FROM STATIC DESCRIPTIONS TO STATE DYNAMICS: MULTI-SCALE PRETRAINING FOR EEG TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) is a temporal signal whose task-relevant structure spans transient events and sustained rhythms. Conventional temporal models learn these features through task-specific supervision, while EEG foundation models seek broader transfer through large-scale unlabelled pretraining. Many use masked content recovery, which can exploit temporal continuity and spatial redundancy without requiring predictions of subsequent activity. Autoregressive EEG approaches introduce future prediction, but typically supervise the next token or chunk without explicitly coupling fine-resolution and longer-range forecasts from the same state. We introduce MSDE, which jointly predicts future teacher states at micro and macro scales through a causal encoder and a shared transition operator conditioned on elapsed time. With a 5.0M-parameter encoder pretrained on 2,115 hours of EEG, MSDE-10% achieves the second-best average fine-tuning rank across 20 OmniEEG-Bench tasks, behind only MSDE-Full among eight evaluated models. Matched controls favour future prediction over content recovery, and joint supervision exceeds both single-scale variants on 18 tasks. Transfer gains also extend to cross-dataset sleep staging without target adaptation. Frozen-state readouts reveal improved access to task-relevant physiological features, and component interventions link these features to classification gains. Spectrum-preserving shifts on FACED further show dependence on temporal alignment beyond whole-window power.

1 INTRODUCTION

Electroencephalography (EEG) records brain electrical activity as a continuous time series, reflecting the combined currents of neuronal populations (Buzsáki et al., 2012). Its fluctuations carry physiological structure: events can elicit transient potentials and modulate the power of ongoing, frequency-specific rhythms (Pfurtscheller & Lopes da Silva, 1999). During sleep, K-complexes reflect transient cortical down-states, while spindles are brief oscillatory bursts arising from thalamocortical activity (Cash et al., 2009; Fernandez & Lüthi, 2020). The duration and rhythmic organisation of a spindle are complementary aspects of the same event. EEG physiology therefore involves both the evolution of local events and the ongoing activity in which they occur.

These physiological properties are represented in the datasets used to evaluate EEG models (Lu et al., 2026). ErrP distinguishes responses to correct and erroneous observed actions, making the timing and succession of response components relevant to decoding (Chavarriaga & Millán, 2010; Chavarriaga et al., 2014). In BETA, responses persist during visual stimulation, and decoding identifies the attended stimulus from the evoked rhythms (Liu et al., 2020; Norcia et al., 2015). SleepEDF extends the context to sleep epochs, where transient events and ongoing rhythmic activity contribute to stage identification (Kemp et al., 2000). Across these settings, useful representations need to preserve different temporal details and integrate the surrounding activity relevant to the task.

To model these task-dependent temporal patterns, supervised EEG models tailor feature extraction and context aggregation to the target decoding task. EEG Conformer combines local convolutions with temporal self-attention for motor-imagery and emotion classification (Song et al., 2023), while SeqSleepNet models within-epoch and across-epoch dependencies for sleep staging (Phan et al., 2019). Task-specific labels guide which temporal features these models retain. Transfer across heterogeneous EEG applications motivates a broader learning signal that is not tied to any one task.

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.

Recent EEG foundation models pursue this goal through large-scale unlabelled pretraining, drawing on approaches developed in language and vision (Kaplan et al., 2020; Hoffmann et al., 2022; He et al., 2022). LaBraM, CBraMod and REVE recover masked EEG patches or codes, achieving strong transfer across clinical and brain-computer interface (BCI) tasks (Jiang et al., 2024; Wang et al., 2025a; El Ouahidi et al., 2026). Architectures also accommodate multiple spatiotemporal scales and recording modalities (Zhou et al., 2025; Xiao et al., 2025). Access to multiple scales, however, does not itself require future prediction across them. With masked modelling, temporal continuity and spatial redundancy can make withheld content recoverable from nearby observations (Kostas et al., 2021; El Ouahidi et al., 2026). Successful recovery therefore need not require encoding how activity develops beyond the observed context. Predicting subsequent activity from the past makes this temporal development an explicit target of pretraining.

Autoregressive EEG models use this predictive approach: Neuro-GPT forecasts subsequent chunk embeddings, while BrainGPT predicts the next signal token (Cui et al., 2024; Yue et al., 2024). However, next-unit prediction at a chosen granularity, even with varying context lengths, does not explicitly couple fine-resolution and longer-range forecasts. This matters for transfer: following transient event progression calls for fine temporal detail, whereas assessing rhythmic persistence requires tracking activity over longer intervals.

We therefore propose MSDE (Multi-Scale State-Dynamics EEG pretraining). MSDE jointly predicts future states at micro and macro scales from a shared EEG state summarising the observed history. This predictive objective encourages the shared state to preserve information about how EEG activity develops over time, with supervision supplied by future states from unlabelled recordings. Forecasting at fine temporal resolution and longer horizons encourages retaining complementary information for downstream tasks with different temporal demands. Requiring both forecasts from the same state can make this information available together, as when sleep staging draws on local events and surrounding activity.

Our contributions are:

- **Multi-scale future-state pretraining.** We introduce MSDE, which trains a shared causal EEG state to support micro- and macro-scale forecasts through one elapsed-time-conditioned transition operator.
- **Transfer across tasks and datasets.** MSDE-Full and MSDE-10% have the two best average fine-tuning ranks across 20 tasks. MSDE-10% exceeds LaBraM, CBraMod, both REVE variants and Neuro-GPT by 2.7–8.0 mean balanced-accuracy points, using roughly $1/14$ of REVE-B's encoder parameters and $1/29$ of its source EEG (Figure 2). Gains extend to cross-dataset sleep staging without target adaptation.
- **Physiological information and temporal alignment.** Frozen-state readouts and component interventions link the transfer gains to task-relevant physiological information. Spectrum-preserving FACED shifts further show that downstream decisions depend on temporal alignment beyond whole-window spectral power.

2 Related work

EEG foundation models achieve strong transfer across tasks through pretraining on large unlabelled EEG corpora. Many use masked modelling: LaBraM predicts discrete codes from a spectral tokenizer, while CBraMod and REVE reconstruct signal patches (Jiang et al., 2024; Wang et al., 2025a; El Ouahidi et al., 2026). REVE's block masking reduces redundant visible context (El Ouahidi et al., 2026). Latent-target approaches replace or supplement signal reconstruction: BENDR discriminates masked representations against negatives (Kostas et al., 2021), EEGPT combines teacher-feature alignment with signal reconstruction, and EEG-JEPA predicts contextual teacher features under structured masking (Wang et al., 2024; Assran et al., 2023; Li et al., 2026). Alongside target design, CSBrain combines cross-scale spatiotemporal tokenisation and sparse attention with masked waveform reconstruction (Zhou et al., 2025), while BrainOmni pairs sensor-aware tokenisation of EEG and magnetoencephalography with masked discrete-code prediction (Xiao et al., 2025). Temporal supervision also varies: temporal-context tasks supervise relative timing or sequence order (Banville et al., 2021), while contrastive predictive coding predicts future latent representations from preceding context (Oord et al., 2018). Neuro-GPT predicts subsequent chunk embeddings from growing

prefixes (Cui et al., 2024); BrainGPT predicts the next signal token with an electrode-wise causal encoder and combines electrode representations for transfer (Yue et al., 2024). MSDE jointly predicts token-level and patch-end states from an exponential moving average (EMA) teacher, using a shared causal state and one transition operator conditioned on elapsed time.

3 MULTI-SCALE STATE-DYNAMICS PRETRAINING

3.1 METHOD OVERVIEW

MSDE jointly predicts future teacher states at micro and macro scales from a shared causal EEG state. Following the predictive-state principle (Shalizi & Crutchfield, 2001), the encoder is trained to retain historical information useful for both prediction scales. Pretraining jointly learns the encoder and a shared transition operator from unlabelled EEG (Figure 1).

The encoder maps an input window to causal states s_t, each summarising observations up to t (Section 3.2). At sampled source times, the operator takes s_t and an elapsed time h to predict $\bar{s}_{t+h}$. An EMA copy of the encoder processes the same window causally to supply this target, using observations up to $t+h$. Micro-scale prediction samples token-state source–target pairs; macro-scale prediction samples patch-end pairs from the same sequence (Section 3.3).

Mean squared prediction errors backpropagate through the operator and student encoder; the teacher receives no gradients and tracks the encoder by EMA. Identity anchoring and variance–covariance regularisation encourage discriminative, varied states, while semigroup consistency aligns direct and composed prediction paths (Section 3.4). After pretraining, the teacher and operator are discarded. A classifier reads mean-pooled non-padding encoder states; fine-tuning (FT) updates both modules, while linear probing (LP) trains only the classifier.

3.2 CAUSAL STATE ENCODER

Given EEG history H_t, the encoder produces $s_t = e_\theta(H_t) \in \mathbb{R}^D$. Input windows are split into one-second patches; per-channel convolutions yield seven tokens per patch from overlapping signal segments, timestamped at kernel ends. Electrode embeddings identify the available channels, and channel-attention layers aggregate their tokens into a learned summary token at each time, accommodating different channel configurations. Causal temporal-attention layers combine these summaries using rotary positions (Su et al., 2024) to produce s_t at each token time, integrating preceding activity without access to observations after t. The final encoded state of each patch also serves as its patch state (Appendix B.3).

3.3 SHARED STATE TRANSITIONS ACROSS TIMESCALES

The transition operator uses a residual update conditioned only on s_t and h:

$$\Phi(s_t, h) = s_t + f_\mu\big([\mathrm{LN}(s_t) \,\|\, \tau(h)]\big), \tag{1}$$

where LN denotes layer normalization, τ embeds elapsed time and f_μ is a multilayer perceptron (MLP) shared across recordings and horizons.

Micro prediction uses token-state pairs at offsets of 0.5–1.0 s, while macro prediction uses patch-end pairs at 1, 2 and 4 s. The corresponding losses, $\mathcal{L}_{\mathrm{micro}}$ and $\mathcal{L}_{\mathrm{macro}}$, minimise mean squared error (MSE) between predicted states and stop-gradient EMA targets. Both losses jointly train the same encoder and transition operator, requiring the shared state to support fine-grained and longer-horizon future prediction.

3.4 INFORMATIVE STATES AND CONSISTENT TRANSITIONS

Future-state prediction jointly learns a state space and a transition rule. Because the EMA teacher follows the student, low prediction error alone does not ensure that states distinguish different EEG windows. A constant encoder, a matching teacher and a transition operator that preserves the constant state form a trivial solution with zero prediction loss and collapsed representations. We therefore complement prediction with constraints on state informativeness and consistency between prediction paths.

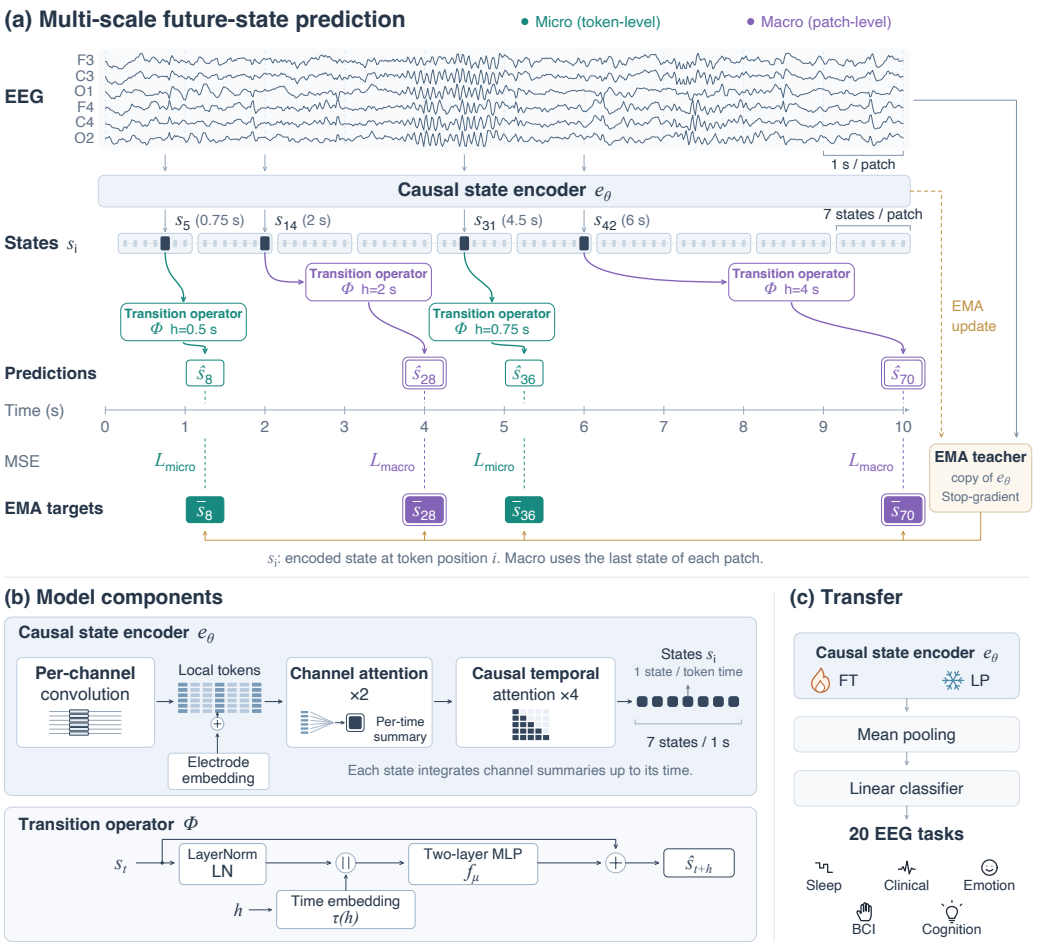

Figure 1: **Multi-scale future-state prediction.** **(a)** Micro- and macro-scale pairs share the encoder and transition operator. Predictions are hollow; stop-gradient EMA targets are solid, with double outlines for macro pairs. **(b)** Encoder and residual transition. **(c)** Mean pooling and linear classification: FT updates the encoder (flame); LP freezes it (snowflake). EEG: ISRUC-S3 stage 2 non-rapid-eye-movement sleep (N2) (Khalighi et al., 2016).

To discourage representation collapse during pretraining, we introduce identity anchoring $\mathcal{L}_{\text{inst}}$ to distinguish different EEG windows. It applies symmetric InfoNCE to the normalised means of non-padding states from the two temporal halves of each window, computed from a single student encoding. The two halves form a positive pair, with other windows in the batch as negatives. This contrastive constraint acts on normalised half-window representations and does not directly control variation among full-window means. These pooled states are read by the downstream classifier, motivating an explicit constraint on their variability. We therefore add variance–covariance regularisation $\mathcal{L}_{\text{vc}}$, which penalises low feature variance and off-diagonal covariance in both non-padding token states and pooled window states (Bardes et al., 2021).

These terms constrain the states supplied to the transition operator, while its prediction paths introduce a further requirement. Direct future prediction starts from an encoded state; a composed prediction also uses a predicted state as an intermediate input. Semigroup consistency $\mathcal{L}_{\text{sg}}$ encourages the two routes to agree by applying Smooth L1 between $\Phi(s_t, a+b)$ and $\Phi(\Phi(s_t, a), b)$, for $a, b > 0$ with $a + b \leq 1$ s.

The complete pretraining objective is

$$\mathcal{L} = \mathcal{L}_{\text{micro}} + \mathcal{L}_{\text{macro}} + 3\mathcal{L}_{\text{inst}} + 2\mathcal{L}_{\text{vc}} + 0.1\mathcal{L}_{\text{sg}}. \tag{2}$$

Implementation details and supporting analysis are provided in Appendices B.3 and J.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We pretrain on the Temple University Hospital (TUH) EEG corpus (Obeid & Picone, 2016), excluding all recordings from the downstream datasets. MSDE-10% and its ablations share a 2,115 h subset and base protocol; MSDE-Full uses the full cleaned corpus (Appendix B.1). All ablations and subsequent cross-dataset and physiological analyses use MSDE-10%, denoted MSDE below. We evaluate FT and LP on 20 OmniEEG-Bench tasks (Lu et al., 2026) spanning sleep, clinical, emotion, BCI and cognition-related settings. Comparators are released Neuro-GPT (Cui et al., 2024), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025a) and REVE-B/L (El Ouahidi et al., 2026) checkpoints, plus MSDE-Scratch as a same-architecture FT reference. Neuro-GPT provides a temporal-prediction baseline; both its EEG encoder and GPT are retained for FT and LP. We report balanced accuracy with shared subject-level splits and validation-based model selection (Appendices A, B.2 and C).

4.2 A SHARED PRETRAINED ENCODER TRANSFERS ACROSS TASKS

We first evaluate the encoder without pretraining to assess its supervised decoding capacity. Under the shared protocol, Scratch reaches 48.1% mean balanced accuracy, above pretrained CBraMod (46.8%) and close to LaBraM (49.4%; Table 1). This competitive baseline may reflect its temporal granularity: where LaBraM and CBraMod use one-second channel–patch tokens (Jiang et al., 2024; Wang et al., 2025a), MSDE aggregates channels into seven states per second, exposing within-second progression to causal temporal attention.

With the architecture fixed, pretraining adds 6.7 points; the 95% confidence interval (CI) is $[4.4, 9.3]$. The 5.7-point median and gains on 19 of 20 tasks support a reusable initialisation across tasks (Table 8). Among eight FT models, MSDE-Full and MSDE-10% have the two best average ranks (2.35 and 3.00); the 10% variant exceeds LaBraM, CBraMod, REVE-B and REVE-L on 14, 17, 13 and 13 tasks, respectively (Table 10). MSDE-10% exceeds the temporal-prediction baseline Neuro-GPT by 6.2 points, with gains on 17 of 20 tasks (Table 8). In Figure 2a, MSDE-10%'s 5.0M-parameter encoder reaches 54.8% using 2,115 h of source EEG, exceeding REVE-B by 2.7 points with roughly $1/14$ of its encoder parameters and $1/29$ of its source data. Full-corpus pretraining further raises mean FT accuracy to 56.1% (Table 1).

Fine-tuning allows the feature extractor to adapt to task labels, so we also evaluate LP with the pretrained feature extractor frozen. MSDE-Full and MSDE-10% attain the two highest mean LP accuracies (45.3% and 43.9%), with average ranks of 2.95 and 3.40 among seven pretrained models (Figure 2b; Table 10). MSDE-10% exceeds Neuro-GPT by 3.9 points, with gains on 16 of 20 tasks. REVE-L has the best average LP rank (2.80); task-wise differences between MSDE-10% and either REVE variant are not significant (Table 9).

4.3 FUTURE PREDICTION IMPROVES ON CONTENT RECOVERY

MSDE differs from masked-recovery pretraining by making future-state prediction its central learning signal. We first test its contribution by removing the prediction losses while retaining identity anchoring and variance–covariance regularisation. MSDE improves mean FT balanced accuracy over this ablation by 4.6 points ($p = 2.4 \times 10^{-5}$), with gains on 18 tasks (Figure 3a; Table 6). We next ask whether conventional masked reconstruction can provide an equally effective training signal. Keeping the causal encoder, 10% pretraining subset and base training protocol fixed, we replace future prediction with raw-signal reconstruction following the masked autoencoder (MAE) objective (He et al., 2022). MSDE outperforms this control at all three mask ratios (0.25, 0.5 and 0.75), with mean gains of 4.3–5.5 points. We also test masked latent infilling, which zeros selected input patches and predicts EMA-teacher states at the masked positions from the visible past. This control trails MSDE by 4.1 points ($p = 6.0 \times 10^{-4}$; Appendix B.1). Its deficit indicates that moving supervision into latent space is insufficient to explain MSDE's transfer advantage. Future prediction requires the same current state to anticipate activity at several later offsets, explicitly rewarding information about how EEG evolves from the observed past. This requirement offers an explanation for the transfer advantage over masked recovery.

Table 1: Balanced accuracy (%): full FT / LP averaged over three split seeds (two for PD31). MSDE-10% uses 10% of the pretraining corpus; MSDE-Full uses all of it. Bold and underline mark the best and second-best means before rounding within each protocol.

	Task	LaBraM	CBraMod	REVE-B	REVE-L	Neuro-GPT	MSDE-Scratch	MSDE-10%	MSDE-Full
		Full FT / LP	Full FT / LP	Full FT / LP	Full FT / LP	Full FT / LP	Full FT	Full FT / LP	Full FT / LP
Sleep	ISRUC-S3	74.8/50.4	57.7/39.9	84.5/77.5	83.7/**78.3**	71.0/57.3	65.3	80.7/70.7	**90.2**/74.6
	SleepEDF	74.8/68.1	71.7/64.4	74.1/68.5	72.4/**71.7**	70.2/62.9	73.4	74.7/67.0	**75.6**/65.2
	HMC	67.2/49.5	66.9/43.0	56.5/43.5	58.6/33.5	60.7/37.0	67.4	**69.8**/**52.3**	69.7/51.4
	ISRUC-S2	21.9/28.1	25.6/26.1	24.1/34.5	23.7/**40.6**	**30.0**/39.9	26.1	26.3/26.1	25.6/30.9
Clinical	TUEV	59.7/**47.4**	57.7/25.8	61.7/39.5	55.2/28.9	**63.3**/37.5	48.4	54.4/38.1	54.6/41.5
	AD65	55.5/50.5	53.7/47.9	54.7/48.8	58.3/53.7	55.7/45.6	44.6	65.2/51.7	**66.6**/**59.1**
	PD31	46.8/43.6	37.2/**51.3**	43.3/42.9	46.0/47.2	52.1/43.1	44.5	52.3/44.5	**58.5**/46.2
Emotion	FACED	21.5/16.6	34.8/11.1	51.6/16.2	47.3/19.6	43.3/14.5	45.8	61.3/**19.7**	**61.5**/17.9
	SEED	61.5/50.1	57.3/43.7	60.9/55.6	60.5/37.7	51.7/47.9	50.8	57.2/**59.1**	**62.0**/54.1
	SEED-IV	33.2/30.8	25.7/27.8	33.0/29.7	30.6/28.6	26.1/27.0	28.7	32.9/**32.7**	**33.4**/30.7
	SEED-V	26.1/24.7	26.5/22.6	24.6/23.5	24.6/23.7	28.5/23.8	25.6	32.6/29.0	**32.7**/**29.4**
	SEED-VII	19.1/**20.2**	19.1/17.9	**22.0**/19.6	21.9/20.1	17.6/17.5	19.0	19.5/18.8	21.2/16.8
BCI	BCIC-IV-2a	37.3/27.3	30.2/26.0	41.8/31.4	44.3/**43.2**	36.1/33.0	41.7	**49.8**/34.6	48.6/38.1
	SHU-MI	51.6/50.6	50.8/50.7	62.4/**55.0**	56.5/53.7	54.6/50.8	**63.9**	63.3/50.7	62.7/52.3
	BCIC-IV-1	50.7/50.2	55.3/50.5	54.2/51.5	53.7/53.2	52.0/53.7	52.3	**57.5**/53.3	55.2/**56.0**
	BETA	31.7/3.6	26.6/2.9	**38.4**/3.6	35.3/**3.6**	17.2/2.8	26.5	31.9/3.5	34.8/3.2
	ErrP	56.4/52.5	52.6/50.5	62.5/52.0	56.8/**63.2**	56.3/49.7	57.5	**71.3**/55.5	70.3/56.2
Other	EEGMAT	73.9/50.0	62.1/50.0	69.9/60.0	**74.9**/53.4	62.8/49.7	64.2	72.5/58.3	72.2/**64.2**
	LEMON-sex	**75.8**/63.0	73.6/57.1	70.8/66.5	73.5/**70.2**	71.6/57.4	65.3	70.2/63.5	71.1/67.0
	LEMON-extr.	48.8/50.7	49.9/50.8	50.9/51.4	51.1/**51.6**	50.8/50.4	50.5	52.3/50.0	**54.9**/50.3
	Mean (20)	49.4/41.4	46.8/38.0	52.1/43.6	51.4/43.8	48.6/40.1	48.1	54.8/43.9	**56.1**/**45.3**

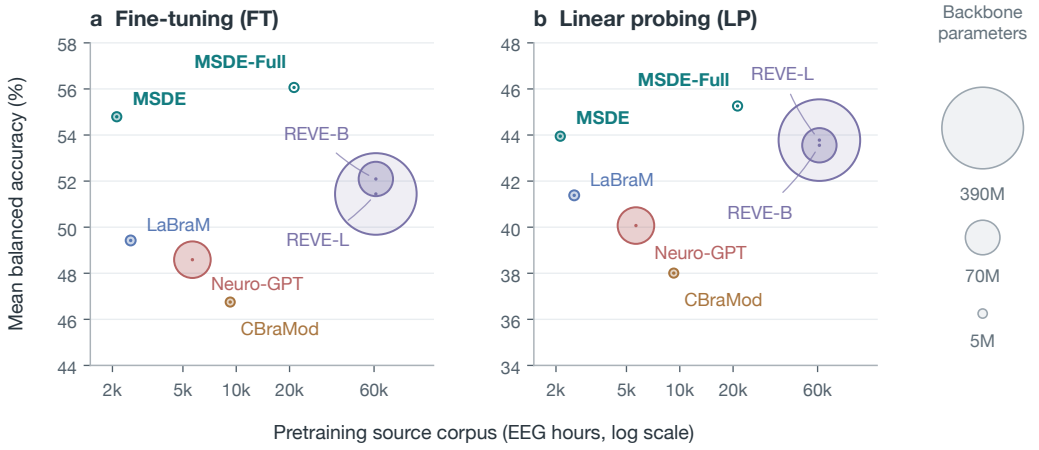

Figure 2: Pretraining scale versus mean balanced accuracy across 20 tasks: (a) FT; (b) frozen-backbone LP. Bubble area denotes backbone parameters on a shared scale (Appendix D).

For this predictive signal to remain useful, the learned states must retain variation: collapsed teacher targets would make prediction trivial. We therefore examine how the auxiliary losses sustain informative states and regularise their transitions. The no-variance–covariance variant scores 2.7 points below MSDE (Appendix C.1). Identity anchoring further distinguishes EEG windows: removing it costs 3.6 points, while overweighting it also hurts transfer, indicating that sample discrimination should complement the prediction objective without dominating it. Beyond maintaining informative states, semigroup consistency aligns direct and composed short-offset predictions; removing it reduces accuracy by 1.9 points.

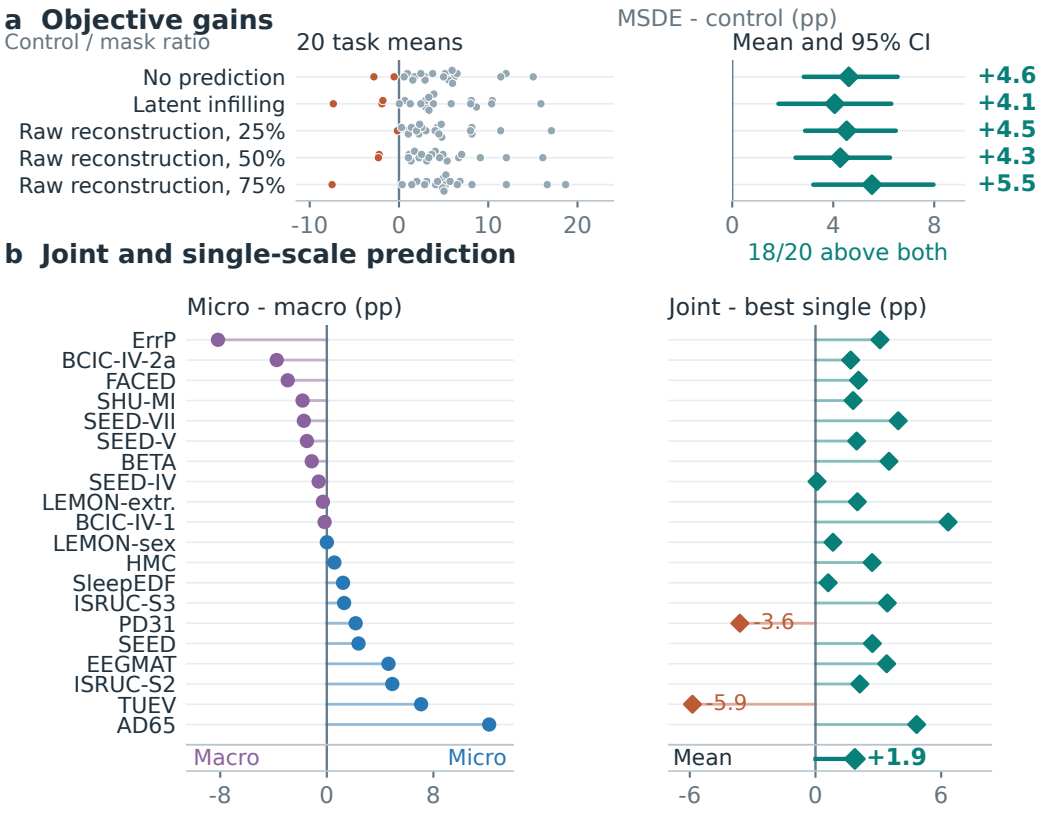

Figure 3: Objective and scale comparisons under FT (balanced-accuracy points). (a) Task gains over controls (left) and equal-task means with 95% task-bootstrap CIs (right). (b) Micro-only minus macro-only (left); joint minus the stronger single-scale reference, selected post hoc from each task's test means (right). The bottom diamond denotes the equal-task mean.

4.4 Joint scales outperform task-wise single-scale choices

To strengthen future-state prediction across timescales, MSDE trains the shared state with both micro- and macro-scale objectives. We assess the benefit of this joint supervision for downstream transfer by retaining only the micro- or macro-scale objective in separate ablations. The preferred scale varies by task: each variant outperforms the other on 10 tasks. Joint prediction improves mean FT by 3.0 points over micro-only and 3.7 over macro-only (both $p < 0.01$; Table 6). It exceeds both on 18 of 20 tasks and averages 1.9 points above a task-wise reference that takes the higher observed FT mean of the two single-scale models (Figure 3b; Appendix I). Because this reference already selects the stronger scale for each task, the remaining FT gain suggests a benefit from learning the two scales together within one representation.

Motor imagery illustrates the benefit of adding macro-scale supervision to micro-scale prediction. It modulates ongoing sensorimotor rhythms (Pfurtscheller & Lopes da Silva, 1999); joint prediction improves over micro-only by 5.4 points on BCIC-IV-2a and 3.6 on SHU-MI, with gains in every seed. Sleep staging likewise combines local events with temporal context (Cash et al., 2009; Fernandez & Lüthi, 2020). On HMC, the advantage holds against both single-scale controls: 2.7 points over micro-only and 3.3 over macro-only, with gains against both in every seed.

4.5 Cross-dataset transfer

We further examine whether future-state prediction helps the encoder capture EEG structure shared across datasets. To test this, we fine-tune on one sleep dataset and evaluate directly on another, without target adaptation. Across six directions among HMC, ISRUC-S3 and SleepEDF, MSDE

averages 44.1% balanced accuracy, 10.5 points above the strongest other foundation model, REVE-L (Table 18). MSDE exceeds REVE-B and CBraMod in five of six directions, and LaBraM and REVE-L in four.

The same-encoder controls help isolate the contribution of the pretraining objective. MSDE exceeds Scratch by 8.7 points and reduces the mean within-direction seed standard deviation from 6.8 to 3.8 points. Replacing future prediction with masked infilling or reconstruction lowers mean accuracy by 5.6 or 3.0 points, respectively (Table 19). MSDE exceeds infilling in all six directions and 17 of 18 source–target–seed pairs, supporting future-state prediction as a useful constraint for transfer across recording configurations.

5 PHYSIOLOGICAL INFORMATION AND TEMPORAL ALIGNMENT

The transfer gains raise two linked questions: which physiological information does pretraining retain, and how does it contribute to classification? We connect frozen-state readouts to component interventions, then use spectrum-preserving shifts to test temporal alignment beyond whole-window power.

5.1 PHYSIOLOGICAL INFORMATION ACCESSIBLE IN FROZEN STATES

To explain MSDE's transfer gains, we evaluate linear readouts from pooled frozen pretrained states to physiological descriptors computed from EEG on held-out subjects, matching architectures and pretraining data across MSDE, infilling and MAE (Appendix L.1). With R^2 averaged over descriptors within each task, MSDE leads on five of six selected tasks, averaging 0.539 versus 0.465 for infilling and 0.489 for MAE (Table 27).

On AD65, dementia-related slowing makes theta/alpha balance relevant to diagnosis (Jeong, 2004; Miltiadous et al., 2023). MSDE reads out the log theta/alpha ratio with $R^2 = 0.794$, versus 0.489 and 0.576 for infilling and MAE, alongside three-seed FT gains of 15.9 and 16.1 balanced-accuracy points, respectively (Table 30). SleepEDF spindle time fraction ($R^2 = 0.900$) and BETA occipital log power in the 16–32 Hz harmonic band ($R^2 = 0.782$) extend the readout gains to sleep events and stimulus-driven rhythms (Figure 4a). These features support N2 identification and frequency-based stimulus decoding, respectively (Fernandez & Lüthi, 2020; Iber, 2007; Norcia et al., 2015). The results connect improved transfer to greater linear access to label-relevant event and spectral information before fine-tuning (Table 28).

5.2 PHYSIOLOGICAL COMPONENTS USED IN DOWNSTREAM DECISIONS

We next test whether classifiers use this information by removing task-related or control components at test time while keeping the fine-tuned models fixed (Appendix L.2). We compare MSDE's advantage over infilling and MAE under task-related and control removal across nine task–component pairs. Target removal reduces MSDE's advantage more than control removal in 16 of these 18 comparisons, with 14 bootstrap 95% intervals entirely above zero (Table 32).

On BETA, removing the labelled stimulus frequency and its harmonics reduces MSDE's advantage over infilling from 3.4 to 0.8 points; removing another stimulus frequency leaves a 3.5-point advantage (Figure 4b). On ErrP, removing response-related 0.5–8 Hz activity (Chavarriaga & Millán, 2010; Chavarriaga et al., 2014) similarly reduces the advantage from 10.4 to 0.8 points, while the 20–28 Hz control preserves it. The two cases connect MSDE's gains to sustained rhythms and transient responses.

Sleep provides a class-specific link to the spindle readout: sigma-band removal lowers MSDE's N2 recall by 24.3–47.5 points across all three sleep datasets, with larger mean losses than either recovery objective (Figure 4c). Stage-specific theta responses and interventions on frozen LP classifiers extend this connection to other sleep features and classification without encoder fine-tuning (Tables 33 and 34).

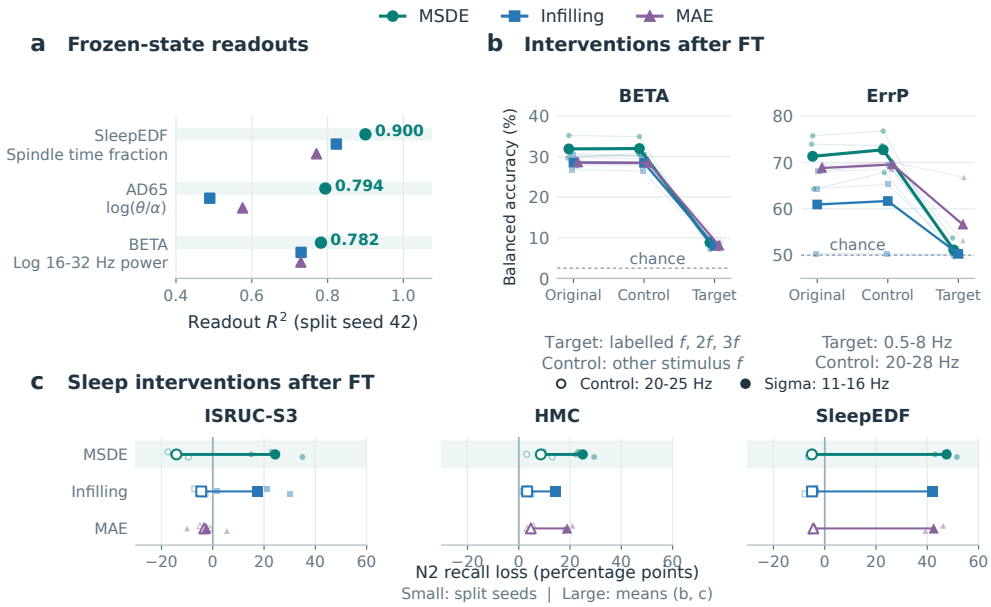

Figure 4: Physiological readouts and interventions: (a) frozen pretrained-state readouts (seed 42); (b) test-time removal after FT: labelled frequency and harmonics versus another frequency (BETA), or 0.5–8 versus 20–28 Hz (ErrP); (c) N2 recall loss after band removal. Small symbols in (b,c) show three split seeds; large symbols show means. Dashed lines in (b) mark chance. Protocols: Appendix L.

5.3 Temporal alignment beyond whole-window spectra

The preceding interventions connect MSDE's classification gains to physiological signal components. We now ask whether the classifier also relies on the temporal alignment of these components relative to the input window, beyond whole-window spectral power. FACED provides a test case: removing frontal 0.5–3 Hz activity lowers MSDE by 15.6 points, compared with 1.1 points for the frontal 3–6 Hz control (Table 31). We therefore circularly shift all channels together while holding classifiers fixed (Appendix K.2). These shifts preserve each channel's whole-window amplitude spectrum and interchannel cross-spectra while changing the signal's alignment with the window boundaries. MSDE initially exceeds Scratch by 15.6 points and infilling by 5.9. Its score drops by 21.5 points at 0.25 s; at 0.5 s, the loss is 41.9 points, versus 18.5 for Scratch, and exceeds infilling's loss by 4.9 (Table 24). To identify which signal components contribute to this sensitivity, we further apply the shifts separately to activity below and above 3 Hz. MSDE's excess accuracy loss relative to Scratch is substantially larger for the low-frequency component (Table 25). These results link MSDE's advantage to temporal alignment, particularly in low-frequency activity, beyond whole-window spectral power.

6 Conclusion

We introduced MSDE, which couples micro- and macro-scale future-state prediction through a shared transition operator conditioned on elapsed time. Across 20 tasks, matched controls support both future prediction over content recovery and joint supervision over either scale alone. The resulting fine-tuning gains extend to sleep staging across datasets without target adaptation.

Linear readouts of frozen MSDE states reveal improved access to task-relevant event and spectral information, while component interventions link these features to classification gains. Spectrum-preserving shifts on FACED further show that classification depends on temporal alignment beyond whole-window power. These findings support joint future prediction as a training principle for transferable EEG representations: a shared state learns to support both fine-resolution and longer-range forecasts, accommodating the different temporal demands of downstream tasks.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and figure preparation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses existing EEG research datasets and involves no new participant recruitment or data collection. Data access and use follow the original providers' terms. EEG recordings and associated clinical and behavioural labels contain potentially sensitive information, requiring appropriate privacy protections in subsequent use. MSDE is developed and evaluated for research purposes; clinical applications would require independent validation.

REPRODUCIBILITY STATEMENT

Detailed experimental hyperparameters and training, evaluation and statistical procedures are provided in Appendices B and C. All training data come from publicly available datasets obtained from their official providers (Appendices A and B.1). We will publicly release the training code, model implementation and model checkpoints upon acceptance.

REFERENCES

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. doi: 10.1371/journal.pone.0256111.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Anahit Babayan, Miray Erbey, Deniz Kumral, et al. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific Data*, 6, 2019.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Klaus-Robert Müller, and Gabriel Curio. The non-invasive Berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

György Buzsáki, Costas A. Anastassiou, and Christof Koch. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, 13(6):407–420, 2012. doi: 10.1038/nrn3241.

Sydney S. Cash, Eric Halgren, Nima Dehghani, Andrea O. Rossetti, Thomas Thesen, Chunmao Wang, Orrin Devinsky, Ruben Kuzniecky, Werner Doyle, Joseph R. Madsen, Edward Bromfield, Loránd Eross, Péter Halász, George Karmos, Richárd Csercsa, Lucia Wittner, and István Ulbert. The human K-complex represents an isolated cortical down-state. *Science*, 324(5930):1084–1087, 2009. doi: 10.1126/science.1169626.

Ricardo Chavarriaga and José del R. Millán. Learning from EEG error-related potentials in non-invasive brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(4):381–388, 2010. doi: 10.1109/TNSRE.2010.2053387.

Ricardo Chavarriaga, Aleksander Sobolewski, and José del R. Millán. Errare machinale est: The use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*, 8:208, 2014. doi: 10.3389/fnins.2014.00208.

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

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

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, 2026.

Laura M. J. Fernandez and Anita Lüthi. Sleep spindles: Mechanisms and functions. *Physiological Reviews*, 100(2):805–868, 2020. doi: 10.1152/physrev.00042.2018.

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

Uri Hasson, Eunice Yang, Ignacio Vallines, David J Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *The Journal of neuroscience*, 28(10):2539–2550, 2008.

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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Conrad Iber. The aasm manual for the scoring of sleep and associated events: rules, terminology, and technical specification. *(No Title)*, 2007.

Jaeseung Jeong. EEG dynamics in patients with Alzheimer's disease. *Clinical Neurophysiology*, 115(7):1490–1505, 2004. doi: 10.1016/j.clinph.2004.01.001.

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, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multitask foundation model for bridging the gap between language and eeg signals. In *International Conference on Learning Representations*, volume 2025, pp. 55436–55457, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A. C. Kamphuisen, and Josefien J. L. Oberyé. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47, 2000.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124, 2016.

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.

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. doi: 10.1088/1741-2552/aace8c.

Jinhao Li, Zhiyuan Ma, Xueqiao Han, Zhongye Xia, Xinche Zhang, Shanghong Xie, Yixuan Liu, Yongjian Li, Runmin Gan, Tianlin Huo, et al. Eeg-jepa: Structured latent prediction for eeg foundation models. *arXiv preprint arXiv:2608.00114*, 2026.

Bingchuan Liu, Xiaoshan Huang, Yijun Wang, Xiaogang Chen, and Xiaorong Gao. BETA: A large benchmark database toward SSVEP-BCI application. *Frontiers in Neuroscience*, 14:627, 2020. doi: 10.3389/fnins.2020.00627.

Ziling Lu, Zongsheng Li, Xinke Shen, Kexin Lou, Yingyue Xin, Xiaoqi Chen, Shinan Wang, Xiang Chen, Jiahao Fan, Chenyu Huang, et al. Omnieeg-bench: A standardized evaluation benchmark for eeg foundation models. *arXiv preprint arXiv:2606.00815*, 2026.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9:531, 2022. doi: 10.1038/s41597-022-01647-1.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, et al. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8, 2023.

John D Murray, Alberto Bernacchia, David J Freedman, Ranulfo Romo, Jonathan D Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, et al. A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience*, 17(12):1661–1663, 2014.

Anthony M. Norcia, L. Gregory Appelbaum, Justin M. Ales, Benoit R. Cottereau, and Bruno Rossion. The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, 15 (6):4, 2015. doi: 10.1167/15.6.4.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Gert Pfurtscheller and Fernando H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999. doi: 10.1016/S1388-2457(99)00141-8.

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

Alexander P. Rockhill, Nicko Jackson, Jobi George, Adam Aron, and Nicole C. Swann. UC San Diego resting state EEG data from patients with Parkinson's disease. OpenNeuro dataset ds002778, version 1.0.5, 2021.

Cosma Rohilla Shalizi and James P Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of statistical physics*, 104(3):817–879, 2001.

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, 2023. doi: 10.1109/TNSRE.2022.3230250.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, et al. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6, 2012.

Anna Tegon, Thorir Mar Ingolfsson, Xiaying Wang, Luca Benini, and Yawei Li. Femba: Efficient and scalable eeg analysis with a bidirectional mamba foundation model. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7. IEEE, 2025.

Guagnyu 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.

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, 2025a.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. EEGMamba: An EEG foundation model with Mamba. *Neural Networks*, 192:107816, 2025b. doi: 10.1016/j.neunet.2025.107816.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *Advances in Neural Information Processing Systems*, 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.

Tongtian Yue, Xuange Gao, Shuning Xue, Yepeng Tang, Longteng Guo, Jie Jiang, and Jing Liu. Braingpt: Unleashing the potential of eeg generalist foundation model by autoregressive pretraining. *arXiv preprint arXiv:2410.19779*, 2024.

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, 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, et al. Electroencephalograms during mental arithmetic task performance. *Data*, 4, 2019.

A DATASETS

The 20 downstream tasks use fixed subject-level train/validation/test splits with nominal 8:1:1 allocation; Table 3 reports their actual sizes. Balanced accuracy is averaged over split seeds 5, 10 and 42. PD31 uses only seeds 5 and 10 because seed 42 contains a single test class; this exclusion applies consistently to all reproduced main and ablation scores. Published benchmark results come from separate runs (Appendix F); temporal-alignment and sleep-event diagnostics use Appendix K, and physiological readouts and component interventions use Appendix L.

Table 2: Processed dataset inventory for the 20 downstream tasks. The channel column reports the number of channels in the processed recordings and the number used for evaluation, respectively. Window length is in seconds. Reported sampling rates are 200 Hz; the TUEV sampling rate is unreported. Subjects and windows refer to the processed dataset before split-specific sampling; the two LEMON tasks use the same recordings with different labels.

Task	Channels	Window	Classes	Subjects	Windows
ISRUC-S3 (Khalighi et al., 2016)	6/6	30	5	10	8,846
SleepEDF (Kemp et al., 2000)	2/2	30	5	153	414,961
HMC (Alvarez-Estevez & Rijsman, 2021)	8/4	30	5	151	137,243
ISRUC-S2	6/6	30	5	8	14,141
TUEV (Harati et al., 2015)	24–32/23	5	6	370	113,353
AD65 (Miltiadous et al., 2023)	19/19	10	3	88	7,013
PD31 (Rockhill et al., 2021)	40/32	1	2	31	9,064
FACED (Chen et al., 2023)	32/32	10	9	123	10,332
SEED (Zheng & Lu, 2015)	62/62	10	3	15	14,895
SEED-IV	62/62	4	4	15	37,575
SEED-V	62/62	1	5	16	117,744
SEED-VII	62/62	10	7	20	27,340
BCIC-IV-2a (Tangermann et al., 2012)	22/22	4	4	9	5,184
SHU-MI (Ma et al., 2022)	32/32	4	2	25	11,988
BCIC-IV-1 (Blankertz et al., 2007)	59/59	4	2	7	1,400
BETA (Liu et al., 2020)	64/64	1	40	70	11,200
ErrP (Chavarriaga & Millán, 2010)	64/64	1	2	6	6,437
EEGMAT (Zyma et al., 2019)	21/19	5	2	36	1,696
LEMON-sex (Babayan et al., 2019)	59/59	1	2	203	191,354
LEMON-extr.	59/59	1	2	203	191,354

Table 3: Train/validation/test split sizes across retained seeds: minima–maxima, or one value when counts match. n is the number of retained seeds; the final column is test subjects per seed. All reported retained test-label distributions contain every task class. PD31 seed 42 is excluded: all 581 test windows belong to class 0, with no class-1 examples.

Task	n	Train windows	Val. windows	Test windows	Test subjects
ISRUC-S3	3	1,600	792–949	996	1
SleepEDF	3	22,790–22,863	40,174–41,150	43,334–43,757	16
HMC	3	23,488–23,715	13,038–13,524	13,580–14,434	15
ISRUC-S2	3	1,200	1,840–1,975	1,720	1
TUEV	3	16,409–16,946	8,966–13,206	9,088–12,908	37
AD65	3	2,790	676–816	697–735	9
PD31	2	1,000	754–947	763–766	3
FACED	3	8,232	1,008	1,092	13
SEED	3	1,440	1,986	993	1
SEED-IV	3	1,920	5,010	2,505	1
SEED-V	3	2,600	14,718	7,359	1
SEED-VII	3	4,480	2,734	2,734	2
BCIC-IV-2a	3	1,120	576	576	1
SHU-MI	3	1,600	962–977	1,442–1,467	3
BCIC-IV-1	3	400	200	200	1
BETA	3	8,960	1,120	1,120	7
ErrP	3	320	1,080–1,102	1,044	1
EEGMAT	3	1,371–1,380	172–181	144–145	3
LEMON-sex	3	6,480	18,828–18,867	19,734–19,869	21
LEMON-extr.	3	6,480	18,828–18,867	19,734–19,869	21

B Training and reproduction details

B.1 Pretraining data, optimisation and hardware

Before pretraining, we removed recordings not originating from TUH and all data belonging to the downstream evaluation datasets, including TUEV. This filtering also excluded the 689 subjects appearing in the TUAB and TUSL evaluation splits. The cleaned corpus contains 1,734,096 windows of 44 s (21,195 h). MSDE and its ablations use a deterministic 10% subset of 173,000 windows (2,115 h); inputs are cropped to the first 16 s. Each ablation repeats pretraining with its specified change to the base protocol. The base protocol uses pretraining seed 0 and specifies 70 epochs (approximately 96k steps), eight H100 GPUs and batch size 16 per GPU. MSDE-Full uses the same architecture and objective on the full corpus for 50 epochs (approximately 332k steps), with 16 H100 GPUs, batch size 16 per GPU and one warm-up epoch. Both use bfloat16 precision; the remaining settings are in Table 4. The full-corpus extension changes both data exposure and optimisation budget (Appendix E).

Objective controls. Masked latent infilling selects one-second patches at a mask ratio of 0.5 and zeros their raw EEG before encoding: $x_{\text{in}} = x \odot (1 - M)$, where M is the patch mask broadcast over signal samples. Masking therefore precedes the convolutional stem rather than removing embeddings after tokenisation. Unmasked patches remain visible under the same causal temporal-attention window as MSDE; a masked position can access preceding context but not future patches. The recovery loss is smooth L1 against stop-gradient EMA-teacher states at token positions belonging to masked patches. Both micro- and macro-scale future-prediction losses are explicitly set to zero, and the transition operator is not trained in this mode. The control retains the encoder, EMA teacher, identity anchoring and variance–covariance regularisation, and uses the same data subset, 70-epoch schedule, batch size and pretraining seed as MSDE.

The masked raw-signal reconstruction control follows the content-recovery objective of masked autoencoders (He et al., 2022). Its head combines a masked-patch token state and electrode embedding in a two-layer network to regress the window-normalised 250 ms segment for that electrode. We test

Table 4: Pretraining settings for MSDE and its ablations.

Parameter	Value
Input	200 Hz, μV/100, 16 s windows, 1 s patches
Stem	250 ms kernel, 125 ms stride, 192 filters, causal
Width	256
Channel layers	2 self-attention layers, 10–20 position embeddings
Temporal layers	4 causal self-attention layers, 8 heads, MLP ratio 4
Positions	rotary, temporal attention only
Attention window	24 tokens per temporal layer
Transition operator	MLP width 768, time embedding 128
Micro-scale horizons	$[0.5, 0.75]$, $[0.75, 1.0]$ s
Macro-scale horizons	$[0.75, 1.25]$, $[1.5, 2.5]$, $[3.5, 4.5]$ s
Pairs per bin / triples per window	16 / 16
Loss weights	micro 1, macro 1, anchoring 3, var–cov 2, semigroup 0.1
Anchoring temperature	0.1
Teacher momentum	0.996
Optimiser	AdamW, lr 10^{-4}, cosine to 10^{-6}, 5 warm-up epochs
Weight decay / clipping	0.05 / 3.0
Batch size	16 per GPU, 8 GPUs (MSDE-Full: 16 GPUs)
Planned epochs	70 (MSDE-Full: 50, one warm-up epoch)
Pretraining seed	0 (distinct from downstream split seeds)
Precision	bfloat16
No var–cov checkpoint	Last available (epoch 52)

mask ratios 0.25, 0.5 and 0.75 with the same causal encoder and planned pretraining schedule as MSDE. The complete objective and horizon comparisons are retained in Appendix C.1.

B.2 Downstream adaptation and hyperparameter selection

FT updates the feature extractor and classifier; LP freezes the pretrained feature extractor and trains only the classifier. For MSDE, valid encoder states are mean-pooled and passed through LayerNorm, dropout and a linear classifier. MSDE and all reproduced baselines use the shared base FT configuration in Table 5 and the same hyperparameter-selection protocol.

Table 5: Shared base fine-tuning configuration for MSDE and reproduced baselines.

Parameter	Value
Optimiser	AdamW
Feature extractor / head learning rate	10^{-4} / 10^{-3}
Weight decay	0.05
Label smoothing	0.1
Learning-rate schedule	5 warm-up epochs, then cosine decay
Gradient clipping	3.0
Batch size	32
Training budget per task	Approximately 60k steps, capped at 150 epochs

Validation-based selection. For each reported model and evaluation setting, validation performance selects hyperparameters and checkpoints; held-out tests provide the reported scores. For all models, FT evaluates a six-candidate learning-rate/layer-decay grid on split seed 42 and reuses the selected configuration for the other seeds. LP evaluates head learning rates 10^{-3}, 3×10^{-3} and 3×10^{-4}.

B.3 Encoder, transition operator and loss implementation

Inputs are sampled at 200 Hz, expressed in μV divided by 100, and split into one-second patches. The per-channel stem uses a 250 ms convolutional kernel, 125 ms stride and 192 filters, followed by per-token LayerNorm and a left-padded convolution to 256 dimensions. Each patch produces seven

tokens, giving 112 for a 16 s input; a token is timestamped at the end of its kernel window. Electrode-name embeddings from a 10–20 vocabulary are added to channel tokens. Two self-attention layers mix a learned summary token with the channel tokens at each time, and the summary-token output supplies the channel-aggregated state. Four pre-norm causal temporal layers use eight heads, MLP ratio four and rotary positions on queries and keys; each layer attends over the preceding 24 tokens. The final token of a one-second patch also supplies its patch state. The encoder accepts one to 64 input channels.

The 1.3M-parameter transition operator uses a 128-dimensional continuous offset embedding and a two-layer mean-prediction MLP of width 768. Its implementation also includes a log-variance head; the future-state losses reported here use MSE on the predicted mean. The EMA teacher has momentum 0.996 and receives no target gradients. Up to 16 source–target pairs are sampled per window for each horizon bin. Micro-scale bins are $[0.5, 0.75]$ and $[0.75, 1.0]$ s; macro-scale bins are $[0.75, 1.25]$, $[1.5, 2.5]$ and $[3.5, 4.5]$ s.

For either scale $r \in \{\mathrm{micro}, \mathrm{macro}\}$, the future-state prediction loss is

$$\mathcal{L}_r = \mathbb{E}_{\mathcal{P}_r} \frac{\|\Phi(s_t, h) - \mathrm{stopgrad}(\bar{s}_{t+h})\|_2^2}{D}, \tag{3}$$

where D is the state dimension and $\mathcal{P}_r$ weights nonempty horizon bins equally and sampled within-window pairs equally within each bin across the batch (Table 4).

Identity anchoring uses symmetric InfoNCE with temperature 0.1 between pooled states from the first and second temporal halves of each window, using other windows in the batch as negatives. Only windows with two nonempty halves participate; n denotes their number in the current loss computation. Normalising both pooled vectors makes this objective depend on their directions, encouraging features shared across temporal halves that distinguish windows.

Variance–covariance regularisation penalises feature standard deviations below a target and squared off-diagonal covariances. The token term measures these statistics over all valid tokens across samples and time; the pooled term measures them across window means. Both terms have unit weight, with target standard deviation $\gamma = 1$ and stabiliser $\epsilon = 10^{-4}$. The variance hinge uses population variance; the covariance penalty uses the $N - 1$ denominator for its N token or pooled observations. A trajectory shared by all windows can produce token variance yet identical pooled vectors, motivating a separate constraint on the downstream representation. The covariance penalty discourages linear redundancy between dimensions.

Semigroup consistency samples up to 16 triples per window, with total offset at most 1 s, to connect direct and composed predictions from the same source state. It averages elementwise Smooth L1 penalties with $\beta = 1$ over the sampled triples and state dimensions. The second step takes the first predicted state as input, and gradients pass through both routes to the shared operator and source encoder state.

C Statistical analysis and numerical precision

For the main FT, LP, objective and horizon comparisons, task means are recomputed from seeds 5 and 10 for PD31 and seeds 5, 10 and 42 for the other 19 tasks. Tasks are equally weighted. Paired comparisons report mean differences and task wins; the 59 task–seed pairs are not treated as independent tasks. The tests are descriptive and are not corrected for multiple comparisons. We retain the recorded per-seed precision and recompute task means before display rounding; separately rounded mean columns are not used for statistical inference. Signed-rank tests use exact conditional sign enumeration, omit zero differences (the Wilcox convention), and assign midranks to tied absolute differences. For directional FT and ablation comparisons, the alternative is the direction stated in the table. LP comparisons are two-sided. Confidence intervals use 100,000 bootstrap resamples of the 20 task means, taking the 2.5th and 97.5th percentiles (random seed 20260916).

For the transfer summaries, write Δ_t for MSDE's retained-seed mean minus a baseline's on task t. W/T/L counts positive, exactly zero and negative Δ_t at the recorded precision; a win is numerical, not a per-task significance claim, and a tie is not defined by non-significance. Median and Q_{25} summarise these paired differences; Q_{25} uses linear interpolation between ordered values (Hyndman–Fan type 7). Ranks in Table 10 use the eight FT models and seven pretrained LP models

in Table 1, with midranks for exact ties; Top-k counts ranks at most k. For K_t classes and balanced accuracy $b_t \in [0, 1]$, the chance-normalised score is $100(b_t - 1/K_t)/(1 - 1/K_t)$, averaged equally across tasks. This aligns the chance reference across class counts without equalising task difficulty. The largest-gain sensitivity analysis removes one task separately for each baseline and reports the remaining mean. These summaries are computed from the retained per-seed scores and dataset class counts using the procedures above.

At the recorded precision, the 1.0–2.0 s horizon and MSDE tie on BCIC-IV-1. The signed-rank test therefore uses 19 nonzero task differences, with 13 wins, one tie and six losses for MSDE. The three horizon variants have one-sided exact p values of 0.00013065, 0.0000314713 and 0.0222797; their two-sided values are 0.00026131, 0.0000629425 and 0.0445595.

The cross-dataset transfer and label-budget analyses use their own experiment records and checkpoints. These results are distinct from those in the main model-selection table. For label-budget results, three seeds are nested within each of six tasks; we report their aggregate means and seed-win counts because individual seed scores are unavailable. Cross-dataset results are computed from per-seed scores, averaging source seeds within each direction and then weighting directions equally. We report sample standard deviations across source seeds and numerical win counts, using complete common directions for aggregate comparisons between models. Since transfer directions share datasets and source models, these summaries are descriptive; the source–target–seed pairs are not treated as independent datasets for significance tests or confidence intervals.

C.1 Complete objective and horizon comparisons

Tables 6 and 7 retain the complete controlled comparisons behind Sections 4.3 and 4.4, including regularisation ablations, LP changes and exact directional tests. Figure 3 visualises the main FT objective and scale contrasts; its gains use the opposite sign to the variant-minus-MSDE differences tabulated here. Removing identity anchoring reduces mean FT accuracy by 3.6 points; increasing its weight by a factor of 3.3 reduces it by 2.5. The no-variance–covariance variant is evaluated from its last available pretraining checkpoint and scores 2.7 points below MSDE. With macro horizons fixed, micro ranges of 0.1–0.5 and 0.1–0.75 s underperform the reference 0.5–1.0 s range by 2.7 and 2.8 points; 1.0–2.0 s is 1.3 points lower (Figure 7).

Table 6: Pretraining controls (20 tasks; three seeds, two for PD31). Δ is variant minus MSDE; intervals are task-bootstrap 95% CIs and p tests whether the variant is worse. Wins count tasks where MSDE is better. Δ LP is the change in mean probe accuracy. Raw reconstruction labels give the mask ratio.

Variant	FT	Δ FT	95% CI	p	Wins	Δ LP
No prediction	50.2	-4.6	[-6.6, -2.8]	2.4×10^{-5}	18/20	-1.6
Masked latent infilling	50.7	-4.1	[-6.3, -1.8]	6.0×10^{-4}	17/20	-2.4
Raw reconstruction (0.25)	50.3	-4.5	[-6.5, -2.9]	1.9×10^{-6}	19/20	-1.3
Raw reconstruction (0.5)	50.5	-4.3	[-6.2, -2.5]	6.7×10^{-5}	18/20	-2.1
Raw reconstruction (0.75)	49.3	-5.5	[-8.0, -3.2]	1.6×10^{-4}	19/20	-3.2
No identity anchoring	51.2	-3.6	[-5.0, -2.2]	3.0×10^{-5}	18/20	-1.4
Anchoring $\times 3.3$	52.3	-2.5	[-3.6, -1.3]	6.0×10^{-4}	18/20	-1.9
No var–cov	52.1	-2.7	[-4.3, -1.1]	6.0×10^{-4}	17/20	-2.5
No semigroup	52.9	-1.9	[-3.0, -0.8]	0.0016	16/20	-1.0
Micro only	51.8	-3.0	[-4.5, -1.4]	0.0016	18/20	-1.1
Macro only	51.1	-3.7	[-5.5, -2.2]	6.7×10^{-6}	19/20	-1.1

Table 7: Micro-scale horizons with macro-scale horizons fixed at 0.75–4.5 s. Differences and task-bootstrap intervals are relative to MSDE; one-sided exact signed-rank tests evaluate whether each variant is worse. W/T/L counts MSDE wins/ties/losses across 20 task means computed from the per-seed scores.

Micro-scale horizon (s)	FT	Δ FT [95% CI]	p	W/T/L	LP
0.5–1.0 (MSDE)	54.8	–	–	–	43.9
0.1–0.5	52.1	-2.7 [-3.8, -1.6]	1.3×10^{-4}	17/0/3	43.8
0.1–0.75	52.0	-2.8 [-4.3, -1.6]	3.1×10^{-5}	17/0/3	42.9
1.0–2.0	53.5	-1.3 [-2.3, -0.3]	0.022	13/1/6	43.2

D COMPLETE TRANSFER RESULTS AND RESOURCES

Tables 8 and 9 separate paired improvements from the model-level summaries in Table 10. The "Best FM per task" row uses the maximum observed FT mean among the five external pretrained models on each task; it is reported as a descriptive reference.

Table 8: Paired FT differences (MSDE − baseline, balanced-accuracy points). W/T/L counts numerical wins/ties/losses across 20 tasks. Task-bootstrap 95% intervals and unadjusted one-sided exact signed-rank tests.

Baseline	Mean Δ	95% CI	Median Δ	Q_{25}	W/T/L	p
Scratch	+6.7	[+4.4, +9.3]	+5.7	+2.3	19/0/1	4.8×10^{-6}
LaBraM	+5.4	[+1.6, +10.1]	+4.0	-0.1	14/0/6	0.0060
CBraMod	+8.0	[+4.4, +11.9]	+5.7	+1.8	17/0/3	2.4×10^{-4}
REVE-B	+2.7	[+0.2, +5.3]	+1.8	-1.0	13/0/7	0.035
REVE-L	+3.4	[+1.0, +5.9]	+2.5	-2.3	13/0/7	0.017
Neuro-GPT	+6.2	[+3.3, +9.0]	+6.2	+1.8	17/0/3	2.9×10^{-4}
Best FM per task	+0.2	[-1.9, +2.4]	+0.0	-3.7	10/0/10	0.478

Table 9: Paired LP differences (MSDE − baseline, balanced-accuracy points). W/T/L counts numerical wins/ties/losses across 20 tasks. Task-bootstrap 95% intervals and unadjusted two-sided exact signed-rank tests.

Baseline	Mean Δ	95% CI	Median Δ	Q_{25}	W/T/L	p
LaBraM	+2.6	[+0.3, +5.2]	+1.5	-0.3	14/0/6	0.036
CBraMod	+5.9	[+3.0, +9.5]	+4.9	+0.8	17/0/3	5.8×10^{-4}
REVE-B	+0.4	[-1.4, +2.1]	+0.7	-1.6	10/0/10	0.498
REVE-L	+0.2	[-3.4, +4.1]	-1.5	-5.2	8/0/12	0.622
Neuro-GPT	+3.9	[+1.1, +6.4]	+4.6	+0.7	16/0/4	0.0017

Table 10: Transfer summaries over 20 tasks: eight FT models and seven pretrained LP models, as in Table 1. CN denotes chance-normalised balanced accuracy (%; Appendix C). Bold and underline mark the best and second-best values within each protocol.

	Model	Mean	Avg. rank	Top-1	Top-2	CN
FT	MSDE-Full	**56.1**	**2.35**	**9/20**	**12/20**	**35.5**
	MSDE-10%	<u>54.8</u>	<u>3.00</u>	<u>4/20</u>	<u>11/20</u>	<u>33.7</u>
	MSDE-Scratch	48.1	5.75	1/20	1/20	23.6
	LaBraM	49.4	4.85	1/20	5/20	25.8
	CBraMod	46.8	6.05	0/20	2/20	21.5
	REVE-B	52.1	4.18	2/20	4/20	29.2
	REVE-L	51.4	4.28	1/20	3/20	28.5
	Neuro-GPT	48.6	5.55	2/20	2/20	24.5
LP	MSDE-Full	**45.3**	<u>2.95</u>	<u>4/20</u>	<u>9/20</u>	**20.5**
	MSDE-10%	<u>43.9</u>	3.40	<u>4/20</u>	5/20	18.2
	LaBraM	41.4	4.08	2/20	4/20	14.2
	CBraMod	38.0	5.78	1/20	1/20	10.0
	REVE-B	43.6	3.60	1/20	6/20	17.7
	REVE-L	43.8	**2.80**	**8/20**	**13/20**	<u>18.4</u>
	Neuro-GPT	40.1	5.40	0/20	2/20	12.4

Table 11: Complete per-task results, including full-corpus pretraining and the MSDE linear probe (three seeds, two for PD31). Full denotes MSDE-Full; LP denotes MSDE's frozen-encoder readout. Bold and underline mark the best and second-best of the eight fine-tuned models before rounding. The average-rank row uses those eight models.

	Task	Scratch	MSDE	Full	LP	LaBraM	CBraMod	REVE-B	REVE-L	Neuro-GPT
Sleep	ISRUC-S3	65.3	80.7	**90.2**	70.7	74.8	57.7	<u>84.5</u>	83.7	71.0
	SleepEDF	73.4	74.7	**75.6**	67.0	<u>74.8</u>	71.7	74.1	72.4	70.2
	HMC	67.4	**69.8**	<u>69.7</u>	52.3	67.2	66.9	56.5	58.6	60.7
	ISRUC-S2	26.1	<u>26.3</u>	25.6	26.1	21.9	25.6	24.1	23.7	**30.0**
Clinical	TUEV	48.4	54.4	54.6	38.1	59.7	57.7	<u>61.7</u>	55.2	**63.3**
	AD65	44.6	<u>65.2</u>	**66.6**	51.7	55.5	53.7	54.7	58.3	55.7
	PD31	44.5	<u>52.3</u>	**58.5**	44.5	46.8	37.2	43.3	46.0	52.1
Emotion	FACED	45.8	<u>61.3</u>	**61.5**	19.7	21.5	34.8	51.6	47.3	43.3
	SEED	50.8	57.2	**62.0**	59.1	<u>61.5</u>	57.3	60.9	60.5	51.7
	SEED-IV	28.7	32.9	**33.4**	32.7	<u>33.2</u>	25.7	33.0	30.6	26.1
	SEED-V	25.6	<u>32.6</u>	**32.7**	29.0	26.1	26.5	24.6	24.6	28.5
	SEED-VII	19.0	19.5	21.2	18.8	19.1	19.1	**22.0**	<u>21.9</u>	17.6
BCI	BCIC-IV-2a	41.7	**49.8**	<u>48.6</u>	34.6	37.3	30.2	41.8	44.3	36.1
	SHU-MI	**63.9**	<u>63.3</u>	62.7	50.7	51.6	50.8	62.4	56.5	54.6
	BCIC-IV-1	52.3	**57.5**	55.2	53.3	50.7	<u>55.3</u>	54.2	53.7	52.0
	BETA	26.5	31.9	34.8	3.5	31.7	26.6	**38.4**	<u>35.3</u>	17.2
	ErrP	57.5	**71.3**	<u>70.3</u>	55.5	56.4	52.6	62.5	56.8	56.3
Other	EEGMAT	64.2	72.5	72.2	58.3	<u>73.9</u>	62.1	69.9	**74.9**	62.8
	LEMON-sex	65.3	70.2	71.1	63.5	**75.8**	<u>73.6</u>	70.8	73.5	71.6
	LEMON-extr.	50.5	<u>52.3</u>	**54.9**	50.0	48.8	49.9	50.9	51.1	50.8
	Mean (20)	48.1	<u>54.8</u>	**56.1**	43.9	49.4	46.8	52.1	51.4	48.6
	Avg. rank	5.75	3.00	2.35	–	4.85	6.05	4.18	4.28	5.55

Table 12: Mean balanced accuracy and average rank over 20 tasks (three seeds, two for PD31). FT compares all eight models; LP compares the seven pretrained models. Bold and underline mark the best and second-best values within each column.

Model	FT mean	FT avg. rank	LP mean	LP avg. rank
MSDE-Full	**56.1**	**2.35**	**45.3**	<u>2.95</u>
MSDE-10%	<u>54.8</u>	<u>3.00</u>	<u>43.9</u>	3.40
MSDE-Scratch	48.1	5.75	–	–
LaBraM	49.4	4.85	41.4	4.08
CBraMod	46.8	6.05	38.0	5.78
REVE-B	52.1	4.18	43.6	3.60
REVE-L	51.4	4.28	43.8	**2.80**
Neuro-GPT	48.6	5.55	40.1	5.40

Table 13: Population standard deviation (points; ddof=0) of balanced accuracy for MSDE (fine-tuning and linear probe) and Scratch. The divisor is 2 for PD31 (seeds 5 and 10) and 3 for other tasks (seeds 5, 10 and 42). All values are recomputed from the same per-seed scores as the main result tables.

Task	MSDE	MSDE LP	Scratch
ISRUC-S3	2.2	2.5	4.7
SleepEDF	1.2	1.2	1.3
HMC	0.7	1.3	0.8
ISRUC-S2	0.4	0.7	2.2
TUEV	5.8	5.4	10.8
AD65	1.8	6.8	9.3
PD31	17.0	15.7	8.7
FACED	2.0	0.5	2.4
SEED	1.4	0.3	3.5
SEED-IV	0.5	0.9	0.4
SEED-V	3.6	2.6	1.3
SEED-VII	0.2	0.2	1.4
BCIC-IV-2a	3.6	2.3	2.9
SHU-MI	4.4	1.0	3.0
BCIC-IV-1	3.7	3.3	1.6
BETA	2.4	0.5	3.0
ErrP	5.0	4.1	6.8
EEGMAT	8.7	6.2	10.8
LEMON-sex	7.1	4.3	6.6
LEMON-extr.	1.7	0.3	0.6
Median	2.3	1.8	2.9

Table 14: Model size, source EEG and pretraining compute. Parameters count retained backbones, including the EEG encoder and GPT for Neuro-GPT; the CBraMod and REVE-Large papers report 4.0M and 408M. Corpus accounting follows Appendix D. Baseline compute is author-reported (– if unavailable; REVE-B is FLOPs-based); MSDE reports measured wall time.

Model	Params (M)	Pretraining EEG (h)	Pretraining compute
MSDE	5.0	2,115	8×H100, 14 h (112 GPU-h)
MSDE-Full	5.0	21,195	16×H100, 17 h (272 GPU-h)
LaBraM	5.8	2,535	–
CBraMod	4.9	≈9,246	4×RTX A5000, ≈5 days
Neuro-GPT	78.4	5,656	–
REVE-B	69.2	61,415	≈260 A100 GPU-h
REVE-L	390.0	61,415	–

Corpus and parameter accounting. Both panels of Figure 2 use the same downstream scores as the transfer tables: we average split seeds within each task (5, 10 and 42, except 5 and 10 for PD31), then give the 20 tasks equal weight. Bubble areas use parameter counts of the retained downstream backbone; pretraining-only components are excluded. Corpus duration describes the source data, not cumulative epoch exposure or training compute. For MSDE, 2,115 and 21,195 h refer to selected 44 s source windows; training uses their first 16 s, corresponding to approximately 769 and 7,707 h of cropped input before repetition over epochs. LaBraM and REVE use reported corpus durations of 2,534.78 and 61,415 h, respectively (Jiang et al., 2024; El Ouahidi et al., 2026). For Neuro-GPT, the reported 5,656 h covers assembled training and held-out validation recordings, rather than the full TUEG archive (Cui et al., 2024). Its 78.4M-parameter backbone includes the EEG encoder, projection, GPT and pooler. For CBraMod, we derive 9,246 h from the reported 1,109,545 retained, non-overlapping 30 s samples after filtering (Wang et al., 2025a). The 27,062 h listed in OmniEEG-Bench describes the original TUEG corpus before this filtering, rather than the retained samples (Lu et al., 2026). Figure 2 compares transfer at each model's reported data and backbone scale; data curation and pretraining protocols differ across models, and training compute is reported separately in Table 14.

Task variation and sensitivity. The positive FT medians against all five external pretrained models show that the average gains extend across tasks (Table 8). Removing each baseline's largest-gain task leaves mean improvements of 2.1–7.1 points (Table 15). Excluding TUEV also preserves the mean gains over Scratch ($+6.8$) and REVE-B ($+3.2$, $p = 0.018$). MSDE-Full and MSDE-10% rank in the top two on 12 and 11 of 20 FT tasks, compared with 4 for REVE-B; chance normalisation preserves the FT ordering. Tables 8 to 10 report the full distributions of task-wise differences and the corresponding FT and LP summaries.

Table 15: FT sensitivity to the largest gain. Each row removes the task with the largest MSDE advantage over that baseline, then averages the remaining 19 paired differences.

Baseline	Removed task	Remaining mean Δ
Scratch	AD65	+6.0
LaBraM	FACED	+3.6
CBraMod	FACED	+7.1
REVE-B	HMC	+2.1
REVE-L	ErrP	+2.8
Neuro-GPT	FACED	+5.6

E FULL-CORPUS EXTENSION

MSDE-Full uses 21,195 h of source recordings instead of the 2,115 h subset. Its mean FT accuracy is 56.1%, versus 54.8% for MSDE, and it improves on 13 of 20 tasks. Its eight-model average FT rank improves from 3.00 to 2.35. The largest FT gains are ISRUC-S3 ($+9.5$), PD31 ($+6.2$) and SEED ($+4.7$). Mean LP rises from 43.9% to 45.3%, with improvements on 13 tasks. AD65 ($+7.5$), EEGMAT ($+5.9$) and ISRUC-S2 ($+4.8$) show the largest LP gains; Figure 5 reports all task-wise FT and LP changes. The seven-model average LP rank improves from 3.40 to 2.95. Both mean FT and LP improve under this extension.

Table 16: Full-corpus paired comparisons, with the same task-level inference as Table 8.

Comparison	Δ	95% CI	Wins	p
MSDE-Full − Scratch	+8.0	[+5.1, +11.1]	18/20	4.8×10^{-6}
MSDE-Full − MSDE	+1.3	[+0.2, +2.6]	13/20	0.049
MSDE-Full − REVE-B	+4.0	[+1.5, +6.5]	17/20	0.0014
MSDE-Full − Neuro-GPT	+7.5	[+4.4, +10.5]	17/20	2.4×10^{-4}
MSDE-Full − strongest baseline per task	+1.5	[-0.6, +3.6]	13/20	0.101

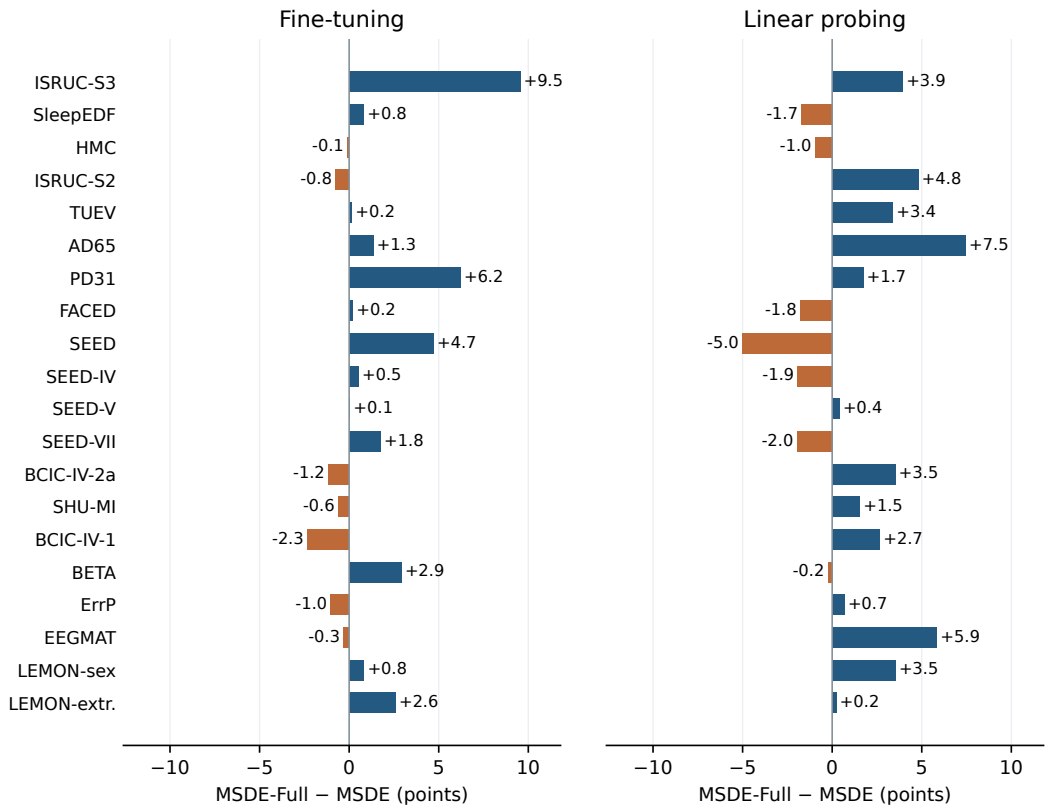

Figure 5: Per-task changes from MSDE to MSDE-Full. Positive values favour full-corpus pretraining. Left: fine-tuning; right: linear probing.

F Context from published benchmark results

We compare with the full-FT (Supplementary Table T7) and frozen-backbone (Table T5) results published with OmniEEG-Bench (Lu et al., 2026). Models include BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), BrainOmni (Xiao et al., 2025), CBraMod (Wang et al., 2025a), EEGConformer (Song et al., 2023), EEGMamba (Wang et al., 2025b), EEGNet (Lawhern et al., 2018), FEMBA (Tegon et al., 2025), LaBraM (Jiang et al., 2024), NeuroGPT (Cui et al., 2024), NeuroLM (Jiang et al., 2025) and REVE (El Ouahidi et al., 2026). Models are ranked on each task among those reporting a result; each model's average covers only its reported tasks. The comparisons cover 13 of our FT tasks and 20 frozen-backbone tasks, with model-specific coverage shown in Table 17.

MSDE-Full and MSDE have the two lowest average FT ranks (2.46 and 3.00). Scratch also ranks highly (5.23), illustrating that architecture, adaptation protocol and task coverage contribute to differences from the published numbers. For frozen backbones, MSDE-Full and MSDE have the two lowest average ranks (3.55 and 4.00), followed by REVE (4.03). The reproduced-model and same-architecture comparisons in Section 4 complement these published results with evaluation under a shared protocol.

Table 17: Average rank against the results published with OmniEEG-Bench (Lu et al., 2026): full fine-tuning (Supplementary Table T7, 13 of our tasks) and frozen backbone (Table T5, 20 tasks). On each task, models are ranked among those that report it; the number of ranked tasks is in parentheses. Our models (italic) are evaluated under our protocol (Section 4.1); published numbers come from the benchmark's training runs.

Model	Fine-tuning	Frozen backbone
MSDE-Full	**2.46 (13)**	**3.55 (20)**
MSDE	3.00 (13)	4.00 (20)
Scratch	5.23 (13)	–
FEMBA	5.88 (12)	6.00 (19)
LaBraM	6.67 (12)	6.71 (17)
CBraMod	6.79 (12)	4.84 (19)
NeuroGPT	7.58 (12)	8.63 (19)
BrainOmni	8.00 (13)	4.72 (20)
REVE	8.46 (12)	4.03 (19)
BIOT	8.54 (13)	6.92 (20)
NeuroLM	9.91 (11)	7.41 (16)
EEGConformer	9.92 (13)	–
EEGMamba	10.25 (12)	8.53 (19)
EEGNet	11.08 (13)	–
BENDR	13.62 (12)	10.63 (19)

G Additional objective and representation analyses

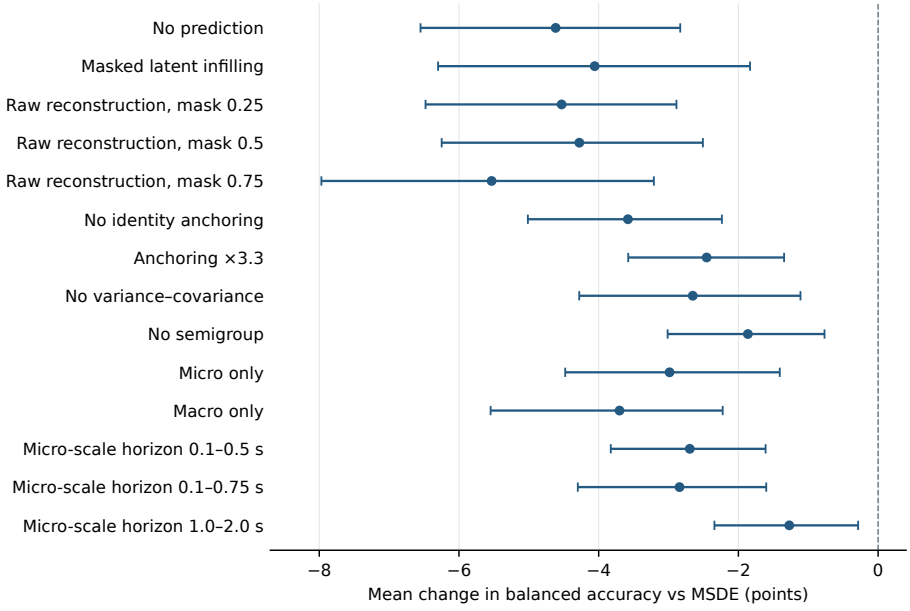

Figure 6: FT changes for objective, regularisation and timescale controls relative to MSDE; 95% task-bootstrap CIs (Tables 6 and 7).

Choice of micro-scale horizons. The single-scale ablations test whether both objectives are useful; here, the macro objective is retained while the micro offsets change (Table 7). The reference 0.5–1.0 s range yields the highest mean FT accuracy among the tested settings, with lower performance for both shorter and longer ranges (Figure 7). This pattern shows that the micro-range choice affects transfer even with macro supervision fixed: extending the micro range to 1.0–2.0 s does not improve mean FT performance. The shortest range nearly matches the reference in LP, so the clearest benefit of the selected range appears after fine-tuning.

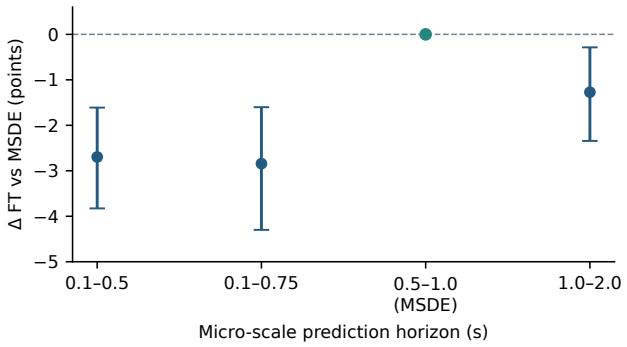

Figure 7: FT changes across micro-scale horizons relative to 0.5–1.0 s; macro horizons are fixed. Error bars: 95% task-bootstrap CIs.

Frozen-state organisation. Frozen states contain sleep-stage information before task-specific label training (Figure 8). Nearest-neighbour balanced accuracy is 0.49 for MSDE versus 0.34 for a random encoder on SleepEDF, with neighbours restricted to other subjects. On ISRUC-S3 it is 0.93 versus 0.68 within the single test subject.

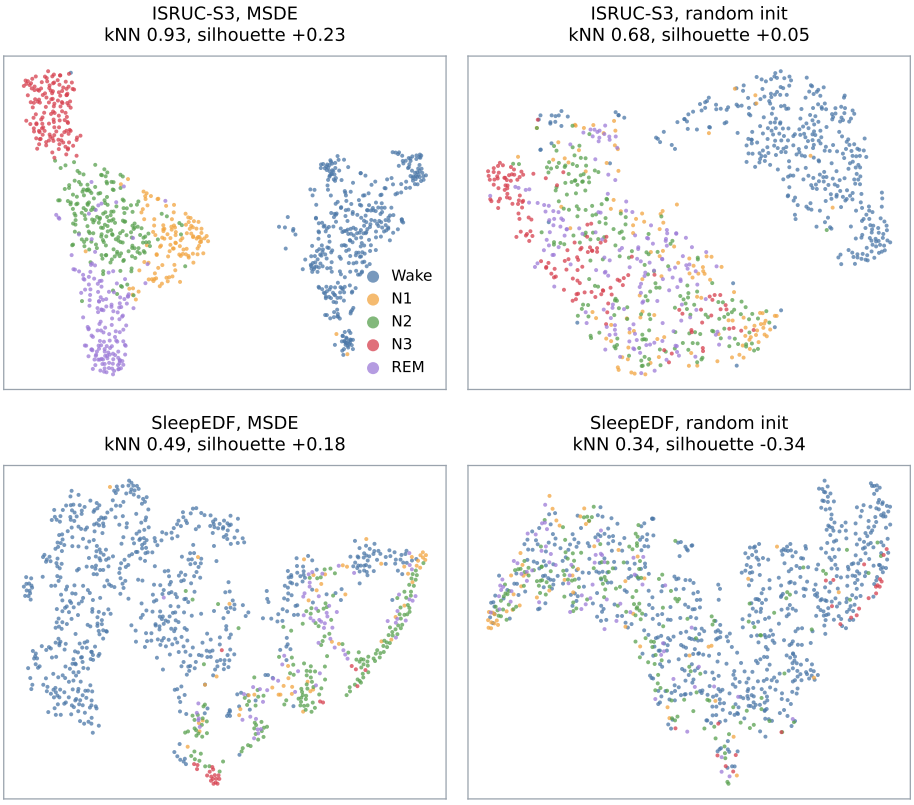

Figure 8: Frozen-state t-SNE, coloured by sleep stage: W, wakefulness; N1–N3, non-rapid-eye-movement sleep stages 1–3; REM, rapid-eye-movement sleep. Scores: nearest-neighbour balanced accuracy ($k = 10$, cosine) and cosine silhouette. SleepEDF neighbours are cross-subject; ISRUC-S3 uses one test subject.

H CROSS-DATASET AND REDUCED-LABEL RESULTS

Cross-dataset sleep staging. The sleep datasets test whether source-task adaptation supports transfer to another recording configuration: six referential channels in ISRUC-S3, four in HMC and two bipolar derivations in SleepEDF. Models fine-tuned on the source dataset are evaluated without target-dataset adaptation on the five shared American Academy of Sleep Medicine (AASM) sleep stages (chance balanced accuracy: 20%). Source training uses seeds 5, 10 and 42; the target test split is fixed at seed 42, so variability reflects source training and split seeds on a common target test set. MSDE uses 10% pretraining in these comparisons.

Table 18 compares five pretrained models with complete three-seed results in all six directions. MSDE achieves the highest mean balanced accuracy (44.1%), followed by REVE-L (33.6%), REVE-B (32.6%), LaBraM (31.6%) and CBraMod (27.0%). MSDE exceeds REVE-B and CBraMod in five directions, and LaBraM and REVE-L in four; the table reports all direction-level means and source-seed variability.

Table 18: Cross-dataset sleep staging with other foundation models. Balanced accuracy (%), mean ± sample SD over three source seeds with a fixed target split. MSDE uses 10% pretraining. Best means are bold and second-best underlined.

Source → target	LaBraM	CBraMod	REVE-B	REVE-L	MSDE
HMC → ISRUC-S3	30.6±9.2	24.9±6.4	<u>35.5</u>±19.4	31.0±4.5	**74.2**±4.7
HMC → SleepEDF	52.2±4.6	40.6±5.7	<u>55.5</u>±2.6	**57.0**±1.8	51.0±4.2
ISRUC-S3 → HMC	<u>23.1</u>±2.6	19.5±0.7	22.3±1.0	23.0±2.8	**25.8**±1.0
ISRUC-S3 → SleepEDF	<u>28.2</u>±10.6	**30.8**±5.2	21.4±1.1	26.5±10.1	25.6±0.3
SleepEDF → HMC	<u>27.3</u>±0.3	27.1±1.6	27.2±2.5	26.6±1.6	**39.7**±1.7
SleepEDF → ISRUC-S3	28.0±7.7	19.2±2.7	34.0±2.7	<u>37.4</u>±3.8	**48.2**±10.9
Mean (six directions)	31.6	27.0	32.6	<u>33.6</u>	**44.1**
MSDE direction wins	4/6	5/6	5/6	4/6	—
MSDE seed-pair wins	14/18	16/18	16/18	12/18	—

Table 19 isolates the pretraining objective within the same encoder across all six directions. Pretraining improves on Scratch by 8.7 points on average, with gains in five directions and 15 of 18 seed pairs. Replacing future prediction with infilling or reconstruction reduces the mean by 5.6 or 3.0 points; MSDE exceeds these controls in six and four directions, respectively, and in 17 and 15 of the 18 seed pairs. The average within-direction seed standard deviation is 3.8 points for MSDE, 6.8 for Scratch, 4.6 for reconstruction and 2.8 for infilling. MSDE thus combines gains over infilling in all six directions with lower source-seed variability than Scratch.

Table 19: Cross-dataset sleep staging with the same encoder. Balanced accuracy (%), mean ± sample standard deviation across three source seeds; the target split is fixed. MSDE uses 10% pretraining. Bold marks the highest mean per direction.

Source → target	MSDE	Infilling	Reconstruction	Scratch
HMC → ISRUC-S3	**74.2**±4.7	66.9±1.8	68.8±4.7	54.1±8.5
HMC → SleepEDF	**51.0**±4.2	35.9±4.0	41.3±9.6	34.5±3.0
ISRUC-S3 → HMC	25.8±1.0	24.2±1.6	22.0±1.5	**25.8**±2.0
ISRUC-S3 → SleepEDF	**25.6**±0.3	21.0±0.4	23.7±1.3	25.5±9.6
SleepEDF → HMC	39.7±1.7	36.2±1.2	**40.6**±6.2	33.6±4.8
SleepEDF → ISRUC-S3	48.2±10.9	47.0±7.9	**50.1**±4.4	38.6±13.1
Mean (six directions)	44.1	38.5	41.1	35.4
Mean seed SD	3.8	2.8	4.6	6.8
MSDE direction wins	—	6/6	4/6	5/6
MSDE seed-pair wins	—	17/18	15/18	15/18

Reduced-label adaptation. Varying label availability then tests how the benefit depends on supervised adaptation. On six tasks, MSDE's mean gains over Scratch at 10%, 20%, 50% and 100%

of training labels are 2.7, 5.2, 6.4 and 7.7 points; at 50%, its gain over infilling is 5.2 (Table 20 and Figure 9). The growing mean advantage over Scratch echoes the FT/LP contrast: supervised adaptation helps realise the benefit of pretraining.

Table 20: Gains of MSDE under reduced training labels (six tasks × three seeds, 18 pairs). Mean gains in points and seed-win counts are descriptive; seeds within a task are not independent tasks (Appendix C).

Labels	vs. Scratch		vs. masked infilling	
	Δ	Seeds won	Δ	Seeds won
10%	+2.7	13/18	+1.3	10/18
20%	+5.2	13/18	+3.1	13/18
50%	+6.4	16/18	+5.2	13/18
100%	+7.7	15/18	+4.6	12/18

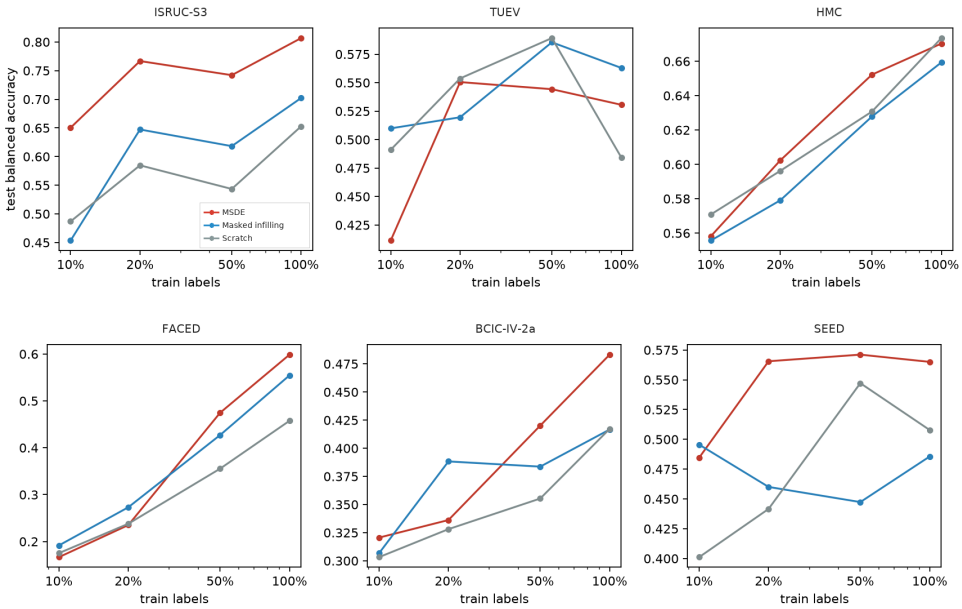

Figure 9: Test balanced accuracy against the fraction of training labels for MSDE, masked infilling and Scratch on six tasks (three seeds).

I PHYSIOLOGICAL RATIONALE AND TARGETED CASE EVIDENCE

The selected task examples in Section 4 relate controlled objective comparisons to known forms of EEG temporal structure. Tables 21 and 22 report all paired seed differences for these comparisons, with the evaluation setting and comparator specified in each row.

Scalp EEG records mixtures of population currents with a major contribution from synaptic activity (Buzsáki et al., 2012). MSDE summarises this recent activity in a causal state and uses elapsed-time conditioning to predict how its contextual representation develops. The range of intrinsic timescales observed in cortical activity motivates supervision over multiple intervals (Murray et al., 2014; Hasson et al., 2008). The micro and macro horizons specify the intervals used in this learning objective.

The joint-scale comparison provides the broadest test of this design. Micro-only has the higher FT mean on 10 tasks and macro-only on 10, including a near tie of 0.003 points on LEMON-sex. Using the higher observed task mean of the two single-scale models as the reference, joint prediction gains 1.9 FT points on average and leads on 18 tasks (Figure 3b). It exceeds both variants in 36 of 59 paired seed comparisons; the corresponding LP margin is −0.2 points, with 11 of 20 task wins.

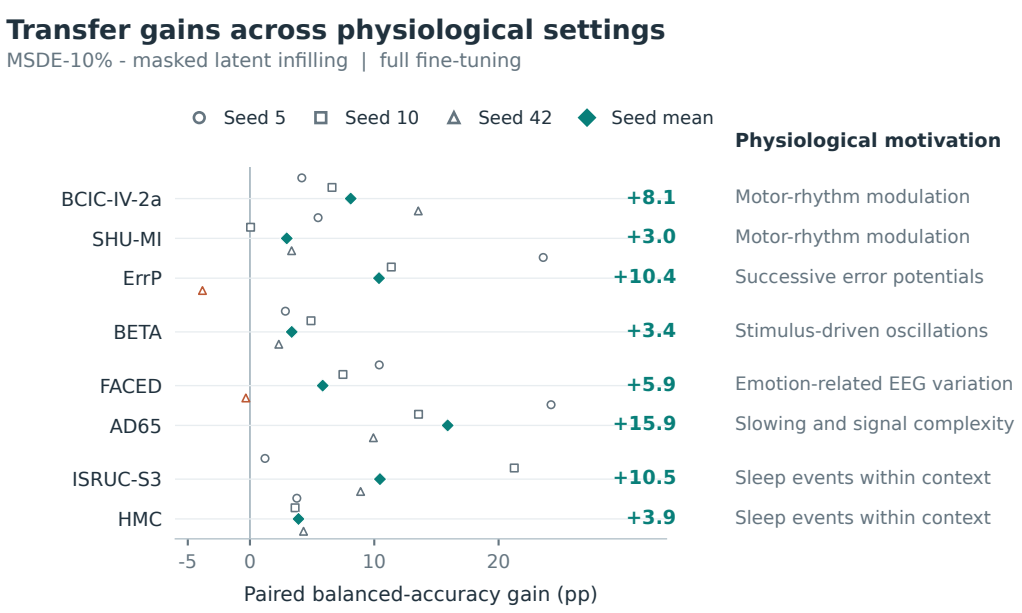

Figure 10: FT gains over latent infilling on eight tasks selected post hoc (MSDE-10%; balanced-accuracy points). Open markers: split seeds 5, 10 and 42; diamonds: task means.

Motor imagery makes the relation between local activity and its development concrete. Event-related desynchronisation (ERD) and event-related synchronisation (ERS) describe frequency-specific modulation of ongoing rhythms, with distinct scalp distributions (Pfurtscheller & Lopes da Silva, 1999). On BCIC-IV-2a and SHU-MI, MSDE exceeds infilling by 8.1 and 3.0 FT points and micro-only by 5.4 and 3.6; all four comparisons favour MSDE in every seed. Both tasks use four-second windows, with one and three test subjects per seed, respectively.

The importance of context also arises when a local event must be read against its background. K-complexes involve isolated cortical down-states, distinguished from down-states in recurring slow oscillations (Cash et al., 2009); spindles are thalamocortical oscillatory bursts with organisation beyond a single event (Fernandez & Lüthi, 2020). This motivates retaining both event and background information for sleep staging (Iber, 2007). ISRUC-S3 gains 10.5 FT and 9.0 LP points over infilling, and HMC gains 2.7 and 3.3 FT points over micro-only and macro-only; all hold in every seed. HMC has 15 test subjects per seed, compared with one for ISRUC-S3. The gains over infilling in all six cross-dataset directions extend this evidence to changed recording conditions (Table 19). MSDE learns context within each input window; sleep-sequence models additionally model successive 30-second epochs (Phan et al., 2019). The exploratory event-stratified probe on ISRUC-S3 in Appendix K.4 relates classification gains to sigma-burst-containing N2 samples and stage-specific confusions. It complements the transfer comparisons by locating a classification benefit within a fixed stage label.

Predictive information need not depend on a long input or on a discrete event sequence. Monitoring error-related potentials (ErrPs) comprise successive potential components when a person observes an erroneous action by an external agent (Chavarriaga & Millán, 2010; Chavarriaga et al., 2014), whereas steady-state visual evoked potentials (SSVEPs) reflect sustained responses at the stimulation frequency and its harmonics (Norcia et al., 2015). Both ErrP and BETA use one-second inputs in our evaluation. On ErrP, gains over infilling (+10.4 FT) and the strongest reproduced foundation model, REVE-B (+8.8), hold in two of three seeds; the gain over micro-only (+11.3) holds in all three. On the 40-target BETA task (Liu et al., 2020), with seven test subjects per seed, gains over infilling (+3.4) and no prediction (+6.0) hold in every seed; MSDE also exceeds its strongest reconstruction control by 3.3. Together, these objective comparisons span transient and sustained responses within the input window.

Table 21: Selected motor-imagery, error-monitoring and SSVEP comparisons: MSDE minus the named comparator in balanced-accuracy percentage points. Means use paired seed differences, without additional significance tests. REVE-B is the highest-FT-mean reproduced foundation model on ErrP; BETA's raw control has the highest FT mean among the three tested mask ratios.

Task	Setting	Comparator	Seed 5	Seed 10	Seed 42	Mean
BCIC-IV-2a	FT	Latent infilling	+4.17	+6.60	+13.54	+8.1
	FT	Micro-only	+3.99	+9.03	+3.30	+5.4
SHU-MI	FT	Latent infilling	+5.48	+0.05	+3.35	+3.0
	FT	Micro-only	+4.44	+5.58	+0.83	+3.6
ErrP	FT	Latent infilling	+23.59	+11.37	-3.82	+10.4
	FT	REVE-B	+12.91	+14.55	-0.99	+8.8
	FT	Micro-only	+11.43	+16.49	+5.83	+11.3
BETA	FT	Latent infilling	+2.85	+4.91	+2.32	+3.4
	FT	No prediction	+4.46	+6.34	+7.05	+6.0
	FT	Raw reconstruction (0.50)	0.00	+7.41	+2.59	+3.3

A fixed label raises a different issue: reproducible temporal structure can remain present within a sustained condition. FACED records video-elicited EEG with emotion-related spectral and spatial information (Chen et al., 2023); its ten-second inputs carry one emotion label while activity continues to fluctuate. MSDE exceeds its strongest reconstruction control by 2.5 FT points and its strongest reproduced foundation model, REVE-B, by 9.7, both in every seed. Removing prediction costs 5.2 points and using micro-only costs 5.0, also in every seed, with 13 test subjects per seed. The stimulus-position correlations and spectrum-preserving interventions in Appendix K make this temporal motivation concrete. Across five emotion datasets, shared time-dependent signal structure is concentrated below 3 Hz; on FACED, MSDE's higher classification score accompanies greater dependence on temporal alignment than Scratch or infilling. The regional and frequency contrasts further locate this dependence within the recorded input.

The clinical example concerns a more persistent condition. AD65 distinguishes Alzheimer's disease, frontotemporal dementia and healthy controls using eyes-closed resting EEG (Miltiadous et al., 2023). AD-associated slowing, reduced fast-rhythm coherence and reduced complexity (Jeong, 2004) motivate learning temporal statistics despite the fixed diagnosis label. MSDE gains 15.9 FT points over infilling in all seeds, 16.1 over the strongest reconstruction control and 7.0 over the strongest reproduced foundation model, REVE-L, in two of three seeds each; there are nine test subjects per seed. The frozen-state readouts in Section 5.1 connect this diagnostic task to information in the representation: theta/alpha balance and alpha-peak frequency can be recovered from frozen states (Table 28). The spectrum-equalisation intervention in Appendix L.2 then tests how amplitude-spectral distinctions contribute to the adapted classifier's decisions.

Table 22: Selected emotion, sleep and diagnostic comparisons, using the same units and averaging as Table 21. Raw controls have the highest FT task mean among mask ratios 0.25, 0.50 and 0.75. The named REVE variants are the highest-FT-mean reproduced foundation models on FACED and AD65. The AD65 LP reconstruction row uses its FT-selected control.

Task	Setting	Comparator	Seed 5	Seed 10	Seed 42	Mean
FACED	FT	Latent infilling	+10.40	+7.48	-0.33	+5.9
	FT	Raw reconstruction (0.25)	+4.04	+3.35	+0.24	+2.5
	FT	REVE-B	+10.78	+8.93	+9.31	+9.7
	FT	No prediction	+5.92	+7.38	+2.23	+5.2
	FT	Micro-only	+7.77	+3.82	+3.37	+5.0
ISRUC-S3	FT	Latent infilling	+1.21	+21.26	+8.90	+10.5
	FT	Raw reconstruction (0.50)	+1.12	+14.71	+11.51	+9.1
	FT	Macro-only	+3.52	+7.49	+3.17	+4.7
	LP	Latent infilling	+14.33	+8.52	+4.27	+9.0
HMC	FT	Latent infilling	+3.77	+3.63	+4.31	+3.9
	FT	Raw reconstruction (0.25)	+1.52	+2.99	+2.18	+2.2
	FT	Micro-only	+1.33	+3.96	+2.82	+2.7
	FT	Macro-only	+3.19	+3.87	+2.75	+3.3
AD65	FT	Latent infilling	+24.22	+13.56	+9.93	+15.9
	FT	Raw reconstruction (0.50)	+22.06	-2.81	+29.10	+16.1
	FT	REVE-L	+16.67	+13.76	-9.58	+7.0
	FT	Micro-only	+13.30	+3.08	-1.89	+4.8
	FT	Macro-only	+24.80	-0.85	+27.10	+17.0
	LP	Latent infilling	-0.45	+5.51	+3.72	+2.9
	LP	Raw reconstruction (0.50)	-8.31	+3.13	+5.18	0.0

J AUXILIARY CONSTRAINTS FOR FUTURE-STATE PREDICTION

This appendix defines the auxiliary constraints introduced in Section 3.4 and derives bounds on pooled-state variability and composed prediction error. The encoder, teacher and transition operator are held fixed in each comparison; target states have finite second moments.

J.1 WINDOW DISCRIMINATION AND POOLED VARIABILITY

Because student and teacher co-evolve, a constant encoder and matching teacher can make future prediction trivial; a transition operator that returns the constant state also satisfies composition consistency. This motivates constraints on both window discrimination and pooled variability.

Anchoring. From a single student encoding of window i, let A_i, B_i be the normalised means of valid states in its first and second temporal halves. Only windows with two nonempty halves enter this loss; let their number in the current loss computation be $n \geq 2$. With temperature τ, the two directional losses are

$$\ell_i^{\rightarrow} = -\log \frac{\exp(A_i^\top B_i/\tau)}{\sum_k \exp(A_i^\top B_k/\tau)}, \qquad \ell_i^{\leftarrow} = -\log \frac{\exp(A_i^\top B_i/\tau)}{\sum_k \exp(A_k^\top B_i/\tau)}. \tag{4}$$

The implemented loss is $\mathcal{L}_{\text{inst}} = (2n)^{-1} \sum_i (\ell_i^{\rightarrow} + \ell_i^{\leftarrow})$. If all A_i equal one vector and all B_i equal another, every candidate in each denominator has the same score, so every directional loss equals $\log n$. An attained loss below $\log n$ excludes that solution. The condition holds for a sample-independent state trajectory when windows share the same valid time positions; different masks need not produce identical pooled vectors.

Pooled variance. For $B \geq 2$ windows with at least two valid tokens in total, let $P_i \in \mathbb{R}^D$ be the full-window mean formed by averaging valid token states. The pooled term includes all batch rows; a fully padded window is assigned the zero vector by the masked pooling operation. Thus B can

differ from the usable-half count n for anchoring. Define

$$\bar{P} = \frac{1}{B}\sum_i P_i, \qquad v_j = \frac{1}{B}\sum_i (P_{ij} - \bar{P}_j)^2, \qquad V_P = \frac{1}{D}\sum_j (\gamma - \sqrt{v_j + \epsilon})_+. \tag{5}$$

Here γ is the target standard deviation, $\epsilon > 0$ is the numerical stabiliser and $(x)_+ = \max(x, 0)$. For each dimension, $\sqrt{v_j + \epsilon} \geq \gamma - (\gamma - \sqrt{v_j + \epsilon})_+$. Averaging over dimensions and using nonnegativity yields

$$\frac{1}{D}\sum_j \sqrt{v_j + \epsilon} \geq (\gamma - V_P)_+. \tag{6}$$

Jensen's inequality gives

$$\frac{1}{D}\sum_j v_j + \epsilon \geq \left(\frac{1}{D}\sum_j \sqrt{v_j + \epsilon}\right)^2 \geq (\gamma - V_P)_+^2. \tag{7}$$

Combining this with $v_j \geq 0$ and the identity

$$\frac{1}{B^2}\sum_{i,k} \|P_i - P_k\|^2 = 2\sum_j v_j \tag{8}$$

gives

$$\frac{1}{B^2 D}\sum_{i,k} \|P_i - P_k\|^2 \geq 2\left[(\gamma - V_P)_+^2 - \epsilon\right]_+. \tag{9}$$

Thus a sufficiently small V_P requires variation between the pooled states retained downstream. For the implemented $\gamma = 1$ and $\epsilon = 10^{-4}$, $V_P < 0.99$ implies strictly positive average squared distance between pooled states. If all P_i are identical, $V_P = 0.99$ regardless of token variation within each window.

The same variance hinge also acts on the collection of valid tokens. For each collection of N observations X_i, the accompanying covariance penalty is $D^{-1}\sum_{j\neq k}\widehat{C}_{jk}^2$, where $\widehat{C} = (N-1)^{-1}\sum_i (X_i - \bar{X})(X_i - \bar{X})^\top$. The variance bound uses the population denominator N, whereas this penalty uses $N - 1$. Token and pooled terms have equal weight; the covariance penalty discourages redundant variation without requiring a full-rank minibatch covariance.

J.2 Direct accuracy and composition consistency

For a source state $S \in \mathbb{R}^D$ and positive offsets a, b with $a + b \leq 1\,\mathrm{s}$, define the direct and composed predictions

$$d = \Phi(S, a + b), \qquad c = \Phi(\Phi(S, a), b). \tag{10}$$

The second route takes a predicted state as its intermediate input; the consistency loss does not use the observed intermediate state. Both routes receive gradients. The implemented elementwise Smooth L1 penalty has $\beta = 1$:

$$\rho(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq 1, \\ |r| - \frac{1}{2}, & |r| > 1. \end{cases} \tag{11}$$

Training averages $D^{-1}\sum_j \rho(c_j - d_j)$ over sampled triples.

Take a fixed distribution q on valid triples with $a + b \in [0.5, 1.0]\,\mathrm{s}$, within the directly supervised micro range, and let Z_{a+b} be the teacher state at the endpoint. Define the direct prediction risk and path discrepancy as

$$R_q = \mathbb{E}_q\|d - Z_{a+b}\|^2/D, \qquad C_q = \mathbb{E}_q\, D^{-1}\sum_j \rho(c_j - d_j). \tag{12}$$

For $|r| \leq 1$, $|r| = \sqrt{2\rho(r)}$; for $|r| > 1$, $\rho(r) = |r| - 1/2$ and $\sqrt{2\rho(r)} \geq 1$. Hence $|r| \leq \sqrt{2\rho(r)} + \rho(r)$ in both cases. Averaging over triples and coordinates and applying Jensen gives

$$\mathbb{E}_q\|c - d\|_1/D \leq \sqrt{2C_q} + C_q. \tag{13}$$

Similarly, Cauchy–Schwarz gives $\mathbb{E}_q\|d - Z_{a+b}\|_1/D \le \sqrt{R_q}$. The triangle inequality therefore yields

$$\mathbb{E}_q\|c - Z_{a+b}\|_1/D \le \sqrt{R_q} + \sqrt{2C_q} + C_q. \tag{14}$$

Direct accuracy and path consistency therefore jointly control mean absolute composed prediction error. The result does not assume that a stochastic process's conditional-mean predictions form an exact semigroup.

Relation to the training samplers. The bound evaluates direct error and consistency on the same distribution q. To connect it to both training objectives, choose triples whose total offset lies in the micro prediction range $[0.5, 1.0]$ s. The training triple sampler also permits shorter total offsets, while the prediction loss first averages pairs within horizon bins and then weights the bins equally. Thus the two raw training means have different sampling distributions.

More precisely, let Q be the training triple distribution and A its subset with directly supervised total offsets. If $p = Q(A) > 0$ and $q = Q(\cdot \mid A)$, nonnegativity gives $C_q \le \mathbb{E}_Q[D^{-1}\sum_j \rho(c_j - d_j)]/p$. Let P_{micro} be the bin-balanced training pair distribution, and q_{pair} the source–endpoint marginal of q. If q_{pair} is absolutely continuous with respect to P_{micro} with density at most κ, then

$$R_q \le \kappa\, \mathbb{E}_{P_{\mathrm{micro}}}\|\Phi(S,h) - Z_h\|^2/D. \tag{15}$$

These relations specify the coverage and reweighting needed to substitute training risks into the bound. When reporting it empirically, evaluating both quantities on the same covered triples avoids estimating density ratios. A sampler containing no covered triples cannot supply this comparison; in particular, deterministic adjacent-token triples can have total offsets below 0.5 s. The consistency analysis concerns the sampled short offsets.

K Temporal alignment and event-related diagnostic analyses

These analyses connect the transfer comparisons to temporal structure in the recorded inputs and to event-related classification errors. Correlation summaries, test-time interventions and frozen-state probes address different parts of this connection; their protocols and effect sizes are reported separately below.

K.1 Shared stimulus-position structure across subjects

For each subject, we correlate the signal at a given stimulus position with the mean signal from the other subjects at the same position. FACED positions identify a window within a shared video; positions in the SEED datasets identify a trial and window index. Controls retain the position key while displacing the signal by 0.25–0.5 s, or use a segment with a different label. The analysis separates components below and above 3 Hz and does not use any model representation. Table 23 shows that the strongest shared structure is in the low-frequency component, with same-position correlations exceeding shifted correlations in every dataset.

Table 23: Cross-subject correlation summaries. The three central columns concern components below 3 Hz; the final column concerns components above 3 Hz. Ranges are descriptive values, not confidence intervals. Each subject is excluded from the mean signal with which it is correlated.

Dataset	Subjects	Same position	Shifted	Different segment	Same, >3 Hz
FACED	123	0.23–0.29	≈ -0.01	≈ 0.01	0.01–0.03
SEED	15	0.09–0.16	0.03–0.06	≈ 0	≈ 0
SEED-IV	15	0.09–0.17	0.01–0.04	≈ 0	≈ 0
SEED-VII	20	0.10–0.19	0.02–0.06	≈ 0	≈ 0.01
SEED-V	16	0.03–0.06	≈ -0.01	≈ 0	≈ 0

K.2 Spectrum-preserving interventions on FACED

The analysis uses fixed fine-tuned MSDE, infilling, raw-reconstruction (MAE) and Scratch classifiers on split seeds 5, 10 and 42. Each split contains 1,092 test windows from 13 subjects. Inter-

ventions are applied only at test time, with unchanged labels and classifiers. For a channel $x_c[n]$ of length N, a common circular displacement by k samples gives

$$x'_c[n] = x_c[(n-k) \bmod N], \qquad X'_c[\omega] = e^{-\mathrm{i}\omega k} X_c[\omega]. \tag{16}$$

Thus $|X'_c[\omega]| = |X_c[\omega]|$, and a common displacement cancels in $X'_c[\omega]\overline{X'_d[\omega]}$. The manipulation tests alignment to the window reference while preserving whole-window amplitude spectra and, for common shifts, cross-spectra.

Table 24: FACED diagnostic scores and losses under common circular shifts. Loss is the original score minus the shifted score; excess loss is MSDE minus Scratch. Scores, changes and paired contrasts are rounded separately. The unshifted MSDE advantage is 15.6 points over Scratch and 5.9 over infilling; at 0.5 s its loss exceeds infilling's by 4.9 points. Changes are percentage points.

Unshifted classification scores (%)			
MSDE	Infilling	MAE	Scratch
61.3	55.4	54.6	45.8

Displacement (s)	MSDE loss	Scratch loss	Excess loss
0.05	2.0	0.2	1.8
0.10	3.9	1.6	2.3
0.25	21.5	9.0	12.5
0.50	41.9	18.5	23.4
1.00	45.5	17.1	28.4

K.3 Frequency, region and interval specificity

To determine whether the whole-window effect is concentrated in a particular frequency range, channel group or short interval, the same 0.5 s displacement is restricted to each of these in turn. The interval analysis tests each of the ten one-second intervals in a FACED input. Frequency-selective interventions shift either the component below 3 Hz or its higher-frequency complement before recombination. Regional interventions shift only the selected channels, thereby also changing their timing relative to the remaining channels. Table 25 reports the complete set of regional and frequency contrasts. The excess loss is larger for the low-frequency component and varies across regions, with frontal and parietal effects. Across the ten individual one-second interventions, excess loss ranges from -2.9 to $+5.7$ points, smaller than the $+23.4$-point whole-window contrast.

Table 25: FACED excess loss under selective 0.5 s circular shifts: MSDE's score loss minus Scratch's, in percentage points. Positive values indicate greater MSDE sensitivity. Entries specifying only a region use the full frequency range.

Shifted component or channel group	Excess loss
All channels, below 3 Hz	$+22.0$
All channels, above 3 Hz	$+4.1$
Frontal, full frequency range	$+6.7$
Frontal, below 3 Hz	$+5.6$
Frontal, above 3 Hz	$+1.8$
Lateral frontal: F7/F8/AF3/AF4	$+2.8$
Parietal	$+4.2$
Central / temporal / occipital	$+0.7$ to $+1.4$
Mid-frontal: Fz/F3/F4	-0.3
Frontopolar: FP1/FP2	-2.6

These selective shifts locate MSDE's excess alignment sensitivity chiefly in the low-frequency component: displacing activity below 3 Hz produces an excess loss close to that of shifting the entire signal. No individual one-second intervention reproduces the whole-window contrast, linking the larger excess loss to perturbing alignment across more of the input. The regional pattern further identifies frontal and parietal channel groups as contributors, with these interventions probing both alignment to the window and timing relative to unshifted channels.

K.4 SLEEP-STAGE ERRORS AND SIGMA-BURST STRATIFICATION

To locate the sleep-staging gains more precisely, we ask which stage distinctions and event-containing windows are better classified from frozen pretrained states. Frozen encoders are evaluated with a common standardised logistic-regression probe and subject-separated folds. ISRUC-S3 uses leave-one-subject-out evaluation over ten subjects and 8,846 samples; the parallel HMC evaluation uses five folds over 40 subjects and 36,931 samples. Samples are stratified by stage and by a simplified detector of sigma bursts, large slow waves and alpha activity. Table 26 compares the four pretrained encoder conditions in both datasets.

Table 26: Overall balanced accuracy (%) of pretrained encoders for the event-analysis probes. These subject-fold diagnostics have their own evaluation protocol and are separate from the benchmark LP table.

Encoder	ISRUC-S3	HMC
MSDE	70.4	47.5
No prediction	64.7	46.8
Masked infilling	66.4	48.1
Raw reconstruction (MAE)	65.5	49.0

The ISRUC-S3 contrasts against the three other pretrained encoders are largest in REM ($+6.8$ to $+12.8$ points) and N1 ($+5.0$ to $+12.6$), with fewer REM-to-N2/N3 and N1-to-W errors. The burst structure of sleep spindles motivates a further comparison within N2 (Fernandez & Lüthi, 2020). Within N2, we subtract each control's classification accuracy from MSDE's separately in detector-positive and detector-negative samples, then compare these two advantages. This contrast is approximately $+7.5$ points, with the same direction against all three pretrained controls. Because the stage label is fixed, the positive contrast locates a larger MSDE advantage in burst-positive N2 windows, beyond a difference in the mixture of sleep stages. Groups use automatic detections, with sigma power and signal quality otherwise uncontrolled; the outcome is sleep-stage classification. The ISRUC-S3 N2 contrasts have reported p values of 0.037–0.23 within approximately 30 exploratory comparisons over ten subjects.

HMC shows a different pattern: N2 burst contrasts are $+5.7$ points against no prediction, $+6.6$ against reconstruction and -0.9 against infilling, while REM contrasts range from -0.8 to -2.7 points. Automatic detections cover 56% of N2 samples in ISRUC-S3 and 7% in HMC, so the strata also differ in their dataset coverage. HMC also has no overall probe advantage over infilling or MAE (Table 26); the benefit and its association with detected bursts therefore depend on the dataset and comparator.

For a complementary view after supervised adaptation, we inspect the separate fine-tuned ISRUC-S3 seed-42 checkpoint. Overall balanced accuracy is 81.0% for MSDE and 71.8% for Scratch; the stagewise gains are 21.9 points in REM and 22.0 in N2, on one held-out subject. This check identifies REM and N2 as stages contributing to the FT gain on that subject, complementing the frozen probes that assess stage information accessible before adaptation.

L PHYSIOLOGICAL READOUTS AND COMPONENT INTERVENTIONS

These diagnostics compare MSDE trained on the 10% TUH subset with masked EMA-state infilling and raw-signal reconstruction (MAE), using the same encoder architecture and pretraining data. Classifier interventions use split seeds 5, 10 and 42; physiological regressions use split seed 42. Both are separate from the cross-fold sleep-event analysis in Appendix K.4. Scores and changes are averaged across the available splits before rounding; control-adjusted bootstrap intervals follow the protocol below.

L.1 READOUTS FROM FROZEN STATES

For each EEG window, the pretrained encoder's pooled frozen state is the input to a ridge-regression readout; the encoder has not been fine-tuned on downstream labels. Targets are computed from

the EEG itself, rather than from the downstream class prediction. Training and test subjects do not overlap; five-fold cross-validation within the training set selects the ridge penalty. Reported values are test R^2 on split seed 42, with the task summary averaging its target-level scores.

Sleep targets include relative band powers, spindle time fraction and slow-wave counts. AD65 targets include posterior alpha-peak frequency, the log theta/alpha ratio and a log slowing index. BETA targets describe log occipital stimulus-frequency and harmonic power; FACED targets include frontal alpha asymmetry and log slow-component power. The six-task comparison uses these task-specific descriptor sets, and the overall mean weights tasks equally.

Table 27: Frozen-state physiological readouts (test R^2, split seed 42), averaged over targets within each task and equally over tasks in the final row. Bold marks the best value per row.

Task	MSDE	Infilling	MAE
SleepEDF	**0.821**	0.752	0.715
AD65	**0.622**	0.507	0.552
ISRUC-S3	**0.587**	0.464	0.521
HMC	**0.560**	0.460	0.489
BETA	**0.291**	0.268	0.282
FACED	0.354	0.338	**0.377**
Mean	**0.539**	0.465	0.489

MSDE leads on five of the six tasks, averaging $R^2 = 0.539$ compared with 0.465 for infilling and 0.489 for MAE. Its advantage spans all three sleep datasets, AD65 and BETA, covering both transient-event descriptors and spectral properties of ongoing activity.

Table 28: Representative physiological readouts from unadapted pretrained encoders (test R^2, split seed 42).

Task	Descriptor	MSDE	Infilling	MAE
SleepEDF	Spindle time fraction	**0.900**	0.823	0.771
SleepEDF	Slow-wave count	**0.697**	0.629	0.594
AD65	Log theta/alpha ratio	**0.794**	0.489	0.576
AD65	Alpha peak frequency	**0.235**	-0.252	0.074
HMC	Alpha relative power	**0.634**	0.388	0.432
BETA	Log occipital power (16–32 Hz)	**0.782**	0.730	0.729

The representative readouts make these improvements concrete: MSDE more accurately retains spindle time fraction and slow-wave counts on SleepEDF, the log theta/alpha ratio and alpha-peak frequency on AD65, and occipital band power on BETA. These properties are accessible before downstream fine-tuning, connecting the pretraining objective to physiological information in the transferable state.

L.2 Fixed-classifier component interventions

The FT diagnostics use fixed fine-tuned classifiers; the LP diagnostics use fixed linear-probe classifiers and frozen encoders. Interventions are applied only at evaluation, with no model update. Band-selective operations are applied over the whole window with zero-phase frequency-domain processing. Table 29 specifies the target and control manipulations.

Table 29: Target and control interventions used in the physiological diagnostics.

Task	Target intervention	Control intervention
BETA	Least-squares removal of sinusoids at the labelled stimulus frequency f and harmonics $2f, 3f$	Remove another stimulus frequency, $f \pm 2$ Hz
BETA	Remove 7.5–33 Hz from occipito-parietal channels	Remove the same band from frontal channels
BCIC-IV-2a	Remove 8–30 Hz from FC/C/CP channels	Remove 31–53 Hz from the same channels
AD65	Replace each channel's amplitude spectrum with its test-cohort mean, retaining phase	Randomise phase, retaining the amplitude spectrum
ErrP	Remove 0.5–8 Hz	Remove 20–28 Hz
ErrP	Subtract the class-average event-related potential (ERP)	Subtract the global-average ERP
Sleep	Separately remove delta (0.3–2), theta (4–8), sigma (11–16) or alpha (8–12 Hz)	Remove 20–25 Hz
FACED	Remove frontal 0.5–3 Hz activity	Remove frontal 3–6 Hz activity

The BETA frequency intervention uses the stimulus label to choose the removed frequency, and class-average ERP subtraction is also label-conditioned. AD65 equalisation replaces each channel's amplitude spectrum with the test-cohort mean while retaining phase. For the band-removal experiments, target and control conditions use the same processing procedure with the frequency or channel selection specified in Table 29.

Table 30: FT diagnostic scores and interventions on stimulus-response, motor-imagery and clinical tasks (three-seed means). Original rows give balanced accuracy (%); intervention rows give changes from those scores in percentage points.

Task	Condition	MSDE	Infilling	MAE
BETA	Original accuracy	31.9	28.5	28.5
	Remove labelled frequency + harmonics	-23.1	-20.5	-20.5
	Control: remove other frequency	+0.1	-0.1	-0.1
	Occipito-parietal 7.5–33 Hz	-26.4	-22.3	-21.2
	Control: frontal 7.5–33 Hz	-0.4	-0.2	-0.1
BCIC-IV-2a	Original accuracy	49.8	41.7	52.1
	Sensorimotor 8–30 Hz	-16.7	-11.3	-24.2
	Control: sensorimotor 31–53 Hz	+1.1	+0.6	-0.1
AD65	Original accuracy	65.2	49.3	49.1
	Amplitude-spectrum equalisation	-32.6	-13.4	-11.9
	Control: phase randomisation	-4.7	-0.9	-3.0
ErrP	Original accuracy	71.3	60.9	68.8
	Remove 0.5–8 Hz	-20.2	-10.6	-12.2
	Control: remove 20–28 Hz	+1.4	+0.7	+0.8
	Subtract class-average ERP	-8.8	-3.7	-7.8
	Control: subtract global-average ERP	-2.4	-0.7	-2.0

Table 31: FT diagnostic scores and interventions on sleep and emotion tasks (three-seed means). Original rows give balanced accuracy (%); intervention rows give changes from those scores in percentage points.

Task	Condition	MSDE	Infilling	MAE
ISRUC-S3	Original accuracy	80.7	70.2	71.6
	Remove 0.3–2 Hz	-32.1	-23.6	-29.7
	Remove 4–8 Hz	-18.0	-3.2	-3.1
	Remove 11–16 Hz (sigma)	-4.8	-2.3	+0.9
	Control: remove 20–25 Hz	+0.6	+0.6	-0.0
HMC	Original accuracy	69.8	65.9	66.7
	Remove 11–16 Hz (sigma)	-14.4	-9.4	-12.8
	Control: remove 20–25 Hz	-8.9	-5.5	-6.9
SleepEDF	Original accuracy	74.5	74.0	72.9
	Remove 11–16 Hz (sigma)	-9.2	-7.6	-7.3
	Control: remove 20–25 Hz	+0.1	-0.1	+0.1
FACED	Original accuracy	61.3	55.5	54.6
	Frontal 0.5–3 Hz	-15.6	-12.6	-11.4
	Control: frontal 3–6 Hz	-1.1	-0.5	-0.3

BETA's target-frequency and spatial controls give converging frequency and regional evidence: occipito-parietal 7.5–33 Hz removal lowers MSDE by 26.4 points, whereas frontal removal lowers it by 0.4. On AD65, spectrum equalisation lowers MSDE by 32.6 points, compared with 13.4 and 11.9 for infilling and MAE; phase randomisation has smaller effects on all three models. The AD65 result complements the theta/alpha and alpha-peak readouts: its frozen states preserve clinically relevant spectral descriptors, and its classification gain depends on amplitude-spectral distinctions.

L.3 MSDE'S ADVANTAGE UNDER TARGET AND CONTROL INTERVENTIONS

Let $G_c = A_{\mathrm{MSDE},c} - A_{\mathrm{baseline},c}$ denote the balanced-accuracy gap under condition c. The control-adjusted reduction in this advantage is

$$E = (G_{\mathrm{original}} - G_{\mathrm{target}}) - (G_{\mathrm{original}} - G_{\mathrm{control}}) = G_{\mathrm{control}} - G_{\mathrm{target}}. \tag{17}$$

A positive value means that the MSDE advantage is smaller under the target intervention than under the control intervention.

Table 32 reports intervals from 2,000 bootstrap resamples. Within a seed, datasets with at least five test subjects use subject resampling; smaller test groups use within-class window resampling. BCIC-IV-2a, ErrP and ISRUC-S3 each contain one test subject per split; their intervals are conditional on those held-out subjects. Results average the three split seeds. The tail fraction is the proportion of bootstrap contrasts at or below zero; the 18 contrasts comprise paired comparisons against two recovery objectives within the tested tasks and components.

Table 32: Control-adjusted changes in MSDE advantage (percentage points), computed as the MSDE–comparator gap under the control intervention minus that under the target intervention. Bootstrap intervals and tail fractions follow Appendix L.3.

Task / component	Comparator	Effect [95% interval]	Tail fraction
BETA: stimulus frequency	Infilling	$+2.7$ $[+0.4, +4.9]$	0.012
	MAE	$+2.7$ $[+1.4, +4.1]$	< 0.001
ErrP: 0.5–8 Hz	Infilling	$+10.2$ $[+8.0, +12.4]$	< 0.001
	MAE	$+8.6$ $[+6.3, +10.9]$	< 0.001
AD65: spectrum equalisation	Infilling	$+15.4$ $[-5.7, +31.7]$	0.085
	MAE	$+19.1$ $[-0.9, +30.7]$	0.037
ISRUC-S3: theta	Infilling	$+14.9$ $[+12.5, +17.3]$	< 0.001
	MAE	$+15.6$ $[+13.0, +18.0]$	< 0.001
ISRUC-S3: sigma	Infilling	$+2.6$ $[+0.7, +4.7]$	0.004
	MAE	$+6.3$ $[+4.3, +8.4]$	< 0.001
HMC: sigma	Infilling	$+1.6$ $[+0.2, +3.0]$	0.017
	MAE	-0.5 $[-1.7, +0.7]$	0.79
SleepEDF: sigma	Infilling	$+1.8$ $[+1.0, +2.7]$	< 0.001
	MAE	$+1.9$ $[+1.3, +2.5]$	< 0.001
FACED: frontal slow band	Infilling	$+2.3$ $[+0.1, +4.2]$	0.020
	MAE	$+3.4$ $[+1.1, +5.5]$	0.003
BCIC-IV-2a: μ/β	Infilling	$+5.8$ $[+2.4, +9.4]$	< 0.001
	MAE	-6.3 $[-9.7, -2.7]$	–

Of the 18 reported contrasts, 16 have positive point estimates and 14 have intervals entirely above zero. The pattern spans stimulus responses, sleep and emotion: removing task-related components reduces MSDE's advantage more than the corresponding control. The complete table provides both objective comparisons for each component, supporting the cross-domain examples in Section 5.2.

L.4 Stage-stratified and frozen-classifier extensions

Class-specific recall changes under sleep-band removal are computed from the per-seed records (Table 33). Stage names have been checked against the official label mapping: classes 0–4 correspond to W, N1, N2, N3 and REM, respectively. Figure 4c shows N2 recall loss under sigma removal and the control band, including the individual split seeds. Sigma removal produces larger N2 recall losses for MSDE than either recovery objective in all three sleep datasets, linking its classification behaviour to a characteristic N2 frequency band (Iber, 2007). On ISRUC-S3, theta removal also has a larger effect on MSDE's REM and N1 recall, whereas strong delta dependence in N3 is shared across objectives.

Table 33: Stage-stratified recall changes after band removal (percentage points), averaged over three split seeds.

Band / stage	Dataset	MSDE	Infilling	MAE
Sigma / N2	ISRUC-S3	-24.3	-17.5	+2.5
Sigma / N2	HMC	-25.0	-14.5	-18.9
Sigma / N2	SleepEDF	-47.5	-42.1	-42.5
Theta / REM	ISRUC-S3	-58.8	-28.3	-36.7
Theta / N1	ISRUC-S3	-37.6	-29.9	-16.1
Theta / N1	HMC	-9.7	-6.5	-11.5
Delta / N3	ISRUC-S3	-72.5	-59.2	-81.5
Delta / N3	HMC	-76.8	-73.9	-77.6
Delta / N3	SleepEDF	-84.1	-86.1	-83.0

The LP interventions examine component dependence without updating the encoder during downstream classification. These interventions are evaluated on tasks whose LP classifiers perform above chance (Table 34). For ISRUC-S3, MSDE's clean advantage over infilling is 9.0 points; the reported delta-band losses of 29.0 and 19.8 points nearly eliminate that gap. For HMC, sigma removal lowers MSDE by 7.3 points, versus 3.5 and 4.3 for infilling and MAE. These results show that classification based on frozen MSDE states already uses the physiological components identified by the interventions, extending the regression evidence to downstream decisions without encoder fine-tuning.

Table 34: Interventions on fixed LP classifiers using frozen encoders (three-seed means). Original rows give balanced accuracy (%); intervention rows give changes from those scores in percentage points.

Task	Condition	MSDE	Infilling	MAE
ISRUC-S3	Original accuracy	70.7	61.7	57.1
	Delta change	-29.0	-19.8	-19.8
SleepEDF	Original accuracy	66.8	63.2	63.3
	Delta change	-29.6	-23.0	-25.9
HMC	Original accuracy	52.3	50.5	51.1
	Sigma change	-7.3	-3.5	-4.3
AD65	Original accuracy	51.7	48.7	51.7
	Spectrum equalisation change	-18.5	-13.7	-17.3