STABILIZING RECURSIVE RETRIEVAL IN IRREGULAR TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented forecasting extends parametric predictors with historical trajectories retrieved from external temporal memory, yet existing methods largely adopt *static memory*, where retrieved evidence remains independent of the model's own predictions. This limitation can fail in continual forecasting, where previous forecasts are stored and later reused, creating a recursive feedback loop among memory, retrieval, and prediction. We present ECHO (**E**volving **C**losed-loop **H**istorical **O**bservations), a framework for studying and controlling recursive retrieval-augmented forecasting with evolving temporal memory. ECHO combines predictive temporal retrieval with a fixed base forecaster and models memory evolution as a dynamical process. We identify a *self-confirming retrieval loop*, in which retrieval errors bias predictions, biased predictions contaminate memory, and the resulting memory reinforces future forecasting errors. We analyze when such perturbations decay or amplify, introduce *grounding* to anchor memory evolution to observed outcomes, and develop a reliability-aware adaptive memory update. Extensive experiments on four public benchmark datasets demonstrate that ECHO achieves competitive performance compared with a wide range of state-of-the-art methods.

1 INTRODUCTION

Accurate forecasting from irregular multivariate time series (IMTS) is challenging because observations are sparse, asynchronous, and recorded at non-uniform time intervals. A growing body of work has addressed these difficulties through continuous-time dynamics, adaptive patching, graph-based interaction modeling, and irregularity-aware representations (Rubanova et al., 2019; Brouwer et al., 2019; Kidger et al., 2020; Yalavarthi et al., 2024; Zhang et al., 2024; Luo et al., 2025; Li et al., 2025; Liu et al., 2026a;b). These methods substantially improve how irregular observations are represented and how temporal and cross-variable dependencies are propagated. Nevertheless, they remain fundamentally constrained by the information contained in the current observation window. When that window is sparse or ambiguous, increasing model capacity alone does not provide direct access to instance-specific historical trajectories that are absent from the current context.

Retrieval-augmented forecasting offers a complementary source of information. Instead of relying exclusively on the observed history and knowledge encoded in model parameters, a forecaster can retrieve relevant historical trajectories from an external temporal repository and use their subsequent evolution as additional predictive evidence. Recent retrieval-based approaches have demonstrated the usefulness of such external temporal context in cross-domain and zero-shot forecasting (Jiang et al., 2026; Ma et al., 2026).

Figure 1 illustrates this intuition under irregular sampling. The observed history contains sparse measurements at non-uniform intervals, making long-horizon extrapolation uncertain. A local forecaster follows the recent observed trend but gradually drifts away as the forecasting horizon extends. In contrast, a retrieval-augmented prediction can exploit a relevant historical trajectory whose subsequent evolution provides information not available in the current window, reducing long-horizon drift in this example. This motivates treating temporal memory as an explicit component of the forecasting system rather than relying solely on increasingly expressive local representations.

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.

Despite this promise, existing retrieval-augmented forecasting largely treats memory as an *exogenous and static resource*. Its information flow is essentially one-way:

$$\mathcal{M} \rightarrow \mathcal{R} \rightarrow \hat{\mathcal{Y}}, \tag{1}$$

where memory $\mathcal{M}$ determines retrieved evidence $\mathcal{R}$ and therefore influences the prediction $\hat{\mathcal{Y}}$, while the forecast itself does not modify the memory. Consequently, existing work primarily asks *what should be retrieved* and *how retrieved evidence should be used*, while assuming that the retrieval memory remains fixed and exogenous to the model's predictions.

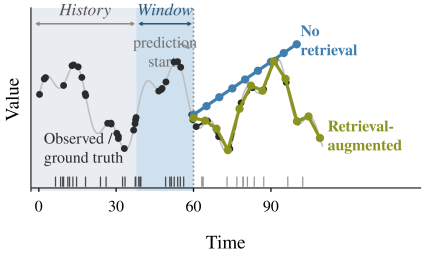


Figure 1: Retrieval reduces long-horizon drift under irregular sampling.

This assumption need not hold under continual forecasting. When the retrieval repository is incrementally updated, previous forecasts may be retained as pseudo-observations or candidate trajectories before their corresponding outcomes become available. This can expand the temporal evidence available to future queries, but an inaccurate stored forecast may also re-enter subsequent predictions through retrieval. Once the corresponding outcomes are observed, they can be used to ground subsequent memory updates. The information flow then becomes recursive:

$$\mathcal{M}_t \rightarrow \mathcal{R}_t \rightarrow \hat{\mathcal{Y}}_t \rightarrow \mathcal{M}_{t+1}. \tag{2}$$

This seemingly small change fundamentally alters the forecasting problem. Memory is no longer a fixed database; it becomes a dynamical state partially shaped by the model's own previous outputs. Similar recursive reuse of model-generated information has been associated with self-reinforcing degradation in broader generative and self-improving systems (Shumailov et al., 2024; Chen et al., 2026). The corresponding stability problem, however, has received little attention in retrieval-augmented forecasting.

Challenge 1: When does recursive retrieval help, remain stable, or degrade? Static retrieval keeps memory fixed during forecasting, whereas recursive retrieval allows model-generated forecasts to re-enter memory as future retrieval evidence. Under static retrieval, a poor retrieved trajectory primarily affects one forecast; under recursive retrieval, its influence can persist across future rounds. A mismatched retrieval can introduce bias into the current prediction. This biased prediction may subsequently be incorporated into memory, making the retrieval of similarly inaccurate evidence more likely in later stages. We refer to this positive-feedback mechanism as a *self-confirming retrieval loop*:

$$\text{retrieval error} \rightarrow \text{prediction error} \rightarrow \text{memory contamination} \rightarrow \text{future retrieval errors}. \tag{3}$$

Importantly, recursion is not necessarily harmful. Accurate forecasts can also expand memory with useful trajectories and improve future retrieval. Recursive retrieval can therefore exhibit qualitatively different regimes, ranging from beneficial memory expansion to bounded degradation and self-reinforcing failure. The central question is thus not merely whether retrieval improves one-step forecasting accuracy, but *what determines the long-term stability of the retrieval–prediction–memory feedback loop?*

Challenge 2: How can memory evolve without amplifying unreliable predictions? Freezing memory eliminates recursive contamination, but also removes the ability to accumulate new temporal evidence. At the opposite extreme, unrestricted write-back allows prediction errors to repeatedly re-enter the retrieval process. Recursive retrieval therefore introduces a fundamental trade-off between *plasticity* and *reliability*: memory should adapt when generated trajectories are useful, but remain anchored when those trajectories are uncertain or inconsistent with reliable evidence. Furthermore, prediction confidence alone is insufficient for this purpose: a plausible but systematically biased trajectory may remain internally consistent while being repeatedly retrieved and reinforced. The second question is therefore *how strongly should model-generated evidence be allowed to shape future memory, and how much external grounding is required to suppress unstable feedback without eliminating useful memory evolution?*

To address these challenges, we propose `ECHO` (**E**volving **C**losed-loop **H**istorical **O**bservations), a framework for studying and controlling retrieval-augmented forecasting with evolving temporal

memory. Rather than introducing a new forecasting backbone, ECHO is built on top of a base IMTS forecaster (Liu et al., 2026b). The forecasting model is trained in the standard way before forecasting and remains fixed thereafter, so that changes in forecasting behavior can be attributed to memory evolution rather than continual parameter adaptation.

ECHO first establishes a predictive temporal-memory interface for irregular observations. Historical contexts are represented at both global and local temporal scales, allowing retrieval to combine whole-trajectory similarity with finer-grained historical matching. Retrieved trajectories are then reweighted according to their predictive utility and incorporated through a bounded residual correction, allowing retrieval to augment a strong IMTS forecaster without replacing its underlying prediction mechanism.

The central object of ECHO is the evolving retrieval memory. At each forecasting round, the current observation queries the existing memory, the retrieved evidence is used by the fixed forecaster to make a prediction, and the resulting trajectory may subsequently influence future memory. The true outcome is revealed only after the current forecast is made and can therefore be used as delayed grounding for later memory updates without leaking information into the current prediction. Repeated over time, this process forms a closed feedback loop between memory, retrieval, and forecasting.

We analyze this feedback process by treating aggregate memory reliability as a dynamical state and characterizing when recursive memory evolution remains stable or amplifies perturbations. We then introduce *grounding*, which incorporates externally observed outcomes into memory evolution to attenuate endogenous feedback from model-generated trajectories. Building on this principle, ECHO estimates the reliability of the current retrieval–prediction state and adaptively controls the extent to which model-generated information is admitted into memory.

Our contributions are summarized as follows: **(i)** We formulate **recursive retrieval-augmented forecasting** with evolving temporal memory, where model-generated forecasts may become future retrieval evidence and transform a one-way retrieval pipeline into a closed-loop dynamical system. **(ii)** We develop **predictive temporal retrieval** for irregular observations, combining global and local historical matching with memory utility and selective residual integration. **(iii)** We identify the **self-confirming retrieval loop**, characterize the local stability of recursive memory dynamics, and establish grounding as a mechanism for attenuating endogenous feedback. **(iv)** We conduct extensive experiments on **four IMTS forecasting benchmarks** and, on selected benchmarks, systematically compare static, recursive, partially grounded, grounded, and adaptive memory regimes through repeated memory-evolution rounds, revealing how different memory-update strategies affect subsequent forecasting performance.

2 Related Work

2.1 Irregular Multivariate Time-Series Forecasting

Irregular multivariate time-series forecasting requires modeling non-uniform temporal intervals, asynchronous observations, and dependencies across variables. Continuous-time approaches model latent evolution between irregular observations using differential-equation-based dynamics (Chen et al., 2023; Rubanova et al., 2019; Brouwer et al., 2019; Kidger et al., 2020; Chen et al., 2018; Mercatali et al., 2024). Graph-based methods instead organize irregular observations into explicit interaction structures to capture temporal and cross-variable dependencies (Liu et al., 2026a; Yalavarthi et al., 2024; Zhang et al., 2024; Luo et al., 2025; Escudero-Arnanz et al., 2025; Luo et al., 2024). Patch- and multi-scale approaches further aggregate irregular observations into adaptive temporal representations, improving the modeling of local and global dynamics under heterogeneous sampling (Li et al., 2025; Liu et al., 2026a;b; Luo et al., 2025; Qiu et al., 2026).

These methods substantially improve how information within the observed history is represented and propagated. ECHO addresses a complementary problem: rather than introducing another IMTS backbone, we study how *external temporal evidence* can be retrieved and, more importantly, how such evidence evolves when retrieval memory becomes recursively coupled with previous forecasts.

Additional related work on retrieval-augmented forecasting and recursive feedback with model-generated data is provided in Appendix A.

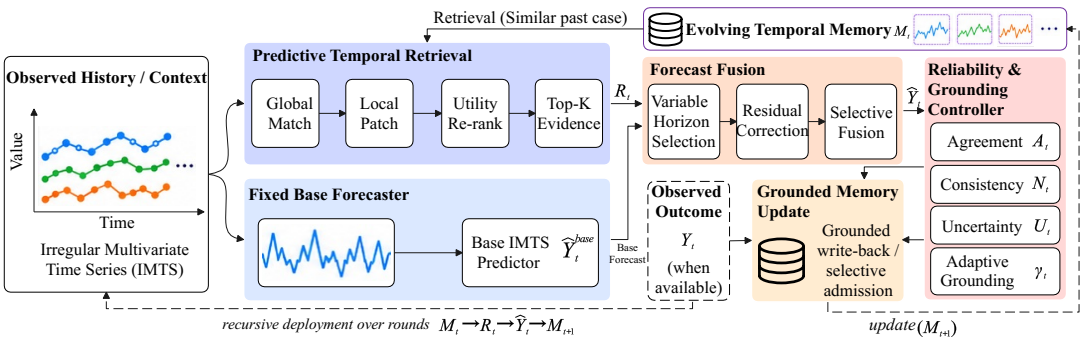


Figure 2: Overview of ECHO.

3 METHODOLOGY

Figure 2 illustrates ECHO. Given an irregular history, a fixed base forecaster produces a local prediction while predictive temporal retrieval provides complementary historical evidence. During forecasting, retrieval memory evolves through model-generated and externally observed trajectories, and a reliability-aware grounding mechanism regulates this recursive feedback.

3.1 PROBLEM DEFINITION

Let $\mathcal{S}_t$ denote the observed irregular history at forecasting round t, $\mathcal{Q}_t$ the set of future query locations, each specifying a forecast time and variable index, and $\mathcal{Y}_t$ the corresponding target values. Given temporal memory $\mathcal{M}_t$, retrieval operator $\mathcal{G}$ extracts:

$$\mathcal{R}_t = \mathcal{G}(\mathcal{S}_t, \mathcal{M}_t), \tag{4}$$

and a forecaster F_θ produces:

$$\hat{\mathcal{Y}}_t = F_\theta(\mathcal{S}_t, \mathcal{Q}_t, \mathcal{R}_t). \tag{5}$$

and a forecaster F_θ produces:

$$\hat{\mathcal{Y}}_t = F_\theta(\mathcal{S}_t, \mathcal{Q}_t, \mathcal{R}_t). \tag{6}$$

Unlike conventional retrieval with externally fixed memory, ECHO allows model-generated trajectories to affect future retrieval:

$$\mathcal{M}_{t+1} = \mathcal{U}\left(\mathcal{M}_t, \mathcal{S}_t, \mathcal{Y}_t, \hat{\mathcal{Y}}_t\right), \tag{7}$$

where $\mathcal{U}$ is the memory-update operator. The true outcome $\mathcal{Y}_t$ is unavailable when $\hat{\mathcal{Y}}_t$ is produced and is used only after the forecasting round for delayed grounding.

Throughout recursive forecasting, the forecasting parameters remain fixed, $\theta_t = \theta$. Hence, the dynamics of interest are induced entirely by the evolving retrieval memory and its effect on subsequent retrieval and prediction.

3.2 PREDICTIVE TEMPORAL RETRIEVAL

ECHO retrieves temporal evidence at both global and local historical scales. Let $\mathbf{q}_t^{\text{glob}}$ and $\{\mathbf{q}_t^{(g)}\}_{g=1}^{G}$ denote the global and local retrieval keys computed from the encoded observed history $\mathcal{S}_t$, where $\mathbf{q}_t^{\text{glob}}$ summarizes the full context and $\mathbf{q}_t^{(g)}$ represents the g-th local temporal region. Memory items store the corresponding keys $\mathbf{k}_m^{\text{glob}}$ and $\{\mathbf{k}_m^{(g)}\}_{g=1}^{G}$.

To reduce similarity collapse caused by dominant common embedding directions, we use centroid-debiased normalized keys:

$$\tilde{\mathbf{z}} = \frac{\mathbf{z} - \bar{\mathbf{k}}}{\|\mathbf{z} - \bar{\mathbf{k}}\|_2}, \tag{8}$$

where $\bar{\mathbf{k}}$ denotes the centroid, i.e., the average of all memory keys.

The global and local similarities are:

$$s_{tm}^{\text{glob}} = \left\langle \tilde{\mathbf{q}}_t^{\text{glob}}, \tilde{\mathbf{k}}_m^{\text{glob}} \right\rangle, \tag{9}$$

$$s_{tm}^{\text{l}} = \frac{1}{G} \sum_{g=1}^{G} \max_{g'} \left\langle \tilde{\mathbf{q}}_t^{(g)}, \tilde{\mathbf{k}}_m^{(g')} \right\rangle. \tag{10}$$

They are combined as:

$$s_{tm}^{\text{hist}} = (1 - \lambda_p) s_{tm}^{\text{glob}} + \lambda_p s_{tm}^{\text{l}}. \tag{11}$$

When irregularity-specific similarity is used,

$$s_{tm}^{\text{base}} = (1 - \lambda_{\text{irr}}) s_{tm}^{\text{hist}} + \lambda_{\text{irr}} s_{tm}^{\text{irr}}. \tag{12}$$

where s_{tm}^{irr} measures similarity between the irregular sampling patterns of the query and memory item using their temporal-gap and observation-mask representations.

Predictive utility. Similarity alone does not guarantee forecasting usefulness. Each memory item therefore maintains a utility score u_m, yielding:

$$s_{tm} = s_{tm}^{\text{base}} + \lambda_u u_m. \tag{13}$$

where u_m denotes the predictive utility of memory item m, estimated from the improvement of retrieval-augmented forecasting over the base prediction. The coefficients λ_p and λ_u are fixed across datasets, while $\lambda_{\text{irr}} = 0$ in all reported experiments; their values are given in Appendix B.3.

The Top-K candidates are selected according to s_{tm} and normalized with temperature τ.

Selective future integration. Different retrieved trajectories may be useful for different variables and forecast horizons. ECHO therefore performs variable–horizon-specific selection over the Top-K retrieved futures, producing $\mathcal{V}_t^*$.

Rather than replacing the base prediction, retrieved evidence provides a residual correction:

$$\hat{\mathcal{Y}}_t = \hat{\mathcal{Y}}_t^{\text{base}} + \beta \, \Delta\mathcal{Y}_t^{\text{ret}}, \tag{14}$$

where $\Delta\mathcal{Y}_t^{\text{ret}}$ denotes the retrieval-derived correction computed from the selected future evidence $\mathcal{V}_t^*$, and $\beta \in [0, 1]$ is a fixed residual-mixing coefficient shared across all reported experiments. Eq. 14 also admits an optional learned gate $\mathbf{G}_t \in [0, 1]$, conditioned on the forecasting and retrieval states, in place of β, which would allow ECHO to adaptively revert toward the base forecaster when retrieved evidence is uninformative; we did not find this adaptation to improve on the fixed β in preliminary experiments and therefore use the simpler fixed coefficient throughout.

3.3 Recursive Memory Dynamics

At forecasting round t, temporal memory contains historical contexts, their subsequent trajectories, and associated utility scores:

$$\mathcal{M}_t = \{(\mathcal{K}_m, \mathcal{V}_m, u_m)\}_{m=1}^{|\mathcal{M}_t|}. \tag{15}$$

We consider five forecasting regimes: *static*, *recursive*, *partially grounded*, *grounded*, and *adaptive*. Static memory performs no write-back. Recursive and grounded memory write $\hat{\mathcal{Y}}_t$ and $\mathcal{Y}_t$, respectively. Partial grounding selects:

$$\tilde{\mathcal{Y}}_t^{(\gamma)} = \begin{cases} \mathcal{Y}_t, & \text{with probability } \gamma, \\ \hat{\mathcal{Y}}_t, & \text{with probability } 1 - \gamma, \end{cases} \tag{16}$$

where $\gamma = 0$ and $\gamma = 1$ recover fully recursive and fully grounded memory.

To characterize recursive behavior independently of a particular memory implementation, let $p_t \in [0, 1]$ denote an abstract scalar state summarizing aggregate memory reliability rather than a directly

observed empirical quantity, and let $f(p_t)$ denote the expected reliability of newly generated evidence. We introduce $\alpha \in (0, 1]$ as an abstract coefficient representing the effective influence of newly admitted evidence on the aggregate memory state:

$$p_{t+1} = (1 - \alpha)p_t + \alpha f(p_t). \tag{17}$$

An equilibrium p^* satisfies $p^* = f(p^*)$. Let $\delta_t = p_t - p^*$ denote a small perturbation around the equilibrium. Linearization around p^* gives:

$$\delta_{t+1} \approx [1 - \alpha + \alpha f'(p^*)]\, \delta_t, \tag{18}$$

and hence local stability requires:

$$|1 - \alpha + \alpha f'(p^*)| < 1. \tag{19}$$

Thus, contractive feedback suppresses perturbations, whereas expansive feedback can repeatedly amplify retrieval and prediction errors, producing a self-confirming retrieval regime.

3.4 Grounded and Adaptive Feedback Control

Grounding limits recursive error amplification by anchoring memory updates to externally observed outcomes. Starting from the recursive dynamics in Eq. 17, we replace the reliability of newly admitted evidence $f(p_t)$ by the grounded mixture $(1 - \gamma)f(p_t) + \gamma q$, where $\gamma \in [0, 1]$ is the grounding rate and $q \in [0, 1]$ is the reliability of grounded evidence. This yields:

$$p_{t+1} = (1 - \alpha)p_t + \alpha\left[(1 - \gamma)f(p_t) + \gamma q\right]. \tag{20}$$

At an equilibrium p^*_γ, the local feedback coefficient is:

$$\lambda_\gamma = 1 - \alpha + \alpha(1 - \gamma)f'(p^*_\gamma), \tag{21}$$

and stability requires $|\lambda_\gamma| < 1$. Grounding therefore attenuates the endogenous feedback contribution by the factor $(1 - \gamma)$.

If f is L-Lipschitz, a sufficient contraction condition is:

$$(1 - \gamma)L < 1, \tag{22}$$

which, for $L > 1$, is satisfied when

$$\gamma > 1 - \frac{1}{L}. \tag{23}$$

This establishes a sufficient grounding threshold and exposes the stability–plasticity trade-off: stronger grounding improves stability but reduces autonomous memory evolution.

Reliability-aware grounding. A fixed γ cannot account for varying prediction reliability across queries and rounds. `ECHO` therefore estimates

$$r_t = \sigma\left(\eta_1 A_t + \eta_2 N_t - \eta_3 U_t\right), \tag{24}$$

where $\sigma(\cdot)$ denotes the sigmoid function, A_t measures forecast–retrieval agreement, N_t retrieval consistency, and U_t uncertainty. The negative sign before U_t reflects that higher uncertainty should reduce the estimated reliability.

The adaptive grounding rate is:

$$\gamma_t = \mathrm{clip}\left(1 - r_t, \gamma_{\min}, \gamma_{\max}\right), \tag{25}$$

where $\gamma_{\min}$ and $\gamma_{\max}$ denote the lower and upper bounds on the adaptive grounding rate, respectively. High-reliability predictions therefore receive weaker grounding, whereas uncertain predictions receive stronger external control.

Once $\mathcal{Y}_t$ becomes available, the write-back target is

$$\tilde{\mathcal{Y}}_t = \begin{cases} \mathcal{Y}_t, & \text{with probability } \gamma_t, \\ \hat{\mathcal{Y}}_t, & \text{with probability } 1 - \gamma_t, \end{cases} \tag{26}$$

followed by

$$\mathcal{M}_{t+1} = \mathcal{U}_{\text{mem}}\left(\mathcal{M}_t, \mathcal{S}_t, \tilde{\mathcal{Y}}_t\right), \tag{27}$$

Here $\mathcal{U}_{\text{mem}}$ denotes the memory admission and retention policy. In the default configuration, new trajectories are appended to an unbounded retrieval pool; under a finite memory budget, the oldest entries are evicted using FIFO replacement. The resulting closed loop is:

$$\mathcal{M}_t \to \mathcal{R}_t \to \hat{\mathcal{Y}}_t \to r_t \to \gamma_t \to \mathcal{M}_{t+1}. \tag{28}$$

3.5 Training Objective

F_θ (the backbone together with the retrieval-fusion residual of Section 3.2) is trained on the training split under the static (non-recursive) supervised forecasting setting of Section 3.1, i.e., without the round-by-round memory of Eq. 7: a single training run produces the frozen parameters θ subsequently used throughout recursive forecasting. Training begins with $\mathcal{R}$ unavailable ($\hat{\mathcal{Y}} = F_\theta(\mathcal{S}, \mathcal{Q})$); after a fixed warm-up period, retrieval memory is constructed from the training split and periodically refreshed for the remainder of training, so that later epochs optimize the same objective with retrieval active, $\hat{\mathcal{Y}} = F_\theta(\mathcal{S}, \mathcal{Q}, \mathcal{R})$. In both phases, given future targets $\mathcal{Y}$ and an observation mask $\mathcal{Y}^{\text{mask}}$ accounting for the irregularly-sampled targets characteristic of IMTS forecasting, we minimize the masked mean squared error over observed target entries:

$$\mathcal{L}(\theta) = \frac{\sum \mathcal{Y}^{\text{mask}} \odot \left(\hat{\mathcal{Y}} - \mathcal{Y}\right)^2}{\sum \mathcal{Y}^{\text{mask}}}, \tag{29}$$

where $\odot$ denotes elementwise multiplication and the sums range over all variables and predicted time steps in the batch. Model selection uses validation performance under the same objective.

4 Experiments

Datasets. We evaluate ECHO on four widely used public benchmarks for irregular multivariate time-series forecasting: **PhysioNet** (Johnson et al., 2014), **MIMIC** (Johnson et al., 2016), **Human Activity** (Rubanova et al., 2019), and **USHCN** (Menne et al., 2015). These datasets span diverse application domains, including healthcare, biomechanics, and climate science, and exhibit substantially different dimensionalities and sampling characteristics. We follow the standard preprocessing and forecasting protocols adopted in prior IMTS forecasting studies (Yalavarthi et al., 2024; Zhang et al., 2023; Li et al., 2025). Dataset statistics and detailed preprocessing procedures are provided in Appendix B.1.

Implementation Details. We train the forecasting backbone using only the training split and freeze the resulting checkpoint for all recursive memory-evolution experiments (Liu et al., 2026b). The initial memory $\mathcal{M}_0$ is constructed exclusively from training observations. We use $K = 2$ retrieved temporal patterns and retrieval temperature $\tau = 0.0012$, selected on the validation set; sensitivity to the retrieval size is reported in Appendix B.5. We set $\gamma_{\min} = 0$ and $\gamma_{\max} = 1$ to allow the adaptive controller to span the full range from recursive to fully grounded updates. The theoretical coefficient α denotes the effective influence of newly admitted evidence. The default ECHO configuration uses an unbounded retrieval pool to isolate the effect of recursive memory evolution from the choice of eviction policy. Under a finite memory budget, we use FIFO replacement and evaluate its sensitivity separately in Table 3. Memory evolution is performed on the validation split after warm-up, while the held-out test set is evaluated with the resulting memory state frozen. Additional training and reproducibility details are provided in Appendix B.3.

4.1 Overall Performance

Table 1 reports the overall forecasting performance across four IMTS benchmarks. ECHO achieves the best result on five of the eight dataset–metric combinations and the second-best result on two others, demonstrating consistently strong performance across heterogeneous datasets. Specifically, ECHO ranks first in both MSE and MAE on Human Activity and MIMIC, and achieves the best MSE on USHCN. Relative to the strongest competing results, it reduces MSE by approximately

Table 1: Overall performance on IMTS forecasting. Results are reported as mean ± standard deviation. The best and second-best results are highlighted in **bold** and with an underline, respectively.

Dataset	Human Activity		USHCN		PhysioNet		MIMIC	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PrimeNet	4.2511±0.0024	1.7018±0.0006	0.4932±0.0010	0.4956±0.0023	0.7953±0.0000	0.6859±0.0000	0.9073±0.0001	0.6614±0.0001
NeuralFlows	0.1722±0.0081	0.3150±0.0084	0.2087±0.0231	0.3157±0.0167	0.4056±0.0030	0.4466±0.0024	0.6085±0.0101	0.5306±0.0066
GRU	0.1387±0.0073	0.2607±0.0092	0.2168±0.0162	0.3180±0.0248	0.6179±0.0045	0.5778±0.0031	0.5895±0.0092	0.5151±0.0048
mTAN	0.0937±0.0027	0.2224±0.0046	0.6839±0.2520	0.5284±0.0603	0.3807±0.0041	0.4290±0.0035	0.9408±0.1126	0.6755±0.0459
SeFT	1.3767±0.0021	0.9753±0.0006	0.3345±0.0020	0.4083±0.0076	0.7796±0.0019	0.6779±0.0028	0.9230±0.0015	0.6628±0.0008
GNeuralFlow	0.4087±0.1691	0.4601±0.0852	0.2229±0.0367	0.3310±0.0351	0.7949±0.0705	0.6621±0.0316	0.8957±0.0209	0.6450±0.0072
GRU-D	0.1769±0.0363	0.3152±0.0284	0.2098±0.0440	0.3053±0.0265	0.3419±0.0026	0.3992±0.0010	0.4759±0.0100	0.4526±0.0055
Raindrop	0.0903±0.0032	0.2092±0.0048	0.2035±0.0336	0.3029±0.0264	0.3480±0.0008	0.4042±0.0017	0.6754±0.1829	0.5444±0.0868
Warpformer	0.0449±0.0009	0.1228±0.0016	0.1889±0.0530	0.2945±0.0524	0.3053±0.0008	0.3657±0.0012	0.4302±0.0035	0.4025±0.0014
tPatchGNN	0.0441±0.0010	0.1239±0.0028	0.1784±0.0272	0.3006±0.0333	0.3118±0.0033	0.3684±0.0029	0.4431±0.0115	0.4077±0.0088
GraFITi	0.0437±0.0005	0.1221±0.0017	0.1643±0.0034	0.2690±0.0197	0.3070±0.0020	0.3631±0.0034	0.4359±0.0455	0.4142±0.0297
Hi-Patch	0.0435±0.0002	0.1204±0.0009	0.1753±0.0238	0.2718±0.0193	0.3072±0.0026	0.3674±0.0036	0.4279±0.0010	0.4033±0.0032
APN	0.0421±0.0001	0.1159±0.0006	0.1591±0.0123	0.2612±0.0150	0.3133±0.0003	0.3661±0.0008	0.4292±0.0027	0.4016±0.0016
ASTGI	0.0420±0.0011	0.1208±0.0036	0.1598±0.0114	**0.2601±0.0146**	**0.3017±0.0006**	**0.3599±0.0008**	0.3909±0.0017	0.3852±0.0004
ECHO	**0.0412±0.0003**	**0.1142±0.0017**	**0.1485±0.0043**	0.2605±0.0061	0.3080±0.0008	0.3620±0.0013	**0.3904±0.0003**	**0.3769±0.0012**

1.9% on Human Activity and 6.7% on USHCN, while improving MAE by about 1.5% on Human Activity and 2.2% on MIMIC. ASTGI provides the closest overall baseline, performing best on PhysioNet and attaining the lowest MAE on USHCN, whereas ECHO is stronger on Human Activity and MIMIC and achieves the lowest MSE on USHCN. On the remaining settings, ECHO achieves comparable results to the strongest baselines. The results indicate that ECHO delivers consistently competitive forecasting performance across datasets with substantially different sampling characteristics. Its gains are consistent with the combination of global–local temporal retrieval, utility-aware re-ranking, and residual integration, which adds historical evidence without replacing the base forecaster.

4.2 Model Analyses

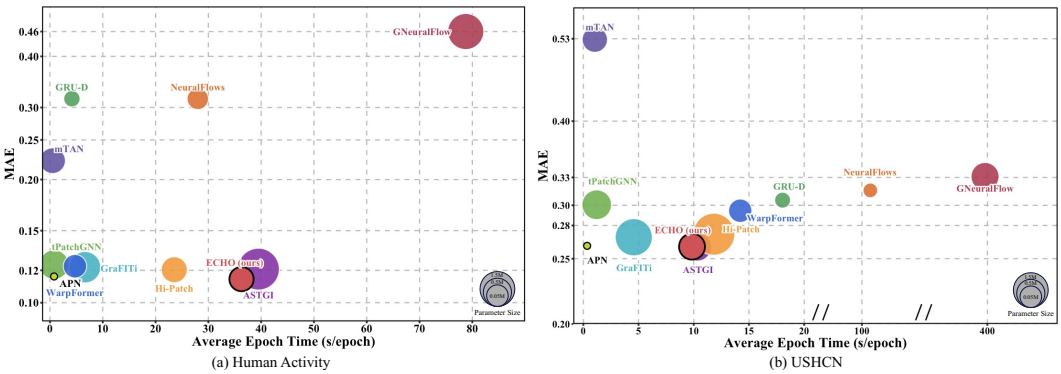
Ablation Studies Table 2 ablates the three reliability signals used by adaptive grounding (A_t, N_t, U_t; each removed by zeroing its weighting coefficient η_i) as well as the adaptive control itself (replaced by a fixed grounding rate). All ablated variants yield mean errors close to the full ECHO configuration, suggesting that no single reliability signal alone accounts for the overall performance.

Table 2: Ablation studies for ECHO. *Fixed Grounding* replaces the adaptive controller with a fixed grounding rate $\gamma = 0.5$.

Dataset	Human Activity		USHCN		PhysioNet		MIMIC	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
w/o Agreement	0.0418±0.0004	0.1162±0.0020	0.1548±0.0050	0.2640±0.0068	0.3092±0.0007	0.3639±0.0012	0.3912±0.0007	0.3802±0.0015
w/o Consistency	0.0419±0.0004	0.1161±0.0019	0.1546±0.0050	0.2636±0.0073	0.3091±0.0007	0.3636±0.0009	0.3909±0.0003	0.3776±0.0010
w/o Uncertainty	0.0418±0.0004	0.1160±0.0021	0.1550±0.0048	0.2640±0.0069	0.3093±0.0007	0.3641±0.0016	0.3908±0.0002	0.3774±0.0011
Fixed Grounding	0.0418±0.0004	0.1161±0.0020	0.1548±0.0051	0.2642±0.0066	0.3092±0.0006	0.3645±0.0017	0.3910±0.0005	0.3798±0.0017
ECHO	**0.0412±0.0003**	**0.1142±0.0017**	**0.1485±0.0043**	**0.2605±0.0061**	**0.3080±0.0008**	**0.3620±0.0013**	**0.3904±0.0003**	**0.3769±0.0012**

Model Efficiency Analysis Figure 3 compares forecasting accuracy and training efficiency on Human Activity and USHCN. The horizontal axis denotes the average training time per epoch, the vertical axis reports MAE, and bubble area represents the number of trainable parameters. Overall, ECHO occupies a favorable accuracy–efficiency region: it achieves near-best forecasting accuracy with a compact parameter budget, while avoiding the substantially higher computational cost of heavier continuous-time models such as GNeuralFlow. Compared with lightweight baselines such as APN, mTAN, and tPatchGNN, ECHO incurs additional computation due to retrieval and memory operations, but this overhead is accompanied by stronger forecasting accuracy.

As shown in Table 1, ASTGI is the closest competitor to ECHO, making it the most informative reference for efficiency comparison. On Human Activity, ECHO achieves better forecasting accuracy while using about **8.4× fewer parameters** than ASTGI, with slightly lower training cost per epoch. On USHCN, the two methods achieve nearly identical MAE, while ECHO attains a lower MSE with about **1.4× fewer parameters** and comparable training time. These results show that ECHO reaches accuracy comparable to or better than its closest competitor with a substantially more compact model, particularly on Human Activity.



(a) Human Activity (b) USHCN

Figure 3: Accuracy–efficiency comparison on Human Activity and USHCN. The bubble area is proportional to the number of trainable parameters.

4.3 Parameter Sensitivity

Effect of Memory Capacity. We study the sensitivity of ECHO to the capacity of the evolving retrieval memory. Forecasting performance remains largely stable across the evaluated finite-capacity settings on both datasets, while the unbounded configuration achieves the best performance. This suggests that ECHO is relatively insensitive to moderate changes in memory capacity, while retaining a larger historical pool can be beneficial in some settings.

Table 3: Effect of memory capacity on Human Activity and USHCN.

Memory Capacity	Human Activity		USHCN	
	MSE	MAE	MSE	MAE
50	0.0418±0.0005	0.1159±0.0016	0.1549±0.0050	0.2640±0.0069
100	0.0417±0.0005	0.1156±0.0017	0.1551±0.0049	0.2644±0.0064
200	0.0418±0.0005	0.1160±0.0018	0.1549±0.0050	0.2640±0.0069
400	0.0418±0.0004	0.1157±0.0016	0.1549±0.0048	0.2641±0.0068
800	0.0418±0.0004	0.1157±0.0018	0.1549±0.0051	0.2638±0.0070
1600	0.0418±0.0005	0.1159±0.0017	0.1551±0.0049	0.2645±0.0064
3200	0.0418±0.0005	0.1159±0.0020	0.1553±0.0047	0.2645±0.0064
∞ (unbounded)	**0.0412±0.0003**	**0.1142±0.0017**	**0.1485±0.0043**	**0.2605±0.0061**

Additional sensitivity analyses of model dimension and retrieval size are provided in Appendix B.5.

5 Conclusion

ECHO studies retrieval-augmented forecasting with recursive memory updates, where model-generated predictions are written back into memory and later reused as retrieval evidence. We show that this closed loop can amplify forecasting errors and induce self-reinforcing degradation in retrieval memory. To address this issue, ECHO combines predictive temporal retrieval with reliability-aware grounding while keeping the base forecaster fixed during forecasting. ECHO is constructed on an in-depth analysis that characterizes the local stability of recursive memory dynamics and establishes a sufficient grounding condition for suppressing error amplification. Empirically, ECHO achieves five best and two second-best results across eight evaluation settings on four IMTS benchmarks, while maintaining a favorable accuracy–efficiency trade-off. These results highlight the importance of treating retrieval memory as an evolving system component and adapting memory feedback for reliable continual forecasting.

A limitation of the current analysis is that the scalar reliability state p_t abstracts the high-dimensional retrieval memory and therefore does not establish global convergence of the complete forecasting system. In addition, our experiments focus on simple append-only or FIFO memory management and do not systematically study utility-aware retention or other selective eviction policies. Future work could extend the stability analysis to richer memory states and investigate adaptive retention strategies under long forecasting horizons.

AI USE STATEMENT

Generative AI tools were used for language polishing, grammar checking, and improving clarity and readability.

REFERENCES

Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. Relaxed recursive transformers: Effective parameter sharing with layer-wise lora. In *ICLR*, 2025.

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

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *NeurIPS*, pp. 7377–7388, 2019.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David A. Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *CoRR*, abs/1606.01865, 2016.

Mingguang Chen, Licheng Wang, and Bo Qu. Recursive self-improvement in AI: from bounded self-refinement to autonomous research loops. *CoRR*, abs/2607.07663, 2026.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, pp. 6572–6583, 2018.

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Contiformer: Continuous-time transformer for irregular time series modeling. In *NeurIPS*, pp. 47143–47175, 2023.

Ranak Roy Chowdhury, Jiacheng Li, Xiyuan Zhang, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. Primenet: Pre-training for irregular multivariate time series. In *AAAI*, pp. 7184–7192, 2023.

Óscar Escudero-Arnanz, Cristina Soguero-Ruíz, and Antonio G. Marques. Explainable spatio-temporal gcnns for irregular multivariate time series: Architecture and application to ICU patient data. *IEEE Trans. Signal Inf. Process. over Networks*, 11:1286–1301, 2025.

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

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

Yue Jiang, Chenxi Liu, Yile Chen, Qin Chao, Shuai Liu, Cheng Long, and Gao Cong. Cross-city time series forecasting with retrieval-augmented large language models. In *WWW*, pp. 4962–4973, 2026.

Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. Data preprocessing and mortality prediction: the physionet/cinc 2012 challenge revisited. In *CinC*, pp. 157–160, 2014.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

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

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

Kenny Ye Liang, Zhongyi Pei, Huan Zhang, Yuhui Liu, Shaoxu Song, and Jianmin Wang. Retrieval-augmented generation with covariate time series. *CoRR*, abs/2603.04951, 2026.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. In *NeurIPS*, pp. 2766–2786, 2024.

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*, pp. 23873–23881, 2026b.

Yicheng Luo, Zhen Liu, Linghao Wang, Binquan Wu, Junhao Zheng, and Qianli Ma. Knowledge-empowered dynamic graph network for irregularly sampled medical time series. In *NeurIPS*, pp. 67172–67199, 2024.

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

Shichao Ma, Zhengyang Zhou, Qihe Huang, Binwu Wang, and Yang Wang. Quizsf: A retrieval-augmented framework for zero-shot time series forecasting. In *WWW*, pp. 1947–1958, 2026.

Matthew J Menne, Claude N Williams Jr, and Russell S Vose. Long-term daily and monthly climate records from stations across the contiguous united states (us historical climatology network). *Technical report, NOAA National Climatic Data Center*, 2015.

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

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Y. 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*, 2025.

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*, 2026.

Yulia Rubanova, Tian Qi Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *NeurIPS*, pp. 5321–5331, 2019.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *ICML*, pp. 19388–19405, 2022.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *ICLR*, 2021.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759, 2024.

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. In *AAAI*, pp. 16255–16263, 2024.

Sean Bin Yang, Ying Sun, Jilin Hu, Zongyi Xu, Kristian Torp, Hua Lu, Bin Yang, and Christian S. Jensen. Refine: Trajectory representation learning via closed-looptranscription. In *KDD*, 2026a.

Xiyuan Yang, Jiaru Zou, Rui Pan, Ruizhong Qiu, Pan Lu, Shizhe Diao, Jindong Jiang, Hanghang Tong, Tong Zhang, Markus J. Buehler, Jingrui He, and James Zou. Recursive multi-agent systems. *CoRR*, abs/2604.25917, 2026b.

Huanyu Zhang, Chang Xu, Yifan Zhang, Zhang Zhang, Liang Wang, and Jiang Bian. Timeraf: Retrieval-augmented foundation model for zero-shot time series forecasting. *IEEE Trans. Knowl. Data Eng.*, 37(9):5654–5665, 2025.

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*, pp. 60179–60196, 2024.

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

A Related Work

A.1 Retrieval-Augmented Time-Series Forecasting

Retrieval-augmented forecasting complements parametric predictors with explicit temporal evidence retrieved from external repositories (Han et al., 2025; Liu et al., 2024; Ning et al., 2025; Zhang et al., 2025; Liang et al., 2026; Jiang et al., 2026). Recent work has explored this paradigm for cross-domain and zero-shot forecasting. `xRAG4TS` retrieves relevant temporal patterns from source domains to support cross-city forecasting (Jiang et al., 2026), while `QuiZSF` retrieves structurally related sequences from a large-scale temporal database and incorporates them through multi-grained interaction (Ma et al., 2026). These methods primarily study how to identify relevant historical evidence and how to integrate it into the current prediction.

A common assumption in this literature is that the retrieval repository is *exogenous*: model predictions depend on retrieved memory, but the memory itself is not shaped by those predictions. `ECHO` removes this assumption. We consider continual forecasting in which model-generated trajectories maybe temporarily retained as retrieval candidates when the corresponding future outcomes are not yet available or are only partially observed. These forecasts are not assumed to be preferable to actual observations; rather, they provide provisional temporal evidence until externally observed outcomes become available for subsequent grounding. Consequently, retrieval memory can become partially shaped by the model's own previous outputs, making the forecasting process recursive. The resulting problem is therefore not only retrieval relevance, but also the long-term stability of the retrieval–prediction–memory feedback loop.

A.2 Recursive Feedback and Model-Generated Data

Feedback between model outputs and future learning signals arises in several forms of closed-loop and self-improving systems (Yang et al., 2026b; Bae et al., 2025). For representation learning, `REFINE` couples generative reconstruction with feedback-driven contrastive learning and analyzes the resulting closed-loop optimization dynamics (Yang et al., 2026a). A different line of work studies repeated reuse of model-generated data and shows that recursive training can progressively distort the data distribution and lead to model collapse (Shumailov et al., 2024). Broader studies of self-improving systems similarly emphasize self-confirming feedback and the importance of reliable external signals for preventing recursive degradation (Chen et al., 2026).

`ECHO` differs from these settings in the *state through which feedback operates*. Unlike closed-loop representation learning, we keep forecasting parameters fixed during recursive forecasting. Unlike recursive training, model-generated trajectories do not affect future predictions through parameter updates; instead, they modify the evidence available for subsequent retrieval. The feedback pathway is therefore: prediction $\rightarrow$ memory $\rightarrow$ retrieval $\rightarrow$ prediction. We study the stability of this recursive retrieval process and introduce external grounding as a mechanism for controlling its evolution.

B Experiments setup details

B.1 Datasets

Table 4 summarizes the statistics of the four benchmark datasets used in our experiments.

PhysioNet. PhysioNet (Johnson et al., 2014) contains irregularly sampled physiological measurements collected from intensive-care patients. Following standard preprocessing, the benchmark contains 11,981 samples and 36 variables, with an average of 308.6 observations per sample.

Table 4: Statistics of the IMTS forecasting benchmarks.

Dataset	# Vars.	# Samples	Avg. Obs.	Max Len.
PhysioNet	36	11,981	308.6	47
MIMIC	96	21,250	144.6	96
Human Activity	12	1,359	362.2	131
USHCN	5	1,114	313.5	337

MIMIC. MIMIC (Johnson et al., 2016) is a high-dimensional clinical benchmark derived from ICU records. The processed dataset contains 21,250 samples and 96 variables, with an average of 144.6 observations per sample.

Human Activity. Human Activity (Rubanova et al., 2019) consists of irregularly sampled 3D positional sensor measurements from human motion sequences. The benchmark contains 1,359 samples with 12 variables and an average of 362.2 observations per sample.

USHCN. USHCN (Menne et al., 2015) contains long-term climate measurements collected across U.S. weather stations. The benchmark contains 1,114 samples and 5 variables, with a maximum sequence length of 337.

B.2 BASELINES

We compare the instantiated `ECHO` model against 14 representative IMTS forecasting baselines spanning recurrent, continuous-time, set-based, graph-based, patch-based, and frequency-aware architectures. `ECHO`'s fixed base forecaster is `APN`, the patch-based IMTS predictor of Liu et al. (2026b) described below; `ECHO` uses this same base forecaster throughout all experiments, and the baselines (including `APN` itself) are evaluated as standalone forecasting models rather than as alternative backbones within `ECHO`.

PrimeNet. PrimeNet (Chowdhury et al., 2023) employs self-supervised pretraining to learn time-sensitive representations from irregularly sampled observations.

NeuralFlows. NeuralFlows (Bilos et al., 2021) models continuous-time latent dynamics using invertible neural flows as an efficient alternative to Neural ODEs.

GRU. GRU (Schirmer et al., 2022) is a standard recurrent architecture that serves as a basic sequential modeling baseline.

mTAN. mTAN (Shukla & Marlin, 2021) uses multi-time attention to map irregular observations onto reference points for temporal representation learning.

SeFT. SeFT (Horn et al., 2020) treats irregular observations as sets and employs attention-based set functions to obtain sampling-invariant representations.

GNeuralFlow. GNeuralFlow (Mercatali et al., 2024) combines continuous-time neural flows with graph neural networks to capture temporal evolution and cross-variable dependencies.

GRU-D. GRU-D (Che et al., 2016) augments recurrent modeling with temporal decay and missingness information to handle irregular observation intervals.

Raindrop. Raindrop (Zhang et al., 2022) employs graph-based attention to model interactions among asynchronously sampled variables.

Warpformer. Warpformer (Zhang et al., 2023) introduces temporal warping into Transformer-based modeling to adapt to non-stationary and irregularly sampled time series.

tPatchGNN. tPatchGNN (Zhang et al., 2024) organizes irregular observations into temporal patches and combines patch-level representations with graph neural networks for forecasting.

GraFITi. GraFITi (Yalavarthi et al., 2024) represents irregular multivariate observations using a bipartite graph and performs graph-based information propagation for forecasting.

Hi-Patch. Hi-Patch (Luo et al., 2025) uses hierarchical patching to capture temporal patterns at multiple granularities under heterogeneous sampling.

APN. APN (Liu et al., 2026b) employs adaptive temporal patching and query-based aggregation to construct compact representations from irregular observations. It also serves as the base forecasting backbone of ECHO.

ASTGI. ASTGI (Liu et al., 2026a) directly represents observations as spatio-temporal points and adaptively constructs causal neighborhood graphs with relation-aware propagation to capture dynamic dependencies across time and variables.

B.3 Implementation Details

Retrieval and Fusion. We retrieve $K = 2$ temporal patterns using temperature $\tau = 0.0012$. The effect of varying K is examined in Appendix B.5, where increasing the retrieval size does not yield a monotonic improvement within the evaluated range. The residual correction in Eq. 14 uses a fixed mixing coefficient $\beta = 0.5$ in all experiments.

Retrieval and Reliability Hyperparameters. The global/local similarity mixing weight in Eq. 11 is fixed at $\lambda_p = 0.50$, and the utility mixing weight in Eq. 13 at $\lambda_u = 0.15$; neither is tuned per dataset. The irregularity-specific similarity weight λ_{irr} is set to 0 in all reported experiments, i.e., $s_{tm}^{\text{base}} = s_{tm}^{\text{hist}}$ throughout. For the reliability signals of Eq. 24, the weighting coefficients are fixed and equal, $\eta_1 = \eta_2 = \eta_3 = 1.0$ (not learned; the ablation rows of Table 2 zero out one η_i at a time rather than optimizing them). Concretely, writing $\bar{y}_t$ for the retrieval-weighted mean of the K retrieved futures, d_t for the mean absolute difference between the base prediction $\hat{y}_t$ and $\bar{y}_t$, s_t for the mean absolute magnitude of $\bar{y}_t$ (used as a per-round normalizing scale), and σ_t for the mean (over time and variables) retrieval-weighted standard deviation of the K retrieved futures around $\bar{y}_t$, agreement and consistency are

$$A_t = \exp\Big(-\frac{d_t}{s_t + \epsilon}\Big), \qquad N_t = \exp\Big(-\frac{\sigma_t}{s_t + \epsilon}\Big), \tag{30}$$

with $\epsilon = 10^{-6}$. U_t uses normalized retrieval entropy in all reported experiments (no external predictive uncertainty estimate is supplied).

Training Schedule. Training proceeds without retrieval for the first 10 epochs. Retrieval memory is then constructed from the training split and refreshed every 5 epochs thereafter.

Memory-Evolution Protocol. After the retrieval warm-up period, memory evolution is performed during validation. At each post-warm-up epoch, the complete validation split is processed once using the fixed iteration order of the validation data loader. All validation batches within the same epoch share the same memory-evolution round index, and the aggregated outcomes from that validation pass are used to update the retrieval memory once at the end of the round.

Accordingly, one memory-evolution round corresponds to one post-warm-up validation epoch rather than to an individual forecasting instance. The number of rounds is therefore determined by the number of post-warm-up epochs completed before early stopping and may vary across datasets and random seeds. All memory-evolution regimes use the same validation sequence and start from the same training-derived initial memory $\mathcal{M}_0$.

After model selection and memory evolution are completed, both the forecasting parameters and the resulting retrieval memory are frozen. The held-out test set is then used only for final evaluation, without any further updates to either model parameters or retrieval memory. Thus, the test results reported in Tables 1 and 5 measure forecasting performance under the final memory state produced by

Table 5: Forecasting performance on the held-out test set under different memory-evolution regimes. Each regime produces a final retrieval-memory state during the post-warm-up validation phase; this memory state is then frozen for test evaluation. Static Memory remains fixed after initialization from the training split. For all evolving-memory regimes, the retrieval pool is unbounded and new trajectories are appended without evicting existing memory entries. Partial Grounding uses $\gamma = 0.5$.

Method	Human Activity		USHCN	
	MSE	MAE	MSE	MAE
No Retrieval	0.0420±0.0006	0.1160±0.0022	0.1554±0.0049	0.2646±0.0069
Static Memory	0.0418±0.0005	0.1158±0.0017	0.1550±0.0048	0.2641±0.0067
Recursive Memory ($\gamma = 0$)	0.0418±0.0004	0.1158±0.0015	0.1569±0.0035	0.2674±0.0058
Partial Grounding ($\gamma = 0.5$)	0.0418±0.0005	0.1160±0.0019	0.1548±0.0051	0.2642±0.0066
Grounded Memory ($\gamma = 1$)	0.0419±0.0005	0.1161±0.0018	0.1550±0.0048	0.2642±0.0067
ECHO	**0.0412±0.0003**	**0.1142±0.0017**	**0.1485±0.0043**	**0.2605±0.0061**

each memory-evolution regime. Memory adaptation therefore uses validation outcomes as delayed supervision during the memory-evolution phase.

Memory Utility. Each memory item is assigned a predictive utility score u_m, initialized to 0 when the item enters memory and updated by an exponential moving average with momentum $\mu = 0.9$:

$$u_m \leftarrow \mu\, u_m + (1 - \mu)\,\mathrm{util},$$

where

$$\mathrm{util} = \mathrm{sigmoid}\left(\frac{e^{\mathrm{base}} - e^{\mathrm{ret}}}{\tau}\right)$$

compares the base forecaster's error e^{base} with the retrieval-augmented error e^{ret} on a query for which memory item m is retrieved. Thus, u_m increases when retrieving item m tends to improve forecasting relative to the base prediction.

Memory Configuration. The initial forecasting memory is constructed exclusively from training observations. The default ECHO configuration uses an unbounded retrieval pool. For finite-memory experiments, we use FIFO replacement to isolate the effect of memory capacity from retention-policy design; the corresponding sensitivity results are reported in Table 3. Since u_m measures predictive usefulness, utility-aware retention is a natural extension but is not explored here. The theoretical coefficient α is an abstract measure of the effective influence of newly admitted evidence and does not correspond to a particular FIFO capacity.

Repeated Runs. Unless otherwise stated, each experiment is repeated with 5 random seeds, and we report mean and standard deviation.

B.4 Additional Experiments

Table 5 summarizes held-out test performance under different memory-evolution regimes. We summarize the main observations below.

B.4.1 Recursive Memory Dynamics

We assess empirical stability by the change in forecasting error under repeated memory evolution relative to Static Memory, which serves as the non-recursive reference. Conventional forecasting baselines are compared in predictive accuracy rather than stability because they do not maintain recursive retrieval memory.

Recursive memory shows dataset-dependent drift relative to static retrieval. Table 5 compares Static Memory against Recursive Memory ($\gamma = 0$), which differ only in whether the retrieval

memory remains fixed after initialization from the training split (static) or is continuously augmented with the model's own past predictions during forecasting (recursive). On Human Activity, the two regimes have nearly identical mean errors (MSE 0.0418 ± 0.0005 vs. 0.0418 ± 0.0004; MAE 0.1158 ± 0.0017 vs. 0.1158 ± 0.0015), suggesting no evident degradation from allowing the model's own forecasts to re-enter memory on this dataset. On USHCN, Recursive Memory has higher mean error than Static Memory (MSE 0.1569 ± 0.0035 vs. 0.1550 ± 0.0048; MAE 0.2674 ± 0.0058 vs. 0.2641 ± 0.0067). This pattern is consistent with the error-amplification mechanism described in Section 3.3, although the aggregate results alone do not directly establish the underlying feedback mechanism. This comparison is made at the level of aggregate forecasting performance (Table 5); a finer-grained, round-by-round trajectory analysis of retrieval quality ρ_t, contamination C_t, and pattern diversity D_t would require per-round logging beyond what the current training and evaluation pipeline records, and we leave such analysis to future work.

B.4.2 Effect of Grounding

Grounding regulates the strength of recursive feedback. Table 5 compares recursive memory ($\gamma = 0$), partially grounded memory ($\gamma = 0.5$), and fully grounded memory ($\gamma = 1$). On USHCN, where Recursive Memory has higher mean error than Static Memory (MSE 0.1569 ± 0.0035 vs. 0.1550 ± 0.0048), introducing grounding reduces the mean error: Partial Grounding achieves an MSE of 0.1548 ± 0.0051, while Grounded Memory achieves 0.1550 ± 0.0048. These results are consistent with the theoretical role of grounding in Eq. 20, which attenuates the influence of model-generated evidence in the recursive feedback dynamics. On Human Activity, where Recursive Memory and Static Memory have nearly identical mean performance, Partial Grounding (MSE 0.0418 ± 0.0005) and Grounded Memory (MSE 0.0419 ± 0.0005) likewise produce similar results. Overall, the mean results suggest a larger effect of grounding on USHCN than on Human Activity, while increasing the fixed grounding rate from 0.5 to 1 does not yield a further monotonic improvement.

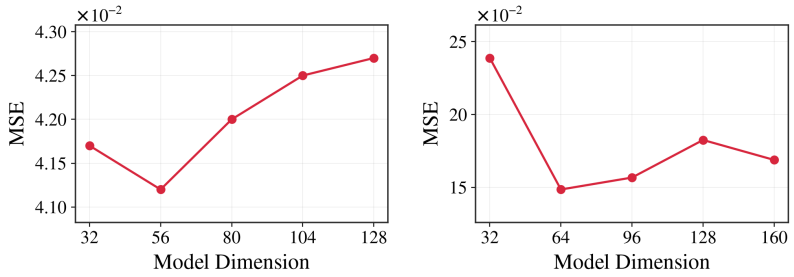
B.4.3 Adaptive Grounding

Adaptive grounding achieves lower mean forecasting error than fixed grounding. Table 2 compares ECHO's adaptive grounding controller against Fixed Grounding ($\gamma = 0.5$) on all four datasets. ECHO achieves lower mean MSE on Human Activity (0.0412 ± 0.0003 vs. 0.0418 ± 0.0004), USHCN (0.1485 ± 0.0043 vs. 0.1548 ± 0.0051), PhysioNet (0.3080 ± 0.0008 vs. 0.3092 ± 0.0006), and MIMIC (0.3904 ± 0.0003 vs. 0.3910 ± 0.0005), with the largest relative gap on USHCN. These results show lower mean forecasting error under the adaptive controller than under the fixed $\gamma = 0.5$ baseline across the four benchmarks. Long-term memory statistics (e.g., contamination C_t) that would more directly explain why adaptive control helps are not currently logged per forecasting round and are left to future work.

B.5 Additional Parameter Sensitivity

We further examine the sensitivity of ECHO to three additional hyperparameters: the shared embedding dimension, the fixed grounding rate used by the *partial/Fixed Grounding* regime, and the retrieval size. The experiments are conducted on Human Activity and USHCN using five seeds, with all other hyperparameters fixed at their reference values. For the embedding-dimension and retrieval-size sweeps, the reference configuration corresponds to the setting used for the main results in Table 1; for the grounding-rate sweep, the reference ($\gamma = 0.5$) corresponds to the *Partial Grounding* row of Table 5.

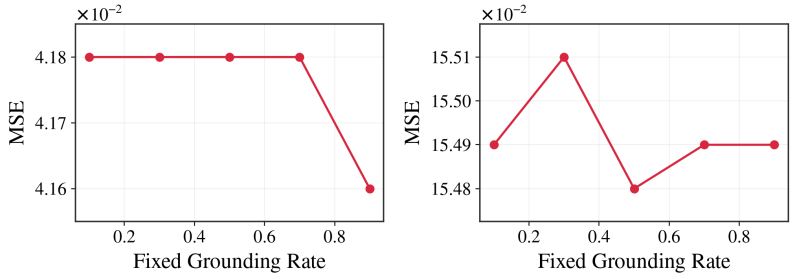
Effect of Model Dimension. Figure 4 examines the shared encoder–retrieval embedding dimension (d_model, with the retrieval dimension tied to it). On Human Activity, the reference dimension achieves the lowest MSE, while both smaller and larger dimensions lead to modest degradation. A similar pattern is observed on USHCN, where reducing the embedding dimension substantially increases both forecasting error and variance across seeds, whereas larger dimensions result in more moderate degradation. These results indicate that ECHO is more sensitive to insufficient representation capacity than to moderate over-parameterization, with the selected dimensions providing a favorable operating point on both datasets.



(a) Human Activity *(b) USHCN*

Figure 4: Sensitivity of forecasting MSE to the shared embedding dimension on Human Activity and USHCN.

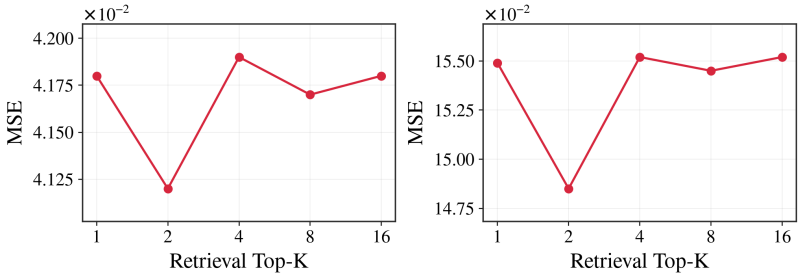
Effect of the Fixed Grounding Rate. Figure 5 varies the fixed grounding rate γ used by the *partial/Fixed Grounding* regime over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. On Human Activity, MSE is essentially flat for $\gamma \leq 0.7$ and decreases slightly at $\gamma = 0.9$; on USHCN, MSE fluctuates within a narrow band (0.1548–0.1551) with no clear trend. Both ranges are well within the run-to-run seed variation observed elsewhere in this sweep, indicating that, within the *partial* regime, forecasting accuracy is not sensitive to the specific value of the fixed grounding rate on either dataset.



(a) Human Activity *(b) USHCN*

Figure 5: Sensitivity of forecasting MSE to the fixed grounding rate γ (*partial* regime) on Human Activity and USHCN.

Effect of Retrieval Size. Figure 6 varies the number of retrieved neighbors (`rift_topk`) around the reference setting. Across the evaluated range, off-reference configurations exhibit relatively small and non-monotonic variations in forecasting error, while the reference setting achieves the lowest MSE on both datasets. This suggests that ECHO is generally robust to moderate changes in retrieval size and does not rely on retrieving a large number of historical trajectories to obtain its forecasting gains.



(a) Human Activity *(b) USHCN*

Figure 6: Sensitivity of forecasting MSE to retrieval size on Human Activity and USHCN.

C Theoretical Analysis

This section provides the formal analysis of the recursive memory dynamics introduced in Section 3.3. We first characterize local stability under recursive retrieval, then analyze the effect of grounding, and finally derive a sufficient condition for global contraction of the induced reliability dynamics.

C.1 Stability of Recursive Retrieval

Recall the recursive reliability dynamics

$$p_{t+1} = (1-\alpha)p_t + \alpha f(p_t), \tag{31}$$

where $p_t \in [0,1]$ denotes aggregate memory reliability, $f(p_t)$ the expected reliability of newly generated evidence, and $\alpha \in (0,1]$ the effective influence of newly admitted evidence.

Define

$$g(p) = (1-\alpha)p + \alpha f(p). \tag{32}$$

For $\alpha > 0$, an equilibrium p^* satisfies

$$p^* = f(p^*). \tag{33}$$

Proposition C.1 (Local Stability of Recursive Retrieval)**.** *Suppose f is continuously differentiable in a neighborhood of an equilibrium p^*. Then p^* is locally asymptotically stable if*

$$|1 - \alpha + \alpha f'(p^*)| < 1, \tag{34}$$

and locally unstable if the magnitude is greater than one.

Proof. Let $\delta_t = p_t - p^*$. A first-order expansion around p^* gives

$$f(p_t) = f(p^*) + f'(p^*)\delta_t + o(|\delta_t|). \tag{35}$$

Using $p^* = f(p^*)$ in Eq. 31,

$$\delta_{t+1} = [1 - \alpha + \alpha f'(p^*)]\,\delta_t + o(|\delta_t|). \tag{36}$$

Thus sufficiently small perturbations contract when the magnitude of the linear coefficient is below one and expand when it exceeds one. □

The derivative $f'(p^*)$ measures the sensitivity of newly generated evidence to the current memory state. Hence, recursive retrieval is locally self-correcting when the induced feedback is contractive and can amplify perturbations when it is expansive. The parameter α affects the propagation of perturbations but not the equilibrium condition $p^* = f(p^*)$.

C.2 Effect of Grounding

With grounding rate $\gamma \in [0,1]$ and grounded-evidence reliability $q \in [0,1]$, the reliability dynamics become

$$p_{t+1} = (1-\alpha)p_t + \alpha\left[(1-\gamma)f(p_t) + \gamma q\right]. \tag{37}$$

Define the corresponding update map

$$g_\gamma(p) = (1-\alpha)p + \alpha\left[(1-\gamma)f(p) + \gamma q\right]. \tag{38}$$

An equilibrium p^*_γ satisfies

$$p^*_\gamma = (1-\gamma)f(p^*_\gamma) + \gamma q. \tag{39}$$

Proposition C.2 (Local Stability under Grounding)**.** *Suppose f is continuously differentiable near p^*_γ. The grounded equilibrium is locally asymptotically stable if*

$$\left|1 - \alpha + \alpha(1-\gamma)f'(p^*_\gamma)\right| < 1. \tag{40}$$

Proof. Differentiating the grounded update map gives

$$g'_\gamma(p) = 1 - \alpha + \alpha(1-\gamma)f'(p). \tag{41}$$

The result follows from the standard local stability condition $|g'_\gamma(p^*_\gamma)| < 1$. □

Grounding attenuates the endogenous feedback contribution by the factor $(1-\gamma)$. It may also shift the equilibrium through Eq. 39, so its nonlinear effect depends on both feedback attenuation and the resulting equilibrium location.

C.3 Sufficient Grounding Condition

Assumption C.1 (Bounded Recursive Sensitivity). *The function $f : [0,1] \to [0,1]$ is L-Lipschitz:*

$$|f(p) - f(p')| \le L|p - p'|, \qquad \forall p, p' \in [0,1]. \tag{42}$$

Theorem C.1 (Sufficient Condition for Grounded Contraction). *Under Assumption C.1, g_γ is a contraction on $[0,1]$ if*

$$(1-\gamma)L < 1. \tag{43}$$

Under this condition, the grounded dynamics admit a unique fixed point and converge to it from any initialization $p_0 \in [0,1]$.

Proof. For any $p, p' \in [0,1]$,

$$|g_\gamma(p) - g_\gamma(p')| \le (1-\alpha)|p - p'| + \alpha(1-\gamma)|f(p) - f(p')| \tag{44}$$

$$\le [1 - \alpha + \alpha(1-\gamma)L]\,|p - p'|. \tag{45}$$

Since $\alpha > 0$, the coefficient is strictly below one whenever $(1-\gamma)L < 1$. Because g_γ maps $[0,1]$ into itself, the contraction mapping theorem guarantees a unique fixed point and convergence from any $p_0 \in [0,1]$. □

Corollary C.1 (Sufficient Grounding Threshold). *If $L > 1$, a sufficient condition for contraction is*

$$\gamma > 1 - \frac{1}{L}. \tag{46}$$

When $L < 1$, fully recursive memory already satisfies this sufficient contraction bound. When $L = 1$, any strictly positive grounding rate makes the bound contractive. Thus stronger recursive sensitivity requires stronger grounding to obtain the same sufficient guarantee.

Scope of the Analysis. The scalar state p_t summarizes the reliability of a high-dimensional retrieval memory rather than the full neural forecasting state. Accordingly, the results characterize the induced memory-reliability dynamics and provide sufficient stability guarantees under the stated assumptions; they do not claim global convergence of the complete forecasting system.