RAGMO-NET: REGIME-AWARE GRAPH MEMORY OPTIMIZATION NETWORK FOR EEG

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic graph representation methods have shown promise for modeling electroencephalography (EEG) connectivity, but existing approaches are limited by heuristic memory updates, costly matrix-valued recurrent state transitions, and a lack of principled adaptation under non-stationary dynamics. We propose **Regime-Aware Graph Memory Optimization Network (RAGMO-Net)**, a framework that adapts graph memory updates based on latent temporal regimes corresponding to evolving connectivity states. RAGMO-Net introduces a regime-conditioned affine graph-memory cell whose state transition is a scalar multiple of the identity, giving a constant-cost update per graph snapshot without matrix-valued gates or nonlinear state transitions. To provide a principled foundation for memory adaptation, we further formulate memory refinement as a regime-aware proximal refinement process, linking dynamic graph learning with non-stationary optimization. Experiments on both EEG and non-EEG datasets demonstrate consistent improvement over strong dynamic graph baselines. These results show that combining regime awareness with optimization-driven memory adaptation yields a scalable and robust approach for learning from non-stationary dynamic graphs.

1 INTRODUCTION

Graph-structured representations are important in electroencephalography (EEG), where multichannel brain activity and inter-channel connectivity provide discriminative cues for neurological disorder detection (Chakladar et al., 2025), cognitive workload prediction (Chakladar et al., 2022), and face perception analysis (Chakladar & Pal, 2024). An EEG dynamic graph is represented as a temporal sequence of graph snapshots, where EEG channels serve as nodes and time-varying channel relationships are represented as edges. However, these graphs are non-stationary, as temporal segments (e.g., event-related potential (ERP) peaks) may reflect transient structural shifts or disorder-related connectivity changes (Tang et al., 2024; Li et al., 2020; Lu et al., 2022). Recent EEG-based Graph Neural Network (GNN) studies further show that functional connectivity and temporal graph patterns are useful for classifying Alzheimer's disease (AD), Frontotemporal Dementia (FTD), and healthy controls (HC) (Eltrass et al., 2026). Therefore, EEG dynamic graph models should encode each graph snapshot while maintaining memory that preserves stable patterns and adapts to temporal changes (Ye et al., 2022).

Dynamic GNNs provide an important foundation for this problem because they are designed to learn from graph structures that evolve over time, making them suitable for modeling both snapshot-level graph representations and temporal memory dynamics (Pareja et al., 2020; Rossi et al., 2020). Temporal Graph Network (TGN) introduced a memory-based approach for learning from dynamic graphs represented by sequences of time-stamped events (Rossi et al., 2020). More recently, a method named PRES studied the scalability of memory-based dynamic Graph Neural Networks and highlighted the difficulty of training memory modules when temporal and structural dependencies require chronological processing (Su et al., 2024). In EEG, a regime can be interpreted as a latent brain-connectivity state, reflecting stable, transitional, or rapidly changing neural dynamics (Michel & Koenig, 2018), which closely relates to dynamic EEG graphs (Cheron et al., 2016; Chodvadiya & Suchithra, 2026). However, detecting temporal change alone is not enough for memory-based prediction; the model must also determine when to retain past information during stable periods and when to adapt quickly under regime shifts. Since evolving graph conditions require continual adaptation, online non-stationary optimization provides a natural framework for adaptive memory

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.

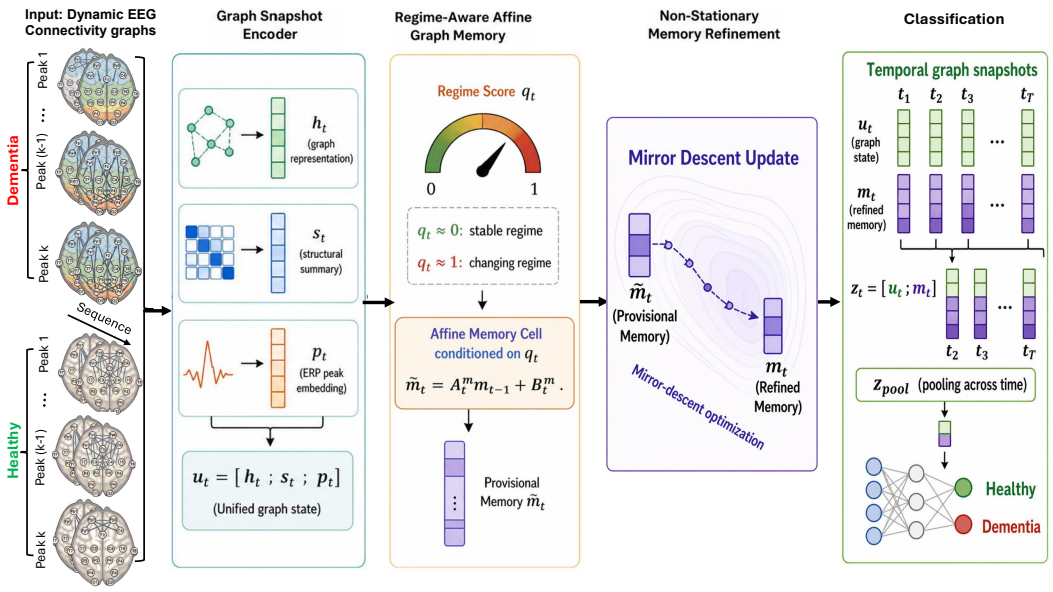


Figure 1: RAGMO-Net: Regime-Aware Graph Memory Optimization Network for EEG-based classification between healthy individuals and individuals with dementia.

modeling (Zhao et al., 2023; 2024). Bregman-proximal updates fit this online non-stationary setting by enabling stable memory adaptation under time-varying objectives (Yuan et al., 2020).

Despite these advances, three challenges remain. **First**, existing dynamic graph methods often use hand-crafted memory updates that are not explicitly conditioned on stable or changing regimes, limiting adaptation to transient or disorder-relevant shifts (Tang et al., 2024; Li et al., 2020; Ye et al., 2022). **Second**, many dynamic graph memory models update memory through matrix-valued recurrent cells or per-event message-passing modules, which incur high inference latency and computational cost as the number of graph snapshots grows (Pareja et al., 2020; Rossi et al., 2020; Su et al., 2024). **Third**, recurrent updates of dynamic graph memory models are often heuristic and lack an optimization-consistent explanation of temporal memory adaptation (Sulem et al., 2024; Xu et al., 2023; Cheron et al., 2016; Chodvadiya & Suchithra, 2026).

To address these limitations, we propose *Regime-Aware Graph Memory Optimization Network (RAGMO-Net)* for EEG-based dynamic graphs (Figure 1). For **Challenge 1**, RAGMO-Net encodes each EEG graph and estimates a regime score from structural drift and memory history, enabling regime-conditioned memory adaptation. For **Challenge 2**, it replaces the matrix-valued recurrent cell with a scalar-gated affine update, reducing the per-snapshot memory transition from a dense linear map to a single scalar multiplication (Table 4). For **Challenge 3**, it employs a regime-aware proximal refinement mechanism, offering an optimization-consistent explanation of memory adaptation under non-stationary EEG dynamics. The main contributions of this work are summarized as follows: 1) a regime-aware dynamic graph framework for modeling non-stationary EEG connectivity patterns; 2) an efficient scalar-gated affine memory update for low-latency temporal graph inference (Table 4), together with a proximal refinement strategy for online non-stationary graph-memory optimization; and 3) extensive experiments on EEG (Table 1) and non-EEG (Table 2) datasets demonstrate that RAGMO-Net improves classification performance and generalizes robustly across diverse domains.

2 Related Works

2.1 Temporal Graph Learning and Memory Mechanisms

Spatial graph convolution with attention-enhanced Bidirectional Long Short-Term Memory (BiLSTM) has been used to model EEG channel topology and temporal memory for emotion recognition (Feng et al., 2022). An EEG graph structure combined with recurrent spatial-temporal modules and a

Gaussian temporal encoder has also been developed to propagate graph information across time slices for emotion recognition (Liu et al., 2024). EEG-DGRN was proposed for subject-independent ERP detection, dynamically updating task-relevant EEG graphs over time to preserve event-related graph context (Zhu et al., 2025). Outside EEG, event-updated node embeddings have been used as memory in dynamic interaction graphs to model communication and association events for link and event-time prediction (Trivedi et al., 2019). Memory-augmented message-passing temporal graph networks have also been studied with a more expressive temporal architecture for event-based graph prediction tasks (Souza et al., 2022).

2.2 Non-stationary Optimization and Online Learning

Adaptive learning is essential because EEG signals are inherently non-stationary across and within sessions, while dynamic graphs evolve as timed events under shifting distributions (Raza et al., 2019; Rossi et al., 2020). To address non-stationarity, the time-varying graphical lasso uses proximal gradient optimization to recover sparse, evolving networks from high-dimensional non-stationary time series with theoretical consistency guarantees (Hallac et al., 2017). Non-stationary online convex optimization (OCO) has also been advanced through dynamic policy regret with switching-cost-aware ensembles and optimal dynamic regret guarantees (Zhao et al., 2023). Furthermore, Sword/Sword++ provides problem-dependent dynamic regret bounds based on gradient variation and small-loss quantities while retaining worst-case optimality (Zhao et al., 2024).

3 Method

The proposed **RAGMO-Net** framework (Figure 1) is divided into three main stages: (a) Regime-Aware Affine Graph Memory, (b) Non-Stationary Memory Refinement, and (c) Classification with RAGMO-Net.The pseudocode for RAGMO-Net and the notations used in the paper are provided in *Appendix* F and *Appendix* A, respectively. A detailed description of each subsection is provided below.

3.1 Regime-Aware Affine Graph Memory

ERP peaks were identified from the averaged epoch-wise EEG signal using global field power (GFP), which measures the standard deviation across EEG channels and reflects the overall strength of brain activation (Chakladar et al., 2021). For each detected ERP peak, we computed a correlation-based connectivity matrix from the multichannel EEG to construct a peak-wise connectivity graph. Let the input be an EEG-based dynamic graph sequence ($\mathcal{X}$) prepared from the ERP-peak index:

$$\mathcal{X} = \{(G_t, p_t)\}_{t=1}^{T}, \qquad G_t = (V, E_t, X_t, A_t), \tag{1}$$

where G_t is the graph at ERP peak index t, V denotes the EEG channels, E_t represents the connectivity edges, X_t contains node features (node strength, mean edge weight, maximum edge weight, and edge-weight standard deviation), and A_t is the connectivity adjacency matrix. Here, p_t denotes a learnable positional embedding for the discrete ERP peak index t. Each graph snapshot (G_t) is first encoded (via f_{enc}) into a graph representation and a graph summary:

$$h_t = f_{\text{enc}}(G_t, p_t), \qquad s_t = \phi(G_t), \qquad u_t = [h_t; s_t; p_t]. \tag{2}$$

Here, h_t captures the learned graph-level representation, s_t summarizes structural properties of the graph, and u_t is the final graph state used by the memory module. Although p_t is already an input to f_{enc}, it is appended explicitly to u_t so that the raw positional index is directly available to the regime scorer g_Δ and the retention gate, without being mixed with graph-structural features in h_t. The graph encoder maps high-dimensional adjacency and node features into a fixed-dimensional state that preserves discriminative connectivity patterns for temporal memory updates. Here, s_t adds explicit structural information, helping the model distinguish stable graph patterns from changing graph states. The structural drift descriptor Δ_t is computed as:

$$\Delta_t = \Gamma(u_t, u_{t-1}) = [\|u_t - u_{t-1}\|_2,\ |\text{dens}(G_t) - \text{dens}(G_{t-1})|], \qquad \Delta_t \in \mathbb{R}^2_{\geq 0}. \tag{3}$$

where $\Gamma(u_t, u_{t-1})$ is the structural drift function capturing changes between u_t and u_{t-1} in terms of embedding and edge density (dens). Edge density $\text{dens}(G_t)$ is the ratio of existing to maximum

possible connections of a graph. Then, the regime score (q_t) for ERP peak index t is estimated using the current graph state u_t, drift descriptor, and the previous memory state m_{t-1} as:

$$q_t = \sigma(g_\Delta([u_t; m_{t-1}; \Delta_t])), \qquad q_t \in (0, 1), \tag{4}$$

where $g_\Delta(\cdot)$ is a learnable neural scoring function, $\sigma(\cdot)$ is the sigmoid activation, and $[\cdot;\cdot]$ denotes vector concatenation. Here, $u_t \in \mathbb{R}^{d_u}$ is the current encoded graph state, $u_{t-1} \in \mathbb{R}^{d_u}$ denotes the previous graph state, and $m_{t-1} \in \mathbb{R}^{d_m}$ denotes the previous memory state. Here, we use $d_u = 96$ and $d_m = 96$ as the dimensions of the encoded graph state and memory state, respectively. A small q_t indicates a stable regime, while a large q_t indicates a changing regime; accordingly, $q_t \approx 0$ favors strong memory retention and $q_t \approx 1$ enables stronger memory adaptation. Thus, q_t acts as a learned regime-change score that controls how the memory should be updated. The provisional memory state ($\widetilde{m}_t$) is updated by an affine graph-memory cell:

$$\widetilde{m}_t = A_t^m m_{t-1} + B_t^m, \tag{5}$$

$$\rho_t = \sigma\big(w_\rho^\top [u_t; q_t] + b_\rho\big), \qquad A_t^m = \rho_t I, \qquad B_t^m = (1 - \rho_t)\Phi(u_t) \tag{6}$$

Here, w_ρ and b_ρ are learnable parameters of the retention gate. Specifically, $w_\rho \in \mathbb{R}^{d_u+1}$ is a weight vector and $b_\rho \in \mathbb{R}$ is a scalar bias, where d_u is the dimension of the encoded graph state u_t and the additional 1 corresponds to the scalar regime score q_t. The sigmoid function maps the gate value to $\rho_t \in (0, 1)$, so that ρ_t controls how much of the previous memory is retained. The function $\Phi(\cdot)$ maps the current graph state u_t into the memory space and produces a candidate memory injection. Thus, $B_t^m = (1 - \rho_t)\Phi(u_t)$ injects new graph information, while $A_t^m = \rho_t I$ retains the previous memory (Identity matrix: I). The superscript m is used to distinguish the memory transition matrix A_t^m from the graph adjacency matrix A_t.

3.2 Non-Stationary Memory Refinement

The provisional memory state $\widetilde{m}_t$ provides a valid regime-aware memory update, but non-stationary graph sequences require the memory to track drifting graph dynamics in a stable and optimization-consistent manner (Su et al., 2024; Feldman & Baskin, 2023). Therefore, after constructing $\widetilde{m}_t$, we recast memory refinement as an online non-stationary optimization problem. At each time step, we define a convex surrogate loss $\ell_t(m; u_t) = \frac{1}{2}\|m - u_t\|_2^2$, which encourages the memory to track the current graph state and admits a closed-form proximal update, and compute its gradient at the provisional memory:

$$g_t = \nabla_m \ell_t(m; u_t)\big|_{m=\widetilde{m}_t} \tag{7}$$

The gradient g_t is treated as frozen during the update at time t. The final refined memory state (m_t) is then obtained through a regime-aware proximal refinement step:

$$m_t = \arg\min_m \left[\underbrace{\eta_t \langle g_t, m\rangle}_{\text{online gradient term}} + \underbrace{D_\psi(m, \widetilde{m}_t)}_{\text{proximity term}} + \underbrace{\lambda(1 - q_t)\|m - m_{t-1}\|_2^2}_{\text{regime-aware stability term}} \right] \tag{8}$$

The first term, $\eta_t\langle g_t, m\rangle$, is the online gradient term, where g_t is the gradient of the convex surrogate loss at the provisional memory $\widetilde{m}_t$, and η_t is the proximal step size. In our implementation, we use a decaying schedule $\eta_t = \eta_0/\sqrt{t}$ with $\eta_0 = 10^{-2}$ selected on the validation set, where $t \in \{1, \ldots, T\}$ is the ERP-peak index and $T = 17$. This term encourages the memory to move in a direction that reduces the current online loss (Yang & Ma, 2023). The second term, $D_\psi(m, \widetilde{m}_t)$, is the proximity term, where $D_\psi(\cdot,\cdot)$ denotes the Bregman divergence induced by ψ. We take $\psi = \frac{1}{2}\|\cdot\|_2^2$, so $D_\psi(m, \widetilde{m}_t) = \frac{1}{2}\|m - \widetilde{m}_t\|_2^2$. The update is therefore a proximal-gradient step, which is the Euclidean special case of online mirror descent; we adopt the proximal form directly because it admits a closed-form solution that preserves the affine structure of the memory. The second term keeps the refined memory m_t close to the provisional affine memory $\widetilde{m}_t$. The third term is the regime-aware stability term, where λ controls the strength of temporal regularization. Under this choice of ψ, the regime-dependent refinement admits the following closed-form solution:

$$m_t = \frac{\widetilde{m}_t - \eta_t g_t + 2\lambda(1 - q_t)m_{t-1}}{1 + 2\lambda(1 - q_t)}. \tag{9}$$

Substituting the provisional affine update $\widetilde{m}_t = A_t^m m_{t-1} + B_t^m$ gives

$$m_t = \bar{A}_t\, m_{t-1} + \bar{B}_t,$$

where the *refined affine transition matrix* ($\bar{A}_t$) and the *refined affine injection vector* ($\bar{B}_t$) are given by

$$\bar{A}_t = \frac{A_t^m + 2\lambda(1-q_t)I}{1+2\lambda(1-q_t)}, \qquad \bar{B}_t = \frac{B_t^m - \eta_t g_t}{1+2\lambda(1-q_t)}.$$

Alternative refinement strategies, such as Sword/Sword++ or meta-learned update rules (Lee et al., 2019), could also adapt the memory, but they require either dynamic-regret machinery or an inner optimization loop at every step. The proximal update used here belongs to the online mirror-descent family, whose behaviour under time-varying objectives is well characterized (Zhao et al., 2024; Eshraghi & Liang, 2022), and it yields a closed-form step that preserves the affine memory structure.

3.3 Classification with RAGMO-Net

RAGMO-Net combines the current graph representation with the memory state: $\overline{z}_t = [u_t; m_t]$. The sequence of joint graph-memory states ($\overline{z}_t$) is then used to compute the attention weight (β_t, in eq. 10) which is assigned to the t-th graph-memory state, with $\beta_t \in (0,1)$ and $\sum_{t=1}^{T} \beta_t = 1$. The learnable scoring function $a(\cdot)$ maps the concatenated vector $[\overline{z}_t; p_t]$ to a scalar attention logit. The exponential function $\exp(\cdot)$ converts each logit into a positive score, while the denominator performs softmax normalization over all T graph snapshots. The resulting pooled representation z_{pool} is computed as a weighted sum of the joint graph-memory states $\overline{z}_t$ using the learned attention weights β_t:

$$\beta_t = \frac{\exp(a([\overline{z}_t; p_t]))}{\sum_{i=1}^{T} \exp(a([\overline{z}_i; p_i]))}, \qquad z_{\text{pool}} = \sum_{t=1}^{T} \beta_t \overline{z}_t. \tag{10}$$

The final class probability is $\widehat{y} = \sigma(w^\top z_{\text{pool}} + b)$, where $\widehat{y}$ denotes the probability of dementia vs. healthy. The training objective combines task, memory-refinement, and regime calibration losses:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{task}}}_{\text{task prediction loss}} + \underbrace{\lambda_{\text{omd}}\mathcal{L}_{\text{omd}}}_{\text{proximal memory-refinement loss}} + \underbrace{\lambda_q \mathcal{L}_q}_{\text{regime calibration loss}} \tag{11}$$

Here, λ_{omd} and λ_q are non-negative coefficients controlling the contributions of the memory-refinement and regime calibration losses, respectively. The loss terms are mentioned below:

$$\mathcal{L}_{\text{task}} = -\left[y \log \widehat{y} + (1-y)\log(1-\widehat{y})\right]. \tag{12}$$

$$\mathcal{L}_{\text{omd}} = \sum_{t=1}^{T} \Big[\eta_t \langle g_t, m_t\rangle + D_\psi(m_t, \tilde{m}_t) + \lambda(1-q_t)\|m_t - m_{t-1}\|_2^2\Big] \tag{13}$$

$$\mathcal{L}_q = \sum_{t=2}^{T} \text{BCE}(q_t, \widetilde{c}_t), \tag{14}$$

where $\widetilde{c}_t$ denotes the pseudo regime-change label constructed from the structural drift magnitude. In Eq. (13), $g_t = \tilde{m}_t - u_t$ is treated as a stop-gradient quantity during backpropagation, consistent with the frozen-gradient assumption in Eq. (7), where g_t is evaluated at the provisional memory $\tilde{m}_t$ and held fixed during the refinement step. We set $\widetilde{c}_t = \mathbb{I}\{\|\Delta_t\|_2 > \tau_\Delta\}$ for $t \geq 2$, where τ_Δ is the median drift magnitude computed on the training split only. BCE denotes the binary cross-entropy loss. The summation starts from $t = 2$ because regime change is defined by comparing consecutive graph states, and no previous graph state is available at $t = 1$. For any fixed network parameters, $\mathcal{L}_{\text{omd}}$ is bounded below, as formally guaranteed by Lemma 1.

Lemma 1 (Structural Boundedness and Stability). *Let $m_0 = 0$, $\lambda \geq 0$, $\eta_t = \eta_0/\sqrt{t}$ with $0 < \eta_0 \leq 1$, $\|u_t\|_2 \leq K_u$, and let Φ be Lipschitz with $\|\Phi(u_t)\|_2 \leq K_\Phi$. Set $K = \max\{K_u, K_\Phi\}$ and $\rho_{\max} := \sigma\big(\|w_\rho\|_2\sqrt{K_u^2+1} + |b_\rho|\big) < 1$. Then (a) $\|\bar{A}_t\|_2 = \frac{\rho_t + 2\lambda(1-q_t)}{1+2\lambda(1-q_t)} \leq 1-\delta$ with $\delta = \frac{1-\rho_{\max}}{1+2\lambda} > 0$, uniformly in t and m_{t-1}; (b) $\|m_t\|_2 \leq K$, $\|g_t\|_2 \leq 2K$, and $\|\bar{B}_t\|_2 \leq M_B := (1+2\eta_0)K$ for all t; and (c) $\mathcal{L}_{\text{omd}} \geq -2\eta_0 K^2(2\sqrt{T}-1) > -\infty$. The proof is given in Appendix B.*

Theorem 1 (Stability and Contraction)**.** *Under the assumptions of Lemma 1 (with any initial memory satisfying $\|m_0\|_2 \le K$), substituting $g_t = \tilde{m}_t - u_t$ into Eq. (9) gives the exact update $m_t = F_t(m_{t-1}; u) := \gamma_t\, m_{t-1} + b_t$, where $\gamma_t = \frac{(1-\eta_t)\rho_t + 2\lambda(1-q_t)}{1+2\lambda(1-q_t)} \le 1-\delta$ and $b_t = \frac{(1-\eta_t)(1-\rho_t)\Phi(u_t)+\eta_t u_t}{1+2\lambda(1-q_t)}$ depend on m_{t-1} only through q_t. Assume g_Δ is L_g-Lipschitz in its memory argument, and let $L_\alpha := \left(\frac{|w_q|}{4} + 2\lambda\right)\frac{L_g}{4}$ and $\kappa := (1-\delta) + 2KL_\alpha$, where w_q is the weight of q_t in w_ρ. Then (a)* BIBO stability: *$\|m_t\|_2 \le K$ for all t whenever $\|m_0\|_2 \le K$; and (b)* contraction: *if $\kappa < 1$, then for any two input sequences u, u' and initial states in the ball of radius K,*

$$\|m_t - m'_t\|_2 \le \kappa^t \|m_0 - m'_0\|_2 + \frac{1}{1-\kappa} \max_{1\le k\le t} \varepsilon_k, \qquad \varepsilon_k := \sup_{\|m\|_2 \le K} \|F_k(m;u) - F_k(m;u')\|_2.$$

The full proof is provided in Appendix B.

Lemma 1 guarantees the retention margin $\delta > 0$ for any fixed network parameters under bounded inputs. Theorem 1 then shows that the memory trajectory is BIBO-stable, and, whenever the regime scorer is not overly sensitive to the memory state ($\kappa < 1$), contractive: the memory forgets its initialization geometrically, and bounded perturbations of the input graph stream produce proportionally bounded deviations of the memory.

4 Experimental Setup

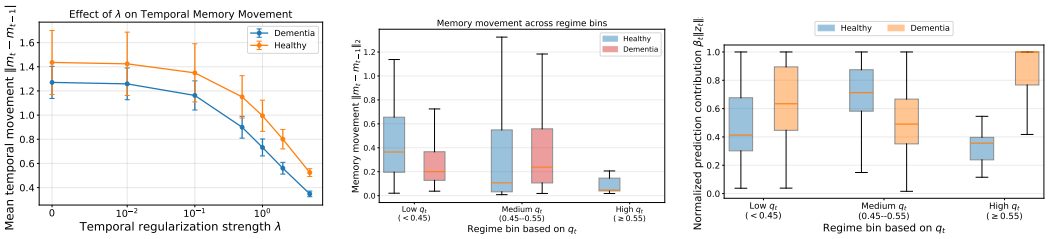
4.1 Datasets

The experimental dataset (Miltiadous et al., 2023) comprises 23 FTD (male=14, female=9) and 29 HC (male=18, female=11) subjects. EEG signals were recorded in a resting-state, eyes-closed condition at a sampling rate of 500 Hz using 19 channels. EEG recordings typically last approximately 12 minutes. In this paper, FTD is referred to under the broader term dementia, while HC and healthy are used interchangeably. We evaluated our model across three complementary settings to assess its generalization of the proposed model over diverse application domains (EEG/ non-EEG). Case 1-*intra-application EEG* setting, we considered EEG-based brain-disorder classification tasks: AD vs. HC (DS1) (Miltiadous et al., 2023), and parkinson's vs. HC (DS2) (Cavanagh et al., 2021). Case 2 considers a *cross-application EEG* setting, including motor imagery vs. resting-state EEG (DS3) (Peterson et al., 2020), pre vs. post-mental-arithmetic EEG (DS4) (Zyma et al., 2019), and RSVP/P300 Speller Brain–Computer Interface (BCI) tasks (DS5) (Won et al., 2022). Case 3, under the *cross-domain non-EEG* setting, we assessed the RAGMO-Net model's robustness and generalization on non-EEG graph and temporal-interaction datasets, including DGraphFin for financial anomaly detection (DS6) (Huang et al., 2022), TGBL-Coin (DS7) (Shamsi et al., 2022), TGBL-Comment (DS8) (Nadiri & Takes, 2022), and TGBL-Flight (DS9) (Strohmeier et al., 2021) from the Temporal Graph Benchmark (Huang et al., 2023), and the MOOC user action dataset (DS10) (Kumar et al., 2019). A detailed description of the datasets (DS1–DS10) is provided in the *Appendix C*.

4.2 RAGMO-Net Training setup

For EEG datasets (experimental, DS1-DS5), RAGMO-Net was evaluated with subject-wise stratified 5-fold cross-validation (CV) to assess inter-subject generalization. In each fold, one subject split was held out for testing, while 30% of the remaining subjects from each class were used for validation, with at least two validation subjects per class. All ERP-peak graphs from the same subject remained in the same fold to prevent subject-level leakage. For each peak, graph summaries included the mean, standard deviation, maximum, minimum, and edge density. Peak-wise features were combined with an 8-dimensional learnable peak embedding, projected to 96 dimensions (d_u, d_m), and aggregated using masked attention pooling with an attention dimension of 64. The classifier used two fully connected layers (64 and 32 units). Across all datasets, training used the BCE loss function with AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 16, and up to 150 epochs. Validation performance-based early stopping used patience 12 and minimum improvement 10^{-4}.

For non-EEG datasets (DS6–DS10), we use leave-one-out CV (LOOCV) over prediction units, such as target nodes, temporal edges, or users, rather than subject-wise stratified CV. This is not literal per-instance LOOCV over all samples; instead, the held-out unit corresponds to the task-specific



(a) (b) (c)

Figure 2: (a): Effect of temporal regularization strength λ on mean temporal memory movement. (b): Memory movement across regime-score bins for healthy and dementia groups (box plot), highlighting group-wise differences in memory adaptation. (c): Prediction contribution across regime bins for healthy and dementia subjects.

prediction target: target nodes for DS6, candidate temporal interactions for DS7, DS8, and DS9, and users for DS10. Here, we use strict chronological splits, training on past interactions and evaluating on future events to prevent information leakage. In all cases, classification results are reported as the mean $\pm$ standard deviation over evaluation splits. The model was implemented using the TensorFlow framework and trained on NVIDIA A100 Tensor Core GPUs, equipped with 40 GB of HBM2 VRAM.

5 Results

5.1 Regime-aware Memory Analysis

Figure 2(a) shows that increasing λ reduces temporal memory movement in both healthy and dementia groups under regime-aware regularization. This confirms that a larger λ strengthens stability around m_{t-1}, while a smaller λ permits more flexible memory adaptation. Figure 2(b) shows memory movement across regime-score bins. Lower q_t strengthens stability regularization, whereas higher q_t allows more adaptive updates. Healthy subjects show broader variation in low and medium regimes, while dementia subjects show larger dispersion in the medium regime. This aligns with prior findings on disrupted connectivity in dementia (Paitel et al., 2025; Ranasinghe & Mapa, 2024) and shows that regime-aware refinement captures distinct temporal dynamics. Figure 2(c) shows the normalized prediction contribution ($\beta_t|z_t|_2$) across low, medium, and high regime-score bins for Healthy and Dementia subjects. Prediction contributions increase across regime-score bins, especially for Dementia, indicating that RAGMO-Net emphasizes high-regime bins linked to abnormal temporal dynamics while Healthy subjects remain more stable.

Figure 3 shows the regime-weighted EEG connectivity difference between dementia and healthy subjects. The connectivity matrices are aggregated using the attention-weighted temporal representation (Eq. 10), where β_t emphasizes informative ERP peaks during temporal pooling. The map reflects both structural differences in A_t and the importance of informative ERP peaks. Stronger frontal (Fp1, Fp2, etc.) and temporal (T3, T4, etc.) connectivity in dementia indicates disorder-relevant connectivity alterations (Ma et al., 2024);

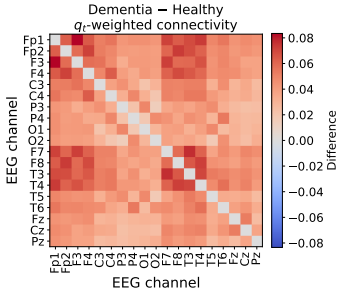
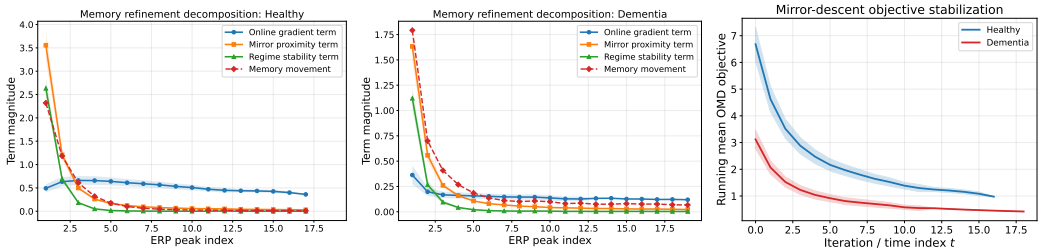


Figure 3: Regime and attention-weighted EEG connectivity difference.

5.2 Non-Stationary Memory Refinement Analysis

Figures 4(a,b) decompose the memory-refinement update into the three terms derived from Eq. 8. The online gradient term drives adaptation toward the current graph state, the proximity term keeps the refined memory close to the provisional affine memory, and the regime stability term penalizes abrupt changes from the previous memory under stable regimes. The healthy group (Figure 4(a)) shows



(a) (b) (c)

Figure 4: Optimization-term decomposition and convergence of the proximal refinement across GFP peaks for (a) healthy and (b) dementia groups. (c) Convergence behavior of the refinement, shown through the running mean objective over GFP peak index for two groups.

larger initial magnitudes but a faster reduction in memory movement, indicating that the refinement quickly stabilizes after the first few peaks. In contrast, the dementia group (Figure 4(b)) exhibits smaller initial terms but a more persistent memory movement and proximity contribution across later peaks, suggesting more prolonged correction under pathological dynamics (Paitel et al., 2025). Figure 4(c) shows convergence of the refinement via the running mean objective over time. The healthy group shows a higher refinement objective with slower stabilization, indicating persistent refinement, whereas the dementia group converges faster to a lower objective.

5.3 Performance Analysis

Table 1: Performance comparison (Accuracy: Acc and F1-score: F1) with baseline models across multiple datasets. Case 1 includes experimental dataset /Exp DS–DS2, and Case 2 includes DS3–DS5. The best results are shaded in gray, and the second-best results are shown in blue; standard deviations are reported in parentheses. Here, SSM denotes a state-space model.

Baseline type	Method	Exp DS (Miltiadous et al., 2023)		DS1 (Miltiadous et al., 2023)		DS2 (Cavanagh et al., 2021)		DS3 (Peterson et al., 2020)		DS4 (Zyma et al., 2019)		DS5 (Won et al., 2022)	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Static graph	GCN (Kipf & Welling, 2016)	80.23(3.55)	79.33(5.01)	85.47(4.45)	84.37(4.24)	86.91(2.23)	85.02(3.01)	77.50(3.45)	75.77(4.03)	83.50(2.34)	81.87(2.78)	86.02(5.14)	84.98(4.77)
	GAT (Veličković et al., 2017)	83.70(2.34)	82.44(4.33)	84.01(5.23)	83.78(6.01)	86.72(2.98)	84.33(3.12)	75.23(3.44)	74.38(4.01)	79.57(4.23)	77.33(3.87)	83.22(7.01)	81.79(6.23)
	BrainNetCNN (Kawahara et al., 2017)	78.53(3.23)	76.32(4.31)	80.93(7.02)	78.45(6.89)	83.46(4.45)	82.48(5.01)	72.88(8.27)	70.34(6.77)	87.68(3.66)	86.23(3.28)	79.27(3.14)	78.10(3.77)
	BrainGNN (Li et al., 2021)	82.81(6.11)	80.03(6.56)	75.33(5.12)	74.76(6.03)	85.45(3.12)	84.77(4.01)	69.23(3.44)	68.38(4.13)	88.89(5.03)	87.55(5.66)	82.66(4.39)	80.71(3.87)
SSM	EEG-Mamba (Gui et al., 2024)	91.20(3.10)	90.05(3.42)	88.75(3.65)	87.60(3.81)	90.85(2.95)	89.40(3.12)	76.95(4.12)	75.80(4.45)	89.10(3.25)	88.05(3.48)	90.25(3.05)	89.10(3.26)
	DG-Mamba (Yuan et al., 2025)	92.10(2.85)	91.05(3.10)	90.25(3.20)	89.30(3.35)	92.05(2.60)	90.80(2.75)	78.10(3.95)	76.95(4.10)	90.85(2.90)	89.70(3.05)	91.60(2.70)	90.55(2.88)
Dynamic graph	ST-GCN (Yan et al., 2018)	87.50(5.29)	86.23(6.01)	87.50(3.33)	86.22(4.23)	92.77(4.12)	90.24(3.98)	79.23(3.44)	78.32(4.18)	87.50(1.24)	86.72(2.03)	89.80(3.33)	88.55(4.14)
	TGAT (Xu et al., 2020)	92.91(2.19)	91.87(3.55)	81.66(3.18)	80.34(4.01)	88.24(4.34)	86.23(3.87)	81.44(6.36)	80.13(5.78)	83.74(4.21)	81.48(4.67)	91.24(2.28)	90.38(1.98)
	TGN (Rossi et al., 2020)	93.34(4.32)	92.17(3.45)	83.34(2.29)	82.23(1.89)	93.54(1.55)	91.88(2.02)	78.34(6.44)	77.56(5.87)	92.12(2.78)	90.88(3.22)	92.11(2.13)	91.83(3.01)
	EvolveGCN (Pareja et al., 2020)	89.46(6.45)	87.55(6.78)	91.67(3.27)	90.34(2.77)	89.56(2.82)	87.45(3.12)	82.76(2.22)	81.82(3.65)	92.44(4.12)	91.88(3.78)	92.02(4.83)	90.72(5.24)
	RAGMO-Net (proposed)	**97.64(1.12)**	**96.73(1.34)**	**93.84(3.89)**	**91.61(2.23)**	**94.98(2.18)**	**92.73(1.78)**	**88.48(2.45)**	**87.56(2.11)**	**95.25(1.49)**	**94.77(2.34)**	**95.29(1.45)**	**94.06(1.87)**

Table 2: Performance comparison (Accuracy, F1-score) with baseline models across multiple non-EEG datasets: DS6–DS10 (case 3). The same highlighting style is used as in Table 1.

Baseline type	Method	**DS6** (Huang et al., 2022)		**DS7** (Shamsi et al., 2022)		**DS8** (Nadiri & Takes, 2022)		**DS9** (Strohmeier et al., 2021)		**DS10** (Kumar et al., 2019)	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Static graph	GCN (Kipf & Welling, 2016)	75.98(2.44)	73.45(3.12)	75.95(3.23)	72.97(4.01)	72.23(5.12)	70.34(4.88)	87.22(3.12)	85.83(3.87)	82.42(2.33)	80.77(3.12)
	GAT (Veličković et al., 2017)	79.43(3.24)	77.23(4.13)	73.01(4.22)	68.84(2.33)	75.87(5.37)	73.11(5.01)	87.97(4.22)	86.34(4.04)	81.19(5.12)	80.08(4.56)
	BrainNetCNN(Kawahara et al., 2017)	75.34(3.23)	74.21(4.01)	83.24(2.89)	82.81(3.12)	84.86(2.98)	82.33(2.34)	92.23(3.57)	91.88(4.12)	77.23(3.21)	75.57(3.47)
	BrainGNN (Li et al., 2021)	73.34(6.23)	72.27(5.78)	87.81(3.33)	85.98(4.11)	84.76(3.56)	82.77(4.10)	91.33(4.57)	90.18(5.02)	81.48(2.39)	79.84(2.10)
SSM	EEG-Mamba (Gui et al., 2024)	82.45(3.06)	80.92(3.44)	86.70(2.91)	84.95(3.18)	86.95(2.74)	85.32(3.05)	91.55(2.68)	90.20(3.02)	83.20(2.91)	81.65(3.24)
	DG-Mamba (Yuan et al., 2025)	83.10(2.84)	81.70(3.21)	87.55(2.66)	85.72(2.94)	87.62(2.58)	86.15(2.87)	92.40(2.41)	91.18(2.76)	84.10(2.63)	82.65(3.01)
Dynamic graph	ST-GCN (Yan et al., 2018)	80.34(4.34)	78.11(5.02)	90.79(4.23)	89.21(5.01)	85.09(1.83)	84.33(1.32)	92.03(1.45)	90.45(2.01)	81.41(3.44)	80.02(2.75)
	TGAT (Xu et al., 2020)	75.57(3.87)	74.82(5.10)	82.62(2.12)	80.77(1.97)	87.58(1.23)	86.22(2.02)	91.03(2.23)	90.65(2.45)	80.02(1.23)	79.22(2.12)
	TGN (Rossi et al., 2020)	85.71(4.21)	84.22(4.02)	79.65(4.32)	76.23(3.87)	79.41(3.27)	78.32(4.01)	91.67(2.40)	90.12(3.11)	82.34(2.33)	80.56(3.01)
	EvolveGCN (Pareja et al., 2020)	83.73(5.34)	82.72(5.23)	88.15(3.56)	86.23(5.12)	88.21(4.01)	87.31(3.87)	93.23(3.12)	92.54(2.78)	84.56(4.23)	82.44(3.98)
	RAGMO-Net (proposed)	**88.20(2.55)**	**87.14(3.17)**	**92.83(1.01)**	**90.18(1.87)**	**93.29(2.34)**	**92.55(3.12)**	**96.72(2.45)**	**94.23(3.12)**	**89.05(1.34)**	**87.94(1.12)**

RAGMO-Net achieves a maximum accuracy of $97.64 \pm 1.12\%$ on the experimental dataset. The performance analysis in Table 1 (Experimental DataSet, DS1–DS5) compares RAGMO-Net with static graph, state-space model (SSM), and dynamic graph baselines. Specifically, for Exp DS and DS2, the second-best model is *TGN*, and RAGMO-Net improves accuracy by **4.30**% and **1.44**%,

respectively. For DS1, DS3, and DS4, the second-best model is *EvolveGCN*, with gains of **2.17**%, **5.72**%, and **2.81**%, respectively. For DS5, the second-best model is again *TGN*, and RAGMO-Net improves accuracy by **3.18**%. Similarly, in Table 2 (DS6–DS10), RAGMO-Net maintains superior performance across all datasets, demonstrating robustness and generalization to non-EEG graph domains. In particular, the second-best model is *TGN* for DS6, where RAGMO-Net improves accuracy by about **2.49**%. For DS7, the second-best model is *ST-GCN*, and RAGMO-Net achieves an improvement of about **2.04**%. For DS8–DS10, *EvolveGCN* is the second-best model, and RAGMO-Net improves accuracy by approximately **5.08**%, **3.49**%, and **4.49**%, respectively. The statistical evidence and additional analyses are provided in *Appendix* E and *Appendix* D, respectively.

5.4 Ablation Study

Table 3 demonstrates the ablation analysis of the RAGMO-Net model. In the *component ablation* group, removing core modeling elements consistently reduces performance. In the *optimization ablation* group, both proximal refinement and the stability penalty are beneficial, since removing them reduces accuracy to 88.44±2.34 and 85.35±1.45, respectively. In the *loss-function ablation* group, removing $\mathcal{L}\mathrm{omd}$, $\mathcal{L}\mathrm{q}$, or using *only* $\mathcal{L}_{\mathrm{task}}$ still gives strong results, but all remain below the full model, showing that the auxiliary losses provide complementary gains beyond task loss alone. Compared with the strongest ablated variant (*only* $\mathcal{L}_{\mathrm{task}}$), RAGMO-Net yields relative improvements of **4.20%** in accuracy and **3.62%** in F1-score.

Table 3: Ablation study of RAGMO-Net on experimental dataset across multiple ablation categories.

Category	Ablation Setting	Acc(%)	F1(%)
Component	Graph only	67.45(5.21)	65.54(4.45)
	W/o Memory	76.33(6.52)	75.12(5.42)
	No Regime Score	78.57(3.45)	75.70(3.64)
	Memory-Only Prediction	80.34(2.34)	80.89(2.15)
	W/o Memory Refinement + $\mathcal{L}_{\mathrm{omd}}$	84.10(3.20)	83.65(3.05)
Optimization	W/o Proximal Refinement	88.44(2.34)	88.06(2.75)
	W/o Stability Penalty ($\lambda = 0$)	85.35(1.45)	84.06(1.53)
Loss Function	W/o $\mathcal{L}_{\mathrm{omd}}$	91.14(3.45)	90.17(3.14)
	W/o $\mathcal{L}_{q}$	93.10(4.23)	92.90(3.61)
	Only $\mathcal{L}_{\mathrm{task}}$	93.44(5.34)	93.11(4.32)
RAGMO-Net	**Full model**	**97.64(1.12)**	**96.73(1.34)**

5.5 Runtime Analysis

Table 4 compares RAGMO-Net with static graph, SSM, and dynamic graph baselines in terms of model size and computational cost. RAGMO-Net achieves the best efficiency, with relative reductions of **25.0%** in parameters, **90.0%** in Mega Floating-Point Operations (MFLOPs), and **18.7%** in inference time compared with the closest baseline (GCN). These reductions mainly arise from the scalar-gated affine memory cell: the transition is a scalar multiple of the identity, so each step requires one scalar gate evaluation and one vector update rather than a dense matrix product or a gated recurrent block. RAGMO-Net reduces per-snapshot computation while preserving temporal memory adaptation.

Table 4: Runtime comparison between RAGMO-Net and other baselines on the experimental dataset.

Baseline type	Model	#Params (M)	MFLOPs	Inference time (ms)
Static graph	GCN (Kipf & Welling, 2016)	0.024	4.112	26.412
	GAT (Veličković et al., 2017)	0.026	6.422	40.791
	BrainNetCNN (Kawahara et al., 2017)	0.077	30.612	28.637
	BrainGNN (Li et al., 2021)	0.039	11.471	65.562
SSM	EEG-Mamba (Gui et al., 2024)	0.858	30.340	276.826
	DG-Mamba (Yuan et al., 2025)	0.637	21.250	268.030
Dynamic graph	ST-GCN (Yan et al., 2018)	0.038	14.013	34.545
	TGAT (Xu et al., 2020)	0.030	7.264	41.622
	TGN (Rossi et al., 2020)	0.031	8.242	45.352
	EvolveGCN (Pareja et al., 2020)	0.416	14.091	322.662
	RAGMO-Net	**0.018**	**0.412**	**21.473**

6 Conclusion

In this work, we proposed RAGMO-Net, a regime-aware graph memory optimization framework for EEG-based dynamic graphs that combines affine memory updates with optimization-consistent mirror-descent refinement. On the experimental dataset, RAGMO-Net achieves an accuracy of $97.64 \pm 1.12\%$ and consistently outperforms diverse baselines across EEG and non-EEG datasets. Runtime analysis provides empirical evidence of improved scalability through the affine memory formulation. Despite these promising results, two **limitations** remain. First, the evaluation is limited by cohort diversity; future work will consider larger multi-site EEG cohorts. Second, RAGMO-Net currently focuses on binary classification and will be extended to multi-class neurological analysis.

AI USE STATEMENT

We only used ChatGPT to assist in generating selected visual elements of Figure 1. AI-assisted figure content was reviewed and verified by the authors. Generative AI tools were not used for research ideation, experimental design, mathematical formulation or analysis, result generation, or manuscript writing. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This study uses only publicly available EEG datasets and does not involve new data collection or direct interaction with human participants. We use the datasets solely for research purposes and cite their original sources and associated publications. We follow the terms and conditions associated with the respective datasets and the principles of the ICLR Code of Ethics, including responsible data use, research integrity, transparency, and respect for privacy.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the RAGMO-Net architecture, pseudocode, mathematical formulation, training settings, datasets, and evaluation protocols in the main paper and *Appendix*. All datasets used in this study are publicly available, and their original sources are cited in the paper. The source code will be made publicly available upon acceptance of the paper.

REFERENCES

James F. Cavanagh, Arun Singh, and Nandakumar S. Narayanan. EEG: Simon conflict in parkinson's. OpenNeuro Dataset, 2021. Available at: `https://openneuro.org/datasets/ds003509`.

Debashis Das Chakladar and Nikhil R Pal. Brain connectivity analysis for EEG-based face perception task. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4):1494–1506, 2024.

Debashis Das Chakladar, Partha Pratim Roy, and Masakazu Iwamura. EEG-based cognitive state classification and analysis of brain dynamics using deep ensemble model and graphical brain network. *IEEE Transactions on Cognitive and Developmental Systems*, 14(4):1507–1519, 2021.

Debashis Das Chakladar, Sumalyo Datta, Partha Pratim Roy, and Vinod A Prasad. Cognitive workload estimation using variational autoencoder and attention-based deep model. *IEEE Transactions on Cognitive and Developmental Systems*, 15(2):581–590, 2022.

Debashis Das Chakladar, Anand Shankar, Foteini Liwicki, Shovan Barma, and Rajkumar Saini. Attention dynamics: Estimating attention levels of adhd using swin transformer. In *International Conference on Pattern Recognition*, pp. 270–283. Springer, 2025.

Guy Cheron, Géraldine Petit, Julian Cheron, Axelle Leroy, Anita Cebolla, Carlos Cevallos, Mathieu Petieau, Thomas Hoellinger, David Zarka, Anne-Marie Clarinval, et al. Brain oscillations in sport: toward EEG biomarkers of performance. *Frontiers in psychology*, 7:246, 2016.

Sujal Chodvadiya and MS Suchithra. Modeling spectral EEG interactions using graph-structured variational representation learning. *Cognitive Neurodynamics*, 20(1):71, 2026.

Ahmed S Eltrass, Youssef Tageldin, and Hania H Farag. A new graph-transformer framework for EEG-based differentiation of alzheimer's disease and frontotemporal dementia. *Biomedical Physics & Engineering Express*, 12(2):025023, 2026.

Nima Eshraghi and Ben Liang. Dynamic regret of online mirror descent for relatively smooth convex cost functions. *IEEE Control Systems Letters*, 6:2395–2400, 2022.

Or Feldman and Chaim Baskin. Leveraging temporal graph networks using module decoupling. *arXiv preprint arXiv:2310.02721*, 2023.

Lin Feng, Cheng Cheng, Mingyan Zhao, Huiyuan Deng, and Yong Zhang. EEG-based emotion recognition using spatial-temporal graph convolutional LSTM with attention mechanism. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5406–5417, 2022.

Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. EEGmamba: Bidirectional state space model with mixture of experts for EEG multi-task classification. *arXiv preprint arXiv:2407.20254*, 2024.

David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 205–213, 2017.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. *Advances in Neural Information Processing Systems*, 36:2056–2073, 2023.

Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. Dgraph: A large-scale financial dataset for graph anomaly detection. *Advances in Neural Information Processing Systems*, 35:22765–22777, 2022.

Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1269–1278, 2019.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10657–10665, 2019.

Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. BrainGNN: Interpretable brain graph neural network for fMRI analysis. *Medical image analysis*, 74:102233, 2021.

Xinzhe Li, Bruno Mota, Toshiyuki Kondo, Slawomir Nasuto, and Yoshikatsu Hayashi. EEG dynamical network analysis method reveals the neural signature of visual-motor coordination. *Plos one*, 15(5):e0231767, 2020.

Chenyu Liu, Xinliang Zhou, Jiaping Xiao, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. VSGT: Variational spatial and gaussian temporal graph models for EEG-based emotion recognition. In *IJCAI*, pp. 3078–3086, 2024.

Runnan Lu, Ying Zeng, Rongkai Zhang, Bin Yan, and Li Tong. SAST-GCN: Segmentation adaptive spatial temporal-graph convolutional network for p3-based video target detection. *Frontiers in Neuroscience*, 16:913027, 2022.

Yuan Ma, Jeffrey Keith Spaneas Bland, and Tsutomu Fujinami. Classification of alzheimer's disease and frontotemporal dementia using electroencephalography to quantify communication between electrode pairs. *Diagnostics*, 14(19):2189, 2024.

Christoph M Michel and Thomas Koenig. EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage*, 180:577–593, 2018.

Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, et al. A dataset of scalp EEG recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6):95, 2023.

Amirhossein Nadiri and Frank W Takes. A large-scale temporal analysis of user lifespan durability on the reddit social media platform. In *Companion Proceedings of the Web Conference 2022*, pp. 677–685, 2022.

Elizabeth R Paitel, Christian BD Otteman, Mary C Polking, Henry J Licht, and Kristy A Nielson. Functional and effective EEG connectivity patterns in Alzheimer's disease and mild cognitive impairment: A systematic review. *Frontiers in aging neuroscience*, 17:1496235, 2025.

Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. EvolveGCN: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5363–5370, 2020.

Victoria Peterson, Catalina Galván, Hugo Hernández, and Ruben Spies. A feasibility study of a complete low-cost consumer-grade brain-computer interface system. *Heliyon*, 6(3), 2020.

PVNN Ranasinghe and Mapa ST Mapa. Functional connectivity and cognitive decline: A review of rs-fmri, EEG, MEG, and graph theory approaches in aging and dementia. *Exploration of Medicine*, 5(6):797–821, 2024.

Haider Raza, Dheeraj Rathee, Shang-Ming Zhou, Hubert Cecotti, and Girijesh Prasad. Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related EEG-based brain-computer interface. *Neurocomputing*, 343:154–166, 2019.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

Kiarash Shamsi, Friedhelm Victor, Murat Kantarcioglu, Yulia Gel, and Cuneyt G Akcora. Chartalist: Labeled graph datasets for utxo and account-based blockchains. *Advances in Neural Information Processing Systems*, 35:34926–34939, 2022.

Amauri Souza, Diego Mesquita, Samuel Kaski, and Vikas Garg. Provably expressive temporal graph networks. *Advances in neural information processing systems*, 35:32257–32269, 2022.

Martin Strohmeier, Xavier Olive, Jannis Lübbe, Matthias Schäfer, and Vincent Lenders. Crowdsourced air traffic data from the opensky network 2019–2020. *Earth System Science Data*, 13(2): 357–366, 2021.

Junwei Su, Difan Zou, and Chuan Wu. Pres: Toward scalable memory-based dynamic graph neural networks. *arXiv preprint arXiv:2402.04284*, 2024.

Déborah Sulem, Henry Kenlay, Mihai Cucuringu, and Xiaowen Dong. Graph similarity learning for change-point detection in dynamic networks. *Machine Learning*, 113(1):1–44, 2024.

Hao Tang, Songyun Xie, Xinzhou Xie, Yujie Cui, Bohan Li, Dalu Zheng, Yu Hao, Xiangming Wang, Yiye Jiang, and Zhongyu Tian. Multi-domain based dynamic graph representation learning for EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 28(9):5227–5238, 2024.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *International conference on learning representations*, 2019.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Kyungho Won, Moonyoung Kwon, Minkyu Ahn, and Sung Chan Jun. EEG dataset for rsvp and p300 speller brain-computer interfaces. *Scientific data*, 9(1):388, 2022.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962*, 2020.

Zhongming Xu, Shaohua Tang, Chuancai Liu, Qiankun Zhang, Heng Gu, Xiaoli Li, Zengru Di, and Zheng Li. Temporal segmentation of EEG based on functional connectivity network structure. *Scientific Reports*, 13(1):22566, 2023.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Zhuang Yang and Li Ma. Adaptive step size rules for stochastic optimization in large-scale learning. *Statistics and Computing*, 33(2):45, 2023.

Mengqing Ye, CL Philip Chen, and Tong Zhang. Hierarchical dynamic graph convolutional network with interpretability for EEG-based emotion recognition. *IEEE transactions on neural networks and learning systems*, 2022.

Deming Yuan, Yiguang Hong, Daniel WC Ho, and Shengyuan Xu. Distributed mirror descent for online composite optimization. *IEEE Transactions on Automatic Control*, 66(2):714–729, 2020.

Haonan Yuan, Qingyun Sun, Zhaonan Wang, Xingcheng Fu, Cheng Ji, Yongjian Wang, Bo Jin, and Jianxin Li. Dg-mamba: Robust and efficient dynamic graph structure learning with selective state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22272–22280, 2025.

Peng Zhao, Yu-Hu Yan, Yu-Xiang Wang, and Zhi-Hua Zhou. Non-stationary online learning with memory and non-stochastic control. *Journal of Machine Learning Research*, 24(206):1–70, 2023.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1–52, 2024.

Jiabin Zhu, Xuanyu Jin, Yuhang Ming, and Wanzeng Kong. EEG-DGRN: dynamic graph representation network for subject-independent ERP detection. *Brain-Apparatus Communication: A Journal of Bacomics*, 4(1):2447576, 2025.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

APPENDIX

A NOTATIONS USED IN THE STUDY

Table 5: Summary of main notations used in RAGMO-Net, listed in order of appearance.

Notation	Definition	Where Used
$\mathcal{X}$	Dynamic EEG graph sequence.	Eq. (1)
G_t	EEG graph snapshot at ERP peak index t.	Eq. (1)
V	Set of graph nodes corresponding to EEG channels.	Eq. (1)
E_t	Set of connectivity edges at ERP peak index t.	Eq. (1)
X_t	Node feature matrix at ERP peak index t.	Eq. (1)
A_t	EEG connectivity adjacency matrix at ERP peak index t.	Eq. (1)
p_t	Learnable positional embedding for ERP peak index t.	Eq. (1), (2)
T	Total number of ERP-peak graph snapshots; $T = 17$.	Eq. (1)
h_t	Graph-level representation computed by the graph encoder for ERP peak index t.	Eq. (2)
s_t	Graph structural summary for ERP peak index t.	Eq. (2)
u_t	Final graph state used by the memory module for ERP peak index t.	Eq. (2)– (4)
d_u	Dimension of the encoded graph state u_t.	Sec. 3.1
d_m	Dimension of the memory state m_t.	Sec. 3.1
$\Gamma()$	Structural drift function.	Eq. (3)
Δ_t	Structural drift descriptor for ERP peak index t.	Eq. (3), (4)
$g_\Delta()$	Learnable neural scoring function for regime estimation.	Eq. (4)
q_t	Regime score for ERP peak index t.	Eq. (4), (8)
A_t^m	Provisional memory transition for ERP peak index t.	Eq. (5), (6)
B_t^m	Provisional memory injection for ERP peak index t.	Eq. (5), (6)
$\tilde{m}_t$	Provisional affine memory state for ERP peak index t.	Eq. (5), (7)
w_ρ, b_ρ	Learnable weight vector and bias scalar of the memory retention gate.	Eq. (6)
ρ_t	Memory retention gate $\in (0, 1)$.	Eq. (6)
g_t	Online surrogate gradient for ERP peak index t.	Eq. (7), (8)
η_t	Proximal step size for GFP peak index t.	Eq. (8)
D_ψ	Bregman divergence used as the proximity term.	Eq. (8)
λ	Temporal stability strength controlling regime-aware memory stabilization.	Eq. (8), (9)
m_t	Refined memory state obtained for ERP peak index t.	Eq. (9)
$\bar{A}_t$	Refined affine transition matrix for ERP peak index t.	Eq. (9), Lemma 1
$\bar{B}_t$	Refined affine injection vector for ERP peak index t.	Eq. (9), Lemma 1
δ	Contraction margin.	Lemma (1), Thm. (1)
α_t	Scalar contraction factor.	Lemma (1)
$\bar{z}_t$	Joint graph-memory state for ERP peak index t.	Eq. (10)
β_t	Attention weight assigned to the t-th graph-memory state.	Eq. (10)
z_{pool}	Attention-pooled temporal representation.	Eq. (10)
λ_{omd}	Coefficient weighting the proximal memory-refinement loss.	Eq. (11)
λ_q	Coefficient weighting the regime calibration loss.	Eq. (11)
$\mathcal{L}_{\text{task}}$	Binary cross-entropy task prediction loss.	Eq. (12)
$\mathcal{L}_{\text{omd}}$	Proximal memory-refinement loss.	Eq. (13)
$\mathcal{L}_q$	Regime calibration loss.	Eq. (14)
$\tilde{c}_t$	Pseudo regime-change label.	Eq. (14)
τ_Δ	Median drift magnitude.	Sec. 3.3
K_u	Uniform bound on the graph-state norm, $\|u_t\|_2 \leq K_u$.	Lemma 1

B Proof of Lemma and Theorem

Lemma 1 (Structural Boundedness and Stability)**.** *Consider the RAGMO-Net memory recursion (Eqs. (5)–(9)) with $m_0 = 0$, $\lambda \geq 0$, and step sizes $\eta_t = \eta_0/\sqrt{t}$ with $0 < \eta_0 \leq 1$. Let $c_t := 2\lambda(1-q_t)$, and assume:*

(A1) Bounded graph states: $\|u_t\|_2 \leq K_u$ *for all t;*

(A2) Lipschitz injection: Φ *is L_Φ-Lipschitz, so that $\|\Phi(u_t)\|_2 \leq K_\Phi := L_\Phi K_u + \|\Phi(0)\|_2$.*

Let $K := \max\{K_u, K_\Phi\}$. Then the following hold.

(a) Uniform contraction. *The refined transition matrix is $\bar{A}_t = \alpha_t I$ with*

$$\alpha_t = \|\bar{A}_t\|_2 = \frac{\rho_t + c_t}{1 + c_t} \in (0,1), \qquad 1 - \alpha_t = \frac{1-\rho_t}{1+c_t} \;\geq\; \delta := \frac{1-\rho_{\max}}{1+2\lambda} \;>\; 0, \quad (15)$$

where $\rho_{\max} := \sigma\big(\|w_\rho\|_2\sqrt{K_u^2+1} + |b_\rho|\big) < 1$. Hence $\|\bar{A}_t\|_2 \leq 1-\delta$ for all t, uniformly in m_{t-1}.

(b) Bounded states. *For all t,*

$$\|\tilde{m}_t\|_2 \leq K, \qquad \|m_t\|_2 \leq K, \qquad \|g_t\|_2 \leq 2K, \qquad \|\bar{B}_t\|_2 \leq M_B := (1+2\eta_0)\,K. \quad (16)$$

(c) Bounded refinement loss. *The online refinement objective $\mathcal{L}_{\mathrm{omd}}$ (Eq. (13)) satisfies*

$$\mathcal{L}_{\mathrm{omd}} \;\geq\; -2\,\eta_0\,K^2\big(2\sqrt{T}-1\big) \;>\; -\infty. \quad (17)$$

Proof. **(a)** By the definition of $\bar{A}_t$ following Eq. (9), $\bar{A}_t = \frac{\rho_t I + c_t I}{1+c_t} = \alpha_t I$, so $\|\bar{A}_t\|_2 = |\alpha_t| = \alpha_t$. Since $\rho_t \in (0,1)$ and $c_t \geq 0$, we have $\alpha_t > 0$ and $1 - \alpha_t = (1-\rho_t)/(1+c_t) > 0$. For the uniform margin, note that $q_t \in (0,1)$ and (A1) give $\|[u_t; q_t]\|_2 \leq \sqrt{K_u^2+1}$, so by the Cauchy–Schwarz inequality $|w_\rho^\top [u_t; q_t] + b_\rho| \leq \|w_\rho\|_2\sqrt{K_u^2+1} + |b_\rho|$. Since σ is increasing, $\rho_t \leq \rho_{\max} < 1$. Combined with $c_t \leq 2\lambda$, this yields $1 - \alpha_t \geq (1-\rho_{\max})/(1+2\lambda) = \delta$. The bound does not depend on m_{t-1}, because m_{t-1} enters ρ_t and c_t only through $q_t \in (0,1)$.

(b) Since $g_t = \tilde{m}_t - u_t$, we have $\tilde{m}_t - \eta_t g_t = (1-\eta_t)\tilde{m}_t + \eta_t u_t$, and the closed form in Eq. (9) can be written as

$$m_t = \frac{1-\eta_t}{1+c_t}\,\tilde{m}_t + \frac{\eta_t}{1+c_t}\,u_t + \frac{c_t}{1+c_t}\,m_{t-1}, \qquad \tilde{m}_t = \rho_t\,m_{t-1} + (1-\rho_t)\,\Phi(u_t).$$

Because $0 < \eta_t \leq \eta_0 \leq 1$, both expressions are convex combinations: the weights are non-negative and sum to one. We argue by induction. The base case is $\|m_0\|_2 = 0 \leq K$. If $\|m_{t-1}\|_2 \leq K$, then (A2) and the triangle inequality give $\|\tilde{m}_t\|_2 \leq \rho_t K + (1-\rho_t)K_\Phi \leq K$. Using (A1) as well, $\|m_t\|_2 \leq K$. Consequently, $\|g_t\|_2 \leq \|\tilde{m}_t\|_2 + \|u_t\|_2 \leq 2K$. Finally, $\|\bar{B}_t\|_2 = \|B_t^m - \eta_t g_t\|_2/(1+c_t) \leq (1-\rho_t)K_\Phi + 2\eta_0 K \leq (1+2\eta_0)K$.

(c) In Eq. (13), the Bregman term $D_\psi(m_t, \tilde{m}_t) = \frac{1}{2}\|m_t - \tilde{m}_t\|_2^2 \geq 0$ and the stability term $\lambda(1-q_t)\|m_t - m_{t-1}\|_2^2 \geq 0$. By the Cauchy–Schwarz inequality and part (b), $\eta_t\langle g_t, m_t\rangle \geq -\eta_t\|g_t\|_2\|m_t\|_2 \geq -2K^2\eta_t$. Using $\sum_{t=1}^T t^{-1/2} \leq 1 + \int_1^T x^{-1/2}\,dx = 2\sqrt{T} - 1$, we obtain

$$\mathcal{L}_{\mathrm{omd}} \;\geq\; -2K^2\sum_{t=1}^{T}\eta_t \;\geq\; -2\,\eta_0\,K^2\big(2\sqrt{T}-1\big) \;>\; -\infty,$$

which completes the proof. □

Remark: The stop-gradient on g_t (Eq. (7)) acts only during backpropagation. In the forward pass, $g_t = \tilde{m}_t - u_t$ still depends on m_{t-1}, and hence so does $\bar{B}_t$. Part (b) therefore establishes boundedness directly from the convex-combination structure, rather than assuming $\|\bar{B}_t\|_2 \leq M_B$. Assumption (A1) holds automatically for any fixed trained network: the correlation-based adjacencies lie in $[-1,1]$, the node features and structural summaries are bounded, p_t is a fixed embedding, and f_{enc} is continuous. Assumption (A2) holds whenever Φ is an affine map or an MLP with Lipschitz activations.

Theorem 1 (Stability and Contraction)**.** *Let the assumptions of Lemma 1 hold, except that the initial memory may be any m_0 with $\|m_0\|_2 \le K$, with δ and K as defined there, and let $c_t = 2\lambda(1-q_t)$. Substituting $g_t = \tilde{m}_t - u_t$ into the closed form in Eq. (9), the refined memory update can be written exactly as*

$$m_t = F_t(m_{t-1};\, u) := \gamma_t\, m_{t-1} + b_t, \qquad \gamma_t = \frac{(1-\eta_t)\rho_t + c_t}{1+c_t}, \qquad b_t = \frac{(1-\eta_t)(1-\rho_t)\,\Phi(u_t) + \eta_t\, u_t}{1+c_t}, \tag{18}$$

where u denotes the input sequence (graph states and structural drifts). Both γ_t and b_t depend on m_{t-1} only through the regime score q_t. Assume further:

(A3) Lipschitz regime scorer: *g_Δ is L_g-Lipschitz in its memory argument, i.e., $|g_\Delta([u_t; m; \Delta_t]) - g_\Delta([u_t; m'; \Delta_t])| \le L_g \|m - m'\|_2$.*

Let w_q denote the entry of w_ρ that multiplies q_t, and define

$$L_\alpha := \Big(\frac{|w_q|}{4} + 2\lambda\Big)\frac{L_g}{4}, \qquad \kappa := (1-\delta) + 2KL_\alpha. \tag{19}$$

Then the following hold.

(a) BIBO stability. *If $\|m_0\|_2 \le K$, then $\|m_t\|_2 \le K$ for all $t = 1, \dots, T$.*

(b) Contraction. *If $\kappa < 1$, then for any two input sequences u, u' satisfying (A1)–(A2) with the same constants, and any initial states with $\|m_0\|_2, \|m_0'\|_2 \le K$, the trajectories $m_t = F_t(m_{t-1}; u)$ and $m_t' = F_t(m_{t-1}'; u')$ satisfy*

$$\|m_t - m_t'\|_2 \;\le\; \kappa^t\, \|m_0 - m_0'\|_2 + \frac{1}{1-\kappa}\, \max_{1\le k\le t} \varepsilon_k, \qquad \varepsilon_k := \sup_{\|m\|_2 \le K} \big\|F_k(m;u) - F_k(m;u')\big\|_2. \tag{20}$$

In particular, for identical inputs ($u = u'$), $\|m_t - m_t'\|_2 \le \kappa^t \|m_0 - m_0'\|_2$, i.e., the memory forgets its initialization geometrically.

Proof. **Exactness of Eq. (18).** From $g_t = \tilde{m}_t - u_t$ we have $\tilde{m}_t - \eta_t g_t = (1-\eta_t)\tilde{m}_t + \eta_t u_t$. Substituting $\tilde{m}_t = \rho_t m_{t-1} + (1-\rho_t)\Phi(u_t)$ into Eq. (9) gives Eq. (18). Note that $\gamma_t \le \alpha_t$, where α_t is the factor in Lemma 1(a), and that the weights of m_{t-1}, $\Phi(u_t)$ and u_t in Eq. (18) are non-negative and sum to one.

Part (a). This is the induction of Lemma 1(b), started from any $\|m_0\|_2 \le K$ instead of $m_0 = 0$. Since m_t is a convex combination of m_{t-1}, $\Phi(u_t)$ and u_t, whose norms are all at most K, we get $\|m_t\|_2 \le K$.

Part (b). *Step 1 (uniform contraction factor).* $1 - \gamma_t = \frac{1-(1-\eta_t)\rho_t}{1+c_t} \ge \frac{1-\rho_t}{1+c_t} \ge \delta$ by Lemma 1(a). Hence $\gamma_t \le 1 - \delta$ for every value of m_{t-1}.

Step 2 (Lipschitz dependence on memory). Fix the input and write $\gamma_t(m)$, $b_t(m)$ for the values obtained with $m_{t-1} = m$. Since $\sigma' \le 1/4$, (A3) gives $|q_t(m) - q_t(m')| \le \frac{L_g}{4}\|m - m'\|_2$. Consequently, $|\rho_t(m) - \rho_t(m')| \le \frac{|w_q|}{4}|q_t(m) - q_t(m')|$ and $|c_t(m) - c_t(m')| = 2\lambda|q_t(m) - q_t(m')|$. Viewing γ_t as a function of $(\rho, c) \in (0,1) \times [0, 2\lambda]$, its partial derivatives

$$\frac{\partial \gamma_t}{\partial \rho} = \frac{1-\eta_t}{1+c} \in [0,1], \qquad \frac{\partial \gamma_t}{\partial c} = \frac{1-(1-\eta_t)\rho}{(1+c)^2} \in [0,1]$$

are bounded by one, so the mean value theorem yields

$$|\gamma_t(m) - \gamma_t(m')| \le |\Delta\rho| + |\Delta c| \le \Big(\frac{|w_q|}{4} + 2\lambda\Big)|\Delta q| \le L_\alpha \|m - m'\|_2.$$

Similarly, $\partial b_t/\partial \rho = -(1-\eta_t)\Phi(u_t)/(1+c)$ and $\partial b_t/\partial c = -b_t/(1+c)$ have norms at most K, because $\|\Phi(u_t)\|_2 \le K$ and $\|b_t\|_2 \le (1-\gamma_t)K \le K$. Integrating along the segment in (ρ, c) gives $\|b_t(m) - b_t(m')\|_2 \le K(|\Delta\rho| + |\Delta c|) \le KL_\alpha \|m - m'\|_2$.

Step 3 (one-step contraction for a fixed input). For $\|m'\|_2 \le K$, adding and subtracting $\gamma_t(m)\, m'$ gives

$$\begin{aligned}\|F_t(m;u) - F_t(m';u)\|_2 &\le \gamma_t(m)\, \|m - m'\|_2 + |\gamma_t(m) - \gamma_t(m')|\, \|m'\|_2 \\ &\quad + \|b_t(m) - b_t(m')\|_2 \\ &\le \big(1 - \delta + KL_\alpha + KL_\alpha\big)\, \|m - m'\|_2 = \kappa\, \|m - m'\|_2.\end{aligned}$$

Step 4 (different inputs and unrolling). Let $d_t := m_t - m'_t$. By part (a), $\|m'_{t-1}\|_2 \le K$. Using Step 3 and the definition of ε_t,

$$\|d_t\|_2 \le \|F_t(m_{t-1};u) - F_t(m'_{t-1};u)\|_2 + \|F_t(m'_{t-1};u) - F_t(m'_{t-1};u')\|_2 \le \kappa\, \|d_{t-1}\|_2 + \varepsilon_t.$$

Since $\kappa < 1$, unrolling this recursion gives

$$\|d_t\|_2 \le \kappa^t \|d_0\|_2 + \sum_{k=1}^{t} \kappa^{t-k} \varepsilon_k \le \kappa^t \|d_0\|_2 + \frac{1}{1-\kappa} \max_{1 \le k \le t} \varepsilon_k,$$

which completes the proof. □

Remark: The decomposition $m_t = \bar{A}_t m_{t-1} + \bar{B}_t$ following Eq. (9) is an exact identity, but $\bar{B}_t$ contains $-\eta_t g_t$, and g_t depends on m_{t-1} in the forward pass. Eq. (18) isolates all memory dependence in γ_t and b_t, which avoids treating $\bar{B}_t$ as an exogenous input. Assumption (A3) holds whenever g_Δ is an MLP with Lipschitz activations. The condition $\kappa < 1$, equivalently $2KL_\alpha < \delta$, is sufficient rather than structural: it holds when the regime scorer is not too sensitive to the memory state relative to the retention margin δ.

C Dataset Details

In this section, we discuss the EEG and non-EEG datasets (DS1–DS10) used in our study.

Alzheimer's disease/healthy controls dataset (DS1): We utilized the publicly available EEG dataset (Miltiadous et al., 2023), which contains resting-state, eyes-closed EEG recordings for dementia analysis. The dataset comprises a total of 88 subjects, including 36 patients with Alzheimer's disease (AD), 23 with frontotemporal dementia (FTD), and 29 healthy controls (HC). The cohort includes both male and female participants with clinically confirmed diagnoses. EEG signals were recorded using a standard 10–20 system with 19 channels at a sampling frequency of 500 Hz. Each recording session lasts approximately 10–12 minutes under a resting-state, eyes-closed condition, ensuring minimal external cognitive influence.

Parkinson's disease/healthy controls dataset (DS2): The dataset (Cavanagh et al., 2021) consists of EEG recordings from 28 Parkinson patients and 28 HC subjects. EEG signals were recorded using 67 channels at a sampling frequency of 500 Hz during a Simon conflict task. Each subject participated in a single EEG recording session (duration: 20–30 minutes) consisting of approximately 240 trials, corresponding to repeated task-based epochs in a reinforcement-learning paradigm.

Motor Imagery dataset (DS3): The dataset (Peterson et al., 2020) consists of EEG recordings from 12 healthy subjects (4 females and 8 males). EEG signals were acquired using 15 channels positioned over the sensorimotor cortex (10–20 system) with a sampling frequency of 125 Hz. Each subject participated in a single session of approximately 1.5 hours, where each session included 4 runs of 40 trials each (20 trials per class), yielding a total of 160 trials per subject. The experiment followed a motor imagery (MI) paradigm with two conditions: imagined hand grasping and rest.

Mental arithmetic dataset (DS4): The dataset (Zyma et al., 2019) consists of EEG recordings from 36 healthy subjects (27 females, 9 males). EEG signals were acquired using a 23-channel system following the International 10–20 electrode placement with a sampling frequency of 500 Hz. The experiment involved a cognitive workload task based on mental arithmetic (serial subtraction) and a resting-state condition (eyes closed). Each recording includes two segments: 180 s of resting-state EEG and 60 s of mental arithmetic activity.

RSVP and P300 Speller EEG dataset (DS5): The dataset (Won et al., 2022) consists of EEG recordings from 55 healthy participants (14 females, 40 males). EEG signals were acquired using 32 electrodes placed according to the international 10–20 system, with a sampling frequency of 512 Hz. Each participant completed eyes-open/eyes-closed resting-state recordings, a rapid serial visual presentation (RSVP) task, and a P300 speller task. The RSVP task included 40 trials per participant, with 40 target and 800 non-target events. The P300 speller included two calibration runs and four test runs, where participants spelled 28 test letters in total.

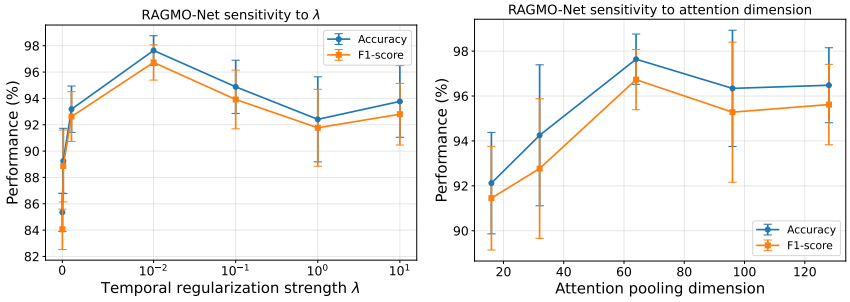
DGraph- anomaly detection dataset (DS6): The DGraph dataset (Huang et al., 2022) is a large-scale real-world dynamic graph from the financial domain designed for graph anomaly detection. It contains approximately 3.7M nodes (users) and 4.3M directed temporal edges representing emergency contact relationships between users. Each node corresponds to a user with a 17-dimensional feature vector derived from profile information, while edges are time-stamped to capture temporal graph dynamics. The dataset includes about 1.2M labeled nodes with ground-truth annotations, where anomalous nodes correspond to fraudsters (15,509) and normal users (1,210,092). In addition, DGraph contains approximately 2.47M background nodes without borrowing behavior, which are not direct prediction targets but preserve graph connectivity. The primary task is node-level anomaly detection in a dynamic graph setting, leveraging structural, feature, and temporal information.

TGBL-Coin (DS7): TGBL-Coin is a temporal cryptocurrency transaction dataset constructed from Stablecoin ERC20 transactions (Shamsi et al., 2022). Each node represents a blockchain address, and each directed edge represents a fund transfer from one address to another at a specific timestamp. The graph contains 638,486 nodes, 22,809,486 temporal edges, and 1,295,720 timestamps. The task is dynamic link prediction, i.e., predicting which destination address a given source address will interact with at a future time. TGBL-Coin models financial transaction behavior and includes weighted, directed edges.

TGBL-Comment (DS8): TGBL-Comment is a large-scale temporal social interaction graph built from Reddit reply interactions between 2005 and 2010 (Nadiri & Takes, 2022). Each node corresponds to a Reddit user, and each directed edge represents one user replying to another user at a specific time. The dataset contains 994,790 nodes, 44,314,507 temporal edges, and 30,998,030 timestamps, making it highly fine-grained temporally. The task is dynamic link prediction, where the goal is to predict whether a user will reply to another user at a given future time. Compared with TGBL-Coin, this dataset captures social communication rather than financial transactions and has many more timestamps.

TGBL-Flight (DS9): TGBL-Flight is a temporal air-traffic graph based on an international flight records dataset from 2019 to 2022 (Strohmeier et al., 2021). Nodes represent airports, and directed temporal edges represent flights between airports on a given day. The dataset contains 18,143 nodes, 67,169,570 temporal edges, and 1,385 timestamps. Node features include airport type, continent, ISO region, longitude, and latitude, while edge attributes include flight numbers. The task is dynamic link prediction, i.e., predicting whether a flight will occur between two airports on a future date. Unlike TGBL-Coin and TGBL-Comment, TGBL-Flight has fewer nodes but the largest number of edges, reflecting dense repeated transportation connections.

MOOC User Action dataset (DS10): The MOOC dataset (Kumar et al., 2019) is a temporal user-action interaction graph collected from an online course platform. Users correspond to students, items correspond to course activities such as videos and assignments, and temporal edges represent timestamped student actions, e.g., viewing a video or submitting an answer. The dataset contains 7,047 users, 98 course items, and 411,749 interactions. It also includes 4,066 dropout labels, making it suitable for user state-change prediction, particularly student dropout prediction from temporal learning behaviors.



(a) (b)

Figure 5: Sensitivity analysis of RAGMO-Net based on different hyperparameters: (a) Effect of temporal regularization strength λ on model performance, showing that moderate regularization ($\lambda = 10^{-2}$) achieves the best accuracy and F1-score, while very small or large values degrade performance. (b) Effect of attention pooling dimension, where the model demonstrates stable performance across different dimensions, with optimal results around 64. Error bars represent standard deviation across folds.

D Experimental Analysis

D.1 Hyperparameter Sensitivity Analysis

The sensitivity analysis on the experimental dataset with respect to the temporal regularization parameter λ is shown in Figure 5(a). As illustrated in the plot, performance initially improves as λ increases, reaching the best accuracy and F1-score around $\lambda = 10^{-2}$. This indicates that moderate regularization effectively balances stability and adaptability in memory updates. However, for larger values of λ, performance gradually degrades, suggesting that overly strong regularization restricts flexible memory adaptation. Overall, the results demonstrate that RAGMO-Net is relatively robust across a range of λ values, with optimal performance achieved at an intermediate regularization strength.

The attention-dimension sensitivity analysis (refer to Figure 5(b)) shows that RAGMO-Net remains stable across different attention hidden sizes, with only minor performance variation from 16 to 96 dimensions. The best accuracy is obtained at dimension 64, suggesting that this setting provides a good balance between expressive temporal weighting and model complexity. The comparable performance across neighboring dimensions also indicates that the model is not highly sensitive to this hyperparameter, supporting the robustness of the attention-pooling module.

E Statistical Analysis

We applied a one-sided exact sign test, which is assumption-free and exact for small samples, to assess whether RAGMO-Net's accuracy improvements over the strongest baseline are statistically significant, under the null hypothesis that wins and losses are equally likely. The number of paired observations reflects each dataset's evaluation protocol: $n = 5$ for EEG datasets (subject-wise 5-fold CV) and $n = 10$ for non-EEG datasets (chronological partition splits), with minimum achievable p-values of 0.031 and 0.001, respectively.

Table 6 reports the results. RAGMO-Net wins strictly on every split for 10 out of 11 datasets ($p \leq 0.031$), and on 9 out of 10 splits for DS6 ($p = 0.011$), confirming that the observed gains are not attributable to random variation in data splits. For non-EEG datasets, RAGMO-Net wins all ten splits for DS7–DS10 ($p = 0.001$) and nine of ten for DS6 ($p = 0.011$), where the accuracy margin is narrowest ($+2.49\%$).

Table 6: One-sided sign test results comparing RAGMO-Net with the best competing baseline per dataset. *Wins* denotes splits where RAGMO-Net achieves higher accuracy; n denotes total splits. For EEG and non-EEG datasets, $n = 5$ and $n = 10$, respectively. All comparisons are significant ($p < 0.05$).

EEG Datasets (5-fold CV, $n = 5$)					**Non-EEG Datasets (10 splits, $n = 10$)**				
Dataset	Best baseline	Wins	n	p-value	Dataset	Best baseline	Wins	n	p-value
Exp DS	TGN (Rossi et al., 2020)	5	5	0.031	DS6	TGN (Rossi et al., 2020)	9	10	0.011
DS1	EvolveGCN (Pareja et al., 2020)	5	5	0.031	DS7	ST-GCN (Yan et al., 2018)	10	10	0.001
DS2	TGN (Rossi et al., 2020)	5	5	0.031	DS8	EvolveGCN (Pareja et al., 2020)	10	10	0.001
DS3	EvolveGCN (Pareja et al., 2020)	5	5	0.031	DS9	EvolveGCN (Pareja et al., 2020)	10	10	0.001
DS4	EvolveGCN (Pareja et al., 2020)	5	5	0.031	DS10	EvolveGCN (Pareja et al., 2020)	10	10	0.001
DS5	TGN (Rossi et al., 2020)	5	5	0.031					

F PSEUDOCODE

Algorithm 1: RAGMO-Net Model

Input: Dynamic graph sequence $\{(G_t, p_t)\}_{t=1}^{T}$, initial memory $m_0 = \mathbf{0}$
Output: Final class probability $\hat{y}$

1 $u_0 \leftarrow \mathbf{0}, \quad m_0 \leftarrow \mathbf{0}$; `// Initialize graph state and memory`
2 **for** $t = 1$ **to** T **do**
3 **Graph Encoding & State Preparation**
4 $h_t, s_t \leftarrow f_{\text{enc}}(G_t, p_t)$; `// Graph encoder + structural summary`
5 $u_t \leftarrow [h_t;\ s_t;\ p_t]$; `// Final graph state`
6 **Regime Score Computation**
7 $\Delta_t \leftarrow \Gamma(u_t, u_{t-1})$; `// Structural drift`
8 $q_t \leftarrow \sigma\big(g_\Delta([u_t;\ m_{t-1};\ \Delta_t])\big)$; `// Regime score` $q_t \in (0,1)$
9 **Provisional Affine Graph-Memory Cell**
10 $\rho_t \leftarrow \sigma\big(w_\rho^\top [u_t;\ q_t] + b_\rho\big)$;
11 $A_t^m \leftarrow \rho_t\, I, \quad B_t^m \leftarrow (1-\rho_t)\,\Phi(u_t)$;
12 $\tilde{m}_t \leftarrow A_t^m m_{t-1} + B_t^m$; `// Provisional memory`
13 **Regime-Aware Proximal Refinement**
14 $g_t \leftarrow \nabla_m \ell_t(m;\ u_t)\big|_{m=\tilde{m}_t}$; `// Frozen surrogate gradient`
15 $m_t \leftarrow \dfrac{\tilde{m}_t - \eta_t g_t + 2\lambda(1-q_t)m_{t-1}}{1+2\lambda(1-q_t)}$; `// Closed-form refinement (Eq. 9)`
16 $u_{t-1} \leftarrow u_t, \quad m_{t-1} \leftarrow m_t$; `// Update for next step`

17 **Classification Step (after memory computation):**
18 **for** $t = 1$ **to** T **do**
19 $z_t \leftarrow [u_t;\ m_t]$; `// Joint graph-memory state`
20 $\beta_t \leftarrow \dfrac{\exp\big(a([z_t;\ p_t])\big)}{\sum_{i=1}^{T} \exp\big(a([z_i;\ p_i])\big)}$; `// Attention weights`
21 $z_{\text{pool}} \leftarrow \sum_{t=1}^{T} \beta_t z_t$; `// Temporal pooling`
22 $\hat{y} \leftarrow \sigma(w^\top z_{\text{pool}} + b)$; `// Final class probability (Dementia vs. Healthy)`
23 **return** $\hat{y}$

The complete pseudocode of the RAGMO-Net model, including memory updating and final classification, is presented in Algorithm 1. The algorithm first initializes the graph state and memory as $u_0 = \mathbf{0}$ and $m_0 = \mathbf{0}$ (line 1), and then iterates over each time step $t = 1, \ldots, T$ (lines 2–16) to compute the regime-aware memory states. Within each iteration, lines 4–5 perform Graph Encoding & State Preparation, lines 7–8 perform Regime Score Computation, lines 10–12 implement the Provisional Affine Graph-Memory Cell, and lines 14–15 perform Regime-Aware Proximal Refinement. Line 16

updates u_{t-1} and m_{t-1} for the next iteration. The classification stage (lines 17–23) then forms the joint graph-memory states (lines 18–19), computes the attention weights, the pooled representation, and the final class probability (lines 20–22), and returns $\hat{y}$ (line 23). This block-wise organization makes the update rule transparent, easy to implement, and directly aligned with the mathematical formulation in Section 3.