281	10.265	3.049	11.769
Pamap2	**0.476**	**3.708**	0.599	5.049	0.646	5.385	0.964	6.365	1.005	6.714	0.861	6.500	0.621	5.093	1.804	9.607	0.688	5.391	–	–
EPA-Air	**0.648**	**5.724**	0.791	6.392	0.778	6.455	1.731	9.692	1.539	8.622	1.059	7.364	1.009	7.602	2.236	10.982	0.838	7.045	2.268	10.832
ClusterTrace	**3.710**	**14.452**	4.199	15.057	4.177	15.677	4.317	15.585	4.492	15.468	4.301	15.665	4.406	15.744	4.576	15.944	4.498	15.585	4.627	16.657
APTC	**0.379**	**4.585**	0.381	4.604	0.383	4.622	0.382	4.616	0.381	4.603	0.382	4.605	0.381	4.604	0.383	4.615	0.381	4.603	0.535	5.561
FNSPID	**0.205**	**2.490**	0.332	3.817	1.069	7.827	0.785	6.496	0.373	4.652	0.662	5.834	0.641	5.959	0.938	6.958	0.885	6.797	1.801	8.453
GDELT	**1.819**	**8.269**	1.859	8.432	1.826	8.376	1.841	8.512	1.837	8.596	1.955	9.264	1.849	8.470	1.865	8.649	1.953	9.233	2.307	9.447
Seabirds	**0.012**	**0.589**	0.032	1.306	0.022	1.041	0.081	2.227	0.097	2.742	0.098	2.284	0.049	1.639	0.134	2.388	0.075	2.285	0.117	2.471
1st Count	**9**	**11**	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2: Ablation studies for ARISE ($\times 10^{-2}$).

Dataset	EPA-Air		CESNET		GDELT		ClusterTrace		Average	
Metrics	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
w/o Dynamic Window	0.94	7.08	2.71	10.98	2.22	9.67	4.86	16.58	2.68	11.07
w/o Three-View Retrieval	1.06	7.41	2.88	11.29	2.39	10.05	5.11	17.18	2.86	11.48
w/o Retrieved Future Values	1.28	7.95	3.08	11.74	2.61	10.76	5.49	17.63	3.11	12.02
w/o Attention Reader	1.57	8.74	3.22	11.82	2.97	11.92	6.34	19.18	3.52	12.91
ARISE (ours)	**0.64**	**5.72**	**2.16**	**8.95**	**1.81**	**8.26**	**3.70**	**14.44**	**2.07**	**9.34**

4.2 Main Results

Comprehensive results are presented in Table 1, with detailed per-horizon results reported in Appendix Table 3. We make the following observations: 1) Compared with forecasters of different architectures, ARISE achieves the strongest overall forecasting performance. It ranks first on 9 of the 11 dataset-average MSEs and all 11 dataset-average MAEs, while remaining competitive on the two remaining MSE cases. The improvements are especially clear on datasets such as FNSPID, Pamap2, EPA-Air, and Seabirds, demonstrating the benefit of retrieval-augmented forecasting across diverse irregular settings. 2) ARISE maintains strong performance across IMTS datasets with markedly different sampling patterns, observation densities, and application domains. This consistency is closely related to its retrieval design: Dynamic Window Construction adapts the historical context to each irregular sample, Multi-Evidence Knowledge Retrieval compares observation patterns, temporal dynamics, and value states from complementary views, and Retrieval-Based Memory Fusion aligns and integrates the paired futures of retrieved cases. Together, these modules allow ARISE to reuse relevant historical evidence while accounting for irregular sampling and incomplete observations.

4.3 Ablation Study

We compare four ablation variants on EPA-Air, CESNET, GDELT, and ClusterTrace—see Table 2. All variants use the same backbone, data splits, training protocol, and checkpoint-selection strategy as the full model, differing only in the component being ablated. We make the following observations: 1) Removing the attention reader causes the largest performance degradation on all four datasets, increasing the average MSE/MAE from $2.07/9.34$ to $3.52/12.91$. Without the reader, the retrieval correction is removed by setting $\Delta h = 0$; see Eq. 14. This result shows that effectively integrating retrieved evidence is important for converting relevant historical cases into forecasting gains. 2) Removing retrieved future values yields the second-largest degradation, with average MSE/MAE increasing to $3.11/12.02$. This confirms that the continuations paired with retrieved histories provide useful predictive information beyond historical similarity alone. 3) Replacing

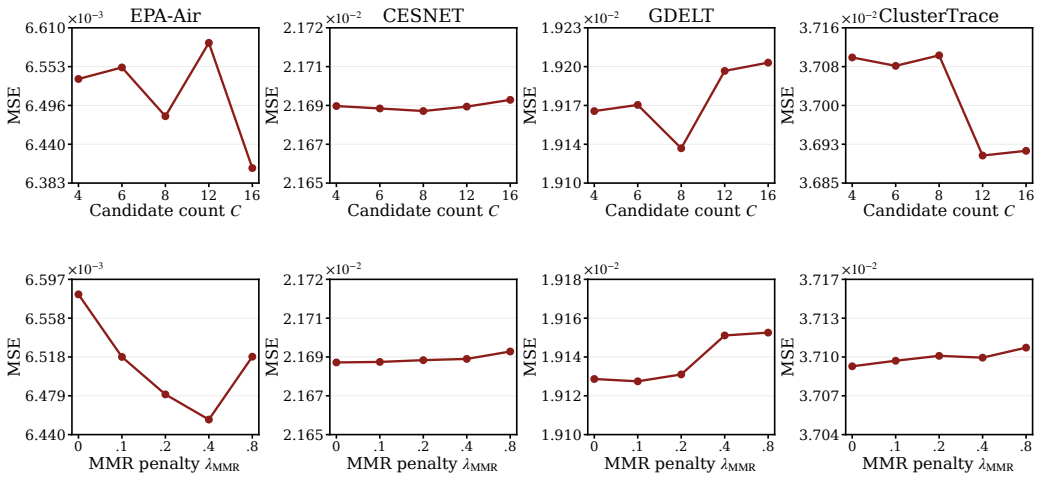

Figure 3: Sensitivity on EPA-Air, CESNET, GDELT, and ClusterTrace (left to right). Rows vary C and λ_{MMR}. Values are unscaled three-horizon mean test MSE (lower is better); tested settings are equally spaced and vertical ranges are independent.

three-view matching with random retrieval consistently degrades both metrics, demonstrating the importance of jointly considering observation patterns, temporal dynamics, and value information when identifying relevant cases under irregular sampling. 4) Replacing Dynamic Window Construction with fixed retrieval windows also reduces forecasting accuracy, indicating that adapting the historical range to each sample helps preserve prediction-relevant evidence while reducing interference from less relevant observations. Overall, the consistent degradation across all four variants confirms the contribution of dynamic context construction, multi-evidence retrieval, paired-future knowledge, and retrieval-based memory fusion to ARISE.

4.4 Parameter Sensitivity

Figure 3 varies C and λ_{MMR} on the same four datasets using unscaled three-horizon mean MSE. Larger C adds candidate left boundaries, not memory entries per sample. EPA-Air and ClusterTrace reach their lowest observed MSE at $C = 16$ (6.405×10^{-3}) and $C = 12$ (3.6903×10^{-2}), whereas CESNET and GDELT favor $C = 8$. EPA-Air is non-monotonic and CESNET changes little, so finer boundary resolution is not uniformly beneficial. For MMR, EPA-Air improves from 6.582×10^{-3} at zero penalty to 6.455×10^{-3} at 0.4, then rises to 6.519×10^{-3} at 0.8; GDELT favors 0.1, while CESNET and ClusterTrace favor zero. These curves indicate a dataset-dependent balance between relevance and diversity. Appendix B extends the analysis to recall budget K and alignment bandwidth σ_t and provides the complete four-parameter plots.

5 Conclusion

In this work, we study retrieval-augmented forecasting for irregular multivariate time series (IMTS), where irregular sampling, asynchronous observations, and uneven observation density make conventional retrieval strategies designed for regularly sampled data less effective. To address these challenges, we propose ARISE, a retrieval-augmented forecasting framework tailored to IMTS. ARISE dynamically constructs informative historical windows by learning sample-specific boundaries, retrieves relevant historical cases from multiple complementary perspectives, and further balances retrieval relevance and diversity through MMR selection. The retrieved histories and their observed continuations are then integrated into the forecasting model through a memory fusion module to provide complementary predictive evidence. Extensive experiments on multiple real-world datasets demonstrate that ARISE consistently improves forecasting accuracy over existing SOTA methods. Code and data are available at `https://anonymous.4open.science/r/ARISE-E162/`.

AI USE STATEMENT

We used ChatGPT to help polish the manuscript and improve the clarity of the writing. All AI-assisted content was carefully reviewed by the authors. All methodological design and experimental analyses were conducted independently by the authors, and the use of AI was limited to auxiliary functions. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on methodological research on retrieval-augmented forecasting for irregular multivariate time series. All experiments use existing benchmark datasets. The retrieval memory stores training histories and their paired future observations; restricting it to the training split does not itself guarantee privacy. The use and sharing of source data, retrieval memories, and related artifacts must respect the original dataset licenses and access restrictions, particularly for human-related data. Retrieved cases may reflect biases in sampling frequency and missingness, and forecasting errors may have harmful consequences in sensitive applications. ARISE is intended for forecasting research rather than autonomous high-stakes decision-making. Deployment in healthcare or other sensitive settings requires further validation, appropriate privacy safeguards, assessment of performance across relevant populations and observation patterns, and human oversight.

REPRODUCIBILITY STATEMENT

The ARISE architecture and training procedure are described in Section 3, and the experimental settings are provided in Section 4.1. Detailed experimental results and additional analyses are provided in the Appendix. Code for the main experiments is available at `https://anonymous.4open.science/r/ARISE-E162/`.

REFERENCES

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. In *NeurIPS*, volume 34, pp. 21325–21337, 2021.

Ella Browning, Mark Bolton, Ellie Owen, Akiko Shoji, Tim Guilford, and Robin Freeman. Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution*, 9(3):681–692, 2018.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pp. 335–336, 1998.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, volume 31, 2018.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *NeurIPS*, volume 32, 2019.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. *arXiv preprint arXiv:2402.06698*, 2024.

DR Easterling. United states historical climatology network daily temperature and precipitation data (1871-1997). Technical report, ORNL Oak Ridge National Laboratory (US), 2002.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *ICML*, volume 267, pp. 21774–21797, 2025.

Yuchen He, Yueyang Cang, Zhiyuan Ning, Ningyu Wang, and Li Shi. RATL: Learning from retrieved residuals for robust multivariate time-series forecasting. *arXiv preprint arXiv:2609.03937*, 2026.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *ICML*, volume 119, pp. 4353–4363, 2020.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *NeurIPS*, volume 33, pp. 6696–6707. Curran Associates, Inc., 2020.

Josef Koumar, Karel Hynek, Tomáš Čejka, and Pavel Šiška. CESNET-TimeSeries24: Time series dataset for network traffic anomaly detection and forecasting. *Scientific Data*, 12(1):338, 2025.

Seunghan Lee, Jaehoon Lee, Jun Seo, Sungdong Yoo, Minjae Kim, Tae Yoon Lim, Dongwan Kang, Hwanil Choi, SoonYoung Lee, and Wonbin Ahn. Not all retrievals are useful: Cross-attention for input-aware RAG in time series forecasting. In *KDD Workshop*, 2026.

Kalev H. Leetaru and Philip A. Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, San Francisco, CA, 2013.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, volume 33, pp. 9459–9474, 2020.

Boyuan Li, Yicheng Luo, Zhen Liu, Junhao Zheng, Jianming Lv, and Qianli Ma. Hyperimts: Hypergraph neural network for irregular multivariate time series forecasting. *arXiv preprint arXiv:2505.17431*, 2025.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. Astgi: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *ICLR*, 2026a.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *AAAI*, volume 40, pp. 23873–23881, 2026b.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-patch: Hierarchical patch gnn for irregular multivariate time series. In *ICML*, volume 267, pp. 41494–41519, 2025.

Giangiacomo Mercatali, Andre Freitas, and Jie Chen. Graph neural flows for unveiling systemic interactions among irregularly sampled time series. In *NeurIPS*, volume 37, pp. 57183–57206, 2024.

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. In *NeurIPS*, volume 38, pp. 163170–163199, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: towards comprehensive and fair benchmarking of time series forecasting methods. In *Proc. VLDB Endow.*, pp. 2363–2377, 2024.

Xiangfei Qiu, Xvyuan Liu, Tianen Shen, Xingjian Wu, Hanyin Cheng, Bin Yang, and Jilin Hu. SEER: Transformer-based robust time series forecasting via automated patch enhancement and replacement. In *ICML*, 2026a.

Xiangfei Qiu, Kangjia Yan, Xvyuan Liu, Xingjian Wu, and Jilin Hu. Bridging time and frequency: A joint modeling framework for irregular multivariate time series forecasting. In *ICML*, 2026b.

Attila Reiss. PAMAP2 physical activity monitoring. UCI Machine Learning Repository, 2012.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *NeurIPS*, volume 32, 2019.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *CinC*, pp. 245–248, 2012.

Vinicius M. A. Souza. Asphalt pavement classification using smartphone accelerometer and complexity invariant distance. *Engineering Applications of Artificial Intelligence*, 74:198–211, 2018.

U.S. Environmental Protection Agency. AirData: Air quality data collected at outdoor monitors across the US. Air Quality System Data Mart, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

Vedrana Vidulin, Mitja Luštrek, Boštjan Kaluža, Rok Piltaver, and Jana Krivec. Localization data for person activity. UCI Machine Learning Repository, 2010.

Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding. MLaaS in the wild: Workload analysis and scheduling in large-scale heterogeneous GPU clusters. In *NSDI*, pp. 945–960, 2022.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. GraFITi: Graphs for forecasting irregularly sampled time series. *AAAI*, 38(15):16255–16263, 2024.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *KDD*, pp. 3273–3285, 2023.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *ICML*, volume 235, pp. 60179–60196, 2024.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting. In *KDD*, pp. 6875–6886, 2026a.

Shiqiao Zhou, Zipeng Wu, Holger Schöner, Edouard Fouché, IAG Wilson, and Shuo Wang. Semantics-enhanced retrieval-augmented time series forecasting. In *ICML Workshop*, 2026b.

Ziyu Zhou, Yiming Huang, Yanyun Wang, Yuankai Wu, James Kwok, and Yuxuan Liang. Revitalizing canonical pre-alignment for irregular multivariate time series forecasting. In *AAAI*, volume 40, pp. 29115–29123, 2026c.

A Additional Full Results

Table 3 reports the complete three-horizon results and the corresponding averages for all datasets. Ranking follows the main table: red marks the best entries and blue with underlining the second best, including ties at the displayed precision.

B Additional Parameter Sensitivity Results

We study four parameters governing historical-context selection, candidate recall, retrieval diversity, and temporal alignment: the candidate window count C, per-view retrieval budget K, MMR penalty λ_{MMR}, and alignment bandwidth σ_t. Figure 4 presents their three-horizon mean test MSE on EPA-Air, CESNET, GDELT, and ClusterTrace, with one parameter per row and one dataset per column. The first and third rows also appear in Figure 3. The tested ranges cover the default setting and several smaller and larger alternatives for each parameter. We discuss each parameter's role and observed response before detailing the evaluation protocol.

Table 3: Detailed forecasting performance on eleven IMTS datasets across three forecasting horizons. Overall performance is evaluated by MSE and MAE ($\times 10^{-2}$). The best and second-best results are highlighted in red bold and blue with an underline, respectively. The first-place count is computed over the 33 forecasting-horizon settings, excluding the Avg rows.

Dataset	Pred	ARISE (ours)		TFMixer (2026)		ASTGI (2026)		APN (2026)		KAFNet (2026)		Hi-Patch (2025)		HyperIMTS (2025)		GNeuralFlow (2024)		GraFITi (2024)		Warpformer (2023)	
		mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
PhysioNet	1	**0.404**	**2.906**	0.619	3.765	0.553	3.882	0.439	3.554	0.654	4.980	0.559	4.004	0.493	3.454	1.220	6.061	0.594	3.972	1.290	6.646
	6	**0.420**	**3.038**	0.486	3.415	0.470	3.692	0.443	3.659	0.497	3.763	0.528	3.963	0.481	3.914	0.929	5.548	0.444	3.604	1.018	5.250
	12	**0.524**	**3.265**	0.569	3.525	0.530	3.805	0.530	3.751	0.550	4.110	0.566	3.859	0.791	5.451	0.972	5.869	0.663	4.538	1.066	5.453
	Avg	**0.449**	**3.070**	0.558	3.568	0.518	3.793	0.471	3.655	0.567	4.284	0.551	3.942	0.588	4.273	1.040	5.826	0.567	4.038	1.125	5.783
HumanActivity	200	0.255	**2.807**	**0.254**	2.814	0.277	3.167	0.259	2.974	0.265	3.054	0.292	3.373	0.261	3.028	1.852	10.641	0.259	2.963	0.396	4.365
	1200	0.358	3.570	**0.354**	**3.568**	0.385	4.001	0.358	3.688	0.373	3.871	0.401	4.149	0.369	3.831	1.877	10.459	0.373	3.949	1.704	10.427
	3000	**0.472**	**4.381**	0.481	4.431	0.509	4.843	0.481	4.534	0.511	4.826	0.492	4.859	0.542	5.153	1.794	10.411	0.502	4.794	2.079	11.317
	Avg	**0.362**	**3.586**	0.363	3.604	0.390	4.004	0.366	3.732	0.383	3.917	0.395	4.127	0.391	4.004	1.841	10.504	0.378	3.902	1.393	8.703
USHCN	1	55.52	25.66	53.76	**25.15**	58.28	37.09	**52.30**	30.97	58.38	35.16	57.19	36.97	66.24	50.59	54.73	35.49	55.93	33.34	55.25	40.90
	6	49.94	24.63	**48.50**	**24.40**	49.03	32.70	60.86	41.02	65.30	43.41	63.62	43.73	68.48	43.97	51.62	34.17	50.98	33.25	50.58	35.67
	12	50.55	**25.00**	**49.18**	27.79	49.78	33.58	51.42	35.70	69.73	47.81	54.16	40.37	72.17	46.68	71.03	46.62	71.83	52.29	53.74	37.28
	Avg	52.00	**25.10**	**50.48**	25.78	52.36	34.46	54.86	35.90	64.47	42.13	58.32	40.36	68.96	47.08	59.13	38.76	59.58	39.63	53.19	37.95
CESNET	24	**2.159**	**8.920**	2.293	9.723	2.164	9.384	2.298	10.211	2.336	10.363	2.340	10.174	2.297	10.050	2.323	10.262	2.234	9.981	2.711	11.545
	168	2.172	**8.963**	2.332	10.035	**2.160**	9.209	2.307	10.166	2.362	10.369	2.443	10.365	2.696	12.280	2.343	10.392	2.307	10.424	2.339	10.162
	336	2.175	**8.976**	2.357	10.259	**2.156**	9.428	2.353	10.101	2.355	10.414	2.523	10.436	2.978	13.340	2.351	10.358	2.303	10.389	4.098	13.600
	Avg	2.169	**8.953**	2.327	10.006	**2.160**	9.340	2.319	10.159	2.351	10.382	2.435	10.325	2.657	11.890	2.339	10.337	2.281	10.265	3.049	11.769
Pamap2	200	**0.211**	**2.394**	0.294	3.279	0.246	2.898	0.225	2.762	0.305	3.201	0.365	4.386	0.308	3.448	1.619	9.100	0.216	2.649	–	–
	1200	**0.484**	**3.895**	0.635	5.481	0.651	5.702	0.754	6.056	0.830	6.138	0.722	6.255	0.609	5.253	1.796	9.644	0.697	5.954	–	–
	3000	**0.733**	**4.833**	0.869	6.387	1.042	7.556	1.914	10.277	1.879	10.803	1.495	8.859	0.946	6.578	1.996	10.076	1.150	7.570	–	–
	Avg	**0.476**	**3.708**	0.599	5.049	0.646	5.385	0.964	6.365	1.005	6.714	0.861	6.500	0.621	5.093	1.804	9.607	0.688	5.391	–	–
EPA-Air	24	**0.606**	**5.485**	0.742	6.125	0.790	6.566	2.093	11.071	1.613	8.813	1.104	7.674	0.851	6.709	2.530	11.791	0.791	6.862	2.001	9.968
	72	**0.639**	**5.724**	0.881	6.632	0.766	6.435	1.454	8.555	1.567	8.668	1.146	7.496	0.841	6.663	2.241	10.811	0.946	7.645	2.461	11.736
	168	**0.699**	**5.965**	0.750	6.419	0.778	6.365	1.647	9.449	1.438	8.385	0.926	6.923	1.336	9.433	1.937	10.343	0.778	6.627	2.341	10.793
	Avg	**0.648**	**5.724**	0.791	6.392	0.778	6.455	1.731	9.692	1.539	8.622	1.059	7.364	1.009	7.602	2.236	10.982	0.838	7.045	2.268	10.832
ClusterTrace	24	**3.404**	**13.655**	3.919	14.730	4.060	15.754	4.250	15.305	4.497	15.502	4.372	15.620	4.470	15.749	4.721	16.318	4.478	15.585	4.600	16.000
	48	**3.882**	**14.759**	4.356	15.305	4.249	15.226	4.465	15.414	4.485	15.479	4.285	15.338	4.316	15.857	4.471	15.674	4.459	15.469	4.710	16.871
	72	**3.845**	**14.943**	4.321	15.137	4.222	16.051	4.237	16.036	4.495	15.422	4.247	16.036	4.432	15.627	4.536	15.840	4.557	15.701	4.571	17.101
	Avg	**3.710**	**14.452**	4.199	15.057	4.177	15.677	4.317	15.585	4.492	15.468	4.301	15.665	4.406	15.744	4.576	15.944	4.498	15.585	4.627	16.657
APTC	50	**0.386**	**4.645**	0.391	4.679	0.390	4.683	0.390	4.693	0.391	4.678	0.392	4.684	0.391	4.678	0.395	4.709	0.391	4.678	0.571	5.791
	100	**0.375**	**4.566**	0.377	4.580	0.383	4.629	0.379	4.594	0.377	4.580	0.377	4.580	0.377	4.581	0.378	4.587	0.377	4.580	0.634	6.150
	200	**0.375**	**4.544**	0.376	4.552	0.376	4.554	0.377	4.562	0.376	4.551	0.376	4.550	0.376	4.552	0.376	4.550	0.376	4.552	0.401	4.743
	Avg	**0.379**	**4.585**	0.381	4.604	0.383	4.622	0.382	4.616	0.381	4.603	0.382	4.605	0.381	4.604	0.383	4.615	0.381	4.603	0.535	5.561
FNSPID	30	**0.048**	**1.349**	0.157	2.770	1.490	7.793	0.606	6.201	0.373	4.626	0.313	4.708	0.839	7.102	0.831	6.591	0.527	5.226	1.940	8.630
	180	**0.154**	**2.398**	0.202	3.214	0.717	7.236	0.318	4.076	0.350	4.427	0.661	5.493	0.555	5.474	0.910	6.690	0.663	5.991	1.771	8.140
	365	0.413	**3.724**	0.638	5.466	1.001	8.451	1.431	9.210	**0.396**	4.902	1.011	7.301	0.528	5.301	1.073	7.592	1.464	9.175	1.691	8.590
	Avg	**0.205**	**2.490**	0.332	3.817	1.069	7.827	0.785	6.496	0.373	4.652	0.662	5.834	0.641	5.959	0.938	6.958	0.885	6.797	1.801	8.453
GDELT	30	1.878	**8.296**	1.913	8.427	**1.876**	8.397	1.902	8.512	1.893	8.591	1.945	8.865	1.882	8.482	1.900	8.583	1.878	8.480	2.100	9.320
	90	**1.877**	**8.308**	1.918	8.484	1.884	8.413	1.902	8.603	1.886	8.614	1.902	8.731	1.914	8.515	1.912	8.622	2.105	9.902	2.690	9.741
	180	**1.701**	**8.203**	1.746	8.384	1.719	8.317	1.719	8.420	1.732	8.583	2.018	10.197	1.751	8.413	1.782	8.742	1.876	9.317	2.130	9.280
	Avg	**1.819**	**8.269**	1.859	8.432	1.826	8.376	1.841	8.512	1.837	8.596	1.955	9.264	1.849	8.470	1.865	8.649	1.953	9.233	2.307	9.447
Seabirds	60	**0.003**	**0.295**	0.016	1.000	0.008	0.623	0.036	1.466	0.083	2.667	0.103	2.951	0.063	1.898	0.111	2.011	0.110	2.839	0.081	2.981
	360	**0.014**	**0.645**	0.034	1.339	0.020	1.168	0.067	1.940	0.094	2.827	0.061	1.890	0.048	1.634	0.131	2.601	0.047	1.797	0.120	2.191
	720	**0.018**	**0.828**	0.046	1.580	0.039	1.332	0.139	3.275	0.114	2.731	0.130	2.011	0.035	1.384	0.161	2.551	0.067	2.220	0.150	2.240
	Avg	**0.012**	**0.589**	0.032	1.306	0.022	1.041	0.081	2.227	0.097	2.742	0.098	2.284	0.049	1.639	0.134	2.388	0.075	2.285	0.117	2.471
1st Count		**24**	**30**	4	3	3	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0

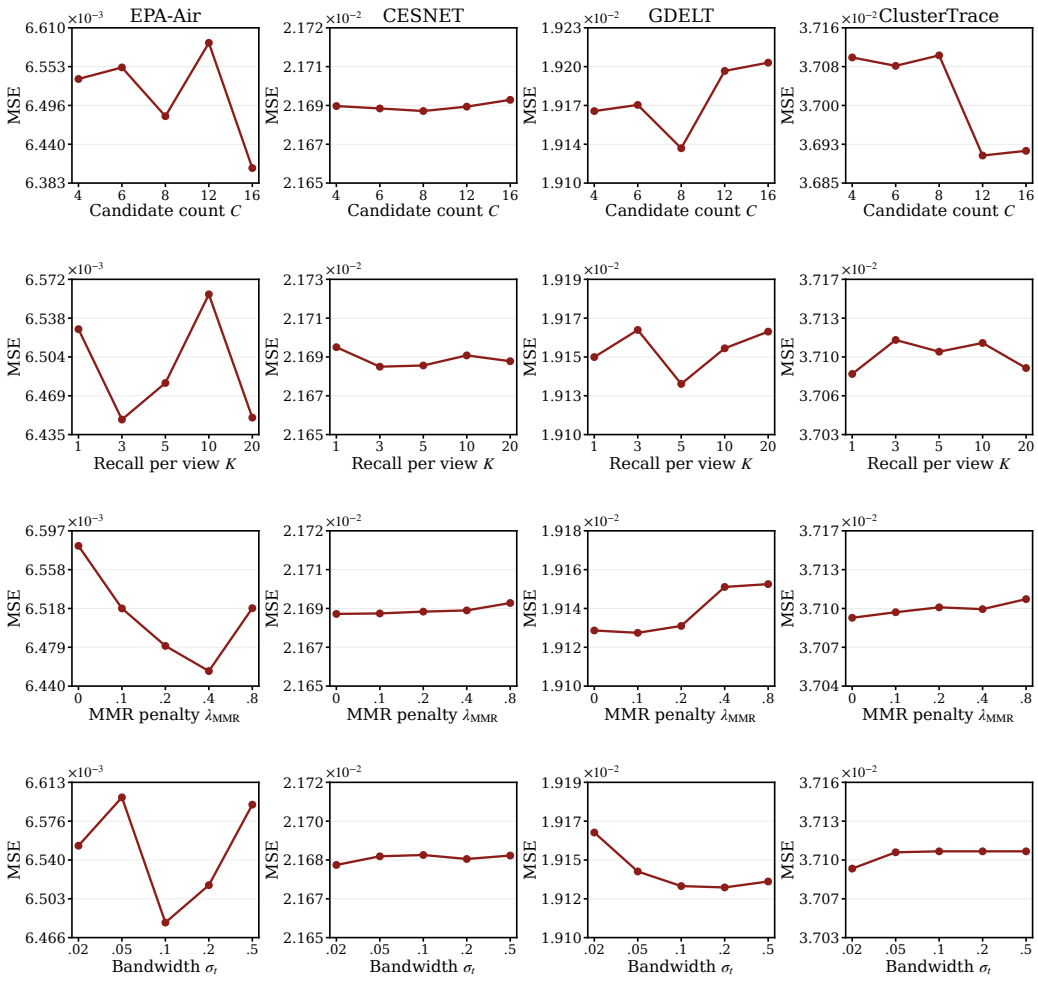

Figure 4: Complete parameter sensitivity of ARISE in actual mean test MSE over three forecasting horizons (lower is better). From top to bottom, the rows vary C, K, λ_{MMR}, and σ_t; columns correspond to EPA-Air, CESNET, GDELT, and ClusterTrace. Tested settings are equally spaced on the horizontal axis. Each panel uses an independent vertical range and five y-axis tick labels, displayed to three decimal places with the indicated scientific-notation multiplier.

Candidate window count C. The first row examines how finely the selector can place a history's left boundary. A small C offers fewer possible contexts, while a larger C provides a finer relative-time grid without changing the right endpoint or the one-entry-per-sample construction. EPA-Air benefits from $C = 16$, reaching MSE 6.405×10^{-3}, but its error first rises to 6.588×10^{-3} at $C = 12$. ClusterTrace achieves its lowest error at $C = 12$ (3.6903×10^{-2}), with little additional change at $C = 16$ (3.6912×10^{-2}). In contrast, CESNET and GDELT favor $C = 8$: CESNET stays within 2.1687×10^{-2}–2.1692×10^{-2}, whereas GDELT increases from 1.9132×10^{-2} at $C = 8$ to 1.9203×10^{-2} at $C = 16$. These responses indicate that finer boundary resolution is useful on some datasets but does not guarantee a better selected context. The non-monotonic and nearly flat curves also rule out a simple relationship between proposal count and forecasting accuracy.

Per-view retrieval budget K. The second row studies the breadth of candidate recall before the three views are combined and filtered. Increasing K makes more histories available to support filtering and MMR, but the reader's final candidate budget R remains a separate parameter. EPA-Air and CESNET reach their lowest observed MSE at $K = 3$, with values of 6.448×10^{-3} and 2.1686×10^{-2}, respectively. GDELT favors $K = 5$ (1.9132×10^{-2}), whereas ClusterTrace favors $K = 1$ (3.7084×10^{-2}). A much larger budget is not consistently better: EPA-Air at $K = 20$ (6.450×10^{-3})

is close to, but does not improve on, $K = 3$. The curves therefore support compact recall pools within the tested settings. A plausible explanation is that larger pools mix useful alternatives with redundant or less relevant histories, limiting the benefit of broader recall.

MMR redundancy penalty λ_{MMR}. The third row evaluates the trade-off between query relevance and candidate diversity. At $\lambda_{\mathrm{MMR}} = 0$, selection has no inter-candidate similarity penalty; increasing it places greater emphasis on avoiding evidence similar to cases already selected. EPA-Air benefits from moderate penalization, with MSE falling from 6.582×10^{-3} at zero to 6.455×10^{-3} at 0.4, then rising to 6.519×10^{-3} at 0.8. GDELT favors 0.1 (1.9128×10^{-2}) and becomes slightly worse at 0.8 (1.9155×10^{-2}). CESNET and ClusterTrace instead attain their lowest errors at zero penalty and remain relatively stable, within 2.1686×10^{-2}–2.1691×10^{-2} and 3.7096×10^{-2}–3.7111×10^{-2}, respectively. These patterns are consistent with diversity being useful when it adds complementary evidence, while excessive penalization may sacrifice relevance. The best observed penalty is therefore dataset-dependent, rather than universally positive or large.

Temporal alignment bandwidth σ_t. The fourth row examines the temporal locality of retrieved future evidence. A narrow Gaussian bandwidth concentrates weight on observations close to each target time; a wider bandwidth combines information from a broader neighborhood and may smooth local changes. The observed response differs across datasets. EPA-Air reaches its lowest MSE at $\sigma_t = 0.1$ (6.481×10^{-3}), with a non-monotonic curve over the full grid. GDELT favors 0.2 (1.9131×10^{-2}), although 0.1 gives a very similar error (1.9132×10^{-2}). CESNET and ClusterTrace attain their lowest errors at 0.02, with MSE 2.1682×10^{-2} and 3.7090×10^{-2}, respectively; ClusterTrace changes little from 0.05 onward. These results are consistent with a dataset-dependent balance between local timing and broader temporal aggregation, rather than a common optimal bandwidth. Numerical bandwidths use each dataset's supplied time coordinates, not a universal duration in seconds.