NEURAL STATE PREDICTION: OBSTRUCTING SHORTCUT LEARNING IN EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models increasingly use masked prediction to learn from unlabeled recordings, but optimizing this objective does not ensure transferable neural representations. A central challenge is that stable positional cues and local correlations can make masked regions predictable without integrating distributed neural context. To reduce this reliance on low-information prediction paths, we introduce **Neural State Prediction (NSP)**, a latent-predictive framework that constrains both the prediction target and the available context. NSP uses a Target Encoder updated by an exponential moving average (EMA) to define latent supervision. **Identity residualization** removes additive effects associated with channel identity and relative time from the targets, while **topology-separated context** excludes their immediate spatial and temporal neighborhood from the visible input. We pretrain NSP on 2.2 million EEG segments from TUEG and evaluate it across 30 downstream datasets spanning clinical diagnosis, sleep staging, emotion recognition, motor imagery, event-related potentials, cognitive-state decoding, and language retrieval. Under full-parameter multi-task fine-tuning on EEG-FM-Bench, NSP achieves **63.94 macro balanced accuracy** across 14 datasets, exceeding the strongest evaluated baseline by **2.35 percentage points**. Controlled component ablations assess the contribution of each mechanism, while matched context controls and held-out interventions characterize the role of context geometry, signal content, and positional information. Jointly designing latent targets and their context offers a promising direction for EEG foundation models that learn from distributed signal structure. Our code is available at anonymous.4open.science/r/Neural-State-Prediction-7289/.

1 INTRODUCTION

Electroencephalography (EEG) records millisecond-scale neural dynamics non-invasively and supports applications ranging from clinical monitoring to emotion recognition and brain–computer interfaces (BCIs). Many EEG systems are still developed for individual tasks and datasets from limited labeled cohorts (Lawhern et al., 2018; Song et al., 2022; 2021). These specialized models remain useful, but often transfer imperfectly across acquisition systems, subject populations, and prediction targets. Inspired by self-supervised learning in language (Brown et al., 2020; Devlin et al., 2019) and vision (Dosovitskiy et al., 2020; He et al., 2022), EEG foundation models (EEG-FMs) instead pretrain reusable encoders on broad unlabeled corpora (Kostas et al., 2021; Jiang et al., 2024b; Wang et al., 2024; 2025; Zhou et al., 2025; Yang et al., 2023; El Ouahidi et al., 2025; Panchavati et al., 2026). EEG-FMs have improved cross-dataset transfer, advancing general-purpose EEG representations.

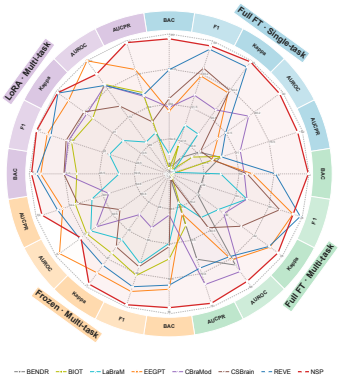

Figure 1: **EEG-FM-Bench macro summary.**

Recent EEG-FMs use four main pretraining paradigms (Figure 2): **(1)** *Contrastive learning.* These methods learn representations by bringing related EEG views closer while separating unrelated views. BIOT (Yang et al., 2023) aligns representations across biosignal segments and modalities, whereas BENDR (Kostas et al., 2021) contrasts masked contextual representations with their corre-

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.

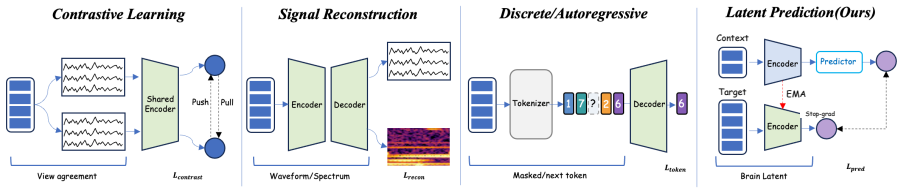

Figure 2: **Four EEG foundation-model pretraining paradigms.** Contrastive learning aligns related views; signal reconstruction recovers waveforms or spectra; and discrete/autoregressive modeling predicts masked or next tokens. In the latent-prediction panel, NSP's Predictor maps Context Encoder states to stop-gradient Target Encoder representations.

sponding latent targets. Their success nevertheless depends on how positive views, negatives, and augmentations define invariance, and view-level agreement does not directly specify which missing neural content should be recovered. **(2)** *Signal reconstruction.* Reconstruction objectives mask part of an EEG recording and recover the omitted waveform, spectrum, or patch. EEGPT (Wang et al., 2024) and REVE (El Ouahidi et al., 2025) reconstruct masked waveforms, CBraMod (Wang et al., 2025) reconstructs EEG patches with criss-cross spatial–temporal attention, and CSBrain (Zhou et al., 2025) reconstructs spectrogram representations. Such objectives preserve signal detail, but may devote capacity to amplitude, noise, and other input-level variation that is not useful downstream. **(3)** *Discrete or autoregressive prediction.* Discrete approaches encode EEG as codebook tokens, whereas autoregressive approaches predict the next token or patch from preceding context. LaBraM (Jiang et al., 2024b), DeWave (Duan et al., 2023), and CodeBrain (Ma et al., 2025) learn or use discrete EEG tokens, while ECHO (Liu et al., 2026), THD-BAR (Yang et al., 2025), and KAST-BAR (Wang et al., 2026) predict sequential tokens or future patches. Discretization can provide compact supervision, although its quality is bounded by the tokenizer or codebook. **(4)** *Latent prediction.* These methods predict learned representations of masked regions, typically through joint-embedding predictive architectures (JEPAs) with a context and a target encoder. EEG2Rep (Mohammadi Foumani et al., 2024) combines latent self-prediction with semantic subsequence preserving (SSP) masking. Brain-JEPA (Dong et al., 2024) models latent brain dynamics, S-JEPA (Guetschel et al., 2024) targets cross-dataset EEG transfer, and Laya (Panchavati et al., 2026), STST-JEPA (Segal et al., 2026), and EEG-JEPA (Li et al., 2026) develop learned-target prediction for EEG with different target construction and masking choices. PATCHCODE (Yu et al., 2026) adds discrete-code supervision from a frozen tokenizer to continuous patch-level latent prediction. These studies make latent prediction a promising alternative to reconstructing every detail of noisy electrophysiological inputs.

Despite these advances, dependence on easily predictable signal structure can hinder the development of EEG representations that generalize across tasks and recording conditions. Pretraining objectives reward agreement with a chosen target, but do not by themselves distinguish transferable neural structure from stable positional cues or local correlations. Motivated by the shared source mixtures that volume conduction induces at neighboring scalp electrodes (Paul, 2005), we examine local interpolation alongside prediction from channel identity, electrode coordinates, and relative time. A masked objective may be reduced through either path without requiring broad neural context. Latent prediction avoids direct waveform reconstruction, but it does not automatically remove these low-information solutions: existing EEG JEPAs define learned targets and masked context, yet generally do not jointly reduce additive channel- and time-conditioned components in the target and exclude the target's nearest spatial and temporal evidence. The central objective-design question is therefore not only whether to predict a waveform, token, or representation, but also what information remains in the prediction target and what evidence is available to infer it.

To address this mismatch between pretraining success and transferable representation learning, we propose **Neural State Prediction (NSP)**, a latent-predictive framework that jointly constrains target content and available context. An EMA-updated Target Encoder first provides stable latent supervision for masked spatiotemporal regions. Identity residualization then removes the additive components associated with channel identity and relative time from each target, reducing position-conditioned predictability. In parallel, topology-separated context withholds the target's nearest-neighbor channels and adjacent time patches from the visible EEG content, requiring the Context Encoder to integrate more distant evidence. The residualized target specifies *what* is predicted, whereas the topology-separated context specifies *from what evidence* the prediction is formed. We

pretrain NSP on 2.2 million EEG segments from TUEG (Obeid & Picone, 2016) and evaluate transfer across 30 downstream datasets spanning clinical diagnosis, motor imagery, emotion, cognitive workload, event-related potentials, language, and other brain-state decoding tasks. The evaluation comprises 16 selected NeuralBench-EEG-Core v1.0 datasets (Banville et al., 2026) (NeuralBench for short), after removing dataset–task pairs also covered by EEG-FM-Bench, together with the 14-dataset EEG-FM-Bench protocol (Xiong et al., 2025). Figure 1 compares representative EEG-FMs from all four pretraining paradigms under EEG-FM-Bench; NSP achieves 63.94 macro balanced accuracy under full-parameter multi-task fine-tuning, outperforming the strongest evaluated baseline by 2.35 percentage points. Matched component and context-control ablations examine how the proposed mechanisms affect transfer, and held-out signal-content interventions and identity probes characterize sensitivity to EEG content and the positional information retained by component variants.

Our contributions are summarized as follows:

- We identify **position-conditioned prediction** and **local interpolation** as two low-information paths in masked EEG latent prediction, motivating joint design of targets and visible context.
- We propose **Neural State Prediction (NSP)**, a latent-predictive **EEG-FM pretraining framework**. Its EMA Target Encoder supplies supervision, while **identity residualization** and **topology-separated context** jointly constrain target content and visible EEG evidence.
- We evaluate NSP across **30 downstream datasets**. Under EEG-FM-Bench multi-task full fine-tuning, it achieves **63.94 macro BAC**, 2.35 percentage points above the strongest evaluated baseline. Matched ablations, signal-content interventions, and identity probes assess the components and learned representations.

2 Neural State Prediction

Neural State Prediction (NSP) learns an EEG representation by predicting the latent states of masked spatiotemporal regions from the remaining recording. As summarized in Figure 3, a Target Encoder operating on the complete token grid defines the latent state to be predicted. Identity residualization removes batch-estimated channel and relative-time components from that state, while topology-separated context controls the evidence available to the Context Encoder. These two constructions have complementary roles: the residualized target specifies *what* is predicted, and the separated context specifies *from what evidence* it is inferred.

2.1 Overview and representation encoder

Let $\mathbf{X} \in \mathbb{R}^{B\times C\times T}$ be a batch of EEG segments, where B is the batch size, C is the number of channels, and T is the number of samples per segment. Each channel is partitioned into P temporal patches, giving the position set $\Omega = [P] \times [C]$; each position carries a d-dimensional state. For example b, $\mathcal{M}_b \subset \Omega$ denotes the target region and $\mathcal{G}_b \subset \Omega$ its spatial–temporal guard. Only EEG content outside $\mathcal{M}_b \cup \mathcal{G}_b$ is visible to the Context Encoder.

The Target and Context Encoders share the same architecture but have separate parameters, denoted by ϕ and θ respectively. Their tokenizers combine three sources of information: a projection of the EEG patch, a learned channel embedding, and adaptive channel position encoding (ACPE) from electrode coordinates. The patch projection is the signal-dependent term. Masking replaces this term before channel and coordinate information is added, so a masked position remains represented in the grid without exposing its EEG content. This ordering keeps masked positions addressable by channel and coordinate while withholding their EEG samples.

Both branches use a Cascaded Spatial–Temporal Transformer (CST). Each block applies spatial multi-head attention (MHA) across channels, then temporal attention across patches within each channel, followed by a feed-forward network (FFN). Temporal attention uses rotary position encoding (RoPE) (Su et al., 2024). We write $\mathcal{S} = \{\mathcal{S}_b\}_{b=1}^{B}$ for a batch of mask sets and $\mathbf{H}_\eta^{(\ell)}(\mathbf{X}; \mathcal{S})$ for the state after block ℓ of encoder $\eta \in \{\theta, \phi\}$ when EEG content at $\mathcal{S}_b$ is masked in example b. The final block is denoted by N_{blk}. The Target branch uses empty mask sets, whereas the Context branch uses $\mathcal{S}_b = \mathcal{M}_b \cup \mathcal{G}_b$.

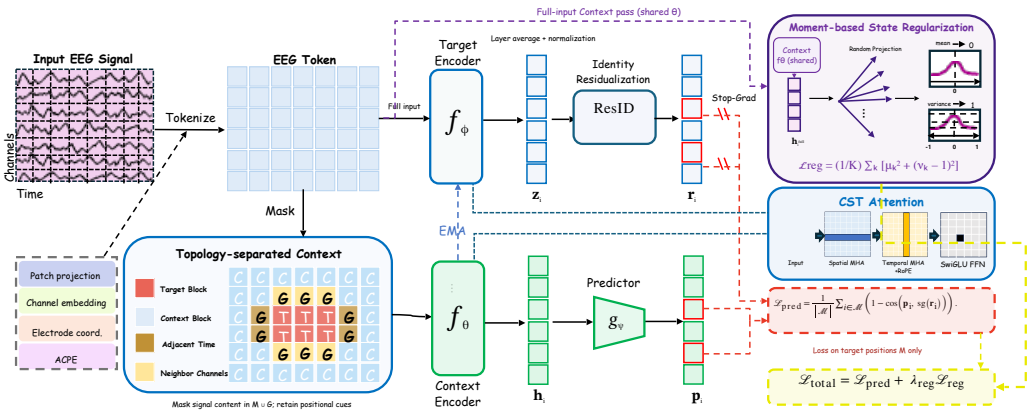

Figure 3: **Neural State Prediction architecture.** Residualized, stop-gradient Target states supervise the masked Context branch through its predictor. A separate complete-input pass through the shared Context Encoder supplies moment-based state regularization. The Target Encoder follows the Context Encoder by EMA; prediction is scored only at target positions.

2.2 Target Construction: Identity-Residualized Latent States

The Target Encoder sees the complete recording. To make its supervision less dependent on any single block, we average a set $\mathcal{I}_T$ of late Target states, apply encoder layer normalization (LN), and standardize each sample–channel slice to zero mean and unit variance over its temporal and feature dimensions:

$$\mathbf{z} = \mathrm{Std}_{P\times d}\left[\mathrm{LN}\left(\frac{1}{|\mathcal{I}_T|}\sum_{\ell\in\mathcal{I}_T}\mathbf{H}_\phi^{(\ell)}(\mathbf{X};\varnothing)\right)\right]. \tag{1}$$

This gives a target latent $\mathbf{z}_{b,p,c} \in \mathbb{R}^d$ for each position. Because this aggregation retains the full patch–channel grid, NSP predicts a state for each masked position rather than a pooled recording-level vector.

Identity residualization removes additive channel and relative-time effects estimated from the current batch. Let $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ be the per-feature mean and variance of $\mathbf{z}$ over all batch, patch, and channel positions. We first place each feature on a common scale,

$$\mathbf{u}_{b,p,c} = \frac{\mathbf{z}_{b,p,c} - \boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2 + \varepsilon}}, \tag{2}$$

where ε is a numerical stabilizer. On this common scale, we estimate the additive component associated with channel identity and relative patch time:

$$\boldsymbol{\mu}_c^{\mathrm{ch}} = \frac{1}{BP}\sum_{b,p}\mathbf{u}_{b,p,c}, \qquad \boldsymbol{\mu}_p^{\mathrm{time}} = \frac{1}{BC}\sum_{b,c}\mathbf{u}_{b,p,c}. \tag{3}$$

The prediction target is the remaining state,

$$\mathbf{r}_{b,p,c} = \mathbf{u}_{b,p,c} - \boldsymbol{\mu}_c^{\mathrm{ch}} - \boldsymbol{\mu}_p^{\mathrm{time}}. \tag{4}$$

This operation is indexed by channel and relative patch time; it does not regress directly on continuous electrode coordinates. We stop gradients through $\mathbf{r}$ before it is used as supervision. Standardization and residualization play different roles: standardization places latent dimensions on comparable scales across the batch, while the two marginal means remove the additive component shared by a channel or a relative time index, leaving residual variation not explained by those two marginals. Electrode coordinates remain available to the encoders; they are not themselves fitted or subtracted by this operation.

The Target Encoder receives no optimizer update. Following self-distillation and joint-embedding predictive learning (Grill et al., 2020; Assran et al., 2023), after each Context Encoder update its parameters follow an exponential moving average:

$$\phi_{s+1} = \tau_s\phi_s + (1-\tau_s)\theta_{s+1}, \tag{5}$$

where τ_s is the scheduled EMA coefficient. EMA provides a slowly varying target while gradients update the Context branch and predictor.

2.3 CONTEXT CONSTRUCTION: TOPOLOGY-SEPARATED EVIDENCE

For each example, target positions $\mathcal{M}_b$ are sampled as connected spatiotemporal blocks rather than as independent tokens. This keeps the prediction target locally coherent. To prevent the Context Encoder from recovering a target by interpolation from its immediate surroundings, we expand every target block with a guard. Let $\mathcal{N}_t$ denote temporal neighborhood expansion and $\mathcal{N}_s$ spatial expansion according to electrode distance. The guard is

$$\mathcal{G}_b = (\mathcal{N}_t(\mathcal{M}_b) \cup \mathcal{N}_s(\mathcal{M}_b)) \setminus \mathcal{M}_b. \tag{6}$$

The prediction loss is evaluated only on $\mathcal{M}_b$; $\mathcal{G}_b$ changes the available evidence but adds no extra targets.

Let $\mathbf{q}^{\theta}_{b,p,c}$ denote the signal-dependent patch embedding in the Context tokenizer and $\mathbf{m}^{\theta}$ a learned mask embedding. With $\mathcal{S}_b = \mathcal{M}_b \cup \mathcal{G}_b$, masking is simply

$$\widetilde{\mathbf{q}}^{\theta}_{b,p,c} = \begin{cases} \mathbf{m}^{\theta}, & (p,c) \in \mathcal{S}_b, \\ \mathbf{q}^{\theta}_{b,p,c}, & \text{otherwise.} \end{cases} \tag{7}$$

Channel and coordinate embeddings are added after this replacement, so the encoder retains position identity but not its EEG content. The final Context state is $\mathbf{h} = \mathbf{H}_{\theta}^{(N_{\mathrm{blk}})}(\mathbf{X}; \mathcal{S})$. A shared position-wise predictor gives $\mathbf{p} = g_{\psi}(\mathbf{h})$ in the Target latent space. Writing nrm for last-dimension ℓ_2 normalization, the loss compares

$$\widehat{\mathbf{r}}_{b,p,c} = \mathrm{nrm}(\mathbf{p}_{b,p,c}), \qquad \overline{\mathbf{r}}_{b,p,c} = \mathrm{nrm}(\mathrm{sg}(\mathbf{r}_{b,p,c})), \tag{8}$$

where sg denotes stop-gradient. The same predictor is applied at every grid position.

2.4 MASKED LATENT PREDICTION

The prediction term is the mean cosine distance between the normalized prediction and the stopped, normalized Target state:

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{\sum_b |\mathcal{M}_b|} \sum_b \sum_{(p,c) \in \mathcal{M}_b} \left(1 - \widehat{\mathbf{r}}^{\top}_{b,p,c} \overline{\mathbf{r}}_{b,p,c}\right). \tag{9}$$

Because the sum ranges only over $\mathcal{M}_b$, the objective asks the Context Encoder to infer target states from EEG evidence outside both the target and its guard. Each comparison remains position matched: the prediction at (p, c) is paired with the residualized Target state from that same token position.

2.5 TRAINING OBJECTIVE AND PARAMETER UPDATES

We accompany latent prediction with a random-projection moment regularizer inspired by SIGReg (Balestriero & LeCun, 2025). It uses a separate, unmasked Context pass $\mathbf{h}^{\mathrm{full}} = \mathbf{H}_{\theta}^{(N_{\mathrm{blk}})}(\mathbf{X}; \varnothing)$. We flatten its token states and project them onto K random unit directions $\mathbf{v}_k$. For $s_{i,k} = (\mathbf{h}^{\mathrm{full}}_i)^{\top} \mathbf{v}_k$, let μ_k and ν_k denote the empirical mean and variance over token index i. Our moment-matching variant encourages zero mean and unit variance:

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{K} \sum_{k=1}^{K} \left[\mu_k^2 + (\nu_k - 1)^2\right]. \tag{10}$$

Unlike the distributional goodness-of-fit test used in SIGReg, this variant matches only the first two projected moments; it does not enforce a Gaussian distribution. It acts on full-input Context features, not Target latents, decoupling regularization from the sampled mask geometry.

The complete objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{reg}}. \tag{11}$$

Gradient updates optimize θ and ψ; the subsequent EMA update in Eq. (5) updates ϕ.

Table 1: Overview of 16 selected NeuralBench-EEG-Core datasets and 14 EEG-FM-Bench datasets (Banville et al., 2026; Xiong et al., 2025). Dataset–task pairs shared by the two benchmarks are excluded from the NeuralBench-EEG-Core selection. Sampling follows each benchmark's dataset-specific preprocessing; NSP uses 200 Hz for EEG-FM-Bench. Benchmark details and complete results are provided in Appendix E.

Task	Dataset	Channels	Window (s)	Samples	Subjects	Classes	Task	Dataset	Channels	Window (s)	Samples	Subjects	Classes
Alzheimer's identification	ADFTD	19	10	7,013	88	3	Motor imagery	Mimul-11	60	5	41,347	11	3
Depression diagnosis	Mumtaz2018	19	5	14,683	64	2		PhysioMI	64	4	9,747	109	4
Parkinson's diagnosis	Singh2021	60	5	39,543	129	2		BCIC-2a	22	4	5,088	9	4
Schizophrenia diagnosis	Albrecht2019	62	5	24,928	77	2		Scherer2015	30	4	3,550	9	5
Abnormal classification	TUAB	23	30	272,320	2,383	2	Motor execution	Srisrisawang2024	60	3	19,157	20	16
Anomalous event detection	TUEV	21	5	113,353	370	6	Visual target detection	Things-EEG-2	63	5	41,570	10	2
Slowing event classification	TUSL	21/22	10	290	28	3	c-VEP visual coding	Thielen2021	8	4	45,000	30	20
Seizure detection	Siena	29	10	50,830	14	2	SSVEP	Wang2017	64	4	8,160	34	40
	CHB-MIT	24	5	707,683	23	2	Error-related negativity	ERP CORE-ERN	30	1	15,972	40	2
Sleep staging	HMC	4	30	136,925	151	5	Face-perception response	ERP CORE-N170	30	1	12,800	40	2
Emotion recognition	SEED	60	10	37,890	15	3	Semantic-violation response	ERP CORE-N400	30	1	4,800	40	2
	SEED-V	60	10	12,318	16	5	Auditory P300	Schreuder2010	60	1	253,800	21	2
	SEED-VII	60	15	19,420	20	7	Mental arithmetic	Zyma2019	20	5	1,659	35	2
Mental workload/stress	Workload	19	10	2,134	36	2	Sentence retrieval	ZuCo	128	3	25,819	12	retr.
	Hinss2023	63	5	15,399	29	3	Word retrieval	Nieuwland2018	74	3	529,730	222	retr.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Pretraining data and preprocessing. We pretrain NSP on the Temple University Hospital EEG Corpus (TUEG), a heterogeneous clinical archive collected from 2002 to 2017 (Obeid & Picone, 2016; Wang et al., 2025). It contains 69,652 EEG files from 14,987 subjects and 26,846 sessions, recorded with variable montages and sampling rates. We map recordings to 19 standard 10–20 channels, resample to 200 Hz, apply 0.5–40 Hz band-pass and 50/60 Hz notch filtering, use common-average referencing, clip at ± 500 μV, and perform recording-level z-normalization. Non-overlapping 6-s windows undergo quality checks after normalization. Downstream preprocessing is benchmark specific: NeuralBench-EEG-Core follows each task's released YAML configuration, while EEG-FM-Bench follows its released event/channel selection, filtering, windowing, montage, and unit-conversion pipeline; NSP uses the task-specific input adapters; the 19-channel, 200-Hz setting specifies pretraining rather than a universal downstream input requirement.

Downstream tasks and metrics. We use NeuralBench-EEG-Core v1.0 (Banville et al., 2026) and EEG-FM-Bench (Xiong et al., 2025) (Table 1). After removing dataset–task pairs shared with EEG-FM-Bench, NeuralBench contributes 16 datasets: four clinical, four BCI/sensorimotor, four event-related, and four cognitive/language tasks. Fourteen classification tasks use balanced accuracy (BAC); the ZuCo and Nieuwland2018 retrieval tasks use per-subject Top-5 accuracy and are reported separately. EEG-FM-Bench contains 14 datasets from 10 paradigms and evaluates single-task and multi-task full fine-tuning, frozen-backbone multi-task transfer, and multi-task low-rank adaptation (LoRA).

Baselines and downstream protocols. The benchmarks use different comparison sets. NeuralBench evaluates BENDR, BIOT, LaBraM, CBraMod, LUNA, and REVE (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024b; Wang et al., 2025; Döner et al., 2025; El Ouahidi et al., 2025) with its unified end-to-end recipe; NSP uses seeds 33–35. EEG-FM-Bench compares BENDR, BIOT, LaBraM, EEGPT, CBraMod, CSBrain, and REVE (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024b; Wang et al., 2024; 2025; Zhou et al., 2025; El Ouahidi et al., 2025) under its four released protocols (Xiong et al., 2025). Its 30-epoch AdamW adaptation uses a 3-epoch warmup, batch size 128, gradient clipping at 1.0, weight decay 0.01, and a one-epoch encoder freeze. Multi-task full fine-tuning uses seeds 42–46; other protocols use seeds 42–44. We report dataset means with sample standard deviation (SD) and average dataset means for each macro.

Pretraining setting. The 42.83M-parameter Large model has eight CST blocks, width 512, eight heads, FFN width 2,048, 0.2-s patches, and a predictor with hidden width 768 and output width 512. Mixed-precision distributed data-parallel training runs on eight NVIDIA H100 GPUs with global batch size 384. AdamW uses learning rate 10^{-4}, weight decay 0.05, gradient clipping at 1.0, 5,000 warmup steps, and cosine decay over 50,000 steps. Component studies hold data, architecture, optimization, and transfer protocol fixed; duration and capacity studies vary only the named factor. Appendix Tables 4–5 provide the complete configuration.

Table 2: **Primary-metric performance on 16 selected NeuralBench-EEG-Core datasets (Banville et al., 2026).** BAC (%) is shown for 14 classification tasks and Top-5 accuracy (%) for two retrieval tasks. Dataset-level values are mean ± sample SD; the macro row averages the 14 BAC means. Red and blue mark the best and second-best mean for each task and for the macro row.

Dataset	Metric	BENDR	BIOT	LaBraM	CBraMod	LUNA	REVE	NSP
Clinical and neurological disease								
Mumtaz2018	BAC	86.22±00.47	79.11±03.39	85.61±03.43	86.04±02.48	85.42±00.98	**87.33**±01.44	**90.43**±06.40
Singh2021	BAC	65.82±05.62	74.18±04.16	**76.64**±04.06	74.37±02.38	64.95±00.49	72.58±01.72	**76.67**±01.70
Albrecht2019	BAC	57.17±00.87	56.18±01.50	63.62±01.93	57.05±03.12	**66.75**±02.18	49.21±10.36	**68.43**±02.71
CHB-MIT	BAC	78.97±01.04	73.58±05.20	**82.09**±00.75	**82.89**±01.58	70.46±02.44	76.51±03.73	81.10±01.39
BCI and sensorimotor decoding								
Thielen2021	BAC	4.96±00.24	4.80±00.25	86.39±00.74	86.37±01.33	**87.06**±00.28	80.51±01.81	**88.83**±03.37
Scherer2015	BAC	21.62±00.66	**26.48**±02.48	21.83±00.17	21.22±01.80	20.95±01.09	26.06±01.07	**32.00**±00.55
Srisrisawang2024	BAC	34.56±00.96	18.00±01.78	52.94±00.88	49.41±00.99	49.92±00.72	**57.90**±01.64	**54.40**±00.95
Wang2017	BAC	13.31±05.51	58.67±01.85	**96.31**±00.45	85.85±07.27	13.08±00.88	65.79±02.09	**96.83**±00.91
Event-related and evoked responses								
ERP CORE-ERN	BAC	78.06±04.59	49.40±01.12	**81.96**±01.52	81.32±01.04	81.57±03.26	**84.61**±01.92	80.07±02.94
ERP CORE-N170	BAC	67.03±00.61	49.84±00.36	72.14±00.55	71.14±00.55	68.74±02.04	**74.03**±05.33	**73.27**±00.91
ERP CORE-N400	BAC	60.45±01.15	50.07±00.64	**64.62**±01.09	60.42±01.13	64.20±01.55	64.51±01.97	**66.67**±00.75
Schreuder2010	BAC	**66.70**±00.16	50.25±00.21	66.05±00.40	63.11±00.62	62.97±00.39	64.70±01.00	**69.87**±00.67
Cognitive state and language decoding								
Zyma2019	BAC	65.28±02.34	66.40±03.04	68.92±02.42	72.02±01.30	72.35±00.41	**72.88**±02.81	**72.50**±05.04
Hinss2023	BAC	45.89±00.19	43.18±04.73	65.46±03.17	61.37±01.70	**66.01**±00.94	65.57±04.49	**72.93**±00.42
ZuCo	Top-5	6.47±01.29	7.67±00.39	8.95±00.26	2.25±00.30	11.32±00.56	**11.39**±00.40	**12.11**±00.39
Nieuwland2018	Top-5	20.73±01.69	5.03±00.09	28.34±01.27	29.73±00.66	27.99±01.37	**33.73**±00.68	**36.27**±00.12
Macro (14 BAC tasks)		53.29	50.01	**70.33**	68.04	62.46	67.30	**73.14**

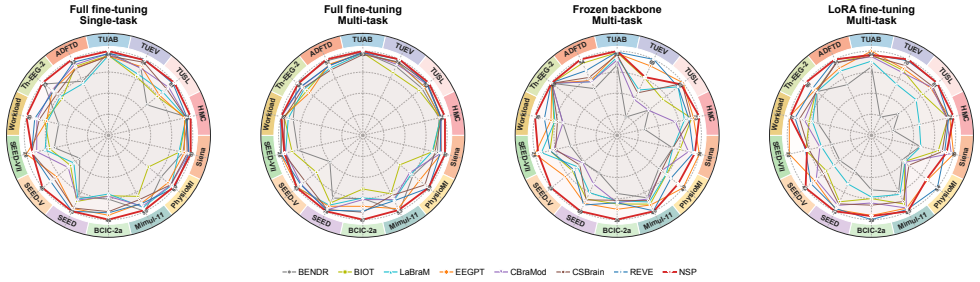

Figure 4: **EEG-FM-Bench radar summary.** Dataset-wise BAC under four adaptation protocols across all 14 datasets; radial positions use per-dataset relative scaling and NSP is highlighted in red.

3.2 Performance Comparison with Baselines

To compare NSP with established EEG foundation models across domains and adaptation settings, we evaluate it on 16 selected NeuralBench-EEG-Core datasets and the 14-dataset EEG-FM-Bench suite. Table 2 reports the NeuralBench-EEG-Core results under its unified end-to-end protocol, while Figure 4 summarizes single-task and multi-task full fine-tuning, plus multi-task frozen and LoRA transfer on EEG-FM-Bench.

On the 14 NeuralBench classification tasks, NSP reaches 73.14 macro BAC, ahead of LaBraM (70.33), CBraMod (68.04), and REVE (67.30). It has the highest mean on nine tasks and ranks among the top two on twelve, spanning clinical, sensorimotor, event-related, and cognitive settings. Representative wins include Mumtaz2018, Singh2021, Thielen2021, Scherer2015, Wang2017, ERP CORE-N400, and Hinss2023. The margins range from near ties on Singh2021 and Wang2017 to 5.52 percentage points on Scherer2015 and 6.92 on Hinss2023, so the macro advantage aggregates both modest and substantial task-level gains. NSP also reaches 12.11 ± 0.39 and 36.27 ± 0.12 Top-5 accuracy on ZuCo and Nieuwland2018, respectively; these retrieval scores are not pooled with BAC.

EEG-FM-Bench tests NSP under several adaptation protocols. Under full-parameter multi-task fine-tuning, NSP averages 63.94 macro BAC, 2.35 percentage points above the strongest evaluated baseline. Gains are largest on several clinical and BCI datasets, including TUSL, HMC, Siena, and BCIC-2a, while emotion and motor-imagery results are more mixed; frozen and LoRA transfer are

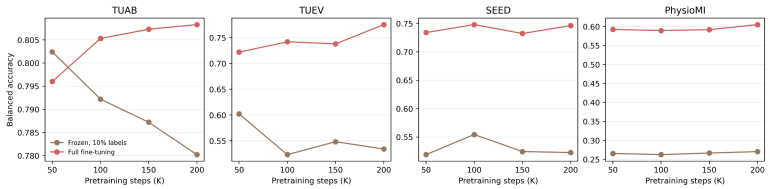

Figure 5: **Task-wise transfer across pretraining duration.** BAC (0–1 scale) across checkpoints under frozen single-task transfer (10% labels) and single-task full fine-tuning.

likewise less uniform than full fine-tuning. TUEV is a near tie under multi-task adaptation, and results on SEED-VII, PhysioMI, and Mimul-11 vary by task. NeuralBench measures single-task transfer across domains, whereas EEG-FM-Bench measures performance across adaptation strategies. Together, they characterize transfer across both tasks and adaptation settings.

3.3 Component and Input/Context Dependence Analysis

Table 3: **Component ablations: frozen single-task transfer with 10% labels.** ResID: identity residualization; Reg.: state regularization. Block mask "–" denotes temporal masking, not unmasked input. Exclusion means additional context removal; Random is size matched. BAC is reported as % (mean ± sample SD); Macro averages four tasks. **Red** and **blue** mark the best and second-best mean per column.

Pretraining design						Frozen Single-Task BAC				
Variant	**Block mask**	**Target**	**ResID**	**Exclusion**	**Reg.**	**TUAB**	**TUEV**	**SEED**	**PhysioMI**	**Macro**
Target construction										
Shared-target baseline	–	Shared	–	None	✓	73.27±00.35	38.37±02.41	40.53±01.78	25.57±01.42	44.44
+ Block masking	✓	Shared	–	None	✓	74.27±00.23	44.17±01.00	38.37±04.12	26.20±00.78	45.75
+ EMA target	✓	EMA	–	None	✓	78.17±00.23	51.40±00.98	**51.97**±00.32	25.97±01.12	51.88
+ Identity residualization	✓	EMA	✓	None	✓	78.73±00.15	59.77±02.17	**51.97**±01.27	**26.50**±00.35	**54.24**
Representation-stability control										
Full NSP w/o regularization	✓	EMA	✓	Topology	–	**80.83**±00.29	48.27±05.12	33.30±00.00	25.13±01.01	46.88
Matched residualization–context comparison										
Equal-size random exclusion	✓	EMA	–	Random	✓	78.92±00.48	58.31±01.54	49.87±01.12	25.38±00.96	53.12
NSP with random exclusion	✓	EMA	✓	Random	✓	79.36±00.51	59.18±01.47	50.94±01.06	25.47±00.93	53.74
NSP w/o identity residualization	✓	EMA	–	Topology	✓	79.71±00.49	**59.86**±01.55	51.58±00.97	25.55±00.98	54.18
Full NSP	✓	EMA	✓	Topology	✓	**80.13**±00.55	**60.70**±01.61	**53.33**±01.00	**28.63**±01.37	**55.70**

Target construction and representation stability. To identify target-side contributions under limited labels, we evaluate frozen single-task transfer with 10% labels on TUAB, TUEV, SEED, and PhysioMI (Table 3). Starting from block masking at 45.75 macro BAC, an EMA Target Encoder raises the score to 51.88, and adding identity residualization further improves it to 54.24. These cumulative changes support both stable latent supervision and removal of additive positional effects. Removing state regularization from Full NSP lowers macro BAC from 55.70 to 46.88, indicating that target and context design do not replace representation-stability control.

Identity residualization and context exclusion. To separate their contributions, matched rows fix the EMA target, regularization, architecture, and visible-token budget. Residualization adds 0.62 percentage points with random exclusion and 1.52 with topology-separated exclusion. Topology separation adds 1.06 percentage points without residualization and 1.96 with it. Both components improve the matched means, more so together.

Signal-content and identity diagnostics. To distinguish signal sensitivity from positional information, we zero inputs, cyclically mismatch recordings within a batch, or expand Context masks while holding Targets fixed (Figure 17). Intervention cosine is normalized by the target-only-masked Context reference. Separate probes show that channel identity remains strongly decodable from Context states and targets without channel residualization. These component diagnostics characterize sensitivity and residual identity information, not channel invariance.

3.4 Training Duration, Capacity, and Label Efficiency

To determine how duration interacts with adaptation, we evaluate frozen and full-parameter single-task transfer at multiple checkpoints on four tasks (Figure 5). Frozen transfer is generally strongest

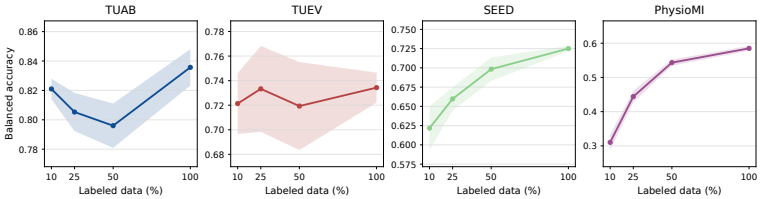

Figure 6: **Label efficiency across downstream tasks.** BAC (0–1 scale) versus labeled-data fraction on TUAB, TUEV, SEED, and PhysioMI; bands show variability across evaluations.

at the shorter budget, whereas full fine-tuning is maintained or improves on several tasks. We also pretrain Tiny and Medium NSP variants and compare them with Large under single-task full fine-tuning (Figure 15). Macro BAC increases with capacity, but the preferred scale varies by task; this comparison is separate from the fixed-architecture component ablation.

To test NSP with limited downstream labels, we fine-tune it with 10%, 25%, 50%, and 100% of the available labels (Figure 6). TUAB and TUEV change by less than 1.5 percentage points between 10% and 100% labels, whereas SEED improves from 62.17 to 72.50 and PhysioMI from 31.03 to 58.50. The TUAB and TUEV curves remain comparatively flat across label fractions.

3.5 Representation geometry and task-conditioned spatial evidence

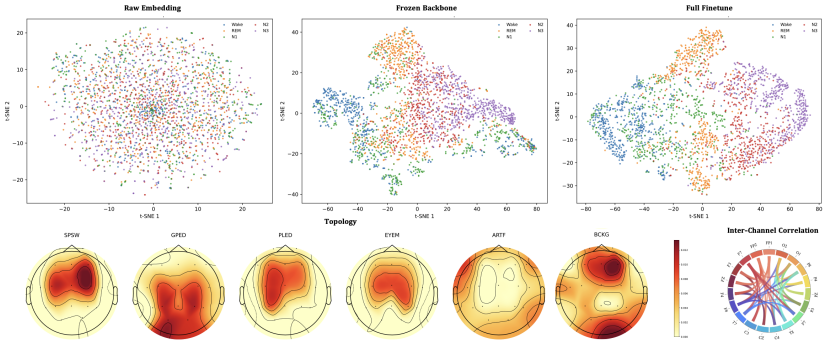

Figure 7: **Representation geometry and spatial evidence.** Top: raw, frozen, and fine-tuned HMC sleep-stage representations visualized by t-SNE. Bottom: TUEV event-class Grad-CAM maps and interchannel correlations. The rows depict different datasets.

To examine geometry and spatial evidence, we visualize HMC embeddings with t-SNE (van der Maaten & Hinton, 2008) and TUEV attribution with Grad-CAM (Selvaraju et al., 2017) (Figure 7). HMC representations show more sleep-stage structure after pretraining and adaptation, although classes still overlap. TUEV attribution patterns differ across SPSW, GPED, PLED, EYEM, ARTF, and BCKG; chords depict selected interchannel correlations.

4 Conclusion

NSP couples EMA-based latent prediction with identity-residualized targets and topology-separated context to reduce additive positional effects and encourage distributed inference. Transfer across 30 datasets and component analyses support jointly designing targets and context for EEG representation learning. Future work could adapt exclusions to signal-dependent correlations. Despite broad benchmark coverage, responsible use requires prospective validation across populations and acquisition settings, together with safeguards for neural-data privacy.

AI USE DISCLOSURE

AI-assisted tools, including OpenAI Codex, were used to support language editing, literature search, code development and debugging, experiment scripting and monitoring, and figure preparation. The authors remain responsible for validating the code, experimental results, and references, and for all scientific claims, interpretations, and final content of this paper.

REFERENCES

Matthew A Albrecht, James A Waltz, James F Cavanagh, Michael J Frank, and James M Gold. Increased conflict-induced slowing, but no differences in conflict-induced positive or negative prediction error learning in patients with schizophrenia. *Neuropsychologia*, 123:131–140, 2019.

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):e0256111, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. doi: 10.48550/arXiv.2506.09985.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pp. 1298–1312. PMLR, 2022.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Elisa Cascardi, et al. Neuralbench: A unifying framework to benchmark neuroai models. *arXiv preprint arXiv:2605.08495*, 2026.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Paolo Detti, Giampaolo Vatti, and G Zabalo Manrique de Lara. EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7), 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

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://openreview.net/forum?id=uazfjnFL0G`.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918, 2023.

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. *Advances in Neural Information Processing Systems*, 38:22541–22577, 2025.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-jepa: towards seamless cross-dataset transfer through dynamic spatial attention. 2024.

Amir Harati, Meysam Golmohammadi, Silvia Lopez, Iyad Obeid, and Joseph Picone. Improved EEG event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4. IEEE, 2015.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Marcel F Hinss, Emilie S Jahanpour, Bertille Somon, Lou Pluchon, Frédéric Dehais, and Raphaëlle N Roy. Open multisession and multi-task EEG cognitive dataset for passive brain-computer interface applications. *Scientific Data*, 10(1), 2023.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1), 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ji-Hoon Jeong, Jeong-Hyun Cho, Kyung-Hwan Shim, Byoung-Hee Kwon, Byeong-Hoo Lee, Do-Yeon Lee, Dae-Hyeok Lee, and Seong-Whan Lee. Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions. *GigaScience*, 9 (10):giaa098, 2020.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. SEED-VII: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 2024a.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024b.

D Jonathan and A Shoeb. Bids chb-mit scalp eeg database, 2023.

Emily S Kappenman, Jaclyn L Farrens, Wendy Zhang, Andrew X Stewart, and Steven J Luck. ERP CORE: An open resource for human event-related potential research. *NeuroImage*, 225:117465, 2021.

Demetres Kostas, Stephane 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.

Elnaz Lashgari, Dehua Liang, and Uri Maoz. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346:108885, 2020. doi: 10.1016/j.jneumeth.2020.108885.

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.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Jinhao Li, Zhiyuan Ma, Xueqiao Han, Zhongye Xia, Xinche Zhang, Shanghong Xie, Yixuan Liu, Yongjian Li, Runmin Gan, Tianlin Huo, and Sen Song. EEG-JEPA: Structured latent prediction for EEG foundation models, 2026.

Wei Lun Lim, Olga Sourina, and Lipo Wang. STEW: Simultaneous task EEG workload data set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2106–2114, 2018. doi: 10.1109/TNSRE.2018.2872924.

Chenyu Liu, Yuqiu Deng, Tianyu Liu, Jinan Zhou, Xinliang Zhou, Ziyu Jia, and Yi Ding. ECHO: Toward contextual seq2seq paradigms in large EEG models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=ClLQ6cLkoR`.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022.

Sebas Lopez, G Suarez, D Jungreis, I Obeid, and Joseph Picone. Automated identification of abnormal adult EEGs. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2017.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. *arXiv preprint arXiv:2506.09110*, 2025.

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

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5544–5555. ACM, 2024. doi: 10.1145/3637528.3671600.

Wajid Mumtaz, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, and Aamir Saeed Malik. A machine learning framework involving eeg-based functional connectivity to diagnose major depressive disorder (mdd). *Medical & biological engineering & computing*, 56(2):233–246, 2018.

Mante S Nieuwland, Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, Sarah Von Grebmer Zu Wolfsthurn, Federica Bartolozzi, Vita Kogan, Aine Ito, et al. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, 7:e33468, 2018.

Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196.

Saarang Panchavati, Uddhav Panchavati, Hiroki Nariai, Corey Arnold, and William Speier. Laya: A LeJEPA approach to EEG via latent prediction over reconstruction. *arXiv preprint arXiv:2603.16281*, 2026.

L Paul. *Electric fields of the brain*. CRC Pres, 2005.

Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on EEG: A systematic comparison. *Journal of Neural Engineering*, 19(6):066020, 2022. doi: 10.1088/1741-2552/aca220.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Reinhold Scherer, Josef Faller, Elisabeth VC Friedrich, Eloy Opisso, Ursula Costa, Andrea Kübler, and Gernot R Müller-Putz. Individually adapted imagery improves brain-computer interface performance in end-users with disability. *PLOS ONE*, 10(5):e0123727, 2015.

Martijn Schreuder, Benjamin Blankertz, and Michael Tangermann. A new auditory multi-class brain-computer interface paradigm: Spatial hearing as an informative cue. *PLOS ONE*, 5(4): e9813, 2010.

Roy Segal, Yoni Svechinsky, and Tomer Fekete. STST-JEPA: Shallow-target spatio-temporal joint embedding prediction architecture for EEG self-supervised learning, 2026.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Arun Singh, Rachel C Cole, Arturo I Espinoza, Aron Evans, Scarlett Cao, James F Cavanagh, and Nandakumar S Narayanan. Timing variability and midfrontal 4 Hz rhythms correlate with cognition in parkinson's disease. *npj Parkinson's Disease*, 7(1):14, 2021.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Nitikorn Srisrisawang and Gernot R Müller-Putz. Simultaneous encoding of speed, distance, and direction in discrete reaching: an EEG study. *Journal of Neural Engineering*, 21(6):066042, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

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.

Jordy Thielen, Pieter Marsman, Jason Farquhar, and Peter Desain. From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface. *Journal of Neural Engineering*, 18, 2021.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.12477.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Eva von Weltin, Tameem Ahsan, Vinit Shah, Dawer Jamshed, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Electroencephalographic slowing: A primary source of error in automatic seizure detection. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2017.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of EEG signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

Haoning Wang, Wenchao Yang, Shuai Shen, and Yang Li. Kast-bar: Knowledge-anchored semantically-dynamic topology brain autoregressive modeling for universal neural interpretation. 2026.

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 *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for ssvep-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2016.

Wei Xiong, Jiangtong Li, Jie Li, and Kun Zhu. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation of eeg foundation models. *arXiv preprint arXiv:2508.17742*, 2025.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Wenchao Yang, Weidong Yan, Wenkang Liu, Yulan Ma, and Yang Li. THD-BAR: Topology hierarchical derived brain autoregressive modeling for EEG generic representations. In *Advances in Neural Information Processing Systems*, 2025.

Kieren Yu, LIU Ziyang, Chang Huang, and Kaishun Wu. Patchcode: Discrete latent predictive learning for eeg foundation model. In *Forty-third International Conference on Machine Learning*, 2026.

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.

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*, 2025.

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

A Related Work

A.1 EEG Foundation Models

EEG foundation models pretrain reusable encoders on heterogeneous, largely unlabeled recordings and adapt them across clinical, BCI, affective, and cognitive tasks. Reconstruction-based models differ in what they ask the encoder to preserve. EEGPT (Wang et al., 2024) and REVE (El Ouahidi et al., 2025) recover masked waveform content, with REVE additionally encoding electrode geometry to support variable layouts. CBraMod (Wang et al., 2025) reconstructs masked EEG patches through separate spatial and temporal attention, whereas CSBrain (Zhou et al., 2025) reconstructs spectrogram representations with specialized spatial–temporal modeling. LaBraM (Jiang et al., 2024b) first converts EEG patches into neural tokens and predicts the masked token identities. DeWave (Duan et al., 2023) and CodeBrain (Ma et al., 2025) likewise introduce discrete codes that replace direct waveform regression with code prediction. These choices determine whether supervision emphasizes sample-level detail, spectral structure, patch content, or a learned vocabulary.

Contrastive EEG foundation models specify transfer through invariance between related views rather than reconstruction. BENDR (Kostas et al., 2021) masks the convolutional sequence and trains a Transformer context representation to identify its corresponding latent target among within-recording negatives. BIOT (Yang et al., 2023) tokenizes channels into a unified biosignal sequence and aligns an original recording with a view perturbed by channel and token dropping. Contrastive objectives can suppress nuisance variation, but the learned invariance depends on positive-pair construction, negative sampling, and augmentations. This is consequential for EEG because the validity of a transformation is task dependent, and an augmentation that is benign for one paradigm may alter task-relevant temporal or spatial structure in another (Lashgari et al., 2020; Rommel et al., 2022).

Autoregressive EEG foundation models instead factor a recording into an ordered prediction problem. ECHO (Liu et al., 2026) uses contextual sequence-to-sequence modeling, THD-BAR (Yang et al., 2025) organizes autoregression with topology-hierarchical structure, and KAST-BAR (Wang et al., 2026) introduces knowledge-anchored semantic dynamics. This family can model long temporal dependencies and supports generative or next-token objectives, but its evidence path is inherently directional. In strongly autocorrelated EEG, next-patch prediction may therefore reward accurate local extrapolation without requiring the representation to integrate spatially distributed evidence.

A.2 Latent Prediction and JEPA

Joint-Embedding Predictive Architectures (JEPA) learn by predicting representations of unobserved regions in a learned embedding space rather than reconstructing the input (LeCun et al., 2022). I-JEPA established masked latent prediction for images (Assran et al., 2023), V-JEPA extended feature prediction to video (Bardes et al., 2024), and data2vec developed a related target-representation objective across speech, vision, and language (Baevski et al., 2022). LeJEPA subsequently provided a theoretically grounded and scalable formulation with an explicit representation regularizer (Balestriero & LeCun, 2025; Van Assel et al., 2025). The same predictive-state principle now reaches world modeling: V-JEPA 2 couples large-scale latent video pretraining with an action-conditioned predictive model for physical-world planning (Assran et al., 2025).

JEPA has also been applied to brain signals. Brain-JEPA predicts masked fMRI representations with functional-coordinate encoding and spatiotemporal masking (Dong et al., 2024). For EEG, S-JEPA introduces spatial block masking for cross-dataset transfer (Guetschel et al., 2024); Laya applies latent prediction at larger scale (Panchavati et al., 2026); STST-JEPA combines shallow EMA targets with an auxiliary reconstruction objective (Segal et al., 2026); and EEG-JEPA structures latent targets by content, support, and encoder depth (Li et al., 2026). PATCHCODE (Yu et al., 2026) combines continuous patch-level latent prediction with auxiliary discrete-code supervision from a frozen tokenizer; we therefore discuss it within the latent-predictive family. These methods establish latent prediction for brain signals, but they do not combine target residualization against channel/time identity with a context guard that excludes the target's immediate spatial and temporal neighborhood.

B Implementation Details

B.1 Architecture

The NSP Large configuration has 42.83M trainable Context-side parameters, including its predictor. The Cascaded Spatial–Temporal Transformer (CST) contains 8 blocks, each with spatial attention ($d_{\text{model}} = 512$, 8 heads) across channels, temporal attention with rotary position encoding (Su et al., 2024) across patches, and a SwiGLU (Swish-gated linear unit) feed-forward sublayer with expansion factor 4. Patch duration is $0.2\,\text{s}$ at $f_s = 200\,\text{Hz}$. The predictor is a two-layer multilayer perceptron (MLP) with hidden width 768, output width 512, and Gaussian error linear unit (GELU) activation, followed by LayerNorm and ℓ_2 normalization. The Target Encoder aggregates the last three blocks ($\mathcal{I}_T = \{5, 6, 7\}$, zero-indexed). The EMA coefficient follows a cosine schedule from 0.996 to 0.9999 over the pretraining duration.

B.2 Identity residualization as an additive projection

The residualization in Eq. (4) has an exact least-squares interpretation. All expressions below act independently on each latent feature, and the vector notation collects those feature-wise problems. Because $\mathbf{u}$ is standardized to have zero global mean over the complete $B \times P \times C$ grid, consider the additive subspace formed by a channel term $\mathbf{a}_c \in \mathbb{R}^d$ and a relative-time term $\mathbf{t}_p \in \mathbb{R}^d$. We identify their zero levels through the constraints $\sum_c \mathbf{a}_c = \mathbf{0}$ and $\sum_p \mathbf{t}_p = \mathbf{0}$, and fit the additive component by

$$(\mathbf{a}^\star, \mathbf{t}^\star) = \underset{\substack{\{\mathbf{a}_c\},\{\mathbf{t}_p\} \\ \sum_c \mathbf{a}_c = \mathbf{0},\ \sum_p \mathbf{t}_p = \mathbf{0}}}{\arg\min} \sum_{b=1}^{B} \sum_{p=1}^{P} \sum_{c=1}^{C} \|\mathbf{u}_{b,p,c} - \mathbf{a}_c - \mathbf{t}_p\|_2^2 . \tag{12}$$

Under uniform weighting and a complete patch–channel grid, the normal equations give

$$\mathbf{a}_c^\star = \frac{1}{BP} \sum_{b,p} \mathbf{u}_{b,p,c} = \boldsymbol{\mu}_c^{\text{ch}}, \qquad \mathbf{t}_p^\star = \frac{1}{BC} \sum_{b,c} \mathbf{u}_{b,p,c} = \boldsymbol{\mu}_p^{\text{time}}. \tag{13}$$

Thus, $\mathbf{r} = \mathbf{u} - \mathbf{a}^\star - \mathbf{t}^\star$ is the orthogonal residual after projecting $\mathbf{u}$ onto this additive identity subspace. In particular, its channel and relative-time marginals vanish:

$$\sum_{b,p} \mathbf{r}_{b,p,c} = \mathbf{0} \quad \forall c, \qquad \sum_{b,c} \mathbf{r}_{b,p,c} = \mathbf{0} \quad \forall p. \tag{14}$$

Consequently, $\sum_{b,p,c} \mathbf{r}_{b,p,c}^\top (\mathbf{a}_c + \mathbf{t}_p) = 0$ for any additive channel–time function, which is the defining orthogonality condition of the projection. This result is deliberately limited: residualization removes the best additive channel and relative-time main effects under the current batch weighting, but it does not remove nonlinear identity information or channel–time interactions, and it does not imply channel invariance.

B.3 Pretraining hyperparameters

Pretraining uses AdamW (Loshchilov & Hutter, 2017) with learning rate 1×10^{-4}, weight decay 0.05, global batch size 384, gradient clipping 1.0, and 5,000 linear-warmup steps followed by cosine annealing. Target blocks cover 50% of the token grid, with temporal spans of 2–10 patches and spatial spans of 3–7 coordinate-near channels. The spatial guard includes the two nearest neighbors by Euclidean distance, and the temporal guard includes adjacent patches. State regularization uses $K = 64$ random unit directions with $\lambda_{\text{reg}} = 0.1$.

TUEG contains 69,652 files from 14,987 subjects and 26,846 sessions recorded with heterogeneous clinical montages and sampling rates (Obeid & Picone, 2016). We map each usable recording to 19 standard electrodes from the international 10–20 system and resample it to 200 Hz. After filtering and common-average referencing, the implementation clips amplitudes to $\pm 500\ \mu$V before recording-level z-normalization, then forms non-overlapping 6-s segments and applies quality checks to the resulting normalized windows. Exceeding $500\ \mu$V is therefore not, by itself, a rule for rejecting a whole segment. The pretraining collection contains 2.2M examples of shape $19 \times 1{,}200$.

C NSP PRETRAINING DETAILS

C.1 OBJECTIVE JUSTIFICATION AND TRAINING ALGORITHM

Predicting latent targets avoids forcing the encoder to reproduce every high-variance detail of the EEG waveform, but it does not determine which information remains in the target. On scalp EEG, channel and relative-time identity provide stable position cues, while volume-conducted neighboring electrodes provide a short path for local interpolation. NSP therefore modifies both sides of the prediction problem: identity residualization changes *what* is predicted, topology-separated context changes *which evidence* is visible, and the state regularizer prevents the Context Encoder from satisfying the objective with a degenerate representation. The full mathematical construction is given in Section 2; Algorithm 1 records the executable order of operations. Here, TARGETCONSTRUCT denotes selected-layer aggregation, standardization, and identity residualization in Section 2; nrm is normalization to unit ℓ_2 norm along the feature dimension. We use $i = (b, p, c)$ for a flattened token index and $\mathcal{M} = \{(b, p, c) : (p, c) \in \mathcal{M}_b\}$, so $|\mathcal{M}| = \sum_b |\mathcal{M}_b|$. The labels $\mathbf{z}_i$, $\mathbf{r}_i$, $\mathbf{h}_i$, and $\mathbf{p}_i$ in Figure 3 use this convention. Here $f_\eta(\mathbf{X}; \mathcal{S}) = \mathbf{H}_\eta^{(N_{\mathrm{blk}})}(\mathbf{X}; \mathcal{S})$; Target construction instead uses the selected intermediate states of f_ϕ. Masking replaces signal-dependent patch content before positional embeddings are added.

Algorithm 1 Neural State Prediction pretraining.

Require: EEG minibatch $\mathbf{X}$; Context Encoder f_θ; Target Encoder f_ϕ; predictor g_ψ

$\{\mathcal{M}_b\}_{b=1}^{B} \leftarrow$ TOPOLOGYMASK($\mathbf{X}$)

$\mathcal{G}_b \leftarrow$ EXPANDGUARD($\mathcal{M}_b$); $\mathcal{S}_b \leftarrow \mathcal{M}_b \cup \mathcal{G}_b$ for each b

$\mathbf{h} \leftarrow f_\theta(\mathbf{X}; \mathcal{S})$ ▷ Mask patch content, not positional cues

$\mathbf{r} \leftarrow \mathrm{sg}\Big(\text{TARGETCONSTRUCT}(\{\mathbf{H}_\phi^{(\ell)}(\mathbf{X}; \varnothing)\}_{\ell \in \mathcal{I}_T})\Big)$

$\mathbf{p} \leftarrow g_\psi(\mathbf{h})$ ▷ Predictor on Context states only

$\mathcal{L}_{\mathrm{pred}} \leftarrow \dfrac{1}{|\mathcal{M}|} \displaystyle\sum_{i \in \mathcal{M}} [1 - \langle \mathrm{nrm}(\mathbf{p}_i), \mathrm{nrm}(\mathbf{r}_i) \rangle]$

$\mathbf{h}^{\mathrm{full}} \leftarrow f_\theta(\mathbf{X}; \varnothing)$ ▷ Shared Context parameters

$\mathcal{L}_{\mathrm{reg}} \leftarrow$ MOMENTREGULARIZER($\mathbf{h}^{\mathrm{full}}$) ▷ Eq. (10)

$\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{reg}}$

$(\theta, \psi) \leftarrow$ ADAMWSTEP($\theta, \psi, \nabla_{\theta,\psi} \mathcal{L}_{\mathrm{total}}$)

$\phi \leftarrow \tau_s \phi + (1 - \tau_s)\theta$ ▷ EMA after the Context update

C.2 HYPERPARAMETER CONFIGURATION

Table 4: Default architecture and pretraining configuration of NSP.

Hyperparameter	Configuration
Input & Tokenization	
Segment duration	6 s
Sampling rate	200 Hz
EEG channels	19
Patch duration	0.2 s
Context Encoder & Predictor	
Encoder depth	8
Hidden dimension	512
Attention heads	8
FFN dimension	2,048
Dropout	0.1
Temporal attention	Bidirectional
Predictor hidden dimension	768
Predictor output dimension	512
Target Encoder	
Target layers	$\{5, 6, 7\}$ (zero-based)
EMA momentum	$0.996 \rightarrow 0.9999$
EMA schedule	Cosine
Target Masking and Context Exclusion	
Target ratio	0.50
Temporal target span	2–10 patches (0.4–2.0 s)
Spatial target span	3–7 coordinate-near channels
Temporal guard radius	1 patch
Spatial guard	2 nearest channels per target channel
Moment-based State Regularization	
Projection directions	64
Regularization weight	$\lambda_{\text{reg}} = 0.1$
Model	
Trainable Context-side parameters	42.83M

Table 5: Pretraining and optimization hyperparameters of NSP.

Hyperparameter	Configuration
Pretraining corpus	TUEG (Obeid & Picone, 2016)
Pretraining examples	2.2M segmented examples
Optimizer	AdamW (Loshchilov & Hutter, 2017)
Learning rate	1×10^{-4}
Adam (β_1, β_2)	$(0.9, 0.999)$
Weight decay	0.05
Training steps	50,000
Warmup steps	5,000
Warmup initial factor	0.1
Learning-rate schedule	Linear warmup followed by cosine annealing
Batch size per GPU	48
Number of GPUs	8 NVIDIA H100 GPUs
Global batch size	384
Gradient accumulation	1
Gradient clipping	Global norm 1.0
Mixed precision	Enabled
Distributed backend	NVIDIA Collective Communications Library (NCCL)
Distributed sampling	Rank-sharded, shuffled, drop-last
Pretraining random seeds	45, 46, 47

C.3 Training Diagnostics

C.3.1 Diagnostic Experiments

We monitor optimization and representation health jointly. The diagnostic sequence compares a temporal shared-target predictor, a spatiotemporal shared-target predictor, an EMA Target Encoder, identity-residualized targets, Full NSP, and NSP without state regularization. These names describe the actual intervention and are used consistently throughout the paper. All configurations share the Large encoder, 50,000 optimization steps, and the same distributed input stream; differences in raw prediction loss are interpreted only when their target and input-mask definitions match.

The control without state regularization reaches a Context-state mean cosine of 0.9422 by step 4,000 and 0.9979 at the endpoint, while its effective rank falls to 113.6. Full NSP instead retains a rank of 470.9 with near-zero mean cosine. In this control, lower prediction loss coincides with highly aligned Context states and reduced effective rank. We therefore assess pretraining with downstream transfer and representation-health measures alongside loss.

Figure 8 reports the optimization trajectory of Full NSP across three independent pretraining runs. Total and prediction loss decrease rapidly during warmup and then continue to decline gradually, while the regularization term approaches zero. At the same time, Context effective rank rises and mean cosine remains near zero, so the lower objective is accompanied by a non-collapsed representation rather than constant states.

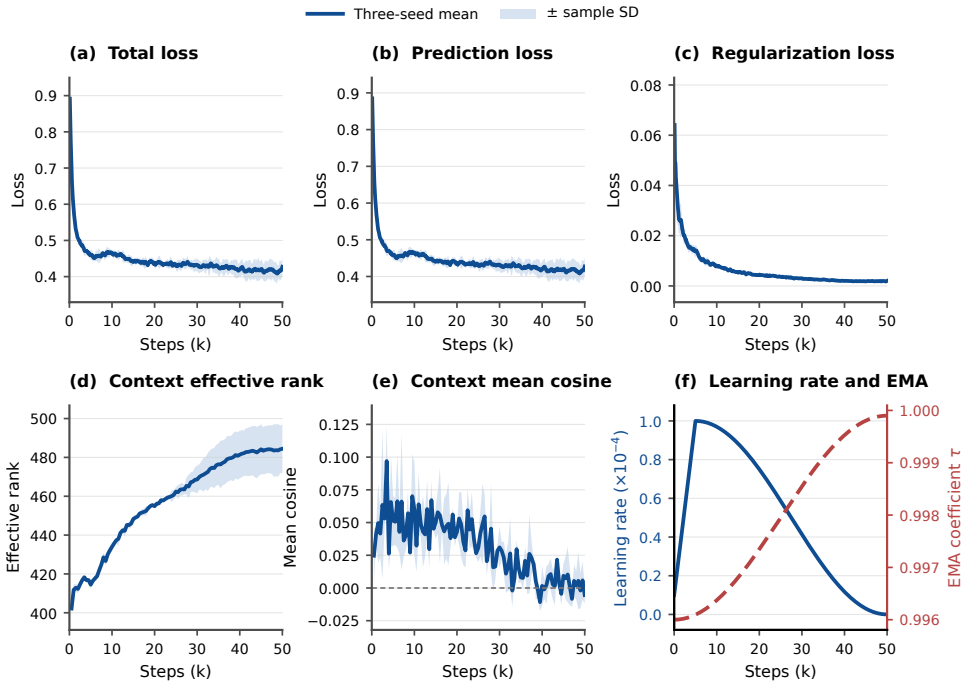

Figure 8: **NSP pretraining diagnostics.** Lines show the across-run mean and shaded regions show sample SD for optimization losses and representation-health statistics. The final panel displays the learning-rate and EMA schedules.

C.4 Downstream Benchmark Requirements

NeuralBench-EEG-Core (Banville et al., 2026) specifies task-level preprocessing, data splits, adaptation settings, and validation metrics. We evaluate 16 selected datasets with single-task adaptation: 14 classification tasks contribute to the BAC macro, while two language tasks use Top-5 retrieval accuracy and are reported separately. EEG-FM-Bench (Xiong et al., 2025) specifies preprocessing and model adapters for 14 datasets and compares single-task full fine-tuning, multi-task full fine-tuning, frozen-backbone transfer, and LoRA. We follow each benchmark's released task definitions and report its results within its own protocol.

C.5 PREPROCESSING PROTOCOLS

EEG-FM-Bench (Xiong et al., 2025): Event selection, filtering, resampling, windowing, channel mapping, and unit conversion follow the benchmark's task/model adapter. The 19-channel, 200-Hz specification above describes NSP pretraining, not a requirement that every downstream source recording have this montage or sampling rate. Downstream tasks can supply different channel counts through their configured adapters; the source montage and the representation presented to the pre-trained encoder must be distinguished.

NeuralBench-EEG-Core (Banville et al., 2026): Each task uses its specified YAML configuration for event selection, filtering, resampling, windowing, and channel mapping. Tasks are not forced to 200 Hz or 19 channels; configurations vary by dataset.

C.6 DOWNSTREAM ADAPTATION

EEG-FM-Bench adaptation: 30 epochs with 3-epoch linear warmup; AdamW (Loshchilov & Hutter, 2017); batch size 128; gradient clipping 1.0; weight decay 0.01; warmup-freeze encoder 1 epoch (allows classification head to stabilize). Learning rate and pooling head hidden dimension are searched per task. Encoder learning rate is set to classifier learning rate $\times$ (1/10, 1/5, or 1/2). Seeds 42–46 for multi-task full fine-tuning; seeds 42–44 for frozen, LoRA, and single-task protocols.

NeuralBench-EEG-Core adaptation: Single-task per dataset; task-specific YAML configurations for epochs, batch size, learning rate, and architecture; seeds 33, 34, 35; checkpoint selection on the task-specific validation metric; test evaluation only on the selected checkpoint.

D DOWNSTREAM EVALUATION DETAILS

D.1 EVALUATION METRICS

Unless stated otherwise, classification and retrieval accuracies, F1, Cohen's κ, and areas under curves are reported as percentages; differences between them are percentage points. Figures 5 and 6 use the equivalent 0–1 BAC scale. MedR is an unscaled median rank, and cosine-based diagnostics are dimensionless.

Let K_{cls} be the number of classes, n_k the support of class k, and N the number of evaluated examples. Balanced accuracy averages class recall,

$$\text{BAC} = \frac{1}{K_{\text{cls}}} \sum_{k=1}^{K_{\text{cls}}} \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}, \tag{15}$$

where TP_k and FN_k are true-positive and false-negative counts. BAC gives each class equal weight under imbalance. With class precision P_k and recall R_k, weighted F1 is

$$\text{F1}_{\text{weighted}} = \sum_{k=1}^{K_{\text{cls}}} \frac{n_k}{N} \frac{2P_k R_k}{P_k + R_k}. \tag{16}$$

Cohen's κ removes agreement expected from the empirical class marginals,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_e = \sum_{k=1}^{K_{\text{cls}}} \frac{n_k}{N} \frac{\hat{n}_k}{N}, \tag{17}$$

where p_o is observed agreement and $\hat{n}_k$ is the number of predictions assigned to class k.

For binary tasks, the area under the receiver-operating-characteristic curve and the area under the precision-recall curve are

$$\text{AUROC} = \int \text{TPR}\, d\text{FPR}, \qquad \text{AUCPR} = \int \text{Precision}\, d\text{Recall}. \tag{18}$$

Here TPR and FPR denote true-positive and false-positive rates. AUROC measures threshold-independent ranking over both classes, while AUCPR is particularly informative when the positive

class is rare. The two language retrieval tasks use

$$\text{Top-5} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[y_i \in \text{Top5}(\hat{\mathbf{p}}_i)], \tag{19}$$

which measures whether the correct item appears among the five highest-scoring candidates and is not numerically comparable to BAC.

D.2 Baseline Models

We compare NSP against six official NeuralBench-EEG-Core foundation-model baselines (Table 2) under the released benchmark protocol (Banville et al., 2026). EEG-FM-Bench uses a distinct comparison set that replaces LUNA with EEGPT and CSBrain (Xiong et al., 2025). All dataset-level entries report a mean and sample standard deviation using the corresponding benchmark's task-specific metric definitions.

BENDR (Kostas et al., 2021) adapts wav2vec-style masked contrastive prediction to EEG. At a masked time step, the Transformer context is trained to identify the corresponding unmasked convolutional representation among within-sequence negatives.

BIOT (Yang et al., 2023) tokenizes each biosignal channel into fixed-length segments and rearranges them into a unified token sequence, allowing one encoder to process heterogeneous channel sets and durations. Its unsupervised objective predicts an original-signal embedding from a channel- and token-dropped view with a contrastive loss.

LaBraM (Jiang et al., 2024b) uses neural tokenization with masked prediction on large-scale heterogeneous EEG. A learned codebook quantizes EEG patches into discrete tokens; a Transformer predicts masked tokens from visible context. Its pretraining corpus contains approximately 2,500 hours collected from about 20 datasets, and the model achieves strong transfer on medical tasks.

CBraMod (Wang et al., 2025) reconstructs masked EEG patches using cascaded spatial–temporal attention. Its criss-cross Transformer applies separate spatial (cross-channel) and temporal (cross-time) attention to the patch grid.

EEGPT (Wang et al., 2024) reconstructs masked EEG content while learning electrode-aware representations through a target-side pretraining design. **CSBrain** (Zhou et al., 2025) operates on EEG spectrograms and reconstructs masked spectral regions with specialized spatial–temporal attention. Both models are evaluated by EEG-FM-Bench but are not part of the released NeuralBench baseline set.

LUNA (Döner et al., 2025) maps variable electrode layouts into a fixed-size latent through learned queries and cross-attention, then applies temporal attention in that topology-agnostic space. It is pretrained on TUEG and Siena with masked-patch reconstruction.

REVE (El Ouahidi et al., 2025) combines masked waveform reconstruction with a four-dimensional positional encoding that accommodates variable electrode locations and recording layouts. Its pretraining corpus aggregates recordings from 92 sources and approximately 25,000 subjects.

The public-model reproduction uses the benchmark's task-specific preprocessing and adaptation configuration with seeds 33, 34, and 35 and checkpoint selection by the task-specific validation metric. NSP follows the same task identities and metric definitions, with validation-selected task-specific readouts.

D.3 Adaptation Protocols

Full fine-tuning: All encoder parameters and classification head are updated. This is the primary protocol for NeuralBench-EEG-Core and EEG-FM-Bench Full FT single-task and multi-task evaluations.

Frozen evaluation: Encoder parameters are frozen; only classification heads are trained. EEG-FM-Bench uses multi-task evaluation, whereas the component ablations train a head separately for each dataset under single-task evaluation.

LoRA: Low-rank adaptation (Hu et al., 2021) updates low-rank matrices injected into attention layers while freezing the base encoder. Rank is set to 8 for all experiments.

All protocols use the corresponding validation metric for checkpoint selection and report the matching test metric only on the selected checkpoint: BAC for classification and Top-5 accuracy for retrieval.

E Complete Benchmark Results

E.1 NeuralBench-EEG-Core

NeuralBench-EEG-Core v1.0 provides task-specific preprocessing, splits, adaptation settings, and evaluation metrics for heterogeneous EEG datasets (Banville et al., 2026). We retain 16 datasets after removing dataset–task pairs also evaluated by EEG-FM-Bench, yielding four clinical and neurological-disease tasks, four BCI and sensorimotor tasks, four event-related or evoked-response tasks, and four cognitive-state or language-decoding tasks. The following four subsections first describe the datasets in each group and then compare model performance. Fourteen classification datasets use BAC as the primary metric, whereas ZuCo and Nieuwland2018 use subject-aggregated Top-5 retrieval accuracy. Every table additionally reports two task-appropriate metrics and marks the best and second-best mean for each metric.

E.1.1 Clinical and Neurological Disease

Mumtaz2018 contains 14,683 five-second windows from 64 participants for major-depression diagnosis (Mumtaz et al., 2018); Singh2021 contains 39,543 windows from 129 participants for Parkinson's-disease classification (Singh et al., 2021); and Albrecht2019 contains 24,928 windows from 77 participants for schizophrenia classification (Albrecht et al., 2019). CHB-MIT (the Dan2023 benchmark task) is the larger seizure-detection task, with 707,683 windows from 23 participants (Jonathan & Shoeb, 2023). All four tasks are binary and report BAC, AUROC, and AUCPR. NSP obtains the highest BAC on the three diagnostic datasets and remains competitive on CHB-MIT, where CBraMod leads BAC while NSP ranks second in AUROC and AUCPR. On Mumtaz2018 and Albrecht2019, NSP also has the highest AUROC; its AUCPR ranks first and second, respectively.

Table 6: NeuralBench clinical and neurological-disease results (%, mean ± sample SD). Red and blue mark the best and second-best mean for each metric.

Mumtaz2018				**Singh2021**			
Model	BAC	AUROC	AUCPR	Model	BAC	AUROC	AUCPR
BENDR	86.22±00.47	92.20±00.26	91.17±00.27	BENDR	65.82±05.62	70.83±06.33	85.33±03.73
BIOT	79.11±03.39	86.20±01.13	82.88±05.00	BIOT	74.18±04.16	81.60±04.45	88.86±02.51
LaBraM	85.61±03.43	88.33±05.14	87.57±03.82	LaBraM	**76.64±04.06**	**87.32±04.69**	**94.19±02.42**
CBraMod	86.04±02.48	92.57±02.82	93.63±01.06	CBraMod	74.37±02.38	83.18±04.65	91.94±02.19
LUNA	85.42±00.98	**94.04±00.69**	**94.43±01.02**	LUNA	64.95±00.49	73.50±01.11	87.02±00.59
REVE	**87.33±01.44**	90.72±01.45	86.62±03.03	REVE	72.58±01.72	82.46±03.04	**92.56±01.81**
NSP	**90.43±06.40**	**96.12±06.78**	**94.81±06.54**	NSP	**76.67±01.70**	**84.21±01.87**	91.26±02.03
Albrecht2019				**CHB-MIT**			
Model	BAC	AUROC	AUCPR	Model	BAC	AUROC	AUCPR
BENDR	57.17±00.87	60.31±01.40	69.42±00.23	BENDR	78.97±01.04	88.53±01.77	99.94±00.01
BIOT	56.18±01.50	57.52±01.38	64.19±02.16	BIOT	73.58±05.20	80.41±02.16	99.81±00.04
LaBraM	63.62±01.93	66.26±04.42	75.04±07.83	LaBraM	**82.09±00.75**	**93.07±00.52**	**99.97±00.00**
CBraMod	57.05±03.12	60.03±04.59	67.79±04.87	CBraMod	**82.89±01.58**	90.84±01.72	99.93±00.03
LUNA	**66.75±02.18**	**71.42±02.22**	**81.81±02.04**	LUNA	70.46±02.44	80.50±06.88	99.87±00.08
REVE	49.21±10.36	50.00±16.05	63.24±12.36	REVE	76.51±03.73	82.74±07.09	99.83±00.11
NSP	**68.43±02.71**	**73.06±02.90**	**79.44±03.15**	NSP	81.10±01.39	**91.64±01.57**	**99.95±00.02**

E.1.2 BCI and Sensorimotor Decoding

This group spans four multiclass BCI paradigms. Thielen2021 provides 45,000 c-VEP windows from 30 participants (Thielen et al., 2021); Scherer2015 contains 3,550 five-class motor-imagery windows from nine participants (Scherer et al., 2015); Srisrisawang2024 contains 19,157 sixteen-class motor-execution windows from 20 participants (Srisrisawang & Müller-Putz, 2024); and Wang2017 contains 8,160 forty-class SSVEP windows from 34 participants (Wang et al., 2016).

NSP leads BAC, weighted F1, and κ on Thielen2021, Scherer2015, and Wang2017. On Srisrisawang2024, REVE is strongest and NSP is second across all three metrics.

Table 7: NeuralBench BCI and sensorimotor results (%, mean ± sample SD). Red and blue mark the best and second-best mean for each metric.

Thielen2021				**Scherer2015**			
Model	BAC	F1	κ	Model	BAC	F1	κ
BENDR	4.96±00.24	1.64±00.37	-0.04±00.26	BENDR	21.62±00.66	19.11±01.61	2.02±00.83
BIOT	4.80±00.25	3.84±00.04	-0.21±00.26	BIOT	**26.48±02.48**	**25.69±03.03**	**8.10±03.10**
LaBraM	86.39±00.74	86.48±00.78	85.67±00.78	LaBraM	21.83±00.17	20.34±00.83	2.29±00.21
CBraMod	86.37±01.33	86.38±01.31	85.66±01.40	CBraMod	21.22±01.80	15.89±05.21	1.52±02.25
LUNA	**87.06±00.28**	**87.05±00.28**	**86.38±00.30**	LUNA	20.95±01.09	14.21±06.01	1.19±01.37
REVE	80.51±01.81	80.55±01.86	79.49±01.91	REVE	26.06±01.07	24.02±01.88	7.57±01.33
NSP	**88.83±03.37**	**88.74±03.36**	**88.05±03.33**	NSP	**32.00±00.55**	**30.13±00.52**	**9.74±00.79**
Srisrisawang2024				**Wang2017**			
Model	BAC	F1	κ	Model	BAC	F1	κ
BENDR	34.56±00.96	32.79±01.28	30.21±01.01	BENDR	13.31±05.51	7.13±04.10	11.09±05.66
BIOT	18.00±01.78	16.24±02.13	12.53±01.89	BIOT	58.67±01.85	58.33±02.05	57.61±01.89
LaBraM	52.94±00.88	53.16±00.70	49.87±00.92	LaBraM	**96.31±00.45**	**96.31±00.44**	**96.21±00.46**
CBraMod	49.41±00.99	48.82±01.23	46.08±01.06	CBraMod	85.85±07.27	85.62±07.59	85.49±07.45
LUNA	49.92±00.72	49.86±01.02	46.63±00.77	LUNA	13.08±00.88	11.22±00.57	10.85±00.90
REVE	**57.90±01.64**	**57.86±01.70**	**55.18±01.73**	REVE	65.79±02.09	65.21±03.10	64.92±03.42
NSP	**54.40±00.95**	**53.75±00.94**	**50.72±00.89**	NSP	**96.83±00.91**	**96.72±00.91**	**96.61±00.90**

E.1.3 Event-Related and Evoked Responses

ERP CORE-ERN, ERP CORE-N170, and ERP CORE-N400 isolate error-related, face-perception, and semantic-violation responses in one-second windows, with 15,972, 12,800, and 4,800 examples, respectively (Kappenman et al., 2021). Schreuder2010 provides a larger auditory-P300 task with 253,800 windows from 21 participants (Schreuder et al., 2010). All four tasks are binary and report BAC, AUROC, and AUCPR. NSP leads all three metrics on ERP CORE-N400 and Schreuder2010, ranks second on ERP CORE-N170, and is below the strongest baselines on ERP CORE-ERN.

Table 8: NeuralBench event-related and evoked-response results (%, mean ± sample SD). Red and blue mark the best and second-best mean for each metric.

ERP CORE-ERN				**ERP CORE-N170**			
Model	BAC	AUROC	AUCPR	Model	BAC	AUROC	AUCPR
BENDR	78.06±04.59	86.11±04.33	96.37±01.35	BENDR	67.03±00.61	71.94±01.25	47.25±02.01
BIOT	49.40±01.12	50.54±06.31	82.92±04.01	BIOT	49.84±00.36	50.20±00.46	25.55±00.46
LaBraM	**81.96±01.52**	**90.10±01.67**	97.18±00.75	LaBraM	72.14±00.55	78.75±01.01	57.33±02.71
CBraMod	81.32±01.04	89.43±00.77	**97.31±00.29**	CBraMod	71.14±00.55	78.72±01.37	57.09±03.52
LUNA	81.57±03.26	89.74±02.40	97.09±01.05	LUNA	68.74±02.04	76.63±01.34	54.25±01.05
REVE	**84.61±01.92**	**91.73±01.41**	**97.76±00.59**	REVE	**74.03±05.33**	**80.88±03.75**	**59.25±05.32**
NSP	80.07±02.94	88.35±03.24	96.86±02.44	NSP	**73.27±00.91**	**79.88±00.99**	**58.41±02.85**
ERP CORE-N400				**Schreuder2010**			
Model	BAC	AUROC	AUCPR	Model	BAC	AUROC	AUCPR
BENDR	60.45±01.15	65.56±01.11	68.95±01.46	BENDR	**66.70±00.16**	**72.33±00.12**	**91.73±00.05**
BIOT	50.07±00.64	49.29±00.43	49.44±01.14	BIOT	50.25±00.21	50.80±00.63	83.76±00.26
LaBraM	**64.62±01.09**	70.43±01.22	72.77±01.87	LaBraM	66.05±00.40	71.99±00.46	91.72±00.22
CBraMod	60.42±01.13	66.54±01.35	66.62±01.33	CBraMod	63.11±00.62	68.28±00.32	90.56±00.11
LUNA	64.20±01.55	70.16±01.71	72.42±02.26	LUNA	62.97±00.39	68.12±00.32	90.38±00.13
REVE	64.51±01.97	**71.31±01.06**	**73.03±00.81**	REVE	64.70±01.00	70.81±00.49	91.15±00.26
NSP	**66.67±00.75**	**72.35±00.81**	**74.11±01.91**	NSP	**69.87±00.67**	**75.71±00.73**	**92.44±00.23**

E.1.4 Cognitive State and Language Decoding

Zyma2019 evaluates binary mental-arithmetic decoding using 1,659 windows from 35 participants (Zyma et al., 2019), while Hinss2023 evaluates three-way cognitive-state decoding using 15,399 windows from 29 participants (Hinss et al., 2023). ZuCo contains 25,819 sentence-reading examples from 12 participants (Hollenstein et al., 2018), and Nieuwland2018 provides 529,730 word-level examples from 222 participants (Nieuwland et al., 2018); these two tasks are evaluated as retrieval rather than classification. NSP ranks second on Zyma2019 BAC, leads all three

Hinss2023 metrics, and obtains the best Top-1, Top-5, and median-rank results on both retrieval datasets. Retrieval values are interpreted within their own metric family and are not pooled into the 14-task BAC macro.

Table 9: NeuralBench cognitive-state and language results (mean ± sample SD). All metrics except median rank (MedR) are percentages. Retrieval uses Top-1, Top-5, and MedR; MedR is an unscaled rank and lower is better. Red and blue mark the best and second-best mean for each metric.

Zyma2019				**Hinss2023**			
Model	BAC	AUROC	AUCPR	Model	BAC	F1	κ
BENDR	65.28±02.34	74.69±05.12	89.33±03.31	BENDR	45.89±00.19	45.01±01.05	18.83±00.28
BIOT	66.40±03.04	72.30±05.30	89.05±01.27	BIOT	43.18±04.73	42.35±05.03	14.77±07.09
LaBraM	68.92±02.42	80.90±03.78	92.18±02.89	LaBraM	65.46±03.17	63.16±04.69	48.20±04.76
CBraMod	72.02±01.30	77.04±00.39	91.57±00.49	CBraMod	61.37±01.70	59.32±02.71	42.06±02.54
LUNA	72.35±00.41	76.85±00.69	90.16±02.05	LUNA	**66.01±00.94**	64.59±01.16	**49.01±01.40**
REVE	**72.88±02.81**	**85.54±01.30**	**94.12±00.86**	REVE	65.57±04.49	**65.02±04.48**	48.35±06.73
NSP	**72.50±05.04**	**81.43±05.67**	**92.27±03.42**	NSP	**72.93±00.42**	**71.26±00.41**	**55.14±00.47**
ZuCo				**Nieuwland2018**			
Model	Top-1	Top-5	MedR↓	Model	Top-1	Top-5	MedR↓
BENDR	1.38±00.09	6.47±01.29	49.33±01.53	BENDR	7.48±00.85	20.73±01.69	26.00±02.65
BIOT	1.91±00.28	7.67±00.39	49.67±02.31	BIOT	1.14±00.08	5.03±00.09	71.00±01.00
LaBraM	2.24±00.24	8.95±00.26	45.00±01.00	LaBraM	11.26±00.86	28.34±01.27	17.33±01.53
CBraMod	0.54±00.17	2.25±00.30	118.33±03.79	CBraMod	11.61±00.22	29.73±00.66	15.33±00.58
LUNA	**3.15±00.15**	11.32±00.56	40.67±01.53	LUNA	10.79±00.67	27.99±01.37	17.00±01.73
REVE	2.95±00.21	**11.39±00.40**	**38.67±01.53**	REVE	**14.44±00.22**	**33.73±00.68**	**12.67±01.16**
NSP	**3.36±00.11**	**12.11±00.39**	**37.22±01.40**	NSP	**15.01±00.03**	**36.27±00.12**	**11.31±00.88**

E.2 EEG-FM-BENCH

The EEG-FM-Bench protocol (Xiong et al., 2025) evaluates 14 datasets across 10 EEG paradigms under four adaptation strategies: single-task full fine-tuning, multi-task full fine-tuning, frozen-backbone multi-task transfer, and multi-task LoRA. Each dataset paragraph introduces the task and recording scale before summarizing NSP across the four strategies. The tables report BAC together with F1 and Cohen's κ for multiclass tasks, or AUROC and AUCPR for binary tasks. Across all 14 datasets, NSP obtains 63.94 macro BAC with multi-task full fine-tuning.

E.2.1 EMOTION RECOGNITION

This domain evaluates affective decoding at progressively finer label granularity. SEED, SEED-V, and SEED-VII move from three to five and seven emotion classes while retaining dense multichannel recordings, allowing the comparison to expose how adaptation quality changes as inter-class boundaries become more subtle.

SEED. SEED evaluates three-class emotion recognition from 60-channel, 10-s EEG windows, comprising 37,890 samples from 15 participants (Zheng & Lu, 2015). NSP obtains 73.40 ± 00.26 BAC with single-task full fine-tuning, 73.62 ± 00.97 with multi-task full fine-tuning, 68.93 ± 00.60 with frozen transfer, and 71.67 ± 00.67 with LoRA. Its best setting is multi-task full fine-tuning, which exceeds the strongest baseline under that setting by 1.00 point; the table reports the corresponding F1 and κ values.

Table 10: Complete SEED results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	59.50±00.42	58.88±00.98	39.42±00.64	BENDR	34.12±00.24	24.48±02.82	01.10±00.32
BIOT	63.87±01.77	63.12±01.94	45.99±02.67	BIOT	57.85±00.19	58.17±00.21	37.17±00.29
LaBraM	61.59±01.71	60.30±01.60	43.28±02.48	LaBraM	52.13±00.05	51.23±00.34	28.33±00.12
EEGPT	70.83±00.28	70.48±00.16	56.41±00.42	EEGPT	61.80±00.65	60.41±01.07	42.89±00.97
CBraMod	62.59±02.40	61.67±02.66	44.36±03.58	CBraMod	44.32±00.28	44.50±00.30	16.76±00.45
CSBrain	70.23±00.25	70.70±00.22	55.70±00.28	CSBrain	58.57±00.05	58.53±00.05	38.00±00.14
REVE	**73.37±00.30**	**73.32±00.41**	**62.77±00.44**	REVE	**63.22±00.13**	**63.35±00.08**	**45.20±00.32**
NSP	**73.40±00.26**	**73.77±00.31**	**60.43±00.45**	NSP	**68.93±00.60**	**69.33±00.60**	**53.73±00.96**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	60.82±00.19	60.84±00.20	41.59±00.16	BENDR	33.30±00.00	17.00±00.00	00.00±00.00
BIOT	65.62±00.92	65.57±00.81	50.11±00.99	BIOT	62.70±00.14	62.60±00.08	44.20±00.22
LaBraM	67.71±00.55	68.11±00.66	52.22±00.83	LaBraM	45.90±00.22	43.07±02.04	18.90±00.29
EEGPT	68.75±00.40	67.42±00.78	52.37±00.28	EEGPT	**69.97±00.05**	**69.27±00.61**	**55.10±00.08**
CBraMod	72.25±01.90	**72.90±01.63**	58.68±02.91	CBraMod	66.07±00.21	65.57±00.38	49.27±00.29
CSBrain	71.92±00.39	71.84±00.20	58.02±00.59	CSBrain	64.67±00.05	64.07±00.17	47.20±00.08
REVE	**72.62±00.92**	72.75±00.76	**62.57±00.94**	REVE	69.23±00.21	68.97±00.46	54.03±00.34
NSP	**73.62±00.97**	**74.02±00.97**	**60.94±01.51**	NSP	**71.67±00.67**	**72.10±00.61**	**57.83±00.98**

SEED-V. SEED-V increases emotion granularity to five classes and provides 60-channel, 10-s windows, with 12,318 samples from 16 participants (Liu et al., 2022). NSP reaches 46.30 ± 01.31 BAC with single-task full fine-tuning, 43.42 ± 00.77 with multi-task full fine-tuning, 38.77 ± 00.55 with frozen transfer, and 41.77 ± 00.21 with LoRA. Single-task full fine-tuning is strongest and leads the best listed baseline by 5.65 points.

Table 11: Complete SEED-V results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	20.20±00.13	10.72±01.57	00.21±00.13	BENDR	20.00±00.00	06.00±01.98	00.00±00.00
BIOT	26.66±00.13	25.85±00.13	09.53±00.08	BIOT	26.27±00.47	24.77±00.90	09.17±00.54
LaBraM	24.35±00.55	23.64±01.21	05.61±01.13	LaBraM	23.23±00.48	19.63±00.91	04.43±00.68
EEGPT	36.43±01.33	33.00±01.94	18.52±00.67	EEGPT	**30.48±00.73**	**28.71±01.45**	**12.90±00.57**
CBraMod	28.34±00.17	27.48±00.89	10.55±00.20	CBraMod	20.28±00.02	10.93±00.03	00.45±00.03
CSBrain	38.23±00.24	39.07±00.34	22.70±00.29	CSBrain	21.97±00.12	14.43±00.62	02.50±00.16
REVE	**40.65±00.51**	**40.35±01.63**	**25.82±00.72**	REVE	25.43±00.33	24.07±00.41	07.07±00.37
NSP	**46.30±01.31**	**44.70±01.65**	**29.93±01.94**	NSP	**38.77±00.55**	**39.50±00.44**	**23.50±00.70**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	21.75±00.66	19.94±02.58	02.90±00.68	BENDR	20.00±00.00	08.80±00.00	00.00±00.00
BIOT	30.54±00.18	30.97±00.26	14.14±00.21	BIOT	28.37±00.12	27.87±00.33	11.33±00.19
LaBraM	**43.26±00.29**	**43.20±00.81**	**28.93±00.50**	LaBraM	23.93±00.53	22.10±00.80	05.17±00.69
EEGPT	43.07±00.60	42.99±00.54	28.62±00.86	EEGPT	36.07±00.85	35.77±01.77	20.10±00.75
CBraMod	40.56±00.77	41.72±00.72	26.22±00.97	CBraMod	35.03±00.27	33.22±00.54	18.88±00.33
CSBrain	38.02±00.23	38.07±01.03	22.48±00.42	CSBrain	30.43±00.17	28.20±00.65	13.17±00.29
REVE	40.84±00.98	40.58±02.00	25.59±01.48	REVE	**38.37±00.21**	**38.87±00.42**	**22.70±00.49**
NSP	**43.42±00.77**	**43.44±00.85**	**28.96±01.05**	NSP	**41.77±00.21**	**41.47±00.31**	**27.03±00.35**

SEED-VII. SEED-VII is a seven-class emotion-recognition task with 60-channel, 15-s windows, totaling 19,420 samples from 20 participants (Jiang et al., 2024a). NSP records 24.50 ± 00.98 BAC with single-task full fine-tuning, 26.50 ± 01.49 with multi-task full fine-tuning, 24.23 ± 00.80 with frozen transfer, and 23.40 ± 01.22 with LoRA. Multi-task full fine-tuning is the strongest NSP setting but remains 1.28 points below the best baseline under the same protocol.

Table 12: Complete SEED-VII results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	17.43±00.60	14.20±00.79	03.98±00.69	BENDR	14.30±00.00	03.63±00.61	00.00±00.00
BIOT	18.89±00.43	16.84±01.12	05.80±00.48	BIOT	20.60±00.45	16.60±00.43	07.50±00.42
LaBraM	20.39±00.41	19.55±03.82	09.42±02.64	LaBraM	23.23±00.48	19.63±00.91	04.43±00.68
EEGPT	**26.42±00.55**	**23.89±01.29**	**14.05±00.99**	EEGPT	**25.20±00.08**	**24.17±00.12**	**12.70±00.08**
CBraMod	23.10±00.93	21.82±02.00	09.70±02.62	CBraMod	19.43±00.72	16.76±00.29	06.97±00.95
CSBrain	23.73±00.17	23.43±00.12	10.40±00.22	CSBrain	18.90±00.29	16.60±00.50	06.27±00.41
REVE	22.77±00.17	**23.67±00.09**	**10.60±00.08**	REVE	20.57±00.17	**20.50±00.33**	08.50±00.29
NSP	**24.50±00.98**	22.27±01.27	10.30±01.35	NSP	**24.23±00.80**	19.13±01.14	**10.07±01.10**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	21.98±00.66	19.31±00.87	09.49±01.03	BENDR	14.30±00.00	04.77±00.33	00.00±00.00
BIOT	23.93±00.39	22.13±00.52	11.38±00.38	BIOT	23.85±00.08	21.95±00.12	11.37±00.11
LaBraM	26.13±00.92	24.74±00.57	13.07±00.93	LaBraM	18.13±00.71	13.30±01.36	04.37±01.23
EEGPT	**27.78±00.47**	25.40±00.12	**15.80±01.12**	EEGPT	**29.83±00.53**	**27.30±01.55**	**17.73±01.03**
CBraMod	26.05±00.72	**25.74±01.22**	13.09±00.86	CBraMod	18.73±00.19	14.87±01.00	05.00±00.36
CSBrain	26.67±01.37	22.68±02.69	11.40±02.04	CSBrain	22.73±00.09	22.27±00.24	10.80±00.08
REVE	**27.62±00.61**	23.93±01.08	14.78±00.44	REVE	**23.93±00.58**	21.67±01.76	**11.48±00.88**
NSP	26.50±01.49	**26.30±01.82**	**14.96±01.77**	NSP	23.40±01.22	**23.30±01.42**	10.10±01.65

E.2.2 Motor Imagery

This domain covers imagined-movement decoding across heterogeneous cohorts, montages, and action taxonomies. PhysioMI provides a large-participant four-class setting, Mimul-11 targets three upper-limb actions, and BCIC-2a supplies the standard four-class BCI protocol.

PhysioMI. PhysioMI evaluates four-class motor imagery from 64-channel, 4-s trials, comprising 9,747 samples from 109 participants (Schalk et al., 2004). NSP achieves 59.23 ± 00.67 BAC with single-task full fine-tuning, 58.68 ± 00.21 with multi-task full fine-tuning, 36.90 ± 00.25 with frozen transfer, and 45.13 ± 00.95 with LoRA. Single-task full fine-tuning performs best for NSP and is 0.80 points below the strongest baseline in that setting.

Table 13: Complete PhysioMI results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	47.78±00.28	42.72±08.93	36.48±08.47	BENDR	25.22±00.24	12.55±02.53	00.27±00.31
BIOT	36.22±00.10	35.65±00.18	14.95±00.15	BIOT	29.80±00.33	29.90±00.43	06.40±00.41
LaBraM	57.27±00.26	57.29±00.25	43.00±00.35	LaBraM	29.63±00.34	28.97±00.77	06.23±00.42
EEGPT	54.16±00.18	53.27±00.92	38.04±01.08	EEGPT	**39.90±01.39**	**39.51±01.92**	**19.87±01.84**
CBraMod	56.74±00.36	56.74±00.39	42.31±00.48	CBraMod	26.90±00.24	22.76±00.17	02.54±00.32
CSBrain	56.57±00.45	56.57±00.53	42.07±00.59	CSBrain	26.80±00.29	24.77±01.25	02.40±00.37
REVE	**60.03±00.58**	**60.08±00.62**	**46.65±00.80**	REVE	32.55±00.18	32.12±00.40	10.08±00.23
NSP	**59.23±00.67**	**59.00±00.78**	**45.63±00.93**	NSP	**36.90±00.25**	**36.47±00.65**	**15.83±00.67**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	44.77±00.31	44.63±00.03	26.33±00.40	BENDR	25.10±00.08	11.90±02.55	00.13±00.12
BIOT	33.02±00.29	32.75±00.36	10.67±00.40	BIOT	26.93±00.05	26.17±00.48	02.57±00.09
LaBraM	43.19±01.03	42.38±01.15	24.25±01.37	LaBraM	27.40±00.16	20.47±01.13	03.20±00.24
EEGPT	50.52±00.80	50.26±00.76	34.00±01.06	EEGPT	**45.83±00.85**	**45.33±00.74**	**27.80±01.10**
CBraMod	41.80±00.69	40.35±00.87	22.42±00.92	CBraMod	27.60±00.22	24.50±02.03	03.50±00.29
CSBrain	55.43±00.37	55.10±00.63	40.48±00.48	CSBrain	30.68±00.28	29.92±00.61	07.57±00.38
REVE	**58.82±00.45**	**58.77±00.56**	**45.10±00.60**	REVE	**53.95±00.33**	**54.18±00.31**	**38.62±00.46**
NSP	**58.68±00.21**	**57.84±00.89**	**41.54±00.81**	NSP	45.13±00.95	44.97±01.98	26.87±00.60

Mimul-11. Mimul-11 is a three-class upper-limb motor-imagery task with 60-channel, 5-s trials, containing 41,347 samples from 11 participants (Jeong et al., 2020). NSP obtains 49.70 ± 00.91 BAC with single-task full fine-tuning, 51.66 ± 00.60 with multi-task full fine-tuning, 40.47 ± 00.26

with frozen transfer, and 44.83 ± 00.80 with LoRA. Multi-task full fine-tuning is its strongest setting and trails the leading baseline by 0.89 points.

Table 14: Complete Mimul-11 results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	**50.44±02.30**	**59.09±01.23**	**36.10±03.49**	BENDR	33.30±00.00	37.90±00.00	00.00±00.00
BIOT	40.55±01.14	49.68±01.29	14.45±02.10	BIOT	36.65±00.13	43.92±00.38	06.90±00.27
LaBraM	47.80±01.68	**57.43±01.27**	**30.71±03.14**	LaBraM	35.50±00.43	42.97±00.88	04.60±00.93
EEGPT	45.19±01.69	54.66±01.38	23.68±02.30	EEGPT	37.11±01.97	**44.95±03.07**	10.06±03.66
CBraMod	47.07±00.09	55.10±00.36	25.13±00.69	CBraMod	33.95±00.11	40.09±00.18	01.51±00.25
CSBrain	42.67±00.77	51.93±00.87	18.13±02.05	CSBrain	34.80±00.59	42.73±01.05	03.10±01.39
REVE	45.41±01.15	50.63±00.71	18.93±01.86	REVE	**40.20±00.32**	**48.02±00.50**	**12.38±00.64**
NSP	**49.70±00.91**	51.33±00.78	30.03±00.89	NSP	**40.47±00.26**	44.67±00.93	**13.80±00.42**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	**52.55±01.52**	**59.61±02.80**	**38.35±04.93**	BENDR	33.30±00.00	37.90±00.00	00.00±00.00
BIOT	40.43±00.63	49.02±00.73	14.63±01.88	BIOT	40.57±00.37	49.53±00.31	12.40±00.57
LaBraM	45.05±00.66	53.42±01.98	22.60±02.20	LaBraM	34.87±00.24	41.67±00.45	03.63±00.40
EEGPT	50.77±00.61	**57.44±01.11**	**31.16±02.43**	EEGPT	43.00±00.36	**52.00±00.50**	**18.90±00.29**
CBraMod	44.53±00.19	52.80±00.57	19.23±00.77	CBraMod	42.90±00.92	50.00±01.22	15.07±01.62
CSBrain	41.64±00.54	50.70±00.61	15.94±01.33	CSBrain	39.10±00.24	48.23±00.26	12.13±00.37
REVE	45.22±00.13	53.73±00.65	21.53±01.16	REVE	**43.77±00.82**	**50.33±01.98**	17.90±02.01
NSP	**51.66±00.60**	55.26±00.81	23.44±00.88	NSP	**44.83±00.80**	50.20±00.96	**18.13±00.59**

BCIC-2a. BCIC-2a provides a standard four-class motor-imagery task with 22-channel, 4-s trials, totaling 5,088 samples from nine participants (Tangermann et al., 2012). NSP reaches 46.50 ± 00.52 BAC with single-task full fine-tuning, 46.58 ± 01.48 with multi-task full fine-tuning, 36.50 ± 01.21 with frozen transfer, and 37.03 ± 00.83 with LoRA. Its best result comes from multi-task full fine-tuning and exceeds the strongest baseline in that setting by 4.08 points.

Table 15: Complete BCIC-2a results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	35.21±00.54	33.25±00.57	13.62±00.72	BENDR	25.17±00.24	10.65±00.92	00.22±00.31
BIOT	33.33±00.79	29.85±00.27	11.10±01.04	BIOT	29.73±00.54	25.10±00.43	06.97±00.79
LaBraM	32.37±00.21	30.73±00.39	09.83±00.30	LaBraM	28.40±00.24	22.83±00.45	04.57±00.33
EEGPT	**44.07±03.27**	38.61±04.23	**25.42±04.36**	EEGPT	**32.02±00.06**	**25.95±01.13**	**09.40±00.08**
CBraMod	33.71±00.86	28.99±01.35	11.61±01.14	CBraMod	29.17±01.03	24.94±01.72	05.56±01.37
CSBrain	42.10±00.87	**40.64±00.62**	22.83±01.13	CSBrain	28.13±00.17	18.57±00.12	04.17±00.26
REVE	42.73±01.10	**42.60±01.77**	**27.08±01.46**	REVE	29.18±00.83	23.37±01.36	05.57±01.12
NSP	**46.50±00.52**	37.13±00.97	23.63±00.95	NSP	**36.50±01.21**	**32.10±02.98**	**15.33±01.95**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	34.98±00.25	30.01±02.27	13.31±00.33	BENDR	25.00±00.00	10.00±00.00	00.00±00.00
BIOT	29.66±00.39	23.61±04.64	05.99±00.45	BIOT	29.00±00.14	19.40±00.14	05.33±00.19
LaBraM	34.58±01.64	31.24±02.71	12.60±02.00	LaBraM	28.30±00.73	17.87±01.01	04.43±00.93
EEGPT	39.23±01.48	33.08±01.86	16.95±01.97	EEGPT	36.70±01.15	32.67±03.16	15.60±01.53
CBraMod	35.50±00.58	23.97±00.25	14.00±00.77	CBraMod	31.00±00.70	22.20±02.19	08.03±00.98
CSBrain	**42.50±01.08**	**39.97±02.34**	**23.32±01.48**	CSBrain	32.00±00.99	20.77±01.11	09.30±01.35
REVE	41.87±03.34	38.68±02.25	22.48±03.47	REVE	**38.63±00.19**	**38.43±00.17**	**18.17±00.25**
NSP	**46.58±01.48**	**40.64±02.07**	**22.76±02.28**	NSP	**37.03±00.83**	**34.50±01.80**	**16.03±01.10**

E.2.3 Clinical EEG Analysis

This domain spans event-level and recording-level clinical interpretation. TUEV and TUSL require localization of transient abnormal events, whereas TUAB evaluates whole-record abnormality, jointly testing robustness to class imbalance and substantial recording heterogeneity.

TUEV. TUEV evaluates six-class abnormal-event classification from 21-channel, 5-s windows, comprising 113,353 samples from 370 participants (Harati et al., 2015). NSP obtains 73.17 ± 01.70 BAC with single-task full fine-tuning, 74.34 ± 01.03 with multi-task full fine-tuning, 52.90 ± 02.16 with frozen transfer, and 69.60 ± 01.25 with LoRA. Multi-task full fine-tuning is strongest and narrowly exceeds the best baseline in that setting by 0.07 points.

Table 16: Complete TUEV results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	65.53±02.40	80.95±02.18	66.62±04.07	BENDR	16.70±00.00	44.10±00.00	00.00±00.00
BIOT	57.07±04.29	74.34±04.00	39.55±04.15	BIOT	52.12±00.55	80.93±00.20	68.85±00.28
LaBraM	59.58±04.31	77.72±03.25	64.06±05.03	LaBraM	41.27±00.12	73.37±00.39	54.97±00.53
EEGPT	66.86±02.54	83.95±00.14	74.45±00.09	EEGPT	**63.08±00.66**	**86.22±00.31**	**76.83±00.71**
CBraMod	65.42±00.90	79.03±01.36	64.13±02.48	CBraMod	32.50±00.04	65.97±00.20	41.58±00.49
CSBrain	71.20±01.34	85.07±00.85	74.67±01.44	CSBrain	38.63±00.17	72.87±00.19	53.40±00.28
REVE	**73.98±01.21**	**90.25±01.42**	**85.07±01.80**	REVE	**69.03±00.17**	**89.53±00.30**	**83.22±00.52**
NSP	**73.17±01.70**	**88.90±01.25**	**81.03±01.00**	NSP	52.90±02.16	81.87±01.40	70.33±02.27
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	67.46±02.59	84.56±01.82	72.85±02.87	BENDR	16.70±00.00	44.10±00.00	00.00±00.00
BIOT	61.35±01.35	80.55±00.46	68.37±00.61	BIOT	59.68±00.17	81.28±00.31	68.33±00.59
LaBraM	71.53±00.19	86.22±00.73	77.42±01.20	LaBraM	47.13±00.59	72.33±01.32	54.33±02.28
EEGPT	70.85±01.16	86.35±00.47	77.69±01.06	EEGPT	61.07±00.98	**86.83±00.24**	**79.60±01.07**
CBraMod	69.41±01.86	83.44±00.31	72.02±00.65	CBraMod	57.80±02.06	77.60±03.19	63.07±04.76
CSBrain	71.82±00.69	87.00±00.64	77.84±01.55	CSBrain	64.02±00.61	79.05±01.32	65.80±02.08
REVE	**74.27±00.87**	**89.37±00.33**	**82.12±00.76**	REVE	**68.53±01.18**	77.83±00.86	58.30±00.99
NSP	**74.34±01.03**	**90.48±00.94**	**81.92±01.80**	NSP	**69.60±01.25**	**90.57±00.22**	**84.83±01.37**

TUSL. TUSL classifies three types of EEG slowing events from 21- or 22-channel, 10-s windows and contains 290 annotated samples from 28 participants (von Weltin et al., 2017). NSP records 81.17 ± 01.12 BAC with single-task full fine-tuning, 81.83 ± 01.25 with multi-task full fine-tuning, 67.07 ± 01.25 with frozen transfer, and 80.33 ± 00.70 with LoRA. Multi-task full fine-tuning is the best NSP configuration and leads the strongest baseline under that protocol by 1.27 points.

Table 17: Complete TUSL results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	47.22±01.48	34.17±02.80	17.88±01.95	BENDR	33.30±00.00	17.50±00.00	00.00±00.00
BIOT	60.05±04.47	56.83±04.49	37.52±06.77	BIOT	47.62±00.05	44.17±00.56	18.75±00.19
LaBraM	66.98±00.44	62.43±01.73	45.08±01.57	LaBraM	63.90±02.52	54.57±02.18	37.63±04.29
EEGPT	62.41±03.10	70.63±02.34	58.92±02.63	EEGPT	**69.13±01.65**	**69.27±02.15**	**59.70±03.25**
CBraMod	74.18±02.69	73.15±02.74	63.67±03.23	CBraMod	33.00±00.00	09.35±01.20	00.00±00.00
CSBrain	71.60±02.63	72.60±02.69	61.91±03.85	CSBrain	**68.17±01.04**	**62.93±01.79**	**47.63±01.93**
REVE	**79.58±01.32**	**74.58±01.25**	**65.07±01.86**	REVE	63.88±00.54	55.37±00.87	38.93±00.97
NSP	**81.17±01.12**	**76.37±01.32**	**67.50±01.10**	NSP	67.07±01.25	59.27±01.15	42.70±00.26
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	73.35±00.98	70.90±01.73	56.65±02.55	BENDR	33.30±00.00	25.70±00.00	00.00±00.00
BIOT	69.90±04.10	71.03±05.19	56.00±07.50	BIOT	62.90±00.28	54.53±00.47	36.93±00.33
LaBraM	75.86±01.15	72.74±01.33	59.60±01.81	LaBraM	54.68±00.47	37.08±01.33	25.47±00.82
EEGPT	76.12±02.05	**84.19±01.04**	**84.30±01.27**	EEGPT	72.59±00.81	71.10±00.03	56.74±00.23
CBraMod	79.56±03.24	76.24±04.45	64.72±05.96	CBraMod	73.90±00.74	67.57±01.41	53.93±01.47
CSBrain	**80.56±00.91**	76.72±00.38	65.57±00.41	CSBrain	75.03±02.53	71.53±03.20	58.12±04.65
REVE	79.85±01.27	77.81±01.89	66.62±02.35	REVE	**80.68±02.81**	**80.78±02.41**	**71.43±03.54**
NSP	**81.83±01.25**	**79.07±01.06**	**68.57±02.06**	NSP	**80.33±00.70**	**80.47±00.48**	**71.27±00.75**

TUAB. TUAB is a binary normal-versus-abnormal recording task represented by 23-channel, 30-s windows, totaling 272,320 samples from 2,383 participants (Lopez et al., 2015). NSP reaches 83.83 ± 00.50 BAC with single-task full fine-tuning, 82.58 ± 00.48 with multi-task full fine-tuning,

81.80 ± 00.26 with frozen transfer, and 80.10 ± 00.29 with LoRA. Single-task full fine-tuning is strongest and improves on the leading baseline in that setting by 1.11 points.

Table 18: Complete TUAB results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	**82.72±00.12**	**89.52±00.14**	**90.68±00.13**	BENDR	68.30±05.85	73.17±00.82	65.93±00.45
BIOT	79.52±00.59	88.27±00.16	88.45±00.07	BIOT	80.30±00.22	87.50±00.14	87.27±00.57
LaBraM	79.50±00.81	86.91±00.51	83.86±01.96	LaBraM	75.87±00.05	84.07±00.05	85.11±00.08
EEGPT	78.66±00.24	87.52±01.01	86.94±02.15	EEGPT	79.53±00.05	88.63±00.12	**88.47±00.05**
CBraMod	81.40±00.33	89.02±00.16	**89.93±00.31**	CBraMod	73.15±00.19	80.41±00.02	79.79±00.15
CSBrain	78.97±00.17	85.33±01.51	82.73±04.69	CSBrain	78.20±00.00	87.00±00.08	87.23±00.12
REVE	81.32±00.35	87.93±01.76	87.38±01.56	REVE	**80.80±00.32**	**88.77±00.41**	88.33±00.55
NSP	**83.83±00.50**	**89.40±01.00**	87.97±01.01	NSP	**81.80±00.26**	**90.80±00.20**	**91.17±00.23**
Model	Multi-task			Model	LoRA		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	**81.72±00.26**	**90.38±00.27**	**90.22±00.15**	BENDR	68.37±00.12	76.77±00.21	71.93±00.41
BIOT	80.85±00.36	88.67±00.62	89.01±00.47	BIOT	**80.67±00.28**	**88.10±00.29**	88.78±00.32
LaBraM	80.47±00.12	88.33±00.59	88.65±00.28	LaBraM	73.93±00.65	82.57±00.48	81.67±00.74
EEGPT	80.68±00.37	**90.10±00.52**	**90.35±00.54**	EEGPT	**83.93±00.50**	**91.93±00.39**	**91.93±00.21**
CBraMod	81.55±00.14	89.57±00.26	89.67±00.02	CBraMod	78.40±00.93	86.07±00.97	87.07±00.95
CSBrain	79.18±00.48	86.32±00.28	85.78±00.25	CSBrain	78.97±00.71	87.07±01.01	87.23±01.09
REVE	81.18±00.46	88.32±00.37	89.33±00.61	REVE	80.43±01.04	88.03±00.54	88.07±01.16
NSP	**82.58±00.48**	88.18±00.40	86.06±01.09	NSP	80.10±00.29	88.00±00.06	**88.93±01.29**

E.2.4 Sleep Staging

Sleep staging evaluates whether the learned representation preserves long-duration physiological structure from a sparse montage. HMC uses standard 30-s epochs and five sleep stages, complementing the shorter event-centered tasks elsewhere in the benchmark.

HMC. HMC evaluates five-stage sleep classification from a sparse four-channel montage using 30-s epochs, comprising 136,925 samples from 151 participants (Alvarez-Estevez & Rijsman, 2021). NSP reaches 75.23 ± 00.60 BAC with single-task full fine-tuning, 75.66 ± 00.60 with multi-task full fine-tuning, 72.97 ± 00.97 with frozen transfer, and 75.53 ± 00.25 with LoRA. Multi-task full fine-tuning performs best and exceeds the strongest baseline under that protocol by 3.48 points.

Table 19: Complete HMC results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	**72.63±00.13**	73.67±00.53	65.86±00.70	BENDR	24.18±00.02	27.02±00.02	07.50±00.05
BIOT	71.01±00.07	72.27±00.11	64.40±00.35	BIOT	66.13±00.21	70.27±00.29	61.03±00.17
LaBraM	70.85±00.44	71.52±00.55	64.47±00.70	LaBraM	59.80±00.00	64.10±00.37	53.40±00.16
EEGPT	69.67±01.24	73.03±00.32	64.64±00.92	EEGPT	**67.83±00.12**	**71.63±00.05**	**62.90±00.08**
CBraMod	71.14±00.14	72.76±00.55	64.86±00.41	CBraMod	51.81±00.55	58.11±00.82	44.95±00.68
CSBrain	71.13±00.26	73.10±00.22	65.20±00.29	CSBrain	65.60±00.00	69.73±00.21	62.00±00.28
REVE	71.82±00.82	**75.08±00.74**	**66.49±00.83**	REVE	63.80±00.08	65.33±00.09	56.27±00.05
NSP	**75.23±00.60**	**78.03±00.67**	**71.57±00.60**	NSP	**72.97±00.97**	**74.93±00.67**	**67.87±00.91**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	70.21±00.72	72.44±00.65	64.57±00.90	BENDR	20.00±00.00	20.30±00.00	00.00±00.00
BIOT	71.45±01.05	72.89±00.34	64.89±00.71	BIOT	70.43±00.19	**74.03±00.19**	**65.90±00.24**
LaBraM	71.63±01.04	72.28±01.08	65.22±01.23	LaBraM	44.10±01.31	47.10±01.56	32.67±01.39
EEGPT	71.30±00.91	72.46±00.96	64.86±00.77	EEGPT	**71.97±00.12**	72.27±00.39	64.63±00.21
CBraMod	72.03±00.14	**75.20±00.29**	**67.43±00.12**	CBraMod	69.13±00.17	71.60±00.73	63.00±00.28
CSBrain	**72.18±00.24**	74.16±00.27	66.43±00.16	CSBrain	69.93±00.09	73.13±00.17	64.70±00.14
REVE	71.93±00.45	72.07±00.73	64.40±00.54	REVE	71.03±00.25	71.77±00.25	63.83±00.26
NSP	**75.66±00.60**	**77.02±00.76**	**70.24±00.87**	NSP	**75.53±00.25**	**76.77±00.67**	**70.10±00.70**

E.2.5 SEIZURE DETECTION

Seizure detection tests rare-event discrimination under marked class imbalance. Siena is therefore reported with AUROC and AUCPR in addition to balanced accuracy so that both ranking quality and positive-class retrieval remain visible.

Siena. Siena is a binary seizure-detection dataset with 29-channel, 10-s windows, totaling 50,830 samples from 14 participants (Detti et al., 2020). NSP obtains 82.67 ± 01.27 BAC with single-task full fine-tuning, 86.90 ± 00.78 with multi-task full fine-tuning, 69.93 ± 00.06 with frozen transfer, and 77.37 ± 02.11 with LoRA. Multi-task full fine-tuning is strongest and leads the best baseline in that setting by 3.62 points; AUROC and AUCPR are included because seizure events are rare.

Table 20: Complete Siena results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				**Parameter-Efficient Multi-Task**			
Model	Single-task			Model	Frozen		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	74.93±01.76	90.90±03.60	99.81±00.08	BENDR	50.00±00.00	61.20±04.16	99.27±00.12
BIOT	72.05±01.15	84.23±03.42	99.72±00.09	BIOT	56.42±00.02	81.20±00.13	**99.70±00.00**
LaBraM	70.97±02.68	91.33±02.00	**99.87±00.05**	LaBraM	50.17±00.24	84.48±01.45	**99.70±00.05**
EEGPT	76.25±01.02	**92.08±01.15**	99.81±00.02	EEGPT	**73.90±01.28**	**97.28±00.43**	**99.93±00.02**
CBraMod	80.64±01.40	**93.86±00.25**	**99.87±00.01**	CBraMod	64.17±01.18	**90.16±00.41**	99.68±00.02
CSBrain	79.25±01.68	91.52±02.96	99.83±00.13	CSBrain	50.00±00.00	56.53±12.17	98.10±00.62
REVE	**80.95±00.32**	91.77±03.26	**99.85±00.07**	REVE	68.60±00.08	77.87±00.37	99.27±00.05
NSP	**82.67±01.27**	91.77±00.76	99.57±00.06	NSP	**69.93±00.06**	75.70±00.44	99.40±00.00
Model	Multi-task			Model	LoRA		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	78.96±03.08	91.31±04.74	99.81±00.11	BENDR	50.00±00.00	80.70±05.21	99.67±00.19
BIOT	74.35±00.71	81.20±00.92	99.26±00.05	BIOT	68.83±00.47	88.20±01.00	99.80±00.00
LaBraM	73.62±00.62	89.43±03.59	99.82±00.07	LaBraM	50.80±00.57	68.57±17.66	99.20±00.65
EEGPT	**83.28±01.56**	**93.44±00.12**	99.86±00.00	EEGPT	**84.57±00.61**	**94.50±00.45**	**99.90±00.00**
CBraMod	82.75±02.10	92.03±01.90	99.83±00.04	CBraMod	**82.13±00.78**	**93.20±01.35**	**99.83±00.05**
CSBrain	76.42±01.43	92.70±00.67	**99.87±00.05**	CSBrain	72.87±01.11	87.80±03.18	99.77±00.11
REVE	81.87±02.03	90.70±01.25	**99.88±00.03**	REVE	75.33±01.23	88.73±01.05	**99.83±00.05**
NSP	**86.90±00.78**	**93.10±01.55**	99.54±00.09	NSP	77.37±02.11	81.97±03.36	99.53±00.06

E.2.6 MENTAL WORKLOAD

Mental workload recognition tests state-level cognitive decoding rather than stimulus-locked or motor responses. The Workload dataset provides a binary stress/workload setting with a conventional 19-channel montage.

Workload. Workload is a binary cognitive-load task based on 19-channel, 10-s windows, comprising 2,134 samples from 36 participants (Lim et al., 2018). NSP achieves 82.53 ± 00.30 BAC with single-task full fine-tuning, 74.28 ± 01.80 with multi-task full fine-tuning, 75.77 ± 00.99 with frozen transfer, and 76.13 ± 01.70 with LoRA. Single-task full fine-tuning is strongest and improves over the best listed baseline under that protocol by 8.41 points.

Table 21: Complete Workload results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	50.00±00.00	52.36±01.25	30.70±00.55	BENDR	50.00±00.00	48.83±01.68	24.83±00.87
BIOT	63.98±00.66	77.68±01.62	**69.30±02.27**	BIOT	58.25±00.50	76.48±00.48	54.17±00.30
LaBraM	61.17±01.01	67.78±01.48	45.83±02.21	LaBraM	57.45±02.83	73.43±00.56	46.72±00.74
EEGPT	62.31±03.98	70.46±04.00	48.13±06.90	EEGPT	61.05±03.33	75.68±00.59	58.23±01.57
CBraMod	71.94±00.60	81.63±01.63	59.73±01.79	CBraMod	50.00±00.00	66.87±00.17	42.70±00.07
CSBrain	69.53±02.65	77.07±04.26	64.67±03.20	CSBrain	55.07±00.94	76.83±00.12	54.73±00.12
REVE	**74.12±01.01**	**83.47±01.74**	64.43±03.71	REVE	**67.10±00.00**	**79.83±00.17**	**61.20±00.85**
NSP	**82.53±00.30**	**87.80±01.68**	**81.73±03.71**	NSP	**75.77±00.99**	**79.73±00.12**	**68.50±01.78**
Model	Multi-task			Model	LoRA		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	63.32±02.29	74.09±01.04	54.78±01.84	BENDR	50.00±00.00	49.20±00.64	25.47±01.06
BIOT	71.37±02.19	72.43±01.94	69.52±00.23	BIOT	58.30±00.71	78.83±00.58	58.60±00.57
LaBraM	67.77±00.95	77.32±02.88	61.13±03.03	LaBraM	55.10±07.21	72.07±02.26	43.80±03.72
EEGPT	72.66±00.90	81.16±03.42	64.30±06.48	EEGPT	**76.87±00.37**	**87.63±01.09**	**76.43±01.52**
CBraMod	74.12±00.79	**83.89±00.93**	**72.58±01.55**	CBraMod	62.80±00.57	**81.63±00.90**	62.80±01.93
CSBrain	71.27±00.75	79.63±01.20	55.10±02.21	CSBrain	67.33±01.42	74.53±00.31	45.48±01.25
REVE	**74.67±01.46**	**82.32±01.47**	68.22±03.92	REVE	71.55±02.13	81.03±01.48	61.33±03.53
NSP	**74.28±01.80**	80.20±01.72	**71.62±01.73**	NSP	**76.13±01.70**	79.40±01.37	**72.87±01.97**

E.2.7 Visual Target Detection

Visual target detection evaluates stimulus-evoked EEG decoding. Things-EEG-2 measures whether the representation can retain short-latency visual information across a high-density 63-channel montage.

Things-EEG-2. Things-EEG-2 evaluates binary visual-target decoding from 63-channel, 5-s windows, with 41,570 samples from 10 participants (Gifford et al., 2022). NSP records 72.77 ± 01.06 BAC with single-task full fine-tuning, 62.50 ± 01.00 with multi-task full fine-tuning, 50.23 ± 00.06 with frozen transfer, and 60.13 ± 00.50 with LoRA. Single-task full fine-tuning performs best and exceeds the strongest baseline in that setting by 1.68 points.

Table 22: Complete Things-EEG-2 results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	**71.09±02.02**	**83.09±00.69**	**52.68±00.45**	BENDR	50.00±00.00	51.17±01.46	11.23±00.39
BIOT	51.10±00.30	62.05±01.53	18.05±01.06	BIOT	50.05±00.00	53.85±00.56	12.43±00.33
LaBraM	53.38±00.39	56.38±01.08	13.65±00.45	LaBraM	50.00±00.00	53.77±00.45	12.57±00.17
EEGPT	62.08±00.58	75.69±01.02	33.32±01.46	EEGPT	50.00±00.00	**59.57±00.42**	**14.18±00.15**
CBraMod	62.40±01.56	77.61±00.32	37.48±00.78	CBraMod	50.00±00.00	53.85±00.07	12.29±00.13
CSBrain	57.20±00.29	62.67±00.53	21.53±00.42	CSBrain	50.00±00.00	51.63±02.31	12.10±00.94
REVE	61.03±00.90	66.77±00.52	22.87±01.72	REVE	**50.47±00.13**	**58.18±00.06**	**14.92±00.06**
NSP	**72.77±01.06**	**89.53±01.72**	**55.67±00.06**	NSP	**50.23±00.06**	57.47±01.27	13.57±00.40
Model	Multi-task			Model	LoRA		
	BAC	AUROC	AUCPR		BAC	AUROC	AUCPR
BENDR	**63.62±01.38**	**77.23±00.88**	**42.29±01.75**	BENDR	50.00±00.00	50.27±00.12	11.03±00.05
BIOT	54.15±00.25	63.01±00.52	20.12±00.33	BIOT	50.67±00.05	**64.13±00.17**	**19.47±00.05**
LaBraM	56.47±00.32	62.82±00.84	22.23±00.64	LaBraM	50.00±00.00	55.27±02.33	13.13±01.13
EEGPT	51.54±00.76	66.64±00.04	20.30±00.65	EEGPT	51.70±00.37	61.83±00.21	14.97±00.31
CBraMod	59.13±00.72	69.58±01.41	31.85±01.21	CBraMod	50.23±00.05	57.60±00.79	14.17±00.31
CSBrain	57.49±00.27	62.47±01.27	24.66±01.93	CSBrain	50.00±00.00	54.43±00.62	13.03±00.21
REVE	59.43±01.20	69.00±00.53	32.07±01.85	REVE	**53.77±00.12**	60.70±00.24	15.70±00.22
NSP	**62.50±01.00**	**76.68±00.61**	**41.80±01.69**	NSP	**60.13±00.50**	**77.17±01.05**	**41.33±01.40**

E.2.8 NEURODEGENERATIVE DISEASE IDENTIFICATION

Neurodegenerative disease identification tests clinically relevant subject-level discrimination. ADFTD separates Alzheimer's disease, frontotemporal dementia, and healthy controls, requiring the representation to capture disease-related differences beyond transient event structure.

ADFTD. ADFTD is a three-class diagnostic task distinguishing Alzheimer's disease, frontotemporal dementia, and healthy controls from 19-channel, 10-s windows; it contains 7,013 samples from 88 participants (Miltiadous et al., 2023). NSP reaches 52.17 ± 00.24 BAC with single-task full fine-tuning, 56.66 ± 01.66 with multi-task full fine-tuning, 54.10 ± 00.95 with frozen transfer, and 50.83 ± 02.12 with LoRA. Multi-task full fine-tuning is strongest and leads the best listed baseline under that setting by 0.95 points.

Table 23: Complete ADFTD results across four adaptation settings (%, mean ± sample SD). **Bold red** and **bold blue** mark the best and second-best mean for each metric and protocol.

Full Fine-Tuning				Parameter-Efficient Multi-Task			
Model	Single-task			Model	Frozen		
	BAC	F1	κ		BAC	F1	κ
BENDR	37.16±02.62	40.38±01.69	07.27±04.67	BENDR	33.30±00.00	22.17±00.75	00.00±00.00
BIOT	47.43±01.03	48.17±01.38	23.77±01.52	BIOT	**51.37±00.09**	**50.47±00.05**	**29.47±00.21**
LaBraM	35.75±02.76	34.47±03.44	11.40±03.67	LaBraM	38.20±00.77	38.40±01.05	09.70±01.27
EEGPT	46.50±02.28	48.77±02.56	21.81±03.72	EEGPT	39.03±02.48	32.02±00.53	07.72±03.37
CBraMod	48.47±01.12	**51.68±01.11**	**31.17±01.74**	CBraMod	38.17±00.13	36.17±00.35	09.12±00.26
CSBrain	46.08±02.61	46.63±03.41	24.27±03.85	CSBrain	41.30±00.28	42.23±00.25	12.40±00.45
REVE	**50.22±00.41**	50.72±01.85	**29.21±02.08**	REVE	46.10±00.24	49.70±00.14	23.20±00.29
NSP	**52.17±00.24**	**52.67±01.97**	27.87±03.25	NSP	**54.10±00.95**	**57.53±02.08**	**37.57±00.55**
Model	Multi-task			Model	LoRA		
	BAC	F1	κ		BAC	F1	κ
BENDR	**55.71±03.76**	**56.97±02.75**	**33.98±04.63**	BENDR	33.87±00.80	18.77±07.78	00.83±01.18
BIOT	52.40±01.98	53.48±02.71	**36.27±03.65**	BIOT	48.77±01.40	51.33±01.88	25.27±01.93
LaBraM	50.27±01.74	52.36±02.23	26.17±03.15	LaBraM	43.07±00.66	45.77±00.76	18.40±01.28
EEGPT	52.91±01.57	54.82±01.30	31.14±01.85	EEGPT	44.93±01.18	45.80±01.31	17.73±01.70
CBraMod	51.95±02.84	54.41±03.13	30.06±04.35	CBraMod	48.83±00.50	**51.57±00.25**	26.27±01.86
CSBrain	48.96±02.65	50.17±03.76	24.96±04.13	CSBrain	48.60±00.45	47.20±00.57	27.33±00.74
REVE	52.12±03.78	51.10±03.90	30.70±06.41	REVE	**50.73±01.28**	24.73±02.80	**56.03±00.40**
NSP	**56.66±01.66**	**60.46±01.88**	32.00±03.66	NSP	**50.83±02.12**	**54.40±02.10**	**31.67±00.63**

F ADDITIONAL ABLATIONS

The pretraining-component ablations use frozen single-task evaluation: the encoder is fixed and a head is trained separately on each of TUAB, TUEV, SEED, and PhysioMI. Unless 100% is specified, heads use 10% of training labels. This differs from EEG-FM-Bench's frozen multi-task protocol. The final study instead varies the head under full fine-tuning.

F.1 CUMULATIVE COMPONENT ABLATION

To determine how transfer changes as prediction targets and available context are constrained, we compare the cumulative configurations and the state-regularization control in Table 3. Macro BAC rises from 44.44 for the shared-target baseline to 45.75 with block masking, 51.88 with EMA, 54.24 with identity residualization, and 55.70 for Full NSP. Task-wise changes are not uniformly positive, and removing regularization lowers the macro to 46.88. The matched context comparisons in Table 3 separately assess exclusion geometry at a fixed visible-token budget.

F.2 GUARD DESIGN

To examine how masking and context exclusion affect transfer, we compare random, spatial-only, and temporal-only exclusion with the complete topology-separated design on four tasks (Figure 9). The exact-temporal condition is a separate temporal-only masking baseline without EMA or residualization, not an isolated guard variant. Table 3 provides matched visible-token controls for the exclusion-geometry comparison.

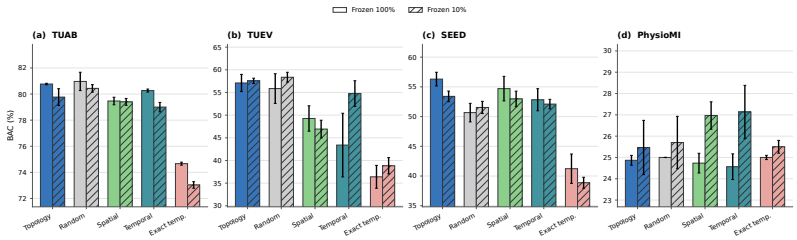

Figure 9: **Context exclusion and temporal-masking controls.** Frozen single-task BAC with 10% and 100% labels. Topology denotes NSP without identity residualization, whereas Random denotes NSP with random exclusion and residualization. Spatial-only and temporal-only exclusion modify available context; the exact-temporal condition is a separate temporal-only baseline without EMA or residualization. These controls are not all isolated changes in guard geometry.

F.3 Residualization design

To isolate which positional marginal is removed by the target construction, we compare no residualization with channel-only, time-only, and joint residualization under matched frozen single-task protocols (Figure 10). Channel and time removal produce different task-wise changes, and neither partial variant reproduces the complete joint configuration uniformly.

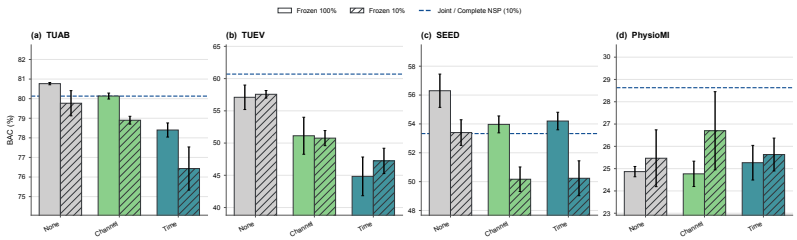

Figure 10: **Partial identity-residualization variants.** Frozen single-task BAC without residualization and with channel-only or time-only residualization using 10% and 100% labels. Dashed lines show complete joint residualization with 10% labels.

F.4 State regularization coefficient

To test sensitivity to moment-based state regularization, we vary λ_{reg} and evaluate the frozen encoder with single-task heads trained on either 10% or 100% of the available labels (Figure 11); these are not full-parameter fine-tuning runs. Under Frozen 100%, the four-task macro is 46.95 without regularization and 55.51, 55.16, 54.74, and 52.68 at $\lambda_{\text{reg}} = 0.025$, 0.05, 0.1, and 0.2, respectively. All tested nonzero weights improve on the unregularized setting in both protocols. The highest observed macro occurs at $\lambda_{\text{reg}} = 0.025$ with 100% labels and at $\lambda_{\text{reg}} = 0.1$ with 10% labels; task-wise responses differ, and these means do not establish a statistically unique optimum.

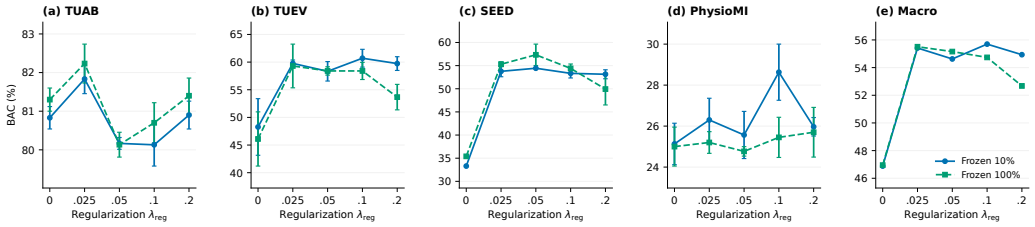

Figure 11: **Sensitivity to moment-based state regularization.** Single-task heads are trained with 10% or 100% of labels while the encoder remains frozen. Task panels show BAC with sample SD, and Macro is the unweighted mean of the four task means. Macro uncertainty is not inferred from task-wise SDs. The highest observed macro occurs at $\lambda_{\text{reg}} = 0.025$ with 100% labels and at $\lambda_{\text{reg}} = 0.1$ with 10% labels.

F.5 TARGET LAYER SELECTION

To identify which encoder depth supplies the most transferable latent target, we compare middle-layer, last-layer, and top-three-layer aggregation under both frozen single-task label budgets (Figure 12, left). Under Frozen 100%, the top-three target obtains 54.25 macro BAC, compared with 52.93 for the last-layer target and 51.24 for the middle-layer target.

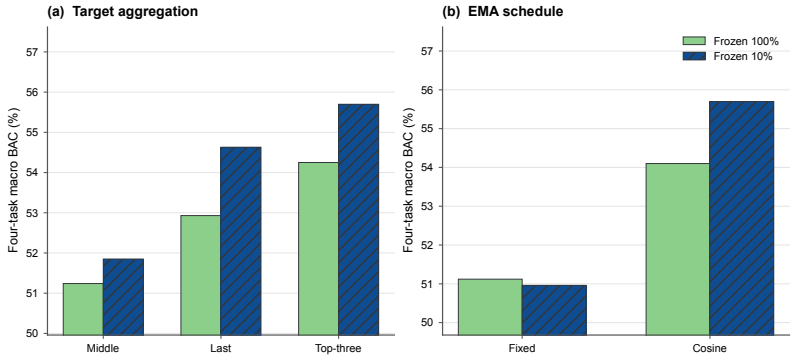

Figure 12: **Target aggregation and EMA schedule.** Aggregate balanced accuracy for middle-layer, last-layer, and top-three-layer targets (left), and fixed versus cosine EMA schedules (right). Bars compare frozen single-task evaluation with 100% and 10% labels.

F.6 EMA SCHEDULE

To test whether a progressively slower Target Encoder improves supervision, we compare fixed and cosine EMA schedules with the remaining configuration held constant (Figure 12, right). The cosine schedule is stronger under both evaluations: 54.10 versus 51.12 macro BAC with Frozen 100%, and 55.70 versus 50.96 with Frozen 10%.

F.7 TASK-WISE COMPARISON OF EMA TARGET AND FULL NSP

To verify that the aggregate component difference is not driven by a single task, we compare the EMA-target configuration with Full NSP on TUAB, TUEV, SEED, and PhysioMI (Figure 13). The task-wise view shows where the complete context design changes transfer rather than compressing the comparison into one macro score.

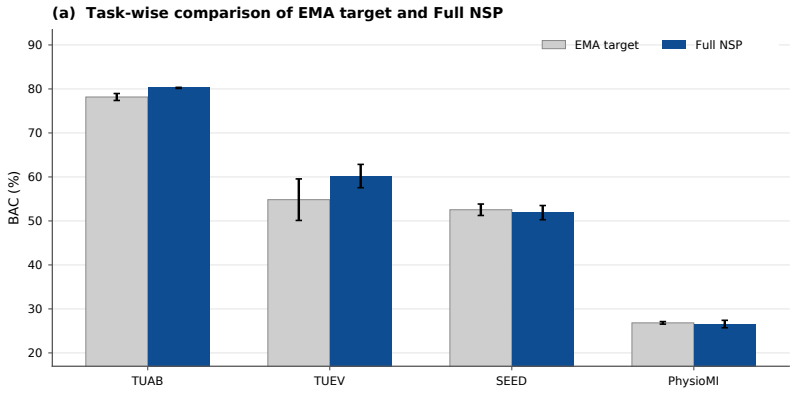

Figure 13: **Task-wise comparison of EMA-target and Full NSP configurations.** Frozen single-task BAC with 10% labels on TUAB, TUEV, SEED, and PhysioMI. The complete method improves some tasks while leaving others similar; error bars summarize variability across repeated evaluations.

F.8 Fine-tuning head design

To determine whether downstream conclusions depend on one readout design, we compare six classification heads while holding fine-tuning and model selection fixed (Figure 14). Global mean–max gives the highest four-task macro BAC (70.82), closely followed by channel mean–max (70.81), dual-stream fusion (70.61), and average pooling (70.43). The task-wise ordering is less uniform: average pooling has the highest mean on TUAB and SEED, global mean–max on TUEV, and channel mean–max on PhysioMI. The results therefore do not identify a uniformly dominant head; small aggregate differences should be interpreted together with the reported variability.

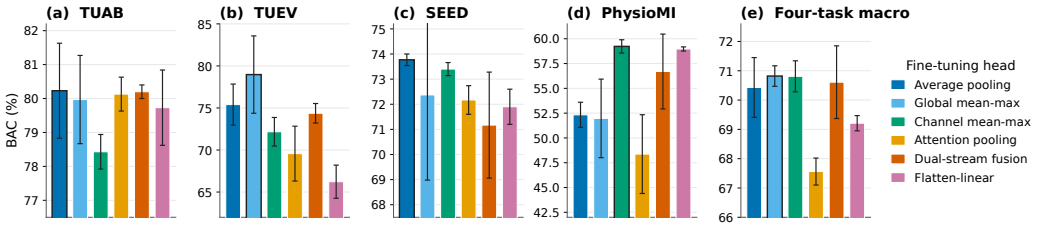

Figure 14: **Effect of fine-tuning head design.** Mean balanced accuracy on TUAB, TUEV, SEED, and PhysioMI and their four-task macro. Error bars show sample standard deviations; colors distinguish six head designs, and dark outlines mark the highest mean in each panel.

G Model Scale and Control Studies

G.1 Model scale

To examine the effect of encoder capacity, we pretrain Tiny (10.74M parameters) and Medium (24.11M) NSP models in addition to Large (42.83M), and evaluate the three scales with average pooling under single-task full fine-tuning (Figure 15). The heatmap covers all 14 EEG-FM-Bench datasets, and the adjacent panel averages their BAC means.

Aggregate trend favors capacity. The reported macro BAC rises from 61.76 for Tiny through 63.28 for Medium to 64.51 for Large. The increase is modest relative to the change in parameter count.

Task-level effects are heterogeneous. Figure 15 shows that Large performs best on HMC, TUAB, Workload, PhysioMI, Things-EEG-2, ADFTD, SEED-VII, and TUSL. Medium is strongest on TUEV, Siena, SEED-V, and Mimul-11, while Tiny leads on SEED and BCIC-2a. The aggregate increase therefore does not imply a uniform task-wise ordering.

Scope of the scale comparison. These independently pretrained models examine capacity under the scale-study protocol. Their full-fine-tuning results are not directly comparable with the frozen, limited-label component ablation in Table 3. The increase in macro BAC therefore characterizes this scale comparison, rather than quantifying a trade-off between model size and objective design.

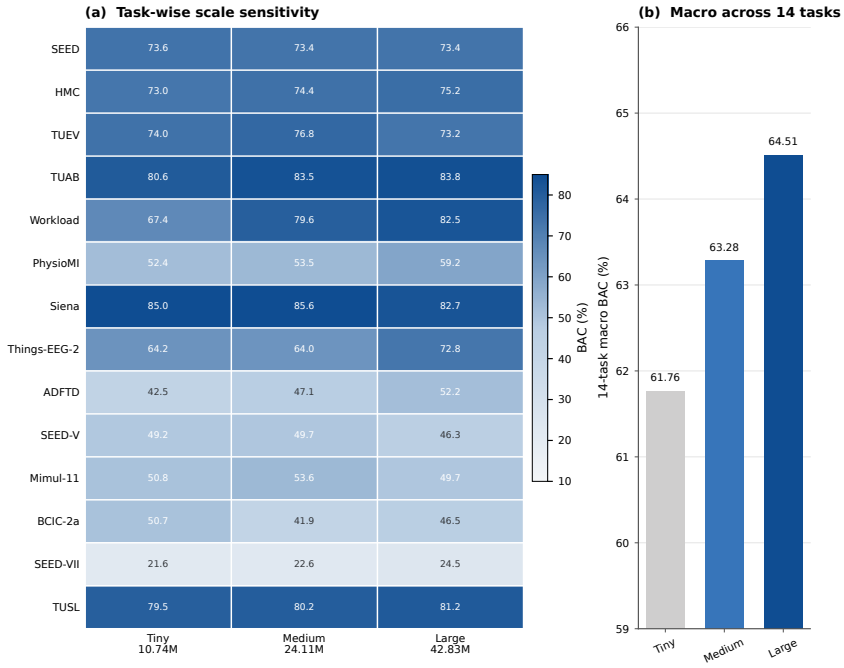

Figure 15: **Model scale produces task-dependent changes.** The heatmap compares balanced accuracy for Tiny, Medium, and Large encoders across the displayed downstream tasks; the bar chart summarizes the accompanying aggregate trend. Larger capacity improves the aggregate score, but the best encoder size varies by dataset.

G.2 Initialization controls

Figure 16 compares the downstream gain of NSP with two controls that disrupt learned parameters: random initialization and weight permutation. Most datasets retain a positive gain over both controls, indicating that transfer depends on learned parameter values and, for many tasks, on their organization within the encoder. The magnitude varies substantially by dataset, and a small number of tasks do not improve over both controls.

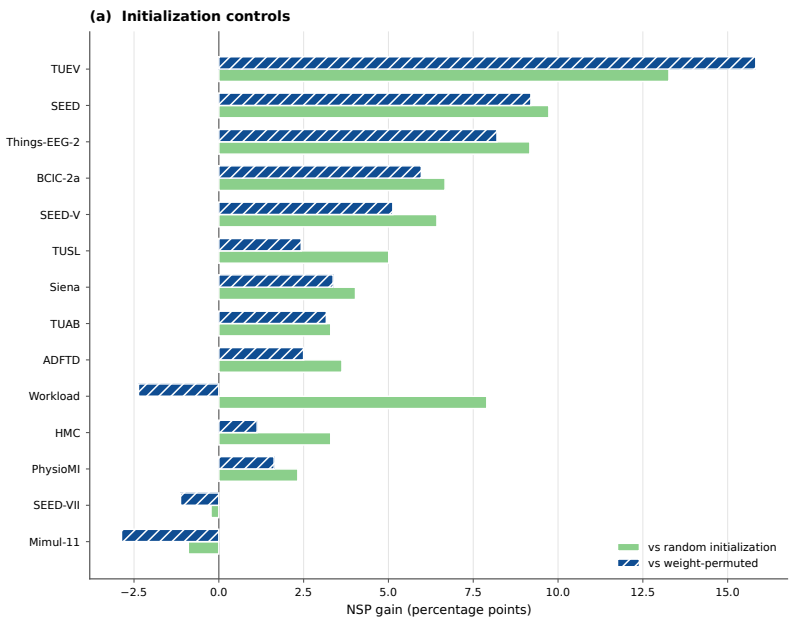

Figure 16: **Initialization and parameter-organization controls.** Dataset-wise NSP gain relative to a randomly initialized encoder and to a weight-permuted encoder. Positive values indicate an advantage for the learned NSP parameters; differences between the two controls show that both parameter values and their arrangement can affect transfer.

G.3 Effect of pretraining recording exclusion

To assess how excluding downstream recordings changes transfer, we compare pretraining with the original TUEG manifest and with a recording-disjoint manifest that removes the evaluated downstream recordings (Table 24). Under the same downstream protocol, BAC changes from 83.83 to 82.87 on TUAB, 73.17 to 71.83 on TUEV, and 81.17 to 79.88 on TUSL. The respective decreases of 0.96, 1.34, and 1.29 percentage points indicate some sensitivity to recording exposure, while much of the observed transfer performance is retained after exclusion. This comparison is defined by recording-level membership in the pretraining manifest.

Table 24: **Recording-disjoint TUEG pretraining control.** Balanced accuracy (%, mean ± sample standard deviation) on the three TUH-derived downstream datasets. The recording-disjoint manifest excludes the evaluated downstream recordings from pretraining.

Pretraining manifest	TUAB	TUEV	TUSL
Original TUEG	83.83±00.50	73.17±01.70	81.17±01.12
Recording-disjoint TUEG	82.87±00.94	71.83±01.53	79.88±00.76

H Mechanism Diagnostics

To complement downstream transfer results, we evaluate sensitivity to signal content and available context, and probe positional information retained by component variants. The interventions hold Target states fixed while changing Context input. They characterize prediction sensitivity and identity decodability, rather than demonstrate the elimination of position-only prediction or local interpolation.

H.1 Context interventions and identity probes

To distinguish signal sensitivity from identity decodability, we evaluate each displayed component variant with Context interventions and linear identity probes (Figure 17). The reference Context

masks target positions only; it is not an entirely unmasked input. For intervention a, relative cosine agreement is $R_a = C_a/C_{\text{ref}}$, where C_a averages prediction–target cosine over the masked target positions and C_{ref} is the same statistic for the reference Context. Targets are computed from the original recordings and held fixed. Zeroing removes EEG content, while sample mismatch cyclically shifts recordings along the batch dimension without permuting channels or time within a recording. Spatial and temporal interventions expand the Context mask around the same targets. Negative ratios indicate negative intervention cosine, not a negative fraction of retained information.

The probe panel reports channel and relative-time classification accuracy from full-input Context states and constructed Target states, with chance levels $1/19$ and $1/30$, respectively. Context channel accuracy remains high because residualization acts on Targets. Time-only residualization likewise leaves channel identity largely decodable; the temporal-only baseline has neither EMA nor residualization. The random-exclusion diagnostic includes residualization and corresponds to *NSP with random exclusion*, not the no-residualization random control in Table 3. The shared-target baseline and Full NSP endpoints are evaluated using the same diagnostic protocol: 512 evaluation segments for interventions and separate sets of 256 segments for probe training and testing, with 64 sampled tokens per segment. Sample-mismatch relative cosine changes from 0.894 to -0.028, while Target channel-probe accuracy changes from 100.00% to 14.43%. The latter remains above chance, supporting reduced, not eliminated, identity decodability. These are single-checkpoint diagnostics rather than estimates of variability across pretraining runs.

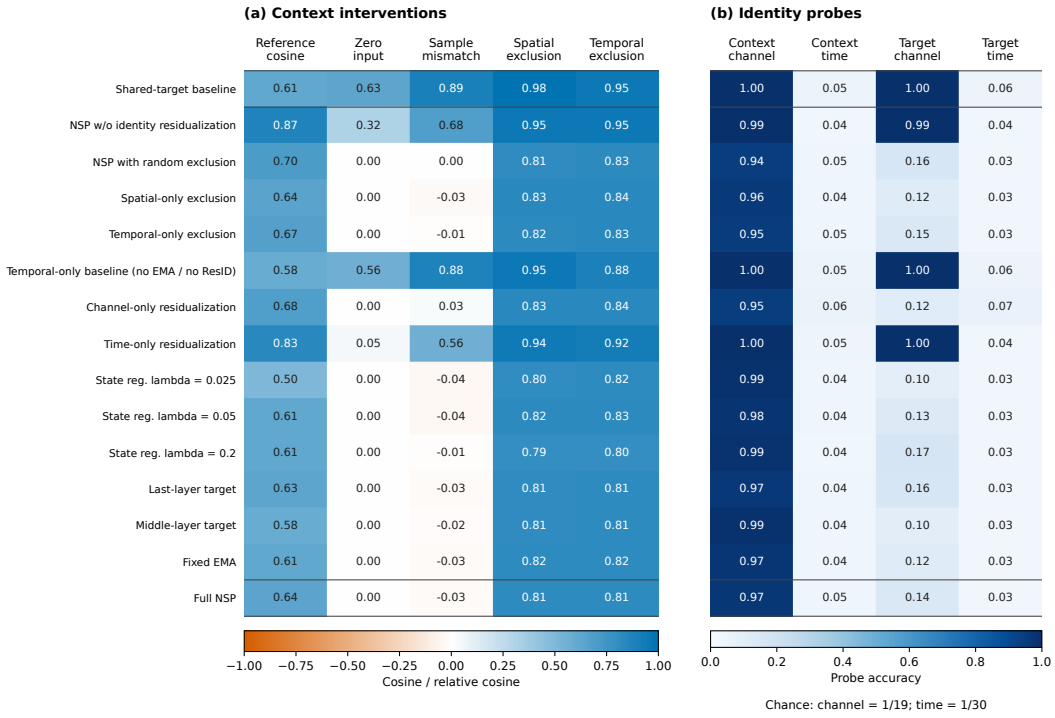

Figure 17: **Context interventions and identity probes from the shared-target baseline to Full NSP.** The first and last rows show the baseline and Full NSP, with component variants between them. (a) Reference cosine uses target-only masking; intervention columns show relative cosine agreement with unchanged Targets. Sample mismatch operates across recordings in a batch. (b) Channel/time probe accuracy for Context and constructed Target states; chance is $1/19$ for channel and $1/30$ for time. Random exclusion includes residualization.

H.2 Component-transfer overview

Figure 18 reports frozen single-task BAC with 10% and 100% labels for guard, residualization, regularization, target-layer, and EMA variants on TUAB, TUEV, SEED, and PhysioMI. The side-by-side label budgets show the magnitude and direction of each component's effect by dataset.

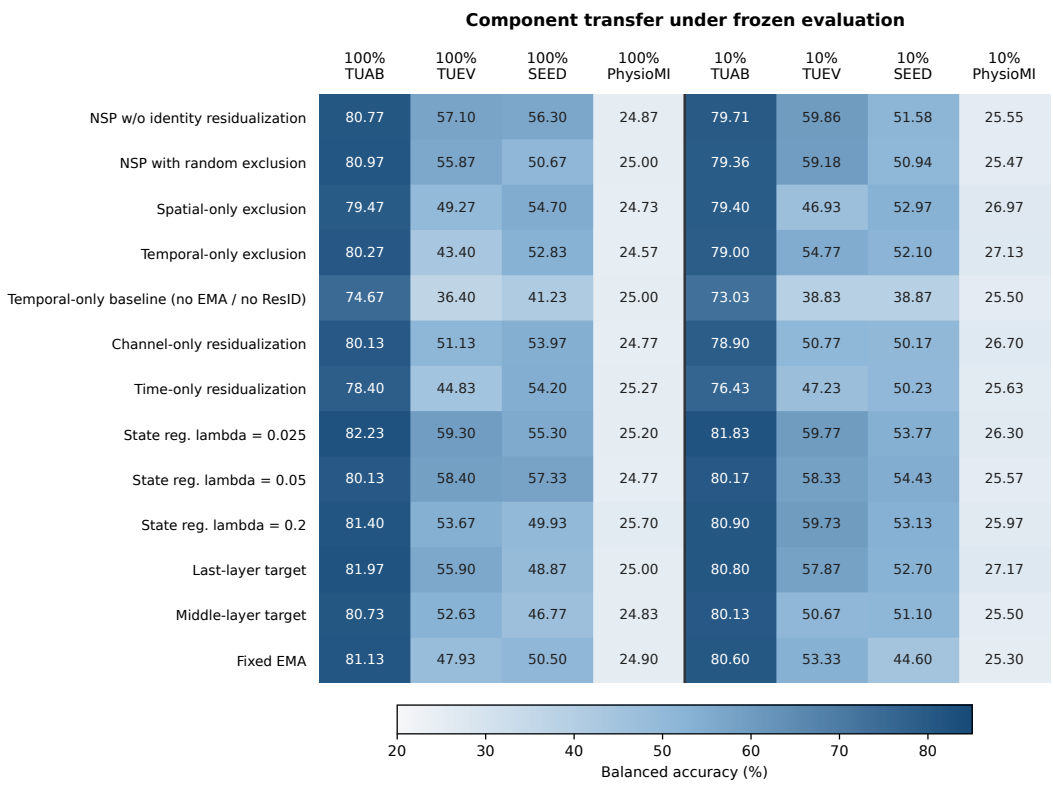

Figure 18: **Complete component-transfer overview.** Frozen single-task BAC for guard, residualization, regularization, target-layer, and EMA variants on TUAB, TUEV, SEED, and PhysioMI. The two column groups use 100% and 10% labels, respectively, and show strongly task-dependent component effects.

I Additional Visualizations

I.1 t-SNE on additional tasks

Beyond the HMC example in the main paper, Figures 19 and 20 compare raw EEG, frozen NSP, and fully adapted NSP representations across ten downstream tasks using t-SNE (van der Maaten & Hinton, 2008).

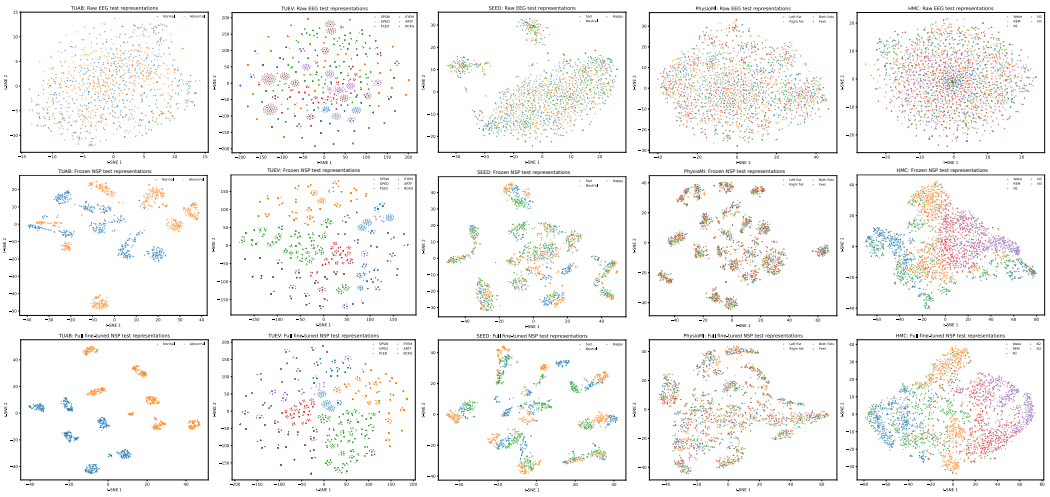

Figure 19: **Representation geometry across five downstream tasks.** Columns show TUAB, TUEV, SEED, PhysioMI, and HMC; rows show raw EEG, frozen NSP, and fully adapted NSP representations.

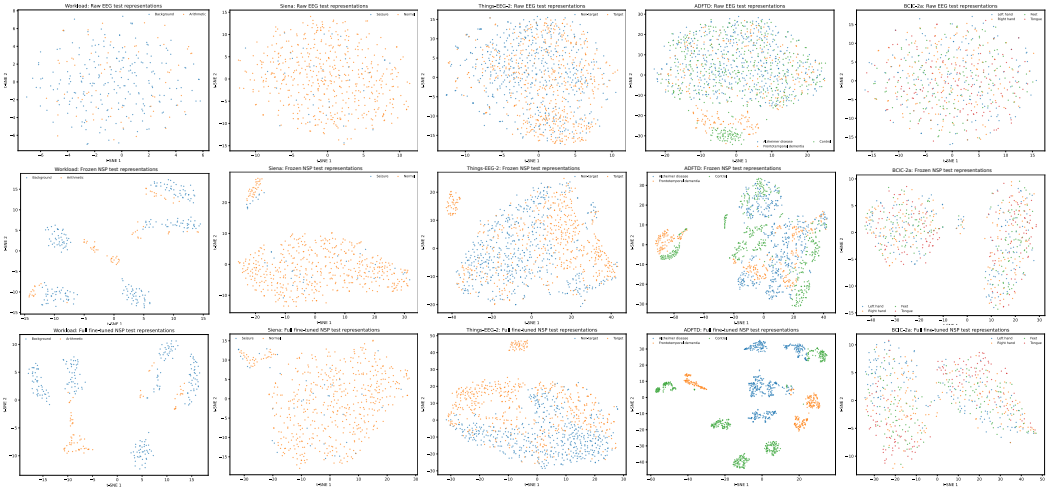

Figure 20: **Representation geometry across additional paradigms.** Columns show Workload, Siena, Things-EEG-2, ADFTD, and BCIC-2a; rows show raw EEG, frozen NSP, and fully adapted NSP representations. Each projection is fit independently and illustrates task-specific changes in class geometry.

I.2 Grad-CAM topology maps

We compute class-conditioned Grad-CAM attribution (Selvaraju et al., 2017) for the 14 EEG-FM-Bench datasets, each summed over time to a per-channel map. Figure 21 shows spatially structured, task-conditioned patterns spanning clinical (TUAB, TUEV, TUSL), sleep (HMC), seizure (Siena), motor-imagery (PhysioMI, BCIC-2a), emotion (SEED), workload (Workload), and visual (Things-EEG-2) tasks. These maps are descriptive and do not identify physiological generators or functional connectivity.

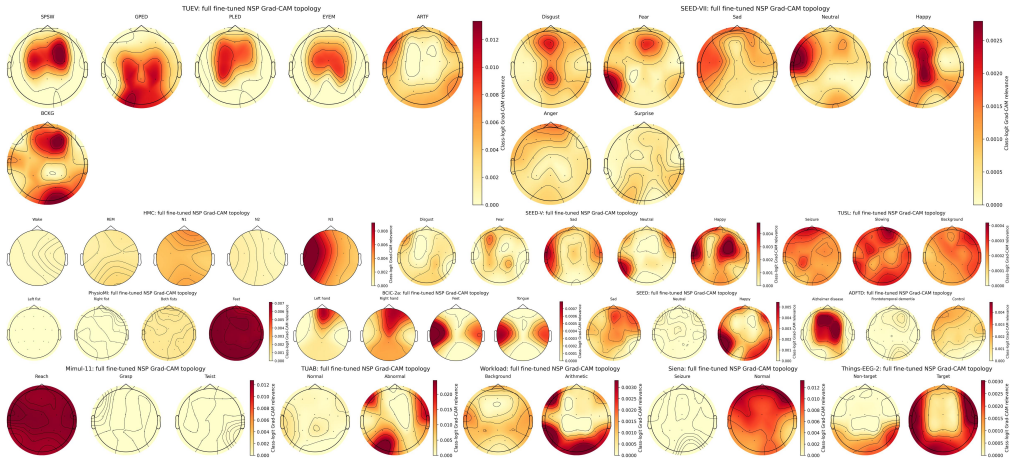

Figure 21: **Class-conditioned Grad-CAM topology maps across 14 EEG-FM-Bench datasets.** Each dataset group contains one topographic attribution map per class, obtained after summing relevance over time. The maps are descriptive and task conditioned; quantitative transfer claims rest on the controlled benchmark evaluations.

I.3 Interchannel correlation structure

To examine task-conditioned interchannel correlations, we display attention-derived correlation scores from task-adapted NSP as chord diagrams across the 14 EEG-FM-Bench datasets (Figure 22). Channel labels follow each dataset's source montage and the benchmark adapter's channel mapping. Ribbon width represents the selected interchannel correlation score, and channels in the same anatomical region share a color.

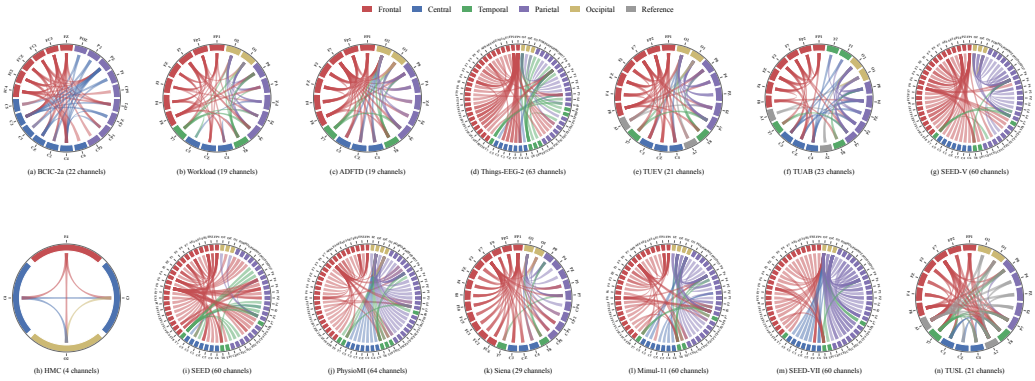

Figure 22: **Interchannel correlation diagrams across 14 EEG-FM-Bench datasets.** Source-montage channel labels are associated with model channels through the benchmark adapter mapping. Ribbon width encodes selected attention-derived interchannel correlation scores, and color groups channels by anatomical region.