NUCLEAR FUSION KNOWLEDGE GRAPH GUIDED EDGE-LOCALIZED MODE RECOGNITION ON EAST

Anonymous authors
Paper under double-blind review

ABSTRACT

Recognizing edge-localized modes (ELMs) and their surrounding plasma states is essential for understanding tokamak operation and supporting plasma monitoring. The task requires distinguishing physically different states with similar diagnostic waveforms, resolving brief events, and learning from strongly imbalanced classes. Human recognition combines diagnostic evidence, physical rules, and experience from previous discharges, providing structured information beyond temporal patterns alone. We propose KGF-ELM, a reliability-aware knowledge fusion framework for point-wise recognition of four ELM-related plasma states on EAST. An ELM knowledge graph abstracts the manual recognition workflow into rules that guide historical case construction from 100 training discharges. The Knowledge Graph Collaborative Fusion Module (KGCFM) brings together physical evidence, prototype activations, relational graph representations, and retrieved cases. It coordinates their predictions using source-quality scores and cross-source agreement, then incorporates knowledge through feature enhancement and gated decision correction. A minority-group safeguard attenuates selected collaborative updates, while joint training aligns the knowledge pathways with a replaceable temporal backbone. Experiments on EAST-ELM640 show Macro-F1 gains of 2.4–10.4 percentage points across five backbones, averaging 4.7 points. With CrossLinear, KGF-ELM achieves 91.7% Accuracy, 81.6% Macro-F1, and 98.0% Macro-AUC, leading the compared temporal models on these metrics. Ablations show that the complete framework outperforms its evaluated fusion and knowledge-component variants. These results establish structured domain knowledge as an effective complement to temporal modeling for ELM-related plasma-state recognition.

1 INTRODUCTION

Magnetic-confinement fusion relies on sustained high-performance plasma operation, where high-confinement mode is frequently accompanied by edge-localized modes (ELMs) (Zohm, 1996; Leonard, 2014). ELM bursts release particles and energy and can impose transient loads on plasma-facing components (Loarte et al., 2003). Their signatures vary with operating conditions, as illustrated by the diversity of small-ELM regimes on EAST (Wang et al., 2014). Tracking ELM-related plasma states therefore requires more than identifying isolated bursts: the surrounding confinement context and transitions between states also matter. Automatic point-wise recognition provides a detailed description of this evolution for plasma monitoring and analysis.

This recognition problem combines three challenges. Different physical states can produce similar local waveforms, brief events must be resolved within longer intervals, and rare ELM classes provide limited supervision. Temporal models capture local patterns, long-range dependencies, and cross-channel interactions (Nie et al., 2022; Wu et al., 2022; Liu et al., 2024a), but these learned representations do not explicitly encode the diagnostic rules and historical references used in manual recognition. Structured domain knowledge can supply this additional context, especially when the waveform alone leaves the plasma state ambiguous.

We organize such knowledge around the manual recognition process: assess confinement context, identify diagnostic activity, and combine evidence to distinguish states. Physical cues describe the current window; prototypes summarize characteristic evidence combinations; typed graph relations

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.

connect evidence, prototypes, and classes; and historical cases provide concrete references from previous discharges. These sources contribute at different levels, yet their usefulness varies with signal quality, case similarity, and agreement. The central challenge is to translate this heterogeneous knowledge into sample-dependent guidance that improves temporal representations and decisions.

We address this challenge with KGF-ELM, a framework centered on the Knowledge Graph Collaborative Fusion Module (KGCFM). An ELM knowledge graph abstracted from manual recognition guides rule-based case construction from 100 training discharges. KGCFM coordinates current physical evidence, relational representations, and retrieved historical context using source-quality scores and prediction agreement. It couples feature enhancement with gated decision correction, while a minority-group safeguard attenuates selected collaborative updates. This shared knowledge interface complements replaceable temporal backbones and improves Macro-F1 across five architectures by 4.7 percentage points on average, reaching 81.6% with CrossLinear.

Our main contributions are:

- We introduce KGCFM, which coordinates heterogeneous ELM knowledge through reliability-aware feature enhancement and source-specific and collaborative decision correction.
- We connect domain recognition rules to temporal modeling through physical evidence, prototypes, a typed relational graph, and training-discharge case memory.
- We evaluate the framework across five temporal backbones and analyze fusion pathways, knowledge components, and resource configurations for four-state point-wise recognition.

2 Related Work

2.1 Time-Series Representation Learning

Temporal models use diverse inductive biases to represent multivariate signals. Transformers model long-range interactions (Vaswani et al., 2017); PatchTST, TimesNet, and iTransformer emphasize temporal patches, multi-periodic structure, and variable-wise tokenization, respectively (Nie et al., 2022; Wu et al., 2022; Liu et al., 2024a). DLinear provides a linear alternative (Zeng et al., 2023), while TimeMixer and SOFTS use multiscale mixing and series–core fusion (Wang et al., 2024a; Han et al., 2024). Timer and Timer-XL investigate generative pre-training and long-context modeling; TimeXer, TimeFilter, and CrossLinear model exogenous interaction, graph filtering, and cross-variable correlation (Liu et al., 2024b; 2025; Wang et al., 2024b; Hu et al., 2025; Zhou et al., 2025). KGF-ELM complements these representation strategies with explicit diagnostic rules and historical evidence through a shared knowledge interface.

2.2 Knowledge-Enhanced and Physics-Informed Learning

Physics-informed learning introduces physical constraints and domain priors into learned models (Raissi et al., 2018). Prototypes and relational encoders represent class and relational structure (Snell et al., 2017; Ji et al., 2021; Schlichtkrull et al., 2018). Retrieval also supplies historical context for temporal prediction: RAFT retrieves similar training sequences and their continuations (Han et al., 2025), while PFRP retrieves from a global memory bank and adaptively combines historical and local predictions (Du et al., 2026). These forecasting methods establish historical retrieval and adaptive combination as relevant precedents.

KGF-ELM addresses four-state point-wise plasma recognition by coupling rule-defined diagnostic evidence and prototypes with a typed evidence–prototype–class graph and historical cases. KGCFM coordinates these sources through quality and agreement cues at both feature and decision levels. In plasma research, learned models address disruption prediction, magnetic control, automatic ELM labeling, and diagnostic event detection (Kates-Harbeck et al., 2019; Degrave et al., 2022; O'Shea et al., 2023; Teo et al., 2024). Our focus is integrating domain recognition knowledge with replaceable temporal models under a common EAST-ELM640 protocol.

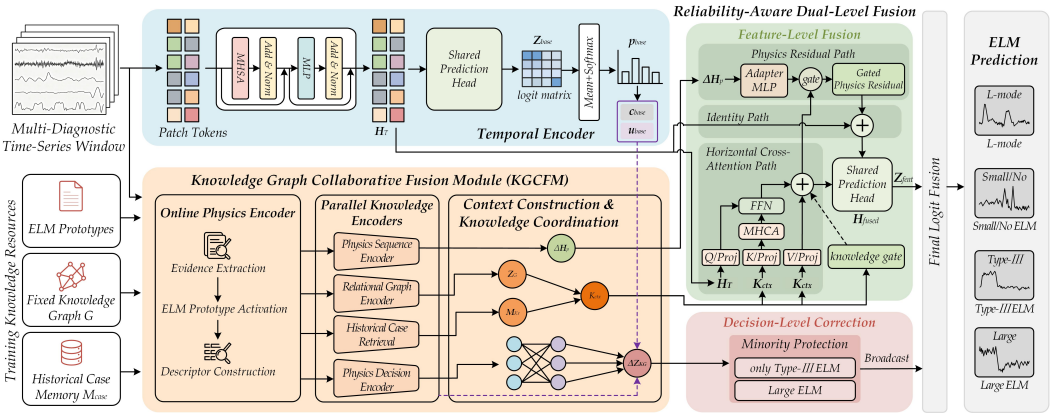

Figure 1: Overall architecture of KGF-ELM. The patch-attention path illustrates one possible backbone realization. KGCFM supplies physical and graph–case information for feature enhancement and knowledge logits for decision correction. The minority-protection block attenuates only the collaborative correction; source-specific corrections bypass this attenuation, as specified in Eq. (15). The resulting correction is broadcast and added to the feature-level logits.

3 Problem Setup and Framework

3.1 Problem Formulation

Given a batch of sliding windows extracted from multi-diagnostic plasma signals, the input and point-wise labels are defined as

$$\boldsymbol{X} \in \mathbb{R}^{B\times T\times C}, \qquad \boldsymbol{Y} = [y_{b,t}] \in \mathcal{Y}^{B\times T}, \qquad \mathcal{Y} = \{0,1,2,3\}, \tag{1}$$

where B, T, and C denote the batch size, window length, and number of diagnostic channels, respectively. Labels 0–3 denote L-mode, Small/No ELM, Type-III ELM, and Large ELM, respectively. The model predicts a class distribution at every time step,

$$\boldsymbol{P}_{\text{final}} = F(\boldsymbol{X}) \in [0,1]^{B\times T\times K}, \qquad \widehat{y}_{b,t} = \operatorname*{argmax}_{k\in\mathcal{Y}} P_{\text{final},b,t,k}, \tag{2}$$

with $K = 4$. The task is challenging because different ELM-related plasma states can exhibit similar local waveforms, while Type-III ELM and Large ELM are substantially underrepresented. We use a temporal model as the main point-wise prediction path and jointly train it with physical, relational, and historical knowledge pathways.

3.2 Overall Framework

Figure 1 presents the overall architecture of KGF-ELM. Given an input window $\boldsymbol{X}$, a replaceable temporal backbone first extracts token-level representations and produces point-wise base logits:

$$\begin{aligned} \boldsymbol{H}_T &= f_\theta(\boldsymbol{X}), \qquad \boldsymbol{H}_T \in \mathbb{R}^{B\times N\times d}, \\ \boldsymbol{Z}_{\text{base}} &= h_\psi(\boldsymbol{H}_T, \mathcal{C}_T) \in \mathbb{R}^{B\times T\times K}, \end{aligned} \tag{3}$$

where N depends on the backbone tokenization, d is the hidden dimension, and $\mathcal{C}_T$ is optional backbone-specific context. Linear maps act on feature dimensions; softmax normalizes over classes. Window-level confidence and predictive-entropy scores are

$$\begin{aligned} \boldsymbol{p}_{\text{base}} &= \text{Softmax}\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Z}_{\text{base}}[:,t,:]\right), \\ c_{\text{base},b} &= \max_{k\in\mathcal{Y}} p_{\text{base},b,k}, \qquad u_{\text{base},b} = -\sum_{k\in\mathcal{Y}} p_{\text{base},b,k}\log p_{\text{base},b,k}. \end{aligned} \tag{4}$$

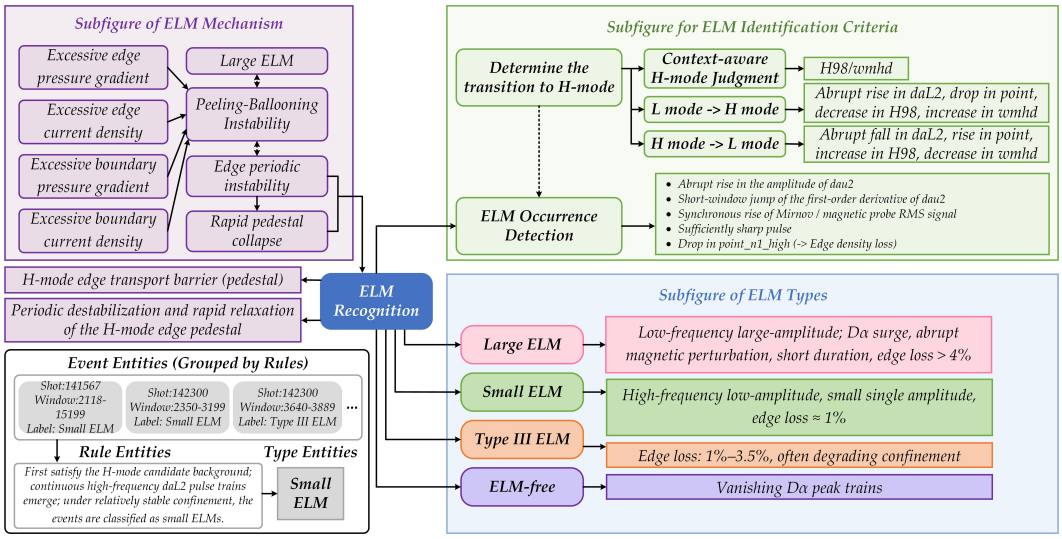

Figure 2: ELM domain-knowledge graph abstracted from the manual recognition workflow, organizing physical mechanisms, diagnostic evidence, and recognition rules. Applying operational rules to 100 training discharges constructs the historical case resource. KGCFM combines a compact relational encoding of this knowledge with retrieved case records.

Averaging logits provides a common window summary for gating and the auxiliary objectives, while classification remains point-wise. Its softmax is the normalized geometric mean of point-wise class probabilities, rather than a class-duration estimate. In mixed-state windows, its entropy can reflect both state heterogeneity and prediction ambiguity.

KGCFM supplies physical evidence, graph context, and retrieved cases for feature enhancement and gated decision correction. Joint training combines point-wise supervision with auxiliary knowledge guidance, so the knowledge objectives also shape the temporal representation. We reserve "baseline" for a separately trained temporal model and use "base output" for the internal prediction before fusion.

4 Reliability-Aware Knowledge Fusion

4.1 Multi-Source Knowledge Representation

Figure 2 organizes the manual ELM recognition workflow: assess confinement context, identify diagnostic activity, and combine evidence to distinguish plasma states. We abstract this workflow into a domain-knowledge graph and operational evidence and prototype rules. Applying the rules to 100 training discharges produces historical case records. KGCFM uses $\mathcal{K}_{\mathrm{ELM}} = \{\mathcal{E}, \boldsymbol{P}_{\mathrm{proto}}, \mathcal{G}, \mathcal{M}_{\mathrm{case}}, \mathcal{I}_{\mathrm{ret}}\}$: evidence definitions, prototype rules, a task-specific typed graph, case memory, and its retrieval index. The typed graph contains 11 evidence, 15 prototype, and four class nodes; historical cases enter as retrieved context tokens. The four outputs follow the dataset categories: L-mode, Small/No ELM, Type-III ELM, and Large ELM. Sections B–C specify the operational rules, fixed graph associations, and case construction.

For each window, the window-level physics analyzer computes an 11-dimensional evidence vector $\boldsymbol{e}$, a 15-dimensional prototype activation $\boldsymbol{a}$, and four-dimensional soft class support $\boldsymbol{s}$, forming $\boldsymbol{q} = [\boldsymbol{e}; \boldsymbol{a}; \boldsymbol{s}] \in \mathbb{R}^{B\times 30}$. It also produces $\boldsymbol{S}_{\mathrm{phy}} \in \mathbb{R}^{B\times T\times 8}$ and physics reliability r_{phy}. The descriptor $\boldsymbol{q}$ drives graph pooling, retrieval, and physics decision encoding, while $\boldsymbol{S}_{\mathrm{phy}}$ supports feature-level fusion.

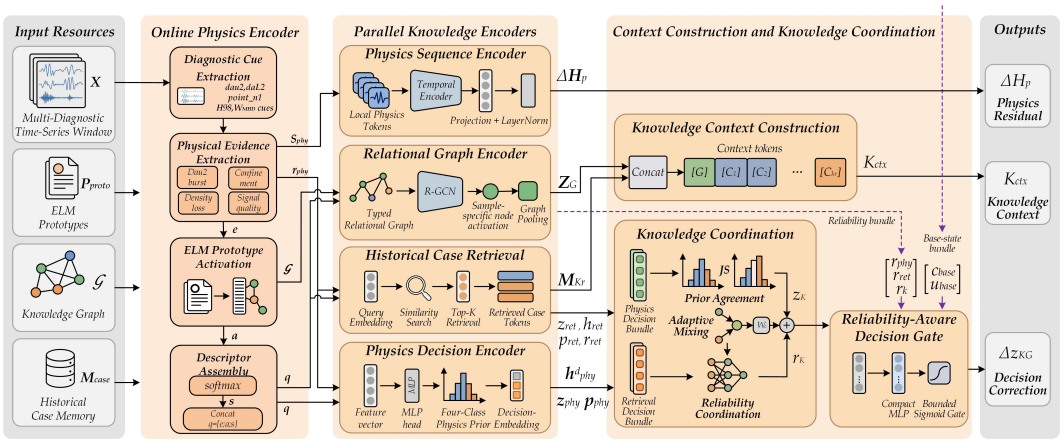

Figure 3: KGCFM provides a physics residual, graph–case context, and reliability-gated decision guidance through interacting pathways. The depicted sigmoid bounds a gate coefficient; the decision correction combines source-specific and collaborative logits as specified in Eqs. (13)–(15).

4.2 Knowledge Graph Collaborative Fusion

Figure 3 illustrates the detailed architecture of the Knowledge Graph Collaborative Fusion Module (KGCFM). Starting from the sample-specific descriptor $\boldsymbol{q}$, KGCFM processes physical, relational, and historical knowledge through interacting pathways and coordinates them according to source agreement and reliability.

The physics pathway provides complementary feature- and decision-level information. A physics sequence encoder maps the local physics sequence $\boldsymbol{S}_{\text{phy}}$ to a token-space residual,

$$\Delta \boldsymbol{H}_p = \text{LN}\left(\boldsymbol{W}_p E_{\text{phy}}(\boldsymbol{S}_{\text{phy}})\right). \tag{5}$$

Here, LN denotes layer normalization. Adaptive pooling matches the backbone token count, yielding $\Delta \boldsymbol{H}_p \in \mathbb{R}^{B\times N\times d}$. The physics decision encoder uses $\boldsymbol{q}$ augmented with reliability, density, and diagnostic-role cues to produce $\boldsymbol{h}^d_{\text{phy}}$ and source logits $\boldsymbol{z}^0_{\text{phy}}$. Superscripts 0 and c distinguish source-specific outputs from their collaborative adaptations. For $j \in \{\text{phy}, \text{ret}\}$ and $\nu \in \{0, c\}$, $\boldsymbol{p}^\nu_j = \text{Softmax}(\boldsymbol{z}^\nu_j)$; the adapters are detailed in Section D.

For relational reasoning, a two-layer relational graph convolutional network (R-GCN) encodes the typed knowledge graph $\mathcal{G}$, while sample-specific activations derived from $\boldsymbol{q}$ determine the graph-pooling weights. The pooled graph representation is projected to the hidden dimension of the temporal backbone:

$$\begin{aligned} \boldsymbol{H}_G &= \text{RGCN}^{(2)}(\mathcal{G}) \in \mathbb{R}^{30\times d_G}, \\ \widetilde{\boldsymbol{Z}}_G &= \text{Pool}_{\boldsymbol{q}}(\boldsymbol{H}_G), \\ \boldsymbol{Z}_G &= \boldsymbol{W}_G \widetilde{\boldsymbol{Z}}_G \in \mathbb{R}^{B\times 1\times d}. \end{aligned} \tag{6}$$

For historical retrieval, $\boldsymbol{q}$ selects the top-K_r training cases by descriptor cosine similarity. The same candidates yield source and collaborative supports; collaborative reweighting additionally uses prototype consistency and agreement with $\boldsymbol{p}^c_{\text{phy}}$, together with offline guidance scores:

$$\left(\boldsymbol{M}_{K_r}, \widetilde{\boldsymbol{p}}^0_{\text{ret}}, \widetilde{\boldsymbol{p}}^c_{\text{ret}}\right) = \mathcal{R}_{\text{case}}\left(\boldsymbol{q}, \boldsymbol{a}, \boldsymbol{p}^c_{\text{phy}}, \mathcal{M}_{\text{case}}, \mathcal{I}_{\text{ret}}\right). \tag{7}$$

Here, $\boldsymbol{M}_{K_r} \in \mathbb{R}^{B\times K_r\times d}$ contains the retrieved case representations. Together with the graph token, they form

$$\boldsymbol{K}_{\text{ctx}} = [\boldsymbol{Z}_G; \boldsymbol{M}_{K_r}] \in \mathbb{R}^{B\times(K_r+1)\times d}. \tag{8}$$

Retrieval-quality cues determine r^0_{ret} and r^c_{ret}. The source encoder maps source support and quality cues to $\boldsymbol{h}^0_{\text{ret}}$ and $\boldsymbol{z}^0_{\text{ret}}$; residual adapters produce their collaborative counterparts using the difference between the two aggregation summaries. The learned priors $\boldsymbol{p}^\nu_{\text{ret}}$ are distinct from the raw case-aggregated supports $\widetilde{\boldsymbol{p}}^\nu_{\text{ret}}$.

KGCFM coordinates the adapted physics and retrieval outputs through prior agreement, adaptive mixing, and reliability scoring:

$$c_{\mathrm{pr}} = 1 - \frac{\mathrm{JS}(\boldsymbol{p}^c_{\mathrm{phy}}, \boldsymbol{p}^c_{\mathrm{ret}})}{\log 2}. \tag{9}$$

Here, JS is the Jensen–Shannon divergence with natural logarithms, so $c_{\mathrm{pr}} \in [0, 1]$. Using $\boldsymbol{h}^d_{\mathrm{phy}}$, $\boldsymbol{h}^c_{\mathrm{ret}}$, the adapted priors, and quality cues including r_{phy}, r^c_{ret}, and c_{pr}, the coordinator produces $\boldsymbol{h}_K$, a mixing coefficient $\eta \in [0, 1]$, reliability r_K, and knowledge logits:

$$\begin{aligned} \boldsymbol{z}_K &= \eta \boldsymbol{z}^c_{\mathrm{phy}} + (1-\eta)\boldsymbol{z}^c_{\mathrm{ret}} + \boldsymbol{W}_I \boldsymbol{h}_K, \\ \boldsymbol{p}_K &= \mathrm{Softmax}(\boldsymbol{z}_K). \end{aligned} \tag{10}$$

The coordinator combines agreement with source-quality cues because the two pathways share diagnostic information. The resulting reliability scores control the contribution of coordinated knowledge. KGCFM supplies $\Delta \boldsymbol{H}_p$ and $\boldsymbol{K}_{\mathrm{ctx}}$ for feature enhancement, together with source-specific and coordinated outputs for decision fusion.

4.3 Reliability-Aware Dual-Level Fusion

KGF-ELM incorporates multi-source knowledge through feature enhancement and decision correction. At the feature level, the temporal states query the graph–case context $\boldsymbol{K}_{\mathrm{ctx}}$, whereas local physical information enters through the separate residual $\Delta \boldsymbol{H}_p$:

$$\begin{aligned} \Delta \boldsymbol{H}_K &= \mathrm{LN}\left(\boldsymbol{W}_K \, \mathrm{MHA}(\boldsymbol{H}_T, \boldsymbol{K}_{\mathrm{ctx}}, \boldsymbol{K}_{\mathrm{ctx}})\right), \\ \boldsymbol{H}_{\mathrm{fused}} &= \boldsymbol{H}_T + s_{\mathrm{phy}} g^f_{\mathrm{phy}} \Delta \boldsymbol{H}_p + s_K g^f_K \Delta \boldsymbol{H}_K, \\ \boldsymbol{Z}_{\mathrm{feat}} &= h_\psi(\boldsymbol{H}_{\mathrm{fused}}, \mathcal{C}_T). \end{aligned} \tag{11}$$

Here, MHA denotes multi-head attention with query, key, and value arguments, respectively.

The physical feature gate uses the temporal summary, physical reliability, and base confidence. The knowledge feature gate uses the temporal summary, coordinated reliability, source retrieval reliability, prior agreement, collaborative retrieval validity, top similarity, retrieval margin, base confidence, and normalized base entropy. The strengths s_{phy} and s_K control the scale of the residual updates. This gated residual formulation allows knowledge contributions to be reduced when their estimated reliability is low.

At the decision level, KGCFM employs source-specific reliability gates together with a collaborative knowledge gate. For $j \in \{\mathrm{phy}, \mathrm{ret}, K\}$ and learned gate functions G_j, the effective gate is

$$g_j = r_j \, \sigma\left(G_j\left(\boldsymbol{h}_j, \boldsymbol{p}_{\mathrm{base}}, c_{\mathrm{base}}, u_{\mathrm{base}}, r_j, \boldsymbol{\xi}_j\right)\right). \tag{12}$$

Here, σ is the sigmoid and $\boldsymbol{\xi}_j$ contains source-specific quality cues. In the source gates, $\boldsymbol{h}_{\mathrm{phy}} = \boldsymbol{h}^d_{\mathrm{phy}}$, $\boldsymbol{h}_{\mathrm{ret}} = \boldsymbol{h}^0_{\mathrm{ret}}$, and $r_{\mathrm{ret}} = r^0_{\mathrm{ret}}$. The collaborative gate uses $\boldsymbol{h}_K$, r_K, and prior agreement.

The source-specific and collaborative decision corrections are

$$\begin{aligned} \Delta \boldsymbol{z}_{\mathrm{src}} &= \alpha_{\mathrm{phy}} g_{\mathrm{phy}} \boldsymbol{z}^0_{\mathrm{phy}} + \alpha_{\mathrm{ret}} g_{\mathrm{ret}} \boldsymbol{z}^0_{\mathrm{ret}}, \\ \Delta \boldsymbol{z}_{\mathrm{collab}} &= \alpha_K g_K \boldsymbol{z}_K. \end{aligned} \tag{13}$$

where α_{phy}, α_{ret}, and α_K control the corresponding decision contributions. This decomposition allows reliable individual sources to contribute directly while the coordinated pathway captures cross-source interactions.

For Type-III ELM and Large ELM, a window-level safeguard is triggered when the base prediction supports the minority group but coordinated knowledge assigns it substantially less probability. Let $\mathcal{Y}_m = \{2, 3\}$, $\widehat{y}^{\mathrm{base}}_b = \arg\max_k p_{\mathrm{base},b,k}$, and $p_{*,b}(\mathcal{Y}_m) = \sum_{k \in \mathcal{Y}_m} p_{*,b,k}$ for $* \in \{\mathrm{base}, K\}$. The protection indicator is

$$\chi_b = \mathbb{I}\left[\widehat{y}^{\mathrm{base}}_b \in \mathcal{Y}_m \text{ or } p_{\mathrm{base},b}(\mathcal{Y}_m) \geq \tau\right] \mathbb{I}\left[p_{K,b}(\mathcal{Y}_m) + m < p_{\mathrm{base},b}(\mathcal{Y}_m)\right]. \tag{14}$$

The collaborative correction is then attenuated and combined with the source-specific correction:

$$\begin{aligned} \gamma_b &= 1 - (1 - \gamma_{\mathrm{protect}})\chi_b, \\ \widetilde{\Delta \boldsymbol{z}}_{\mathrm{collab},b} &= \gamma_b \Delta \boldsymbol{z}_{\mathrm{collab},b}, \\ \Delta \boldsymbol{z}_{\mathrm{KG}} &= \Delta \boldsymbol{z}_{\mathrm{src}} + s_{\mathrm{C}} \widetilde{\Delta \boldsymbol{z}}_{\mathrm{collab}}, \end{aligned} \tag{15}$$

where τ is the base-support threshold and m the probability-gap margin, $\gamma_{\text{protect}} \in [0, 1]$ controls attenuation, and s_{C} controls collaborative strength. The scalar γ_b scales the entire collaborative vector, while $\Delta \boldsymbol{z}_{\text{src}}$ is unchanged.

Finally, the combined decision correction is broadcast over the temporal dimension and added to the feature-level logits:

$$\begin{aligned} \boldsymbol{Z}_{\text{final}} &= \boldsymbol{Z}_{\text{feat}} + \lambda_{\text{dec}} \, \text{Broadcast}(\Delta \boldsymbol{z}_{\text{KG}}), \\ \boldsymbol{P}_{\text{final}} &= \text{Softmax}(\boldsymbol{Z}_{\text{final}}). \end{aligned} \tag{16}$$

Here, λ_{dec} scales the window-level class correction. The feature pathway provides temporally resolved logits, while the broadcast correction adjusts class preferences across the window. This separates local representation enhancement from contextual decision adjustment. The safeguard acts on the collaborative correction according to the combined support of the minority group.

4.4 Training Objective

KGF-ELM is optimized with a composite objective for point-wise classification, auxiliary knowledge supervision, and fusion regularization:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{know}} + \mathcal{L}_{\text{safe}}. \tag{17}$$

Here, $\mathcal{L}_{\text{cls}}$ combines class-balanced cross-entropy, focal loss, and Soft Macro-F1 loss. $\mathcal{L}_{\text{know}}$ supervises knowledge heads and guides window-averaged base logits; $\mathcal{L}_{\text{safe}}$ combines gate supervision, gate regularization, prediction consistency, and a minority-class penalty. Learnable encoders and fusion modules are optimized jointly; rule definitions and stored knowledge resources remain fixed. Definitions and coefficients are given in Section E.

5 Experiments

5.1 Experimental Setup

Experiments use EAST-ELM640 (Si et al., 2025), with 640 discharges and 18 diagnostic signals. The 448/96/96 training/validation/test sets are verified to be disjoint at the discharge level; windows of length $T = 100$ are extracted after splitting. Knowledge statistics, historical memory, and the retrieval index use 100 training discharges, which also participate in temporal-model training.

All models use the same input signals and discharge split. Training uses AdamW with batch size 192, learning rate 10^{-4}, cosine annealing, at most 12 epochs, and early-stopping patience three. For each run, checkpoint selection is based exclusively on validation Macro-F1, and the checkpoint with the highest validation Macro-F1 is subsequently evaluated on the test set. The main tables report the test performance of these validation-selected checkpoints from the historical experiments. Three-seed summaries for the full CrossLinear and Timer-XL models are reported in Appendix G.1. Experiments run on an NVIDIA RTX 4090. Macro-F1 is the primary metric because it gives equal weight to each class's precision–recall balance under class imbalance. Accuracy, Macro-Precision, Macro-Recall, and one-vs-rest Macro-AUC provide complementary summaries. Precision and recall distinguish false-positive and false-negative behavior; their operational costs depend on the downstream application. Offline evaluation counts overlapping window positions separately (Section A). CrossLinear is the default backbone; cross-backbone experiments use the same fusion design.

5.2 Main Results

Table 1 shows that KGF-ELM leads the compared models in Accuracy, Macro-Precision, Macro-F1, and Macro-AUC, reaching 91.7%, 82.9%, 81.6%, and 98.0%, respectively. Against CrossLinear, the strongest baseline in Macro-F1, knowledge fusion raises Macro-F1 by 2.4 points and Macro-Precision by 2.7 points while also improving recall. TimeXer has the highest Macro-Recall in this table; KGF-ELM combines stronger precision with the highest Macro-F1, demonstrating improved class-balanced recognition.

Table 1: Single-run test performance of validation-selected checkpoints on EAST-ELM640 (%). P, R, F1, and AUC are macro-averaged. Best and second-best reported results are bold and underlined, respectively.

Model	Acc.	Macro-P	Macro-R	Macro-F1	Macro-AUC
KGF-ELM	**91.7**	**82.9**	<u>80.5</u>	**81.6**	**98.0**
Timer-XL	88.6	78.8	76.8	77.3	96.5
TimeFilter	86.0	72.9	76.6	74.7	95.0
CrossLinear	<u>90.1</u>	<u>80.2</u>	79.2	<u>79.2</u>	<u>97.5</u>
Timer	82.9	73.4	64.0	68.0	92.0
iTransformer	84.7	72.3	70.3	70.8	94.5
TimeMixer	85.5	80.0	52.7	56.9	86.9
TimeXer	88.0	75.4	**81.7**	77.8	96.4
SOFTS	86.7	77.6	69.5	72.7	95.2
PatchTST	85.8	75.5	72.9	73.2	94.9

Table 2: Single-run test performance of temporal backbones with and without KGCFM using validation-selected checkpoints on EAST-ELM640 (%). P, R, F1, and AUC denote macro-averaged metrics. The better result within each backbone pair is shown in bold.

Backbone	Variant	Acc.	Macro-P	Macro-R	Macro-F1	Macro-AUC
Timer-XL	Base	88.6	**78.8**	76.8	77.3	96.5
	+KGCFM	**90.9**	77.4	**83.4**	**79.8**	**97.6**
CrossLinear	Base	90.1	80.2	79.2	79.2	97.5
	+KGCFM	**91.7**	**82.9**	**80.5**	**81.6**	**98.0**
Timer	Base	82.9	73.4	64.0	68.0	92.0
	+KGCFM	**89.9**	**77.6**	**79.5**	**78.4**	**96.9**
TimeXer	Base	88.0	75.4	81.7	77.8	96.4
	+KGCFM	**90.5**	**78.2**	**84.5**	**80.8**	**97.6**
TimeFilter	Base	86.0	72.9	76.6	74.7	95.0
	+KGCFM	**91.1**	**78.4**	**82.1**	**80.1**	**97.3**

5.3 Cross-Backbone Generality

KGCFM improves Macro-F1 on all five backbones by 2.4–10.4 percentage points, averaging 4.7 points (Table 2). Timer gains the most, from 68.0% to 78.4%, while CrossLinear reaches the highest final score of 81.6%. The common improvement across these architectures demonstrates that the knowledge interface complements different temporal representations.

Accuracy, Macro-Recall, and Macro-AUC also improve for every backbone. Timer-XL gains 6.6 points in Macro-Recall with a 1.4-point reduction in Macro-Precision, yielding a 2.5-point Macro-F1 increase. The other four backbones improve both precision and recall. Thus, the framework improves class-balanced recognition through backbone-dependent changes in the precision–recall balance.

5.4 Ablation Study

The complete model achieves the highest Macro-F1 among the evaluated variants on both Timer-XL and CrossLinear (Table 3). Relative to the complete configuration, the knowledge-free variants are lower by 2.5 and 2.4 points, the variants without feature fusion by 1.7 and 2.3 points, and those without decision fusion by 0.3 and 1.5 points. The results support combining representation enhancement with decision-level knowledge in the evaluated configurations.

Among knowledge-component variants, excluding physical evidence or historical retrieval gives the largest Macro-F1 reductions: $1.1/1.9$ and $0.9/2.0$ points on Timer-XL/CrossLinear, respectively. Excluding prototypes reduces the scores by $0.5/0.8$ points, and excluding the graph by $0.6/0.4$ points. Disabling reliability-aware control yields reductions of $0.2/0.4$ points. The complete model

Table 3: Ablation results in Macro-F1 (%) on Timer-XL and CrossLinear.

Overall modules			*KGCFM components*		
Variant	Timer-XL	CrossLinear	Variant	Timer-XL	CrossLinear
w/o Knowledge	77.3	79.2	w/o Physical	78.7	79.7
w/o Feature	78.1	79.3	w/o Prototype	79.3	80.8
w/o Decision	79.5	80.1	w/o Graph	79.2	81.2
w/o Reliability	79.6	81.2	w/o Historical	78.9	79.6
Full Model	**79.8**	**81.6**	Full Model	**79.8**	**81.6**

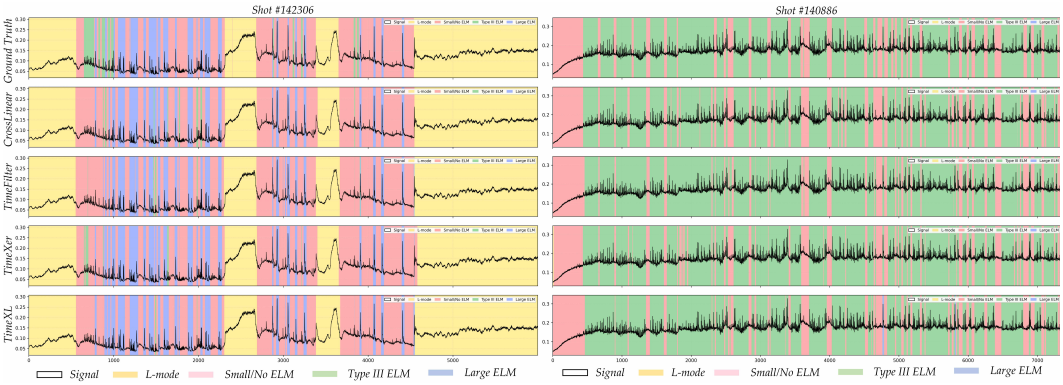

Figure 4: Qualitative comparison of point-wise ELM recognition on two selected test discharges. The first row shows the ground-truth labels, followed by CrossLinear, TimeFilter, TimeXer, and Timer-XL predictions, each with KGCFM. The black curves denote the *dau2* diagnostic signal, while the background colors represent the four ELM-related plasma states.

outperforms every evaluated variant on both backbones, supporting the combined use of physical, relational, and historical knowledge.

5.5 Qualitative Analysis

Figure 4 illustrates complementary state patterns on two test discharges. The left example combines extended L-mode intervals with repeated Small/No ELM–Large ELM transitions; the right contains frequent Type-III ELM–Small/No ELM transitions. Across four KGCFM-enhanced backbones, predictions track the broad state evolution and resolve many short intervals. Differences are concentrated around event durations and transition boundaries. These examples characterize the enhanced models' temporal behavior, complementing the paired backbone comparisons in Table 2.

6 Conclusion

We present KGF-ELM, a framework centered on KGCFM for integrating domain recognition knowledge with point-wise temporal modeling. Physical evidence, prototype rules, relational structure, and historical cases guide prediction through reliability-aware feature enhancement and gated decision correction.

Experiments on EAST-ELM640 show Macro-F1 gains across five backbones, averaging 4.7 percentage points and reaching 81.6% with CrossLinear. The complete model achieves the highest reported Macro-F1 among the evaluated ablation variants, while sensitivity analyses examine knowledge-resource configurations and fusion strength. These results support structured domain knowledge as an effective complement to temporal representations for ELM-related plasma-state recognition.

AI USE STATEMENT

We used generative AI tools, including ChatGPT/Codex, to assist with translation, manuscript editing, LaTeX formatting, and checking consistency among the manuscript text, equations, and reported experimental results. The authors are responsible for verifying AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

Sections 3–4 describe the model, and Section 5.1 specifies the experimental protocol. The appendix details preprocessing, knowledge construction, fusion, losses, and implementation. The experimental dataset cannot be publicly released. Source code can be provided by the authors upon request.

REFERENCES

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Dazhao Du, Tao Han, and Song Guo. Predicting the future by retrieving the past. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):20896–20904, 2026. doi: 10.1609/aaai.v40i25.39230. URL https://ojs.aaai.org/index.php/AAAI/article/view/39230.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. SOFTS: Efficient multivariate time series forecasting with series-core fusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/754612bde73a8b65ad8743f1f6d8ddf6-Abstract-Conference.html.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21774–21797. PMLR, 2025. URL https://proceedings.mlr.press/v267/han25d.html.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. *arXiv preprint arXiv:2501.13041*, 2025.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568(7753):526–531, 2019.

Anthony W Leonard. Edge-localized-modes in tokamaks. *Physics of Plasmas*, 21(9), 2014.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024b.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 83982–84006, 2025.

A Loarte, G Saibene, R Sartori, M Becoulet, L Horton, T Eich, A Herrmann, M Laux, G Matthews, S Jachmich, et al. Elm energy and particle losses and their extrapolation to burning plasma experiments. *Journal of Nuclear Materials*, 313:962–966, 2003.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Finn H O'Shea, Semin Joung, David R Smith, and Ryan Coffee. Automatic identification of edge localized modes in the diii-d tokamak. *APL Machine Learning*, 1(2), 2023.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378(C), 2018.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Hao Si, Xiao Wang, Fan Zhang, Xiaoya Zhou, Dengdi Sun, Wanli Lyu, Qingquan Yang, and Jin Tang. Hgts-former: Hierarchical hypergraph transformer for multivariate time series analysis. *arXiv preprint arXiv:2508.02411*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

NX Teo, VH Hall-Chen, K Barada, RH Ng, L Gu, AK Yeoh, QT Pratt, X Garbet, and TL Rhodes. Using convolutional neural networks to detect edge localized modes in diii-d from doppler backscattering measurements. *Review of Scientific Instruments*, 95(7), 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

HQ Wang, GS Xu, HY Guo, BN Wan, L Wang, R Chen, SY Ding, N Yan, XZ Gong, SC Liu, et al. Edge-coherent-mode nature of the small edge localized modes in experimental advanced superconducting tokamak. *Physics of Plasmas*, 21(9), 2014.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a7ac8a21e5a27e7ab31a5f42a0117bdb-Abstract-Conference.html.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498, 2024b.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4120–4131, 2025.

Hartmut Zohm. Edge localized modes (elms). *Plasma Physics and Controlled Fusion*, 38(2):105–128, 1996.

APPENDIX

This appendix details data processing, knowledge resources, retrieval, fusion, training, and supplementary analyses. Superscripts 0 and c denote source-specific and collaboratively adapted decision outputs, respectively.

A DATA PROCESSING AND EVALUATION

EAST-ELM640 contains 640 discharges, partitioned into 448 training, 96 validation, and 96 test discharges. Discharge identifiers were checked to verify pairwise disjointness before extracting windows. The 100 knowledge-construction discharges are a subset of the training set and also participate in temporal-model training. Labels 0–3 denote L-mode, Small/No ELM, Type-III ELM, and Large ELM. Small/No ELM combines small-ELM and no-ELM intervals; the labels define the four recognition categories used throughout the experiments.

The 18 channels, in implementation order, are `H98`, `q95`, `betap`, `pcrl01`, `tritop`, `nel`, `wmhd`, `point_n1`, `drsep`, `dau2`, `daL2`, `pxuv58`, `pxuv7`, `sxr20u`, `g106`, `pxuv32`, `eng`, and `cmph5t`. The loader uses the channels in each row of the supplied discharge records. It converts entries to numeric values, rounds `wmhd` to three decimals, replaces infinities with missing values, and applies forward filling, backward filling, then zero filling within each discharge. This loader adds no temporal resampling, channel-lag correction, or explicit denoising filter. It also adds no dataset-level standardization; any backbone-specific input transformation is part of that model.

The window length is $T = 100$ samples and the stride is one sample. A discharge of length $L \geq T$ contributes $L-T+1$ windows. The primary classification loss uses all T point-wise labels, including minority-class positions within mixed-state windows. Window-level auxiliary classifiers use the majority label, with ties resolved by the smallest class index. These auxiliary terms learn a window-level class summary, while point-wise supervision trains the temporally resolved output. The soft knowledge-guidance targets are rule or model distributions, as specified in Section E.

The reported evaluation pools predictions over windows and time steps. Its unit is a time point observed within a particular window context. For zero-based index t in a discharge of length $L \geq T$, the number of occurrences is $m(t) = \max\{0, \min(t, L-T) - \max(0, t-T+1) + 1\}$; interior points with complete coverage occur T times. Accuracy is computed over these window positions, Macro-Precision, Macro-Recall, and Macro-F1 average the four class-wise scores, and Macro-AUC averages one-vs-rest AUCs. All compared models use the same protocol. Overlapping predictions are retained separately, so the expanded position count represents dependent window observations.

B WINDOW-BASED PHYSICAL EVIDENCE AND PROTOTYPES

B.1 DIAGNOSTIC EVIDENCE

For each observed window, the physics analyzer computes diagnostic evidence directly from the input signals. This runtime computation uses window statistics and event-aligned context; labels enter only the training objectives and offline resource construction. Let $\epsilon = 10^{-6}$, $[x]_a^b = \min(b, \max(a, x))$, and $\langle x \rangle$ denote a temporal mean. For a diagnostic sequence x, the robust standardized signal is

$$z_t(x) = \frac{x_t - \mathrm{median}(x)}{1.4826\,\mathrm{median}(|x - \mathrm{median}(x)|) + \epsilon}. \tag{A1}$$

The median and median absolute deviation use `torch.nanmedian` along the time axis. For an even number of finite samples, this selects the lower of the two central order statistics.

Let $z_t^u = z_t(\mathtt{dau2})$ and $z_t^l = z_t(\mathtt{daL2})$. The nonnegative upper-channel peak is $A = \max(0, \max_t z_t^u)$, the burst scores are $b = [A/8]_0^1$ and $b_l = [\max_t z_t^l/8]_0^1$, and their agreement is $c_D = [1 - |b - b_l|]_0^1$. The peak-occupancy and frequent-activity cues are defined below. They summarize threshold-exceeding samples; they are not calibrated estimates of ELM frequency in hertz.

$$d = \left[\frac{8}{T}\sum_t \mathbb{I}[z_t^u > 3]\right]_0^1, \qquad f = \left[\frac{1}{6}\sum_t \mathbb{I}[z_t^u > 3]\right]_0^1 \left(1 - [A/15]_0^{0.6}\right). \tag{A2}$$

The confinement-context scores are $H = 0.75[(\langle \texttt{H98} \rangle - 0.85)/(0.25+\epsilon)]_0^1 + 0.25[(\langle \texttt{wmhd} \rangle - 8 \times 10^4)/(7 \times 10^4 + \epsilon)]_0^1$ and $L = 1 - H$. They provide soft context for the prototype rules; they are not confinement labels or standalone L–H/H–L transition detectors.

B.2 Event-aligned density proxy

Let $t_\star = \arg\max_t z_t^u$ and $n_t = \texttt{point_n1}_t$. The density calculation requires $\max_t z_t^u \geq 3$, $t_\star - 40 \geq 0$, and $t_\star + 35 < T$, using zero-based indexing. With half-open sample intervals, it compares the 65th percentile before the event with the 35th percentile after the event:

$$\begin{aligned} n_{\text{pre}} &= Q_{0.65}(n_{[t_\star - 40, t_\star - 5)}), \qquad n_{\text{post}} = Q_{0.35}(n_{[t_\star + 3, t_\star + 35)}), \\ D &= \max\left(100\frac{n_{\text{pre}} - n_{\text{post}}}{|n_{\text{pre}}| + \epsilon}, 0\right), \qquad \ell = [D/25]_0^1. \end{aligned} \tag{A3}$$

The proxy uses an event peak with sufficient pre- and post-event context, corresponding to $40 \leq t_\star \leq 64$ for $T = 100$. It returns $D = 0$ when the peak or context is unavailable, a quantile is nonfinite, or $n_{\text{pre}} \leq \epsilon$. Available decreases are grouped as weak ($2 \leq D < 7$), moderate ($7 \leq D < 15$), and strong ($D \geq 15$). These thresholds describe a diagnostic density-decrease proxy; the four recognition labels remain defined by the dataset. The accompanying density-quality code records the event/context status used by the physics decision encoder.

The remaining cues are post-imputation finiteness Q, drift $v = \mathbb{I}[(\max n - \min n)/(|\langle n \rangle| + \epsilon) > 0.20](1 - \ell)$, and mixed evidence $m = [0.3\min(b, L) + 0.4\min(H, \ell)]_0^1$. For finite loader outputs, $Q = 1$ and the quality contribution to physical reliability becomes $0.45(1 - v)$. The query evidence vector is $\boldsymbol{e} = [b, d, \ell, H, L, f, 0, 0, m, Q, v]$. Its boundary and recovery coordinates, together with prototype coordinates 13 and 14, are set to zero at runtime. The common vocabulary also represents historical case fields and supplies the node identities used by the relational encoder.

B.3 Prototype activation and physical reliability

The 15 prototype activations are deterministic soft rules. Let U_i denote their raw scores in Table A1; the descriptor stores $a_i = [U_i]_0^1$. The class-support calculation uses the raw scores before this clipping. The last four prototypes represent ambiguity or quality conditions rather than additional output classes.

Table A1: Prototype scores computed from the current input window.

i	Prototype interpretation	Raw score U_i
1	L-mode, weak burst and loss	$L(1-b)(1-\ell)$
2	L-mode, ambiguous transition	$L(0.5 + 0.5m)$
3	Stable H-mode, weak loss	$H(1-\ell)(1-0.5f)$
4	Frequent small activity, weak loss	$Hf(1-\ell)$
5	Weak burst without loss	$H[1 - \lvert b - 0.4\rvert]_0^1(1-\ell)$
6	Type-III-like burst and loss	$H[1 - \lvert b - 0.55\rvert]_0^1(0.4+\ell)$
7	Type-III-like frequent activity	$Hf(0.2+\ell)$
8	Type-III-like mixed transition	$m(0.3+\ell)$
9	Strong burst and density decrease	$Hb\ell$
10	Large sparse burst and decrease	$H[A/12]_0^1\ell(1-0.5d)$
11	Large burst, weak density evidence	$Hb(0.2+0.3\ell)$
12	Mixed evidence	m
13	Recovery tail	0
14	Rule–label conflict	0
15	Finiteness or drift cue	$[1 - Q + v]_0^1$

$$
\begin{aligned}
\boldsymbol{v}_s &= [0.80U_1 + 0.45U_2,\ 0.70U_3 + 0.75U_4 + 0.55U_5,\\
&\qquad 0.70U_6 + 0.70U_7 + 0.50U_8,\ 0.90U_9 + 0.75U_{10} + 0.40U_{11}],\\
s_k &= \frac{v_{s,k} + 0.03}{\sum_{j=0}^{3}(v_{s,j} + 0.03)}, \qquad \boldsymbol{q} = [\boldsymbol{e}; \boldsymbol{a}; \boldsymbol{s}] \in \mathbb{R}^{30}.
\end{aligned}
\tag{A4}
$$

Physical reliability combines post-imputation finiteness, drift, event-aligned density evidence, burst strength, and confinement context into a fusion-control score. Let I_D indicate an event-aligned density decrease of at least 2%. The score is

$$
r_{\mathrm{phy}} = [0.45Q(1-v) + 0.25I_D + 0.20b + 0.10H]_0^1 \begin{cases} 0.55, & b < 0.15 \text{ and } \ell < 0.05, \\ 1, & \text{otherwise.} \end{cases}
\tag{A5}
$$

The query role starts at weak-matched (ID 1). It is updated to strong-clean (ID 0) when $\max_i a_i > 0.75$ and $r_{\mathrm{phy}} > 0.65$, to noisy-or-drift-dominated (ID 6) when $Q < 0.8$ or $v > 0.8$, and finally to explanation-only (ID 7) when $\max_i a_i < 0.25$. Later assignments take precedence. For finite preprocessed inputs, the noisy-or-drift condition reduces to $v > 0.8$. The shared role vocabulary additionally reserves IDs 2–5 and 8 for mixed evidence, label-boundary transition, recovery tail, rule–label conflict, and unknown, respectively.

The density vocabulary contains 11 IDs. Runtime assignment starts at unknown (ID 10) and follows Table A2. IDs 6–9 are retained for resource compatibility. Learned 8-dimensional role and density embeddings, together with r_{phy} and $[D/100]_0^1$, augment the 30-dimensional descriptor to form the physics decision encoder's 48-dimensional input.

Table A2: Density-quality vocabulary and runtime assignment. v_p and v_b denote a valid burst peak and sufficient event context.

ID	Meaning	Runtime condition
0	Strong aligned decrease	$v_p \wedge v_b,\ D \geq 15$
1	Moderate aligned decrease	$v_p \wedge v_b,\ 7 \leq D < 15$
2	Weak aligned decrease	$v_p \wedge v_b,\ 2 \leq D < 7$
3	Subthreshold / numerical fallback	$v_p \wedge v_b,\ D < 2$
4	No qualifying burst	$\neg v_p$
5	Insufficient event context	$v_p \wedge \neg v_b$
6	Slow-drift invalidity	Reserved
7	Unphysical-drop invalidity	Reserved
8	Nonpositive density reference	Reserved
9	No density loss	Reserved
10	Unknown	Initial fallback

The current routine initializes D to zero. Numerical or reference-density failures that leave $D = 0$ with valid peak and context conditions therefore receive ID 3; they are not separately assigned reserved IDs 6–9.

B.4 Local physical sequence

For feature fusion, define $u_t = [z_t^u/8]_{-3}^3$ and $l_t = [z_t^l/8]_{-3}^3$. The scaled density decrease is

$$
\delta_t = \left[-\frac{n_t - n_{t-1}}{\max(\mathrm{median}\,|n|, \epsilon)} \right]_0^3, \qquad \delta_0 = 0.
$$

The eight channels of S_{phy} are listed below. Invalid numerical values are sanitized before encoding.

$$
\begin{aligned}
S_{\mathrm{phy},t} = \big[& u_t, l_t, \mathbb{I}[z_t^u > 3], [1 - |u_t - l_t|/3]_0^1, \delta_t,\\
& \sigma(z_t^u - 3)\delta_t, [(\mathtt{H98}_t - 0.85)/0.25]_{-2}^2,\\
& [(\mathtt{wmhd}_t - 8 \times 10^4)/(7 \times 10^4)]_{-2}^2 \big].
\end{aligned}
\tag{A6}
$$

A point-wise multilayer perceptron maps these channels to the backbone hidden dimension, followed by adaptive average pooling to the required token count N. This operation provides a common tensor interface. For backbones whose token axis combines channels and patches, matching N should not be interpreted as an exact one-to-one physical alignment of every token with a unique time interval.

C Relational Knowledge and Historical Retrieval

The ELM domain-knowledge graph supplies the recognition vocabulary and rule associations. Its computational encoding has 30 nodes: 11 evidence, 15 prototype, and four class nodes, connected by 100 directed edges of types supports, contradicts, indicates, and associated-with. The rule formulas, graph topology and edge weights, stored case records, and retrieval index remain fixed during temporal-model training. The 100 construction discharges supply rule-evaluated cases rather than mined graph topology.

For each input window, evidence values and prototype activations are recomputed, and the query selects and weights historical cases. Node embeddings, relation-specific transformations, case encoders, and fusion networks are learned jointly with the temporal backbone. Contradiction is represented by a relation type with nonnegative edge weights and a learned transformation.

Node and node-type embeddings initialize a two-layer R-GCN. For node i, relation-specific messages are normalized by its total incoming edge weight across all relations. Omitting dropout from the notation, one layer has the following form, where $w_{ji} \geq 0$ and r_{ji} denotes the relation type:

$$\widehat{h}_i = \mathrm{LN}\left(W_0 h_i + b_0 + \frac{\sum_{j \to i} w_{ji} W_{r_{ji}} h_j}{\max(1, \sum_{j \to i} w_{ji})} \right), \qquad \widetilde{Z}_G = \sum_{i=1}^{30} \frac{q_i}{\max(\epsilon, \sum_j q_j)} H_{G,i}. \tag{A7}$$

GELU follows the first graph layer, and an additional layer normalization follows the second. The graph hidden dimension is 128; the pooled representation is projected to the backbone dimension. All 30 nodes participate in graph encoding. The nonnegative query descriptor $\boldsymbol{q}$, including inferred class support, then determines the pooling weights. A zero query coordinate removes that node's direct pooling contribution, while its messages remain part of the relational encoding. A linear head on the projected graph token supplies auxiliary graph supervision.

C.1 Explicit graph construction

Index evidence nodes as $e_1, \ldots, e_{11}$ in the order of $\boldsymbol{e}$, prototype nodes as $p_1, \ldots, p_{15}$ in Table A1, and class nodes by labels $k = 0, 1, 2, 3$. Table A3 lists all forward associations. A positive signed evidence coefficient creates a supports edge $e_j \to p_i$; a negative coefficient creates a contradicts edge, using the absolute coefficient as its nonnegative weight. These coefficients specify semantic associations, not derivatives of the nonlinear prototype formulas. Each listed prototype–class pair creates an indicates edge. Every forward edge of weight w has an associated-with reverse edge of weight $0.25w$. Thus 30 supports, nine contradicts, and 11 indicates edges produce 50 reverse edges, giving exactly 100 directed edges. These rule-schema coefficients encode the prescribed strengths of semantic associations and remain fixed during model training. The 0.25 reverse-edge multiplier provides weaker reverse links. Node embeddings and relation-specific R-GCN transformations are learned, allowing the encoded associations to interact with the classification objective.

C.2 Case-memory contract

The offline resource contains 28,688 window records constructed from 100 training discharges. Fixed-coverage windows use length 100 and stride 100, supplemented by event-centered and label-boundary windows. Each record stores diagnostic evidence, prototype scores, class support, a guidance weight, a role, and discharge/interval provenance. Training annotations support offline boundary handling and rule–label consistency assessment. Current-window queries are computed from signals without using query labels.

Table A3: Complete forward-edge specification of the computational graph. Reverse edges follow the rule given in the text. Coefficients are model settings and retain their specified precision.

Prototype	Signed evidence associations	Class association
1	$e_5 : +0.90,\ e_1 : -0.55,\ e_3 : -0.55$	$k = 0,\ w = 0.80$
2	$e_5 : +0.75,\ e_9 : +0.45$	$k = 0,\ w = 0.45$
3	$e_4 : +0.80,\ e_3 : -0.60,\ e_6 : -0.20$	$k = 1,\ w = 0.70$
4	$e_4 : +0.75,\ e_6 : +0.90,\ e_3 : -0.55$	$k = 1,\ w = 0.75$
5	$e_4 : +0.70,\ e_1 : +0.35,\ e_3 : -0.50$	$k = 1,\ w = 0.55$
6	$e_4 : +0.70,\ e_1 : +0.65,\ e_3 : +0.45$	$k = 2,\ w = 0.70$
7	$e_4 : +0.65,\ e_6 : +0.90,\ e_3 : +0.35$	$k = 2,\ w = 0.70$
8	$e_9 : +0.80,\ e_3 : +0.45$	$k = 2,\ w = 0.50$
9	$e_4 : +0.75,\ e_1 : +0.95,\ e_3 : +0.95$	$k = 3,\ w = 0.90$
10	$e_4 : +0.70,\ e_1 : +0.80,\ e_3 : +0.90,\ e_2 : -0.20$	$k = 3,\ w = 0.75$
11	$e_4 : +0.65,\ e_1 : +0.85,\ e_3 : +0.25$	$k = 3,\ w = 0.40$
12	$e_9 : +0.95,\ e_7 : +0.80$	—
13	$e_8 : +0.95$	—
14	$e_9 : +0.50,\ e_{10} : -0.20$	—
15	$e_{11} : +0.95,\ e_{10} : -0.80$	—

The retained index contains 6,493 cases from 96 of the 100 construction discharges: 1,047 strong-clean and 5,446 weak-matched cases. Stored majority annotations give class counts of 899, 5,385, 180, and 29; the argmax of stored class support gives 1,021, 5,394, 63, and 15. The bank preserves the selected cases' empirical composition rather than enforcing class quotas. Retrieval supplies diagnostic support and contextual representations; point-wise class-imbalance objectives are defined separately in Section E.

The archived index retains all strong-clean and weak-matched records in the construction resource. Each retained record has a fixed guidance weight

$$g_i = \begin{cases} 1.00, & \text{strong-clean}, \\ 0.25, & \text{weak-matched}. \end{cases}$$

These values are read from the stored records and enter candidate weighting and case-token construction. Annotation-mixed guidance targets in the offline records are separate from the class-support vectors used by retrieval. The index builder's default guidance threshold is zero, which retains this same set of accepted-role records.

Queries and historical cases share a 30-coordinate vocabulary. Query evidence coordinates 7 and 8 and prototype coordinates 13 and 14 are zero; historical records retain their stored values. Cosine search normalizes the complete descriptor without masking these coordinates. Historical-only values therefore contribute to the case-vector norm, while their direct dot-product contribution with a zero query coordinate is zero.

Descriptors are ℓ_2-normalized, and queries select the top $K_r = 5$ cases by cosine similarity. The default configuration permits training queries to retrieve cases from the same training discharge. Validation and test discharges are excluded from the historical memory, so their queries have no same-discharge candidates. Appendix G.2 examines training-time same-discharge exclusion as a separate control. The stride-one windows used for temporal learning and evaluation are separate from the offline resource-construction windows.

C.3 Candidate aggregation and reliability

The same cosine-selected candidates support two aggregations. For candidate i, let c_i be descriptor similarity, $a_i^{\cos}$ prototype cosine similarity, $\boldsymbol{s}_i$ stored class support, $v_i \in \{0, 1\}$ its accepted-role indicator, and $g_i \in [0, 1]$ its clipped guidance score. Using the adapted physical prior $\boldsymbol{p}^c_{\text{phy}}$ defined in Appendix D, the prior agreement and candidate weights are

$$
\begin{aligned}
A_i &= 1 - \frac{\mathrm{JS}(\boldsymbol{p}^c_{\mathrm{phy}}, \boldsymbol{s}_i)}{\log 2}, \\
w_i^0 &= v_i g_i \,\mathrm{Softmax}_i(10 c_i), \\
w_i^c &= v_i g_i \,\mathrm{Softmax}_i\left(\frac{c_i + 0.20 a_i^{\cos} + 0.20 A_i + 0.10 g_i}{0.10}\right), \\
\widetilde{\boldsymbol{p}}^{\nu}_{\mathrm{ret}} &= \frac{\sum_i w_i^{\nu} \boldsymbol{s}_i}{\sum_i w_i^{\nu}}, \qquad \nu \in \{0, c\}.
\end{aligned}
\tag{A8}
$$

The Jensen–Shannon divergence uses natural logarithms. If the weight sum is at most 10^{-6}, the aggregate support is replaced by the uniform distribution. Candidate weighting is performed after cosine top-K_r selection and cannot introduce a case outside that candidate set. In the implementation, the physical prior and query prototypes are detached when computing collaborative reweighting.

$$
\begin{aligned}
V^{\nu} &= \mathbb{I}[c_1 \geq 0.85]\mathbb{I}[c_1 - c_2 \geq 0.03]\mathbb{I}[\textstyle\sum_i w_i^{\nu} > 10^{-6}], \\
r^0_{\mathrm{ret}} &= V^0 \max_k \widetilde{p}^0_{\mathrm{ret},k}, \\
r^c_{\mathrm{ret}} &= V^c \max_k \widetilde{p}^c_{\mathrm{ret},k} \left(0.5 + 0.5 \frac{\sum_i w_i^c A_i}{\max(\epsilon, \sum_i w_i^c)}\right).
\end{aligned}
\tag{A9}
$$

Here $c_1 \geq c_2$ are the two highest descriptor similarities. When only one candidate is retrieved, the implementation uses c_1 itself as the margin. Thus, a $K_r = 1$ comparison also changes the effective margin test. Reliability gates the decision contribution; an invalid window-level retrieval does not by itself mask every case token in feature attention. The attention mask checks the candidate role and numerical similarity, while the graph token remains available.

Each case token is encoded from its 30-dimensional normalized descriptor and four scalars: similarity, guidance score, accepted-role indicator, and the unnormalized collaborative weight w_i^c. A two-layer MLP with GELU, dropout, and layer normalization maps this 34-dimensional input to dimension d. The resulting case tokens are concatenated with the projected graph token to form $K_{\mathrm{ctx}} \in \mathbb{R}^{B \times (K_r+1) \times d}$.

D Knowledge Coordination and Dual-Level Fusion

The implementation retains source-specific decision heads while learning separate collaborative adaptations. Let sg denote stop-gradient and write $h_{\mathrm{phy}} = \boldsymbol{h}^d_{\mathrm{phy}}$ for the physical decision representation in the main text. The physics encoder produces h_{phy} and z^0_{phy} from the augmented descriptor. A lower-channel adapter uses burst and agreement cues to obtain

$$
z^c_{\mathrm{phy}} = \mathrm{sg}(z^0_{\mathrm{phy}}) + A_D([b_l, c_D]), \qquad \boldsymbol{p}^{\nu}_{\mathrm{phy}} = \mathrm{Softmax}(z^{\nu}_{\mathrm{phy}}), \quad \nu \in \{0, c\}. \tag{A10}
$$

For retrieval, define the seven-dimensional vectors $u^{\nu} = [\widetilde{\boldsymbol{p}}^{\nu}_{\mathrm{ret}}; r^{\nu}_{\mathrm{ret}}; c_1; \mathrm{margin}]$. The source encoder produces $(h^0_{\mathrm{ret}}, z^0_{\mathrm{ret}})$ from u^0, while residual adapters account for the difference between collaborative and source aggregation:

$$
\begin{aligned}
\delta u &= \mathrm{sg}(u^c - u^0), \\
h^c_{\mathrm{ret}} &= \mathrm{sg}(h^0_{\mathrm{ret}}) + A_h(\delta u), \qquad z^c_{\mathrm{ret}} = \mathrm{sg}(z^0_{\mathrm{ret}}) + A_z(\delta u).
\end{aligned}
\tag{A11}
$$

The output layers of A_D, A_h, and A_z are initialized to zero. Source heads retain their own auxiliary supervision, while stop-gradient separates their representations from the residual adaptation paths. Consequently, the logits used in the source-specific correction are z^0_{phy} and z^0_{ret}, whereas the mixture inside the coordinator uses z^c_{phy} and z^c_{ret}. The corresponding priors are not interchangeable with the raw rule support $\boldsymbol{s}$ or raw retrieved support $\widetilde{\boldsymbol{p}}^{\nu}_{\mathrm{ret}}$.

The coordinator combines $\mathrm{sg}(h_{\mathrm{phy}})$, h^c_{ret}, their absolute difference and element-wise product, the two adapted priors, and nine quality cues. The quality cues are physical reliability, collaborative retrieval reliability, prior agreement, top similarity, retrieval margin, candidate prior agreement, prototype consistency, lower-channel burst support, and cross-channel agreement. It produces a hidden representation h_K, a mixing coefficient $\eta \in [0, 1]$, and the following logits and reliability:

$$\begin{aligned} c_{\mathrm{pr}} &= 1 - \mathrm{JS}(\boldsymbol{p}^c_{\mathrm{phy}}, \boldsymbol{p}^c_{\mathrm{ret}})/\log 2, \\ z_K &= \eta z^c_{\mathrm{phy}} + (1-\eta) z^c_{\mathrm{ret}} + W_I h_K, \\ \widehat{r}_K &= \sigma R([h_K; r_{\mathrm{phy}}; r^c_{\mathrm{ret}}; c_{\mathrm{pr}}; \mathrm{margin}]), \\ r_K &= \widehat{r}_K \left(0.25 + 0.75 \frac{r_{\mathrm{phy}} + r^c_{\mathrm{ret}}}{2}\right)(0.5 + 0.5 c_{\mathrm{pr}}). \end{aligned} \tag{A12}$$

The adapted priors share diagnostic inputs, and collaborative retrieval also uses the physical prior. The coordinator therefore combines their agreement with source reliability and retrieval-quality cues. Reliability is detached for feature-fusion inputs and auxiliary-loss weighting, while decision gating retains a differentiable dependence on the coordinator score.

D.1 Feature and decision gates

Let $\overline{H}_T$ denote mean-pooled temporal tokens, c_{base} the base confidence, and u_{base} its unnormalized entropy. The feature gates use the following inputs; the knowledge feature gate uses source retrieval reliability r^0_{ret} together with collaborative retrieval validity V^c.

$$\begin{aligned} g^f_{\mathrm{phy}} &= r_{\mathrm{phy}}\, \sigma F_{\mathrm{phy}}([\overline{H}_T; r_{\mathrm{phy}}; 1 - c_{\mathrm{base}}]), \\ g^f_K &= r_K(0.5 + 0.5 r^0_{\mathrm{ret}})\, \sigma F_K([\overline{H}_T; r_K; r^0_{\mathrm{ret}}; c_{\mathrm{pr}}; V^c; \\ &\qquad\qquad c_1; \mathrm{margin}; c_{\mathrm{base}}; u_{\mathrm{base}}/\log 4]), \\ s_{\mathrm{phy}} &= 0.20 \tanh(a^f_{\mathrm{phy}}), \qquad s_K = 0.20 \tanh(a^f_K). \end{aligned} \tag{A13}$$

The raw feature-strength parameters a^f_{phy} and a^f_K are initialized to 0.01. Thus, feature fusion starts with small, nonzero residual strengths. Its strengths are bounded in magnitude and may be signed. The temporal tokens query the graph–case context by cross-attention, and the shared prediction head maps the resulting H_{fused} to Z_{feat} as in the main text.

Decision gates receive detached base probabilities, their confidence and unnormalized entropy, a source representation, reliability, and a source-specific quality cue. For the physical, retrieval, and collaborative gates, these quality cues are normalized density decrease, top retrieval similarity, and prior agreement, respectively. With $\alpha_{\mathrm{phy}} = \alpha_{\mathrm{ret}} = \alpha_K = 0.02$, the full-fusion prediction is

$$\begin{aligned} \Delta z_{\mathrm{src}} &= \alpha_{\mathrm{phy}} g_{\mathrm{phy}} z^0_{\mathrm{phy}} + \alpha_{\mathrm{ret}} g_{\mathrm{ret}} z^0_{\mathrm{ret}}, \\ \Delta z_{\mathrm{KG}} &= \Delta z_{\mathrm{src}} + s_C \gamma \alpha_K g_K z_K, \\ Z_{\mathrm{final},t} &= Z_{\mathrm{feat},t} + \lambda_{\mathrm{dec}} \Delta z_{\mathrm{KG}}, \qquad s_C = [a_C]^{0.25}_0. \end{aligned} \tag{A14}$$

The collaborative-strength parameter a_C is initialized to zero. The default minority safeguard uses the group $\{2, 3\}$, base-support threshold $\tau = 0.25$, probability-gap margin $m = 0.05$, and attenuation factor $\gamma_{\mathrm{protect}} = 0$. As defined in the main text, the safeguard checks both base minority support and whether the coordinated prior reduces the combined probability of this group. It suppresses only the collaborative correction when triggered; source-specific and feature-level contributions remain active.

The correction is constant over time within a window, giving $Z_{\mathrm{final},t,k} - Z_{\mathrm{final},t',k} = Z_{\mathrm{feat},t,k} - Z_{\mathrm{feat},t',k}$ for each class k. It adjusts class preferences through a shared offset, while feature fusion supplies temporal variation. Gate and strength coefficients control the correction scale; the source-logit magnitude also contributes. The minority safeguard attenuates the collaborative vector according to group-level support, while retaining source-specific and feature contributions.

E EXPANDED TRAINING OBJECTIVE

We expand the three groups in $\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{know}} + \mathcal{L}_{\text{safe}}$. Let i index the BT positions in a minibatch, $y_i \in \{0, 1, 2, 3\}$ be the class label, $y_{i,k} = \mathbb{I}[y_i = k]$ its one-hot encoding, and $p_{i,k} = P_{\text{final},i,k}$. For class counts n_k from the original training sequences before overlapping-window expansion, the effective-number weights are

$$\widetilde{w}_k = \frac{1-\beta}{1-\beta^{\max(n_k,1)}}, \qquad w_k = \frac{\widetilde{w}_k}{\frac{1}{4}\sum_{j=0}^{3}\widetilde{w}_j}, \qquad \beta = 0.99999. \tag{A15}$$

$$\begin{aligned}
\mathcal{L}_{\text{CB}} &= -\frac{\sum_i w_{y_i} \log p_{i,y_i}}{\sum_i w_{y_i}}, \\
\mathcal{L}_{\text{focal}} &= -\frac{1}{BT}\sum_i (1-p_{i,y_i})^{1.5} \log p_{i,y_i}, \\
\mathcal{L}_{\text{F1}} &= 1 - \frac{1}{4}\sum_{k=0}^{3} \frac{2\sum_i p_{i,k} y_{i,k}}{\sum_i p_{i,k} + \sum_i y_{i,k} + \epsilon}, \\
\mathcal{L}_{\text{cls}} &= \mathcal{L}_{\text{CB}} + 0.30\mathcal{L}_{\text{focal}} + 0.20\mathcal{L}_{\text{F1}}.
\end{aligned} \tag{A16}$$

Class weights are applied to cross-entropy, whereas the focal term in this configuration is unweighted. Soft Macro-F1 is evaluated over the minibatch. For the window-level terms, let $\overline{Z}_j = T^{-1}\sum_t Z_{j,t}$ for $j \in \{\text{base}, \text{feat}, \text{final}\}$, and let y_b^w be the window majority target. Define a reliability-weighted average $\mathcal{W}_r[v] = \sum_b r_b v_b / (\sum_b r_b + \epsilon)$; the implementation returns zero for an effectively zero total weight.

The soft-guidance loss uses temperature $\vartheta = 2$ on the base mean logits only. The target distribution is clipped away from zero and renormalized, but is not temperature-scaled. Using $\mathcal{C}_r(z) = \mathcal{W}_r[\text{CE}(z_b, y_b^w)]$, define

$$\mathcal{D}_r(\boldsymbol{\pi}) = \mathcal{W}_r\left[\vartheta^2 \,\text{KL}\left(\boldsymbol{\pi}_b \,\middle\|\, \text{Softmax}(\overline{Z}_{\text{base},b}/\vartheta)\right)\right], \qquad \vartheta = 2. \tag{A17}$$

$$\begin{aligned}
\mathcal{L}_{\text{know}} = 0.001\{&\mathcal{D}_{r_{\text{phy}}}(\boldsymbol{s}) + \mathcal{D}_{r^0_{\text{ret}}}(\widetilde{\boldsymbol{p}}^0_{\text{ret}}) + \mathcal{D}_{r_K}(\text{sg}(\boldsymbol{p}_K))\} \\
&+ 0.005\mathcal{C}_{r_{\text{phy}}}(z^0_{\text{phy}}) + 0.005\mathcal{C}_{r^0_{\text{ret}}}(z^0_{\text{ret}}) \\
&+ 0.002\mathcal{C}_{r_K}(z_K) + 0.001\mathcal{C}_{r_K}(z_G) \\
&+ 0.001\mathcal{W}_{r_{\text{phy}} r^0_{\text{ret}}}[\text{JS}(\boldsymbol{p}^0_{\text{phy}}, \boldsymbol{p}^0_{\text{ret}})].
\end{aligned} \tag{A18}$$

Here z_G is the graph auxiliary-head output. The first two soft targets are rule support and raw source retrieval support. The consistency term compares source priors with superscript 0, whereas coordination uses adapted priors with superscript c. Reliability weights are detached in these losses. The soft-guidance terms supervise $\overline{Z}_{\text{base}}$, integrating knowledge into temporal learning alongside the feature- and decision-fusion pathways.

The gate-helpfulness target compares base and coordinated window-level cross-entropies. With zero comparison margin and temperature 0.5, it supervises the collaborative gate logit $\ell_{K,b}$ through binary cross-entropy with logits, encouraging stronger gating when the coordinated prediction has lower loss:

$$\begin{aligned}
t_b^{\text{help}} &= \text{sg}\left[\sigma\left(\frac{\text{CE}(\overline{Z}_{\text{base},b}, y_b^w) - \text{CE}(z_{K,b}, y_b^w)}{0.5}\right)\right], \\
\mathcal{L}_{\text{help}} &= \mathcal{W}_{r_K}[\text{BCEWithLogits}(\ell_{K,b}, t_b^{\text{help}})].
\end{aligned} \tag{A19}$$

Let $\overline{p}_j = \text{Softmax}(\overline{Z}_j)$ and $I_b^m = \mathbb{I}[y_b^w \in \{2, 3\}]$. For windows with a minority majority label, the training penalty discourages a reduction in that class's probability relative to the detached base prediction. At inference, the safeguard uses predicted support for the combined minority group. These mechanisms operate alongside point-wise classification supervision. The remaining regularizers are

$$
\begin{aligned}
\mathcal{L}_{\text{minority}} &= \frac{\sum_b I_b^m [\text{sg}(\overline{p}_{\text{base},b,y_b^w}) - \overline{p}_{\text{final},b,y_b^w}]_+}{\sum_b I_b^m + \epsilon}, \\
\mathcal{L}_{\text{preserve}} &= \mathcal{W}_{1-r_K}[\text{KL}(\text{sg}(\overline{p}_{\text{base},b}) \| \overline{p}_{\text{feat},b})], \\
\mathcal{L}_{\text{gate}} &= \frac{1}{B} \sum_b (g_{\text{phy},b}^2 + g_{\text{ret},b}^2 + \text{sg}(s_C) g_{K,b}^2), \\
\mathcal{L}_{\text{safe}} &= 0.01\mathcal{L}_{\text{help}} + 0.01\mathcal{L}_{\text{minority}} + 0.005\mathcal{L}_{\text{preserve}} + 0.0001\mathcal{L}_{\text{gate}}.
\end{aligned}
\tag{A20}
$$

F Implementation and Optimization

CrossLinear is the default temporal backbone in the main paper. KGCFM interfaces with each backbone through token features and its point-wise classification head; token layouts may differ across backbones. Table A4 specifies the CrossLinear full-fusion configuration with $\lambda_{\text{dec}} = 0.50$. Its channel–patch token layout contains $18 \times 5 = 90$ tokens for a 100-sample window and patch length 20. The CrossLinear mixing coefficients are trainable parameters initialized to the listed values.

Table A4: CrossLinear full-fusion implementation settings.

Setting	Value
Window length / stride	$100/1$
Batch size	192
Temporal hidden / feed-forward dimensions	$256/512$
Patch length	20
Channel–patch tokens	90
Graph layers / hidden dimension	$2/128$
Feature attention heads / dropout	$4/0.1$
Knowledge decision hidden dimension	128
CrossLinear mixing initialization	$\alpha = 1.0,\ \beta = 0.5$
Optimizer / initial learning rate	AdamW / 10^{-4}
Weight decay / gradient-norm limit	$0/1.0$
Maximum epochs / early-stopping patience	$12/3$
Learning-rate schedule	Cosine
Feature fusion starts	Epoch 1
Decision fusion starts	Epoch 3
Retrieval candidates / temperature	$5/0.10$
Decision-fusion scale	0.50
Hardware reported in the main paper	NVIDIA RTX 4090

Feature fusion is active from the first epoch, while the source and collaborative decision gates are disabled during the first two epochs and activated from the third. Auxiliary knowledge objectives remain available during this warm-up. The optimizer uses the combined loss above, with global gradient-norm clipping at 1.0. Validation Macro-F1 determines checkpoint selection, and the selected checkpoint is restored for final testing. At inference, the window-level physics analyzer and retrieval query use signals only.

The independent temporal baselines in the main comparisons are trained separately. In KGF-ELM, the internal base output shares the jointly optimized backbone and prediction head, providing the pre-fusion reference used by the gates and auxiliary objectives.

F.1 Computational environment

Experiments were implemented in Python 3.12.11 with PyTorch 2.5.1, CUDA 12.1, and cuDNN 9.1.0. NumPy 1.26.4 and pandas 2.2.3 supported data processing, and scikit-learn 1.5.2 provided evaluation metrics. Each training or evaluation run used one NVIDIA GeForce RTX 4090 GPU on a Linux server. Separate experiments could run on different GPUs; an individual run did not use distributed training. The default data loader used ten workers and pinned memory, with automatic mixed precision disabled.

G Additional Analysis of the Reported Results

Table A5 summarizes the changes computed from the main-paper backbone scores. Macro-F1 gains are 2.5, 2.4, 10.4, 3.0, and 5.4 percentage points, averaging 4.7 points to one decimal place. Accuracy, Macro-Recall, and Macro-AUC improve across all five architectures.

Table A5: Changes after adding KGCFM to each backbone, in percentage points.

Backbone	ΔAcc.	ΔMacro-P	ΔMacro-R	ΔMacro-F1	ΔMacro-AUC
Timer-XL	+2.3	-1.4	+6.6	+2.5	+1.1
CrossLinear	+1.6	+2.7	+1.3	+2.4	+0.5
Timer	+7.0	+4.2	+15.5	+10.4	+4.9
TimeXer	+2.5	+2.8	+2.8	+3.0	+1.2
TimeFilter	+5.1	+5.5	+5.5	+5.4	+2.3

The largest Macro-F1 gain occurs on Timer, while CrossLinear achieves the highest final score. Timer-XL's 6.6-point recall gain offsets a 1.4-point precision reduction; the other four backbones improve both. This pattern shows that a shared knowledge interface can complement different temporal representations.

The ablations further distinguish the observed contributions. The variants without feature fusion have larger score differences than those without decision fusion on both evaluated backbones. Among knowledge sources, physical evidence and historical cases show the largest differences, with prototypes, graph encoding, and reliability-aware control contributing additional gains in the reported configurations. Figure 4 complements these comparisons by illustrating the enhanced models' state sequences and transition behavior on two discharges.

G.1 Three-seed test evaluation

We report three full-model runs with random seeds 2024, 2025, and 2026 for CrossLinear and Timer-XL. Table A6 presents the individual test scores and their mean and sample standard deviation. Summary statistics are computed from the archived one-decimal scores and then rounded to one decimal place.

CrossLinear-based KGF-ELM achieves a mean Macro-F1 of 81.3% with a standard deviation of 0.4 percentage points; individual scores range from 80.9% to 81.6%. Timer-XL-based KGF-ELM achieves a mean Macro-F1 of 79.6% with a standard deviation of 0.2 percentage points; individual scores range from 79.5% to 79.8%. The single-run results reported in the main paper correspond to seed 2025 for CrossLinear and seed 2026 for Timer-XL.

Table A6: Test performance over three random seeds (%). The summary rows report the mean and sample standard deviation across seeds. Statistics are calculated from the archived one-decimal scores.

Backbone	Seed	Acc.	Macro-P	Macro-R	Macro-F1	Macro-AUC
CrossLinear	2024	91.5	80.5	81.7	80.9	97.8
	2025	91.7	82.9	80.5	81.6	98.0
	2026	91.4	81.9	81.1	81.3	97.9
	Mean ± SD	91.5 ± 0.2	81.8 ± 1.2	81.1 ± 0.6	81.3 ± 0.4	97.9 ± 0.1
Timer-XL	2024	90.2	77.4	82.6	79.5	97.3
	2025	90.5	76.8	84.0	79.5	97.5
	2026	90.9	77.4	83.4	79.8	97.6
	Mean ± SD	90.5 ± 0.4	77.2 ± 0.3	83.3 ± 0.7	79.6 ± 0.2	97.5 ± 0.2

Every recorded full-model run exceeds the corresponding reported temporal baseline in Macro-F1. Relative to the single-run baseline references of 79.2% for CrossLinear and 77.3% for Timer-XL, the mean full-model differences are approximately 2.1 and 2.3 percentage points, respectively. These results characterize full-model repeatability; the baseline references are not a paired three-seed baseline evaluation.

G.2 TRAINING-TIME SAME-DISCHARGE EXCLUSION

The historical case memory is constructed exclusively from training discharges, and the training, validation, and test discharge sets are mutually disjoint. Consequently, validation and test queries have no same-discharge cases in the memory. Applying a same-discharge exclusion criterion to these queries does not remove any additional candidates.

Training queries differ because their source discharges can also contribute historical cases. We therefore examine a configuration that excludes same-discharge cases during training retrieval. This changes the training retrieval context while preserving the separation between the historical memory and the validation and test discharges.

Table A7: Reported test performance with same-discharge cases excluded during training retrieval (%). Each row reports one ablation run.

Backbone	Acc.	Macro-P	Macro-R	Macro-F1	Macro-AUC
CrossLinear	91.1	78.8	83.7	80.8	97.9
Timer-XL	90.1	77.0	82.7	78.9	97.4

With same-discharge cases excluded during training retrieval, Macro-F1 reaches 80.8% on CrossLinear and 78.9% on Timer-XL. These scores exceed the reported temporal baseline references of 79.2% and 77.3% by 1.6 percentage points for both backbones. The observed gains therefore persist when training queries retrieve historical context exclusively from other discharges.

Because the memory contains only training discharges, same-discharge exclusion does not change candidate eligibility for validation or test queries. Its influence on test performance arises through the training retrieval context and the resulting learned parameters.

G.3 MINORITY-PROTECTION ABLATION

Table A8 reports test Macro-F1 with the collaborative minority-group safeguard disabled during training and evaluation. This configuration sets $\gamma_b = 1$ for all windows while retaining the minority-class training penalty $\mathcal{L}_{\text{minority}}$. Each backbone has one recorded ablation run.

Table A8: Test Macro-F1 with the collaborative minority-group safeguard disabled (%). The minority-class training penalty remains active. Each value corresponds to one ablation run.

Backbone	Macro-F1
CrossLinear	79.8
Timer-XL	78.6

The recorded ablation scores are 79.8% for CrossLinear and 78.6% for Timer-XL, below the corresponding full-model scores across the three seeds in Table A6. This pattern is consistent with a beneficial role for the collaborative minority-group safeguard in aggregate class-balanced recognition. Macro-F1 summarizes all four classes; class-specific precision and recall are needed to distinguish the effects on Type-III and Large ELM separately.

H SENSITIVITY ANALYSES

The sensitivity experiments use Timer-XL and follow the discharge-level split, input channels, window construction, optimizer, and validation-based checkpoint-selection protocol described in the main paper. Each configuration is trained for at most 12 epochs with early-stopping patience three. We vary the number of knowledge-construction discharges, the retrieved-candidate count, or the decision-fusion scale, while retaining the other configured settings. These analyses characterize sensitivity within the Timer-XL configuration; CrossLinear remains the default backbone for the main comparison. Scores are reported as percentages, and differences as percentage points.

Table A9: Effect of the number of knowledge-construction discharges on Timer-XL, with $K_r = 5$ and $\lambda_{\text{dec}} = 0.50$.

Discharges	Acc.	Macro-P	Macro-R	Macro-F1	Macro-AUC
25	88.8	78.7	77.5	77.6	96.6
50	89.1	78.5	78.3	77.9	96.7
75	90.0	78.0	80.8	78.8	97.2
100	90.9	77.4	83.4	79.8	97.6

H.1 KNOWLEDGE-RESOURCE SIZE

Among the evaluated resources, the 100-discharge configuration achieves the highest Macro-F1, 79.8%, compared with 77.6% for the 25-discharge configuration (Table A9). Accuracy, Macro-Recall, Macro-F1, and Macro-AUC increase across the reported configurations. The endpoint comparison gives a 5.9-point recall gain and a 1.3-point precision reduction, resulting in a 2.2-point Macro-F1 gain. These results characterize the constructed resources, whose case counts and composition depend on the selected training discharges.

H.2 RETRIEVAL-CONFIGURATION SENSITIVITY

Table A10: Effect of retrieval candidate count on Timer-XL using 100 knowledge-construction discharges and $\lambda_{\text{dec}} = 0.50$.

K_r	Acc.	Macro-P	Macro-R	Macro-F1	Macro-AUC
1	90.6	77.1	82.9	79.4	97.4
3	90.8	77.3	83.2	79.6	97.5
5	90.9	77.4	83.4	79.8	97.6
10	90.8	77.3	83.3	79.7	97.6

Macro-F1 rises from 79.4% at $K_r = 1$ to 79.8% at $K_r = 5$, then remains close at 79.7% with ten candidates (Table A10). The five-case configuration gives the highest score in this sweep, while increasing the candidate count to ten yields no further improvement. For $K_r = 1$, the reliability margin uses the top similarity itself; for multiple candidates, it uses the top-two similarity gap.

H.3 DECISION-FUSION SCALE

Table A11: Effect of decision-fusion scale on Timer-XL using 100 knowledge-construction discharges and $K_r = 5$.

λ_{dec}	Acc.	Macro-P	Macro-R	Macro-F1	Macro-AUC
0.00	90.1	79.3	79.6	79.4	97.3
0.25	90.6	77.0	83.0	79.6	97.4
0.50	90.9	77.4	83.4	79.8	97.6
0.75	90.3	76.7	83.2	79.0	97.5

The decision-scale sweep reaches its highest Macro-F1 of 79.8% at $\lambda_{\text{dec}} = 0.50$, compared with 79.4% at zero and 79.0% at 0.75 (Table A11). Moderate decision correction therefore gives a better result than either omitting the offset or applying the largest tested scale. At zero, the model retains feature fusion and its configured auxiliary objectives. This sweep reports its own trained configurations; the pathway variants are listed in Table 3.

I COMPUTATIONAL COST

We profiled the complete CrossLinear-based KGF-ELM model on one NVIDIA GeForce RTX 4090 using batch size 192. Physical analysis, graph encoding, historical retrieval, feature fusion, and

decision correction remained enabled. The recorded model-side measurement covers 100 timed batches, or 19,200 windows, as summarized in Table A12.

Table A12: Model-side batched inference cost of the complete CrossLinear-based KGF-ELM model.

Measurement	Value
Timed batches	100
Processed windows	19,200
Batch size	192
Timed duration	5.7 s
Throughput	3,350.6 windows/s
Amortized time per window	0.3 ms
Peak allocated GPU memory	144.9 MB

The complete model processes $3{,}350.6$ windows/s at batch size 192, with an amortized cost of 0.3 ms per window and peak allocated GPU memory of 144.9 MB. Rates are computed from the unrounded measurements before display rounding. These values characterize batched model-side inference, including all knowledge and fusion pathways. Diagnostic acquisition, signal synchronization, preprocessing, and communication lie outside this measurement.

J Instance-Level Retrieval and Prediction Behavior

J.1 Case protocol and probability summaries

Two test windows illustrate complementary outcomes of knowledge enhancement. Discharge 140086, window 10244–10343, is entirely L-mode. Discharge 142356, window 4092–4191, is a mixed transition window whose majority class is Type-III ELM with purity 0.41. Both queries satisfy the retrieval-validity criterion, and the displayed historical cases come from training discharges. These examples examine model behavior within selected windows and are not an additional aggregate benchmark.

The probability summaries in this case study average point-wise softmax probabilities. For stage $j \in \{\mathrm{base}, \mathrm{final}\}$, define

$$\pi_{j,b,k} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{Softmax}(Z_{j,b,t,:})_k. \tag{A21}$$

This summary differs from the gating distribution $\boldsymbol{p}_{\mathrm{base}} = \mathrm{Softmax}(T^{-1} \sum_t Z_{\mathrm{base},t})$ in the main method. Averaging probabilities and applying softmax to averaged logits generally give different results. The case-study class is $\arg\max_k \pi_{j,b,k}$; it does not imply that every point within a mixed window has that predicted class.

J.2 Physical support and retrieved cases

Both queries activate drift-dominated and low-confinement cues, together with noisy-or-drift and transitional prototypes. The recorded low-confinement activations are 0.943 and 0.904 for discharges 140086 and 142356, respectively. Their physical support is dominated by L-mode, even though the second window has a Type-III majority label. Table A13 reports the physical support and reliability summaries. High physics–retrieval agreement can therefore coexist with a mismatch to the majority label of a mixed window.

Table A14 lists the three highest-similarity cases for each query. Their source discharges differ from the test queries. The top similarities are 0.8699 and 0.8546; the corresponding gaps to the second candidate are approximately 0.034 and 0.036, both exceeding 0.03. Candidate weights need not decrease with cosine rank because reweighting also uses prototype consistency, physical-prior agreement, offline guidance, and admissibility.

Table A13: Recorded physical class support and reliability summaries for the two queries.

Quantity	140086	142356
L-mode support	0.721	0.680
Small/No ELM support	0.139	0.188
Type-III ELM support	0.098	0.096
Large ELM support	0.041	0.037
Physical reliability	0.129	0.110
Retrieval reliability	0.608	0.524
Prior agreement	0.973	0.947
Coordinated reliability	0.248	0.210

Table A14: Top-three retrieved cases. The last column gives the L-mode support of each historical case.

Query	Rank	Source	Window	Cosine	Weight	s_{L}
140086	1	140370	5420–5519	0.8699	0.0418	0.627
140086	2	140624	5791–5890	0.8359	0.0679	0.655
140086	3	140624	6000–6099	0.8184	0.0590	0.590
142356	1	140370	5420–5519	0.8546	0.0517	0.627
142356	2	140355	7000–7099	0.8186	0.0361	0.678
142356	3	140624	4235–4334	0.8124	0.0395	0.367

J.3 Reinforcement and unresolved ambiguity

Table A15 illustrates how fusion changes the mean point-wise class probabilities. In the homogeneous L-mode window, the mean L-mode probability rises from 98.1% to approximately 100.0%, reinforcing the already correct window summary.

Table A15: Mean point-wise class probabilities in two selected windows (%). Values are rounded independently to one decimal place.

Shot	Output	L-mode	Small/No	Type-III	Large
140086	Pre-fusion	98.1	1.9	0.0	0.0
140086	Final	100.0	0.0	0.0	0.0
142356	Pre-fusion	34.8	38.9	26.3	0.1
142356	Final	15.0	42.8	42.2	0.0

In the mixed window, Type-III probability rises from 26.3% to 42.2%, while L-mode probability falls from 34.8% to 15.0%. Small/No ELM remains slightly higher at 42.8%, leaving the window-summary argmax unchanged. The complete model therefore redistributes support toward Type-III while retaining ambiguity between the two competing states.

Together, the examples show confidence reinforcement in a homogeneous window and probability redistribution in a mixed window. Although the retrieved cases predominantly support L-mode, the mixed window's final output assigns most probability to Small/No and Type-III ELM. This contrast illustrates why the framework treats historical cases as contextual evidence to be integrated with temporal and physical representations, rather than as labels to copy. The aggregate value of the historical pathway is reflected in the lower Macro-F1 of the variants without retrieval in Table 3.

K Scope, Limitations, and Reproducibility

The evaluation establishes KGF-ELM's effectiveness for four-state recognition on EAST-ELM640 under the specified discharge split and overlapping-window protocol. Its diagnostic rules and historical resources are tailored to this setting. Reliability scores serve as fusion-control quantities rather than calibrated correctness probabilities. Window-level auxiliary targets and the group-level safeguard complement the point-wise prediction objective.

The current pipeline uses complete windows, including post-event context and backward filling when needed. Cross-device adaptation, causal streaming, and finer event-local protection provide concrete directions for extending the framework.

The dataset cannot be publicly released. Source code can be provided by the authors upon request. The appendix specifies the signal order, preprocessing, knowledge representations, fusion equations, objectives, and implementation settings. Numerical reproduction uses these specifications together with authorized data access and the matching split, resource, configuration, and checkpoint artifacts.