SEQNET: SUPPORT-AWARE IRREGULAR FORECASTING WITH EVENT MEMORY AND QUERY ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Irregular observations are common in healthcare, activity monitoring, and climate analysis. Forecasting from such data requires compressing histories with varying lengths and irregular time intervals into fixed-length representations. However, although historical compression can capture overall temporal trends, it may overlook a small number of locally predictive observations, such as abnormal values and abrupt changes, making these critical local signals difficult to fully exploit. Moreover, when the target variable has insufficient historical observations and information from other variables is needed, different forecasting targets may require different historical evidence and information sources, making fixed information-selection strategies difficult to adapt to changing observation conditions.We propose **SEQNet**, a forecasting model that preserves both global historical trends and prediction-relevant local evidence. Specifically, SEQNet summarizes historical observations to learn variable-level temporal patterns while constructing a candidate set of local events and dynamically retrieving observations that are relevant to the current forecasting query. When the target variable is insufficiently observed, SEQNet further adapts to its current observation support by assigning different weights to the target variable's own history, information from other variables, and general patterns learned from data.Experiments on PhysioNet, Activity, USHCN, and MIMIC-III show that SEQNet achieves the lowest MSE among the compared methods on all four datasets. On PhysioNet, SEQNet reduces MSE by **24.49%** relative to Hi-Patch, the strongest baseline.

1 INTRODUCTION

Irregular multivariate time-series forecasting is important in healthcare, human activity monitoring, and climate analysis, where the central challenge is to make full use of historical information from irregular observations (Johnson et al., 2016; Menne et al., 2009; Silva et al., 2012). To address this challenge, researchers have developed a wide range of forecasting models. Recent approaches organize historical observations into temporal patches, hypergraphs, or recursive temporal representations (Li et al., 2025; 2026a; Zhang et al., 2024). For historical aggregation, mTAN uses time attention to map observation sequences of different lengths into fixed-length representations (Shukla & Marlin, 2021). Despite the strong results achieved by existing methods, we believe that two aspects of retaining and using historical information deserve further study.

❶ *Compressing history may weaken local observations.* A common approach is to use attention-based aggregation or temporal patch encoding to convert histories with different numbers of observations and irregular time intervals into compact representations. Specifically, mTAN uses time embeddings and attention to map variable-length observations into fixed-length representations, while t-PatchGNN organizes irregular observations into temporal patches and models local temporal features together with relationships among variables (Shukla & Marlin, 2021; Zhang et al., 2024). However, when many observations are compressed together, a model may preserve the overall trend while losing local characteristics, such as how large a sudden change was and when it occurred. For example, in the precipitation records of USHCN, the overall evolution of precipitation and a small number of prominent local peaks may provide different predictive value (Menne et al., 2009). The overall trend reflects changes over a longer time range, whereas the magnitude and timing of a local peak, together with its difference from neighboring observations, may provide additional predictive

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.

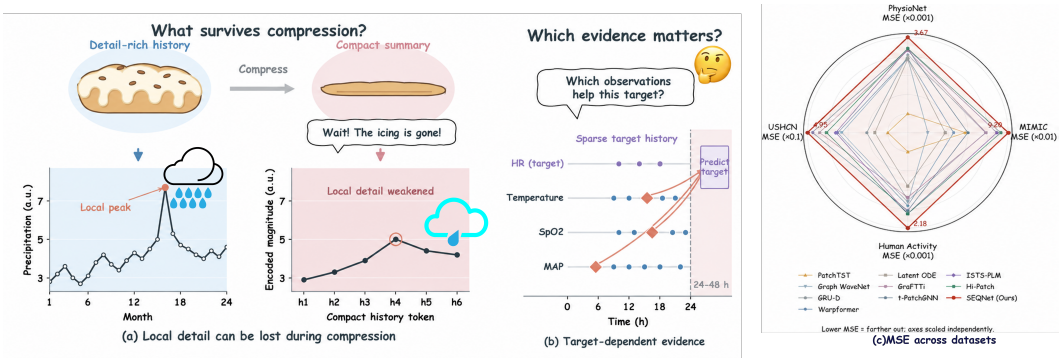


Figure 1: Motivation and forecasting results of SEQNet. (a) A detail-rich precipitation history contains a local peak that becomes less explicit after compression into a small set of history tokens; the bread illustration depicts the loss of local detail. (b) When heart-rate (HR) history is sparse, observations of temperature, oxygen saturation (SpO_2), and mean arterial pressure (MAP) can provide target-dependent evidence. The dashed line separates the 0–24 h history from the 24–48 h prediction window; orange markers highlight candidate observations and arrows indicate possible information paths. (c) MSE comparison across PhysioNet, MIMIC-III, Human Activity, and USHCN. The four radial axes use independent scales, with lower MSE farther from the center.

cues. Figure 1(a) illustrates both views of the same history: the highlighted precipitation peak contains local detail beyond the overall pattern, while aggregation into a few history tokens preserves the broad trajectory but makes the event's magnitude and timing less explicit. Consequently, once history is compressed, the model may have difficulty making full use of important local information, which in turn makes predictions that depend on such information more difficult.

❷ *How can we use local observations that are valuable for the target prediction?* A locally salient observation is not necessarily equally informative for every forecasting target. Retaining local observations is useful only when those observations provide effective evidence for the current prediction. If the target variable has sufficient historical records, the model can look for useful local observations within its own history. In irregular data, however, the target variable may have only a few records, may not have been observed for a long time, or may even be completely unobserved, making information from other variables necessary.For example, in MIMIC-III clinical records, a target measurement may have no recent observation while other variables are still observed. In this case, the model needs to determine which observations from other variables can provide useful evidence for the current prediction (Johnson et al., 2016). Figure 1(b) illustrates this situation using sparse heart-rate records together with observations of temperature, oxygen saturation, and mean arterial pressure. The arrows indicate different information sources that may be used for the same forecasting target, raising a further question: which local observations are truly relevant to the current prediction?When the target variable's own history is insufficient, borrowing information from other variables should not simply amount to using all available observations. Instead, the model needs to distinguish observations that provide useful clues about the target's future from observations that may be locally salient but are less relevant to the current prediction. Therefore, the model must not only preserve local information from historical observations, but also determine, according to the specific forecasting target and currently available observations, which local observations from the target itself or from other variables should receive greater attention.

To address these limitations, we propose **SEQNet**, a **S**upport-aware **E**vent-memory and **Q**uery-routing **Net**work for irregular multivariate forecasting. Our approach contains two key components: Soft Temporal Memory and Retained Events, and Query-Conditioned Support-Aware Routing.The first component preserves both overall temporal trends and important local information. Soft Temporal Memory maps irregular histories with different lengths and sampling intervals to shared relative-time positions, forming compact historical representations that capture overall temporal changes and relationships among variables. In parallel, Retained Events constructs a candidate set of local observations according to recency, deviation from the variable's typical value, and local changes between consecutive visible observations. These local details are preserved separately from the compressed historical representation and remain directly accessible during prediction.The second component determines which local observations and information sources are most useful for the

current forecasting target. For each forecasting query, the model first uses the target variable, prediction time, and current observation condition to dynamically retrieve the local observations that are most relevant to that prediction from the retained event candidates. Support-Aware Routing then uses the target variable's current observation support to adaptively balance three information sources: *self*, representing the target variable's own history; *cross*, representing information provided by other variables; and *prior*, representing general patterns learned from the training data. When the target variable has sufficient historical observations, the model can rely more strongly on its own history; when its observations are sparse, stale, or completely absent, the model can place greater reliance on cross-variable information and learned priors. In this way, SEQNet first preserves both global trends and local details, and then adapts the use of this information to the specific forecasting target and its current observation condition. Figure 1(c) summarizes the four-dataset MSE comparison, with independently scaled axes to show the relative ordering on each task.

Our contributions are fourfold:

- ❶ We propose SEQNet, which preserves both global historical trends and local observations, and adapts their use to each forecasting target.
- ❷ We design Soft Temporal Memory and Retained Events to summarize irregular histories while keeping candidate local observations separately available for prediction.
- ❸ We introduce Query-Conditioned Support-Aware Routing to retrieve local observations relevant to each forecasting target and adaptively balance information from the target's own history, other variables, and learned priors according to the available observation support.
- ❹ We evaluate SEQNet on PhysioNet, Activity, USHCN, and MIMIC-III, with ablations on USHCN to examine the effects of local-event retention and query-conditioned support-aware routing.

2 Related Work

In recent years, deep learning methods have been widely applied to irregular multivariate time series forecasting in domains such as healthcare, human activity analysis, and environmental monitoring (Johnson et al., 2016; Silva et al., 2012). Among these tasks, irregular multivariate time series forecasting (IMTSF) has attracted increasing attention due to its fundamental importance in both academic research and practical applications (Zhang et al., 2024; Li et al., 2025). According to their research objectives, existing IMTSF methods mainly focus on two directions: historical aggregation and compression, and adaptive information selection under changing observation support.

Historical Aggregation and Compression. Early methods mainly model irregular sampling processes through recurrent structures and continuous-time dynamics. GRU-D incorporates observation masks and time intervals into recurrent networks through decay mechanisms, enabling hidden states to adapt to missing observations (Che et al., 2018). Latent ODE further models latent states as continuous-time dynamics through ordinary differential equations, allowing the model to handle observations with irregular time intervals (Rubanova et al., 2019). Attention mechanisms are then introduced to improve the modeling of importance among observations at different timestamps. Multi-Time Attention Networks map irregular observations with varying numbers of measurements into fixed-length representations through time embeddings and attention mechanisms (Shukla & Marlin, 2021). Recent studies further incorporate patch and graph structures to enhance irregular history representations. t-PatchGNN organizes irregular observations into temporal patches and employs graph neural networks to capture local temporal patterns and relationships among variables (Zhang et al., 2024). HyperIMTS utilizes hypergraph representations to model high-order dependencies among variables (Li et al., 2025). ReIMTS learns recursive multi-scale representations while preserving original timestamps (Li et al., 2026a). In addition, APN investigates adaptive aggregation strategies for irregular observations, while TFMixer jointly models temporal and frequency information for irregular time series (Liu et al., 2026b; Qiu et al., 2026). GraFITi represents observations as edges of a sparse bipartite graph and forecasts missing edge values (Yalavarthi et al., 2024). Hi-Patch builds hierarchical patch graphs to extract dependencies at different temporal scales (Luo et al., 2025), while ASTGI represents individual observations as points and uses adaptive graph neighborhoods for propagation and query-based prediction (Liu et al., 2026a). ISTS-PLM adapts pretrained language models through time-aware and variable-aware representations (Zhang et al., 2025). Across these

approaches, recurrent states, attention-based aggregation, and temporal patches provide different ways to summarize irregular histories. However, compressing observations into shared summaries can preserve broad temporal trends while obscuring the magnitude and timing of individual changes. This loss of local detail may limit forecasts of rare or abrupt events whose predictive cues are concentrated in a few observations.

Adaptive Information Selection under Changing Observation Support. Beyond how historical information is represented, determining which historical observations and information sources should be used for the current forecasting target is also important. In real-world data, observation support often varies across samples, variables, and time. For example, in clinical monitoring, different variables may have very different sampling frequencies, while some target variables may remain unobserved for long periods or even be completely missing. Therefore, models may need to obtain complementary information from other variables and general patterns learned from data. Existing methods have considered the use of target-related information. GraFITi establishes relationships among observations through a graph structure and predicts values at specified missing positions (Yalavarthi et al., 2024), while ASTGI incorporates queries into graph neighborhood propagation and prediction (Liu et al., 2026a). Group Distributionally Robust Optimization increases the training weights of difficult environments to optimize worst-case performance across environments (Sagawa et al., 2019). In forecasting with exogenous variables, GCGNet uses graph consistency to guide a generative predictor and studies missing-value perturbations of exogenous inputs (Li et al., 2026b). Different from these works, we combine observation-deletion training with query-conditioned support-aware routing, enabling the model to learn from incomplete histories and select relevant local observations and information sources for each specific forecasting target.

3 Methodology

For sample b, let $\mathcal{H}_b = \{(x_{bj}, t_{bj}, c_{bj}, m_{bj})\}_{j=1}^{N_b}$ denote a packed historical record over V variables and interval $[t_b^s, t_b^e]$. Here $x_{bj} \in \mathbb{R}$ is a measurement, t_{bj} its time, $c_{bj} \in \{1, \dots, V\}$ its variable, and m_{bj} indicates a visible, non-padding event. The query set $\mathcal{Q}_b = \{(\tau_{bi}, c_{bi}^q)\}_{i=1}^{Q_b}$ specifies prediction times and variables, with optional observed static covariates. We learn $\widehat{y}_{bi} = f_\theta(\mathcal{H}_b, \mathcal{Q}_b)_i$. Future values y_{bi}, validity masks o_{bi}, and nonnegative target weights a_{bi} enter supervision and evaluation only; $a_{bi} = 0$ at invalid targets. An observation view changes m_{bj} while retaining the same queries and target values.

3.1 Structure Overview

We develop SEQNet to preserve both temporal trends and local observations and to retrieve the information relevant to each forecast query. Figure 2 presents the prediction path and its four functional stages. Given an irregular history and a set of future queries, the model first encodes visible measurements together with their observation support. It then constructs a dual history memory containing compressed global patterns and separately retained local events. For each forecast query, support-aware routing reads the available source memories and combines self, cross-variable, and prior information before producing the prediction. During training, original and observation-deleted histories share the same queries and targets and provide complementary supervision to the shared forecasting network.

The forecasting network consists of two complementary components: Soft Temporal Memory and Retained Events, and Support-Aware Query Routing. First, an event encoder maps each visible measurement to a token, while support descriptors summarize when and how densely each variable was observed. Soft Temporal Memory aggregates these tokens at shared relative-time anchors; a parallel path retains selected real observations. Two interaction blocks update the soft slots across time and variables, while the retained observations remain available through the parallel path. Next, each forecast query reads the memory of each source variable and forms self, cross-variable, and prior representations. A support-conditioned router combines these representations before the output head produces the requested forecast. During training, original and deleted histories share the same targets; event supervision and constrained environment weighting provide additional training signals. At inference, the forecasting network processes the observed history in one forward pass. Algorithms 1 and 2 give the full procedures.

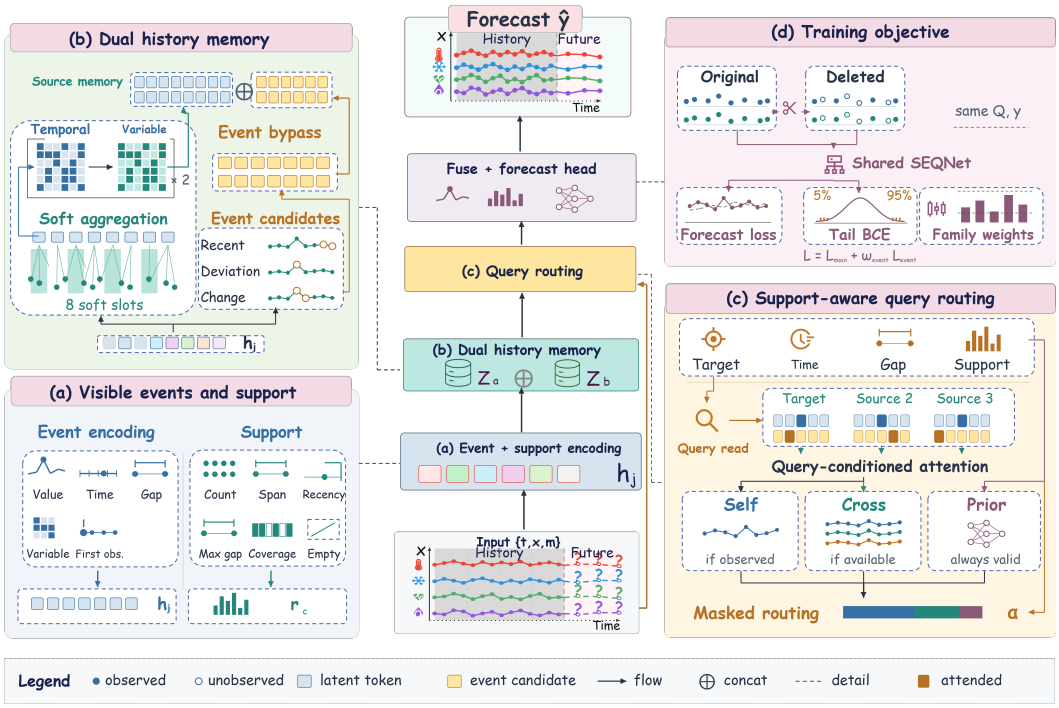


Figure 2: SEQNet architecture and training. The central path shows the complete flow from an irregular history and future queries to the forecast. (a) Visible events are encoded from their values, times, gaps, and variable identities, while count, span, recency, maximum gap, coverage, and empty-history status form variable-level support descriptors. (b) Soft temporal aggregation produces compact global memories Z_a, while selected real observations form an event bypass Z_b; only the soft slots enter temporal–variable interactions, and both paths form each source memory. (c) Each query uses its target, prediction time, forecast gap, and current support to read the source memories and route among self, cross-variable, and prior information before fusion and forecasting. (d) Original and observation-deleted histories share the same queries, targets, and SEQNet parameters. Training combines forecasting loss, tail-event supervision, and constrained observation-family weighting.

3.2 Visible Events and Observation Support

The event encoder represents both a measurement and its sampling context. Given the visible history $\mathcal{H}_b$, we first compute its relative timestamp and the elapsed physical time since the preceding visible measurement of the same variable:

$$u_{bj} = \frac{t_{bj} - t_b^s}{\max(t_b^e - t_b^s, \epsilon)}, \qquad g_{bj} = \log\left(1 + \frac{\Delta t_{bj}}{\tau_0}\right). \tag{1}$$

Here τ_0 is the dataset time unit, and Δt_{bj} is the elapsed time since the preceding visible event of the same variable; it is zero for the first event. An MLP combines the value, a learnable time embedding, the gap, and a variable embedding, with a learned first-observation offset, to produce $h_{bj} \in \mathbb{R}^d$. Hidden and padded payloads are masked before the learned transformations and the resulting tokens are zeroed.

Event values alone do not indicate whether a variable has been densely observed or has a long gap before the forecast. We therefore additionally compute $r_{bv} \in \mathbb{R}^6$ from its visible timestamps. Its components are log observation count, normalized span, recency, maximum gap, occupied-bin coverage, and an empty-variable indicator; Appendix B gives the boundary conventions. These descriptors condition memory construction, variable attention, and query routing. Thus two histories with similar measured values can still lead to different representations when their sampling support differs.

3.3 Soft Temporal Memory and Retained Events

This module converts variable-length event sequences into aligned summaries that can exchange information across variables. Specifically, each variable receives K slots centered at fixed relative-time anchors $a_k = (k+1/2)/K$, for $k = 0, \dots, K-1$. For one attention head, the logit from visible event j of variable v to slot k is

$$\ell_{bvkj} = \frac{q_{bvk}^\top k_{bj}}{\sqrt{d_h}} - \frac{(u_{bj} - a_k)^2}{2\sigma_{bvk}^2}, \qquad z_{bvk} = \mathrm{Project}\left(\sum_{j \in \mathcal{I}_{bv}} \mathrm{softmax}_j(\ell_{bvkj})\, v_{bj} \right), \tag{2}$$

where $\mathcal{I}_{bv}$ contains visible events of variable v, d_h is the head dimension, and k_{bj}, v_{bj} are learned token projections. Slot queries depend on slot identity, the variable summary, and support. A bounded learned bandwidth $\sigma_{bvk} \in [0.02, 1]$ depends on support, the variable summary, and slot identity, separately for each head. Head outputs are concatenated, projected, and normalized to form $Z \in \mathbb{R}^{B \times V \times K \times d}$. Variables with no visible events have zero memories and an invalid-source flag.

Aggregation gives a compact representation. A parallel memory keeps selected input observations directly accessible to the decoder. We construct a candidate-event memory containing up to R real observations per variable. The default allocation retains two recent events, three large absolute deviations from a training-set median, and three large changes from the previous visible event, excluding already selected positions. Deviation scores use a frozen training-set scale, and change scores also account for elapsed physical time (Appendix B). The selected tokens retain their encoded values and times and receive selection-type embeddings. They bypass the interaction stack and are concatenated with the updated slots at decoding. For each forecasting query, attention retrieves relevant information from the temporal summaries and candidate events.

3.4 Factorized Interaction and Source-Wise Queries

The soft memories provide shared temporal positions at which variables can exchange information. Each of two blocks applies attention across the K slots within a variable, followed by attention across variables at each anchor. Each attention sublayer uses residual connections, normalization, and a feed-forward update; temporal logits include a learned relative-slot bias, and variable queries and keys include support. The resulting interaction cost for attention scores is $O(B(VK^2 + KV^2)d)$ per block, compared with $O(BV^2K^2d)$ for dense attention over all slots. This statement concerns the interaction stack; encoder and decoder costs are given in Appendix D.

The decoder uses both the updated soft slots and retained observations. To keep their source identity explicit, query i first encodes its relative time, forecast gap, target-variable identity, and target support into h_{bi}^q. Let M_{bv} concatenate the updated slots and valid retained events of source variable v. The query reads each source independently:

$$s_{biv} = \mathrm{Attn}(h_{bi}^q, M_{bv}, M_{bv}), \quad h_{bi}^{\mathrm{self}} = s_{bi,c_{bi}^q}, \quad h_{bi}^{\mathrm{cross}} = \sum_{v \in \mathcal{S}_{bi}} \beta_{biv} s_{biv}, \tag{3}$$

where $\mathcal{S}_{bi}$ contains nonempty source variables other than the target and $\beta_{biv} = \mathrm{softmax}_{v \in \mathcal{S}_{bi}}((h_{bi}^q)^\top s_{biv}/\sqrt{d})$. A prior MLP forms h_{bi}^{prior} from the query representation and available static features. The prior uses support information from the query.

A target with little recent history may need a different balance of information sources from a well-observed target. To learn this balance, a router maps the query, target support, and the three route representations to masked softmax weights α_{bi}:

$$h_{bi}^{\mathrm{mix}} = \sum_{r \in \{\mathrm{self}, \mathrm{cross}, \mathrm{prior}\}} \alpha_{bi}^r h_{bi}^r, \qquad \widehat{y}_{bi} = \mathrm{Head}\big(\mathrm{Fuse}(h_{bi}^{\mathrm{mix}}, \ell_{bi}, \eta_{bi}, \mathrm{age}_{bi}, s_b)\big). \tag{4}$$

Here ℓ_{bi} is the latest visible target-variable value (zero if absent), η_{bi} indicates its availability, and s_b denotes optional static features. The self route requires target-variable history, the cross route requires another valid variable, and the prior is always valid. The continuous head predicts a mean and auxiliary lower/upper-tail logits. For the short-horizon Activity task, the mean additionally uses ℓ_{bi} as a residual anchor; the other three tasks disable this residual. An optional nonnegative-event head is described in Appendix B.4.

3.5 Paired Views and Constrained Environment Weighting

The training objective should expose the predictor to different observation conditions without changing the requested forecast. For this purpose, each sample supplies its original history $\mathcal{V}_b^0$ and one difficult view $\mathcal{V}_b^h$. We sample uniformly from five deletion families: random points, channel subsets, contiguous time blocks, early observations, and late observations. Retention is sampled from $\{0.8, 0.6, 0.4, 0.2\}$, and all query targets are shared between the two views. The masks are reproducible from dataset, sample identity, family, retention, and replica identifiers.

Let p_b/q_b correct the actual sample probability q_b to the target probability p_b, and let $p_g = 1/5$ and q_g denote base and adversarial family probabilities. Writing $\ell_b^v = \sum_i a_{bi}(\widehat{y}_{bi}^v - y_{bi})^2$, the main loss is

$$\mathcal{L}_{\text{main}} = \frac{1}{B}\sum_b \frac{p_b}{q_b}\left[\frac{1}{2}\ell_b^0 + \frac{1}{2}\frac{q_{g_b}}{p_{g_b}}\ell_b^h\right]. \tag{5}$$

Target weights specify the supervised measure; reported benchmark reductions are defined separately in Appendix E.3. Family sampling remains uniform, while the loss ratio q_g/p_g changes the emphasis on difficult views. Following the group-robust learning principle of emphasizing difficult groups (Sagawa et al., 2019), we optimize family weights under explicit box and divergence constraints. After a ten-epoch uniform warm-up, smoothed observed family losses $\overline{L}_g$ determine

$$\max_q \sum_g q_g \overline{L}_g \quad \text{subject to} \quad \sum_g q_g = 1, \quad \tfrac{1}{2}p_g \le q_g \le 2p_g, \quad D_{\text{KL}}(q\|p) \le \delta. \tag{6}$$

The default maximum KL radius is 0.05. The box bounds preserve coverage of every family while limiting concentration on any one family. Appendix C details the solver and schedule.

3.6 Event Supervision and Training Objective

To supervise event-related outputs, $\mathcal{L}_{\text{event}}$ applies binary cross-entropy to lower- and upper-tail labels defined by training-set 5th and 95th percentiles, on both views. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \omega_{\text{event}}\mathcal{L}_{\text{event}}. \tag{7}$$

Every ten optimizer steps, a gradient controller adjusts ω_{event} using gradient norms at the shared fusion layer and an auxiliary gradient budget of 0.1. Appendix C gives the coefficient update. At inference, one forward pass processes the observed history and queries through Equations (1)–(4).

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on four widely used real-world irregular time-series benchmarks spanning activity monitoring, clinical forecasting, and climate forecasting: Activity, PhysioNet 2012, USHCN, and MIMIC-III (Silva et al., 2012; Menne et al., 2009; Johnson et al., 2016). Table 1 summarizes their statistics and forecasting protocols. Activity and PhysioNet 2012 use training-only channel-wise min–max scaling, whereas USHCN retains its processed raw coordinates. For MIMIC-III, we use a fixed 0–24 h history to forecast the 24–48 h window under a 60/20/20 instance split. Further dataset details are provided in Appendix E.

Table 1: Statistics and forecasting protocols of the datasets. Counts refer to forecasting samples.

Dataset	Variables	Train	Validation	Test	History → target
Activity	12	6,443	2,148	2,148	3 seconds → 1 second
PhysioNet 2012	41	7,187	2,396	2,392	24 hours → next 24 hours
USHCN	5	15,715	5,224	5,259	24 months → next month
MIMIC-III	96	14,073	4,692	4,692	0–24 h → 24–48 h

Settings. All experiments use a fixed architecture with hidden dimension 64, four attention heads, eight soft-memory slots, at most eight retained events per variable, and two interaction blocks.

We train with AdamW for at most 50 epochs, select checkpoints by clean validation MSE, and report results over five random seeds. MSE and MAE use the dataset-specific reductions defined in Appendix E.3; further optimization and dataset settings are detailed in Appendix E.

Baselines. We compare SEQNet with recurrent, continuous-time, Transformer, graph-based, and recent irregular-series forecasting methods. Table 2 reports the complete baseline comparison.

Table 2: Mean Absolute Error (MAE) and Mean Squared Error (MSE) on four real-world irregular time-series datasets. Lower values are better. The best results are marked in red, and the second-best results in blue. Metric scales are shown beneath each dataset name.

Method	PhysioNet		MIMIC		Human Activity		USHCN	
	MSE $\times 10^{-3}$	MAE $\times 10^{-2}$	MSE $\times 10^{-2}$	MAE $\times 10^{-2}$	MSE $\times 10^{-3}$	MAE $\times 10^{-2}$	MSE $\times 10^{-1}$	MAE $\times 10^{-1}$
DLinear	41.86 $\pm$ 0.05	15.52 $\pm$ 0.03	14.93 $\pm$ 0.09	44.76 $\pm$ 0.19	4.03 $\pm$ 0.01	4.21 $\pm$ 0.01	6.21 $\pm$ 0.00	3.88 $\pm$ 0.02
TimesNet	16.48 $\pm$ 0.11	6.14 $\pm$ 0.03	14.71 $\pm$ 0.12	34.63 $\pm$ 0.22	3.12 $\pm$ 0.01	3.56 $\pm$ 0.02	5.58 $\pm$ 0.05	3.60 $\pm$ 0.04
PatchTST	12.00 $\pm$ 0.23	6.02 $\pm$ 0.14	11.30 $\pm$ 0.30	34.77 $\pm$ 0.11	4.29 $\pm$ 0.14	4.80 $\pm$ 0.09	5.75 $\pm$ 0.01	3.57 $\pm$ 0.02
Crossformer	6.66 $\pm$ 0.11	4.81 $\pm$ 0.11	12.36 $\pm$ 0.10	34.96 $\pm$ 0.09	4.29 $\pm$ 0.20	4.89 $\pm$ 0.17	5.25 $\pm$ 0.04	3.27 $\pm$ 0.09
Graph WaveNet	6.04 $\pm$ 0.28	4.41 $\pm$ 0.11	12.76 $\pm$ 0.33	34.67 $\pm$ 0.14	2.89 $\pm$ 0.03	3.40 $\pm$ 0.05	5.29 $\pm$ 0.04	3.16 $\pm$ 0.05
MTGNN	6.26 $\pm$ 0.18	4.46 $\pm$ 0.07	13.04 $\pm$ 0.14	33.96 $\pm$ 0.19	3.03 $\pm$ 0.03	3.53 $\pm$ 0.03	5.39 $\pm$ 0.05	3.34 $\pm$ 0.02
StemGNN	6.86 $\pm$ 0.28	4.76 $\pm$ 0.19	13.78 $\pm$ 0.33	34.72 $\pm$ 0.21	8.81 $\pm$ 0.37	6.90 $\pm$ 0.02	5.75 $\pm$ 0.09	3.40 $\pm$ 0.09
CrossGNN	7.22 $\pm$ 0.36	4.96 $\pm$ 0.12	14.69 $\pm$ 0.12	37.16 $\pm$ 0.24	3.03 $\pm$ 0.10	3.48 $\pm$ 0.08	5.66 $\pm$ 0.04	3.53 $\pm$ 0.05
FourierGNN	6.84 $\pm$ 0.35	4.65 $\pm$ 0.08	14.21 $\pm$ 0.18	38.70 $\pm$ 0.17	2.99 $\pm$ 0.02	3.42 $\pm$ 0.02	5.82 $\pm$ 0.06	3.62 $\pm$ 0.07
GRU-D	5.59 $\pm$ 0.09	4.08 $\pm$ 0.05	10.52 $\pm$ 0.17	28.69 $\pm$ 0.03	2.94 $\pm$ 0.05	3.53 $\pm$ 0.06	5.54 $\pm$ 0.38	3.40 $\pm$ 0.28
SeFT	9.22 $\pm$ 0.18	5.40 $\pm$ 0.08	10.40 $\pm$ 0.22	27.39 $\pm$ 0.07	12.20 $\pm$ 0.17	8.43 $\pm$ 0.07	5.80 $\pm$ 0.19	3.70 $\pm$ 0.11
RainDrop	9.82 $\pm$ 0.08	5.57 $\pm$ 0.06	13.69 $\pm$ 0.34	31.79 $\pm$ 0.11	14.92 $\pm$ 0.14	9.45 $\pm$ 0.05	5.78 $\pm$ 0.22	3.67 $\pm$ 0.17
Warpformer	5.94 $\pm$ 0.35	4.21 $\pm$ 0.12	11.76 $\pm$ 0.13	34.15 $\pm$ 0.07	2.79 $\pm$ 0.04	3.39 $\pm$ 0.03	5.25 $\pm$ 0.05	3.23 $\pm$ 0.05
mTAN	6.23 $\pm$ 0.24	4.51 $\pm$ 0.17	11.10 $\pm$ 0.09	31.09 $\pm$ 0.12	3.22 $\pm$ 0.07	3.81 $\pm$ 0.07	5.33 $\pm$ 0.05	3.26 $\pm$ 0.10
Latent ODE	6.05 $\pm$ 0.57	4.23 $\pm$ 0.26	12.08 $\pm$ 0.12	32.13 $\pm$ 0.13	3.34 $\pm$ 0.11	3.94 $\pm$ 0.12	5.62 $\pm$ 0.03	3.60 $\pm$ 0.12
CRU	8.56 $\pm$ 0.26	5.16 $\pm$ 0.08	11.07 $\pm$ 0.08	33.12 $\pm$ 0.11	6.97 $\pm$ 0.78	6.30 $\pm$ 0.47	6.09 $\pm$ 0.17	3.54 $\pm$ 0.18
Neural Flow	7.20 $\pm$ 0.07	4.67 $\pm$ 0.04	13.12 $\pm$ 0.05	32.13 $\pm$ 0.07	4.05 $\pm$ 0.13	4.46 $\pm$ 0.09	5.35 $\pm$ 0.05	3.25 $\pm$ 0.05
GraFITi	5.11 $\pm$ 0.19	3.96 $\pm$ 0.09	10.49 $\pm$ 0.13	27.38 $\pm$ 0.03	3.03 $\pm$ 0.14	3.45 $\pm$ 0.10	5.07 $\pm$ 0.03	2.97 $\pm$ 0.04
t-PatchGNN	4.98 $\pm$ 0.08	3.72 $\pm$ 0.03	11.20 $\pm$ 1.20	27.77 $\pm$ 0.05	2.66 $\pm$ 0.03	3.15 $\pm$ 0.02	5.00 $\pm$ 0.04	3.08 $\pm$ 0.04
ISTS-PLM	4.92 $\pm$ 0.05	3.65 $\pm$ 0.04	10.08 $\pm$ 0.23	26.99 $\pm$ 0.06	2.58 $\pm$ 0.03	3.12 $\pm$ 0.04	5.23 $\pm$ 0.07	2.99 $\pm$ 0.07
Hi-Patch	4.86 $\pm$ 0.03	3.62 $\pm$ 0.07	9.92 $\pm$ 0.14	26.86 $\pm$ 0.04	2.57 $\pm$ 0.02	3.11 $\pm$ 0.03	5.14 $\pm$ 0.05	2.96 $\pm$ 0.04
SEQNet (Ours)	3.67 $\pm$ 0.05	3.20 $\pm$ 0.04	9.20 $\pm$ 0.06	26.36 $\pm$ 0.04	2.18 $\pm$ 0.02	2.68 $\pm$ 0.03	4.95 $\pm$ 0.05	3.01 $\pm$ 0.07

4.2 Forecasting Results

Table 2 reports forecasting MSE and MAE on four datasets. SEQNet achieves the lowest MSE on PhysioNet, MIMIC-III, Activity, and USHCN. Relative to the strongest MSE baseline on each dataset in Table 2, SEQNet reduces MSE by 24.49%, 7.26%, 15.18%, and 1.00%, respectively. Table 6 reports the results in unscaled units.

4.3 Observation-Deletion Robustness

We evaluate SEQNet on PhysioNet, USHCN, and Activity under five observation-deletion families and four retention levels, giving 20 test environments per evaluated dataset. Table 3 reports the average and worst MSE over these environments. Relative to clean evaluation, average MSE increases by 22.70% on PhysioNet, 6.26% on USHCN, and 416.48% on Activity. MIMIC-III has fewer eligible test targets and is evaluated only in the clean forecasting comparison.

4.4 Ablation Study

We analyze the contribution of event preservation, event selection, and source-wise routing from three complementary perspectives. First, we progressively construct SEQNet on PhysioNet by adding soft temporal memory, retained events, query-conditioned retrieval, and the complete training configuration. Second, we compare the combined event selector with single-criterion and random-event alternatives on USHCN. Third, we replace support-aware source routing with four restricted fusion strategies on USHCN and MIMIC-III. Figure 3 summarizes the results. Panel (a) reports MSE reductions relative to the reference configuration; panels (b)–(d) report MSE increases relative to the corresponding full model. Appendix F provides the numerical results and exact interventions.

Table 3: SEQNet test MSE under 20 observation-deletion environments. Average and worst MSE summarize five deletion families at four retention levels. The clean MSE is recorded for the same evaluation.

Dataset	Clean MSE	Average MSE	Worst MSE	Increase
PhysioNet	0.00367	0.00450303	0.00674313	22.70%
USHCN	0.495	0.52597860	0.58131493	6.26%
Activity	0.00218	0.01125928	0.06988143	416.48%

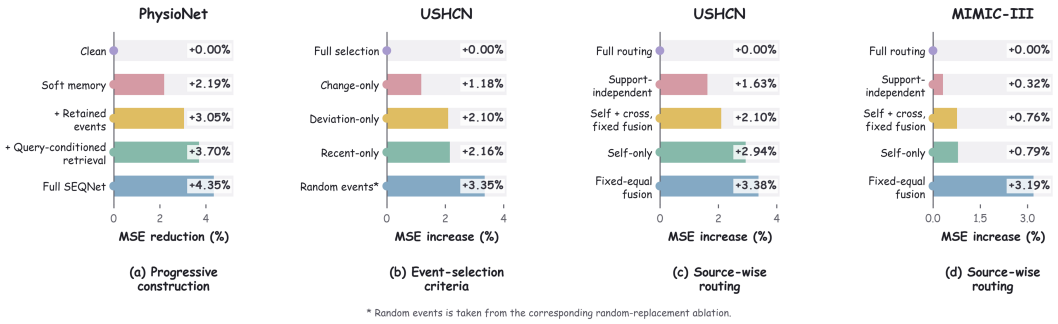


Figure 3: SEQNet model analyses. (a) Progressive construction on PhysioNet: soft temporal memory, retained events, query-conditioned retrieval, and the full model successively reduce MSE relative to the clean reference configuration. (b) USHCN event selection: the complete selection rule outperforms single-criterion and random-event alternatives. (c,d) Source-wise routing on USHCN and MIMIC-III: removing support dependence or changing source fusion increases MSE relative to full routing. Panel (b)'s random-event value comes from the matched random-replacement ablation.

Progressive construction. Figure 3(a) shows progressive improvements from soft temporal memory, retained events, and query-conditioned retrieval. Full SEQNet achieves the largest PhysioNet MSE reduction, supporting the complementary roles of compressed history, directly accessible events, and the complete routing and training design.

Event-selection criteria. Figure 3(b) compares the complete selector with change-only, deviation-only, recent-only, and random-event alternatives under the same event budget. Full selection performs best, showing that the three criteria provide complementary signals and that event identity, rather than token count alone, contributes to forecasting.

Source-wise routing. Figures 3(c) and 3(d) show that full support-aware routing outperforms support-independent, fixed-fusion, self-only, and fixed-equal alternatives on USHCN and MIMIC-III. The results support adapting self, cross-variable, and prior information to the target's current observation support.

5 Conclusion

We presented SEQNet for irregular multivariate forecasting under changing observation support. Soft temporal memory and retained events model global historical patterns and local observations through complementary paths, while query-conditioned retrieval and support-aware routing dynamically combine self, cross-variable, and prior information. Paired-view training and constrained family weighting further expose the model to different forms and degrees of historical deletion. Across four forecasting datasets, SEQNet achieves the lowest MSE among the compared methods, and the progressive construction, event-selection, and routing ablations support its key design choices. By preserving directly retrievable local events alongside compressed history and organizing predictive evidence according to the target's current observation support, SEQNet captures global temporal patterns without discarding critical local cues, especially when the target history is sparse, stale, or entirely absent.

REFERENCES

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.

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):160035, 2016.

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.

Boyuan Li, Zhen Liu, Yicheng Luo, and Qianli Ma. Learning recursive multi-scale representations for irregular multivariate time series forecasting. *arXiv preprint arXiv:2602.21498*, 2026a.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Guo Guo, and Bin Yang. Gcgnet: Graph-consistent generative network for time series forecasting with exogenous variables. In *International Conference on Learning Representations*, volume 2026, pp. 27904–27926, 2026b.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Guo Guo, Bin Yang, and Jilin Hu. Astgi: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 23940–23958, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 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 *Forty-second International Conference on Machine Learning*, 2025.

Matthew J Menne, Claude N Williams Jr, and Russell S Vose. The us historical climatology network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, 90(7): 993–1008, 2009.

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. *arXiv preprint arXiv:2602.00582*, 2026.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. In *... International Conference on Learning Representations*, volume 2021, pp. 14897, 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 *2012 computing in cardiology*, pp. 245–248. IEEE, 2012.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16255–16263, 2024.

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.

Weijia Zhang, Chenlong Yin, Hao Liu, and Hui Xiong. Unleashing the power of pre-trained language models for irregularly sampled time series. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3831–3842, 2025.

A Notation and Algorithms

Table 4 collects the symbols used in the method and algorithms. Algorithm 1 describes the inference path, including the empty-source cases. Algorithm 2 adds the training-only operations to that same predictor.

Table 4: Notation. Sample indices are b, historical event indices j, query indices i, variable indices v, and slot indices k.

Symbol	Meaning
B, V, N_b, Q_b	Batch size, number of variables, packed history size, and query count
$\mathcal{H}_b, \mathcal{Q}_b$	Visible historical record and requested time–variable queries
$x_{bj}, t_{bj}, c_{bj}, m_{bj}$	Historical value, physical time, variable identity, and visibility
y_{bi}, o_{bi}, a_{bi}	Training/evaluation target, validity, and target weight
u_{bj}, g_{bj}, τ_0	Relative time, log physical gap, and dataset time unit
$r_{bv}, \mathcal{I}_{bv}$	Six-dimensional support and the visible-event set of variable v
K, R, d, d_h	Slot count, maximum retained events, model dimension, head dimension
a_k, σ_{bvk}, Z	Fixed anchor, learned bandwidth (head index suppressed), and soft memory
$M_{bv}, s_{biv}, h^q_{bi}$	Source memory, query-conditioned source summary, and query token
β_{biv}, α_{bi}	Cross-source attention weights and three-route fusion weights
$\ell_{bi}, \eta_{bi}, s_b$	Latest visible target-variable value, availability, and static covariates
$\mathcal{V}^0_b, \mathcal{V}^h_b$	Original and difficult histories
p_b, q_b, p_g, q_g	Target/actual sample probabilities and base/adversarial family probabilities
$\theta, \theta_{\text{ref}}$	Model parameters and reference fusion-layer weight
ω_{event}	Event-supervision coefficient

Algorithm 1 SEQNet forward pass

Require: History $\mathcal{H}$, queries $\mathcal{Q}$, optional static features s, parameters θ
Ensure: Predicted means and auxiliary head outputs
Sanitize hidden/padded values, timestamps, and variable identifiers
Compute u, g from visible same-variable predecessors using Eq. (1)
Encode visible events as h; compute variable support r and valid-source flags
Construct K soft slots per variable using Eq. (2)
Select up to R distinct recent, high-deviation, and high-change real events
for each of the two interaction blocks **do**
    Apply within-variable temporal attention and its residual/FFN update
    Apply support-conditioned variable attention at each anchor and its residual/FFN update
    Zero invalid variable outputs
end for
Concatenate updated slots and retained events into each source memory M_v
for each requested query (τ_i, c^q_i) **do**
    Encode query time, forecast gap, target identity, and target support into h^q_i
    Read every nonempty M_v to obtain s_{iv} using Eq. (3)
    Construct self, cross, and prior representations
    Mask unavailable self/cross routes and normalize the three route logits
    Fuse the weighted route sum with last-value, availability, age, and static features
    Apply the output head; add the last-value mean residual if enabled
end for
return all query predictions

B Representation and Decoder Details

B.1 Visible-event encoding and support conventions

The event encoder uses a 16-dimensional time representation, a 16-dimensional variable embedding, one value coordinate, and one physical-gap coordinate. The time representation concatenates learned linear combinations of u and u^2 with learned sinusoidal features. Two linear layers with GELU produce a d-dimensional token, followed by a first-observation offset and layer normalization. Every

Algorithm 2 Paired-view training with constrained family weighting

Require: Training data, validation data, fitted training statistics, $p_g = 1/5$

Initialize model parameters θ and $q_g \leftarrow p_g$
for epoch $e = 1, \ldots, 50$, subject to the configured stopping rule **do**
  for batch drawn with actual sample probabilities q_b **do**
    Set sample correction $w_b \leftarrow p_b / q_b$
    Draw a family and retention for each sample and generate $\mathcal{V}_b^h \subseteq \mathcal{V}_b^0$
    Predict with the same model on $\mathcal{V}^0$ and $\mathcal{V}^h$
    Compute $\mathcal{L}_{\text{main}}$ using Eq. (5) and paired-view event loss
    Update the event-supervision coefficient when due
    Backpropagate Eq. (7), clip gradients, and update θ with AdamW
    Accumulate importance-corrected difficult-view losses by family
  end for
  Update smoothed family losses and q under Eq. (6) after warm-up
  Evaluate clean validation MSE; save the model if it improves
  Every five epochs, also record the five-family validation proxy
end for
return the model checkpoint selected by clean validation MSE

view recomputes predecessor gaps and support from its own visible events. Deleting an event updates memory membership and the available predecessor relation.

For a nonempty variable with n observations, duration $D = \max(t^e - t^s, \epsilon)$, and visible timestamps $t_1 \le \cdots \le t_n$, the support descriptor is

$$r_v = \left[\log(1+n), \frac{t_n - t_1}{D}, \frac{t^e - t_n}{D}, \frac{\max_{j>1}(t_j - t_{j-1})}{D}, \text{cov}_8, 0 \right]. \tag{8}$$

For a singleton, the normalized maximum gap is one and coverage is zero. For an empty variable, $r_v = [0, 0, 1, 1, 0, 1]$. The coverage statistic is the fraction of eight equal-width, left-closed and right-open history bins containing a visible observation; it is computed for variables with at least two observations. These conventions make lack of support explicit without introducing synthetic measurements.

B.2 Soft-memory bandwidth and event selection

The variable summary used by soft memory is computed from encoded visible events of that variable. Slot queries are linear projections of this summary, the learned slot embedding, and support. For head h, the frozen implementation computes bandwidth as

$$\sigma_{bvkh} = 0.02 + 0.98\,\text{sigmoid}\left(W_{\sigma,h}[\bar{h}_{bv}; r_{bv}] + b_{\sigma,h} + \text{mean}(e_k^{\text{slot}})\right), \tag{9}$$

where $\bar{h}_{bv}$ is the variable summary and e_k^{slot} is the learned slot embedding. This is a shared slot-dependent offset to the per-head bandwidth logits, rather than a separate unrestricted bandwidth network for every slot. The anchors remain fixed at the interval midpoints.

The event selector uses training-fitted medians μ_v^{med} and positive scales s_v^{rob}. After selecting the two latest visible events, it selects up to three remaining events by

$$A_{bj} = \frac{|x_{bj} - \mu_{c_{bj}}^{\text{med}}|}{\max(s_{c_{bj}}^{\text{rob}}, 10^{-6})}. \tag{10}$$

It then selects up to three further events with visible predecessors by

$$C_{bj} = \frac{|x_{bj} - x_{b,j^-}|}{\max(s_{c_{bj}}^{\text{rob}}, 10^{-6})\sqrt{1 + (t_{bj} - t_{b,j^-})/\tau_0}}, \tag{11}$$

where j^- is the preceding visible event of the same variable. Events with no predecessor cannot enter by the change criterion. Selection is discrete and deduplicated in recent–amplitude–change order; unused slots are masked. The selected encoded tokens receive embeddings identifying which criterion selected them. The criterion combines value change and elapsed time.

B.3 Routing and missing-source cases

The query time features consist of relative query time and forecast gap, projected to 16 dimensions. A separate target-variable embedding and projected target support feed a two-layer query MLP with normalization. Multihead query attention reads each source memory with token-level validity masks. An empty source returns a zero summary. When the target has no observations but another variable is present, cross and prior remain available. When no variable is observed, the router places all weight on the prior. Last-value features are zeroed when unavailable, with an explicit availability indicator provided to the shared fusion MLP.

B.4 Output types

For continuous variables, a shared mean head produces the regression output and a two-logit head produces the lower/upper-tail auxiliary predictions. The continuous event labels are $\mathbb{1}[y < Q_{0.05,v}]$ and $\mathbb{1}[y > Q_{0.95,v}]$, using quantiles fitted on training targets. Both original and difficult views receive the same labels and equal view weights. The event loss is summed over the two binary predictions, multiplied by valid target weights and sample correction, and averaged over the batch.

C Training Details

C.1 Observation families and sampling correction

All perturbations delete only originally visible historical observations. Random-point deletion spreads removal across events; channel-subset deletion removes variable support; time-block deletion removes a contiguous temporal region; early and late deletion preferentially remove observations from opposite ends of the history. Retention denotes the requested retained fraction, with discrete counts determined by the generator. The query set, future values, and target validity are held fixed across views. A SHA256-derived stable seed identifies each dataset/sample/family/retention/replica combination, with the epoch providing the training replica.

Stratified sample sampling can oversample tail or event strata. Its correction p_b/q_b is distinct from the family loss ratio q_g/p_g: the former preserves a chosen sample risk, whereas the latter deliberately changes family emphasis. For a fixed per-sample loss ℓ_b, positive q_b on the support of p_b yields the identity

$$\mathbb{E}_{b\sim q}\left[\frac{p_b}{q_b}\ell_b\right] = \sum_b p_b \ell_b. \tag{12}$$

Target weights a_{bi} come from the dataset adapter and target validity. Evaluation metrics follow the reductions in Section E.3.

C.2 Constrained family solver

The family-loss accumulator uses the difficult views already processed for training, with sample importance weights. An exponential moving average with decay 0.9 smooths each observed family mean; a newly observed family is initialized from its first value. Uniform weights are used during the first ten epochs. At subsequent epoch ends, the KL radius ramps to 0.05 over five epochs, and the new family weights are used in following optimization steps.

For the box-plus-simplex problem, the solver starts at the lower bounds and allocates the remaining mass in decreasing loss order. If the resulting vector satisfies the KL constraint, it solves Eq. (6). Otherwise, the constrained solution has the clipped exponential form

$$q_g = \operatorname{clip}\left(p_g \exp\left(\frac{\overline{L}_g - \nu}{\tau}\right), 0.5p_g, 2p_g\right), \tag{13}$$

for an appropriate normalization offset ν and positive temperature τ. Nested bisection enforces the simplex and KL constraints; small floating-point simplex residuals are redistributed only within remaining box slack. When all family losses coincide, $q = p$ is returned.

C.3 GRADIENT-BUDGET CONTROLLER

Let θ_{ref} be the first linear weight in the decoder's shared fusion MLP. For the event-supervision term $a = \mathrm{event}$, define its weighted eligibility coverage c_a as eligible event-target weight divided by total valid target weight. At a controller update, let

$$G_0 = \|\nabla_{\theta_{\mathrm{ref}}} \mathcal{L}_{\mathrm{main}}\|_2, \quad G_a^{\mathrm{cond}} = \left\| \nabla_{\theta_{\mathrm{ref}}} \frac{\mathcal{L}_a}{c_a + \epsilon} \right\|_2, \quad G_a^{\mathrm{raw}} = \|\nabla_{\theta_{\mathrm{ref}}} \mathcal{L}_a\|_2. \tag{14}$$

Moving averages with decay 0.9 define a provisional coefficient

$$\omega_a^* = \mathrm{clip}\left(\rho_a \frac{\mathrm{EMA}(G_0)}{\mathrm{EMA}(G_a^{\mathrm{cond}}) + \epsilon_G}, 0, 10 \right), \qquad \rho_a = 0.1. \tag{15}$$

The coefficient itself is smoothed with decay 0.9 after initialization. Zero-coverage terms receive zero coefficients at an update. A common rescaling enforces a total budget of $0.3\,\mathrm{EMA}(G_0)$ against the sum of coefficient-weighted moving-average raw norms.

D STRUCTURAL PROPERTIES AND COMPUTATIONAL SCOPE

Proposition 1 (Hidden-payload invariance)**.** Fix the visible events, query set, static inputs, history boundaries, and model parameters. Under deterministic evaluation, changes to the values, times, or variable identifiers of hidden/padded positions do not change the predictions, provided those positions remain excluded before feature construction and event selection.

Proof. The event encoder replaces all hidden payloads before learned transformations and zeros the resulting hidden tokens. Visible predecessor gaps, support descriptors, variable summaries, and selected events are computed from visible positions only, so each is unchanged. Soft-memory attention excludes the same hidden positions and therefore produces the same valid slots. All subsequent blocks, query reads, route masks, and output heads receive identical valid inputs. Their deterministic composition gives identical predictions. □

The hypothesis includes fixed history boundaries, because changing them changes relative times and support. The implementation's checks cover hidden-payload and padding invariance, empty histories, single-event inputs, batch gathering for retained events, and the DRO constraints.

To state the computational scope precisely, let N and Q be padded historical and query lengths. The factorized stack builds $VK^2 + KV^2$ attention scores per sample and head per block, while fully connected attention over VK tokens builds V^2K^2 scores. With L blocks, its arithmetic cost is

$$O\big(BL\big[(VK^2 + KV^2)d + VKd^2\big]\big), \tag{16}$$

including projection and feed-forward terms up to fixed constants. The current dense masked soft-memory implementation constructs source–slot–event logits for all variables before masking, so that stage costs $O(BVKNd)$ rather than assuming a sparse $O(BKNd)$ implementation. Source-wise decoding adds attention work $O(BQV(K+R)d)$, together with projection costs. Retained events are absent from the factorized stack, but still incur encoding, selection, and decoder costs.

Training evaluates the model on original and difficult histories. The gradient controller measures the event-loss gradients periodically. Inference uses one forward pass on the observed history.

E DATASET PROTOCOLS AND EVALUATION MEASURES

E.1 DATA CONSTRUCTION

Activity. The task uses 12 coordinate channels and 10,739 windows, with 6,443 training, 2,148 validation, and 2,148 test windows. A three-second history predicts the following second. Channel min–max statistics are fitted using the training partition and then frozen. The batch size is 32, the physical-gap time unit is 0.1 in the adapter's time coordinates, and last-value mean residuals are enabled. The test split contains valid scalar forecasting targets. We use a window-level split.

Table 5: Shared architecture and training settings. Dataset-specific batches, time units, and residual choices are stated above. Results are summarized over five random seeds.

Setting	Value
Hidden dimension / attention heads / interaction blocks	64 / 4 / 2
Soft slots / maximum retained events per variable	8 / 8
Recent / amplitude / change selection quotas	2 / 3 / 3
Soft-memory bandwidth interval	$[0.02, 1]$
Optimizer / initial learning rate / weight decay	AdamW / 10^{-3} / 10^{-4}
Gradient clipping / maximum epochs / number of seeds	1 / 50 / 5
Original / difficult supervised-view weights	0.5 / 0.5
Retention levels	$\{0.8, 0.6, 0.4, 0.2\}$
DRO lower/upper family multipliers / maximum KL	0.5 / 2 / 0.05
Gradient-controller period / EMA	10 steps / 0.9
Per-term / total auxiliary gradient budgets	0.1 / 0.3
Maximum auxiliary coefficient	10

Table 6: SEQNet test MSE and MAE over five random seeds. Values are reported as mean $\pm$ standard deviation in unscaled units.

Dataset	Test MSE	Test MAE
Activity	0.00218 ± 0.00002	0.0268 ± 0.0003
PhysioNet 2012	0.00367 ± 0.00005	0.0320 ± 0.0004
USHCN	0.495 ± 0.00005	0.301 ± 0.007
MIMIC-III	0.0920 ± 0.0006	0.2636 ± 0.0004

PhysioNet 2012. The 41-variable configuration begins from the full list of 12,000 patient identifiers. Identifiers are shuffled with split seed 42 and assigned in proportions 60/20/20 before eligibility filtering. The resulting train/validation/test sample counts are 7,187/2,396/2,392. The first 24 hours form history, and the following 24 hours provide targets. Scaling is training-only per-channel min–max, batch size is 32, and the physical-gap time unit is one hour. The test split contains 469,482 valid targets, and the last-value mean residual is disabled.

USHCN. The processed collection contains 1,114 stations and five variables. Station splitting precedes window construction, yielding 15,715/5,224/5,259 samples without stations shared across partitions. Each sample uses 24 months to forecast the following month. The current adapter retains its processed raw coordinates without an additional min–max transform. The batch size is 64, the physical-gap unit is 30 days, and the test split contains 39,064 valid targets. Last-value mean residuals are disabled.

MIMIC-III. The MIMIC-III dataset contains 23,457 records with 96 variables. We use a fixed 60/20/20 train/validation/test split at the forecasting-instance level, with the same partition membership retained throughout training and evaluation. The historical window is 0–24 h and the target window is 24–48 h; the two windows do not overlap, and observations outside the specified 48-hour interval are not forecast targets. The split seed is 42, giving 14,073 training, 4,692 validation, and 4,692 test forecasting samples. Min–max statistics are fitted on the training partition. Batch size is eight and the physical-gap unit is one hour. An entry is observed when its value is finite and nonzero.

E.2 Optimization and checkpoint selection

Table 5 lists the shared settings. Early stopping for PhysioNet 2012 and MIMIC-III starts at epoch 11, uses ten consecutive non-improving clean-validation epochs, and is protected by a minimum of 25 epochs. MIMIC-III uses early stopping and selects checkpoints by validation MSE. Its learning rate uses a plateau scheduler with patience ten and factor 0.5. The maximum training budget is 50 epochs.

We select checkpoints using clean validation MSE.

Table 7: PhysioNet progressive construction in Figure 3(a). Reductions are relative to the clean reference configuration.

Configuration	MSE reduction
Clean reference	0.00%
Soft temporal memory	2.19%
Memory + retained events	3.05%
Memory + events + query-conditioned retrieval	3.70%
Full SEQNet	4.35%

Table 8: Source-wise routing comparisons. MSE increases are relative to full routing on each dataset.

Configuration	USHCN	MIMIC-III
Full routing	0.00%	0.00%
Support-independent	1.63%	0.32%
Self + cross, fixed fusion	2.10%	0.76%
Self-only	2.94%	0.79%
Fixed-equal fusion	3.38%	3.19%

Table 9: USHCN component ablations. Metrics are unscaled channel-macro errors; changes use full SEQNet as reference.

Variant	Test MSE	Test MAE	MSE increase
Full SEQNet	0.495000	0.301000	–
Remove retained events	0.501103	0.306145	1.23%
Random events	0.511604	0.309000	3.35%
Remove query routing	0.504021	0.304085	1.82%

E.3 Metric definitions

Let $\mathcal{O}_b$ contain valid scalar future targets of sample b, and let $e_{bi} = \widehat{y}_{bi} - y_{bi}$. Activity and PhysioNet 2012 first average squared error over each sample's targets and then over valid samples:

$$\mathrm{MSE}_{\mathrm{sample}} = \frac{1}{|\mathcal{B}_+|} \sum_{b \in \mathcal{B}_+} \frac{1}{|\mathcal{O}_b|} \sum_{i \in \mathcal{O}_b} e_{bi}^2, \qquad \mathcal{B}_+ = \{b : |\mathcal{O}_b| > 0\}. \tag{17}$$

This is the sample-micro reduction used in the reported results, rather than pooling all scalar errors across patients before division. For channel-macro evaluation, define $\mathcal{O}_v = \{(b,i) : i \in \mathcal{O}_b, c_{bi}^q = v\}$ and $\mathcal{C}_+ = \{v : |\mathcal{O}_v| > 0\}$. Then

$$\mathrm{MSE}_{\mathrm{channel}} = \frac{1}{|\mathcal{C}_+|} \sum_{v \in \mathcal{C}_+} \frac{1}{|\mathcal{O}_v|} \sum_{(b,i) \in \mathcal{O}_v} e_{bi}^2. \tag{18}$$

This is the USHCN headline metric. For MIMIC-III, MSE pools all observed scalar targets before division:

$$\mathrm{MSE}_{\mathrm{entry}} = \frac{\sum_b \sum_{i \in \mathcal{O}_b} e_{bi}^2}{\sum_b |\mathcal{O}_b|}. \tag{19}$$

MAE replaces squared error by absolute error under the same reduction, and headline RMSE is the square root of headline MSE.

F Details of the Component Ablations

Shared protocol. The three USHCN representation ablations use the proposed backbone, not a substituted t-PatchGNN backbone. All use the same 15,715/5,224/5,259 station-disjoint forecasting samples, 39,064 valid test targets, raw coordinates, batch size 64, and a maximum of 50 epochs. The checkpoint with the lowest clean validation MSE is evaluated once on the clean test split using channel-macro MSE and MAE. Event auxiliary supervision, stratified sampling, paired views, and constrained family weighting are used in all variants. Checkpoints are selected by clean validation MSE.

Table 10: USHCN event-selection results. MSE and MAE are unscaled; relative changes are measured against the full model.

Selection	Test MSE	Test MAE	MSE change (%)
Recent-only	0.50568021	0.30886359	+2.16
Deviation-only	0.50540020	0.30449949	+2.10
Change-only	0.50084060	0.30314251	+1.18
Random events	0.51160400	0.30900000	+3.35
Full selection	0.495000	0.301000	0.00

Remove Retained Events. Decoder memory contains only the updated soft slots, $M_{bv} = Z_{bv}$. Event encoding, memory construction, factorized interaction, the query decoder, and the output heads retain their full-model definitions. In particular, the tail-label auxiliary loss remains active.

Random Events. The selector first determines the full model's valid retained-token slots. Within each variable it then assigns pseudorandom scores to visible observations and chooses replacement indices without replacement. The original valid-slot mask and selection-type embeddings are retained, so the intervention changes event identity while preserving slot count and type allocation. Each run uses its own seeded local generator for event replacement.

Remove Query Routing. Per-source attention in Equation (3) is retained. For the set $\mathcal{C}_b$ of nonempty source variables, define

$$\bar{s}_{bi} = \frac{1}{|\mathcal{C}_b|} \sum_{v \in \mathcal{C}_b} s_{biv}, \qquad \bar{r}_b = \frac{1}{|\mathcal{C}_b|} \sum_{v \in \mathcal{C}_b} r_{bv}. \tag{20}$$

Both quantities are zero when $\mathcal{C}_b$ is empty. The self and cross inputs to the mixture are both set to $\bar{s}_{bi}$, and the router receives $\bar{r}_b$ in place of target-specific support. Both historical routes are available whenever any source is nonempty. The query encoder still receives the target identity and target support; the prior, mixture MLP, last-value features, fusion layer, and output heads are retained. This experiment tests source differentiation at readout.

AI Use Statement

Generative AI tools assisted language revision, LaTeX organization, consistency checking, figure-caption and reference preparation, and pseudocode drafting. The authors are responsible for the final manuscript, citations, methodological descriptions, implementation, experimental protocols, and reported results.