When Masked Reconstruction Becomes Spatial Interpolation: What EEG Foundation Models Miss

Anonymous authors
Paper under double-blind review

Abstract

EEG foundation models pretrained by masked reconstruction often trail smaller supervised models, and their gains may depend on task-specific heads or fine-tuning rather than on the pretrained representation. This paper investigates what masked reconstruction transfers to downstream EEG tasks, and why it behaves differently from images. A standard criss-cross Transformer is pretrained on 5,200 hours of EEG data and evaluated on seven benchmarks under linear probing and fine-tuning with a standardised head. When predicting masked EEG patches, a linear map from the visible electrodes at the same time step recovers 66–84% of the encoder's improvement over a simple predictor, so the pretext task can largely be solved by spatial interpolation. Accordingly, during pretraining, spatial attention rapidly converges to a largely trial-invariant map concentrated on neighbouring electrodes, which the same architecture trained on task labels does not form. In contrast, ViT-MAE on images exhibits content-dependent attention under the same analysis. On BCI IV-2a, residual-stream probes show high subject decodability (0.93) but weak class decodability (0.30); across datasets, per-electrode band power is strongly represented before mean pooling but partly lost by it, while channel covariance is substantially less accessible. In a released EEG foundation model, 99% of the pooled representation's energy lies in a component shared across trials, comparable to random initialisation, and falls to 30% after fine-tuning. Replacing learned spatial attention with fixed maps has little effect on downstream performance on most benchmarks, whereas adding covariance and differential entropy features improves performance by up to 11 percentage points on per-subject motor imagery. Together, these results diagnose a modality-specific spatial shortcut in masked EEG pretraining and motivate objectives that explicitly target input-dependent cross-channel structure.

1 Introduction

Electroencephalography (EEG) is a widely used non-invasive technique for measuring brain activity, offering high temporal resolution at relatively low acquisition cost. It supports a broad range of applications, including sleep monitoring (de Gans et al., 2024), stress assessment (Vanhollebeke et al., 2022), and clinical neurological analysis (Zhang et al., 2023). EEG decoding has traditionally relied on classical machine-learning approaches built on hand-crafted signal features (Aggarwal & Chugh, 2022). Although effective in individual paradigms, such pipelines typically require substantial task-specific feature engineering and adaptation, limiting their ability to generalise across datasets and applications (Hong et al., 2025).

The rise of deep learning (LeCun et al., 2015) shifted EEG decoding towards end-to-end representation learning, with convolutional, recurrent, graph-based, and Transformer architectures achieving strong performance across a range of paradigms (Lawhern et al., 2018; Song et al., 2022; Craik et al., 2019). However, these models are typically trained separately for each dataset or task, limiting their ability to exploit the growing volume of heterogeneous EEG data. This has motivated EEG foundation models (FMs), which aim to learn transferable representations through large-scale self-supervised pretraining (Sun et al., 2025). This paradigm is particularly attractive for EEG, where labeled data is expensive to acquire and no single dataset covers the full diversity of recording con-

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.

ditions and paradigms. Accordingly, a growing body of work has explored EEG FMs (Wang et al., 2024; 2025; Yi et al., 2023; Yang et al., 2023; Ho & Armanfard, 2023; Jiang et al., 2024; Xiao et al., 2025; Dong et al., 2024; Zhou et al., 2025; Ouahidi et al., 2025; Döner et al., 2025; Ma et al., 2026; Fang et al., 2025; Yu et al., 2026). However, EEG presents domain-specific challenges that differ substantially from those of images and text. Its signal-to-noise ratio varies across subjects, sessions, and recording setups, complicating cross-dataset generalisation (Kim et al., 2023), while its structure has dependencies across time and across electrodes (Trifonov, 2016). Despite this, many EEG FMs inherit architectural and pretraining choices originally developed for vision or language. A less examined question is whether the pretraining objective itself is well matched to EEG statistics. Masked reconstruction is effective in vision because recovering missing content requires exploiting image structure (He et al., 2022); in EEG, however, strong spatial redundancy between neighbouring electrodes can make a masked segment highly predictable from nearby channels. This may allow the objective to be solved largely through spatial interpolation, reducing the pressure to learn input-dependent cross-channel interactions that could be useful downstream. Whether pretrained encoders actually learn such input-dependent interactions has not been measured.

This paper investigates what masked reconstruction transfers to downstream EEG tasks and whether spatial interpolation offers a shortcut for predicting masked signals from neighbouring electrodes. The core contributions are as follows.

- Masked EEG reconstruction is shown to be largely solvable by a fixed spatial map, and spatial attention converges early to such a trial-invariant map.
- The resulting representation separates subjects more sharply than tasks and makes channel covariance barely available; supplying covariance and differential entropy (DE) features improves transfer, whereas replacing spatial attention with a fixed map costs little.
- The analysis is extended to four released EEG FMs, suggesting that trial invariance accompanies masked objectives with different prediction targets.

2 Related Work

Shortcut learning in self-supervised pretraining: Self-supervised objectives can be solved through shortcuts that satisfy the pretext task without learning the structure most useful downstream (Kolesnikov et al., 2019; Geirhos et al., 2020). This has been observed for both contrastive learning (Robinson et al., 2021) and masked reconstruction, where lower reconstruction error need not imply better representations (Chen et al., 2022). In vision, masked modelling can induce local attention (He et al., 2022), although locality itself does not preclude strong transfer (Xie et al., 2023), and pixel reconstruction may still overemphasise low-level detail (Liu et al., 2023). In EEG, locality interacts with strong spatial redundancy between neighbouring electrodes. SPOTR (Gui et al., 2026) restricts access to neighbouring patches, while Li et al. (2025) show that adjacent-channel correlation can favour local signal-similarity features. However, how local redundancy shapes spatial attention during masked pretraining remains unclear.

EEG foundation models: Recent EEG FMs differ mainly in their prediction targets, architectures, and masking strategies. BENDR (Kostas et al., 2021) and EEGPT (Wang et al., 2024) combine masked prediction with contrastive or teacher–student objectives, while LaBraM (Jiang et al., 2024), CodeBrain (Ma et al., 2026), Patchcode (Yu et al., 2026), and EEG2Rep (Mohammadi Foumani et al., 2024) replace raw-signal reconstruction with discrete or latent targets. Architectural variants address montage and spatio-temporal structure through channel-wise tokenisation (Yang et al., 2023), criss-cross attention (Wang et al., 2025), or cross-scale sparse attention (Zhou et al., 2025), while REVE (Ouahidi et al., 2025) uses structured masking to reduce local redundancy. Despite these increasingly specialised designs, systematic evaluations report inconsistent downstream gains over simpler masked-reconstruction Transformers (Yang et al., 2026), and weak alignment between pretraining and downstream gradients (Xiong et al., 2025). Tang et al. (2026) further show that several hand-crafted EEG features are linearly decodable from mean-pooled foundation-model representations. The analysis here instead probes representations before and after channel pooling, revealing that pooling itself can discard channel-specific spectral information.

3 EXPERIMENTAL SETUP

3.1 MODELS AND CROSS MODEL ANALYSIS

The primary encoder is a criss-cross Transformer (Zhang & Yan, 2023; Wang et al., 2025), whose separate spatial and temporal attention branches expose electrode-to-electrode attention directly. It has 8 layers of width 256 (6.36M parameters) and is pretrained by masked reconstruction with 60% of tokens masked (see Appendices B and C).

To test whether the findings are biased by this design, the diagnostics are repeated on four released FMs (Section 2): CBraMod (Wang et al., 2025), which shares the criss-cross attention; ST-EEGFormer (Yang et al., 2026) and LaBraM (Jiang et al., 2024), which attend jointly over all tokens and respectively reconstruct raw patches and predict discrete codes; and BIOT (Yang et al., 2023), trained contrastively to match the embedding of a masked input to that of the complete input (see Appendix D). EEGNet (Lawhern et al., 2018), DeepConvNet (Schirrmeister et al., 2017), CTNet (Zhao et al., 2024) and EEG Conformer (Song et al., 2022), trained from scratch on each dataset, serve as supervised references (see Appendix E).

3.2 DATASETS

Pretraining datasets: The primary encoder is pretrained on 17 EEG datasets totalling approximately 5,200 hours (see Appendix F), none of which is used for downstream evaluation. Following Yang et al. (2026), recordings are resampled to 128 Hz and segmented into fixed-length windows (see Appendix F.2).

Downstream datasets: All models are evaluated on seven datasets spanning distinct paradigms (see Table 1; full description in Appendix G.1). For each FM the data is resampled and normalised following their original implementation (see Appendix D). The native rate and z-normalisation is used for supervised models.

Table 1: Downstream evaluation datasets. Frequencies are native sampling rates.

Dataset	Paradigm	Subjects	Channels	Freq. (Hz)	Segment (s)	Classes	Primary metric
BCI IV-2a (Tangermann et al., 2012)	Motor imagery	9	22	250	4.0	4	Accuracy
FACED (Chen et al., 2023)	Emotion	123	32	256	10.0	9	Balanced acc.
Mumtaz (Mumtaz et al., 2017)	Depression	63	19	256	5.0	2	Balanced acc.
SEED-VIG (Zheng & Lu, 2017)	Vigilance	21	17	200	5.0	Regression	Pearson r
ISRUC (Khalighi et al., 2016)	Sleep staging	10	6	200	30.0	5	Balanced acc.
Mental arithmetic (MA) (Zyma et al., 2019)	Mental arithmetic	36	19	256	5.0	2	Balanced acc.
TUEV (Obeid & Picone, 2016)	Event type	370	21	250	5.0	6	Balanced acc.

3.3 DIAGNOSTIC MEASURES

The analysis uses three groups of diagnostics, each addressing one part of the proposed spatial shortcut: the prediction of masked EEG patches, the structure of spatial attention, and the information available in the encoder's internal activations. Full definitions are given in Appendix H. Mental arithmetic (MA) appears in every diagnostic as a common reference, and the remaining datasets rotate across them (FACED, Mumtaz, BCI IV-2a, ISRUC), so that the behaviour is not shown on a single fixed subset.

Reconstruction sufficiency: Masked reconstruction hides some EEG patches and trains the model to predict them from the visible ones (He et al., 2022). To test how much of the encoder's reconstruction ability requires the encoder at all, three restricted predictors are evaluated on the same masked patches: a *distance interpolator*, which weights the visible electrodes at the same time patch by their distance to the masked one; a *linear spatial predictor* from all visible electrodes at that time patch; and a *linear temporal predictor* from the adjacent patches of the masked electrode. The predictors are fitted on the training subjects of each downstream dataset and evaluated on held-out subjects, and the encoder is applied as pretrained, without seeing these datasets. Each is scored by its *recovered fraction*, the share of the encoder's MSE improvement over a zero predictor that is recovered by the predictor: 0 means no better than zero prediction, and 1 matches the encoder (see Appendix H.5).

Spatial-attention structure: The same-time-step channel attention maps (see Appendix H.1) are analysed with two measures, where a sample is one trial at one time patch.

- *Static share* (σ): the fraction of attention energy, above the uniform baseline, contained in a single pattern shared across samples (see Appendix H.2). It is a variant of the attention variability proposed by Vig & Belinkov (2019), which measures how far a head's attention deviates from its mean over inputs. This study formulates it as an energy fraction so that it is comparable across montages and models. $\sigma = 1$ means the departure from uniform attention is the same on every trial; $\sigma \approx 0$ means it changes with the input. It tests whether spatial attention adapts to the signal or acts as a fixed mixing of channels.
- *Distance dependence* (R^2): the fraction of variance in the row-centred mean attention logits explained by a smooth function of inter-electrode distance, compared against a channel-permutation null (see Appendix H.3). Inspired by the mean attention distance of Raghu et al. (2021) which summarises how far a head attends. Here, R^2 measures how much of the attention map can be explained by distance. A high R^2 means the average attention map is set by scalp geometry and mostly encodes electrode proximity.

Representation content: Linear probes (Alain & Bengio, 2016) are fitted to the encoder's internal activations (residual stream) at every layer, to test which EEG properties are linearly decodable (see Appendix I). Classification probes predict task class and subject identity on the same window-level split and are scored by balanced accuracy. Two EEG properties are used as regression targets: the DE of each electrode in five frequency bands, which is its log band power under a Gaussian assumption (Duan et al., 2013), and the covariance between electrodes. Regression probes predict each electrode's DE individually, and the covariance from the tokens of all electrodes together; they are scored by the R^2 of their predictions on held-out data, which, unlike the distance R^2 above, measures how much of the target the activations make linearly available.

3.4 Downstream evaluation protocol

The diagnostics above examine the encoder itself: its attention, its reconstructions and its internal activations. The protocol below measures how much downstream EEG tasks can use the learned representations (hyperparameters in Appendix J).

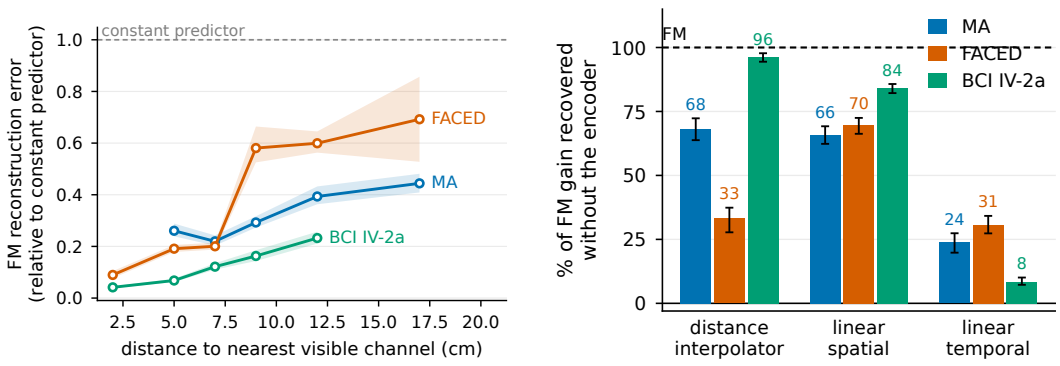
Training modes: Linear probing is the main protocol, as it measures what the frozen representation makes available: for the primary encoder and all released models, the encoder is frozen and only a readout is trained. Fine-tuning, with all layers trained under layer-wise learning-rate decay, measures how much the released models' representations change when adapted to the task. Supervised models trained from scratch provide a reference. Each pretrained model keeps its native token pooling, followed by the same linear readout; the primary encoder averages its tokens, except on ISRUC, where it uses its CLS token. BIOT's convolutional channel adapter is also trained, to map each montage onto its 18 input channels.

Splits: Cross-subject evaluation is the primary protocol, following prior EEG FM evaluations (Jiang et al., 2024; Yang et al., 2023). Each of three seeds draws a different split of subjects into 70% training, 10% validation and 20% test sets, with no subject shared between sets. TUEV keeps its predefined test set, and SEED-VIG uses a $60/20/20$ split because model selection is unstable with a 10% validation set. BCI IV-2a is evaluated per subject: the model is trained on session 1 and tested on session 2 of the same subject, over three training seeds. Results are reported as mean $\pm$ standard deviation over the three seeds, using the primary metric in Table 1; secondary metrics are defined in Appendix K.

4 Diagnosis of masked EEG reconstruction

4.1 A Spatial Shortcut in Masked EEG Reconstruction

Masked reconstruction is largely solvable from spatial context: Masked modelling rewards any context that predicts the missing observation, and in EEG volume conduction makes nearby



(a) (b)

Figure 1: Masked-patch reconstruction in the primary encoder on FACED, MA, and BCI IV-2a. **(a)** Reconstruction error relative to a zero prediction versus distance to the nearest visible electrode at the same time patch, adjusted for the number of visible channels. **(b)** Recovered fraction, defined as the share of the encoder's error reduction over a zero prediction reproduced by (1) a distance interpolator, and (2) a linear map using the simultaneously visible electrodes (linear spatial), and by (3) a linear map using adjacent patches from the masked channel's (linear temporal).

electrodes strongly redundant (Nunez & Harth, 1982). The encoder uses this: after adjusting for the number of visible channels, its reconstruction error grows with the distance to the nearest visible electrode (see Figure 1a). As shown in Figure 1b, the linear spatial predictor reaches a recovered fraction of 66–84%. The distance interpolator recovers 68% on MA and 96% on BCI IV-2a, but only 33% on FACED. This suggests that the gap between the two spatial predictors captures cross-channel structure beyond nearby electrodes. The linear temporal predictor recovers at most 31%. The linear spatial predictor applies the same weights on every trial, so most of the encoder's gain is available from a trial-invariant spatial map, leaving little pressure to learn input-dependent interactions between channels (Varela et al., 2001; Fries, 2005).

Spatial attention converges to a trial-invariant map: On downstream data, 87–90% of the non-uniform spatial-attention energy of the pretrained encoder is shared across trials (see static share σ in Table 2), and inter-electrode distance explains 70–87% of the variance of this shared map (R^2; channel-permutation null ≤ 0.05). Two controls show that this map comes from pretraining. At initialisation, attention is close to uniform, and its small departures from uniform are neither shared across trials nor shaped by distance ($\sigma = 0.01$–0.02, $R^2 = 0.05$–0.15). Trained from initialisation on the task labels instead, the same architecture learns non-uniform attention that changes with the input ($\sigma = 0.04$–0.25, against 0.84–0.88 for the pretrained encoder on the same held-out windows) and is not distance-shaped ($R^2 = 0.05$–0.07).

Applying the same diagnostic to four released EEG FMs (ST-EEGFormer, LaBraM, CBraMod and BIOT) gives static shares of 0.62–0.90 across all four, despite substantial architectural differences (see Table 2). Distance dependence is less uniform. The encoders that predict the masked signal or its discrete codes with a learned per-electrode embedding (Primary encoder, ST-EEGFormer, LaBraM) show strongly distance-shaped maps ($R^2 = 0.63$–0.87, permutation null ≤ 0.05). CBraMod predicts the masked signal too but encodes channel identity through a convolution over channel index, and is less explained by distance ($R^2 = 0.26$–0.64). BIOT also masks its input, dropping channels and tokens, but its target is the embedding of the unmasked signal rather than the signal itself; its map is trial-invariant but not distance-shaped ($R^2 \leq 0.04$, at the permutation null). Distance dependence varies across models, but trial invariance appears under every objective that masks the input.

All five encoders are pretrained to recover missing input, as signal, discrete codes or embeddings, and all five route spatial attention in a trial-invariant way, whereas the same architecture is not trial-invariant at initialisation or after supervised training. The trial-invariant map is therefore produced by the pretraining objective, not by the architecture, the input or a supervised objective.

Table 2: Spatial attention of EEG encoders (deep-half layer means). Brackets: 95th percentile of the channel-permutation null of R^2. *init.*: pretraining initialisation; *supervised*: trained from initialisation on the task labels; both means over three seeds. †Attention over its 16 bipolar channels.

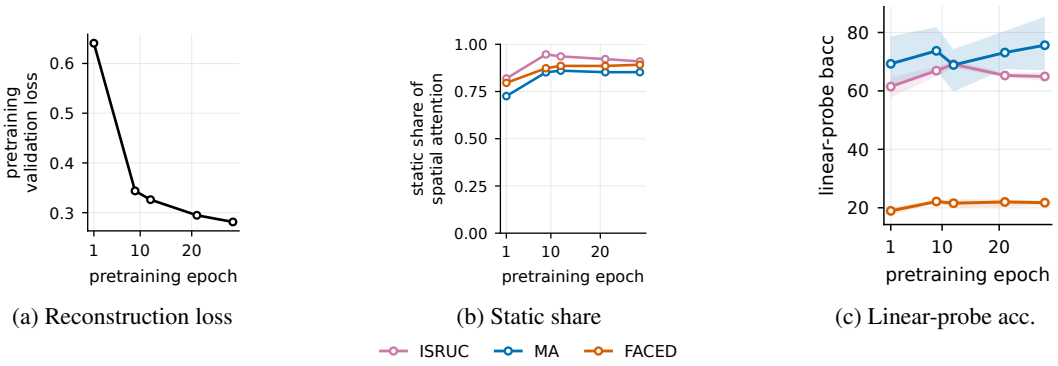
model	objective	channel identity	data	static share	distance R^2
Ours	masked recon.	embedding table	MA	0.87	0.87 [0.05]
			FACED	0.90	0.70 [0.02]
			Mumtaz	0.90	0.87 [0.05]
Ours, init.	none	embedding table	MA	0.02	0.05 [0.03]
			FACED	0.02	0.15 [0.02]
			Mumtaz	0.01	0.06 [0.03]
Ours, supervised	task labels	embedding table	MA	0.13	0.06 [0.04]
			FACED	0.25	0.05 [0.02]
			Mumtaz	0.04	0.07 [0.03]
ST-EEGFormer	masked recon.	embedding table	MA	0.80	0.69 [0.04]
			FACED	0.77	0.63 [0.02]
			Mumtaz	0.77	0.72 [0.05]
LaBraM	masked code pred.	per-electrode table	MA	0.90	0.77 [0.05]
			FACED	0.89	0.74 [0.02]
			Mumtaz	0.89	0.76 [0.05]
CBraMod	masked recon.	conv. over channel index	MA	0.79	0.64 [0.04]
			FACED	0.63	0.39 [0.02]
			Mumtaz	0.62	0.26 [0.04]
BIOT†	contrastive	embedding table	MA	0.90	0.02 [0.03]
			FACED	0.77	0.02 [0.03]
			Mumtaz	0.84	0.04 [0.03]

4.2 Attention Map Invariance Across Training Duration

One might argue that the trial-invariant map described in Section 4.1 reflects insufficient training. To test it, the same encoder is measured at five checkpoints of the run, epochs 1, 9, 12, 21 and 28 of 30, on identical held-out windows (see Figure 2). Looking at Figure 2b, from a near-uniform map at initialisation (see Table 2), spatial attention becomes largely trial-invariant within the first epoch, and reaches its final values by epoch 9 and then stays fixed; the attention maps plotted on the scalp show the same (see Figure 8 in Appendix L). Over the same run, the pretraining validation loss falls from 0.64 to 0.28 and is still decreasing at epoch 28 (see Figure 2a), while transfer does not follow: after epoch 9, linear-probe accuracy shows no gain beyond the split-to-split spread on any of the three datasets (see Figure 2c). Further optimisation of the objective improves reconstruction but not transfer, which points to the pretraining objective rather than to under-training.

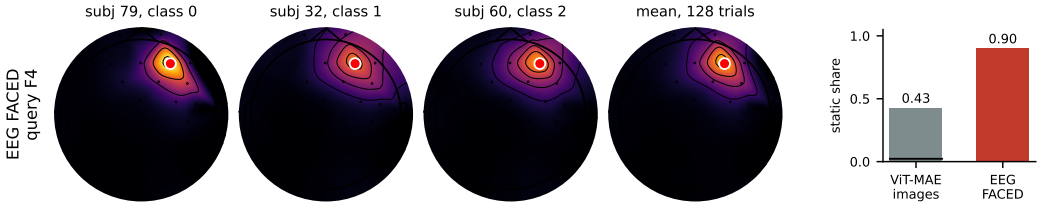
4.3 Masked Reconstruction in EEG and Vision

In vision, the limitation of masked reconstruction concerns *content*: pixel-space reconstruction spends capacity on low-level detail and background texture, which is why the field has moved toward latent-space objectives such as DINO (Caron et al., 2021) and JEPA (LeCun et al., 2022) that predict semantic features rather than pixels (Liu et al., 2023; Wei et al., 2024). Figure 3b compares the static share of ViT-MAE (He et al., 2022) on 256 Imagenette (Howard, 2019) images with that of this study's EEG encoder on FACED, both averaged over the deeper half of the encoder. In ViT-MAE, 43% of the non-uniform attention energy is shared across images (permutation null 2%); in the EEG encoder, 90% is shared across trials, and the attention of a given electrode closely matches the population mean on every trial (see Figure 3a). Targets other than the raw signal do not remove the trial-invariant map: LaBraM (Jiang et al., 2024), which predicts discrete codes, and BIOT (Yang et al., 2023), which matches the embedding of the complete input, are both trial-invariant (see Table 2). Whether a DINO- or JEPA-style predictive latent target behaves differently on EEG remains open. Appendix M shows that the map also persists under a substantially different masking strategy and in a larger model.



(a) Reconstruction loss (b) Static share (c) Linear-probe acc.

Figure 2: Pretraining trajectory of the primary encoder at epochs 1, 9, 12, 21, and 28 of a 30-epoch run. **(a)** Validation reconstruction loss. **(b)** Spatial-attention static share.**(c)** Cross-subject linear-probe balanced accuracy on MA, ISRUC, and FACED; points show the mean and ranges show the minimum and maximum over three subject splits.



(a) (b)

Figure 3: Spatial-attention patterns in EEG and images. **(a)** Attention from electrode F4 in the EEG encoder on three FACED trials from different subjects and classes, alongside the mean over 128 trials. **(b)** Static share averaged over the deeper half of each encoder for ViT-MAE on 256 Imagenette images (permutation null 0.02), and the EEG encoder on FACED (see Table 2).

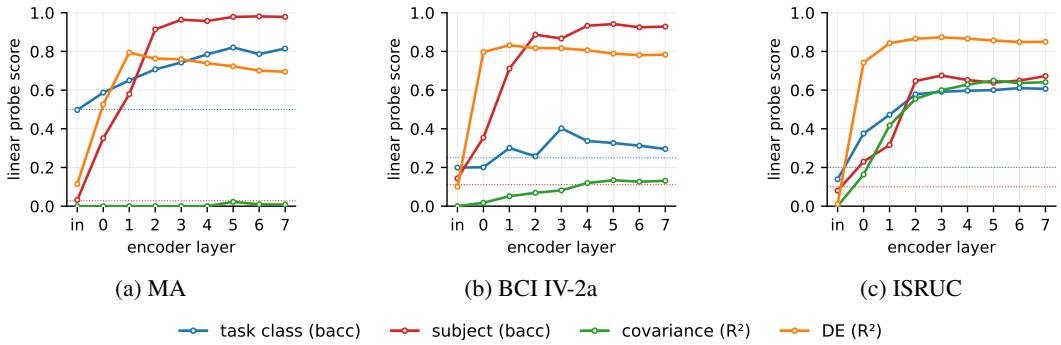
5 Consequences on Learning

5.1 What does the encoder learn?

The preceding analysis shows that the spatial map is nearly trial-invariant, but does not show what the encoder represents. Therefore, linear probes are fitted to the residual stream at every layer (see Figure 4; the method and per-stage results are provided in Appendix I).

Subject identity is more decodable than task content: Task class and subject identity are probed on the same window-level split. As shown in Figure 4, at the last layer, subject identity is almost perfectly decodable on MA (0.98, chance 0.03) and BCI IV-2a (0.93, chance 0.11), and reaches 0.67 on ISRUC (chance 0.10), while task class is less decodable on every dataset (MA 0.81, BCI IV-2a 0.30, ISRUC 0.61).. The representation separates who was recorded more sharply than what they were doing. On BCI IV-2a, whose two sessions per subject were recorded on different days, a probe fitted on session-1 windows identifies the subject of session-2 windows at 0.924 balanced accuracy (chance 0.111). Thus, subject decodability is not an artifact of the session (see Appendix N). Kommineni et al. (2026) report the same relative ordering between subject and task decodability on BCI datasets, using identical train/test splits for both probes.

DE is decodable while covariance varies by datasets: Each electrode's DE is linearly decodable from its own token from the first layer onwards ($R^2 = 0.70$ on MA, 0.78 on BCI IV-2a, 0.85 on ISRUC at the last layer). The covariance between electrodes, represented by the off-diagonal entries of the log-Euclidean tangent of the channel covariance (Arsigny et al., 2007; Barachant et al., 2011), is almost absent on MA ($R^2 = 0.01$) and low on BCI IV-2a (0.13; see Figure 4). This is consistent



(a) MA (b) BCI IV-2a (c) ISRUC

Figure 4: Linear probes performance at the input and after each encoder layer. Task class and subject identity probes (balanced accuracy) use the same window-level split; dotted lines show chance levels. DE (R^2) is probed from each electrode's own token; covariance (R^2) is probed from the concatenated per-electrode tokens, on the off-diagonal entries of the covariance tangent. Probe details are given in Appendix I.

with the trial-invariant map of Section 4.1. With input-dependent weights $A_{cc'}(\mathbf{X})$, the spatial step combines electrodes through the logits $\mathbf{q}_c^\top \mathbf{k}_{c'}$, which are products of the two electrodes' features. With the trial-invariant map $\bar{A}$ these products become constants, and the output for electrode c reduces to a fixed mixture of value vectors, $\mathbf{o}_c = \sum_{c'} \bar{A}_{cc'} \mathbf{v}_{c'}$. This spreads each electrode's features to its neighbours but cannot form the products $x_c x_{c'}$ that covariance requires; such terms can only come from the temporal attention or the MLP layers. ISRUC is the exception ($R^2 = 0.64$), plausibly because its six electrodes give only 15 cross-channel entries, against 231 on BCI IV-2a.

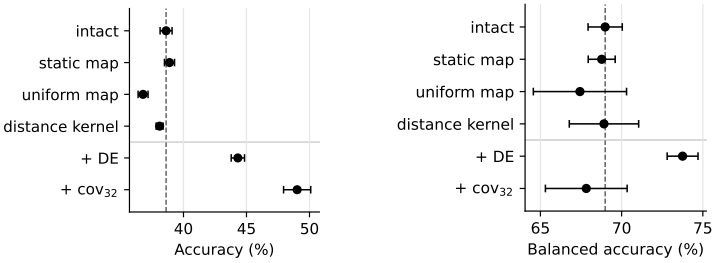
Pooling can obscure electrode-specific information: The DE probes above read each electrode's own token, but on MA and BCI IV-2a the linear head receives the mean over electrodes for the downstream task. After averaging, the DE probe drops on MA from $R^2 = 0.70$ to 0.30 (see Figure 6 in Appendix I): averaging tokens erases which electrode carried which spectrum. On BCI IV-2a, where the spectral modulation is spatially broad (McFarland et al., 2000), the mean retains it ($R^2 = 0.78$; see Figure 7 in Appendix I). On MA, DE is therefore represented but not read out.

5.2 Replacing spatial attention and supplying missing statistics

Section 4.1 found a trial-invariant spatial map, and Section 5.1 found channel covariance was barely available and DE only partly read out. The first is tested by substituting a fixed map, the second by supplying these statistics to the head of the frozen encoder: DE per electrode, and cov_{32}, the log-Euclidean tangent of the channel covariance reduced to 32 dimensions by PCA (see Appendix O). All datasets and the released FMs are reported in Appendices P.1 and P.2.

Spatial substitution: At test time, the post-softmax electrode-to-electrode attention of every layer is replaced by a fixed map: the *static map*, the model's own mean map; the *distance kernel*, a smooth function of inter-electrode distance fitted to the mean logits; or the *uniform map*, equal weight on all electrodes. The integrated CLS token is unchanged, and the layers that follow were trained with the learned attention, so the substitution is if anything penalised (see Appendix O). On both datasets the static map and the distance kernel stay within half a point of the intact encoder, and only the uniform map, which removes routing between electrodes altogether, costs close to two points (see Figures 5a and 5b). The same holds on FACED (below one point) and TUEV (up to 2.5), and on MA and Mumtaz all differences lie within the split-to-split spread (see Appendix P.1). Based on the predictions in Section 4.1, the learned spatial attention can be replaced by a fixed map at small cost.

Supplying missing statistics: On BCI IV-2a, where the probes found covariance almost absent, cov_{32} is the largest gain in Figure 5a (+10.4 points), and it also adds +2.9 on TUEV. On ISRUC, the one dataset whose representation already carries covariance ($R^2 = 0.64$ in Section 5.1), cov_{32} adds nothing (see Figure 5b): covariance helps exactly where the probe says it is missing. DE helps on both (+5.7 and +4.8), although the pooled vector already carries most of it. Both features also



(a) BCI IV-2a (b) ISRUC

Figure 5: Downstream interventions on the frozen encoder. **(a)** BCI IV-2a (motor imagery, per subject). **(b)** ISRUC (sleep staging, cross-subject). Learned electrode-to-electrode attention is replaced at test time with the model's mean map (*static map*), a map fitted to inter-electrode distance (*distance kernel*), or equal weights (*uniform map*). Separate probes receive DE or cov_{32} features. Error bars show the standard deviation across three seeds.

improve the released FMs (see Appendix P.2). Without any encoder, the better of cov_{32} and DE alone comes within two points of, or exceeds, the encoder with that feature supplied to the head on six of the seven datasets (see Appendix Q): the frozen representation adds little beyond fixed statistics of the input. Trained from initialisation on the task labels, the same architecture, whose attention is input-dependent, matches or exceeds this linear probe on FACED and Mumtaz but not on MA (see Table 17), so the trial-invariant map is not needed to reach the frozen encoder's accuracy.

5.3 Fine-tuning reshapes the frozen representation

Fine-tuning ST-EEGFormer, a 26M-parameter masked-reconstruction encoder, per subject on BCI IV-2a raises its accuracy by 16 points (see Table 18). Two quantities of the pooled representation that feeds the linear head describe what changes. The first is the *representation static share*: the fraction of the representation's energy carried by its mean over trials, i.e. by a component that is identical on every trial and therefore cannot separate them (see Appendix H.6). The second is the Fisher ratio of between-class to within-class scatter (Fukunaga, 2013), which is near zero when the class means coincide and grows as they separate (see Appendix H.7). With the encoder frozen, the representation static share is 98.8%, as high as at random initialisation (98.6%), and the Fisher ratio is 0.04 (0.01 at random initialisation): almost all of the representation is the same on every trial, and pretraining does not change this. After fine-tuning, the representation static share falls to 30% and the Fisher ratio rises to 0.61. Across the nine subjects, those whose representation becomes more class-separable after fine-tuning also reach higher accuracy (Spearman $\rho = 0.93$ between Fisher ratio and fine-tuned accuracy; $\rho = -0.74$ for the representation static share, see Table 33). Fine-tuning does not read the pretrained representation better; it reshapes it. The primary encoder agrees: when fine-tuned from its pretrained weights, it is matched by the same architecture trained from initialisation on FACED and Mumtaz, and exceeds it only on MA (see Table 17).

6 Conclusion

In vision, recovering a masked patch requires information from its surroundings; in EEG, a masked electrode is largely predictable from its neighbours, and a fixed spatial map identical on every trial accounts for much of the reconstruction. Pretraining accordingly drives spatial attention within the first epoch to a trial-invariant map shaped by scalp distance, which further pretraining and a larger model do not change. The map appears in every released encoder trained on masked input, whether it predicts the signal, discrete codes or an embedding, but not in the same architecture trained on task labels, whose attention stays input-dependent. Downstream, the learned map can be replaced by a fixed one at small cost, the representation separates subjects more sharply than tasks and makes channel covariance barely available, and supplying covariance and DE features improves transfer. These findings and diagnostics give future EEG foundation models a concrete target: pretraining objectives that cannot be solved predominantly through fixed spatial interpolation.

AI USE STATEMENT

In this work, we used generative AI tools for design or provide feedback on research methodology or experiments. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results; and provide critical ingredients for proving mathematical claims,assist in the writing of proofs, generating synthetic data sets are not applicable to this work.

Additionally, we used generative AI tools for create or edit software code, edit a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the main author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text describes the pretraining setup, evaluation protocol, and interventions used in this study. The appendices provide further details on dataset preprocessing and splits, model and probe configurations, training settings, the construction of fixed attention maps and input statistics, and the analyses of released foundation models. We report variation across seeds or subject splits where applicable. Code is provided in the supplementary materials.

REFERENCES

Swati Aggarwal and Nupur Chugh. Review of machine learning techniques for eeg based brain computer interface. *Archives of Computational Methods in Engineering*, 29(5):3001–3020, 2022.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Diego Alvarez-Estevez and Roselyne Rijsman. Haaglanden Medisch Centrum sleep staging database. *PhysioNet*, March 2022. doi: 10.13026/t79q-fr32. URL `https://doi.org/10.13026/t79q-fr32`. Version 1.1.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2011.

Lies Bollens, Bernd Accou, Marlies Gillis, Wendy Verheijen, Tom Francart, et al. Sparrkulee: A speech-evoked auditory response repository of the ku leuven, containing eeg of 85 participants. *Data*, 9(8):94, 2024. doi: 10.3390/data9080094.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European conference on computer vision*, pp. 108–124. Springer, 2022.

Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. Eeg datasets for motor imagery brain–computer interface. *GigaScience*, 6(7):gix034, 2017.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

Carlijn J de Gans, P Burger, ES Van den Ende, J Hermanides, PWB Nanayakkara, RJBJ Gemke, F Rutters, and DJ Stenvers. Sleep assessment using eeg-based wearables–a systematic review. *Sleep Medicine Reviews*, 76:101951, 2024.

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

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=uazfjnFL0G.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pp. 81–84. IEEE, 2013.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. NeurIPT: Foundation model for neural interfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=D4hcJPkJ3y.

Pascal Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in cognitive sciences*, 9(10):474–480, 2005.

Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Yiyu Gui, Mingzhi Chen, Yuesheng Zhu, Guibo Luo, and Yuchao Yang. Spotr: Spatio-temporal pooling one-token reconstruction for universal physiological signal self-supervised learning. *arXiv preprint arXiv:2606.21973*, 2026.

Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A Smith, and Roy Schwartz. How much does attention actually attend? questioning the importance of attention in pretrained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1403–1416, 2022.

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.

Thi Kieu Khanh Ho and Narges Armanfard. Self-supervised learning for anomalous channel detection in EEG graphs: Application to seizure analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):7866–7874, 2023. doi: 10.1609/aaai.v37i7.25952. URL https://doi.org/10.1609/aaai.v37i7.25952.

Jiahui Hong, Siqing Li, Muqing Jian, and Luming Yang. Bidirectional time-frequency pyramid network for enhanced robust eeg classification. *arXiv preprint arXiv:2510.10004*, 2025.

Jeremy Howard. Imagenette. https://github.com/fastai/imagenette, 2019.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=96f9N8986n.

Murat Kaya, Mustafa Kemal Binli, Erkan Ozbay, Hilmi Yanar, and Yuriy Mishchenko. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Scientific data*, 5(1):180211, 2018.

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*, 124:180–192, 2016.

Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The nmt scalp eeg dataset: An open-source annotated dataset of healthy and pathological eeg recordings for predictive modeling. *Frontiers in neuroscience*, 15:755817, 2022.

Brian Kim, Brian A Erickson, Guadalupe Fernandez-Nunez, Ryan Rich, Georgios Mentzelopoulos, Flavia Vitale, and John D Medaglia. Eeg phase can be predicted with similar accuracy across cognitive states after accounting for power and signal-to-noise ratio. *eneuro*, 10(9), 2023.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929. IEEE, 2019.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Simon Bock Segaard, Jeppe Roden Münster, Andreas Peter Juhl Hansen, Takfarinas Medani, Tiantian Feng, Richard Leahy, and Shrikanth Narayanan. Aperiodic and low-frequency spectral bias in reconstruction based eeg foundation models. *arXiv preprint arXiv:2605.26434*, 2026.

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.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

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

Ang Li, Zikai Wang, Liuyin Yang, Zhenyu Wang, Tianheng Xu, Honglin Hu, and Marc M Van Hulle. Comet: A contrastive-masked brain foundation model for universal eeg representation. *arXiv preprint arXiv:2509.00314*, 2025.

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.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. Improving pixel-based mim by reducing wasted modeling capability. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5338–5349. IEEE, 2023.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Towards decoupled interpretability and multi-scale architecture for EEG foundation model. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10007532.

Dennis J McFarland, Laurie A Miner, Theresa M Vaughan, and Jonathan R Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain topography*, 12(3): 177–186, 2000.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5544–5555, 2024.

Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (eeg)-based computer-aided technique to diagnose major depressive disorder (mdd). *Biomedical Signal Processing and Control*, 31:108–115, 2017.

Paul L Nunez and Erich Harth. Electric fields of the brain: the neurophysics of eeg, 1982.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Robert Oostenveld and Peter Praamstra. The five percent electrode system for high-resolution eeg and erp measurements. *Clinical neurophysiology*, 112(4):713–719, 2001.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG–adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=ZeFMtRBy4Z.

Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B Esper, Dung Truong, Arnaud Delorme, Michael P Milham, and Scott Makeig. Hbn-eeg: The fair implementation of the healthy brain network (hbn) electroencephalography dataset. *bioRxiv*, pp. 2024–10, 2024.

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

James R Stieger, Stephen A Engel, and Bin He. Continuous sensorimotor rhythm based brain computer interface learning in a large population. *Scientific Data*, 8(1):98, 2021.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 57(11):1–43, 2025.

Ling Tang, Qian Chen, Jilin Mei, Houshi Xu, Quanshi Zhang, Jing Shao, Na Zou, Xia Hu, and Dongrui Liu. What do eeg foundation models capture from human brain signals? *arXiv preprint arXiv:2605.11410*, 2026.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Mikhail Trifonov. The structure function as new integral measure of spatial and temporal properties of multichannel eeg. *Brain informatics*, 3(4):211–220, 2016.

Gert Vanhollebeke, Stefanie De Smet, Rudi De Raedt, Chris Baeken, Pieter van Mierlo, and Marie-Anne Vanderhasselt. The neural correlates of psychosocial stress: A systematic review and meta-analysis of spectral analysis eeg studies. *Neurobiology of stress*, 18:100452, 2022.

Francisco Varela, Jean-Philippe Lachaux, Eugenio Rodriguez, and Jacques Martinerie. The brainweb: phase synchronization and large-scale integration. *Nature reviews neuroscience*, 2(4):229–239, 2001.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pp. 63–76, 2019.

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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Yibing Wei, Abhinav Gupta, and Pedro Morgado. Towards latent masked image modeling for self-supervised visual representation learning. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

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

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 (NeurIPS)*, 2025. URL https://openreview.net/forum?id=cjHQj0tCy6.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14475–14485. IEEE, 2023.

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

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

Liuyin Yang, Qiang Sun, Ang Li, and Marc M Van Hulle. Are eeg foundation models worth it? comparative evaluation with traditional decoders in diverse bci tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

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

Jiayan Zhang, Junshi Li, Zhe Huang, Dong Huang, Huaiqiang Yu, and Zhihong Li. Recent progress in wearable brain–computer interface (bci) devices based on electroencephalogram (eeg) for medical applications: A review. *Health data science*, 3:0096, 2023.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. Ctnet: a convolutional transformer network for eeg-based motor imagery classification. *Scientific reports*, 14(1):20237, 2024.

Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of neural engineering*, 14(2):026017, 2017.

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 (NeurIPS)*, 2025. URL https://openreview.net/forum?id=agcXjEHmyW.

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

A NOTATION

Table 3: Summary of notation.

Notation	Meaning
EEG input and model	
C	Number of electrodes in a montage.
c, c'	Electrode indices.
$\mathbf{X} \in \mathbb{R}^{C\times T}$	EEG window of T samples; x_c is the signal of electrode c.
P	Number of samples in a patch (16 at 128 Hz, i.e. 125 ms).
n	Time-patch index.
H	Number of spatial attention heads.
$\mathbf{q}_c$, $\mathbf{k}_{c'}$, $\mathbf{v}_{c'}$	Query, key and value vectors of electrodes c and c' in a spatial attention layer.
$A_{cc'}(\mathbf{X})$	Input-dependent spatial attention weight from electrode c to c'.
$\bar{A}$	Trial-invariant (mean) spatial attention map.
$\mathbf{o}_c$	Output of the spatial attention step for electrode c.
Spatial-attention diagnostics	
s, S	Sample index and number of samples; a sample is one trial at one time patch.
h	Spatial-head index.
i, j	Query- and key-electrode indices.
$L^{(s,h)} \in \mathbb{R}^{C\times C}$	Same-time-step electrode-to-electrode attention logits of sample s and head h.
$A^{(s,h)}$	Attention weights, $\mathrm{softmax}(L^{(s,h)})$, normalised within the block.
U	Uniform map, $U_{ij} = 1/C$.
$Z^{(s)}$	Non-uniform part of the attention map of sample s, $A^{(s)} - U$.
$\bar{Z}$	Mean of $Z^{(s)}$ over samples.
$\lVert\cdot\rVert_F$	Frobenius norm, over heads and both electrode axes.
σ	Static share: fraction of non-uniform attention energy carried by $\bar{Z}$.
$\bar{L}_{ij}$	Logits averaged over heads and samples.
d_{ij}	Euclidean distance between electrodes i and j (metres).
$\varphi(d_{ij})$	Distance basis $\left(d_{ij}, d_{ij}^2, e^{-d_{ij}/0.05}, e^{-d_{ij}/0.10}\right)$.
Φ	Design matrix of φ over the $C(C-1)$ ordered electrode pairs.
$\mathcal{P}$	Row-centring operator (within each query row, self-connections excluded).
β, $\hat{\beta}$, λ	Distance-regression coefficients, their fit, and the ridge penalty ($\lambda = 10^{-6}$).
R^2	Distance dependence: variance of the row-centred mean logits explained by φ.
π	Permutation of electrode labels for the null distribution of R^2.
Reconstruction diagnostics	
x, $\hat{x} \in \mathbb{R}^P$	Masked target patch and its prediction.
E_{zero}	Mean squared error of predicting every masked patch as zero.
E_{FM}	Mean squared error of the pretrained encoder.
m, E_m	Restricted predictor and its mean squared error on the same masked patches.
$\mathrm{recovered}(m)$	Recovered fraction, $(E_{\mathrm{zero}} - E_m)/(E_{\mathrm{zero}} - E_{\mathrm{FM}})$.
δ	Width of the Gaussian distance interpolator.
Representation diagnostics	
$z \in \mathbb{R}^q$	Pooled representation received by the linear head.
μ	Mean of z over trials.

Continued on the next page

Table 3 continued from the previous page

Notation	Meaning
σ_z	Representation static share, $\|\mu\|_2^2/\mathbb{E}\|z\|_2^2$.
$y_s,\ k$	Class label of trial s, and class index.
$\mu_k,\ N_k$	Mean representation and number of trials of class k.
$S_b,\ S_w$	Between- and within-class scatter matrices.
F	Fisher ratio, $\mathrm{tr}(S_b)/\,\mathrm{tr}(S_w)$.
Features and metrics	
DE	Differential entropy per electrode and band, $\frac{1}{2}\log(2\pi e\,\text{band power})$.
cov_{32}	Log-Euclidean tangent of the trace-normalised channel covariance, reduced to 32 dimensions by PCA.
K	Number of classes.
κ	Cohen's kappa.
r	Pearson correlation (SEED-VIG).
ρ	Spearman rank correlation.

B ARCHITECTURE DETAILS

In this section, the architecture of the model used for analysis is described in more depth.

B.1 PREPROCESSING

Recordings from different sources differ in hardware gain, reference and subject physiology. Each window is normalised as a whole, so that the amplitude ratios between electrodes are preserved. Electrodes whose variance exceeds 10 times, or falls below 0.1 times, the median electrode variance of the window are rescaled to that bound, which limits the effect of bad electrodes. The window is then centred by its global mean and divided by 1.4826 times its global median absolute deviation, computed over all electrodes and samples; the median absolute deviation is robust to blinks, muscle artefacts and electrode pops. During pretraining, values are finally clipped to $[-20, 20]$. The primary encoder uses the same normalisation downstream, without the clipping. Filtering and resampling of the pretraining corpus are described in Appendix G.3.

B.2 PATCHING, TOKENIZATION, AND MASKING

Given a batch $\mathbf{X} \in \mathbb{R}^{B\times C\times T}$, the temporal dimension is segmented into $N = \lfloor T/P \rfloor$ non-overlapping patches of size P, producing $\mathbf{X}_{\text{patch}} \in \mathbb{R}^{B\times N\times C\times P}$. Each patch is linearly projected to dimension D: $\mathbf{Z}^{(0)}_{n,c} = \mathbf{W}_{\text{proj}}\,\mathbf{X}_{\text{patch}}[n, c] + \mathbf{b}_{\text{proj}}$. Learnable channel embeddings $\mathbf{e}^{\text{chan}}_c$ and sinusoidal temporal embeddings $\mathbf{e}^{\text{temp}}_n$ are added to each token.

Following BERT-style masked prediction Devlin et al. (2019), each token (n, c) is independently replaced with a learnable `[MASK]` vector with probability ρ: $\tilde{\mathbf{Z}}^{(0)}_{n,c} = (1-m_{n,c})\,\mathbf{Z}^{(0)}_{n,c} + m_{n,c}\,\mathbf{z}_{\texttt{[MASK]}}$, $m_{n,c} \sim \text{Bernoulli}(\rho)$. The reconstruction objective is computed only at masked positions:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{P\,|\mathcal{M}|} \sum_{(n,c)\in\mathcal{M}} \|\hat{\mathbf{X}}_{n,c} - \mathbf{X}_{n,c}\|_2^2, \qquad \mathcal{M} = \{(n, c) : m_{n,c} = 1\}. \tag{1}$$

One learnable CLS token per time patch is appended to the electrode tokens; it is never masked, takes part in spatial and temporal attention like an electrode, and its mean over time patches is the readout on ISRUC.

B.3 PARALLEL SPATIAL-TEMPORAL ATTENTION

Each encoder block uses parallel spatial-temporal attention followed by a feed-forward network. The H attention heads are split in half: $H/2$ *spatial heads* attend across channels at a fixed time step, and $H/2$ *temporal heads* attend across time steps within a fixed channel. This factorisation replaces $\mathcal{O}((NC)^2)$ joint attention with $\mathcal{O}(CN^2 + NC^2)$ Wang et al. (2025); Zhang & Yan (2023). A single linear projection produces $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, which then is split along the head dimension into spatial and temporal halves $(\mathbf{Q}_s, \mathbf{K}_s, \mathbf{V}_s)$ and $(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t)$.

Branch fusion. The spatial and temporal outputs are then concatenated along the head dimension and projected: $\mathbf{Z}^{\text{attn}} = \text{Concat}[\mathbf{O}_s, \mathbf{O}_t]\mathbf{W}_O$, with $\mathbf{O}_s, \mathbf{O}_t \in \mathbb{R}^{B\times L\times D/2}$. A residual connection and feed-forward network complete the block.

C PRETRAINING SETTINGS

Pretraining hyperparameters are summarised in Table 4.

Table 4: Pretraining hyperparameters.

Parameter	Value
Optimizer	AdamW
Learning rate	5×10^{-4} (max)
Weight decay	0.05
LR schedule	OneCycleLR (cosine annealing)
Warmup	15% of total steps
Initial LR	5×10^{-5} (max / 10)
Final LR	5×10^{-8}
Gradient clipping	1.0 (norm)
Epochs	30
Batch size	64–256 (channel-adaptive)
Samples per dataset	50,000 / epoch
Sampling rate	128 Hz
Window length	6 s (48 patches of 125 ms)
Precision	Mixed (BF16)
Masking strategy	Random (per token)
Mask ratio	60%
Loss	MSE (masked patches only)

Batch sizes are adapted to channel count: 64 for $\geq$64 channels, 128 for 32–63 channels, 192 for 16–31 channels, and 256 for fewer than 16 channels, keeping memory usage roughly constant on a 40 GB A100.

D BENCHMARK FOUNDATION MODELS

Four publicly available EEG foundation models with distinct architectures and pretraining objectives are evaluated from their official checkpoints. For each, windows are resampled to the model's native rate and normalised as in its published pipeline before entering the encoder. All models share the same readout: a LayerNorm on the output tokens, the model's native token pooling, and Dropout(0.1) followed by a linear layer. Table 5 summarises the implementation details.

Table 5: Implementation details of the benchmark foundation models. K: number of classes; D: encoder width. The readout has $DK + K$ trainable parameters in linear probing.

Model	Rate (Hz)	Normalisation	Pooling	D	Layers	Pretraining data
LaBraM Jiang et al. (2024)	200	$\times 10^{-2}$ (units of 0.1 mV)	mean over patch tokens	200	12	~2,500 h, 20 datasets
BIOT† Yang et al. (2023)	200	per-channel 95th percentile	mean over tokens	256	4	see text
ST-EEGFormer-s Yang et al. (2026)	128	per-channel z-score	mean over patch tokens	512	8	8M segments, 11 datasets
CBraMod Wang et al. (2025)	200	$\times 10^{-3}$ (mV)	mean over patch tokens	200	12	TUEG

†BIOT's channel convolution (input channels $\rightarrow$ 18) is trained in every mode.

LaBraM Jiang et al. (2024). The Large Brain Model (LaBraM) is a BEiT-v2-style model pretrained on about 2,500 hours of EEG from 20 datasets. Raw EEG is cut into channel-wise patches of 200 samples, which a vector-quantised tokenizer trained on Fourier-domain targets maps to discrete neural codes; a 12-layer Transformer (width 200, 10 heads) is then pretrained to predict the codes of masked patches from their context. Spatial position embeddings are tied to the 10–20 system, so each downstream montage is mapped to the pretrained electrode indices. The `labram-base` checkpoint is used.

BIOT Yang et al. (2023). The Biosignal Transformer (BIOT) tokenises each channel independently into STFT-based patches and concatenates the channels into one sequence, with channel and positional embeddings added to each token; a 4-layer linear-attention encoder (width 256, 8 heads) processes the sequence. A learnable 1-D convolution maps each downstream montage onto the 18 input channels of the pretrained encoder and is trained in every mode. Two released checkpoints are used. The downstream results use `EEG-six-datasets-18-channels`, pretrained contrastively on MGH and SHHS and then supervised on the training sets of TUAB, TUEV, CHB-MIT and IIIC; its TUEV score is therefore not a clean transfer number. The attention diagnostic (Table 2) concerns the pretraining objective and uses the purely contrastive `EEG-SHHS+PREST-18-channels`.

ST-EEGFormer Yang et al. (2026). ST-EEGFormer is a Vision Transformer pretrained with a masked-autoencoder objective on raw EEG. EEG is divided into spatio-temporal patches with sinusoidal temporal and learned per-channel spatial embeddings; 75% of the tokens are masked, the encoder processes the visible ones, and a light decoder reconstructs the masked patches under an MSE loss. Pretraining used 11 public datasets (over 8 million segments) at 128 Hz with per-channel z-standardisation. The small model (ST-EEGFormer-s, 25.8M encoder parameters) is used, closest in size to the other models.

CBraMod Wang et al. (2025). CBraMod is pretrained by patch-based masked reconstruction on the Temple University Hospital EEG corpus. EEG is cut into patches of one second (200 samples), and a 12-layer criss-cross Transformer (width 200, 8 heads) models spatial and temporal dependencies in two parallel branches, half of the heads attending across channels within a time patch and half across time within a channel. Positional information is supplied by an asymmetric conditional positional encoding, a convolution over the channel and time axes, rather than by an electrode table.

Training strategies. Each foundation model is evaluated by *linear probing*, with the encoder frozen and only the readout trained, and by *fine-tuning*, with all layers trained under layer-wise learning-rate decay (Appendix J). Linear probing measures what the pretrained representation makes available; fine-tuning measures how far task-specific adaptation changes it.

E Benchmark Deep Learning Models

Four supervised architectures are trained from scratch as reference points: two convolutional networks and two convolution–Transformer hybrids. Each runs at the dataset's native sampling rate on per-channel z-scored windows, with kernel and pooling sizes scaled to the sampling rate as noted below, and is trained with the shared recipe of Appendix J. Table 6 summarises the architectures.

EEGNet Lawhern et al. (2018). A compact CNN that mirrors the filter-bank common spatial pattern pipeline: a temporal convolution with $F_1 = 8$ filters of length $f_s/2$, a depthwise spatial convolution with $D = 4$ filters per temporal filter spanning all channels, and a separable convolution, each followed by batch normalisation, ELU, temporal downsampling and dropout ($p = 0.5$). The flattened features feed a linear classifier.

DeepConvNet Schirrmeister et al. (2017). A deeper CNN whose first block applies a temporal convolution (25 filters) followed by a spatial convolution across all channels, and whose next three blocks are temporal convolutions with 50, 100 and 200 filters. Each block ends with batch normalisation, ELU, max pooling and dropout ($p = 0.5$). The temporal kernel is 5 samples at 128 Hz and 10 at 250/256 Hz. The flattened features feed a linear classifier.

CTNet Zhao et al. (2024). An EEGNet-style convolutional stem ($F_1 = 20$, $D = 2$, dropout 0.3) turns the window into a sequence of 40-dimensional tokens. Learnable positional encodings are added, and a 6-layer Transformer encoder (4 heads, post-norm) processes the tokens. Its output is added to the stem tokens, flattened, and classified by dropout ($p = 0.5$) and a linear layer.

EEG Conformer Song et al. (2022). A shallow convolutional stem (temporal convolution with 40 filters of length 25, spatial convolution across all channels, batch normalisation, ELU, average

pooling, dropout, and a 1×1 projection to 40 dimensions) produces tokens for a 6-layer pre-norm Transformer encoder with 10 heads. The pooling kernel and stride (75 and 15 for a 1,000-sample window) scale with the window length. The flattened sequence is classified by a three-layer MLP ($256 \to 32 \to K$, ELU, dropout 0.5 and 0.3).

Table 6: Supervised baselines, all trained from scratch at the dataset's native sampling rate. K: number of classes.

Model	Type	Depth	Width / heads	Classifier
EEGNet	CNN (depthwise / separable)	3 conv stages	32 filters / –	Linear
DeepConvNet	CNN	4 conv blocks	25→200 filters / –	Linear
CTNet	CNN + Transformer	2 conv + 6 layers	40 / 4	Dropout + linear
EEG Conformer	CNN + Transformer	1 conv + 6 layers	40 / 10	MLP ($256 \to 32 \to K$)

F Pre-training Datasets

The pretraining corpus comprises 17 EEG datasets, 16 public and one non-public, from more than 3,500 participants. It covers motor imagery and execution, event-related potentials, steady-state visually evoked potentials, auditory and naturalistic stimulation, affective processing, sleep, clinical EEG and sustained attention, with channel counts from 3 to 129. Table 7 summarises the datasets.

Table 7: Pretraining datasets. Native rate is the sampling frequency before preprocessing; all recordings are resampled to 128 Hz. Data scale is given in the natural unit of each dataset.

Dataset	Paradigm	Subjects	EEG ch.	Native rate (Hz)	Data scale
EEG-MI-BCI Cho et al. (2017)	Motor imagery (2-class)	52	64	512	∼10k trials
HGD Schirrmeister et al. (2017)	Motor execution (4-class)	14	128	500	∼13.5k trials
PhysioNet MI Schalk et al. (2004)	Motor imagery / execution	109	64	160	14 runs/subject
BCI IV-2b Tangermann et al. (2012)	Motor imagery (2-class)	9	3	250	5 sessions/subject
Large-MI-Classic Kaya et al. (2018)	Motor imagery (up to 6-class)	13	19	200	∼50k trials
Large-MI-5F Kaya et al. (2018)	Finger motor imagery (5-class)	13	19	200–1,000	∼18k trials
P300 Won et al. (2022)	P300 / RSVP	105	32	512	∼376k trials
SSVEP Liu et al. (2020)	SSVEP (40-class)	70	64	250	∼56k trials
Stieger2021 Stieger et al. (2021)	Online motor-imagery BCI	62	62	1,000	∼600 h
SparrKULee Bollens et al. (2024)	Auditory speech	85	64	1,024	∼170 h
SEED-V Liu et al. (2021)	Emotion	16	62	1,000	48 sessions
HBN (R1–R3) Shirazi et al. (2024)	Resting / visual-cognitive	471	129	500	∼404 h
NMT Khan et al. (2022)	Clinical EEG (normal / abnormal)	2,417	19	200	∼625 h
HMC Alvarez-Estevez & Rijsman (2022)	Clinical sleep / PSG	151	4	256	151 nights
TUEP Obeid & Picone (2016)	Clinical EEG (epilepsy)	200	19	250–1,000	∼2k recordings
TUSZ Obeid & Picone (2016)	Seizure monitoring	675	17–19	≥250	∼1,476 h
Driving (non-public)	Vigilance / driving	9	32	250	continuous

Motor imagery and execution. EEG-MI-BCI Cho et al. (2017) contains left- and right-hand motor imagery from 52 participants. HGD Schirrmeister et al. (2017) adds executed movements (left hand, right hand, feet, rest) in 4 s trials from 14 participants with 128 electrodes. The PhysioNet Motor Movement/Imagery dataset Schalk et al. (2004) contains executed and imagined hand and foot movements from 109 participants. BCI IV-2b Tangermann et al. (2012) uses only three bipolar channels around C3, Cz and C4, the sparsest montage in the corpus. The classical and five-finger paradigms of the large motor-imagery dataset of Kaya et al. (2018) complete this group.

Evoked and frequency-tagged responses. The P300 dataset Won et al. (2022) combines P300-speller and RSVP experiments from 105 participants, with more than 375,000 stimulus presentations. The SSVEP dataset Liu et al. (2020) contains visual target selection over 40 stimulation frequencies from 70 participants.

Continuous, naturalistic and cognitive recordings. Stieger2021 Stieger et al. (2021) contributes more than 600 hours of online BCI cursor control from 62 participants over multiple sessions. SparrKULee Bollens et al. (2024) contains continuous recordings of 85 participants listening to natural

speech. SEED-V Liu et al. (2021) contains EEG from 16 participants watching emotion-eliciting film clips in three sessions. The first three releases of HBN-EEG Shirazi et al. (2024) add about 471 children and adolescents recorded with a 129-channel montage during resting state and passive and active visual-cognitive tasks (movie viewing, surround suppression, contrast-change detection, sequence learning, symbol search).

Clinical EEG and sleep. NMT Khan et al. (2022) contains 2,417 clinical recordings from unique patients labelled normal or abnormal by neurologists (about 625 hours). TUEP Obeid & Picone (2016), a subset of the Temple University Hospital EEG Corpus, contains recordings from 100 participants with epilepsy and 100 without. TUSZ Obeid & Picone (2016) is the seizure subset of the same corpus, with seizure onset, offset, channel and type annotated. HMC Alvarez-Estevez & Rijsman (2022) contains 151 whole-night polysomnographic recordings from a sleep clinic, with four EEG derivations and expert sleep-stage labels.

Sustained attention. A non-public dataset of 9 participants recorded with a 32-channel cap at 250 Hz during monotonous simulated driving adds continuous vigilance data.

F.1 Data Preprocessing

Preprocessing is minimal and identical across sources. Power-line noise is removed at 50 and 60 Hz with an IIR notch filter (`mne`), a 0.1–64 Hz IIR band-pass filter is applied, and signals are resampled to 128 Hz. No artefact rejection or channel interpolation is applied.

F.2 Data Segmentation

Datasets differ in trial duration and total volume, so the hop between windows depends on the dataset.

Default windowing. Most trial-based datasets (EEG-MI-BCI, PhysioNet MI, BCI IV-2b, Large-MI-Classic, P300, SparrKULee, HMC) are cut into 6 s windows with a 0.5 s hop.

Large and continuous sources. Large-MI-5F, HGD, SEED-V, HBN, TUEP, TUSZ, NMT and the driving dataset use 6 s windows with a 2.5 s hop to limit redundancy. Stieger2021, which alone spans about 600 hours, uses a 5 s hop.

SSVEP. SSVEP epochs last 5 s, shorter than the default window, so 2 s windows with a 0.125 s hop are extracted.

Segment dropping. Trailing segments shorter than the window are discarded; no zero-padding is applied.

F.3 Corpus Statistics and Balanced Sampling

The 17 datasets are stored as 20 training parts: the two Large-MI paradigms form one part, HBN is kept as its three releases, TUSZ is split by channel count (19 and 17 channels) and TUEP by reference montage. The corpus totals about 5,200 hours, with individual datasets ranging from about a dozen to more than a thousand hours. A 6 s window at 128 Hz gives $N = 48$ patches of 16 samples (125 ms); SSVEP windows give $N = 16$.

To keep large datasets from dominating the updates, a balanced interleaved sampler draws, at the start of each epoch, at most 50,000 windows uniformly at random from each part (all windows of smaller parts), i.e. at most 10^6 windows per epoch. Each batch comes from a single part, so channel count is constant within a batch, and parts are interleaved round-robin. The draw is repeated every epoch, so larger datasets contribute different windows across epochs.

G Downstream Datasets

G.1 Datasets description

Seven downstream datasets cover motor imagery, emotion recognition, depression detection, vigilance estimation, sleep staging, cognitive load and clinical event classification, with 6 to 32 channels, 2 to 9 classes and one regression task (Table 1).

BCI IV-2a Tangermann et al. (2012). 22-channel EEG at 250 Hz from 9 subjects performing motor imagery of the left hand, right hand, both feet and tongue. Each subject recorded two sessions on different days of 288 trials each (72 per class). A trial consists of a fixation cross (2 s), a cue (1.25 s) and a 4 s imagery period, which is the evaluated window (5,040 trials in total).

FACED Chen et al. (2023). 32-channel EEG from 123 subjects watching 28 video clips that elicit nine emotions: amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness and neutral. Recordings (250 or 1,000 Hz depending on the cohort) are resampled to 256 Hz and cut into three 10 s windows per clip (10,332 windows), giving a 9-class problem.

Mumtaz Mumtaz et al. (2017). 19-channel resting-state EEG at 256 Hz from patients with major depressive disorder (DSM-IV) and healthy controls; 63 subjects are available after export. Recordings are cut into 5 s windows (7,143 windows) for binary classification (depression vs. control).

SEED-VIG Zheng & Lu (2017). 17-channel EEG at 200 Hz recorded during simulated driving: 23 recordings from 21 subjects. Vigilance is given by PERCLOS, the percentage of eye closure measured by eye tracking, every 8 s. One 5 s window is taken per label (885 per recording, 20,355 in total), and the continuous label is regressed.

ISRUC-Sleep, subgroup III Khalighi et al. (2016). Full-night polysomnography from 10 healthy subjects with 6 EEG channels (F3, C3, O1, F4, C4, O2, referenced to the contralateral mastoid) at 200 Hz. Recordings are scored in 30 s epochs into five AASM stages (W, N1, N2, N3, REM; 8,889 epochs). N2 dominates and N1 is rare.

Mental arithmetic Zyma et al. (2019). 19-channel EEG from 36 subjects, cleaned with ICA by the dataset authors, with about 180 s of rest followed by 60 s of serial subtraction. Recordings are resampled to 256 Hz and cut into non-overlapping 5 s windows (1,707 windows); the task is rest vs. arithmetic.

TUEV Obeid & Picone (2016). Clinical EEG from the Temple University Hospital corpus, annotated by neurophysiologists into six classes: spike and sharp wave, generalised periodic epileptiform discharge, periodic lateralised epileptiform discharge, eye movement, artefact and background. 21 channels at 250 Hz from 370 subjects (290 in the official training set, 80 in the evaluation set) are cut into 5 s windows (83,692 windows).

G.2 Evaluation protocol

All datasets except BCI IV-2a and TUEV are split by subject into $70/10/20\%$ training, validation and test subjects; SEED-VIG uses $60/20/20\%$. For BCI IV-2a, each subject is trained on session 1, with 20% of its trials held out for validation, and tested on session 2. TUEV keeps its official training and evaluation sets, with 10% of the training subjects used for validation. Every result is repeated over three subject splits (seeds 42, 4 and 92), and the split seed also seeds training.

G.3 Data Preprocessing

The notch (50 and 60 Hz) and 0.1–64 Hz band-pass filters of the pretraining corpus (Appendix F.1) are applied. Each model then receives the windows at its own sampling rate and normalisation (Appendix D); the supervised baselines receive them at the native rate, z-scored per channel.

H DIAGNOSTIC DETAILS

H.1 WHAT IS MEASURED

All attention diagnostics use the *same-time-step channel block* of the attention map. For a layer, a spatial head h and a sample s (one trial at one time patch), $L^{(s,h)} \in \mathbb{R}^{C\times C}$ are the query–key logits among the C electrodes at that time patch, and $A^{(s,h)} = \mathrm{softmax}(L^{(s,h)})$ the weights, normalised within the block. Statistics are pooled over S samples and reported per layer; tables give the mean over the deeper half of the layers. BIOT uses linear attention without logits; its block is read from the implied row-stochastic map A, with $\log A$ as logits, over its 16 bipolar derivations, whose positions are the midpoints of their two electrodes.

H.2 STATIC SHARE

Every row of $A^{(s)} \in \mathbb{R}^{H\times C\times C}$ sums to one, so all maps contain the uniform component $U_{ij} = 1/C$. It is removed first, $Z^{(s)} = A^{(s)} - U$. With $\bar{Z} = \frac{1}{S}\sum_s Z^{(s)}$ the pattern shared across samples, and since the residual $Z^{(s)} - \bar{Z}$ has zero mean,

$$\frac{1}{S}\sum_s \|Z^{(s)}\|_F^2 = \|\bar{Z}\|_F^2 + \frac{1}{S}\sum_s \|Z^{(s)} - \bar{Z}\|_F^2,$$

with the Frobenius norm taken over heads and both channel axes. The static share is

$$\sigma = \frac{\|\bar{Z}\|_F^2}{\frac{1}{S}\sum_s \|Z^{(s)}\|_F^2} = 1 - \frac{\frac{1}{S}\sum_s \|Z^{(s)} - \bar{Z}\|_F^2}{\frac{1}{S}\sum_s \|Z^{(s)}\|_F^2}.$$

$\sigma = 1$ if every sample produces the same non-uniform map, and $\sigma \to 0$ for independent zero-mean departures from uniform as S grows; $\sigma = 0$ for exactly uniform attention. Unlike the attention variability of Vig & Belinkov (2019), σ is an energy fraction with the uniform floor removed, so it is comparable across channel counts. For ViT-MAE, a chance level is obtained by permuting the patch positions independently per image before averaging (0.02); for the EEG encoder the initialisation rows of Table 2 play this role.

H.3 DISTANCE DEPENDENCE (R^2)

The logits are averaged over spatial heads and then over samples, $\bar{L}_{ij} = \frac{1}{S}\sum_s \frac{1}{H}\sum_h L_{ij}^{(s,h)}$; self-connections are excluded. With d_{ij} the Euclidean distance in metres between electrodes i and j on the `standard_1005` montage (Oostenveld & Praamstra, 2001), distance is represented by the basis $\varphi(d_{ij}) = (d_{ij},\, d_{ij}^2,\, e^{-d_{ij}/0.05},\, e^{-d_{ij}/0.10})$, stacked over the $C(C-1)$ ordered pairs into $\Phi \in \mathbb{R}^{C(C-1)\times 4}$.

Adding a constant to all logits of a query row leaves the softmax unchanged, so both $\bar{L}$ and each column of Φ are row-centred, $(\mathcal{P}\bar{L})_{ij} = \bar{L}_{ij} - \frac{1}{C-1}\sum_{j'\neq i}\bar{L}_{ij'}$. The regression therefore asks whether relative attention within a query row follows relative distance. A ridge fit ($\lambda = 10^{-6}$, for numerical stability only),

$$\hat{\beta} = \arg\min_\beta \ \|\mathcal{P}\bar{L} - (\mathcal{P}\Phi)\beta\|_2^2 + \lambda\|\beta\|_2^2, \qquad R^2 = 1 - \frac{\|\mathcal{P}\bar{L} - (\mathcal{P}\Phi)\hat{\beta}\|_2^2}{\|\mathcal{P}\bar{L}\|_2^2},$$

gives the fraction of variance of the row-centred mean logits explained by a fixed function of distance.

Null and confidence intervals. For each of 200 permutations π of the electrode labels, the distances are replaced by $d_{\pi(i)\pi(j)}$, with $\bar{L}$ unchanged, and the regression is refitted; the 95th percentile of the null R^2 is reported. It is 0.03–0.05 for 19 electrodes and 0.02 for 32, against 0.70–0.87 for the pretrained encoder. Confidence intervals come from 200 bootstrap resamples of the maps.

H.4 PROTOCOL

Static share and distance R^2 describe the frozen encoder and use no labels or split. For each dataset, training and validation data are pooled, 6 subjects are drawn with up to 8 windows each (seed 0), and 64 windows are forwarded in evaluation mode without masking. Each window is cropped to its first 5 s and given to every model in its own downstream input format (sampling rate, normalisation, patching). A window therefore yields 40 maps for the primary encoder and ST-EEGFormer (125 ms patches), 9 for BIOT (STFT tokens with a 0.5 s hop) and 5 for LaBraM and CBraMod (1 s patches). All maps of a layer are pooled. The training trajectory (Figure 2) uses the same estimators on the first 6 s of up to 64 windows, with per-window MAD scaling, on ISRUC, MA and FACED.

H.5 RECONSTRUCTION SUFFICIENCY

For each dataset, 8 windows per subject (at most 256) are taken, cropped to their first 6 s at 128 Hz and divided by 1.4826 times their median absolute deviation. 30% of the subjects are held out. The pretrained encoder reconstructs the windows under random masking of 60% of the tokens, and every predictor is scored on the same masked patches.

Zero baseline. The context-free reference predicts every masked patch $x \in \mathbb{R}^P$ by $\hat{x} = 0$, with error $E_{\text{zero}} = \mathbb{E}\big[\frac{1}{P}\|x\|_2^2\big]$. The signals are band-passed above 0.1 Hz and hence close to zero-mean, so zero uses no information from other electrodes, other time patches or the patch itself.

Recovered fraction. With E_{FM} the error of the encoder and E_m that of a predictor m,

$$\text{recovered}(m) = \frac{E_{\text{zero}} - E_m}{E_{\text{zero}} - E_{\text{FM}}}.$$

It is 0 for a predictor no better than zero and 1 for one that matches the encoder; values above 1 or below 0 are possible.

Predictors. All three are fitted on the training subjects and evaluated on the held-out subjects.

- **Distance interpolator.** The masked patch is the average of the visible electrodes at the same time patch, weighted by $\exp(-(d/\delta)^2)$. A single width δ is chosen from 2–20 cm on the training subjects.
- **Linear spatial predictor.** For each electrode, a ridge regression predicts its masked patch from the visible electrodes at the same time patch.
- **Linear temporal predictor.** A ridge regression shared across electrodes predicts the masked patch from the adjacent patches $n - 1$ and $n + 1$ of the same electrode, when visible.

Ridge penalties are chosen from $\{10^{-1}, \ldots, 10^3\}$ on the last 20% of the training windows. Confidence intervals come from 200 bootstrap resamples of the held-out subjects.

H.6 REPRESENTATION STATIC SHARE

The definition of Appendix H.2 is applied to the pooled representation $z \in \mathbb{R}^q$ consumed by the linear head. With $\mu = \mathbb{E}[z]$ its mean over trials, $\mathbb{E}\|z\|_2^2 = \|\mu\|_2^2 + \mathbb{E}\|z - \mu\|_2^2$, and

$$\sigma_z = \frac{\|\mu\|_2^2}{\mathbb{E}\|z\|_2^2}.$$

There is no uniform part to remove, and z is not centred, since centring would remove the component being measured. $\sigma_z = 1$ when every trial has the same representation. A high σ_z does not require a learned LayerNorm offset: a randomly initialised network, whose LayerNorm bias is zero, already reaches 0.986, and fine-tuning, which trains that bias, lowers it to 0.295 (Table 33).

H.7 FISHER RATIO

Class separation of z is measured by the Fisher ratio (Fukunaga, 2013),

$$F = \frac{\mathrm{tr}(S_b)}{\mathrm{tr}(S_w)}, \qquad S_b = \sum_k N_k(\mu_k - \mu)(\mu_k - \mu)^\top, \qquad S_w = \sum_k \sum_{s:\, y_s = k} (z_s - \mu_k)(z_s - \mu_k)^\top,$$

with μ_k and N_k the mean and number of trials of class k. F is near zero when the class means coincide and grows as they separate relative to the spread within each class.

I RESIDUAL-STREAM PROBES

Setup. For each dataset, up to 16 windows per subject (at most 320) are taken, cropped to their first 6 s at 128 Hz and divided by 1.4826 times their median absolute deviation. At the input and after every encoder block, the pre-norm residual stream is averaged over time patches, giving one vector per electrode, and these are averaged over electrodes to give the pooled vector that the linear head receives. The probes are:

- **class**: logistic regression on the task label, balanced accuracy, on 30% held-out subjects (subject-disjoint), and on 30% held-out windows (window-level);
- **subject**: logistic regression on subject identity, balanced accuracy, on the same window-level split;
- **covariance**: ridge regression to the $C(C+1)/2$ entries of the log-Euclidean tangent of the window covariance, from the pooled vector or from the concatenated per-electrode vectors, on all entries, the diagonal or the off-diagonal;
- **DE**: ridge regression to each electrode's differential entropy in five bands, from that electrode's own vector, from the pooled vector, or from the integrated CLS token after the last block.

Regression probes are scored by R^2 on held-out subjects. The inverse regularisation of the logistic regression ($\{0.01, 0.1, 1\}$) and the ridge penalty (10^0–10^5) are selected on the last 20% of the training rows; nothing is tuned on the test rows. Negative R^2 is shown as 0 in the figures.

At the last layer, values not shown in Figure 4 are, for MA, BCI IV-2a and ISRUC: subject-disjoint class 0.74, 0.32, 0.64; diagonal covariance from the per-electrode vectors 0.56, 0.62, 0.87; covariance from the pooled vector 0.06, 0.32, 0.60.

Per-stage probes. Figures 6 and 7 repeat the probes at every stage of every block: the spatial value projection, the spatial and temporal branch outputs, the attention and MLP writes to the residual stream, and the block output. On MA, per-electrode DE drops from the spatial value projection to the spatial branch output ($0.70 \to 0.62$ at L1, $0.73 \to 0.66$ at L3, $0.63 \to 0.51$ at L7), and pooling over electrodes costs $0.69 \to 0.30$ at the last layer. On BCI IV-2a neither step loses DE ($0.75 \to 0.78$, $0.76 \to 0.78$, $0.69 \to 0.74$; pooling $0.78 \to 0.78$). This is consistent with band power on BCI IV-2a varying jointly over the sensorimotor areas of both hemispheres (Pfurtscheller & Da Silva, 1999), so that averaging neighbours or all electrodes preserves it, whereas on MA it differs more between electrodes.

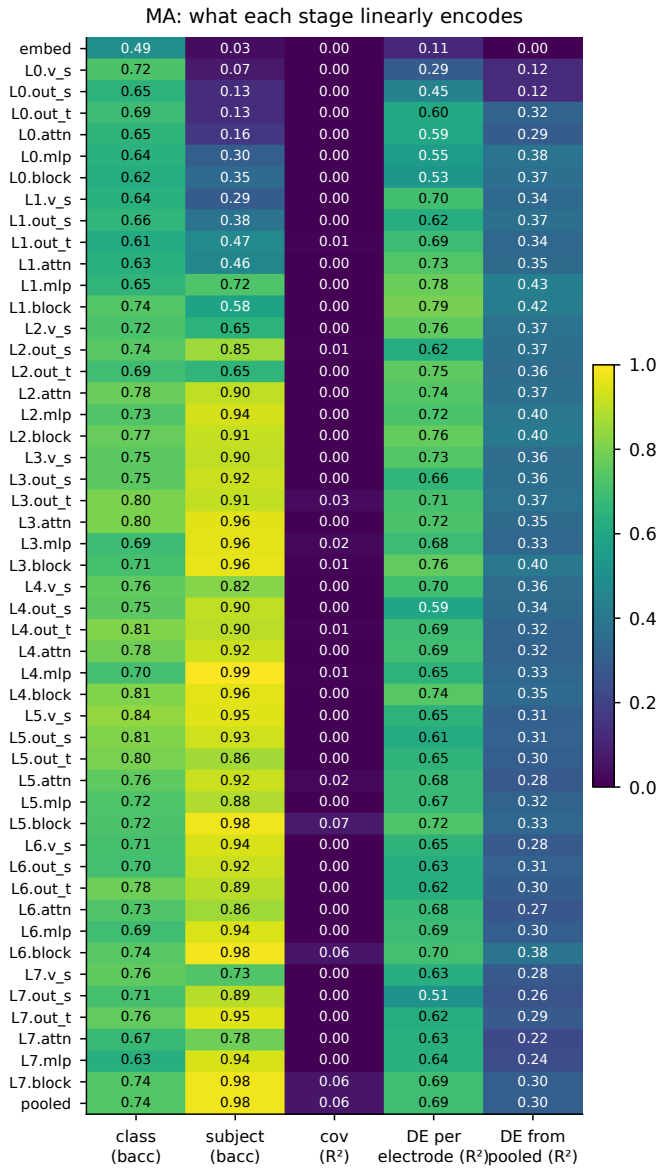


Figure 6: Per-stage probe scores on MA. Rows are ordered by depth: `v_s` spatial value projection, `out_s` and `out_t` branch outputs, `attn` and `mlp` writes to the residual stream, `block` block output. The last two columns compare DE probed per electrode with DE probed from the electrode-averaged vector.

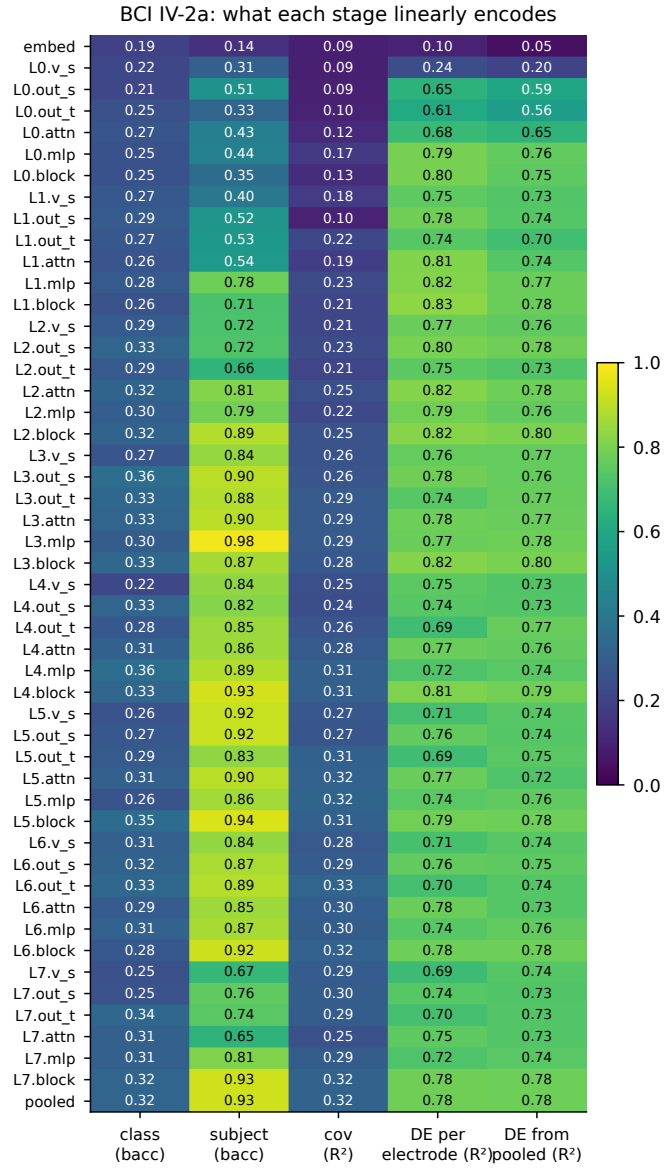


Figure 7: Per-stage probe scores on BCI IV-2a; rows and columns as in Figure 6.

J TRAINING STRATEGY

All models are trained with one configuration per evaluation mode (Table 8), following the ST-EEGFormer benchmark Yang et al. (2026); the recipe is not tuned per model, apart from the overrides listed below.

Linear probing. The pretrained encoder is frozen and only the readout (Appendix D) is trained, together with BIOT's channel adapter.

Fine-tuning. All encoder layers are trained with the readout. The learning rate of block l of an L-block encoder is scaled by 0.75^{L-l}, so earlier blocks move less. The learning rate follows a per-iteration linear warm-up and cosine decay, gradients are clipped at norm 1.0, and mixup ($\alpha = 0.2$, applied to 90% of the batches) is used for classification.

Supervised baselines. EEGNet, DeepConvNet, CTNet and EEG Conformer are trained from scratch with the settings of the last column of Table 8.

Common settings. Classification uses cross-entropy with label smoothing 0.1 and class weights proportional to $1/\sqrt{n_k}$ for class counts n_k, capped at a ratio of 10. For SEED-VIG the loss is $0.5 \cdot \mathrm{MSE} + 0.5 \cdot (1 - r)/2$, with r the Pearson correlation Yang et al. (2026). Early stopping and model selection use the validation primary metric (Table 1) and start after the warm-up epochs.

Overrides. CBraMod is fine-tuned with a learning rate of 10^{-4}. Linear probing uses 20 warm-up epochs on TUEV and a patience of 30 on BCI IV-2a. The supervised baselines use a learning rate of 10^{-3} on TUEV, with inverse-frequency class weights capped at a ratio of 15.

Table 8: Training hyperparameters for each evaluation mode, shared across models within a mode (overrides in the text).

Hyperparameter	**Linear probe**	**Fine-tuning**	**Supervised baselines**
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-3}	5×10^{-4}	3×10^{-3}
Weight decay	0.05	0.05	0.05
Batch size	64	64	64
Epochs	100	100	100
Warm-up epochs	10	10	10
LR schedule	Cosine	Cosine	Cosine
Label smoothing	0.1	0.1	0.1
Early-stopping patience	20	30	10
Layer-wise LR decay	–	0.75	–
Gradient clipping	–	1.0	–

K EVALUATION METRICS

K.1 CLASSIFICATION

Accuracy and balanced accuracy. Accuracy is the primary metric on BCI IV-2a, whose classes are balanced. All other classification datasets use balanced accuracy, the mean per-class recall,

$$\mathrm{BAcc} = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \tag{2}$$

with K classes, so that each class counts equally whatever its prevalence (e.g. background windows in TUEV, N2 epochs in ISRUC). Macro recall, also reported in the full tables, is the same quantity.

Cohen's κ. $\kappa = (p_o - p_e)/(1 - p_e)$, with p_o the accuracy and p_e the agreement expected by chance from the class frequencies; 0 is chance level and 1 perfect agreement. It is the usual metric in sleep staging.

Macro F1, precision and recall. Precision, recall and their harmonic mean F1 are computed per class and averaged with equal weights.

AUROC. The area under the ROC curve is computed one-vs-rest from the predicted probabilities and macro-averaged. It does not depend on a decision threshold.

K.2 Regression

For SEED-VIG, whose target is the continuous PERCLOS index, Pearson's r between predictions and targets is the primary metric, following Zheng & Lu (2017). It is invariant to offset and scale, so it measures whether predictions follow the fluctuations of vigilance. Mean squared error (MSE) and mean absolute error (MAE), in PERCLOS units, measure the size of the errors.

L Additional Figure of attention evolution during training

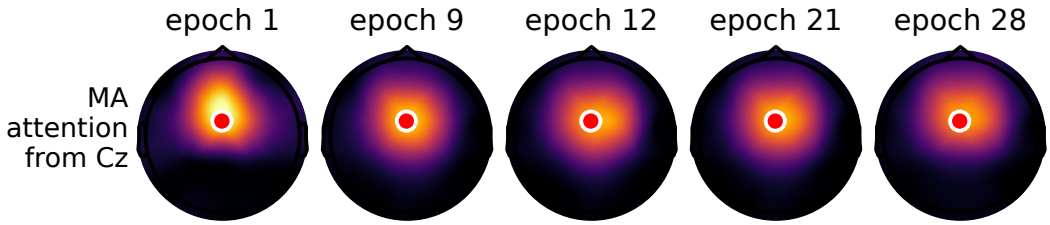


Figure 8: Evolution of the spatial attention map from Cz.

M Robustness to masking strategy and model size

Two variations of the primary encoder test whether the spatial shortcut depends on the masking strategy or on model capacity (Table 9).

Correlation-aware masking. Random masking almost always leaves a masked electrode's neighbours visible, so spatial interpolation is available by construction. An identical model is therefore pretrained with a masking that removes it. For each window, the Pearson correlation between electrodes is computed from the raw signal, 20% of the electrodes are drawn as seeds, and each seed is masked together with its most correlated partner and that partner's most correlated partner; two hops are used because masking only the nearest correlate would leave the second nearest available for the same interpolation. Whole electrodes are masked across all time patches, so a masked electrode has no visible counterpart anywhere in the window. A second round masks a further 15% of the remaining tokens at random to keep temporal reconstruction pressure. Montages with fewer than 8 electrodes fall back to random masking of 50% of the tokens. The map does not become input-dependent: its static share rises slightly, to 0.92 on all three datasets. It remains distance-shaped, with R^2 falling modestly to 0.67–0.75, far above the permutation null. This is consistent with useful context lying further away when an electrode's nearest correlates are masked with it, so that a monotone decay in distance describes the fixed map less well.

Model size. A base model with embedding dimension 512 instead of 256 (8 layers, 25M parameters in the encoder blocks against 6.36M) is pretrained with the same objective and random masking. It reaches a lower validation loss (0.21 against 0.28 at epoch 28) but forms the same map: static share 0.88–0.92 and $R^2 = 0.75$–0.89 at every checkpoint. Between epochs 12 and 28 the static share and R^2 each change by at most 0.02, so the larger model, like the primary encoder (Section 4.2), fixes the map early and keeps it as reconstruction improves.

Table 9: Static share and distance R^2 (deep-half layer means) under correlation-aware masking and for a larger model. Protocol as in Table 2. 95th percentile of the channel-permutation null of R^2: 0.05 (MA), 0.02 (FACED), 0.05 (Mumtaz).

		MA		FACED		Mumtaz	
model	masking	static	R^2	static	R^2	static	R^2
Primary (6.36M)	random	0.87	0.87	0.90	0.70	0.90	0.87
Primary (6.36M)	corr.-aware	0.92	0.73	0.92	0.67	0.92	0.75
Base (25M), epoch 12	random	0.88	0.87	0.89	0.76	0.90	0.87
Base (25M), epoch 21	random	0.89	0.88	0.90	0.75	0.91	0.88
Base (25M), epoch 28	random	0.90	0.88	0.90	0.75	0.92	0.89

N SUBJECT IDENTITY PERSISTS ACROSS RECORDING SESSIONS

The subject probe of Section 5.1 splits at the window level, so windows from the same subject and the same recording session appear on both sides of the split. A high score is therefore compatible with a per-session fingerprint (electrode impedance, cap placement, amplifier state on that day) rather than a property of the person. BCI IV-2a separates the two: each of its nine subjects was recorded on two different days, distributed as separate sessions. The same 9-way logistic probe is fitted on the pooled last-layer representation (256-d), with 144 windows per subject per session (1,296 per session), in three ways: *within-session*, a random $70/30$ split of session-1 windows, which reproduces the protocol of Section 5.1; *cross-session*, fitted on all session-1 windows and evaluated on all session-2 windows; and *shuffled*, the cross-session fit with permuted subject labels. Balanced accuracy is reported over the nine subjects (chance $1/9 = 0.111$).

The within-session probe reaches 0.977 and the cross-session probe 0.924, against 0.103 for permuted labels. A session fingerprint cannot transfer across sessions, so the probe reads the subject, not the recording.

O REPLACING THE SPATIAL KERNEL AT INFERENCE

If spatial attention is a trial-invariant proximity map, replacing it with a fixed map should cost nothing. Both mechanisms below act on the pretrained encoder at downstream time only; pretraining is untouched and no encoder weight changes.

Spatial substitution. At every layer, the post-softmax attention among the C electrodes is replaced by a fixed map $\mathbf{M}^{(\ell)} \in \mathbb{R}^{H\times C\times C}$, one per spatial head, calibrated once on the training split, as in constant-attention probes of language models (Hassid et al., 2022). With $\bar{\mathbf{S}}^{(\ell)}$ the model's spatial logits averaged over training windows and time patches, three maps are used:

- **static**: $\mathbf{M} = \mathrm{softmax}(\bar{\mathbf{S}})$, the model's mean map, with all trial dependence removed;
- **distance kernel**: the row-centred off-diagonal of $\bar{\mathbf{S}}$ is regressed, per head, onto the basis $\{d, d^2, e^{-d/5\,\mathrm{cm}}, e^{-d/10\,\mathrm{cm}}\}$ of inter-electrode distance and replaced by the fit, so only the distance-determined part of the mean map remains;
- **uniform**: $\mathbf{M}_{ij} = 1/C$, which removes routing between electrodes.

Only the electrode-to-electrode block is replaced. Rows and columns involving the integrated-CLS query keep the model's per-trial weights, and each electrode row keeps its own mass on the CLS, with the fixed block renormalised to the remainder. The readout pathway is therefore unchanged.

Classical side features. For each window $\mathbf{X} \in \mathbb{R}^{C\times T}$, the covariance $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}^\top/T$ is divided by its mean diagonal, $10^{-4}\mathbf{I}$ is added, and the log-Euclidean tangent vector $\mathrm{vec}_\triangle(\log \mathbf{\Sigma}) \in \mathbb{R}^{C(C+1)/2}$ is reduced by a PCA fitted on the training split to $k \in \{16, 32\}$ dimensions. Differential entropy is computed per channel from the power spectrum in five bands (1–4, 4–8, 8–13, 13–30, 30–50 Hz), giving $C \times 5$ values. Each feature is layer-normalised and concatenated to the pooled representation before the linear classifier; a d-dimensional feature adds $d(K+2)$ trainable parameters for K classes and nothing to the encoder.

Side features are evaluated under linear probing only. Fine-tuning is designed to change the encoder, with layer-wise learning-rate decay and a readout that receives a small share of the update (Bao et al., 2021; He et al., 2022; Kumar et al., 2022), so a gain from a feature concatenated to that readout would measure the recipe rather than the representation. Under linear probing the readout is the only trained component, and the question of what the pretrained representation makes available is well posed (Alain & Bengio, 2016).

P ADDITIONAL RESULTS

P.1 VANILLA CRISS CROSS ADDITIONAL RESULTS

All the results with the different spatial kernels and added side features on the vanilla criss cross transformer under linear probe are shown here.

Table 10: Primary encoder on BCI IV-2a, per subject (trained on session 1, tested on session 2; 4 classes), linear probe. Mean over the nine subjects; mean ± s.d. over seeds 42/92/4.

Configuration	Accuracy	F1	AUC ROC	Kappa	Precision
Ours	38.62 ± 0.39 %	0.341 ± 0.006	0.663 ± 0.003	0.182 ± 0.005	0.383 ± 0.006
+ static map	38.89 ± 0.33 %	0.344 ± 0.001	0.666 ± 0.001	0.185 ± 0.004	0.382 ± 0.010
+ uniform map	36.79 ± 0.32 %	0.319 ± 0.008	0.672 ± 0.002	0.157 ± 0.004	0.380 ± 0.009
+ distance kernel	38.11 ± 0.22 %	0.343 ± 0.002	0.666 ± 0.002	0.175 ± 0.003	0.377 ± 0.006
+ DE	44.32 ± 0.43 %	0.423 ± 0.006	0.708 ± 0.001	0.258 ± 0.006	0.462 ± 0.008
+ cov_{16}	45.66 ± 1.02 %	0.411 ± 0.011	0.743 ± 0.002	0.275 ± 0.014	0.466 ± 0.013
+ cov_{32}	49.02 ± 0.88 %	0.450 ± 0.008	$\mathbf{0.759} \pm 0.003$	0.320 ± 0.012	0.515 ± 0.025
+ cov_{16} + DE	45.55 ± 1.42 %	0.425 ± 0.020	0.742 ± 0.012	0.274 ± 0.019	0.488 ± 0.010
+ cov_{32} + DE	$\mathbf{49.95} \pm 0.42$ %	$\mathbf{0.473} \pm 0.007$	0.756 ± 0.003	$\mathbf{0.333} \pm 0.006$	$\mathbf{0.524} \pm 0.007$

Table 11: Primary encoder on ISRUC, cross-subject (sleep staging, 5 classes). Mean ± s.d. over subject-split seeds 42/92/4.

Configuration	Balanced accuracy	F1	AUC ROC	Kappa	Precision
Linear probe					
Ours	68.99 ± 0.86 %	0.689 ± 0.012	0.922 ± 0.002	0.632 ± 0.006	0.697 ± 0.009
+ static map	68.77 ± 0.68 %	0.694 ± 0.006	0.919 ± 0.003	0.634 ± 0.012	0.704 ± 0.002
+ uniform map	67.43 ± 2.34 %	0.679 ± 0.028	0.925 ± 0.003	0.620 ± 0.026	0.703 ± 0.028
+ distance kernel	68.91 ± 1.74 %	0.693 ± 0.020	0.922 ± 0.005	0.635 ± 0.015	0.706 ± 0.019
+ DE	$\mathbf{73.75} \pm 0.78$ %	$\mathbf{0.734} \pm 0.002$	$\mathbf{0.941} \pm 0.003$	$\mathbf{0.686} \pm 0.011$	$\mathbf{0.743} \pm 0.004$
+ cov_{16}	67.29 ± 2.14 %	0.674 ± 0.031	0.922 ± 0.003	0.614 ± 0.018	0.686 ± 0.036
+ cov_{32}	67.82 ± 2.05 %	0.674 ± 0.027	0.919 ± 0.006	0.614 ± 0.017	0.690 ± 0.019
+ cov_{16} + DE	72.90 ± 1.65 %	0.725 ± 0.020	0.940 ± 0.002	0.675 ± 0.017	0.730 ± 0.020
+ cov_{32} + DE	72.33 ± 0.78 %	0.719 ± 0.016	0.937 ± 0.002	0.664 ± 0.008	0.726 ± 0.014
Fine-tuning					
Ours	73.02 ± 1.01 %	0.732 ± 0.021	0.924 ± 0.010	0.689 ± 0.016	0.756 ± 0.027

Table 12: Primary encoder on MA, cross-subject (rest vs. arithmetic, 2 classes). Mean ± s.d. over subject-split seeds 42/92/4.

Configuration	Balanced accuracy	F1	AUC ROC	Kappa	Precision
Linear probe					
Ours	70.28 ± 7.64 %	0.701 ± 0.086	0.815 ± 0.068	0.406 ± 0.174	0.719 ± 0.121
+ static map	73.87 ± 9.31 %	0.727 ± 0.100	0.825 ± 0.096	0.459 ± 0.195	0.726 ± 0.099
+ uniform map	64.38 ± 6.42 %	0.637 ± 0.059	0.760 ± 0.068	0.288 ± 0.129	0.678 ± 0.123
+ distance kernel	71.67 ± 5.84 %	0.694 ± 0.035	0.808 ± 0.069	0.394 ± 0.074	0.689 ± 0.029
+ DE	$\mathbf{77.35} \pm 5.72$ %	0.747 ± 0.066	0.851 ± 0.067	0.500 ± 0.128	0.737 ± 0.064
+ cov_{16}	73.26 ± 3.50 %	0.710 ± 0.065	0.799 ± 0.074	0.431 ± 0.114	0.715 ± 0.067
+ cov_{32}	73.19 ± 5.38 %	0.717 ± 0.042	0.848 ± 0.070	0.439 ± 0.085	0.720 ± 0.044
+ cov_{16} + DE	76.77 ± 2.38 %	$\mathbf{0.759} \pm 0.027$	0.830 ± 0.024	$\mathbf{0.524} \pm 0.047$	$\mathbf{0.775} \pm 0.054$
+ cov_{32} + DE	74.38 ± 2.16 %	0.710 ± 0.048	$\mathbf{0.865} \pm 0.066$	0.434 ± 0.084	0.716 ± 0.053
Fine-tuning					
Ours	74.54 ± 9.12 %	0.696 ± 0.086	0.816 ± 0.087	0.410 ± 0.165	0.694 ± 0.077

Table 13: Primary encoder on FACED, cross-subject (emotion, 9 classes; chance 11.1%). Mean ± s.d. over subject-split seeds 42/92/4.

Configuration	Balanced accuracy	F1	AUC ROC	Kappa	Precision
Linear probe					
Ours	21.77 ± 0.86 %	0.206 ± 0.011	0.646 ± 0.004	0.122 ± 0.008	0.220 ± 0.009
+ static map	21.54 ± 0.47 %	0.205 ± 0.010	0.636 ± 0.008	0.118 ± 0.005	0.214 ± 0.002
+ uniform map	20.94 ± 0.56 %	0.196 ± 0.002	0.635 ± 0.005	0.110 ± 0.006	0.210 ± 0.003
+ distance kernel	21.35 ± 1.19 %	0.201 ± 0.017	0.636 ± 0.007	0.117 ± 0.013	0.221 ± 0.003
+ DE	$\mathbf{23.29} \pm 0.22$ %	0.220 ± 0.005	$\mathbf{0.664} \pm 0.001$	$\mathbf{0.139} \pm 0.003$	0.236 ± 0.015
+ cov_{16}	21.37 ± 1.06 %	0.204 ± 0.008	0.644 ± 0.006	0.118 ± 0.013	0.223 ± 0.012
+ cov_{32}	20.50 ± 0.71 %	0.196 ± 0.014	0.628 ± 0.007	0.107 ± 0.009	0.225 ± 0.007
+ cov_{16} + DE	23.22 ± 0.68 %	$\mathbf{0.224} \pm 0.006$	$\mathbf{0.664} \pm 0.005$	0.137 ± 0.008	$\mathbf{0.237} \pm 0.011$
+ cov_{32} + DE	23.01 ± 0.58 %	0.220 ± 0.005	0.660 ± 0.004	0.136 ± 0.008	0.234 ± 0.009
Fine-tuning					
Ours	48.29 ± 1.25 %	0.486 ± 0.011	0.824 ± 0.010	0.415 ± 0.014	0.499 ± 0.006

Table 14: Primary encoder on Mumtaz, cross-subject (depression vs. control, 2 classes). Mean ± s.d. over subject-split seeds 42/92/4.

Configuration	Balanced accuracy	F1	AUC ROC	Kappa	Precision
Linear probe					
Ours	85.84 ± 10.64 %	0.863 ± 0.096	0.954 ± 0.033	0.728 ± 0.190	0.876 ± 0.077
+ static map	88.03 ± 8.46 %	0.885 ± 0.075	0.961 ± 0.039	0.770 ± 0.150	0.895 ± 0.063
+ uniform map	$\mathbf{91.37} \pm 4.36$ %	$\mathbf{0.918} \pm 0.034$	$\mathbf{0.986} \pm 0.010$	$\mathbf{0.837} \pm 0.067$	$\mathbf{0.930} \pm 0.020$
+ distance kernel	85.40 ± 11.06 %	0.858 ± 0.102	0.952 ± 0.046	0.717 ± 0.203	0.868 ± 0.090
+ DE	85.22 ± 7.42 %	0.858 ± 0.060	0.952 ± 0.018	0.718 ± 0.118	0.877 ± 0.034
+ cov_{16}	85.91 ± 6.90 %	0.859 ± 0.066	0.936 ± 0.039	0.719 ± 0.132	0.861 ± 0.064
+ cov_{32}	86.27 ± 5.36 %	0.861 ± 0.051	0.936 ± 0.028	0.723 ± 0.103	0.864 ± 0.053
+ cov_{16} + DE	86.40 ± 4.62 %	0.869 ± 0.035	0.949 ± 0.013	0.738 ± 0.070	0.881 ± 0.019
+ cov_{32} + DE	86.11 ± 5.31 %	0.866 ± 0.042	0.947 ± 0.010	0.733 ± 0.084	0.879 ± 0.025
Fine-tuning					
Ours	88.90 ± 7.96 %	0.898 ± 0.066	0.964 ± 0.022	0.798 ± 0.129	0.926 ± 0.031

Table 15: Primary encoder on TUEV (6 event types; official training and evaluation sets). Mean ± s.d. over seeds 42/92/4. Background dominates the class distribution, so balanced accuracy is the operative metric.

Configuration	Balanced accuracy	F1	AUC ROC	Kappa	Precision
Linear probe					
Ours	34.36 ± 0.31 %	0.235 ± 0.020	0.894 ± 0.043	0.375 ± 0.013	0.250 ± 0.019
+ static map	33.03 ± 1.66 %	0.233 ± 0.015	0.888 ± 0.033	0.326 ± 0.010	0.228 ± 0.019
+ uniform map	31.83 ± 3.17 %	0.227 ± 0.007	0.835 ± 0.066	0.341 ± 0.004	0.223 ± 0.019
+ distance kernel	32.52 ± 2.25 %	0.224 ± 0.008	0.867 ± 0.065	0.355 ± 0.001	0.232 ± 0.019
+ DE	38.45 ± 0.73 %	0.299 ± 0.007	0.869 ± 0.069	0.419 ± 0.006	0.281 ± 0.002
+ cov_{16}	36.01 ± 0.69 %	0.285 ± 0.010	0.873 ± 0.076	0.397 ± 0.003	0.283 ± 0.009
+ cov_{32}	37.24 ± 0.84 %	0.285 ± 0.006	$\mathbf{0.913} \pm 0.010$	0.400 ± 0.004	0.283 ± 0.012
+ cov_{16} + DE	38.25 ± 0.90 %	0.304 ± 0.031	0.880 ± 0.034	0.420 ± 0.026	$\mathbf{0.309} \pm 0.009$
+ cov_{32} + DE	$\mathbf{39.13} \pm 1.51$ %	$\mathbf{0.305} \pm 0.014$	0.862 ± 0.068	$\mathbf{0.421} \pm 0.012$	0.290 ± 0.007
Fine-tuning (seed 42 only)					
Ours	33.59 %	0.260	0.850	0.344	0.238

Table 16: Primary encoder on SEED-VIG, cross-subject (vigilance regression, continuous PERCLOS target; 21 subjects, 23 recordings). Mean ± s.d. over subject-split seeds 42/92/4. One partition (seed 92) yields $r \approx 0.25$ for every configuration and dominates the spread.

Configuration	Pearson r	MSE	MAE
Linear probe			
Ours	0.490 ± 0.224	0.0685 ± 0.0170	0.208 ± 0.027
+ static map	0.492 ± 0.176	0.0593 ± 0.0072	0.195 ± 0.015
+ uniform map	0.451 ± 0.148	0.0706 ± 0.0135	0.210 ± 0.025
+ distance kernel	0.503 ± 0.170	$\mathbf{0.0585} \pm 0.0090$	$\mathbf{0.194} \pm 0.019$
+ DE	0.479 ± 0.185	0.0842 ± 0.0244	0.238 ± 0.032
+ cov_{16}	0.508 ± 0.260	0.0734 ± 0.0187	0.219 ± 0.020
+ cov_{32}	0.490 ± 0.220	0.0795 ± 0.0133	0.229 ± 0.006
+ cov_{16} + DE	$\mathbf{0.519} \pm 0.223$	0.0824 ± 0.0360	0.232 ± 0.048
+ cov_{32} + DE	0.489 ± 0.208	0.0813 ± 0.0263	0.236 ± 0.039
Fine-tuning			
Ours	0.552 ± 0.148	0.0567 ± 0.0103	0.189 ± 0.027

Table 17: Primary architecture trained from initialisation on the task labels, against the pretrained encoder under linear probing and fine-tuning, on the same subject splits (mean ± s.d. over seeds 42/92/4). Static share: spatial attention of the model trained from initialisation and of the pretrained encoder, measured on the same held-out windows (Table 2).

Dataset	Training	Balanced accuracy	F1	AUC ROC	Kappa	Precision	Static share
MA	From initialisation	62.50 ± 2.55 %	0.612 ± 0.007	0.678 ± 0.019	0.234 ± 0.015	0.618 ± 0.006	0.13
	Linear probe	70.28 ± 7.64 %	0.701 ± 0.086	0.815 ± 0.068	0.406 ± 0.174	0.719 ± 0.121	0.87
	Fine-tuning	74.54 ± 11.17 %	0.696 ± 0.105	0.816 ± 0.107	0.410 ± 0.202	0.694 ± 0.094	–
FACED	From initialisation	49.43 ± 1.19 %	0.494 ± 0.015	0.833 ± 0.007	0.427 ± 0.015	0.497 ± 0.016	0.25
	Linear probe	21.77 ± 0.86 %	0.206 ± 0.011	0.646 ± 0.004	0.122 ± 0.008	0.220 ± 0.009	0.88
	Fine-tuning	48.29 ± 1.53 %	0.486 ± 0.014	0.824 ± 0.012	0.415 ± 0.017	0.499 ± 0.007	–
Mumtaz	From initialisation	85.27 ± 7.66 %	0.858 ± 0.061	0.949 ± 0.026	0.720 ± 0.119	0.883 ± 0.033	0.04
	Linear probe	85.84 ± 10.64 %	0.863 ± 0.096	0.954 ± 0.033	0.728 ± 0.190	0.876 ± 0.077	0.84
	Fine-tuning	88.90 ± 9.75 %	0.898 ± 0.081	0.964 ± 0.027	0.798 ± 0.157	0.926 ± 0.038	–

P.2 Released and supervised baselines

Tables 18–31 give the full results of the supervised baselines and of the released foundation models under linear probing and fine-tuning, and the change from adding the side features to each foundation model under linear probing. All foundation models use the shared readout of Appendix D in both modes. Fine-tuned CBraMod falls below its own linear probe on most datasets. The released CBraMod pipeline fine-tunes with a larger head on the flattened patch representations; the smaller shared readout is the most likely cause, as no other was found, and the rows are reported as obtained.

Table 18: BCI IV-2a, per subject (trained on session 1, tested on session 2), averaged over the nine subjects. Mean ± s.d. over seeds 42/92/4; chance 25%.

Model	Accuracy	F1	AUC ROC	Kappa	Recall	Precision
Supervised baselines (trained from scratch)						
EEGNet	50.77 ± 0.54 %	0.505 ± 0.006	0.755 ± 0.005	0.344 ± 0.007	0.508 ± 0.005	0.519 ± 0.004
DeepConvNet	45.37 ± 0.07 %	0.443 ± 0.006	0.725 ± 0.003	0.272 ± 0.001	0.454 ± 0.001	0.473 ± 0.013
CTNet	53.57 ± 1.67 %	0.529 ± 0.021	0.782 ± 0.007	0.381 ± 0.022	0.536 ± 0.017	0.552 ± 0.014
EEG Conformer	43.66 ± 1.26 %	0.430 ± 0.015	0.691 ± 0.015	0.249 ± 0.017	0.437 ± 0.013	0.452 ± 0.017
Foundation models, linear probe						
LaBraM	28.40 ± 0.12 %	0.266 ± 0.002	0.533 ± 0.003	0.045 ± 0.002	0.284 ± 0.001	0.286 ± 0.005
BIOT	37.41 ± 0.61 %	0.350 ± 0.010	0.641 ± 0.004	0.165 ± 0.008	0.374 ± 0.006	0.383 ± 0.012
CBraMod	30.95 ± 0.42 %	0.271 ± 0.008	0.577 ± 0.003	0.079 ± 0.006	0.310 ± 0.004	0.315 ± 0.008
ST-EEGFormer	41.36 ± 0.77 %	0.368 ± 0.012	0.707 ± 0.003	0.218 ± 0.010	0.414 ± 0.008	0.404 ± 0.011
Foundation models, fine-tuning						
LaBraM	29.70 ± 0.13 %	0.286 ± 0.006	0.551 ± 0.000	0.063 ± 0.002	0.297 ± 0.001	0.304 ± 0.004
BIOT	40.33 ± 0.58 %	0.383 ± 0.007	0.657 ± 0.006	0.204 ± 0.008	0.403 ± 0.006	0.428 ± 0.008
CBraMod	25.59 ± 0.23 %	0.137 ± 0.019	0.529 ± 0.008	0.008 ± 0.003	0.256 ± 0.002	0.119 ± 0.040
ST-EEGFormer	56.85 ± 0.47 %	0.554 ± 0.007	0.793 ± 0.009	0.425 ± 0.006	0.569 ± 0.005	0.576 ± 0.012

Table 19: Effect of the side features on BCI IV-2a under linear probing: intact accuracy, and the change from adding DE or cov_{32} (paired over seeds 42/92/4, mean ± s.d.).

Model	intact	+DE	$+\text{cov}_{32}$
LaBraM	28.40 ± 0.12 %	$+10.84 \pm 0.20$ %	$+17.62 \pm 0.99$ %
BIOT	37.41 ± 0.61 %	$+4.26 \pm 1.18$ %	$+9.85 \pm 0.36$ %
CBraMod	30.95 ± 0.42 %	$+10.25 \pm 0.45$ %	$+16.13 \pm 0.82$ %
ST-EEGFormer	41.36 ± 0.77 %	$+2.80 \pm 0.99$ %	$+8.09 \pm 0.42$ %

Table 20: FACED, cross-subject: subjects split $70/10/20$ into training, validation and test sets. Mean $\pm$ s.d. over three subject splits (seeds 42/92/4); chance 11.1%.

Model	Balanced accuracy	F1	AUC ROC	Kappa	Recall	Precision
Supervised baselines (trained from scratch)						
EEGNet	49.96 ± 2.53 %	0.495 ± 0.030	0.849 ± 0.010	0.433 ± 0.030	0.499 ± 0.026	0.499 ± 0.029
DeepConvNet	62.27 ± 0.73 %	0.625 ± 0.004	0.911 ± 0.004	0.572 ± 0.006	0.623 ± 0.007	0.637 ± 0.001
CTNet	56.14 ± 0.80 %	0.560 ± 0.008	0.871 ± 0.016	0.502 ± 0.008	0.561 ± 0.008	0.566 ± 0.008
EEG Conformer	68.74 ± 1.12 %	0.684 ± 0.014	0.931 ± 0.001	0.642 ± 0.015	0.687 ± 0.011	0.684 ± 0.014
Foundation models, linear probe						
LaBraM	17.52 ± 0.76 %	0.157 ± 0.013	0.608 ± 0.001	0.072 ± 0.009	0.175 ± 0.008	0.181 ± 0.013
BIOT	18.94 ± 1.08 %	0.165 ± 0.011	0.611 ± 0.009	0.090 ± 0.013	0.189 ± 0.011	0.178 ± 0.021
CBraMod	17.62 ± 0.79 %	0.160 ± 0.009	0.605 ± 0.001	0.073 ± 0.009	0.176 ± 0.008	0.171 ± 0.009
ST-EEGFormer	22.33 ± 2.30 %	0.212 ± 0.019	0.651 ± 0.010	0.126 ± 0.028	0.223 ± 0.023	0.221 ± 0.022
Foundation models, fine-tuning						
LaBraM	45.72 ± 1.59 %	0.459 ± 0.019	0.801 ± 0.006	0.387 ± 0.017	0.457 ± 0.016	0.470 ± 0.018
BIOT	28.20 ± 1.57 %	0.279 ± 0.015	0.651 ± 0.029	0.183 ± 0.033	0.282 ± 0.016	0.286 ± 0.021
CBraMod	13.98 ± 0.60 %	0.088 ± 0.019	0.547 ± 0.019	0.033 ± 0.005	0.140 ± 0.006	0.130 ± 0.055
ST-EEGFormer	37.63 ± 3.01 %	0.377 ± 0.034	0.741 ± 0.020	0.295 ± 0.035	0.376 ± 0.030	0.387 ± 0.041

Table 21: Effect of the side features on FACED under linear probing: intact balanced accuracy, and the change from adding DE or cov_{32} (paired over seeds 42/92/4, mean $\pm$ s.d.).

Model	intact	+DE	$+\text{cov}_{32}$
LaBraM	17.52 ± 0.76 %	$+4.83 \pm 0.31$ %	-0.14 ± 0.95 %
BIOT	18.94 ± 1.08 %	$+3.65 \pm 2.73$ %	$+0.08 \pm 0.46$ %
CBraMod	17.62 ± 0.79 %	$+5.59 \pm 1.65$ %	$+0.79 \pm 0.49$ %
ST-EEGFormer	22.33 ± 2.30 %	$+3.02 \pm 2.05$ %	-0.13 ± 1.30 %

Table 22: TUEV: official training and evaluation sets, 10% of the training subjects for validation. Mean $\pm$ s.d. over seeds 42/92/4.

Model	Balanced accuracy	F1	AUC ROC	Kappa	Recall	Precision
Supervised baselines (trained from scratch)						
EEGNet	35.24 ± 1.25 %	0.264 ± 0.018	0.729 ± 0.019	0.311 ± 0.025	0.352 ± 0.013	0.249 ± 0.012
DeepConvNet	36.39 ± 1.26 %	0.303 ± 0.002	0.743 ± 0.003	0.370 ± 0.015	0.364 ± 0.013	0.283 ± 0.005
CTNet	32.60 %	0.228	0.718	0.150	0.326	0.287
EEG Conformer	39.54 ± 3.96 %	0.279 ± 0.057	0.719 ± 0.039	0.274 ± 0.043	0.395 ± 0.040	0.260 ± 0.039
Foundation models, linear probe						
LaBraM	35.73 ± 1.22 %	0.272 ± 0.010	0.853 ± 0.072	0.400 ± 0.002	0.357 ± 0.012	0.283 ± 0.009
BIOT	36.92 ± 1.06 %	0.280 ± 0.013	0.816 ± 0.012	0.352 ± 0.032	0.369 ± 0.011	0.306 ± 0.050
CBraMod	37.79 ± 1.36 %	0.277 ± 0.022	0.851 ± 0.026	0.399 ± 0.014	0.378 ± 0.014	0.278 ± 0.022
ST-EEGFormer	40.92 ± 1.38 %	0.302 ± 0.012	0.803 ± 0.042	0.386 ± 0.018	0.409 ± 0.014	0.272 ± 0.010
Foundation models, fine-tuning						
LaBraM	31.68 ± 1.75 %	0.312 ± 0.011	0.678 ± 0.009	0.329 ± 0.048	0.317 ± 0.018	0.345 ± 0.011
BIOT	39.02 ± 2.93 %	0.308 ± 0.005	0.646 ± 0.045	0.374 ± 0.021	0.390 ± 0.029	0.304 ± 0.003
CBraMod	35.81 ± 2.18 %	0.274 ± 0.022	0.602 ± 0.030	0.324 ± 0.032	0.358 ± 0.022	0.280 ± 0.033
ST-EEGFormer	40.92 ± 1.38 %	0.302 ± 0.012	0.803 ± 0.042	0.386 ± 0.018	0.409 ± 0.014	0.272 ± 0.010

Table 23: Effect of the side features on TUEV under linear probing: intact balanced accuracy, and the change from adding DE or cov_{32} (paired over seeds 42/92/4, mean $\pm$ s.d.).

Model	intact	+DE	$+\text{cov}_{32}$
LaBraM	35.73 ± 1.22 %	$+1.45 \pm 1.51$ %	$+1.67 \pm 1.45$ %
BIOT	36.92 ± 1.06 %	$+0.27 \pm 0.81$ %	-0.92 ± 1.25 %
CBraMod	37.79 ± 1.36 %	$+0.32 \pm 1.85$ %	$+0.46 \pm 1.00$ %
ST-EEGFormer	40.92 ± 1.38 %	-0.49 ± 0.83 %	-0.03 ± 0.23 %

Table 24: ISRUC, cross-subject: subjects split $70/10/20$ into training, validation and test sets. Mean $\pm$ s.d. over three subject splits (seeds 42/92/4).

Model	Balanced accuracy	F1	AUC ROC	Kappa	Recall	Precision
Supervised baselines (trained from scratch)						
EEGNet	66.77 ± 1.67 %	0.660 ± 0.024	0.923 ± 0.006	0.603 ± 0.023	0.668 ± 0.017	0.670 ± 0.028
DeepConvNet	67.64 ± 1.81 %	0.667 ± 0.030	0.931 ± 0.007	0.615 ± 0.038	0.676 ± 0.018	0.689 ± 0.023
CTNet	66.32 ± 3.25 %	0.664 ± 0.027	0.919 ± 0.009	0.628 ± 0.020	0.663 ± 0.033	0.677 ± 0.023
EEG Conformer	72.58 ± 3.60 %	0.702 ± 0.033	0.943 ± 0.006	0.669 ± 0.014	0.726 ± 0.036	0.729 ± 0.017
Foundation models, linear probe						
LaBraM	63.61 ± 7.83 %	0.624 ± 0.073	0.916 ± 0.021	0.578 ± 0.077	0.636 ± 0.078	0.657 ± 0.052
BIOT	74.14 ± 1.63 %	0.728 ± 0.019	0.936 ± 0.002	0.678 ± 0.013	0.741 ± 0.016	0.730 ± 0.011
CBraMod	60.56 ± 5.23 %	0.583 ± 0.040	0.884 ± 0.004	0.541 ± 0.040	0.606 ± 0.052	0.625 ± 0.028
ST-EEGFormer	67.37 ± 2.68 %	0.659 ± 0.035	0.919 ± 0.003	0.609 ± 0.011	0.674 ± 0.027	0.686 ± 0.018
Foundation models, fine-tuning						
LaBraM	74.73 ± 1.34 %	0.748 ± 0.012	0.953 ± 0.007	0.706 ± 0.021	0.747 ± 0.013	0.771 ± 0.004
BIOT	74.58 ± 2.24 %	0.727 ± 0.027	0.935 ± 0.017	0.697 ± 0.025	0.746 ± 0.022	0.730 ± 0.020
CBraMod	34.74 ± 4.49 %	0.243 ± 0.077	0.697 ± 0.018	0.134 ± 0.049	0.347 ± 0.045	0.310 ± 0.084
ST-EEGFormer	71.73 ± 0.11 %	0.715 ± 0.005	0.912 ± 0.010	0.675 ± 0.006	0.717 ± 0.001	0.721 ± 0.007

Table 25: Effect of the side features on ISRUC under linear probing: intact balanced accuracy, and the change from adding DE or cov_{32} (paired over seeds 42/92/4, mean $\pm$ s.d.).

Model	intact	+DE	$+\text{cov}_{32}$
LaBraM	63.61 ± 7.83 %	$+5.29 \pm 5.07$ %	$+3.25 \pm 3.54$ %
BIOT	74.14 ± 1.63 %	-0.47 ± 1.07 %	-1.83 ± 4.40 %
CBraMod	60.56 ± 5.23 %	$+6.50 \pm 4.38$ %	-0.77 ± 2.81 %
ST-EEGFormer	67.37 ± 2.68 %	$+4.99 \pm 1.14$ %	$+1.74 \pm 1.17$ %

Table 26: Mumtaz, cross-subject: subjects split $70/10/20$ into training, validation and test sets. Mean $\pm$ s.d. over three subject splits (seeds 42/92/4).

Model	Balanced accuracy	F1	AUC ROC	Kappa	Recall	Precision
Supervised baselines (trained from scratch)						
EEGNet	87.96 ± 4.71 %	0.873 ± 0.053	0.917 ± 0.061	0.751 ± 0.103	0.880 ± 0.047	0.895 ± 0.037
DeepConvNet	89.27 ± 2.29 %	0.886 ± 0.024	0.979 ± 0.005	0.776 ± 0.047	0.893 ± 0.023	0.901 ± 0.018
CTNet	88.64 ± 3.58 %	0.882 ± 0.037	0.951 ± 0.019	0.765 ± 0.073	0.886 ± 0.036	0.887 ± 0.034
EEG Conformer	90.73 ± 3.30 %	0.912 ± 0.036	0.971 ± 0.009	0.823 ± 0.071	0.907 ± 0.034	0.913 ± 0.037
Foundation models, linear probe						
LaBraM	81.42 ± 7.63 %	0.818 ± 0.068	0.925 ± 0.040	0.638 ± 0.134	0.814 ± 0.076	0.831 ± 0.055
BIOT	88.66 ± 6.23 %	0.888 ± 0.056	0.961 ± 0.033	0.777 ± 0.112	0.887 ± 0.062	0.894 ± 0.051
CBraMod	87.17 ± 5.96 %	0.872 ± 0.055	0.927 ± 0.045	0.745 ± 0.110	0.872 ± 0.060	0.882 ± 0.057
ST-EEGFormer	88.30 ± 7.44 %	0.883 ± 0.072	0.960 ± 0.042	0.766 ± 0.143	0.883 ± 0.074	0.887 ± 0.072
Foundation models, fine-tuning						
LaBraM	90.06 ± 4.22 %	0.895 ± 0.045	0.963 ± 0.032	0.792 ± 0.088	0.901 ± 0.042	0.906 ± 0.036
BIOT	89.40 ± 2.44 %	0.889 ± 0.025	0.949 ± 0.020	0.780 ± 0.050	0.894 ± 0.024	0.899 ± 0.022
CBraMod	92.76 ± 0.34 %	0.926 ± 0.003	0.932 ± 0.011	0.852 ± 0.006	0.928 ± 0.003	0.926 ± 0.003
ST-EEGFormer	89.21 ± 1.34 %	0.888 ± 0.012	0.935 ± 0.020	0.777 ± 0.025	0.892 ± 0.013	0.894 ± 0.017

Table 27: Effect of the side features on Mumtaz under linear probing: intact balanced accuracy, and the change from adding DE or cov_{32} (paired over seeds 42/92/4, mean $\pm$ s.d.).

Model	intact	+DE	$+\text{cov}_{32}$
LaBraM	81.42 ± 7.63 %	-0.74 ± 6.95 %	-0.33 ± 5.00 %
BIOT	88.66 ± 6.23 %	$+1.13 \pm 0.63$ %	-0.48 ± 2.32 %
CBraMod	87.17 ± 5.96 %	-2.83 ± 3.57 %	-0.98 ± 3.08 %
ST-EEGFormer	88.30 ± 7.44 %	$+0.17 \pm 2.75$ %	-3.64 ± 5.40 %

Table 28: MA, cross-subject: subjects split $70/10/20$ into training, validation and test sets. Mean $\pm$ s.d. over three subject splits (seeds 42/92/4).

Model	Balanced accuracy	F1	AUC ROC	Kappa	Recall	Precision
Supervised baselines (trained from scratch)						
EEGNet	65.83 ± 3.87 %	0.633 ± 0.019	0.715 ± 0.028	0.276 ± 0.048	0.658 ± 0.039	0.631 ± 0.022
DeepConvNet	52.00 ± 3.46 %	0.242 ± 0.071	0.728 ± 0.060	0.021 ± 0.037	0.520 ± 0.035	0.297 ± 0.296
CTNet	74.64 ± 2.92 %	0.707 ± 0.040	0.824 ± 0.034	0.427 ± 0.069	0.746 ± 0.029	0.698
EEG Conformer	71.22 ± 3.18 %	0.649 ± 0.008	0.802 ± 0.013	0.332 ± 0.028	0.712 ± 0.032	0.703 ± 0.033
Foundation models, linear probe						
LaBraM	62.08 ± 6.74 %	0.627 ± 0.076	0.759 ± 0.082	0.262 ± 0.154	0.621 ± 0.067	0.660 ± 0.115
BIOT	61.82 ± 8.60 %	0.608 ± 0.076	0.739 ± 0.045	0.219 ± 0.154	0.618 ± 0.086	0.604 ± 0.071
CBraMod	64.55 ± 5.00 %	0.633 ± 0.040	0.753 ± 0.058	0.274 ± 0.084	0.646 ± 0.050	0.638 ± 0.047
ST-EEGFormer	69.68 ± 3.69 %	0.684 ± 0.022	0.823 ± 0.032	0.383 ± 0.053	0.697 ± 0.037	0.730 ± 0.107
Foundation models, fine-tuning						
LaBraM	77.89 ± 4.19 %	0.729 ± 0.042	0.863 ± 0.023	0.471 ± 0.080	0.779 ± 0.042	0.722 ± 0.037
BIOT	67.21 ± 1.28 %	0.653 ± 0.023	0.729 ± 0.029	0.312 ± 0.038	0.672 ± 0.013	0.648 ± 0.019
CBraMod	56.31 ± 5.65 %	0.375 ± 0.168	0.697 ± 0.065	0.082 ± 0.077	0.563 ± 0.057	0.439 ± 0.272
ST-EEGFormer	76.15 ± 1.87 %	0.754 ± 0.016	0.870 ± 0.032	0.509 ± 0.032	0.761 ± 0.019	0.749 ± 0.014

Table 29: Effect of the side features on MA under linear probing: intact balanced accuracy, and the change from adding DE or cov_{32} (paired over seeds 42/92/4, mean $\pm$ s.d.).

Model	intact	+DE	$+\text{cov}_{32}$
LaBraM	62.08 ± 6.74 %	$+7.74 \pm 4.52$ %	$+11.01 \pm 12.27$ %
BIOT	61.82 ± 8.60 %	$+1.59 \pm 8.36$ %	$+8.73 \pm 12.96$ %
CBraMod	64.55 ± 5.00 %	$+5.62 \pm 2.12$ %	$+10.24 \pm 10.84$ %
ST-EEGFormer	69.68 ± 3.69 %	$+4.90 \pm 5.13$ %	$+1.75 \pm 5.80$ %

Table 30: SEED-VIG, cross-subject: subjects split $60/20/20$ into training, validation and test sets. Mean $\pm$ s.d. over three subject splits (seeds 42/92/4).

Model	Pearson r	MSE	MAE
Supervised baselines (trained from scratch)			
EEGNet	0.581 ± 0.006	0.0904 ± 0.0232	0.251 ± 0.035
DeepConvNet	0.616 ± 0.027	0.0874 ± 0.0196	0.252 ± 0.031
CTNet	0.602 ± 0.015	0.0831 ± 0.0244	0.236 ± 0.032
EEG Conformer	0.650 ± 0.032	0.0680 ± 0.0139	0.213 ± 0.028
Foundation models, linear probe			
LaBraM	0.450 ± 0.088	0.0708 ± 0.0111	0.216 ± 0.020
BIOT	0.490 ± 0.142	0.0656 ± 0.0095	0.200 ± 0.024
CBraMod	0.337 ± 0.119	0.0759 ± 0.0238	0.224 ± 0.043
ST-EEGFormer	0.527 ± 0.164	0.0574 ± 0.0130	0.192 ± 0.037
Foundation models, fine-tuning			
LaBraM	0.720 ± 0.006	0.0820 ± 0.0658	0.120 ± 0.068
BIOT	0.613 ± 0.040	0.0570 ± 0.0036	0.188 ± 0.005
CBraMod	0.396 ± 0.111	0.0843 ± 0.0169	0.226 ± 0.033
ST-EEGFormer	0.736 ± 0.005	0.0414 ± 0.0010	0.160 ± 0.001

Table 31: Effect of the side features on SEED-VIG under linear probing: intact Pearson r, and the change from adding DE or cov_{32} (paired over seeds 42/92/4, mean $\pm$ s.d.).

Model	intact	+DE	$+\text{cov}_{32}$
LaBraM	0.450 ± 0.088	-0.059 ± 0.071	$+0.095 \pm 0.068$
BIOT	0.490 ± 0.142	-0.016 ± 0.035	$+0.059 \pm 0.024$
CBraMod	0.337 ± 0.119	$+0.025 \pm 0.109$	$+0.138 \pm 0.132$
ST-EEGFormer	0.527 ± 0.164	-0.021 ± 0.004	$+0.015 \pm 0.044$

Q COVARIANCE AND DIFFERENTIAL ENTROPY WITHOUT AN ENCODER

Table 32 reports the side features alone, without an encoder: cov_{32} (the log-Euclidean tangent of the trace-normalised covariance, reduced by a PCA fitted on the training split to 32 dimensions), DE (differential entropy per channel in five bands), and both concatenated. Each is followed by a logistic regression (ridge regression for SEED-VIG) whose regularisation is chosen on the validation split. The features use the same windows (global MAD, 128 Hz) and splits as the linear probe of the primary encoder, shown for comparison.

Table 32: Side features without an encoder against the linear probe of the primary encoder. Balanced accuracy (%; accuracy for BCI IV-2a, whose classes are balanced), Pearson r for SEED-VIG; mean $\pm$ s.d. over seeds 42/92/4. BCI IV-2a uses the fixed session split, so the feature baselines have no spread across seeds.

Dataset	**cov_{32}**	**DE**	**cov_{32} + DE**	**Ours, linear probe**
BCI IV-2a	47.99 ± 0.00	47.34 ± 0.00	46.80 ± 0.00	38.62 ± 0.47
FACED	15.21 ± 0.31	21.45 ± 0.63	21.96 ± 0.16	21.77 ± 0.86
Mumtaz	73.93 ± 8.45	85.04 ± 2.44	86.38 ± 4.18	85.84 ± 10.64
MA	78.88 ± 8.85	62.79 ± 3.08	75.71 ± 5.61	70.28 ± 7.64
ISRUC	42.52 ± 6.38	71.99 ± 1.96	71.61 ± 2.90	68.99 ± 1.05
TUEV	33.73 ± 2.78	33.49 ± 0.92	38.90 ± 0.74	34.36 ± 0.31
SEED-VIG	0.480 ± 0.189	0.310 ± 0.184	0.445 ± 0.240	0.490 ± 0.224

R LINEAR PROBING VERSUS FINE-TUNING PER SUBJECT (ST-EEGFORMER, BCI IV-2A)

Table 33: Pooled-representation geometry of ST-EEGFormer on BCI IV-2a, per subject. *Static share*: fraction of representation energy in the trial-mean vector (Appendix H.6). *Fisher*: between- to within-class scatter (trace ratio; Appendix H.7). Random: same architecture, re-initialised.

	Accuracy		Repr. static share			Fisher		
P	probe	FT	pretrained	FT	random	pretrained	FT	random
1	0.556	0.677	0.991	0.182	0.988	0.070	0.725	0.019
2	0.281	0.385	0.979	0.494	0.980	0.012	0.446	0.015
3	0.510	0.795	0.982	0.225	0.986	0.031	1.573	0.012
4	0.253	0.288	0.992	0.686	0.988	0.029	0.040	0.025
5	0.319	0.476	0.991	0.212	0.985	0.010	0.177	0.011
6	0.368	0.476	0.988	0.292	0.986	0.015	0.183	0.009
7	0.403	0.670	0.986	0.194	0.987	0.025	0.787	0.008
8	0.420	0.653	0.987	0.199	0.984	0.023	0.660	0.008
9	0.542	0.712	0.993	0.168	0.988	0.134	0.892	0.020
mean	0.406	0.570	0.988	0.295	0.986	0.039	0.609	0.014