# M$^2$-EEG: A MACRO–MICRO PRETRAINING PARADIGM FOR EXPANDING THE TEMPORAL SCOPE OF EEG FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Electroencephalography (EEG) foundation models have advanced transferable representation learning, yet pretraining and evaluation remain centered largely on short local segments. For recording-scale decisions, aggregating independently encoded windows leaves local representations unaware of context elsewhere in the recording. We introduce M$^2$-EEG, a Macro–Micro pretraining paradigm built on the principle that waveform resolution and temporal reach should be modeled separately. A high-resolution Micro pathway preserves local waveform evidence, while a lower-bandwidth Macro pathway models recording-wide context. Aligned fusion produces context-aware temporal and recording-level representations within a single whole-recording backbone pass. Using a 5,300-hour quality-controlled EEG corpus, M$^2$-EEG is progressively pretrained from local representation learning to long-context modeling and joint Macro–Micro adaptation using a unified family of multi-scale spectral prediction objectives. Across six long-context EEG datasets, frozen M$^2$-EEG representations outperform those of the compared EEG foundation models under aligned downstream readouts, while retaining strong short-window transfer with the same complete backbone. Controlled context-span experiments further show that expanding contextual support improves predictions even when local EEG evidence is held fixed. Together, these results support Macro–Micro pretraining as a scalable approach to extending EEG foundation learning from local segments toward recording-scale interpretation.

## 1 INTRODUCTION

Electroencephalography (EEG) foundation models increasingly learn transferable representations across datasets and tasks through large-scale self-supervised pretraining (Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024). Advances in montage flexibility and spatial modeling further extend their use across heterogeneous recordings (Wang et al., 2025; El Ouahidi et al., 2025; Döner et al., 2025). Their temporal scope, however, remains less developed: much of this progress is established with short local segments as the unit of learning. For example, LaBraM uses a pretraining budget corresponding to 4 or 8 seconds for 64- or 32-channel inputs, respectively, whereas CBraMod pretrains on 30-second segments (Jiang et al., 2024; Wang et al., 2025). This concerns the context used to learn representations, rather than an inability to accept variable-length inputs: BIOT explicitly accommodates variable sample lengths, and REVE supports arbitrary signal lengths through its positional encoding (Yang et al., 2023; El Ouahidi et al., 2025). The question is how to extend transferable EEG representations from local segments to the temporal scope of a recording.

Recording-scale interpretation can require information extending well beyond an isolated window. Sleep staging benefits from relationships among successive epochs (Phan et al., 2019), while clinical session modeling draws on patterns distributed across a recording (Cao et al., 2026). Yet many foundation-model benchmarks emphasize segment-level decoding, including motor imagery, emotion recognition, and event classification (Wang et al., 2025; El Ouahidi et al., 2025). When a short-window backbone is adapted by independently encoding windows and pooling their features or predictions, context enters only after local encoding. A learned sequence model can contextualize those embeddings, as in existing sleep-staging evaluation pipelines (Wang et al., 2025), but such task-specific adaptation does not by itself establish recording-scale foundation pretraining. Our goal



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.

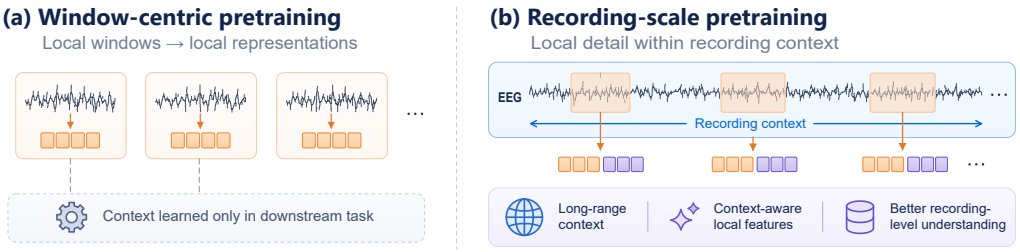


Figure 1: **Extending the context of EEG pretraining.** **(a)** Window-centric pretraining can be followed by downstream context aggregation. **(b)** M$^2$-EEG forms recording-wide context within the pretrained backbone before task readout, while preserving local waveform evidence. Waveforms are schematic.

is to make recording context part of the reusable pretrained representation, rather than introduce it only during downstream task adaptation (Figure 1). Our primary evaluation therefore tests recording context already available from a frozen pretrained backbone, rather than context learned by an additional task-specific recording-level sequence model.

Extending this contextual scope requires reconciling local waveform fidelity with long temporal reach. Transient morphology and local spectral structure call for detailed representations, whereas minutes-to-hours context makes dense global attention over high-resolution tokens costly (Vaswani et al., 2017). Compressing every temporal scale into a single coarse representation, in turn, risks discarding local detail. Recent work already spans cross-scale EEG representations, long-sequence architectures, and session-scale clinical modeling (Zhou et al., 2025; Hong et al., 2025; Cao et al., 2026; Pradeepkumar et al., 2026). Appendix B provides a fuller comparison of these routes and their evidence boundaries. This progress motivates a specific design question: *how can a signal-only foundation model preserve detailed local evidence while learning reusable context across the temporal scope of an entire recording?* Our premise is that **waveform resolution and temporal reach should be modeled separately, then reconnected through learned cross-scale interactions**.

We introduce M$^2$-EEG, a Macro–Micro pretraining paradigm built around this separation. The high-resolution Micro pathway captures local waveform structure within bounded temporal regions, while the lower-bandwidth Macro pathway propagates coarse context across the recording. Aligned fusion produces context-aware temporal and recording-level representations in a single whole-recording backbone pass. Multiple supervised intervals share this recording context, while a direct Micro pathway preserves fine-grained local evidence. Separating the two scales avoids dense global attention over high-resolution waveform tokens. Short-window and whole-recording tasks use the same complete architecture, retaining Macro, Micro, and aligned fusion at either input scale rather than switching to a separate long-context model or a Micro-only short-window encoder.

We train this hierarchy on approximately 5,300 hours of quality-controlled EEG using temporally aligned multi-scale spectral prediction objectives. Learning proceeds from local representation learning to long-context modeling and then to joint Macro–Micro adaptation. Our experiments examine three questions: *(i)* whether recording-scale representations improve transfer when decisions depend on long temporal context, *(ii)* whether the resulting gains can be attributed to broader contextual support rather than local evidence alone, and *(iii)* whether extending temporal scope preserves conventional short-window transfer. Across six long-context datasets, frozen M$^2$-EEG representations outperform the compared foundation-model representations under aligned task readouts. The same complete backbone retains strong short-window transfer under task-specific downstream adaptation. Controlled context-span experiments further show that broader recording context improves predictions even when local EEG evidence is held fixed.

Our contributions are threefold:

- We formulate recording-scale EEG foundation learning as the problem of extending temporal reach without sacrificing local waveform resolution, motivating a separation between these two representational requirements.

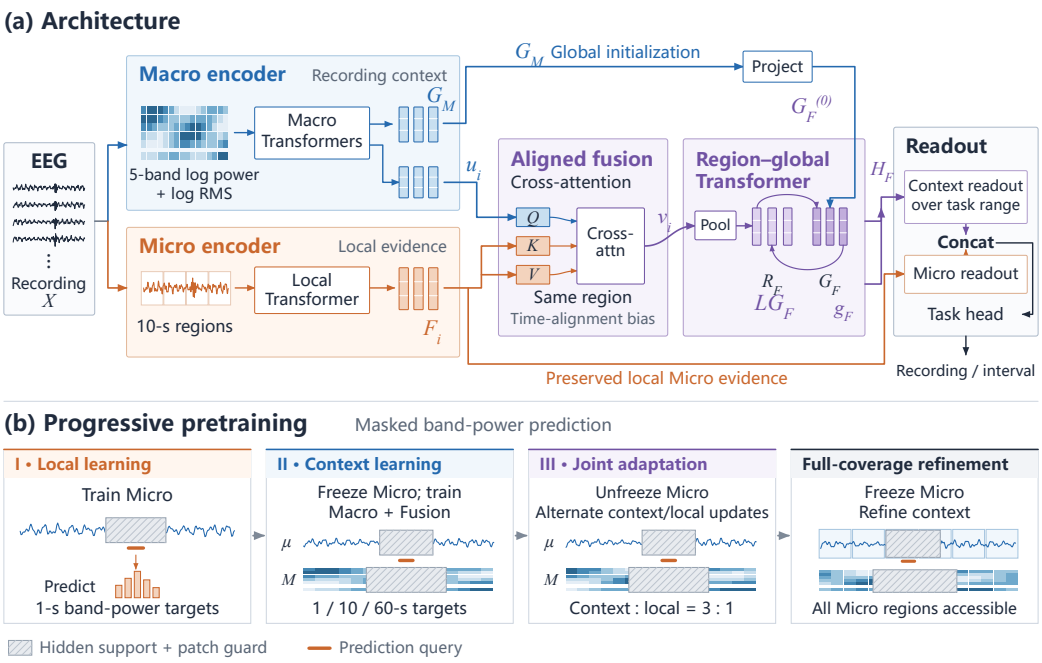


Figure 2: **Architecture and pretraining of M$^2$-EEG. (a)** Aligned Fusion and recording-scale contextualization connect parallel Macro and Micro pathways; readouts retain a direct Micro path. **(b)** Local, contextual, and joint learning precede full-coverage contextual refinement. Inputs are schematic. Raw-signal supports are masked in both pathways; full coverage refers to Micro access.

- We introduce M$^2$-EEG, a progressively pretrained Macro–Micro backbone shared by short-window and whole-recording tasks. A single whole-recording forward pass forms context-aware temporal and recording-level representations alongside preserved high-resolution local evidence.
- We evaluate M$^2$-EEG across six long-context EEG datasets and complementary short-window tasks, and show through controlled context-span experiments that broader recording context improves predictions even when local EEG evidence is held fixed.

## 2 Method

### 2.1 Overview

M$^2$-EEG separates waveform resolution from temporal reach through two synchronized pathways. After resampling to $f_s = 200$ Hz, an EEG input $X$ with $C$ available channels and duration $S$ seconds is divided into $N = \lceil S/L \rceil$ regions of length $L = 10$ seconds. Within each region, the Micro pathway preserves fine-grained waveform structure, while the Macro pathway propagates a lower-bandwidth temporal context across regions. Temporally aligned Fusion then lets contextual Macro states retrieve fine-grained Micro evidence before task readout, producing local, contextual temporal, and recording-level representations. The same backbone supports both short-window and whole-recording inputs; only the temporal range exposed to the downstream readout changes (Figure 2).

### 2.2 Macro–Micro Architecture

**High-resolution local encoding.** Micro applies high-resolution attention locally within each bounded 10-s region. Let $\widetilde{X}_{i,c,p}$ denote patch $p$ from channel $c$ in region $i$, and let $\eta_c$ encode its electrode identity and metadata. A shared projection gives $E_{i,c,p} = \phi_\mu(\widetilde{X}_{i,c,p}, \eta_c)$, followed by learned pooling over valid electrodes, $A_{i,p} = \mathrm{SetPool}_c(E_{i,c,p})$. The region representation and its

local summary are

$$F_i = \mathrm{Enc}_\mu(A_i + P_\mu), \qquad m_i = \mathrm{Pool}_{\mathrm{valid}}(F_i), \tag{1}$$

where $\mathrm{Enc}_\mu$ is a region-bounded temporal Transformer. Inputs are padded only to complete the final region; padded Micro patches are masked as attention keys, and padded seconds are excluded from valid aggregation and downstream readout. The local summary $m_i$ excludes physical padding but retains SSL mask tokens at valid time positions. Fusion accesses the full high-resolution states $F_i$, while $m_i$ serves as a compact local summary for pretraining and range-level Micro readout.

**Coarse recording context.** The Macro pathway constructs a low-bandwidth recording view directly from EEG rather than pooling Micro outputs. For second $t$ and available channel $c$, $D_{i,t,c}$ contains five-band log power over a four-second support and the log RMS of the channel waveform after subtracting its median within the same support. Band validity and electrode identity inform pooling across channels. Spectral support and masking conventions are detailed in Appendix C.1. We form

$$A^M_{i,t} = \mathrm{SetPool}_c\big(\phi_M(D_{i,t,c}, \eta_c)\big) + P^M_{i,t},$$

and stack the second-level states as $A^M_i = (A^M_{i,t})_t$. Macro processing then follows

$$u_i = \mathrm{Enc}_{M,\mathrm{local}}(A^M_i), \qquad R^{(0)}_{M,i} = \mathrm{Pool}_{\mathrm{valid}}(u_i), \qquad (R_M, G_M) = \mathrm{LG}_M(R^{(0)}_M, G^{(0)}_M). \tag{2}$$

Here, $u_i$ are region-local second-level states, $R^{(0)}_M$ initializes one state per valid region, and $G^{(0)}_M$ is a learned global-token bank. $\mathrm{LG}_M$ is a sparse region–global Transformer: region queries attend to neighboring regions and global tokens, after which global queries attend to all updated regions and global tokens. The global bank propagates recording-wide information, yielding contextualized region states $R_M$ and global states $G_M$.

**Temporally aligned cross-scale fusion.** Cross-scale fusion projects $u_i$ into queries $Q_i$ that attend to Micro states $F_i$ within the same region. For second center $\tau_t$ and Micro-patch center $\tau_p$, we introduce the temporal alignment bias

$$B_{t,p} = -\frac{|\tau_t - \tau_p|}{\lambda},$$

where $\lambda > 0$ controls the temporal decay bandwidth. The fused second-level states are

$$Q_i = u_i W_Q, \qquad \Delta_i = \mathrm{CrossAttn}(Q_i, F_i, F_i; B_i), \qquad v_i = (Q_i + \Delta_i) \odot M^{\mathrm{sec}}_i, \tag{3}$$

where invalid Micro patches are masked as keys and $M^{\mathrm{sec}}_i$ removes invalid physical seconds.

Pooling $v_i$ initializes fused region states $R^{(0)}_{F,i} = \mathrm{Pool}_{\mathrm{valid}}(v_i)$, while projecting $G_M$ initializes the fused global bank $G^{(0)}_F = \mathrm{Project}(G_M)$. A two-layer $\mathrm{LG}_F$ block further propagates context before returning region information to each valid second:

$$(R_F, G_F) = \mathrm{LG}_F(R^{(0)}_F, G^{(0)}_F), \qquad H_{F,i,t} = \mathrm{LN}_H(v_{i,t} + R_{F,i}), \qquad g_F = \mathrm{LN}_g\left(\frac{1}{G}\sum_{k=1}^{G} G_{F,k}\right). \tag{4}$$

After removing the padded tail, $H_F$ contains one context-aware state per physical second, while $g_F$ summarizes the fused global-token bank and is not obtained by simply averaging $H_F$. The original Micro states $F_i$ retain their region-bounded forward receptive field after contextualization.

**Multi-granularity representations and task readout.** The backbone exposes three reusable representation levels: local Micro states $F$, contextual temporal states $H_F$, and the recording-level representation $g_F$. For a supervised temporal range $\mathcal{R}$, we denote the contextual readout by $c_{\mathcal{R}} = \mathrm{Read}_F(H_F, g_F; \mathcal{R})$. Recording-level tasks typically use $g_F$, whereas interval-level tasks aggregate the corresponding $H_F$ states and may additionally condition on $g_F$. Separately, $m_{\mathcal{R}} = \mathrm{Read}_\mu(F; \mathcal{R})$ summarizes range-aligned Micro evidence. Our main Concat-Dense configuration retains both scales:

$$z_{\mathcal{R}} = [c_{\mathcal{R}}; m_{\mathcal{R}}], \qquad \widehat{y}_{\mathcal{R}} = h_{\mathcal{T}}(z_{\mathcal{R}}). \tag{5}$$

The read operators specify task-dependent aggregation rather than additional backbone networks, and $h_{\mathcal{T}}$ is the task head. Their exact instantiations for the long-context evaluations are reported in Appendix D. Multiple supervised intervals within one long forward pass share its recording-level context.

**Recording-scale computation.** For $N$ regions, local region radius $r$, and $G$ global tokens, the region–global attention connectivity is

$$\mathcal{O}(N(2r+1+G)+G(N+G)), \tag{6}$$

which grows linearly with recording length for fixed $r$ and $G$. Meanwhile, high-resolution Micro computation remains restricted to fixed-size 10-s regions.

### 2.3 Progressive Multi-Scale Self-Supervised Pretraining

Pretraining progresses from local learning to contextual learning and joint adaptation, followed by full-coverage refinement (Figure 2b). All phases use signal-derived spectral supervision, without clinical or behavioral labels.

**Shared support-aware multi-scale spectral prediction.** Let $P_{j,c,b}$ denote linear power in frequency band $b$ for channel $c$, measured on a four-second support centered at second $j$. For a query beginning at $t$ and spanning $s$ seconds, we define

$$y_{t,c,b,s} = \log\left(\frac{1}{s}\sum_{j=t}^{t+s-1} P_{j,c,b} + \epsilon\right), \qquad \widetilde{y}_{t,c,b,s} = \frac{y_{t,c,b,s} - \mu_{s,b}}{\sigma_{s,b}}, \qquad s \in \{1, 10, 60\}. \tag{7}$$

Linear powers are averaged before taking the logarithm, and $\mu_{s,b}, \sigma_{s,b}$ are training-set statistics. To prevent supervision leakage, we hide each query's complete raw-signal support, $[t-2, t+s+1)$, from both pathways, including a patch-aligned guard. Targets are computed before masking; hidden positions remain distinct from padding or missing data. Appendix C.1 specifies descriptor validity and masking.

**Stage I: local representation learning.** Stage I trains Micro on masked short EEG windows to predict one-second spectral targets using a channel-conditioned local head. The head combines query-aligned patch states with their valid-token summary (Appendix C.2, Equation 9).

**Stage II: context representation learning.** Stage II freezes Micro and trains Macro and Fusion on long crops with queries at $s \in \{1, 10, 60\}$ seconds. A separate contextual prediction head combines query-aligned $H_F$ with the shared $g_F$ (Appendix C.2, Equation 10). Let $\mathcal{L}_s$ average valid SmoothL1 errors within queries and then across queries at scale $s$ (Appendix C.3). For the scales $\mathcal{A}$ present in a crop, the objective is

$$\mathcal{L}_{\text{context}} = \frac{\sum_{s\in\mathcal{A}} w_s \mathcal{L}_s}{\sum_{s\in\mathcal{A}} w_s}, \qquad (w_1, w_{10}, w_{60}) = (0.5, 0.3, 0.2). \tag{8}$$

Context learning begins with content-independent budgeted Micro access, then introduces a learned Macro-conditioned Router under the same objective (Appendix C.4).

**Stage III: joint Macro–Micro adaptation.** Stage III unfreezes Micro and alternates long-context and local updates, allowing both scales to co-adapt while retaining the Stage-I learning signal. The update ratio and optimization paths are specified in Appendix C.4.

**Full-coverage contextual refinement.** Finally, Micro is frozen while the contextual pathway is refined under full Micro coverage with a training-only Bridge (Appendix C.4, Equation 12). Downstream inference uses full coverage and unit-gate Dense Fusion, without Router selection or Bridge weighting. Appendix D.1 reports checkpoint lineage and stage budgets.

## 3 Experiments

We evaluate recording-scale transfer on six long-context datasets and preservation of local transfer on five short-window tasks. Section 4 tests the contribution of contextual support through controlled context-span and representation analyses.

### 3.1 Pretraining and Implementation Setup

We pretrain M$^2$-EEG on approximately 5,300 hours of quality-controlled EEG from 9,437 training recordings, with an additional 498 recordings reserved for self-supervised validation. The corpus spans ten public sources in the TUH, Sleep-EDFx, NMT, FACED, and Siena domains (Obeid & Picone, 2016; Kemp, 2013; Khan et al., 2022; Chen et al., 2023; Detti, 2020). No clinical or behavioral labels from these datasets are used during self-supervised pretraining. Appendix D.1 and Table 6 report the source composition, train–validation separation, and signal preparation.

The final pretrained checkpoint is selected by contextual self-supervised validation loss (Table 8). All main downstream M$^2$-EEG evaluations start from this checkpoint, retaining the complete Macro–Micro backbone with both pathways and aligned fusion for short-window and whole-recording tasks; task readouts and downstream adaptation follow the respective evaluation protocols.

### 3.2 Evaluation Protocol

**Compared backbones.** We compare M$^2$-EEG with six pretrained EEG foundation models: CBraMod, REVE-Large, EEGPT, LaBraM, BIOT, and CSBrain (Wang et al., 2025; El Ouahidi et al., 2025; Wang et al., 2024; Jiang et al., 2024; Yang et al., 2023; Zhou et al., 2025).

**Primary comparison scope.** The long-context comparison evaluates frozen-backbone representation transfer: what recording context is available from pretraining under aligned task readouts? M$^2$-EEG forms a single contextual sequence through Macro and Fusion before task readout; supervised intervals share this backbone-encoded context. All models receive the same task recordings and supervised ranges, with aligned splits, targets, readout families, validation rules, and scorers, while retaining model-native signal interfaces. This includes the shared three-neighbor temporal head for seizure detection, but no additional recording-level contextualizer. Training such a contextualizer for each baseline is a complementary evaluation of downstream temporal adaptation, rather than context supplied by pretraining.

**Reporting.** Tables 1–5 report mean $\pm$ standard deviation; our experiments use three independently initialized downstream runs from one pretrained checkpoint per backbone. Published CSBrain short-window results retain their source summaries (Appendix D.4). This variation reflects downstream optimization rather than uncertainty across pretraining seeds. Best means are bold, second-best means are underlined, and light-blue rows identify M$^2$-EEG; arrows indicate each metric's preferred direction. Appendix D.2 and Tables 11–12 specify the complete comparison contract and task readouts.

We follow official splits where available and audit long-context test-cohort exposure to the SSL train and validation pools. Table 7 distinguishes dataset exposure from recording and subject matches: test recordings are disjoint, while subject-level evidence depends on available repository identities; CHB-MIT and SEED-VIG are absent from pretraining.

### 3.3 Main Results on Long-Context EEG

The six primary datasets span four complementary uses of temporal context: whole-recording classification, contextualized local classification, continuous state regression, and event detection. Appendix D.3 reports the dataset splits, task heads, validation rules, and scorers for all six evaluations.

**Whole-recording abnormality classification.** TUAB (Obeid & Picone, 2016; Schirrmeister et al., 2017) and NMT (Khan et al., 2022) assign one label to each complete usable recording, making them the most direct tests of recording-level representations. Under the aligned frozen-backbone protocol, M$^2$-EEG achieves the highest mean AUROC, AUPRC, and balanced accuracy on both cohorts (Table 1). Its balanced-accuracy margin over the strongest comparison is 0.0035 on TUAB and 0.0399 on NMT.

Table 1: Frozen-backbone whole-recording abnormality classification on TUAB and NMT.

| | TUAB | | | NMT | | |
|---|---|---|---|---|---|---|
| Model | AUROC ↑ | AUPRC ↑ | BAcc ↑ | AUROC ↑ | AUPRC ↑ | BAcc ↑ |
| CBraMod | 0.8787±0.0094 | 0.8787±0.0097 | 0.7814±0.0093 | 0.8564±0.0092 | 0.8811±0.0094 | 0.7871±0.0091 |
| REVE-Large | 0.8970±0.0088 | 0.9044±0.0092 | 0.8297±0.0088 | 0.8777±0.0086 | 0.8953±0.0088 | 0.7799±0.0094 |
| EEGPT | 0.9120±0.0083 | 0.8870±0.0090 | 0.8237±0.0085 | 0.8738±0.0088 | 0.8857±0.0091 | 0.7975±0.0087 |
| LaBraM | 0.9045±0.0085 | 0.9080±0.0093 | 0.8321±0.0082 | 0.8702±0.0089 | 0.8718±0.0093 | 0.7816±0.0092 |
| BIOT | 0.8870±0.0092 | 0.8929±0.0096 | 0.7868±0.0097 | 0.8084±0.0098 | 0.8311±0.0099 | 0.7656±0.0095 |
| CSBrain | 0.8472±0.0080 | 0.8423±0.0110 | 0.7662±0.0100 | 0.8702±0.0070 | 0.8752±0.0090 | 0.7801±0.0110 |
| $M^2$-EEG | **0.9187±0.0082** | **0.9153±0.0094** | **0.8356±0.0074** | **0.9148±0.0075** | **0.9182±0.0079** | **0.8374±0.0083** |

**Full-record sleep staging.** Sleep-EDFx (Kemp, 2013) evaluates local 30-s decisions after each complete PSG recording has been contextualized as a single sequence. Table 2 reports participant-equal test metrics, preventing participants with longer recordings from dominating the evaluation. Under frozen-backbone transfer, $M^2$-EEG leads on all six metrics, improving subject-macro F1 from 0.5638 for the strongest comparison to 0.5880. Section 4 tests this context effect while holding local evidence fixed.

Table 2: Frozen-backbone Sleep-EDFx staging. Metrics are computed per participant, then averaged.

| Model | Accuracy ↑ | BAcc ↑ | Macro-F1 ↑ | Kappa ↑ | AUROC ↑ | AUPRC ↑ |
|---|---|---|---|---|---|---|
| CBraMod | 0.8270±0.0074 | 0.5961±0.0108 | 0.5631±0.0087 | 0.6539±0.0092 | 0.9321±0.0070 | 0.6425±0.0093 |
| REVE-Large | 0.8208±0.0076 | 0.6165±0.0104 | 0.5638±0.0091 | 0.6516±0.0094 | 0.9469±0.0066 | 0.6673±0.0089 |
| EEGPT | 0.8283±0.0072 | 0.5955±0.0107 | 0.5637±0.0089 | 0.6576±0.0091 | 0.9368±0.0069 | 0.6372±0.0092 |
| LaBraM | 0.8163±0.0078 | 0.5837±0.0110 | 0.5460±0.0084 | 0.6325±0.0095 | 0.9272±0.0073 | 0.6242±0.0094 |
| BIOT | 0.8193±0.0077 | 0.5787±0.0109 | 0.5395±0.0086 | 0.6367±0.0096 | 0.9354±0.0071 | 0.6206±0.0093 |
| CSBrain | 0.8120±0.0080 | 0.5754±0.0200 | 0.5482±0.0240 | 0.6325±0.0180 | 0.9355±0.0050 | 0.6265±0.0210 |
| $M^2$-EEG | **0.8374±0.0068** | **0.6287±0.0102** | **0.5880±0.0092** | **0.6885±0.0087** | **0.9541±0.0063** | **0.6846±0.0088** |

**Full-session vigilance regression.** SEED-VIG (Zheng & Lu, 2017) evaluates continuous state estimation by predicting each 8-s PERCLOS target after contextualizing the complete approximately 118-min session. Under frozen-backbone transfer (Table 3), $M^2$-EEG reduces group-equal out-of-fold RMSE by 14.0% relative to the strongest comparison (0.2014 vs. 0.2342) and obtains higher Pearson correlation and concordance correlation. Because the 21 filename-prefix groups have not been independently verified as unique participants, these results are treated as grouped-session evidence rather than a definitive subject-independent estimate.

Table 3: Frozen-backbone SEED-VIG grouped regression. Lower RMSE and MAE are better.

| Model | Group RMSE ↓ | Group MAE ↓ | Pearson $r$ ↑ | CCC ↑ |
|---|---|---|---|---|
| CBraMod | 0.2532±0.0097 | 0.2113±0.0089 | 0.4074±0.0118 | 0.2227±0.0134 |
| REVE-Large | 0.2810±0.0103 | 0.2417±0.0094 | 0.5800±0.0105 | 0.3593±0.0126 |
| EEGPT | 0.2342±0.0092 | 0.1983±0.0085 | 0.5098±0.0110 | 0.3602±0.0127 |
| LaBraM | 0.2481±0.0095 | 0.2084±0.0087 | 0.3891±0.0121 | 0.2426±0.0136 |
| BIOT | 0.2354±0.0094 | 0.1936±0.0084 | 0.4631±0.0113 | 0.3158±0.0131 |
| CSBrain | 0.2492±0.0090 | 0.2087±0.0080 | 0.4143±0.0200 | 0.2025±0.0160 |
| $M^2$-EEG | **0.2014±0.0081** | **0.1732±0.0076** | **0.7542±0.0072** | **0.5689±0.0115** |

**Continuous seizure event detection.** TUSZ (Shah et al., 2018) and CHB-MIT (Shoeb & Guttag, 2010) test whether recording-scale representations improve decisions that remain localized in time. Table 4 reports Event-F1 together with window-level AUROC and balanced accuracy. With frozen backbones and the same temporal head, $M^2$-EEG leads on all three metrics, increasing Event-F1 from 0.3584 to 0.5302 on TUSZ and from 0.6159 to 0.7759 on CHB-MIT.

Table 4: Frozen-backbone seizure detection with the same temporal head and event scorer.

| | TUSZ | | | CHB-MIT | | |
|---|---|---|---|---|---|---|
| Model | Event-F1 ↑ | Win. AUROC ↑ | Win. BAcc ↑ | Event-F1 ↑ | Win. AUROC ↑ | Win. BAcc ↑ |
| CBraMod | 0.3305±0.0118 | 0.7898±0.0104 | 0.7096±0.0108 | 0.3014±0.0132 | 0.8839±0.0094 | 0.7925±0.0099 |
| REVE-Large | 0.3356±0.0124 | 0.8735±0.0092 | 0.7624±0.0099 | 0.6159±0.0135 | 0.9300±0.0087 | 0.8347±0.0090 |
| EEGPT | 0.3584±0.0121 | 0.8556±0.0095 | 0.7733±0.0100 | 0.4839±0.0119 | 0.9395±0.0084 | 0.8099±0.0092 |
| LaBraM | 0.2591±0.0116 | 0.7950±0.0103 | 0.7161±0.0107 | 0.1538±0.0128 | 0.8762±0.0097 | 0.8199±0.0101 |
| BIOT | 0.1698±0.0113 | 0.5895±0.0117 | 0.5619±0.0121 | 0.0635±0.0124 | 0.7936±0.0108 | 0.7431±0.0105 |
| CSBrain | 0.1485±0.0120 | 0.6106±0.0140 | 0.5743±0.0130 | 0.4194±0.0200 | 0.8368±0.0100 | 0.7307±0.0120 |
| $M^2$-EEG | **0.5302±0.0121** | **0.8771±0.0089** | **0.7957±0.0096** | **0.7759±0.0118** | **0.9573±0.0071** | **0.8612±0.0079** |

### 3.4 Short-Window Transfer

We evaluate mental arithmetic (STRESS) (Zyma et al., 2019), motor imagery (BCIC-IV-2a and PhysioNet-MI) (Tangermann et al., 2012; Schalk, 2009), emotion recognition (SEED) (Zheng & Lu, 2015), and major depressive disorder classification (Mumtaz) (Mumtaz, 2016; Mumtaz et al., 2017). Our short-window implementations follow CBraMod's backbone-plus-classifier adaptation design (Wang et al., 2025), with dataset-specific heads for each foundation model. We report test performance of validation-selected probing or fine-tuning configurations; $M^2$-EEG retains its complete Macro–Micro backbone (Appendix D.4).

Table 5: Short-window balanced accuracy under dataset-specific downstream adaptation protocols.

| Model | STRESS | BCIC-IV-2a | SEED | PhysioNet-MI | Mumtaz MDD |
|---|---|---|---|---|---|
| CBraMod | 0.7256±0.0132 | 0.5138±0.0066 | 0.7272±0.0066 | 0.6417±0.0091 | 0.9560±0.0056 |
| REVE-Large | 0.7660±0.0355 | 0.6396±0.0095 | 0.7365±0.0072 | 0.6480±0.0140 | 0.9644±0.0097 |
| EEGPT | 0.5597±0.0171 | 0.5846±0.0070 | 0.7122±0.0022 | 0.5319±0.0148 | 0.8475±0.0233 |
| LaBraM | 0.6909±0.0125 | 0.4869±0.0085 | 0.7318±0.0019 | 0.6173±0.0122 | 0.9409±0.0079 |
| BIOT | 0.6875±0.0186 | 0.4748±0.0093 | 0.7097±0.0024 | 0.6153±0.0154 | 0.9358±0.0052 |
| CSBrain | 0.7558±0.0106 | 0.5657±0.0071 | 0.7302±0.0035 | 0.6304±0.0090 | 0.9643±0.0155 |
| $M^2$-EEG | **0.7805±0.0124** | **0.6582±0.0118** | **0.7405±0.0088** | **0.6612±0.0121** | **0.9721±0.0086** |

Across these dataset-specific protocols, $M^2$-EEG obtains the highest mean balanced accuracy in all five comparisons. Its margin over the strongest comparison ranges from 0.0040 on SEED to 0.0186 on BCIC-IV-2a. The SEED margin is small relative to run-to-run variation and should therefore be interpreted cautiously. Together, these results support preserved short-window transfer after recording-scale pretraining.

## 4 Analysis

We examine training progression, contextual support, and readout design through stage-wise probes and controlled comparisons

**Progressive pretraining.** We compare frozen stage checkpoints using the same linear-probe protocol (Figure 4a). On Sleep-EDFx, Joint subject-macro F1 rises from 0.508 after context learning to 0.567 after joint adaptation and 0.583 after full-coverage refinement. The final-versus-context gain is 0.075 (95% CI $[0.039, 0.111]$); the additional refinement gain remains uncertain. On PhysioNet-MI, a Micro-only probe improves from 0.705 to 0.718 balanced accuracy between local pretraining and the final checkpoint (difference 0.013, $[0.001, 0.025]$; Figure 4b). Thus, along this training lineage, contextual transfer improves without degrading the tested local representation.

**Context and fusion.** Joint has the best point estimate across six tasks (Figure 3), although its larger readout dimension limits capacity-matched interpretation. Holding the 30-s Sleep-EDFx target and its Micro evidence fixed, expanding external support to the full recording improves Joint F1 by 0.0350 (95% CI $[0.0190, 0.0529]$; Figure 4c), with boundary-crossing descriptors masked. We then compare Aligned Joint with Macro–Micro late fusion from the same final backbone, matching the 512-D interface and linear-head protocol. Aligned Joint reaches 0.583 versus 0.562 F1, a paired

gain of 0.021 ($[0.003, 0.039]$; Figure 4d). This supports the fused readout under the tested protocol, rather than isolating a retrained architecture variant. Further controls retain context gains under identical signal coverage (Appendix E.4). To test downstream aggregation of the same frozen local representation on NMT, we compare Micro mean, attention, and temporal-Transformer readouts. They reach at most 0.7481 BAcc versus 0.8175 for Macro–Micro late fusion and 0.8224 for Full Joint (Appendix E.5). Thus, the tested downstream aggregators do not reproduce the Macro-aware gains on NMT.

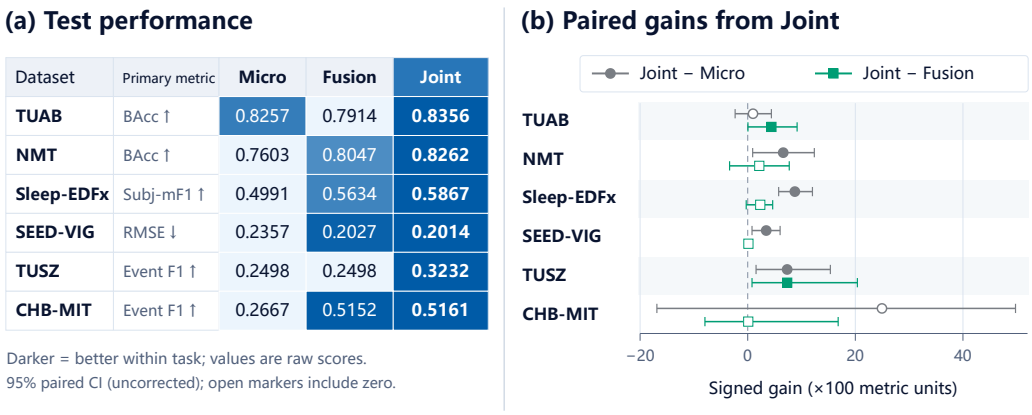


Figure 3: **Complementarity of local and contextual readouts.** **(a)** Raw scores with row-scaled colors (darker is better; RMSE ↓). **(b)** Sign-aligned Joint gains ($\times 100$), with uncorrected 95% paired cluster-bootstrap CIs; filled markers exclude zero. Joint is 512-D; individual branches are 256-D.

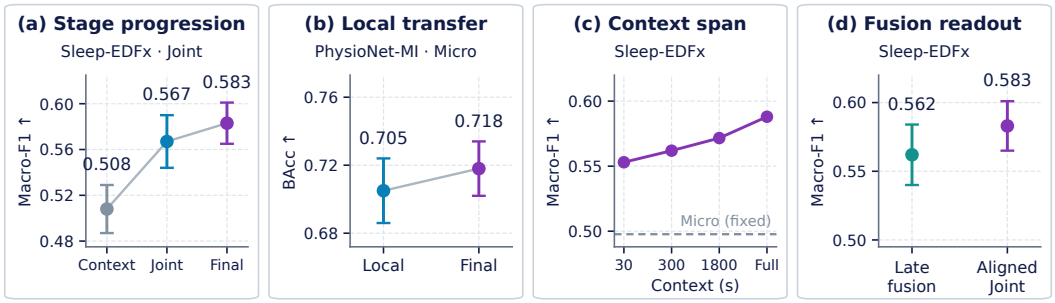


Figure 4: **Training progression, contextual support, and fusion readout.** **(a)** Sleep-EDFx stage probes. **(b)** PhysioNet-MI Micro transfer. **(c)** Fixed-target context control with unchanged Micro evidence. **(d)** Final-backbone readouts with matched 512-D interfaces. Points/bars in (a,b,d): mean ± sample SD of three downstream runs. Panel (c): separate context-control point estimates. Sleep-EDFx Macro-F1 is subject-averaged; paired CIs are reported in Appendix E.

## 5 Conclusion

We introduced M$^2$-EEG, a Macro–Micro pretraining paradigm that separates high-resolution local waveform modeling from lower-bandwidth recording-scale contextualization through aligned fusion. Across six long-context datasets, its frozen representations improve transfer under aligned readouts; the same complete backbone retains strong short-window transfer. Stage-wise probes show improved contextual transfer alongside preserved local discrimination, while controlled analyses show that broader recording context improves a fixed local Sleep-EDFx decision. Aligned Joint outperforms late fusion on the matched Sleep-EDFx readout comparison, whereas late fusion remains competitive on NMT. These findings support reusable recording context within the pretrained backbone, not an architecture-matched ranking of complete downstream systems. Comparisons with separately trained recording-level contextualizers, independent pretraining replications, and untouched external cohorts remain open.

## CODE OF ETHICS

This work studies representation learning for electroencephalography (EEG) using publicly available research datasets. We use the datasets only for the research purposes described in this paper and follow their respective access, licensing, and usage requirements. The proposed model is evaluated as a research method for EEG representation learning and is not intended to provide autonomous clinical diagnosis, treatment recommendations, or medical decision making.

We report known limitations of the evaluation protocols, including uncertainty in cross-repository subject identity, dataset-specific downstream adaptation, and analyses based on a single pretrained checkpoint. No clinical or behavioral labels are used in self-supervised pretraining. We have aimed to avoid overstating clinical implications and to distinguish representation-learning results from evidence of clinical utility. To the best of our knowledge, the work complies with the ICLR Code of Ethics.

## REPRODUCIBILITY STATEMENT

We provide methodological and experimental details required to reproduce the reported results in the main paper and appendices. Section 2 specifies the Macro–Micro architecture, temporally aligned cross-scale fusion, multi-scale self-supervised objectives, and progressive pretraining procedure. The appendices document the spectral targets and masking rules, training-time access mechanisms, pretraining corpus composition, checkpoint lineage, model-native input interfaces, downstream readouts, dataset splits, validation procedures, task-specific evaluation protocols, and statistical analyses.

All main long-context comparisons use fixed pretrained checkpoints and aligned recording identities, splits, targets, readout families, validation rules, and evaluation scorers while retaining each pretrained model's native signal interface. We explicitly report that downstream-run variation does not represent pretraining-seed uncertainty and describe the identifiable overlap-removal procedure and the remaining limits of cross-repository subject de-duplication. Architecture and optimization hyperparameters are provided in the appendix. Code and configuration files will be released upon publication.

## REFERENCES

Peng Cao, Ali Mirzazadeh, Jong Woo Lee, Aleksandar Videnovic, and Dina Katabi. CLEF: EEG foundation model for learning clinical semantics. *arXiv preprint arXiv:2605.10817*, 2026. URL https://arxiv.org/abs/2605.10817.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10:740, 2023. doi: 10.1038/s41597-023-02650-w. URL https://doi.org/10.1038/s41597-023-02650-w.

Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-NTFM: A unified foundation model for EEG signal representation learning. *arXiv preprint arXiv:2509.24222v2*, 2025. URL https://arxiv.org/abs/2509.24222v2. Version 2, revised March 3, 2026; first posted in 2025.

Paolo Detti. Siena scalp EEG database. *PhysioNet*, 2020. doi: 10.13026/5d4a-j060. URL https://doi.org/10.13026/5d4a-j060. Version 1.0.0.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/66969a9e6bd7a26dfeccea7227178ca7-Abstract-Conference.html.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG—adapting to any setup with large-scale pretraining on 25,000 subjects.

In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/20a917f77773ac0fa8bea2bdd6606b66-Abstract-Conference.html.

Jiazhen Hong, Geoffrey Mackellar, and Soheila Ghane. SAMBA: Toward a long-context EEG foundation model via spatial embedding and differential Mamba. *arXiv preprint arXiv:2511.18571*, 2025. URL https://arxiv.org/abs/2511.18571.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Bob Kemp. Sleep-EDF Database Expanded. *PhysioNet*, 2013. doi: 10.13026/C2X676. URL https://physionet.org/content/sleep-edfx/1.0.0/. Version 1.0.0.

Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The NMT scalp EEG dataset: An open-source annotated dataset of healthy and pathological EEG recordings for predictive modeling. *Frontiers in Neuroscience*, 15:755817, 2022. doi: 10.3389/fnins.2021.755817. URL https://doi.org/10.3389/fnins.2021.755817.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://doi.org/10.3389/fnhum.2021.653659.

Wajid Mumtaz. MDD patients and healthy controls EEG data (new). figshare dataset, version 2, 2016. URL https://doi.org/10.6084/m9.figshare.4244171.v2.

Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (EEG)-based computer-aided technique to diagnose major depressive disorder (MDD). *Biomedical Signal Processing and Control*, 31: 108–115, 2017. doi: 10.1016/j.bspc.2016.07.006. URL https://doi.org/10.1016/j.bspc.2016.07.006.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196. URL https://doi.org/10.3389/fnins.2016.00196.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019. doi: 10.1109/TNSRE.2019.2896659. URL https://doi.org/10.1109/TNSRE.2019.2896659.

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. doi: 10.1109/TBME.2022.3147187. URL https://doi.org/10.1109/TBME.2022.3147187.

Huy Phan, Kristian P. Lorenzen, Elisabeth Heremans, Oliver Y. Chén, Minh C. Tran, Philipp Koch, Alfred Mertins, Mathias Baumert, Kaare B. Mikkelsen, and Maarten De Vos. L-SeqSleepNet: Whole-cycle long sequence modeling for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4748–4757, 2023. doi: 10.1109/JBHI.2023.3303197. URL https://doi.org/10.1109/JBHI.2023.3303197.

Jathurshan Pradeepkumar, Zheng Chen, and Jimeng Sun. Neural signals generate clinical notes in the wild, 2026. URL https://arxiv.org/abs/2601.22197.

Gerwin Schalk. EEG motor movement/imagery dataset. *PhysioNet*, 2009. doi: 10.13026/C28G6P. URL https://doi.org/10.13026/C28G6P. Version 1.0.0.

Robin Tibor Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2017. URL https://isip.piconepress.com/conferences/ieee_spmb/2017/papers/l01_01.pdf.

Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018. doi: 10.3389/fninf.2018.00083. URL https://doi.org/10.3389/fninf.2018.00083.

Ali Shoeb and John Guttag. Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th International Conference on Machine Learning*, 2010. URL https://icml.cc/Conferences/2010/papers/493.pdf.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017. doi: 10.1109/TNSRE.2017.2721116. URL https://doi.org/10.1109/TNSRE.2017.2721116.

Siyi Tang, Jared A. Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel L. Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *The Tenth International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2104.08336.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055. URL https://doi.org/10.3389/fnins.2012.00055.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

Nan Zhang, Wei-Long Zheng, Wei Liu, and Bao-Liang Lu. Continuous vigilance estimation using LSTM neural networks. In *Neural Information Processing (ICONIP 2016)*, volume 9948 of *Lecture Notes in Computer Science*, pp. 530–537, 2016. doi: 10.1007/978-3-319-46672-9_59. URL https://doi.org/10.1007/978-3-319-46672-9_59.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497. URL https://doi.org/10.1109/TAMD.2015.2431497.

Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. *Journal of Neural Engineering*, 14(2):026017, 2017. doi: 10.1088/1741-2552/aa5a98. URL https://doi.org/10.1088/1741-2552/aa5a98.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7e199ad8ae40eb19b2980f61f659cb07-Abstract-Conference.html.

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. doi: 10.3390/data4010014. URL https://doi.org/10.3390/data4010014.

# A Large Language Model Usage Statement

Large language models (LLMs) were used as research-assistance tools during the development and preparation of this work. Their use included literature retrieval and discovery, brainstorming and discussion of research ideas and experimental presentation, drafting and revising portions of manuscript text, improving clarity and organization, and assisting with language polishing.

LLMs were not treated as sources of scientific evidence or as substitutes for the underlying experiments. All technical claims, mathematical descriptions, experimental results, numerical values, citations, and conclusions included in the paper were reviewed and verified by the authors against the underlying implementation, experimental outputs, datasets, and/or cited literature. The authors determined the final research design, conducted and interpreted the experiments, selected the reported results, and take full responsibility for the content of the paper.

# B Related Work

This appendix expands the prior-work positioning in Section 1, separating input-length flexibility, task-specific sequence modeling, and recording-scale foundation pretraining.

## B.1 EEG Foundation Models

From BENDR onward, EEG foundation models have improved transfer through self-supervised objectives, heterogeneous datasets, montage adaptation, and larger pretraining corpora (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024; 2025). More recent work broadens spatiotemporal and electrode modeling across acquisition setups (Zhou et al., 2025; El Ouahidi et al., 2025; Döner et al., 2025). Together, these lines have advanced representation learning across common EEG benchmarks and input configurations. Yet the temporal unit presented to the pretraining objective is usually a local segment: CBraMod and CSBrain use 30-s samples, and REVE uses 10-s crops (Wang et al., 2025; Zhou et al., 2025; El Ouahidi et al., 2025). This unit differs from both the duration of the source recordings and the length an encoder can accept. BIOT processes variable-length inputs, and REVE supports arbitrary lengths through positional encoding, but input flexibility alone does not establish recording-scale pretraining (Yang et al., 2023; El Ouahidi et al., 2025). Learning shared context across a recording is therefore a separate design problem from scaling datasets or accommodating different montages.

## B.2 Long-Context EEG Tasks and Modeling

Temporal context is valuable when EEG labels depend on transitions or evidence distributed across time. In sleep staging, sequence-based models have progressed from neighboring epochs in DeepSleepNet, SeqSleepNet, and SleepTransformer toward approximately whole-cycle context in L-SeqSleepNet (Supratak et al., 2017; Phan et al., 2019; 2022; 2023). Similar needs arise in recording-level abnormality assessment on TUH/TUAB, continuous seizure analysis, and vigilance estimation (Obeid & Picone, 2016; Schirrmeister et al., 2017; Shoeb & Guttag, 2010; Tang et al., 2022; Zhang

et al., 2016). Their required temporal scale varies with the task. These studies establish context as a modeling requirement across clinical and behavioral settings, but their representations and temporal ranges are designed around specific tasks: sleep stages, pathological recordings, seizure events, or evolving vigilance. The open question for EEG foundation learning is how to acquire reusable contextual representations that support both local decoding and recording-level interpretation without relying on a different task-specific temporal model for each application.

### B.3 Adapting Foundation Models to Long EEG Recordings

Existing EEG foundation models handle longer recordings through several routes. A local encoder can process successive windows, followed by pooling or a task-specific sequence head; CBraMod, CSBrain, and REVE use the latter for sleep staging (Wang et al., 2025; Zhou et al., 2025; El Ouahidi et al., 2025). Such heads learn cross-window context during downstream adaptation, after the local representation has already been pretrained. BIOT and REVE also accommodate variable-length inputs (Yang et al., 2023; El Ouahidi et al., 2025). SAMBA uses a U-shaped convolutional/Mamba architecture and reports pretraining on 100-s TUAB segments, with transfer to shorter tasks (Hong et al., 2025). CLEF builds session representations from 1,280-s spectrogram inputs, reinjects patch detail, and evaluates both signal-only reconstruction and clinical-text alignment; longer sessions are pooled across windows at evaluation (Cao et al., 2026). CELM aggregates EEG epochs for clinical report generation (Pradeepkumar et al., 2026). These studies already establish multiscale or long-context modeling. M$^2$-EEG takes a complementary approach by explicitly separating high-resolution local waveform modeling from coarse recording-scale contextualization and coupling them during signal-only pretraining. This preserves a reusable local representation while allowing local and interval states to be formed within shared recording context. Task-specific cross-window contextualization is a complementary adaptation setting, distinct from evaluating recording context already supplied by a pretrained backbone.

## C Additional Pretraining Details

This appendix provides the masking rules, prediction heads, loss reduction, and training-time access mechanisms summarized in Section 2.3.

### C.1 Spectral Support and Shared Masking

The linear band power $P_{j,c,b}$ is measured on a four-second support centered at second $j$. Equation 7 averages these powers before taking the logarithm; a longer-scale target is not a single Fourier transform over the full query interval. The normalization statistics $\mu_{s,b}$ and $\sigma_{s,b}$ are estimated from the training set separately for each scale and frequency band. Invalid channel–band combinations do not contribute to the loss. The Macro input's log-RMS descriptor is not itself a prediction target.

A query beginning at $t$ and spanning $s$ seconds depends on raw signal over $[t-2, t+s+1)$, beyond its nominal query timestamps. Supervision is constructed from the original signal before the union of these supports is hidden from both pathways. A guard aligned to Micro patch boundaries is included in the hidden support. Micro centers waveforms using only visible, physically valid samples and replaces patches touching hidden support with mask tokens. Macro emits a second-level descriptor only when its complete four-second support exists and remains visible. Boundary positions and descriptors overlapping hidden support instead receive a mask token. Hidden positions remain on the temporal axis and are distinct from padding or missing data.

### C.2 Local and Contextual Prediction Heads

**Local prediction.** In Stage I, each masked short window passes through Micro alone, producing patch states $F$ and their valid-token summary $m$. For a one-second query at $t$, the prediction head receives

$$z_t^{\mathrm{local}} = \tfrac{1}{2} m + \tfrac{1}{2} \operatorname{Pool}(F_{\mathrm{query}(t)}). \tag{9}$$

Here, $F_{\mathrm{query}(t)}$ contains the Micro patch states aligned with the target second. A channel-conditioned local prediction head maps $z_t^{\mathrm{local}}$ to the standardized one-second target $\widetilde{y}_{t,c,b,1}$. SmoothL1 errors are averaged over valid channel–band targets.

**Contextual prediction.** In Stage II, a crop-level forward pass through Micro, Macro, and Fusion produces $H_F$ and $g_F$, shared by all sampled queries. Each query combines its corresponding contextual interval with the shared global representation:

$$z_{t,s}^{\text{context}} = \text{LN}(W_{\text{mix}} [\text{Pool}(H_F[t : t + s]); g_F] + b_{\text{mix}}) . \tag{10}$$

An independent channel-conditioned contextual head predicts the corresponding spectral targets from $z_{t,s}^{\text{context}}$. This self-supervised prediction path is separate from the downstream Concat-Dense readout.

### C.3 Loss Reduction Across Queries and Scales

For query $q$, let $\widetilde{y}_{q,c,b}$ denote its standardized target, $\mathcal{V}_q$ its valid channel–band pairs, and $\mathcal{Q}_s$ the set of queries at scale $s$. We first average SmoothL1 errors within each query and then within each scale:

$$\ell_q = \frac{1}{|\mathcal{V}_q|} \sum_{(c,b)\in\mathcal{V}_q} \text{SmoothL1}(\widehat{y}_{q,c,b}, \widetilde{y}_{q,c,b}), \qquad \mathcal{L}_s = \frac{1}{|\mathcal{Q}_s|} \sum_{q\in\mathcal{Q}_s} \ell_q. \tag{11}$$

Equation 8 combines the scale-specific losses using fixed weights $(w_1, w_{10}, w_{60}) = (0.5, 0.3, 0.2)$. Weights are renormalized over the scales $\mathcal{A}$ present in each crop. Their relative contribution therefore does not depend on the number of queries sampled at each scale.

### C.4 Training-Time Access and Adaptation

**Context warm-up and routed access.** Stage II first optimizes Macro, Fusion, and the contextual head with content-independent uniform or random budgeted Micro access. This warm-up separates early context learning from optimization of a learned access policy. A Macro-conditioned Router then replaces content-independent access and is optimized jointly with the contextual pathway under the same objective.

During Stage IIb and the long branch of Stage III, the Macro-conditioned Router operates under an approximately 20% region budget. Candidate Micro regions are still evaluated to provide routing gradients, so this access budget is not interpreted as a proportional reduction in pretraining computation.

**Joint adaptation.** Stage III unfreezes Micro and alternates three long-context updates with one local update, rather than summing both objectives within every update. Long updates optimize $\mathcal{L}_{\text{context}}$ through Micro, Macro, Fusion, and the Router. Local updates retain the Stage-I objective through Micro and its local head, preserving a local learning signal during joint adaptation.

**Bridge continuation.** Starting from the selected Stage-III checkpoint, the final full-coverage continuation freezes Micro and retains the same contextual loss. A training-only Bridge assigns a continuous Macro-conditioned weight $a_i$ to the cross-attention residual:

$$v_i = (Q_i + a_i\Delta_i) \odot M_i^{\text{sec}}. \tag{12}$$

Macro, Fusion, the contextual head, and the Bridge are optimized while Micro remains frozen. The main full-recording evaluations discard Bridge scoring and use the unit gate $a_i = 1$.

## D Experimental Details

This appendix provides the corpus, checkpoint, interface, and task-level details supporting Sections 3.1–3.4.

### D.1 Pretraining Corpus and Checkpoint Lineage

**Frozen corpus.** Table 6 reports the data that actually entered the self-supervised pool rather than the advertised size of each public repository. The TUH domain combines six sources—TUAB, TUAR, TUEG, TUEP, TUEV, and TUSZ—so the five sampling domains correspond to ten concrete sources. The training-hour total includes every quality-controlled interval, including valid tails

shorter than 10 seconds; the 1,909,457 complete Stage-I windows instead correspond to 5,304.0 hours. Clinical, sleep-stage, affective, and event labels were not read by the pretraining objective.

Table 6: Frozen self-supervised corpus. Hours denote quality-controlled training intervals before repeated sampling across epochs or stages.

| Sampling domain | Sources | Train records | SSL-val records | Train hours |
|---|---|---|---|---|
| TUH family | 6 | 7,194 | 371 | 2,025.4250 |
| Sleep-EDFx | 1 | 148 | 8 | 2,870.3667 |
| NMT | 1 | 1,964 | 104 | 199.4608 |
| FACED | 1 | 99 | 6 | 126.3994 |
| Siena | 1 | 32 | 9 | 90.2000 |
| Total | 10 | 9,437 | 498 | 5,311.8520 |

Within each sampling domain, available source identity keys were deterministically separated between SSL training and validation. The resulting train/validation identity-key intersections are empty within domains, but a key is not a globally harmonized medical identity: in particular, the same individual may not be identifiable across different repositories or TUH subcorpora.

**Downstream exposure audit.** Table 7 summarizes the six long-context evaluation cohorts against the frozen SSL training and validation pools. Dataset exposure denotes the presence of that dataset in pretraining, not exposure to its test cohort. Exact evaluation-record matches are absent in the audit. Subject matching uses harmonized TUH alpha identifiers and Sleep-EDFx participant identities where available; NMT's record-derived keys do not establish patient separation.

Table 7: Long-context evaluation exposure to the frozen SSL corpus (train + validation). Record entries give matched/evaluated recordings; subject entries give matched/evaluated identities searched against the mapped SSL identities. SEED-VIG covers all sessions evaluated by outer-fold predictions.

| Dataset | Dataset in SSL | Record matches | Subject-level audit | Supported scope |
|---|---|---|---|---|
| TUAB | Yes | 0/276 | 0/253 under mapped TUH alpha IDs | No matches among auditable TUH identities; unmapped identities unresolved. |
| TUSZ | Yes | 0/1,188 | 0/51 under mapped TUH alpha IDs | No matches among auditable TUH identities; unmapped identities unresolved. |
| Sleep-EDFx | Yes | 0/20 | 0/10 under participant IDs | Subject-unseen within Sleep-EDFx repository identities. |
| NMT | Yes | 0/184 | Patient identity unresolved | Record-disjoint; available keys are record-derived. |
| CHB-MIT | No | 0/145 | Dataset absent from SSL | Dataset-unseen transfer. |
| SEED-VIG | No | 0/23 | Dataset absent from SSL | Dataset-unseen; 21 filename-prefix groups are not verified person IDs. |

The TUH subject checks cover harmonized alpha identities across mapped subcorpora; numeric or pseudo-identities without a verified mapping, including unresolved TUEV identities, remain outside that subject-level check. Repository identifiers do not establish global medical identity across datasets. Thus, zero matches under mapped keys are distinct from exhaustive cross-repository subject separation.

All input was converted to physical microvolts and resampled to 200 Hz. Macro descriptors and spectral targets use physical microvolts, whereas Micro receives the same waveforms scaled by $1/100$. Stage-I caching retained 1,909,457 non-overlapping complete 10-s windows. Long-context stages read continuous crops from the source recordings rather than concatenating cached windows;

their requested crop lengths extend to 3,600 seconds, subject to the duration of a quality-controlled continuous interval.

**Checkpoint lineage.** The evaluated checkpoint is inherited through the complete progressive chain in Table 8. The budget column gives the completed optimization budget of a stage, while the selected column identifies the checkpoint inherited by the next stage; these values are not interchangeable.

Table 8: Checkpoint lineage used by all downstream model lineages. Validation-selected intermediate checkpoints, rather than the final optimization step of each stage, are inherited.

| Stage | Optimization role | Micro status | Budget | Selected |
|---|---|---|---|---|
| I | Local masked spectral prediction on 10-s windows | trained | 50 epochs | epoch 50 |
| IIa | Long-context prediction with uniform/random Micro access | frozen | 8,000 steps | step 5,900 |
| IIb | Macro-conditioned routed context learning | frozen | 12,000 steps | step 10,700 |
| III | Alternating three long updates with one local update | trained | 8,000 steps | step 5,800 |
| Final | Full-coverage contextual continuation with the training-only Bridge | frozen | 2,000 steps | step 800 |

The final pretrained checkpoint is selected at step 800 of the full-coverage continuation. Downstream inference loads these weights but replaces the training-time access mechanism with full-coverage Dense Fusion and a unit residual gate.

**Model configuration.** Table 9 specifies the backbone used throughout the progressive chain. Micro, Macro, and Fusion contain 5,092,865, 1,940,993, and 3,621,120 parameters, respectively (10,654,978 in total, excluding prediction and task heads). Micro regions are executed in subbatches of eight while Macro and Fusion retain the shared crop-level context.

Table 9: Pretraining model configuration. FFN denotes feed-forward hidden width.

| Component | Configuration |
|---|---|
| Micro | 200 Hz input; 10-s regions; 50-sample (0.25-s) patches, kernel = stride = 50; width 256, 6 layers, 8 heads, FFN 1024; GELU, post-norm, dropout 0. |
| Macro local encoder | 11 input features per second/channel (five band powers, log-RMS, five validity bits); width 128, 1 layer, 4 heads, FFN 512; pre-norm, ReLU, dropout 0.1. |
| Macro region–global | 4 layers; 16 global tokens; local radius 16 regions on either side. |
| Aligned Fusion | Projection $128 \rightarrow 256$; 8-head cross-attention; temporal alignment scale $\lambda = 2$ s; 2 region–global layers. |
| Region–global blocks | FFN width $4D$; GELU; attention and FFN dropout 0.1. |
| Electrode interface | Vocabulary capacity 512 with separate positive/negative electrode embeddings; visible-sample median centering; no adapters or measured 3D electrode coordinates used. |

**Optimization and sampling.** All stages use AdamW with $\epsilon = 10^{-8}$, gradient-norm clipping at 1, and BF16 autocast with FP32 parameters and optimizer states. Table 10 gives the stage-specific settings; training budgets and inherited checkpoints remain as in Table 8.

Table 10: Stage-specific pretraining settings. Batch sizes are effective global batches over four GPUs; long-crop probabilities apply before conditioning sampling on available continuous signal duration.

| Setting | Stage I | Stage IIa | Stage IIb | Stage III | Final |
|---|---|---|---|---|---|
| Learning rate | $10^{-4}$ initially; $10^{-5}$ continuation | $10^{-4}$ | $10^{-4}$ | Micro: $2\times10^{-5}$; others: $2\times10^{-4}$ | $10^{-4}$ |
| AdamW $\beta_1, \beta_2$ | 0.9, 0.999 | 0.9, 0.95 | 0.9, 0.95 | 0.9, 0.95 | 0.9, 0.95 |
| Weight decay | 0.01 | 0.05 | 0.05 | 0.05 | 0.05 |
| Warmup | 5% of initial 10-epoch budget | 400 steps | 600 steps | 400 steps | 100 steps |
| Global batch | 128 windows | 8 crops | 8 crops | 8 long crops / 128 short windows | 8 crops |
| Crop length (s) | 10 | 300 / 600 / 1200 | 600 / 1200 / 3600 | Long: 1200 / 3600; short: 10 | 1200 / 3600 |
| Length probabilities | 1 | 0.5 / 0.3 / 0.2 | 0.4 / 0.4 / 0.2 | Long: $4/9, 5/9$ | 0.4 / 0.6 |
| Validation frequency | Each epoch | Every 100 steps | Every 100 steps | Every 100 steps | Every 100 steps |
| Validation crop (s) | 10 | 300 | 600 | Long: 1200; short: 10 | 1200 |

Stage I traverses and shuffles all 1,909,457 cached windows per epoch, using 32 windows per GPU. Its 50 epochs comprise seven cosine-scheduled segments ending at epochs 10, 15, 25, 35, 40, 45, and 50. Each continuation retains the preceding model and AdamW states but restarts the cosine schedule at $10^{-5}$ without additional warmup. Thus, this is not one 50-epoch cosine schedule. For later stages, the learning-rate multiplier is $\max\{0.1, \min[1, (k+1)/W]\, c(k)\}$, where $k$ is the update index, $W$ is the warmup budget, and $c(k)$ is the cosine decay factor; the floor also applies during warmup.

Long updates accumulate two sequential batch-one crops per GPU, synchronizing gradients on the second crop, for eight crops per optimizer step. Stage III alternates three long updates with one short update rather than summing both losses at every step. Long-crop sampling uses base domain weights 0.612245/0.204082/0.102041/0.051020/0.030612 for TUH/SleepEDFx/NMT/FACED/Siena, renormalized over domains eligible for the requested duration. Sampling then selects a subject, a record, and a continuous QC interval; crop starts lie on a 10-s grid. Short recordings are neither concatenated nor padded to meet a longer requested crop. Stage I uses a full-window traversal instead of these domain weights.

Stage-I validation uses 2,048 short windows. Long-stage validation uses two fixed crops per sampling domain when eligible (ten crops per gating mode); Stage III additionally monitors 64 short windows, but selects checkpoints using the long-context loss. The recorded seeds are 20260918 for Stages IIa–III and 20260920 for the final stage. The Stage-I implementation defaults to seed 19, but its historical run metadata does not separately record that seed.

**Compute environment.** Pretraining used single-node, four-GPU PyTorch distributed data parallelism with NCCL. The documented training-machine snapshot contains four NVIDIA GeForce RTX 4090 GPUs, each reporting 49,140 MiB (approximately 48 GiB) of memory, an AMD EPYC 7543 CPU (32 cores/64 threads), and 251 GiB system RAM. The GPUs communicate through PCIe/system interconnects without NVLink. The reproduction-environment snapshot comprises Ubuntu 24.04.4 LTS, Python 3.13.13, PyTorch 2.9.1+cu128, CUDA runtime 12.8, cuDNN 9.10.2, NCCL 2.27.5, NumPy 2.5.2, SciPy 1.18.0, MNE 1.12.1, and pyEDFlib 0.1.42. These software versions describe the inspected environment, not an archived immutable container for every historical stage. Recorded training-loop times are 6.68 hours for the seven Stage-I segments combined and 10.29 hours for the complete Stage-IIb run (26.71 and 41.17 allocated GPU-hours, respectively). These are partial costs, not total pretraining time; complete timings for the remaining stages and peak-memory measurements are unavailable.

### D.2 COMMON LONG-CONTEXT EVALUATION CONTRACT

All six long-context tasks freeze the pretrained backbone and train only the task readout. Each reported mean and standard deviation comes from three independently initialized downstream runs from one fixed pretrained checkpoint per backbone; it is not an average over pretraining seeds. Feature standardization, regularization, checkpoint selection, and decision thresholds use only the training and validation partitions of the frozen run. The test partition is scored after these choices are fixed; SEED-VIG instead uses grouped nested cross-validation and reports pooled out-of-fold predictions.

**Model-native signal interfaces.** The representation-transfer comparison retains each model's native numerical input interface. Table 11 summarizes the retained model-native interfaces. Within a task, recording identities, time ranges, labels, split roles, readout family, and scorer are aligned, while amplitude scaling, montage reconstruction, sampling interface, and token pooling remain compatible with the corresponding pretrained model.

Table 11: Frozen-backbone interfaces in the long-context suite. These transformations are part of each pretrained model's input contract.

| Backbone | Retained input interface | Representation supplied to the aligned task readout |
|---|---|---|
| $M^2$-EEG | 200-Hz physical $\mu$V Macro view and $\mu\mathrm{V}/100$ Micro view | Full-record Dense Fusion states plus the range-aligned Micro summary |
| CBraMod | Calibrated physical microvolts under its native encoder | Pooled local patch/token features over the target range |
| REVE-Large | Per-record, per-channel standardization and clipping to $\pm 15$ | 1,216-D final-token pool, aggregated over the target range |
| EEGPT | Native 256-Hz target-encoder interface; deterministic referential reconstruction when required | Official summary tokens pooled over valid patches and the target range |
| LaBraM | Calibrated microvolts divided by 100 with checkpoint electrode identifiers | Valid-duration mean of final normalized patch tokens |
| BIOT | Per-window, per-channel absolute-q95 scaling and clipping, with TCP16 construction when required | Official transformer sequence mean over the target range |
| CSBrain | Model-native 10-s EEG feature interface | Features pooled over the target range and supplied to the same task readout |

**$M^2$-EEG task readouts.** Every long recording is passed through Macro and Fusion as a single contextual sequence. Padding completes only the final 10-s region; invalid patches and seconds are excluded from pooling and task targets. Table 12 specifies the exact fused and Micro evidence exposed to each task head.

Table 12: M$^2$-EEG readouts for the six long-context tasks. Concatenation retains both contextual and local evidence; the backbone remains frozen.

| Task | Supervision unit | Contextual readout | Local Micro readout |
|---|---|---|---|
| TUAB | one label per recording | $g_F$ | valid-duration mean over the recording |
| NMT | one label per recording | $g_F$ | valid-duration mean over the recording |
| Sleep-EDFx | one label per 30-s epoch | $\mathrm{Read}_F(H_F, g_F; 30\ \mathrm{s})$ | mean of the three aligned 10-s region summaries |
| SEED-VIG | one target per 8-s query | $\mathrm{Read}_F(H_F, g_F; 8\ \mathrm{s})$ | independently encoded, exactly aligned 8-s summary |
| TUSZ | one score per 12-s window | mean of aligned $H_F$ states | mean of aligned Micro states |
| CHB-MIT | one score per 10-s window | mean of aligned $H_F$ states | mean of aligned Micro states |

The two 256-D readouts are concatenated before the task head. TUSZ and CHB-MIT do not concatenate $g_F$ directly, but their $H_F$ states have already been formed within the full-recording Fusion context.

### D.3 Task-Specific Long-Context Protocols

**Whole-recording abnormality classification.** The TUAB cohort contains 2,993 recordings divided into 2,175/542/276 training/validation/test recordings under subject separation. Each usable recording is represented with 16 TCP bipolar leads at 200 Hz. The NMT cohort contains 2,403 recordings divided into 1,997/222/184 recordings, using 16 referential leads at 200 Hz. For both datasets, a training-only `StandardScaler` precedes a class-balanced logistic-regression head with $C = 1$. The binary threshold maximizing validation balanced accuracy is then fixed for test evaluation. No recording is truncated to an initial fixed-duration excerpt.

**Full-record sleep staging.** Sleep-EDFx contains 79 recordings from 40 participants. The split contains 20/10/10 participants and 40/19/20 recordings for training/validation/test, respectively. The complete physical EDF timeline is retained, including long wake periods; the Fpz–Cz and Pz–Oz channels are resampled from 100 to 200 Hz. Movement and unknown epochs remain on the timeline but do not enter the loss or metrics. After merging N3 and N4, the task contains 182,661 labeled 30-s epochs over Wake, N1, N2, N3, and REM. A subject-weighted linear softmax head is selected on validation macro-F1. The primary metric first computes macro-F1 within each participant and then averages across participants, preventing long recordings from dominating the reported value.

**Full-session vigilance regression.** SEED-VIG contains 23 approximately 118-min sessions with 17 EEG channels and 885 eight-second PERCLOS targets per session. The evaluation uses five outer and three inner grouped folds over 21 filename-prefix groups. The inner folds select ridge regularization using group-weighted validation performance; outer-fold predictions are clipped to $[0, 1]$ and combined only after every group has served as held-out data. RMSE, MAE, Pearson correlation, and concordance correlation are calculated per group before averaging. Because the filename prefixes have not been independently verified as unique participant identifiers, this protocol is evidence under grouped session separation rather than a confirmed subject-independent test.

**Continuous seizure event detection.** TUSZ contains 4,295 recordings with subject-disjoint train/validation/test partitions of 1,597/1,510/1,188 recordings and 316/55/51 participants. Each recording is encoded independently, after which non-overlapping 12-s windows are classified. CHB-MIT contains 676 recordings split into 372/159/145 recordings from 14/5/4 non-overlapping patient groups; 10-s windows are evaluated every 5 seconds, and the test partition contains 39 reference seizures.

For both datasets and every compared backbone, the frozen per-window representation is passed to the same three-neighbor temporal head: the previous, current, and next windows from the same

recording are concatenated and processed by a 128-unit MLP with GELU activation and dropout 0.2. Boundary windows replicate the nearest in-recording neighbor. Each downstream rerun is trained independently, and its event threshold is selected by validation Event-F1. Positive windows are projected back to the physical timeline and merged under the frozen event scorer. Predicted and reference events are matched one-to-one when temporal intersection-over-union is at least 0.1; unmatched predictions and references become false positives and false negatives. Window AUROC/BAcc and event F1 are computed separately. In the available TUSZ annotation export, intervals without an event label are treated as background; annotation coverage has not been independently re-audited and is therefore a boundary of this evaluation.

### D.4 Short-Window Protocols

Table 13 summarizes the short-window datasets and evaluation splits. For each dataset, we select the downstream configuration using validation data and report its test performance, rather than selecting by test score. CSBrain short-window results are taken from Zhou et al. (2025), except for three-class SEED, which uses the CSBrain* entry in Table 1 of Chen et al. (2025) (arXiv v2).

Table 13: Short-window datasets included in the complete cross-model comparison. Counts are train/validation/test windows or trials.

| Dataset | Target and duration | Split counts | Participant separation | Signal preparation |
|---|---|---|---|---|
| STRESS | rest vs. arithmetic, 5 s | 1,707 total; 28/4/4 participants | subject-disjoint; four test participants | 20 channels, 500→200 Hz |
| BCIC-IV-2a | four-class imagery, 4 s | 2,880/1,152/1,152 | A01–05/A06–07/A08–09 | 22 channels, CAR, 0.5–99.5 Hz, unlabeled Euclidean alignment |
| SEED | three-class emotion, 5 s | 10,125 each | same 15 participants across sessions 1/2/3 | 60 mapped channels; unlabeled session standardization |
| PhysioNet-MI | four-class imagery, 4 s | 6,300/1,734/1,803 | S001–070/S071–089/S090–109 | average reference, 0.3-Hz high-pass, 60-Hz notch |
| Mumtaz MDD | HC vs. MDD, 5 s | 4,891/1,041/1,211 | subject-disjoint; 43/9/11 participants | 19 channels, 256→200 Hz, 0.3–30 Hz, 50-Hz notch |

**Downstream adaptation.** Following CBraMod's downstream implementation design (Wang et al., 2025), our short-window runs pair each pretrained backbone with a dataset-specific classifier. Configurations are selected from the evaluated probing and fine-tuning settings using validation performance; classifier architectures can differ across tasks. Selected $M^2$-EEG configurations combine backbone representations with auxiliary waveform or spatial features; SEED uses participant-specific heads after STRESS adaptation. PhysioNet-MI additionally uses prediction ensembling, test-time augmentation, and unlabeled participant adaptation. BCIC-IV-2a fine-tunes the backbone; the other selected configurations freeze it during task-head fitting.

STRESS has only four held-out participants, and SEED separates sessions rather than participants. The fixed evaluation partitions were available during iterative development; the results are development-test evidence under the reported protocols rather than a new untouched blind evaluation.

## E Analysis Protocols and Statistical Scope

### E.1 Common Analysis Contract

All analyses freeze the evaluated backbone and use unit-gate Dense Fusion where applicable. The stage-wise comparison uses the selected checkpoints along the training lineage; the remaining analyses use the final pretrained checkpoint. Only the analysis-specific probe, aggregator, or task head is refitted. Downstream seeds and training budgets differ from the main evaluations, so absolute

scores need not match Tables 1–4. Only within-analysis contrasts are interpreted. The readout-complementarity and context-control analyses use 5,000 paired cluster-bootstrap resamples at the reporting unit, with per-comparison, uncorrected 95% intervals conditional on one pretrained checkpoint. The aggregation comparison reports the mean and sample standard deviation across three downstream runs and computes paired intervals within each seed.

### E.2 Stage-Wise Transfer and Fusion Readouts

The stage comparison uses the selected context-learning, joint-adaptation, and final checkpoints in Table 8. All are evaluated with full Micro coverage and a unit residual gate. Sleep-EDFx uses a 512-D Joint readout formed by concatenating the mean of target-aligned $H_F$ states with the mean Micro summary over the three aligned 10-s regions; $g_F$ is not concatenated directly. The PhysioNet-MI comparison uses only the valid-token Micro mean within each 4-s trial from the local and final checkpoints, without auxiliary branches or test-time adaptation. The data splits follow Appendix D.3 and Table 13. Both comparisons fit a linear classifier with training-only feature standardization and validation selection under the same search protocol within each task.

Table 14: Frozen stage probes. Values are mean ± sample SD across three downstream runs. The two columns use different datasets, metrics, and readouts; dashes denote settings outside the respective comparison.

| Checkpoint | Sleep-EDFx Joint F1 | PhysioNet-MI Micro BAcc |
|---|---|---|
| Local pretraining | — | $0.7050 \pm 0.0190$ |
| Context learning | $0.5080 \pm 0.0210$ | — |
| Joint adaptation | $0.5670 \pm 0.0230$ | — |
| Full-coverage refinement | $0.5830 \pm 0.0180$ | $0.7180 \pm 0.0160$ |

For the final-backbone Sleep-EDFx comparison, late fusion replaces the target-aligned $H_F$ mean with the mean contextual Macro region state $R_M$ over the same target, retaining the identical Micro summary. Both concatenations are 512-dimensional and use the same linear-head protocol. Late fusion obtains $0.5620 \pm 0.0220$ subject-macro F1 versus $0.5830 \pm 0.0180$ for Aligned Joint; the latter reuses the final-stage evaluation in Table 14. This is a readout comparison on jointly pretrained representations, not a no-fusion pretraining ablation.

Table 15: Paired effects for the stage and readout comparisons. Intervals use 5,000 paired subject-bootstrap resamples, averaging each resampled difference over three downstream runs; they are per-comparison, uncorrected 95% intervals conditional on the trained lineage.

| Comparison | Difference | 95% CI |
|---|---|---|
| Sleep: joint adaptation − context learning | 0.0590 | $[0.0220, 0.0960]$ |
| Sleep: final − context learning | 0.0750 | $[0.0390, 0.1110]$ |
| Sleep: final − joint adaptation | 0.0160 | $[-0.0120, 0.0440]$ |
| PhysioNet-MI: final Micro − local Micro | 0.0130 | $[0.0010, 0.0250]$ |
| Sleep: Aligned Joint − late fusion | 0.0210 | $[0.0030, 0.0390]$ |

The stages differ in cumulative training as well as optimization regime; their probe trajectory does not isolate schedule effects at equal budgets. Full coverage improves the Sleep-EDFx point estimate after joint adaptation, but that incremental interval includes zero.

### E.3 Readout Definitions and Representation Scope

For a supervised range $\mathcal{R}$, the three readouts in the complementarity analysis are

$$z_{\text{Micro}} = \text{Read}_{\mu}(F; \mathcal{R}), \qquad z_{\text{Fusion}} = \text{Read}_{F}(H_F, g_F; \mathcal{R}), \qquad z_{\text{Joint}} = [z_{\text{Micro}}; z_{\text{Fusion}}]. \quad (13)$$

Micro and Fusion are 256-dimensional, whereas Joint is their 512-dimensional concatenation. Fusion already includes the learned Micro-to-Fusion interaction. The complementarity analysis therefore tests whether a direct local bypass adds information under the stated probe, not whether

Macro alone is sufficient. The nonzero paired effects in Figure 3 occur for Joint relative to both branches on TUSZ, to Micro on NMT, Sleep-EDFx, and SEED-VIG, and to Fusion on TUAB. On the four-participant CHB-MIT test cohort, Joint–Fusion is 0.0010 with a wide 95% interval of $[-0.0795, 0.1681]$. Because Joint doubles the readout dimension, this experiment establishes complementarity under the stated probe rather than superiority to every capacity-matched alternative.

### E.4 Context Controls

Figure 5 reports the fixed-target context control. The context-partitioning control instead preserves the complete Sleep-EDFx waveform and Micro input while resetting Macro/Fusion context at 30-, 300-, or 1,800-s block boundaries (Figure 6a). Joint rises from 0.5522 under independent 30-s context to 0.5880 under unified full-record context (Figure 6b), a paired difference of 0.0358 (95% CI $[0.0206, 0.0531]$). The full-record differences relative to 300 and 1,800 s are 0.0263 ($[-0.0026, 0.0606]$) and 0.0148 ($[-0.0092, 0.0402]$); Fusion is non-monotonic, with a full-record minus 30-s difference of 0.0409 ($[-0.0047, 0.0792]$).

The fixed-target context control retains one local target and its Micro representation, expands only recording-anchored external support, and masks coarse descriptors crossing a boundary. Against strict-local support, full-record Joint improves subject-macro F1 by 0.0350 (95% CI $[0.0190, 0.0529]$); at full support, it exceeds Micro by 0.0903 and Fusion by 0.0342, with both intervals excluding zero. Its full-record Joint differences relative to 300- and 1,800-s support are 0.0261 ($[-0.0023, 0.0601]$) and 0.0164 ($[-0.0080, 0.0436]$). Both controls use Sleep-EDFx and are not independent cross-dataset replications.

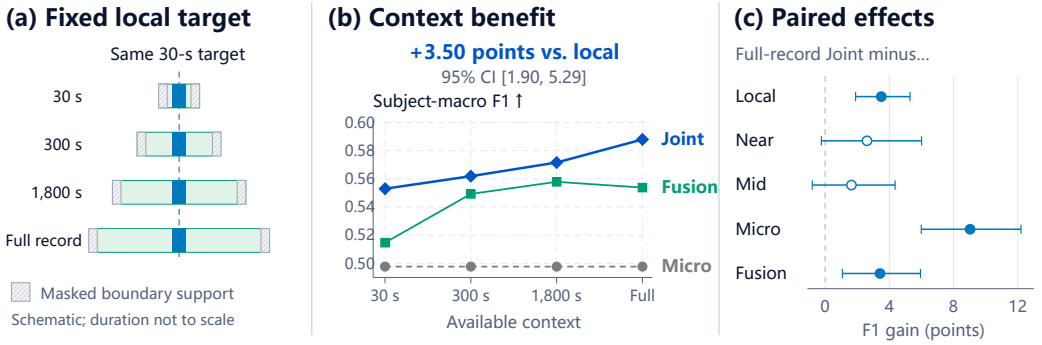


Figure 5: **Recording-wide context improves a fixed Sleep-EDFx decision.** **(a)** Fixed target and Micro evidence; hatched edges mask boundary support. **(b)** Subject-macro F1. **(c)** Paired effects with 95% cluster-bootstrap CIs; filled markers exclude zero. The full-record score belongs to this context-control protocol, separately from the stage/readout comparison.

### E.5 Aggregation Alternatives

The aggregation comparison uses the frozen pretrained backbone on NMT. It compares Micro mean, attention, and temporal-Transformer aggregation with two Macro-aware alternatives. Full Joint reaches $0.8224 \pm 0.0159$ balanced accuracy versus 0.7089–0.7481 for Micro-only aggregation (Figure 7a); its paired interval excludes zero against attention in all three seeds and against mean or temporal-Transformer aggregation in two (Table 16). Macro–Micro late fusion reaches $0.8175 \pm 0.0115$, only 0.0049 below Full Joint, and is slightly higher on AUROC and AUPRC (Figure 7b–c); all three Full Joint–late-fusion intervals cross zero. These results support retaining Macro context, but not native Fusion over every Macro-aware combination.

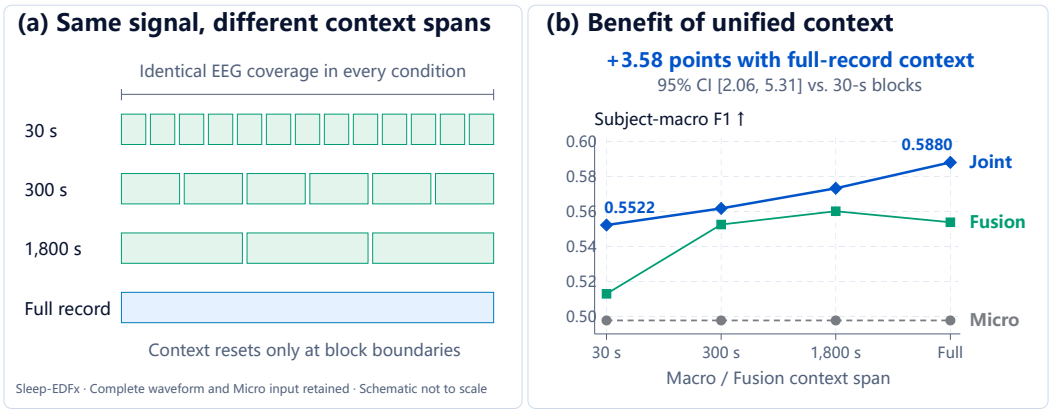


Figure 6: **Unified full-record context outperforms independently contextualized short blocks on Sleep-EDFx.** **(a)**, Complete waveform and Micro evidence are retained while only Macro/Fusion context boundaries change; the schematic is not to scale. **(b)**, Subject-macro F1 across context spans; the annotation gives the paired full-record minus 30-s Joint difference and 95% cluster-bootstrap interval. Connecting lines show point estimates, not uncertainty bands.

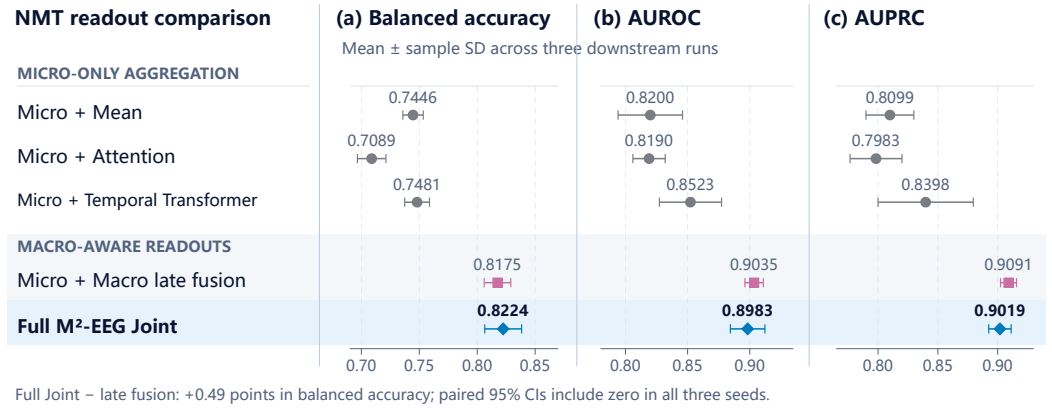


Figure 7: **Micro-only downstream aggregation does not match Macro-aware readouts on NMT.** Balanced accuracy **(a)**, AUROC **(b)**, and AUPRC **(c)** show mean and sample standard deviation over three independently initialized downstream runs. Shading separates Micro-only and Macro-aware readouts; Full Joint leads balanced accuracy, but its 0.0049 advantage over late fusion is unresolved by the seed-level paired intervals.

Table 16: Seed-wise paired NMT balanced-accuracy difference (Full Joint minus alternative) with 95% confidence interval.

| Alternative | Seed 11 | Seed 29 | Seed 47 |
|---|---|---|---|
| Micro + Mean | $0.1039\ [0.0488, 0.1640]$ | $0.0706\ [0.0064, 0.1369]$ | $0.0590\ [-0.0044, 0.1250]$ |
| Micro + Attention | $0.1442\ [0.0760, 0.2154]$ | $0.0921\ [0.0226, 0.1626]$ | $0.1041\ [0.0378, 0.1700]$ |
| Micro + Temporal Transformer | $0.0909\ [0.0277, 0.1574]$ | $0.0766\ [0.0109, 0.1412]$ | $0.0553\ [-0.0109, 0.1261]$ |
| Micro + Macro late fusion | $0.0326\ [-0.0109, 0.0766]$ | $-0.0171\ [-0.0556, 0.0204]$ | $-0.0007\ [-0.0336, 0.0319]$ |