C-SHARE: CROSS-SUBJECT HETEROGENEITY-AWARE RETRIEVAL AUGMENTATION FOR EARLY CAP PHASE A FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cyclic alternating pattern (CAP), an established EEG marker of NREM sleep instability and fragmentation, comprises recurrent Phase A activations alternating with Phase B intervals. Forecasting Phase A onsets could enable anticipatory monitoring and inform future interventions aimed at improving sleep quality. Realizing this potential requires overcoming three challenges: sparse event annotations, physiologically heterogeneous precursors, and generalization to unseen subjects. Parametric forecasters pool these limited observations into stable population-level predictions. Retrieval augmentation can use the same training data differently, retrieving source events as empirical corrections, but the retrieved observations do not necessarily constitute independent or transferable evidence. We introduce C-SHARE, a model-agnostic plug-in retrieval-augmentation module for frozen EEG forecasters. C-SHARE groups overlapping pre-onset windows into onset-indexed episodes, learns future-conditioned multi-subspace keys for predictive matching, and gates each correction using event- and subject-level support. In addition, we contribute NapCAP-128Ch, to our knowledge the first high-density CAP EEG dataset and the largest CAP-annotated EEG cohort by subject count, comprising 120 subjects. Across eight backbones and three horizons on this cohort and the public CAP Sleep Database, C-SHARE improves over the corresponding frozen backbone in AUPRC in all 48 dataset–backbone–horizon settings and in AUROC in 42 of 48, yielding more consistent gains than general-purpose retrieval baselines.

1 INTRODUCTION

Sleep staging is the gold standard for characterizing macroscopic sleep architecture, including sleep efficiency, wake after sleep onset, total sleep time, and sleep-stage transitions (Stanchina et al., 2005; Eldele et al., 2021; Phan et al., 2022), but provides limited resolution of sleep microstructure. We therefore study early forecasting of cyclic alternating pattern (CAP) Phase A events (Terzano et al., 1990). CAP comprises recurrent Phase A activations and Phase B background intervals, and CAP rate is an established marker of NREM sleep instability and fragmentation that complements macro-level sleep measures (Terzano et al., 1985; 2001; Parrino et al., 2012; 2026). To our knowledge, existing automated CAP methods focus on detecting Phase A after event-specific amplitude, frequency, or waveform patterns are already observable, treating the problem as offline labeling (Hartmann & Baumert, 2019; You et al., 2022; Murarka et al., 2022; Avila et al., 2025). In contrast, early Phase A forecasting predicts whether an onset will occur in a 10-second interval beginning 10, 20, or 30 seconds after the observation window, before its morphology becomes directly visible. Such advance warning could eventually support proactive monitoring and closed-loop interventions, for example through phase-locked auditory stimulation (Ngo et al., 2015; Besedovsky et al., 2017).

Forecasting must instead extract predictive structure from noisy, nonstationary pre-event EEG. Yet observation windows preceding a Phase A onset can resemble non-event windows, leaving no consistent precursor morphology visually apparent (Figure 1a). Phase A itself is not expressed through a single canonical waveform. Established scoring rules encompass manifestations ranging from high-amplitude slow transients and K-complex sequences to mixed-frequency bursts and cortical arousals (Terzano et al., 2001; Ferri et al., 2005; Parrino et al., 2026). Within a subject, distinct Phase A man-

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.

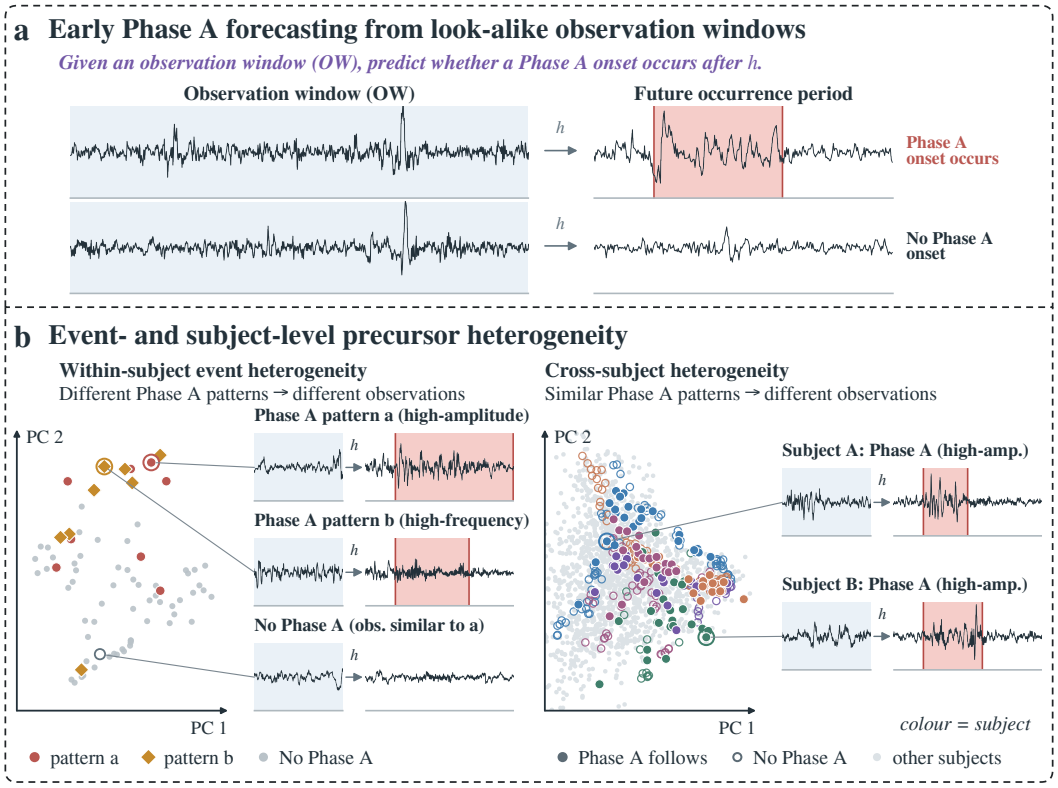


Figure 1: Early Phase A forecasting and why a single precursor template fails. **(a)** Given an observation window, the task predicts whether a Phase A onset occurs in a future occurrence period after horizon h. Windows followed by an onset and by no onset can look alike. **(b)** Precursors are heterogeneous. *Within subjects*, distinct Phase A patterns follow distinct precursors, and non-event windows can mimic them. *Across subjects*, similar Phase A patterns follow different precursors (colour: subject). Points are PCA projections of real windows.

ifestations can be preceded by different, visually non-specific patterns (Figure 1b). Across subjects, similar Phase A events can arise from different EEG backgrounds and pre-onset expressions (also Figure 1b). Early forecasting therefore cannot rely on a single precursor template shared across events and subjects.

These heterogeneous precursors are sparsely represented in available EEG cohorts. A parametric forecaster pools limited observations into a stable population-level prediction rule, while retrieval augmentation preserves direct access to individual source events at inference time. Existing methods integrate retrieved examples through input conditioning, representation mixing, or prediction-level fusion (Han et al., 2025; Du et al., 2026; Ning et al., 2025). To our knowledge, however, retrieval augmentation has not previously been applied to EEG event-onset forecasting. In this setting, deployment requires cross-subject retrieval. Every reference for an unseen subject comes from another individual, so a close match need not preserve its association with future Phase A risk (Hartmann & Baumert, 2023; Jemal et al., 2024) (Figure 1b). Uncontrolled use of retrieved evidence can therefore introduce negative transfer rather than improve the forecast (Liang et al., 2024; Ding et al., 2026). The central challenge is therefore to identify which source events provide transferable evidence and how strongly that evidence should influence the population forecast.

We propose C-SHARE (**C**ross-**S**ubject **H**eterogeneity-**A**ware **R**etrieval Augmentation for **E**arly CAP Phase A forecasting), a model-agnostic plug-in for forecasting Phase A onsets in unseen subjects. It augments a frozen forecaster with a cross-subject episodic memory. C-SHARE separates retrieval augmentation into two decisions: which source episodes provide transferable predictive evidence and how strongly that evidence should alter the base forecast. Onset-centered memory and future-conditioned multi-subspace retrieval address the first decision by reducing event redundancy and

identifying partial predictive matches across subjects. A lightweight reliability controller addresses the second by gating each retrieval-proposed correction according to the correction itself and its event- and subject-level support.

We evaluate C-SHARE on NapCAP-128Ch, a 120-subject high-density EEG cohort acquired in our laboratory with expert annotations of CAP Phase A events, and further evaluate it on the public CAP Sleep Database (Terzano et al., 2001; Pollard et al., 2026). Across both datasets, eight backbones, and three horizons, C-SHARE improves AUPRC in all 48 dataset–backbone–horizon settings and AUROC in 42 of 48, with more consistent gains than general-purpose retrieval baselines.

Our contributions are summarized as follows:

- We contribute NapCAP-128Ch, our newly collected 30-minute afternoon-nap high-density EEG (128 channels, GSN-HydroCel-129 net) from 120 healthy adults aged 19–79 years under a standardized protocol with institutional ethics approval and written informed consent. Two experienced experts independently annotated Phase A onsets and offsets following established CAP criteria (Terzano et al., 2001; Parrino et al., 2026) (inter-rater $\kappa = 0.69$ on 2-second windows), and disagreements were resolved into consensus labels. To our knowledge, NapCAP-128Ch is the first high-density CAP EEG dataset and the largest CAP-annotated EEG cohort by subject count.[1]
- To our knowledge, this is the first application of retrieval augmentation to EEG event-onset forecasting, providing a complementary approach to sparse, heterogeneous cross-subject prediction. Parametric models provide stable population-level predictions, while retrieval keeps individual source events explicitly accessible at inference time to form query-specific empirical corrections.
- We propose C-SHARE, a model-agnostic retrieval-augmentation plug-in for EEG forecasters that requires neither a large foundation model nor collaboration between large and small models. Its retrieval encoder and memory can be reused without modification across backbones. Only a linear controller is refitted for each forecaster.

2 RELATED WORK

Retrieval-augmented forecasting complements parametric models with explicit memories of observed histories and futures. RAFT retrieves similar training subsequences and exposes their continuations to a forecaster (Han et al., 2025); PFRP learns predictive similarity over a global memory bank (Du et al., 2026); and TS-RAG integrates retrieved segments with time-series foundation models (Ning et al., 2025). PIR performs instance-aware post-hoc forecast revision (Liu et al., 2025), whereas SARAF uses dataset-level stationarity to control retrieval selection and aggregation before combining the base and retrieval forecasts with fixed equal weights (Zhou et al., 2026). Meanwhile, case-based physiological prediction has retrieved related patients or trajectories for clinical forecasting (Ebadollahi et al., 2010).

These methods, however, are designed primarily for continuous-value forecasting or general post-hoc revision rather than sparse, cross-subject event-onset forecasting. In our setting, multiple overlapping windows can redundantly represent the same rare event, while deployment is strictly subject-disjoint. Cross-subject physiological heterogeneity can therefore make apparently similar references non-transferable. To our knowledge, C-SHARE is the first retrieval-augmented method for EEG event-onset forecasting. It addresses event redundancy and cross-subject non-transferability through onset-centered episodic memory, predictive cross-subject retrieval, and reliability-controlled empirical correction.

3 METHOD

3.1 PROBLEM FORMULATION AND METHOD OVERVIEW

Let $x_{u,t} \in \mathbb{R}^{C \times L}$ denote a multichannel EEG history of subject u ending at anchor time t. Following EEG event-forecasting formulations (Slimen et al., 2020; Jemal et al., 2024), for $h \in \mathcal{H} =$

[1] Code and dataset are provided in the supplementary material.

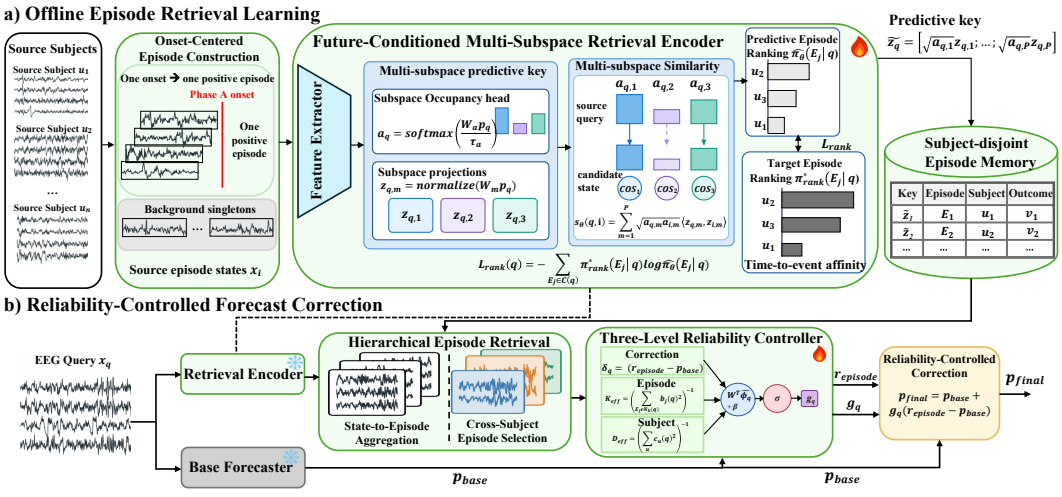


Figure 2: Overview of C-SHARE. **(a) Offline episode retrieval learning:** onset-centered episodes and future-conditioned multi-subspace ranking yield predictive keys for a subject-disjoint episodic memory. **(b) Reliability-controlled forecast correction:** a query from an unseen subject hierarchically retrieves cross-subject episodes, and a three-level controller gates the proposed correction $r_{\text{episode}} - p_{\text{base}}$ by the correction itself and its episode/subject support, revising rather than replacing the frozen base forecast. Flames/snowflakes: trainable/frozen; green: new components.

$\{10, 20, 30\}$ seconds we define

$$y_{u,t}^{(h)} = \mathbb{1}[\exists (s,e) \in \mathcal{A}_u : s \in (t+h, t+h+10]]\,, \tag{1}$$

where $\mathcal{A}_u$ is the set of annotated Phase A intervals and s denotes onset time. Targets are defined independently across horizons; the complete sampling protocol is in Appendix C. Under leave-one-subject-out (LOSO) evaluation, the held-out target subject $u^\star$ is excluded from every fitted component and the retrieval memory; the remaining subjects form the source set.

For a target query $q = (u^\star, t)$, let $p_{\text{base},q}^{(h)}$ denote the base prediction, $r_{\text{episode},q}^{(h)}$ the retrieved evidence, and $g_q^{(h)} \in [0,1]$ the learned gate. The final prediction is

$$p_{\text{final},q}^{(h)} = p_{\text{base},q}^{(h)} + g_q^{(h)} \left(r_{\text{episode},q}^{(h)} - p_{\text{base},q}^{(h)} \right), \tag{2}$$

so retrieval proposes the correction $r_{\text{episode},q}^{(h)} - p_{\text{base},q}^{(h)}$ rather than replacing the base forecast. Figure 2 shows the two stages. *Offline episode retrieval learning* constructs the episodic memory and learns which episodes are predictive analogues. *Reliability-controlled forecast correction* retrieves query-specific evidence and gates its effect. The frozen encoder and memory are reused across backbones; only the controller is refitted.

3.2 Offline Episode Retrieval Learning

The first stage requires a retrieval unit organized around sparse events and a similarity criterion that captures predictive relevance rather than morphological resemblance.

Onset-centered episode construction. Unlike window-indexed forecasting memories (Han et al., 2025; Du et al., 2026), we group positive windows that contain the same source onset o in their occurrence period into one episode:

$$E_{u,o}^{(h)} = \{x_{u,t_m} : o \in (t_m + h, t_m + h + 10]\}_{m=1}^{M_{u,o}}\,. \tag{3}$$

Background windows remain singleton episodes. The resulting memory is $\mathcal{M}^{(h)} = \{(E_j, u_j, v_j)\}_{j=1}^{N_E}$, where u_j and $v_j \in \{0,1\}$ identify the source subject and outcome. This preserves multiple precursor states while giving each event one outer-level vote.

Future-conditioned multi-subspace retrieval learning. For each horizon, an independent retrieval encoder comprises a feature extractor $f_\theta^{(h)}$, a subspace occupancy head, and P subspace projection heads. The feature extractor maps the query's sequence of two-second spectral segments to a pooled representation $p_q^{(h)} = f_\theta^{(h)}(x_q)$. To avoid one global criterion for heterogeneous EEG expressions, the occupancy head produces $a_q^{(h)} = \mathrm{softmax}(W_a^{(h)} p_q^{(h)}/\tau_a)$, while each projection produces $z_{q,m}^{(h)} = \mathrm{normalize}(W_m^{(h)} p_q^{(h)})$. For a query q and a candidate state i, their multi-subspace similarity is

$$s_\theta(q,i) = \sum_{m=1}^{P} \sqrt{a_{q,m}^{(h)} a_{i,m}^{(h)}} \left\langle z_{q,m}^{(h)}, z_{i,m}^{(h)} \right\rangle . \tag{4}$$

Subspaces assigned high occupancy by both the query and candidate contribute most. Each window exports the equivalent composite key $\tilde{z}_q^{(h)} = [\sqrt{a_{q,1}} z_{q,1}; \ldots; \sqrt{a_{q,P}} z_{q,P}]$.

Future annotations supervise retrieval relations rather than a prediction head. Each state has a Gaussian-smoothed time-to-event distribution $\bar{d}_q^{(h)}$ with a no-event coordinate. For a source query q, only cross-subject episodes $\mathcal{C}(q) = \{E_j : u_j \neq u_q\}$ are candidates. The future affinity of a candidate state $x_{jm} \in E_j$ is

$$\alpha^\star(q, x_{jm}) = \exp\left(-\frac{\|\bar{d}_q^{(h)} - \bar{d}_{jm}^{(h)}\|_1}{\tau_f} \right). \tag{5}$$

We aggregate the strongest state-level matches without favoring long episodes using the length-normalized operator

$$\mathrm{Agg}_{\tau,R}(\mathcal{V}) = \tau \log \left[\frac{1}{r} \sum_{v \in \mathrm{Top}_r(\mathcal{V})} \exp(v/\tau) \right], \qquad r = \min(R, |\mathcal{V}|). \tag{6}$$

Applying the operator and normalizing over $\mathcal{C}(q)$ yields the target ranking $\pi_{\mathrm{rank}}^\star(E_j \mid q)$ from future affinities and the predicted ranking $\widehat{\pi}_\theta(E_j \mid q)$ from Eq. 4. We optimize the listwise objective (Swezey et al., 2021)

$$\begin{aligned} \mathcal{L}_{\mathrm{rank}}(q) &= -\sum_{E_j \in \mathcal{C}(q)} \pi_{\mathrm{rank}}^\star(E_j \mid q) \log \widehat{\pi}_\theta(E_j \mid q), \\ \mathcal{L}_{\mathrm{enc}} &= \mathbb{E}_q[\mathcal{L}_{\mathrm{rank}}(q)] + \beta_{\mathrm{bal}} \mathcal{L}_{\mathrm{bal}} + \beta_{\mathrm{ent}} \mathcal{L}_{\mathrm{ent}}, \end{aligned} \tag{7}$$

Here, $\mathcal{L}_{\mathrm{bal}}$ prevents batch-level subspace collapse, whereas $\mathcal{L}_{\mathrm{ent}}$ encourages sparse, low-entropy occupancy for each state. Appendix C details the feature-extractor architecture, training targets, and regularizers.

3.3 Reliability-Controlled Forecast Correction

At inference, the frozen retrieval encoder maps an unseen-subject query to a predictive key used to retrieve evidence from the fixed source memory; the controller then gates its effect on the base forecast.

Hierarchical episode retrieval. After centering and renormalizing composite keys by the source-memory mean, let s_{jm} denote query–state cosine similarity. State-to-episode aggregation retains the strongest states without giving longer episodes more outer-level weight:

$$S_\theta^{\mathrm{ret}}(q, E_j) = \mathrm{Agg}_{\tau_{\mathrm{in}}, M_{\mathrm{in}}}(\{s_{jm} : x_{jm} \in E_j\}) . \tag{8}$$

Cross-subject episode selection then applies a per-subject cap, selects the top-K episodes $\mathcal{N}_K(q)$, and aggregates their outcomes:

$$b_j(q) = \frac{e^{S_\theta^{\mathrm{ret}}(q,E_j)/\tau_{\mathrm{out}}}}{\sum_{E_l \in \mathcal{N}_K(q)} e^{S_\theta^{\mathrm{ret}}(q,E_l)/\tau_{\mathrm{out}}}}, \qquad r_{\mathrm{episode},q}^{(h)} = \sum_{E_j \in \mathcal{N}_K(q)} b_j(q) v_j. \tag{9}$$

$r_{\mathrm{episode},q}^{(h)}$ is query-conditional evidence, not an independently calibrated probability.

Three-level reliability control. Similarity determines which episodes contribute, but not whether their evidence should alter the forecast. The controller uses the proposed correction $\delta_q^{(h)} = r_{\text{episode},q}^{(h)} - p_{\text{base},q}^{(h)}$ and its episode- and subject-level support. Let $c_u(q) = \sum_{j:u_j=u} b_j(q)$ be the total weight assigned to source subject u. The effective numbers of episodes and source subjects are

$$K_{\text{eff},q}^{(h)} = \left(\sum_{E_j \in \mathcal{N}_K(q)} b_j(q)^2 \right)^{-1}, \qquad D_{\text{eff},q}^{(h)} = \left(\sum_u c_u(q)^2 \right)^{-1}. \tag{10}$$

They measure weight concentration at the two levels: larger values indicate support from more episodes and source subjects, respectively. After source-based standardization, a horizon-specific linear controller maps the three descriptors to the gate:

$$\widetilde{\phi}_q^{(h)} = \text{Std}\left([\delta_q^{(h)}, K_{\text{eff},q}^{(h)}, D_{\text{eff},q}^{(h)}] \right), \qquad g_q^{(h)} = \sigma\left((w_h)^\top \widetilde{\phi}_q^{(h)} + \beta_h \right). \tag{11}$$

Binary cross-entropy on Eq. 2 trains the controller without query-level retrieval-utility labels.

Controller training and inference. We train and freeze the base forecaster and retrieval encoder on source subjects, then fit the controller on source-subject queries whose retrieval features are out-of-fold: each query retrieves only from source subjects outside its own subject group. Inference requires only target EEG, without target labels or subject-specific adaptation; Appendix C details optimization and leakage prevention.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two EEG cohorts: our newly collected NapCAP-128Ch dataset and the PhysioNet CAP Sleep Database (Terzano et al., 2001; Pollard et al., 2026), to our knowledge the only widely used public CAP-annotated EEG benchmark.

NapCAP-128Ch. We recorded 30-minute afternoon-nap high-density EEG (128 channels, GSN-HydroCel-129 net) from 120 healthy adults.

PhysioNet CAP Sleep Database. We use the full overnight recordings of 16 healthy controls. Because no multichannel montage is shared across all recordings, we use one central derivation per subject (C4–A1, or C3–A2 when unavailable).

Baselines and evaluation. As no dedicated CAP forecasting model exists, we evaluate eight frozen backbones spanning general time-series models (PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023), TSLANet (Eldele et al., 2024), S-Mamba (Wang et al., 2025)), EEG/medical models (EEGNet (Lawhern et al., 2018), MEDformer (Wang et al., 2024)), and CAP detectors retargeted to forecasting (GTransU (You et al., 2022), MuRAT (Yaman et al., 2025)). We also compare with cross-subject adaptations of RAFT (Han et al., 2025) and SARAF (Zhou et al., 2026) that preserve their retrieval mechanisms and use a source-fitted query-level fusion gate. All methods follow Leave-One-Subject-Out (LOSO) evaluation. We report AUROC and AUPRC; Appendices A and C provide dataset statistics and implementation details.

4.2 Experimental Results

4.2.1 Results on NapCAP-128Ch

Across eight backbones and three horizons, C-Share improves AUPRC in all 24 settings (mean gain: 6.19 points) and AUROC in 22 of 24 (mean gain: 3.15 points), with significant gains in all 24 AUPRC settings and 18 of 24 AUROC settings (paired Wilcoxon tests, Holm-corrected; Appendix I). It outperforms RAFT in 23 of 24 AUPRC settings and SARAF in all 24, and in AUROC it outperforms RAFT in 20 of 24 and SARAF in 22 of 24. Results across convolutional, Transformer, and state-space backbones support the plug-in use of C-Share across architectures and the value of its event-structured, reliability-controlled retrieval. Full values and significance tests are reported in Appendix Table 4 and Appendix I.

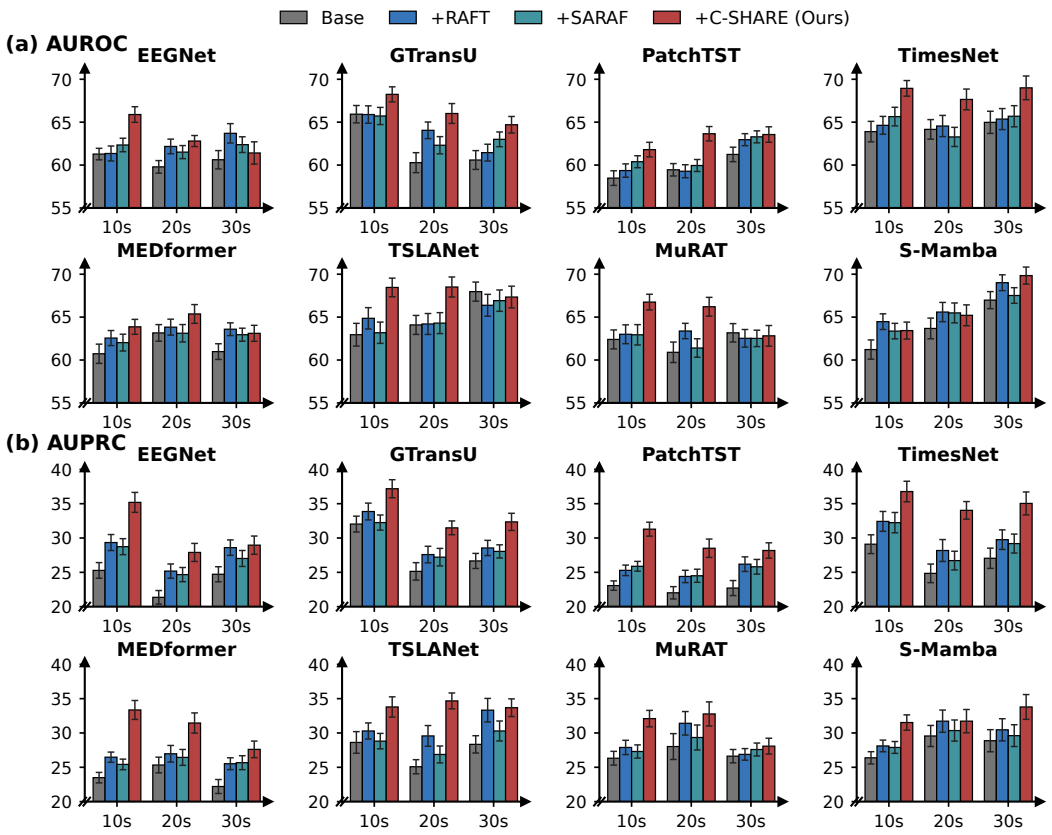


Figure 3: Leave-one-subject-out (LOSO) performance on NapCAP-128Ch (120 subjects). Grouped bars compare each frozen backbone with RAFT, SARAF, and C-SHARE at three horizons: **(a)** AUROC and **(b)** AUPRC. Each subject serves once as the unseen test subject. Bars show mean performance across held-out subjects, and error bars denote SEM across those subjects. Numerical results are in Appendix Table 4.

4.2.2 RESULTS ON THE CAP SLEEP DATABASE

Results on the public CAP Sleep Database show the same overall trend. C-SHARE improves AUPRC in all 24 settings (mean gain: 1.64 points) and AUROC in 20 of 24 (mean gain: 1.87 points; Figure 4). It also outperforms RAFT and SARAF on both metrics throughout the 24-setting grid. Gains are smaller than on NapCAP-128Ch. This difference may reflect the single-channel montage, which provides less spatial information for matching precursor patterns, and the smaller pool of 15 source subjects per fold, which limits subject-level diversity even though each subject contributes hundreds of events. The result shows that C-SHARE remains effective with single-channel overnight EEG, while suggesting that richer spatial information and more diverse source subjects may provide additional benefit.

4.3 ABLATION STUDY

Table 1 isolates the contribution of C-SHARE's three design choices: onset-centered episodic memory, the reliability controller, and the multi-subspace retrieval encoder. All variants use the same frozen TimesNet backbone and evaluation protocol as Table 4. Ablations on the other backbones are reported in Appendix H.

Retrieval unit. Flat window-level retrieval lets overlapping windows from one onset occupy several retrieved slots and never outperforms episode retrieval at $g{=}1$ or $g{=}0.5$, supporting each onset as a single retrieval unit.

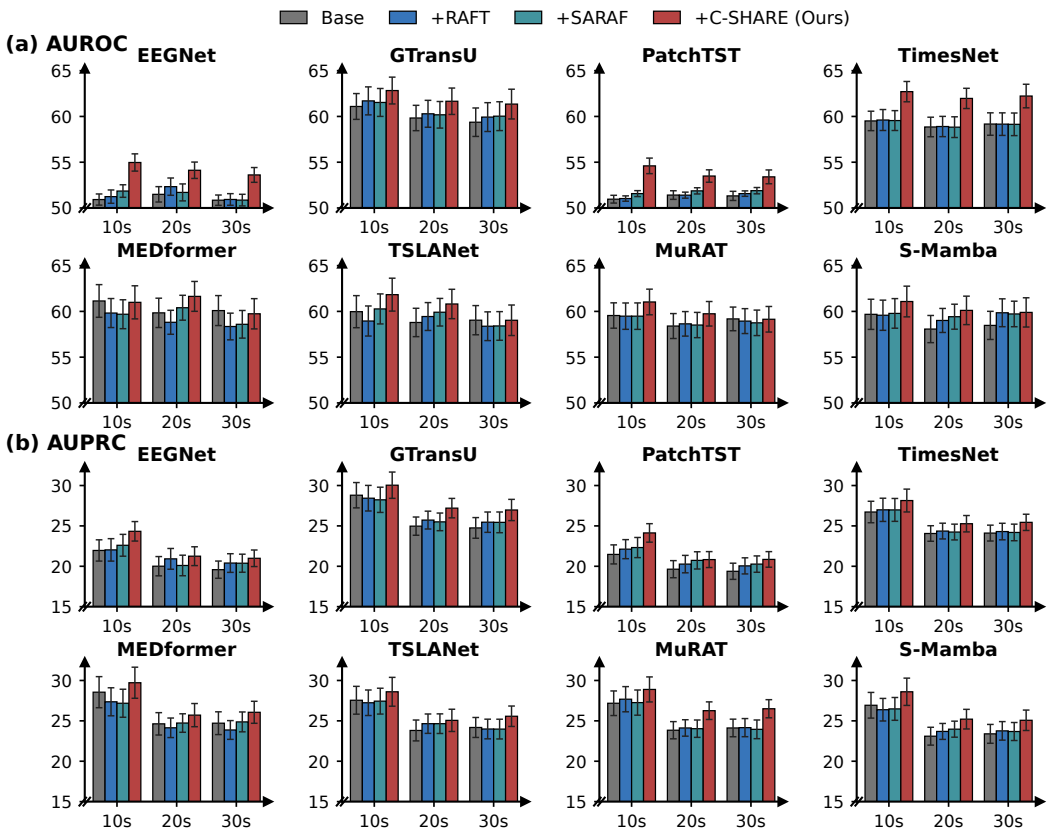


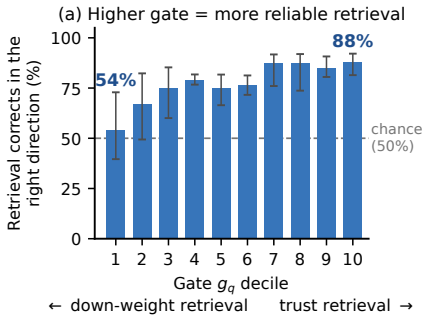
Figure 4: Leave-one-subject-out (LOSO) performance on the public CAP Sleep Database (16 subjects): **(a)** AUROC and **(b)** AUPRC. Each subject serves once as the unseen test subject. Bars show mean performance across held-out subjects, and error bars denote SEM across those subjects. Numerical results are in Appendix Table 5.

Table 1: Component ablation on frozen TimesNet across three prediction horizons. Columns correspond to the three design choices; the controller column gives the gate g in Eq. 2. The last row is the complete C-SHARE. AUROC and AUPRC are percentages reported as mean ± standard deviation across held-out subjects. Ablations on the other backbones are reported in Appendix H.

Retrieval unit	Reliability controller	Multi-subspace encoder	10s		20s		30s	
			AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
–	g=0 (TimesNet only)	–	63.90±13.08	29.10±14.99	64.16±12.49	24.85±14.85	64.98±14.16	27.05±16.16
Window	g=1 (retrieval only)	✓	55.82±12.90	28.61±12.61	51.98±13.38	20.12±11.98	54.12±10.86	20.42±11.13
Episode	g=1 (retrieval only)	✓	56.85±11.91	28.83±12.99	55.94±11.71	24.18±13.47	54.42±10.01	21.52±10.79
Window	g=0.5 (fixed)	✓	60.15±13.88	31.13±17.30	59.58±10.06	23.44±14.62	60.92±14.13	24.66±13.45
Episode	g=0.5 (fixed)	✓	62.10±13.11	31.13±17.83	62.54±10.05	26.77±16.20	62.39±14.01	26.40±14.45
Episode	global scalar g	✓	64.50±11.61	35.84±17.06	65.95±11.11	29.95±17.00	66.65±13.40	30.92±15.81
Episode	query-level g_q	✗	68.67±11.91	36.71±15.98	63.71±11.01	32.32±14.93	67.25±12.71	33.79±17.99
Episode	query-level g_q	✓	**68.94**±9.97	**36.77**±16.44	**67.65**±13.24	**34.03**±14.03	**69.00**±15.10	**35.04**±18.30

Reliability controller. Used alone (g=1) or blended at a fixed g=0.5, retrieval does not reliably improve TimesNet; a learned global g improves both metrics at every horizon, and the query-level gate g_q performs best throughout. Because per-subject AUROC and AUPRC are invariant to monotone recalibration of the base forecast, the global-g gain shows that retrieval adds ranking information, which g_q exploits further; RAFT and SARAF use the same gate type yet gain less (Table 4).

Multi-subspace encoder. With g_q kept, a single-subspace encoder lowers both metrics at all horizons, most at 20 and 30 seconds ($-3.94/-1.75$ AUROC, $-1.71/-1.25$ AUPRC points) and only marginally at 10 seconds. At 20 seconds, the multi-subspace advantage grows with the number of positive episodes in memory (Appendix E).



(b) Controller-feature ablation

Controller features	Benefit AUROC	AUPRC
Correction only	0.589 [0.479, 0.687]	35.67
Episode support only	0.539 [0.428, 0.629]	34.66
Source-subject support only	0.572 [0.471, 0.656]	35.86
Correction + Episode support + Source-subject support	**0.648** [0.553, 0.710]	**36.77**

Figure 5: Query-level controller reliability (TimesNet, 10s). **(a)** Percentage of queries for which retrieval moves the base prediction toward the true label (c_q=1), grouped by gate-value decile. **(b)** Controller-input ablation; Benefit AUROC evaluates identification of c_q=1, and AUPRC the final forecast. Error bars and bracketed AUROCs show subject-clustered bootstrap 95% CIs.

4.4 Analysis

4.4.1 Does the Controller Identify Useful Retrieval?

To test whether the controller identifies useful retrieval, we define $c_q = \mathbb{1}[(r_{\text{episode},q} - p_{\text{base},q})(y_q - p_{\text{base},q}) > 0]$, a post-hoc indicator of whether retrieval moves the base prediction toward the true outcome. On TimesNet at the 10-second horizon, beneficial proposals increase from approximately 54% in the lowest gate-value decile to 88% in the highest (Figure 5). The complete controller achieves a retrieval-benefit AUROC of 0.648 (95% CI $[0.553, 0.710]$) and the highest AUPRC, whereas every single-descriptor controller remains statistically indistinguishable from chance. The learned gate therefore gives more weight to retrieval when its proposed correction is more likely to help.

5 Limitations

This study has two limitations. First, NapCAP-128Ch comprises controlled afternoon naps rather than full-night polysomnography. Although this enabled high-density acquisition and expert CAP annotation, naps differ from overnight sleep in architecture, circadian context, homeostatic pressure, and stage distribution. Evaluation on the CAP Sleep Database uses a single EEG derivation, leaving full-night high-density validation for future work. Second, this offline study shows that upcoming Phase A onsets are predictable from pre-event EEG, not that acting on these forecasts is clinically beneficial. Future closed-loop studies should evaluate safety, false-alarm tolerance, intervention timing, and effects on sleep stability and clinical outcomes.

6 Conclusion

Early CAP Phase A forecasting must contend with event sparsity and cross-subject physiological heterogeneity. To our knowledge, this work is the first to extend retrieval augmentation to EEG event-onset forecasting. C-SHARE provides a model-agnostic retrieval-augmentation plug-in for frozen EEG forecasters with an explicit memory of source-subject events. Its onset-centered organization, future-conditioned multi-subspace matching, and event- and subject-aware gating allow retrieved evidence to be incorporated without target-subject adaptation. We further contribute NapCAP-128Ch, a 120-subject, expert-annotated high-density CAP EEG dataset. Across two datasets, eight backbones, and three horizons, C-SHARE improves over the corresponding frozen backbone in AUPRC in all 48 evaluated settings and in AUROC in 42 of 48, while providing more consistent gains than general-purpose retrieval baselines. Together, these findings support retrieval augmentation as a complementary strategy for EEG event forecasting within the two cohorts and three horizons evaluated here.

AI USE STATEMENT

We used generative AI tools only to aid and polish writing. They were not used for ideation, method development, implementation, experiments, analysis, or the generation of any result. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work involves human-subject EEG recordings. The NapCAP-128Ch study was approved by the ethics committee of the authors' institution. Written informed consent, including consent to share de-identified data for research, was obtained from all participants before participation. The study was conducted in accordance with the Declaration of Helsinki. The shared data replace participant names and dates with random identifiers.

REPRODUCIBILITY STATEMENT

The model and all training objectives are formalized with explicit equations in the main text, and all implementation details, including dataset descriptions, metrics, model, and experiment configurations, are given in the Appendix. Code and dataset are provided in the supplementary material.

REFERENCES

S Avila, JS Murguía, B Strizek, J Jimenez-Cruz, Anna M Bianchi, Alfonso Alba, E Arce-Santana, and Martín O Mendez. Lstm-based approach for a-phase detection from single-channel eeg in whole sleep recordings. *Biomedical Signal Processing and Control*, 109:108022, 2025.

Luciana Besedovsky, Hong-Viet V Ngo, Stoyan Dimitrov, Christoph Gassenmaier, Rainer Lehmann, and Jan Born. Auditory closed-loop stimulation of eeg slow oscillations strengthens sleep and signs of its immune-supportive function. *Nature communications*, 8(1):1984, 2017.

Tao Ding, Yuhao Zhang, Xingwei Zhao, Jian Shi, Jun Liu, Zhouping Tang, Bo Tao, and Han Ding. Rssda: A riemannian source selection and domain alignment framework for cross-subject multi-class motor imagery classification. *Pattern Recognition*, pp. 113934, 2026.

Dazhao Du, Tao Han, and Song Guo. Predicting the future by retrieving the past. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20896–20904, 2026.

Shahram Ebadollahi, Jimeng Sun, David Gotz, Jianying Hu, Daby Sow, and Chalapathy Neti. Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. In *AMIA annual symposium proceedings*, volume 2010, pp. 192, 2010.

Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, and Xiaoli Li. Tslanet: rethinking transformers for time series representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Raffaele Ferri, Oliviero Bruni, Silvia Miano, and Mario G Terzano. Topographic mapping of the spectral components of the cyclic alternating pattern (cap). *Sleep Medicine*, 6(1):29–36, 2005.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GUDnecJdJU.

Simon Hartmann and Mathias Baumert. Automatic a-phase detection of cyclic alternating patterns in sleep using dynamic temporal information. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(9):1695–1703, 2019.

Simon Hartmann and Mathias Baumert. Subject-level normalization to improve a-phase detection of cyclic alternating pattern in sleep eeg. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1–4. IEEE, 2023.

Imene Jemal, Lina Abou-Abbas, Khadidja Henni, Amar Mitiche, and Neila Mezghani. Domain adaptation for eeg-based, cross-subject epileptic seizure prediction. *Frontiers in neuroinformatics*, 18:1303380, 2024.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Zilin Liang, Zheng Zheng, Weihai Chen, Zhongcai Pei, Jianhua Wang, and Jianer Chen. Manifold embedded instance selection to suppress negative transfer in motor imagery-based brain–computer interface. *Biomedical Signal Processing and Control*, 88:105556, 2024.

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 35604–35629. Curran Associates, Inc., 2025. doi: 10.52202/085713-1196. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/331c41353b053683e17f7c88a797701d-Paper-Conference.pdf.

Shruti Murarka, Aditya Wadichar, Ankit Bhurane, Manish Sharma, and U Rajendra Acharya. Automated classification of cyclic alternating pattern sleep phases in healthy and sleep-disordered subjects using convolutional neural network. *Computers in Biology and Medicine*, 146:105594, 2022.

Hong-Viet V Ngo, Arjan Miedema, Isabel Faude, Thomas Martinetz, Matthias Mölle, and Jan Born. Driving sleep slow oscillations by auditory closed-loop stimulation—a self-limiting process. *Journal of Neuroscience*, 35(17):6630–6638, 2015.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 163170–163199. Curran Associates, Inc., 2025. doi: 10.52202/085713-5448. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/eed25c037bc08afcbefab6f7a6b700e0-Paper-Conference.pdf.

Liborio Parrino, Raffaele Ferri, Oliviero Bruni, and Mario G Terzano. Cyclic alternating pattern (cap): the marker of sleep instability. *Sleep medicine reviews*, 16(1):27–45, 2012.

Liborio Parrino, Heidur Gretarsdottir, Robert Thomas, Ivana Rosenzweig, Oliviero Bruni, Gulcin Benbir Senel, Fábio Mendonça, Antonio G. Ravelo García, Erna Sif Arnardóttir, and Raffaele Ferri. Atlas and updated rules for the scoring of cyclic alternating pattern (cap) in human sleep. a consensus report by a taskforce of the european sleep research society. *Journal of Sleep Research*, 35(4):e70283, 2026. doi: https://doi.org/10.1111/jsr.70283. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.70283.

Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Itaf Ben Slimen, Larbi Boubchir, and Hassene Seddik. Epileptic seizure prediction based on eeg spikes detection of ictal-preictal states. *Journal of biomedical research*, 34(3):162, 2020.

Michael L Stanchina, Muhanned Abu-Hijleh, Bilal K Chaudhry, Carol C Carlisle, and Richard P Millman. The influence of white noise on sleep in subjects exposed to icu noise. *Sleep medicine*, 6(5):423–428, 2005.

Robin Swezey, Aditya Grover, Bruno Charron, and Stefano Ermon. Pirank: Scalable learning to rank via differentiable sorting. *Advances in Neural Information Processing Systems*, 34:21644–21654, 2021.

Mario Giovanni Terzano, D Mancia, MR Salati, G Costani, A Decembrino, and Liborio Parrino. The cyclic alternating pattern as a physiologic component of normal nrem sleep. *Sleep*, 8(2):137–145, 1985.

Mario Giovanni Terzano, Liborio Parrino, Giuseppe Fioriti, Bruno Orofiamma, and Henri Depoortere. Modifications of sleep structure induced by increasing levels of acoustic perturbation in normal subjects. *Electroencephalography and clinical neurophysiology*, 76(1):29–38, 1990.

Mario Giovanni Terzano, Liborio Parrino, Adriano Sherieri, Ronald Chervin, Sudhansu Chokroverty, Christian Guilleminault, Max Hirshkowitz, Mark Mahowald, Harvey Moldofsky, Agostino Rosa, et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep. *Sleep medicine*, 2(6):537–553, 2001.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Xiaocui Yang, Han Zhao, Daling Wang, and Yifei Zhang. Is mamba effective for time series forecasting? *Neurocomputing*, 619:129178, 2025.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Suleyman Yaman, Hasan Güler, Abdulkadir Sengur, Abdul Hafeez-Baig, and U Rajendra Acharya. Murat-cap-net: A novel multi-input residual attention network for automated detection of a-phases and subtypes in cyclic alternating patterns. *Biomedical Signal Processing and Control*, 110:108221, 2025.

Jiaxin You, Yu Ma, and Yuanyuan Wang. Gtransu-cap: Automatic labeling for cyclic alternating patterns in sleep eeg using gated transformer-based u-net framework. *Computers in Biology and Medicine*, 147:105804, 2022.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting, 2026. URL https://arxiv.org/abs/2606.04135.

A Dataset Details and Statistics

This appendix documents the two cohorts used in our experiments. Table 2 summarizes the recordings and annotated events, and Table 3 gives the number of forecasting samples and their class balance for each horizon.

A.1 NapCAP-128Ch

Participants. We recruited 120 healthy adults aged 19–79 years (55 male, 65 female; mean age 35.1 $\pm$ 11.9 years). Participants were screened for sleep, neurological, and psychiatric disorders. Each participant contributed one recording, so subjects and recordings are the same unit in every split. The study was approved by the ethics committee of the authors' institution. Written informed consent, including consent to share de-identified data for research, was obtained from all participants before participation. The study was conducted in accordance with the Declaration of Helsinki.

Acquisition protocol. All naps began at 13:00 in our laboratory under controlled temperature and lighting. Participants followed a regulated sleep schedule on the preceding night. EEG was recorded for 30 minutes at 250 Hz with a 128-channel saline-electrode GSN-HydroCel-129 net.

Annotation. Two experienced scorers independently annotated Phase A events (onset, offset, and A1/A2/A3 subtype) following established CAP scoring criteria (Terzano et al., 2001; Parrino et al., 2026). Inter-rater agreement on Phase A, computed as Cohen's κ over non-overlapping 2-second windows labeled as Phase A or not, was $\kappa = 0.69$. The scorers jointly resolved disagreements to produce consensus labels.

A.2 PhysioNet CAP Sleep Database

The CAP Sleep Database (Terzano et al., 2001; Pollard et al., 2026) contains 108 overnight polysomnographic recordings with expert sleep-stage and Phase A annotations. It combines healthy controls with clinical groups whose conditions alter CAP expression. Prior automated CAP studies therefore commonly analyze a single diagnostic subgroup rather than pool all recordings (Hartmann & Baumert, 2019; You et al., 2022; Murarka et al., 2022). Following this practice, we use the 16 healthy-control recordings (`n1`–`n16`). The recordings come from different acquisition sites, so their EEG derivations and sampling rates differ; age and sex are taken from the spreadsheet distributed with the database (`gender-age.xlsx`). Native sampling rates are 100–512 Hz, and all signals are resampled to 128 Hz. We use the C4–A1 derivation for 12 subjects and C3–A2 for the 4 subjects without C4–A1 (`n6`–`n9`). Phase A onsets and subtypes are taken directly from the database annotations. Because these recordings are whole nights, each subject contributes about 20 times more forecasting samples than a NapCAP-128Ch nap, and positive windows are more common. As a public database, it serves as a second cohort with a different population, recording setting, and montage. All models are trained and evaluated within this cohort under the same leave-one-subject-out protocol; no model trained on NapCAP-128Ch is transferred to it.

Table 2: Recording-level statistics. Per-subject values are mean $\pm$ standard deviation (range).

	NapCAP-128Ch (ours)	CAP Sleep DB (healthy)
Subjects / recordings	120 / 120	16 / 16
Sex (M/F)	55/65	7/9
Age (years)	35.1 $\pm$ 11.9 (19–79)	32.2 $\pm$ 5.6 (23–42)
Recording type	afternoon nap	overnight PSG
EEG channels used	128	1 (C4–A1 / C3–A2)
Recording length (min)	30 (fixed)	533.6 $\pm$ 62.4 (490–735)
Total recorded time (min)	3,600	8,538
Phase A events (total)	1,602	6,569
A1 / A2 / A3	1,184 / 167 / 251	4,017 / 1,341 / 1,211
Phase A per subject	13.4 $\pm$ 5.7 (3–27)	410.6 $\pm$ 99.8 (259–546)
Phase A duration (s)	4.8 $\pm$ 2.4 (2–24)	7.7 $\pm$ 5.6 (2–65)

Differences between the cohorts. The two cohorts differ in the ways that matter for cross-subject retrieval. NapCAP-128Ch has many subjects but few events per subject: a subject contributes on average 13 Phase A events and 318 windows, so the evidence needed to forecast an unseen subject

Table 3: Forecasting samples per horizon. A sample is a 60-second NREM window that satisfies the exclusion rules in Appendix C (stride 2 s). It is positive if a Phase A onset occurs in $(t+h, t+h+10]$. Every window belongs to exactly one held-out subject, so the counts sum over LOSO test folds. Pooled prevalence is computed over all windows; subject-mean prevalence averages the per-subject prevalence, which is the AUPRC of a no-skill classifier under our subject-averaged metric.

	NapCAP-128Ch			CAP Sleep DB		
Horizon h	10 s	20 s	30 s	10 s	20 s	30 s
Windows (total)	38,180			98,384		
Windows per subject	318 ± 169			6,149 ± 815		
Positive windows	5,131	4,724	4,385	20,367	18,311	18,134
Negative windows	33,049	33,456	33,795	78,017	80,073	80,250
Neg. : pos. ratio	6.4	7.1	7.7	3.8	4.4	4.4
Pooled prevalence (%)	13.44	12.37	11.49	20.70	18.61	18.43
Subject-mean prevalence (%)	14.87	13.55	12.24	20.92	18.83	18.66

must come mainly from other subjects. The CAP Sleep Database has the reverse profile, with only 15 source subjects per fold but hundreds of events per subject. Its healthy controls are also a younger and narrower age group (32.2 ± 5.6 years, range 23–42, vs. 35.1 ± 11.9 years, range 19–79). The two cohorts also differ in spatial coverage (128 channels vs. one derivation) and class balance. The positive rate decreases with the horizon in both cohorts.

B Full Main-Result Tables

Tables 4 and 5 report the mean ± standard deviation underlying the main-text Figures 3 and 4, respectively. The parenthesized value under each method is the relative change (%) from the corresponding frozen backbone.

Table 4: Full results for early CAP Phase A forecasting on NapCAP-128Ch (120 subjects). AUROC and AUPRC are percentages (mean ± standard deviation across held-out subjects); the parenthesized value is the relative change (%) from the corresponding frozen backbone (blue: gain, orange: loss). RAFT (Han et al., 2025) and SARAF (Zhou et al., 2026) use source-fitted query-level fusion. "base" below each AUPRC column is the subject-mean positive-class prevalence (the AUPRC of a no-skill classifier) across the 120 held-out subjects.

Method	10s		20s		30s	
	AUROC	AUPRC (base=14.87)	AUROC	AUPRC (base=13.55)	AUROC	AUPRC (base=12.24)
EEGNet (2018)	61.28±7.35	25.28±12.47	59.79±8.03	21.36±10.84	60.62±11.69	24.72±12.01
+ RAFT	61.35±9.54 (+0.11%)	29.34±12.78 (+16.06%)	62.17±9.34 (+3.98%)	25.18±11.38 (+17.88%)	63.70±12.37 (+5.08%)	28.60±12.34 (+15.70%)
+ SARAF	62.34±8.62 (+1.73%)	28.74±12.77 (+13.69%)	61.51±8.36 (+2.88%)	24.67±11.32 (+15.50%)	62.38±10.03 (+2.90%)	27.02±12.74 (+9.30%)
+ C-SHARE	65.89±9.98 (+7.52%)	35.19±15.92 (+39.20%)	62.80±7.06 (+5.03%)	27.90±14.40 (+30.62%)	61.41±14.24 (+1.30%)	28.96±14.54 (+17.15%)
GTransU (2022)	65.93±11.08	32.03±12.65	60.28±12.75	25.14±13.91	60.59±12.00	26.66±12.04
+ RAFT	65.89±11.11 (-0.06%)	33.86±13.41 (+5.71%)	64.05±10.60 (+6.25%)	27.59±13.20 (+9.75%)	61.45±10.73 (+1.42%)	28.55±12.16 (+7.09%)
+ SARAF	65.72±10.97 (-0.32%)	32.25±12.08 (+0.69%)	62.31±11.01 (+3.37%)	27.21±14.00 (+8.23%)	63.00±9.41 (+3.98%)	28.05±10.51 (+5.21%)
+ C-SHARE	68.24±9.62 (+3.50%)	37.18±14.40 (+16.08%)	66.01±12.68 (+9.51%)	31.48±11.02 (+25.22%)	64.70±10.54 (+6.78%)	32.35±13.70 (+21.34%)
PatchTST (2023)	58.48±9.34	23.07±7.46	59.45±7.91	22.02±9.82	61.24±9.24	22.71±11.99
+ RAFT	59.36±8.51 (+1.50%)	25.29±8.40 (+9.62%)	59.28±8.30 (-0.29%)	24.40±9.85 (+10.81%)	62.96±7.65 (+2.81%)	26.19±11.65 (+15.32%)
+ SARAF	60.39±7.66 (+3.27%)	25.88±7.82 (+12.18%)	59.95±7.66 (+0.84%)	24.50±10.46 (+11.26%)	63.29±7.66 (+3.35%)	25.81±11.73 (+13.65%)
+ C-SHARE	61.80±9.33 (+5.68%)	31.29±11.13 (+35.63%)	63.65±9.20 (+7.06%)	28.51±14.65 (+29.47%)	63.56±9.87 (+3.79%)	28.18±12.30 (+24.09%)
TimesNet (2023)	63.90±13.08	29.10±14.99	64.16±12.49	24.85±14.85	64.98±14.16	27.05±16.16
+ RAFT	64.63±11.41 (+1.14%)	32.42±15.85 (+11.41%)	64.55±13.56 (+0.61%)	28.18±17.49 (+13.40%)	65.36±13.42 (+0.58%)	29.76±15.65 (+10.02%)
+ SARAF	65.65±11.89 (+2.74%)	32.24±16.22 (+10.79%)	63.28±12.23 (-1.37%)	26.71±14.92 (+7.48%)	65.69±13.50 (+1.09%)	29.19±15.13 (+7.91%)
+ C-SHARE	68.94±9.97 (+7.89%)	36.77±16.44 (+26.36%)	67.65±13.24 (+5.44%)	34.03±14.03 (+36.94%)	69.00±15.10 (+6.19%)	35.04±18.30 (+29.54%)
MEDformer (2024)	60.73±12.28	23.49±8.36	63.16±10.60	25.33±12.53	60.98±10.04	22.20±11.15
+ RAFT	62.56±9.61 (+3.01%)	26.47±8.14 (+12.69%)	63.83±10.18 (+1.06%)	26.97±13.26 (+6.47%)	63.59±8.13 (+4.28%)	25.52±9.47 (+14.95%)
+ SARAF	62.03±10.72 (+2.14%)	25.42±8.43 (+8.22%)	63.12±11.12 (-0.06%)	26.44±12.70 (+4.38%)	62.95±8.31 (+3.23%)	25.67±11.12 (+15.63%)
+ C-SHARE	63.87±9.51 (+5.17%)	33.35±15.11 (+41.98%)	65.37±11.96 (+3.50%)	31.45±16.10 (+24.16%)	63.11±10.23 (+3.49%)	27.61±13.25 (+24.37%)
TSLANet (2024)	62.95±14.50	28.62±17.27	64.09±12.12	25.08±11.30	67.97±12.21	28.34±13.80
+ RAFT	64.86±13.62 (+3.03%)	30.29±12.85 (+5.84%)	64.20±13.35 (+0.17%)	29.57±16.54 (+17.90%)	66.38±13.95 (-2.34%)	33.32±18.76 (+17.57%)
+ SARAF	63.18±13.60 (+0.37%)	28.81±12.34 (+0.66%)	64.30±13.37 (+0.33%)	26.87±13.50 (+7.14%)	66.92±13.64 (-1.54%)	30.29±15.94 (+6.88%)
+ C-SHARE	68.46±11.88 (+8.75%)	33.79±16.14 (+18.06%)	68.51±12.78 (+6.90%)	34.67±12.66 (+38.24%)	67.34±13.86 (-0.93%)	33.68±14.09 (+18.84%)
MuRAT (2025)	62.40±12.10	26.32±10.98	60.90±13.07	28.02±20.63	63.17±11.76	26.62±10.65
+ RAFT	63.00±12.02 (+0.96%)	27.89±11.43 (+5.97%)	63.38±9.82 (+4.07%)	31.41±18.74 (+12.10%)	62.53±11.24 (-1.01%)	26.88±9.24 (+0.98%)
+ SARAF	62.94±12.98 (+0.87%)	27.30±10.55 (+3.72%)	61.40±11.78 (+0.82%)	29.34±19.93 (+4.71%)	62.52±10.49 (-1.03%)	27.59±10.34 (+3.64%)
+ C-SHARE	66.75±10.01 (+6.97%)	32.09±13.13 (+21.92%)	66.21±11.98 (+8.72%)	32.77±19.27 (+16.95%)	62.82±13.17 (-0.55%)	28.08±12.64 (+5.48%)
S-Mamba (2025)	61.22±12.29	26.38±9.78	63.68±13.20	29.56±16.82	66.98±10.88	28.87±17.64
+ RAFT	64.47±10.01 (+5.31%)	28.11±9.30 (+6.56%)	65.59±12.30 (+3.00%)	31.72±17.70 (+7.31%)	69.01±10.09 (+3.03%)	30.46±17.67 (+5.51%)
+ SARAF	63.37±9.90 (+3.51%)	27.89±9.48 (+5.72%)	65.49±12.69 (+2.84%)	30.35±16.87 (+2.67%)	67.52±9.97 (+0.81%)	29.61±17.40 (+2.56%)
+ C-SHARE	63.43±10.82 (+3.61%)	31.52±12.23 (+19.48%)	65.21±13.33 (+2.40%)	31.72±18.62 (+7.31%)	69.84±10.88 (+4.27%)	33.79±19.77 (+17.04%)

Table 5: Full leave-one-subject-out validation results on the CAP Sleep Database, a second, public cohort (16 subjects). AUROC and AUPRC are percentages (mean ± standard deviation across held-out subjects); the parenthesized value is the relative change (%) from the corresponding frozen backbone (blue: gain, orange: loss). RAFT (Han et al., 2025) and SARAF (Zhou et al., 2026) use source-fitted query-level fusion. "base" below each AUPRC column is the subject-mean positive-class prevalence (the AUPRC of a no-skill classifier) across the 16 held-out subjects.

Method	10s		20s		30s	
	AUROC	AUPRC (base=20.92)	AUROC	AUPRC (base=18.83)	AUROC	AUPRC (base=18.66)
EEGNet (2018)	50.92±2.44	21.96±5.27	51.49±3.36	20.01±4.77	50.85±2.26	19.58±4.27
+ RAFT	51.24±2.91 (+0.63%)	22.03±5.56 (+0.32%)	52.33±3.78 (+1.63%)	20.92±5.14 (+4.55%)	50.93±2.54 (+0.16%)	20.40±4.62 (+4.19%)
+ SARAF	51.85±2.74 (+1.83%)	22.60±5.41 (+2.91%)	51.70±3.71 (+0.41%)	20.10±5.09 (+0.45%)	50.86±2.54 (+0.02%)	20.38±4.47 (+4.09%)
+ C-SHARE	54.97±3.77 (+7.95%)	24.33±4.83 (+10.79%)	54.12±3.58 (+5.11%)	21.25±4.66 (+6.20%)	53.61±3.23 (+5.43%)	20.99±4.10 (+7.20%)
GTransU (2022)	61.09±5.64	28.80±6.27	59.83±5.54	24.97±4.52	59.37±6.22	24.75±5.12
+ RAFT	61.70±6.12 (+1.00%)	28.43±6.34 (-1.28%)	60.29±5.88 (+0.77%)	25.72±4.42 (+3.00%)	59.93±6.30 (+0.94%)	25.46±5.04 (+2.87%)
+ SARAF	61.53±6.10 (+0.72%)	28.23±6.25 (-1.98%)	60.18±5.81 (+0.58%)	25.50±4.37 (+2.12%)	60.03±6.30 (+1.11%)	25.44±5.13 (+2.79%)
+ C-SHARE	62.83±5.86 (+2.85%)	30.04±6.51 (+4.31%)	61.66±5.77 (+3.06%)	27.20±4.82 (+8.93%)	61.35±6.50 (+3.34%)	26.97±5.26 (+8.97%)
PatchTST (2023)	50.96±1.70	21.47±4.72	51.41±1.88	19.64±4.24	51.32±2.00	19.38±4.01
+ RAFT	51.03±1.17 (+0.14%)	22.12±4.72 (+3.03%)	51.40±1.32 (-0.02%)	20.27±4.32 (+3.21%)	51.57±1.19 (+0.49%)	20.05±3.99 (+3.46%)
+ SARAF	51.57±1.30 (+1.20%)	22.32±4.99 (+3.96%)	51.87±1.36 (+0.89%)	20.73±4.27 (+5.55%)	51.89±1.38 (+1.11%)	20.27±4.05 (+4.59%)
+ C-SHARE	54.60±3.40 (+7.14%)	24.13±4.58 (+12.39%)	53.49±2.71 (+4.05%)	20.83±3.95 (+6.06%)	53.40±3.02 (+4.05%)	20.85±3.86 (+7.59%)
TimesNet (2023)	59.51±4.25	26.72±5.32	58.85±4.27	24.06±3.85	59.17±4.90	24.11±3.93
+ RAFT	59.61±4.58 (+0.17%)	26.99±5.75 (+1.01%)	58.90±4.43 (+0.08%)	24.36±3.90 (+1.25%)	59.16±4.93 (-0.02%)	24.30±4.10 (+0.79%)
+ SARAF	59.55±4.34 (+0.07%)	26.98±5.67 (+0.97%)	58.83±4.53 (-0.03%)	24.24±3.91 (+0.75%)	59.14±4.94 (-0.05%)	24.19±4.11 (+0.33%)
+ C-SHARE	62.71±4.44 (+5.38%)	28.14±5.66 (+5.31%)	61.97±4.42 (+5.30%)	25.27±4.06 (+5.03%)	62.23±5.15 (+5.17%)	25.44±4.02 (+5.52%)
MEDformer (2024)	61.14±7.14	28.55±7.70	59.84±6.41	24.63±5.54	60.09±6.55	24.71±5.64
+ RAFT	59.82±6.34 (-2.16%)	27.36±6.94 (-4.17%)	58.81±5.24 (-1.72%)	24.14±4.84 (-1.99%)	58.35±5.83 (-2.90%)	23.87±4.65 (-3.40%)
+ SARAF	59.69±6.32 (-2.37%)	27.18±6.94 (-4.80%)	60.40±5.44 (+0.94%)	24.73±4.58 (+0.41%)	58.59±6.04 (-2.50%)	24.87±4.92 (+0.65%)
+ C-SHARE	60.99±7.23 (-0.25%)	29.72±7.70 (+4.10%)	61.64±6.53 (+3.01%)	25.71±5.74 (+4.38%)	59.74±6.61 (-0.58%)	26.06±5.45 (+5.46%)
TSLANet (2024)	59.97±7.01	27.55±6.89	58.79±6.20	23.81±5.17	59.04±6.38	24.18±4.95
+ RAFT	58.95±6.57 (-1.70%)	27.24±6.32 (-1.13%)	59.44±6.09 (+1.11%)	24.65±4.80 (+3.53%)	58.38±6.28 (-1.12%)	23.99±4.87 (-0.79%)
+ SARAF	60.27±6.56 (+0.50%)	27.44±6.40 (-0.40%)	59.91±6.03 (+1.91%)	24.65±4.86 (+3.53%)	58.41±6.26 (-1.07%)	23.98±4.89 (-0.83%)
+ C-SHARE	61.83±7.19 (+3.10%)	28.60±7.14 (+3.81%)	60.81±6.42 (+3.44%)	25.06±5.51 (+5.25%)	59.03±6.67 (-0.02%)	25.57±5.06 (+5.75%)
MuRAT (2025)	59.55±5.54	27.18±6.09	58.40±5.45	23.84±4.26	59.18±5.17	24.12±4.36
+ RAFT	59.48±5.76 (-0.12%)	27.68±6.24 (+1.84%)	58.64±5.36 (+0.41%)	24.12±4.07 (+1.17%)	58.94±5.44 (-0.41%)	24.16±4.49 (+0.17%)
+ SARAF	59.48±5.78 (-0.12%)	27.26±6.24 (+0.29%)	58.51±5.52 (+0.19%)	24.03±4.30 (+0.80%)	58.75±5.57 (-0.73%)	23.95±4.63 (-0.70%)
+ C-SHARE	61.03±5.60 (+2.49%)	28.89±6.20 (+6.29%)	59.74±5.36 (+2.29%)	26.26±4.36 (+10.15%)	59.14±5.59 (-0.07%)	26.50±4.49 (+9.87%)
S-Mamba (2025)	59.68±6.59	26.93±6.36	58.07±5.94	23.10±4.42	58.47±6.15	23.39±4.67
+ RAFT	59.58±6.60 (-0.17%)	26.38±5.59 (-2.04%)	59.01±5.25 (+1.62%)	23.69±3.95 (+2.55%)	59.85±6.09 (+2.36%)	23.76±4.59 (+1.58%)
+ SARAF	59.78±6.47 (+0.17%)	26.48±5.65 (-1.67%)	59.42±5.46 (+2.32%)	23.97±3.97 (+3.77%)	59.72±5.58 (+2.14%)	23.68±4.46 (+1.24%)
+ C-SHARE	61.08±6.71 (+2.35%)	28.60±6.76 (+6.20%)	60.11±6.22 (+3.51%)	25.21±4.86 (+9.13%)	59.89±6.39 (+2.43%)	25.07±5.06 (+7.18%)

C EXPERIMENTAL DETAILS

Task construction and preprocessing. Each example contains a 60-second EEG history ending at anchor time t. Anchors are sampled every 2 seconds, beginning at the first position with a complete history. Sampling ends when the longest-horizon occurrence period, $(t+30, t+40]$, would exceed the recording, so all horizons share the same anchors. Each NapCAP-128Ch recording lasts 1,800 seconds, whereas CAP Sleep Database recordings span the full night. Anchors are retained only when the history lies within a continuous NREM interval. EEG is resampled to 128 Hz and filtered with a 0.5–30 Hz fourth-order band-pass filter and a 50 Hz notch filter in both datasets; NapCAP-128Ch uses all 128 channels. We exclude windows overlapping annotated artifacts, anchors inside a Phase A interval, and anchors with any Phase A onset in the shared exclusion interval $(t, t+10)$, which precedes all occurrence periods. For $h \in \{10, 20, 30\}$ seconds, the target y_h indicates whether any Phase A onset occurs in the associated occurrence period $(t+h, t+h+10]$. Targets are evaluated independently by horizon: they identify whether an onset falls within the specified period, not when the next onset occurs. Consequently, for $h \in \{20, 30\}$, an earlier onset in the intervening gap is permitted and multiple targets may be positive for the same anchor. To avoid near-boundary false negatives, we additionally exclude anchors whose next onset falls in a 10-second censor buffer $(t+40, t+50]$ immediately after the longest-horizon occurrence period. The public CAP Sleep Database uses the same forecasting construction.

CAP-timing control. Because Phase A recurs within CAP sequences and a 60-second history may contain earlier onsets, we test whether recurrence timing alone explains the forecasting targets. A CAP-timing prior scores each evaluated NapCAP-128Ch window using only the annotated Phase A onsets inside its history: the time since the last onset (2-second bins, with a separate bin when the history contains no onset) and the number of onsets. For each held-out subject, the positive rate in each bin is estimated on the source subjects and shrunk toward the source prevalence. This prior uses expert annotations of the history, which are unavailable at deployment, and is therefore a diagnostic control rather than a baseline. Across the 120 held-out subjects (Table 6), the prior is at chance level

Table 6: CAP-timing prior on NapCAP-128Ch, computed from annotated onsets in the 60-second history. AUROC and AUPRC are percentages (mean ± standard deviation across the 120 held-out subjects); "Base AUPRC" is the subject-mean positive-class prevalence (the AUPRC of a no-skill classifier), as in Table 4.

Horizon	AUROC	AUPRC	Base AUPRC
10s	50.55±11.49	18.38±9.58	14.87
20s	49.63±10.85	16.58±8.69	13.55
30s	51.06±11.51	15.99±10.54	12.24

Table 7: Base-forecaster learning rate and weight decay (η/λ) per backbone. All backbones use batch size 64 and $E/P/W = 100/20/5$ (maximum epochs, early-stopping patience, and warm-up epochs) on both cohorts.

Backbone	NapCAP-128Ch (η/λ)	CAP Sleep Database (η/λ)
EEGNet	$10^{-4}/5\times10^{-3}$	$10^{-4}/5\times10^{-3}$
GTransU	$3\times10^{-4}/10^{-3}$	$10^{-4}/5\times10^{-3}$
PatchTST	$10^{-4}/5\times10^{-3}$	$10^{-4}/5\times10^{-3}$
TimesNet	$10^{-4}/10^{-3}$	$10^{-4}/10^{-3}$
MEDformer	$10^{-4}/5\times10^{-3}$	$10^{-4}/5\times10^{-3}$
TSLANet	$10^{-4}/10^{-3}$	$10^{-4}/10^{-3}$
MuRAT	$3\times10^{-4}/10^{-3}$	$10^{-4}/5\times10^{-3}$
S-Mamba	$3\times10^{-4}/10^{-3}$	$3\times10^{-4}/10^{-3}$

in AUROC and falls below every frozen backbone in Table 4 in both AUROC and AUPRC at all three horizons. CAP recurrence timing therefore does not account for the forecasting signal.

Feature-extractor architecture and episode sampling. The feature extractor receives Fourier features from consecutive two-second EEG segments. Spectral projection applies frequency-wise normalization and a learned linear map to each channel token. Given projected tokens $e_{q,t,c}$, channel-attentive pooling uses $\omega_{q,t,c} = \mathrm{softmax}_c(g_{\mathrm{ch}}(e_{q,t,c}))$ and $h_{q,t} = \sum_c \omega_{q,t,c} e_{q,t,c}$. Depthwise-separable residual temporal convolutions with increasing dilation process $\{h_{q,t}\}$, after which learned attention pooling produces $p_q^{(h)}$. During training, we sample positive episodes uniformly and then sample trajectory states within each selected episode; background singletons fill the remaining batch positions. This prevents long positive episodes from dominating the retrieval objective.

Ranking distributions and subspace regularization. The label-side and embedding-side episode distributions used in Eq. 7 are

$$\begin{aligned} A^\star(q, E_j) &= \mathrm{Agg}_{\tau_A, M_{\mathrm{in}}}(\{\alpha^\star(q, x) : x \in E_j\}), \\ \pi^\star_{\mathrm{rank}}(E_j \mid q) &= \frac{A^\star(q, E_j)}{\sum_{E_l \in \mathcal{C}(q)} A^\star(q, E_l)}, \\ \widehat{\pi}_\theta(E_j \mid q) &= \frac{\exp(S^{\mathrm{train}}_\theta(q, E_j)/\tau_e)}{\sum_{E_l \in \mathcal{C}(q)} \exp(S^{\mathrm{train}}_\theta(q, E_l)/\tau_e)}. \end{aligned} \tag{12}$$

With $\bar{a}_{\mathrm{batch}} = B^{-1} \sum_q a_q^{(h)}$, the two label-free regularizers are $\mathcal{L}_{\mathrm{bal}} = \mathrm{KL}(\bar{a}_{\mathrm{batch}} \| \mathrm{Unif}(P))$ and $\mathcal{L}_{\mathrm{ent}} = B^{-1} \sum_q H(a_q^{(h)})$.

Optimization details. All base forecasters are trained with a sigmoid focal loss whose positive class is reweighted by the training negative-to-positive ratio, using AdamW and cosine-annealed learning rates, batch size 64, and $E/P/W = 100/20/5$ (maximum epochs, early-stopping patience, and minimum warm-up epochs before stopping, respectively) for every backbone on both cohorts. Table 7 reports the per-backbone learning rate and weight decay. The minimum learning rate is 10^{-5} with a 50-epoch cosine period, except for the NapCAP-128Ch GTransU and MuRAT runs, which use a 100-epoch period and a zero minimum. Base-forecaster runs use random seed 123.

The horizon-specific retrieval encoder is optimized with AdamW using learning rate 3×10^{-4}, weight decay 10^{-4}, and the same batch size 64 and $E/P/W = 100/20/5$ schedule as the base

forecasters (Table 7). Its random seed is 0. The linear controller is optimized in full batch with Adam (learning rate 0.05, weight decay 10^{-3}) for 300 epochs, and the final parameters are used.

Detector retargeting. GTransU and MuRAT were designed for per-second Phase A detection from raw single-channel EEG. We keep their encoders but replace the dense per-second segmentation head with a pooled binary head that predicts the horizon-specific target and share encoder weights across channels with mean channel pooling so that both cohorts' montages are supported.

Retrieval-baseline adaptation. RAFT and SARAF use the same frozen backbone outputs, source bank, prediction horizons, and selection environment as C-SHARE. We keep each method's retrieval mechanism, which matches source windows by the similarity of their observed EEG histories. Keys are computed from the backbone representation of each history, and values are binary per-horizon outcomes. Because RAFT's original fusion learns to combine retrieval with the local representation and assumes same-series retrieval, we adapt it with a per-query linear gate on retrieval reliability and retrieval–base disagreement, fitted with the same source-subject protocol as the C-SHARE controller. SARAF's original rule is a fixed equal-weight (0.5/0.5) blend. The main tables give SARAF the same query-level gate as RAFT, thereby comparing retrieval evidence without confounding the result with a query-independent fusion rule. Its fixed fusion configuration is reported separately in Appendix G. For RAFT and SARAF, retrieval budgets are selected per backbone using only source-subject validation data.

D ADDITIONAL CONTROLLER RELIABILITY ANALYSIS

The diagnostic in Sec. 4.4.1 uses the same source-fitted controller and full 120-subject NapCAP-128Ch cohort as the main experiments. The adopted controller's 10s AUPRC matches the main-text TimesNet result exactly (36.77%). Confidence intervals for the query-level direction-correctness and retrieval-benefit diagnostics are computed by a subject-clustered bootstrap with 2,000 resamples.

Repeating the diagnostic at the other horizons gives directionally consistent results. At $h = 30$s, retrieval-benefit AUROC is 0.679 (95% CI $[0.580, 0.748]$); at $h = 20$s, the effect is weaker and its CI ($[0.483, 0.698]$) includes chance. At $h = 10$s, the realized correction on harmful queries, $|g_q \delta_q|$, averages 0.143 ($[0.085, 0.229]$), compared with 0.170 ($[0.130, 0.219]$) under fixed $g = 0.5$. The learned gate thus tends to apply smaller corrections to harmful queries, although the difference at this horizon is modest and the confidence intervals overlap.

E MULTI-SUBSPACE ANALYSIS UNDER A FINITE EVENT MEMORY

We examine how the benefit of multi-subspace retrieval changes with the amount of positive event evidence available in memory. At the 20-second horizon, we vary the positive-episode budget $B \in \{16, 32, 64, 128, 256, 512, \text{full}\}$ and evaluate retrieval without subspaces and with multi-subspace keys against the same sampled memory, thereby isolating how effectively each representation uses the available finite event evidence.

Figure 6 shows that the multi-subspace advantage emerges from 128 positive episodes and widens as more of the finite memory becomes accessible. With the full bank of approximately 900 positive episodes, it increases AUPRC from 20.47% to 24.18%, an 18% relative gain. This result indicates that multi-subspace matching extracts more predictive value from a limited event memory than a single global retrieval geometry, although it cannot compensate for an extremely small event bank. Its benefit therefore comes from using observed event experience more efficiently, rather than from creating evidence absent from the source cohort.

F CAN RETRIEVAL RECOVER POSITIVE EVIDENCE MISSED BY THE BASE FORECASTER?

Section 4.4.1 examined when the controller trusts retrieval. We next examine the evidence contributed when the base forecaster underestimates a positive query. Figure 7 shows such a query at the 20s horizon for TimesNet with the multi-subspace retrieval encoder. This post-hoc case illustrates

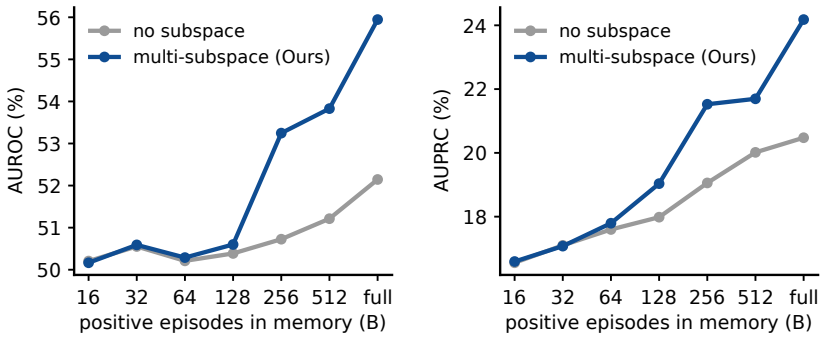


Figure 6: Retrieval-only AUROC/AUPRC (20s horizon) for no-subspace vs. multi-subspace (C-SHARE) retrieval keys, as the number of positive precursor episodes B accessible in memory is varied. The two encoders are evaluated against the identical sampled memory at every (repeat, B) pair, isolating the effect of multi-subspace matching from memory-sampling variance.

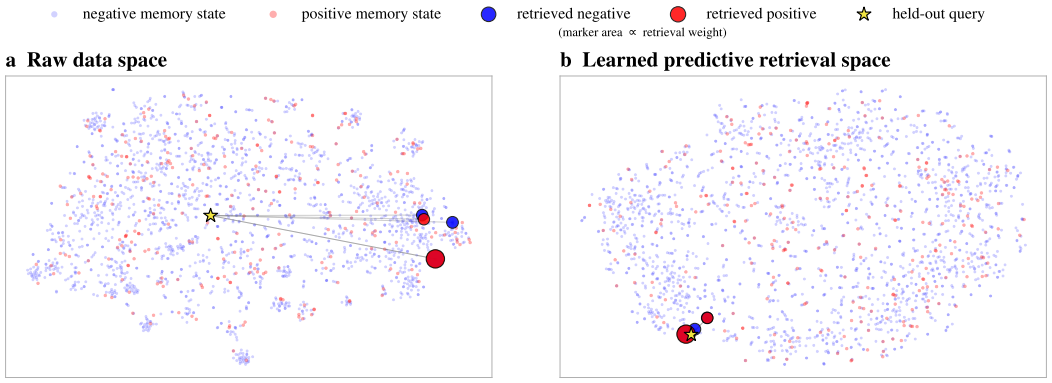


Figure 7: Retrieval supplies positive evidence missed by the base forecaster for a held-out positive query at the 20s horizon. The highest-weight episodes are distant in raw beta-topographic space (**a**) but close to the query in the learned predictive retrieval space (**b**); marker area denotes retrieval weight, hue the future Phase A label, and the star the query.

the correction mechanism; the held-out comparisons in the main text provide the population-level evidence.

The high-weight source episodes are not local neighbors of the query in the raw beta-topographic space (Fig. 7a). The learned predictive-key space places the same episodes in the query's immediate neighborhood (panel b). Panel (a) uses a prevalence-preserving background sample, and both panels show the same five highest-weight episodes from the top-10 retrieval set. Their PCA-initialized t-SNE projections are fitted separately, so distances are interpretable within, but not across, panels. Only two of the five displayed episodes are positive, but one is the highest-weight retrieval. Across the full top-10 set, the mean weight of a positive episode is $1.31\times$ that of a negative episode. This weighting raises the retrieved positive evidence to $r_{\text{episode}} = 0.247$, above the base prediction $p_{\text{base}} = 0.175$, and therefore proposes an upward correction to the final forecast. Relative to negative queries in the same held-out fold, the score moves from the 18.8th to the 88.7th percentile. The correction therefore comes from preferentially weighting sparse positive evidence, rather than from retrieving a positive majority.

G SARAF WITHOUT LEARNED QUERY-LEVEL FUSION

Table 4 reports SARAF fused through the same source-fitted query-level gate given to RAFT, rather than SARAF's fixed, query-independent equal-weight (0.5/0.5) rule. A fixed blend is better suited

Table 8: SARAF (Zhou et al., 2026) without learned query-level fusion, on NapCAP-128Ch (120 subjects): the retrieval side is identical to Table 4, but fusion is SARAF's original fixed, query-independent equal-weight (0.5/0.5) blend rather than the query-gated version used there. AUROC and AUPRC are percentages (mean ± standard deviation across held-out subjects); improvement is computed against the base-only row for that backbone. Blue and orange denote positive and negative changes, respectively.

Method	10s		20s		30s	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
EEGNet (2018)	61.28±7.35	25.28±12.47	59.79±8.03	21.36±10.84	60.62±11.69	24.72±12.01
+ SARAF	60.57±8.99 (-1.16%)	26.16±12.47 (+3.48%)	59.01±9.80 (-1.30%)	21.86±10.92 (+2.34%)	61.17±11.01 (+0.91%)	25.11±11.91 (+1.58%)
GTransU (2022)	65.93±11.08	32.03±12.65	60.28±12.75	25.14±13.91	60.59±12.00	26.66±12.04
+ SARAF	64.47±11.02 (-2.21%)	30.19±10.62 (-5.74%)	60.35±11.95 (+0.12%)	24.22±11.78 (-3.66%)	60.56±11.29 (-0.05%)	24.92±9.77 (-6.53%)
PatchTST (2023)	58.48±9.34	23.07±7.46	59.45±7.91	22.02±9.82	61.24±9.24	22.71±11.99
+ SARAF	57.18±9.51 (-2.22%)	22.01±6.78 (-4.59%)	58.10±8.42 (-2.27%)	21.65±10.69 (-1.68%)	59.75±10.23 (-2.43%)	21.55±11.25 (-5.11%)
TimesNet (2023)	63.90±13.08	29.10±14.99	64.16±12.49	24.85±14.85	64.98±14.16	27.05±16.16
+ SARAF	61.75±12.99 (-3.36%)	27.31±14.47 (-6.15%)	60.91±13.57 (-5.07%)	22.99±13.01 (-7.48%)	62.73±17.00 (-3.46%)	26.53±15.01 (-1.92%)
MEDformer (2024)	60.73±12.28	23.49±8.36	63.16±10.60	25.33±12.53	60.98±10.04	22.20±11.15
+ SARAF	58.68±12.20 (-3.38%)	23.71±8.12 (+0.94%)	58.52±11.31 (-7.35%)	22.60±11.93 (-10.78%)	57.70±9.97 (-5.38%)	21.14±10.78 (-4.77%)
TSLANet (2024)	62.95±14.50	28.62±17.27	64.09±12.12	25.08±11.30	67.97±12.21	28.34±13.80
+ SARAF	56.63±12.01 (-10.04%)	23.00±8.73 (-19.64%)	55.47±11.76 (-13.45%)	20.86±11.43 (-16.83%)	60.17±14.39 (-11.48%)	24.50±14.82 (-13.55%)
MuRAT (2025)	62.40±12.10	26.32±10.98	60.90±13.07	28.02±20.63	63.17±11.76	26.62±10.65
+ SARAF	61.44±12.76 (-1.54%)	25.75±10.25 (-2.17%)	60.31±12.03 (-0.97%)	27.36±19.50 (-2.36%)	62.59±11.59 (-0.92%)	24.72±11.16 (-7.14%)
S-Mamba (2025)	61.22±12.29	26.38±9.78	63.68±13.20	29.56±16.82	66.98±10.88	28.87±17.64
+ SARAF	61.51±11.45 (+0.47%)	25.48±9.56 (-3.41%)	63.72±12.89 (+0.06%)	28.62±16.17 (-3.18%)	65.14±9.38 (-2.75%)	26.06±15.81 (-9.73%)

Table 9: Component ablation on frozen GTransU across three prediction horizons (120 held-out subjects; mean ± standard deviation across subjects, percentages; bold: best in column). Columns follow Table 1; the last row is the complete C-SHARE. The window-level retrieval variants of Table 1 are not repeated here. The retrieval-only row does not depend on the backbone and is therefore identical to that of Table 1.

Retrieval unit	Reliability controller	Multi-subspace encoder	10s		20s		30s	
			AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
–	g=0 (GTransU only)	–	65.93±11.08	32.03±12.65	60.28±12.75	25.14±13.91	60.59±12.00	26.66±12.04
Episode	g=1 (retrieval only)	✓	56.85±11.91	28.83±12.99	55.94±11.71	24.18±13.47	54.42±10.01	21.52±10.79
Episode	g=0.5 (fixed)	✓	64.20±12.02	31.38±11.94	59.12±11.76	25.21±10.71	58.26±13.92	24.63±10.76
Episode	global scalar g	✓	68.16±9.43	36.85±13.79	**66.01±12.28**	29.95±11.10	61.96±13.38	30.22±12.17
Episode	query-level g_q	✗	66.91±10.44	37.02±15.48	65.97±10.14	29.63±15.34	62.05±8.24	30.15±12.82
Episode	query-level g_q	✓	**68.24±9.62**	**37.18±14.40**	**66.01±12.68**	**31.48±11.02**	**64.70±10.54**	**32.35±13.70**

to SARAF's original setting, which retrieves within the same series and domain from a dense training pool. It is less suitable for forced cross-subject retrieval, where evidence quality varies across queries. For completeness, Table 8 reports SARAF with its original fusion rule, a fixed, query-independent equal-weight (0.5/0.5) blend of the retrieval and backbone predictions with no learned parameters. This configuration improves AUROC in only 4 and AUPRC in only 4 of the 24 backbone–horizon combinations, and both in just 1 (mean change -2.10 and -1.44 percentage points, respectively), materially worse than the query-gated version used in the main text.

H COMPONENT ABLATION ON ADDITIONAL BACKBONES

Table 1 ablates C-SHARE on TimesNet. Table 9 repeats the retrieval-and-fusion ablation on frozen GTransU. Among the non-TimesNet backbones, GTransU has the highest absolute AUPRC at the 10s horizon in Table 4. Both ablations use the full 120-subject NapCAP-128Ch cohort, LOSO folds, source memory, and controller implementation. Table 10 extends the controller ablation to all seven non-TimesNet backbones.

On GTransU, the query-level gate g_q has the highest AUPRC at all three horizons and the highest AUROC at 10s and 30s. At 20s, its AUROC ties the global scalar g to two decimal places (66.01). Retrieval alone (g=1) remains below the frozen backbone in every column. The fixed blend (g=0.5) also fails to exceed the backbone in five of six columns. Retrieved evidence therefore helps only when weighted by a learned gate. Replacing the multi-subspace encoder with a single-subspace encoder lowers both metrics at all horizons. The multi-subspace gains are $+1.33$, $+0.04$, and $+2.65$ AUROC points and $+0.16$, $+1.85$, and $+2.20$ AUPRC points at 10s, 20s, and 30s, respectively.

Table 10: Controller ablation across backbones (120 held-out NapCAP-128Ch subjects; mean ± standard deviation across subjects, percentages; bold: best of the four variants per backbone and column). All retrieval variants use episode retrieval with the multi-subspace encoder; the controller column follows Table 1. The query-level g_q rows are the +C-SHARE rows of Table 4.

Backbone	Reliability controller	10s		20s		30s	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
EEGNet	g=0 (base only)	61.28±7.35	25.28±12.47	59.79±8.03	21.36±10.84	60.62±11.69	24.72±12.01
	g=0.5 (fixed)	61.48±10.78	30.36±14.66	57.84±11.64	23.69±13.86	58.41±12.20	23.40±11.31
	global scalar g	64.24±9.65	34.67±15.49	62.69±6.87	26.95±14.84	**62.25±13.52**	**29.16±12.09**
	query-level g_q (C-SHARE)	**65.89±9.98**	**35.19±15.92**	**62.80±7.06**	**27.90±14.40**	61.41±14.24	28.96±14.54
GTransU	g=0 (base only)	65.93±11.08	32.03±12.65	60.28±12.75	25.14±13.91	60.59±12.00	26.66±12.04
	g=0.5 (fixed)	64.20±12.02	31.38±11.94	59.12±11.76	25.21±10.71	58.26±13.92	24.63±10.76
	global scalar g	68.16±9.43	36.85±13.79	**66.01±12.28**	29.95±11.10	61.96±13.38	30.22±12.17
	query-level g_q (C-SHARE)	**68.24±9.62**	**37.18±14.40**	**66.01±12.68**	**31.48±11.02**	**64.70±10.54**	**32.35±13.70**
PatchTST	g=0 (base only)	58.48±9.34	23.07±7.46	59.45±7.91	22.02±9.82	61.24±9.24	22.71±11.99
	g=0.5 (fixed)	59.66±9.40	28.04±11.97	60.13±6.04	23.64±13.91	60.40±8.79	23.05±11.30
	global scalar g	61.23±8.79	**31.32±12.56**	62.47±6.74	26.56±13.33	62.51±8.52	26.20±11.63
	query-level g_q (C-SHARE)	**61.80±9.33**	31.29±11.13	**63.65±9.20**	**28.51±14.65**	**63.56±9.87**	**28.18±12.30**
MEDformer	g=0 (base only)	60.73±12.28	23.49±8.36	63.16±10.60	25.33±12.53	60.98±10.04	22.20±11.15
	g=0.5 (fixed)	59.17±11.95	30.43±15.25	59.71±9.09	24.64±14.85	59.30±9.78	23.29±11.34
	global scalar g	63.75±9.89	33.02±14.71	**66.19±10.20**	29.45±15.22	**63.59±9.04**	27.50±12.79
	query-level g_q (C-SHARE)	**63.87±9.51**	**33.35±15.11**	65.37±11.96	**31.45±16.10**	63.11±10.23	**27.61±13.25**
TSLANet	g=0 (base only)	62.95±14.50	28.62±17.27	64.09±12.12	25.08±11.30	**67.97±12.21**	28.34±13.80
	g=0.5 (fixed)	62.58±13.18	28.60±11.97	60.74±13.62	26.84±11.07	61.21±11.24	23.81±10.51
	global scalar g	67.30±13.14	33.59±16.56	67.77±13.80	33.35±11.63	67.88±13.55	32.96±13.88
	query-level g_q (C-SHARE)	**68.46±11.88**	**33.79±16.14**	**68.51±12.78**	**34.67±12.66**	67.34±13.86	**33.68±14.09**
MuRAT	g=0 (base only)	62.40±12.10	26.32±10.98	60.90±13.07	28.02±20.63	63.17±11.76	26.62±10.65
	g=0.5 (fixed)	63.73±11.67	29.78±12.96	62.38±12.62	27.66±14.06	62.12±13.63	23.65±10.88
	global scalar g	65.41±10.44	**32.17±13.94**	65.96±11.64	**33.00±19.93**	**63.89±12.85**	**28.15±11.47**
	query-level g_q (C-SHARE)	**66.75±10.01**	32.09±13.13	**66.21±11.98**	32.77±19.27	62.82±13.17	28.08±12.64
S-Mamba	g=0 (base only)	61.22±12.29	26.38±9.78	63.68±13.20	29.56±16.82	66.98±10.88	28.87±17.64
	g=0.5 (fixed)	62.80±11.10	**32.05±13.91**	62.30±12.02	24.82±16.05	65.17±9.52	26.82±14.19
	global scalar g	63.00±11.39	31.19±12.26	64.29±13.32	31.09±17.96	68.25±9.40	31.67±18.37
	query-level g_q (C-SHARE)	**63.43±10.82**	31.52±12.23	**65.21±13.33**	**31.72±18.62**	**69.84±10.88**	**33.79±19.77**

Across the seven backbones (42 backbone–horizon–metric cells), the query-level gate g_q exceeds the fixed blend (g=0.5) in 41 cells and the frozen base (g=0) in 40. It exceeds the global scalar g in 31 cells and ties it in one. The query-level controller is therefore not uniformly superior to a learned global scalar gate. The global scalar gate is better in four of six MuRAT cells, two of six MEDformer and EEGNet cells, and one of six TSLANet and PatchTST cells. The query-level gate is at least as good in all six GTransU and S-Mamba cells. On GTransU, its AUROC margins over the global scalar gate are $+0.08$, $+0.00$, and $+2.74$ points at 10s, 20s, and 30s. The corresponding AUPRC margins are $+0.33$, $+1.53$, and $+2.13$ points.

I STATISTICAL SIGNIFICANCE OF MAIN RESULTS

Table 4 reports per-setting AUROC/AUPRC as mean ± standard deviation across the 120 held-out subjects of the NapCAP-128Ch cohort, with improvement direction summarized by a win/loss count in the main text. For each backbone–horizon setting, Table 11 adds the mean paired improvement, subject win count, and paired one-sided Wilcoxon signed-rank test. The alternative hypothesis is C-SHARE > base, with $n = 120$ subjects.

All tests use the per-subject predictions underlying Table 4. We first verified that these predictions reproduce the reported means and standard deviations. At the uncorrected $p < 0.05$ threshold, AUPRC improves in all 24 settings and AUROC in 19. Of the five non-significant AUROC cells, two have no mean improvement: TSLANet and MuRAT at the 30-second horizon, where AUROC changes by -0.63 and -0.35 points (the two AUROC losses in Table 4). The other three (EEGNet at 30 seconds and MEDformer at 20 and 30 seconds) have positive mean gains ($+0.79$, $+2.20$, and $+2.13$ points) but improve only 60, 56, and 57 of the 120 subjects, respectively, so these mean gains are driven by a subset of subjects rather than a consistent per-subject shift.

Correcting for multiple comparisons with the Holm–Bonferroni procedure over all 48 tests (24 settings × 2 metrics, family-wise $\alpha = 0.05$) leaves all 24 AUPRC improvements significant; for AUROC, MEDformer at the 10-second horizon ($p = 0.009$) no longer passes, leaving 18 of 24 sig-

nificant. Pooling all 24 settings with a subject-clustered bootstrap of the 120-subject cohort (2,000 resamples) gives a mean improvement of $+6.19$ AUPRC points (95% CI $[+5.46, +6.78]$) and $+3.15$ AUROC points (95% CI $[+2.28, +3.60]$). Both intervals exclude 0, consistent with the pooled result reported in Sec. 4.2.1.

Table 11: Statistical significance of C-SHARE's improvement over the frozen backbone in every backbone–horizon setting of Table 4 (NapCAP-128Ch, full 120-subject LOSO evaluation). "Win" counts subjects (out of 120) where C-SHARE improves over base; p is from a paired one-sided Wilcoxon signed-rank test ($n = 120$), uncorrected; † marks the one cell with uncorrected $p < 0.05$ that is not significant after Holm–Bonferroni correction over all 48 tests.

Backbone	Horizon	AUROC			AUPRC		
		Δ (pt)	Win	p	Δ (pt)	Win	p
EEGNet	10s	+4.61	79/120	<0.001	+9.91	105/120	<0.001
	20s	+3.01	77/120	<0.001	+6.53	89/120	<0.001
	30s	+0.79	60/120	0.351	+4.24	86/120	<0.001
GTransU	10s	+2.31	89/120	<0.001	+5.15	82/120	<0.001
	20s	+5.73	93/120	<0.001	+6.34	79/120	<0.001
	30s	+4.11	85/120	<0.001	+5.69	94/120	<0.001
PatchTST	10s	+3.32	80/120	<0.001	+8.22	101/120	<0.001
	20s	+4.20	92/120	<0.001	+6.49	102/120	<0.001
	30s	+2.32	70/120	<0.001	+5.47	116/120	<0.001
TimesNet	10s	+5.04	88/120	<0.001	+7.67	109/120	<0.001
	20s	+3.49	77/120	0.002	+9.18	110/120	<0.001
	30s	+4.02	83/120	<0.001	+7.99	96/120	<0.001
MEDformer	10s	+3.14	70/120	0.009†	+9.86	113/120	<0.001
	20s	+2.20	56/120	0.215	+6.13	100/120	<0.001
	30s	+2.13	57/120	0.693	+5.41	109/120	<0.001
TSLANet	10s	+5.51	91/120	<0.001	+5.16	102/120	<0.001
	20s	+4.42	94/120	<0.001	+9.59	108/120	<0.001
	30s	-0.63	52/120	0.753	+5.34	87/120	<0.001
MuRAT	10s	+4.35	71/120	<0.001	+5.77	108/120	<0.001
	20s	+5.32	73/120	<0.001	+4.75	82/120	<0.001
	30s	-0.35	59/120	0.663	+1.46	71/120	0.001
S-Mamba	10s	+2.21	92/120	<0.001	+5.14	100/120	<0.001
	20s	+1.52	74/120	<0.001	+2.16	112/120	<0.001
	30s	+2.86	90/120	<0.001	+4.92	91/120	<0.001