WHAT MASKING GEOMETRY WORKS BEST FOR EEG FOUNDATION MODELS?
A CONTROLLED EVALUATION ACROSS MAE AND JEPA

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models hold promise for scalable brain-signal decoding across clinical and cognitive neuroscience applications, yet their pre-training pipelines remain poorly understood. Among design choices, the masking strategy is particularly critical: it determines what the network must predict and from which context. Yet it has never been ablated in isolation, as each new model bundles a new masking strategy with a new backbone and objective. In this paper, we formalize the design choices for spatio-temporal masking strategies and train various models with a single pipeline under varying masking configurations across two SSL frameworks (MAE and JEPA). We then systematically evaluate the resulting 58 pre-trained models on the 12 datasets of OpenEEGBench under a linear probe. Both frameworks agree on an optimal masking configuration and on shared failure modes. Outside these, performance is robust: 11 MAE and 9 JEPA configurations are statistically indistinguishable from the best. We further identify a novel JEPA-specific failure mode, tagged *bias-inflation collapse*, invisible to standard detectors. With a well-chosen mask, our pipeline reaches REVE-level downstream performance at a fraction of REVE's pre-training compute.

1 INTRODUCTION

Self-supervised foundation models for electroencephalography (EEG) have multiplied rapidly in the past few years. Models such as REVE (Ouahidi et al., 2025b), LaBraM (Jiang et al., 2024), BIOT (Yang et al., 2023), BrainBERT (Wang et al., 2023), and Neuro-GPT (Cui et al., 2024) are reporting state-of-the-art performance on batteries of downstream decoding tasks, ranging from clinical applications such as abnormality detection (Obeid & Picone, 2016) or sleep staging (Khalighi et al., 2016) to cognitive neuroscience such as motor imagery (Brunner et al., 2008) or emotion recognition (Liu et al., 2022). The large majority of these models share a common design principle: a masked-prediction pretext task, in which a portion of the input is hidden and the network is trained to recover it from the visible context. To define this task, a critical ingredient is the choice of the masking strategy, which defines what is the context and what is the target in this predictive task. The literature uses a remarkably diverse set of strategies: random patch masking, whole-channel masking, contiguous time-window masking, and various spatio-temporal block schemes. Yet every new model bundles together multiple design changes at once, rarely ablating each choice in isolation: a different backbone, a different pre-training corpus, a different pretext objective, and—crucially—a different masking strategy. This makes it impossible to attribute downstream gains to any single component.

In adjacent self-supervised fields for other modalities, the masking strategy has long been recognised as a crucial design choice: MAE (He et al., 2022) ablates the masking ratio, I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024) ablate multi-block masking shapes and counts, and ST-MAE (Feichtenhofer et al., 2022) establishes a spatio-temporal masking precedent for video. As no comparable study exists for EEG, the community lacks principled guidance on how to apply masking.

This motivates the central question of this paper: *What masking geometry works best for EEG foundation models?* and *What's its impact on their downstream performance?* A robust answer may have a double impact: First, it provides an actionable default for practitioners regarding an important design choice that is currently made by inheritance or guesswork. Second, it sheds light on the deeper

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.

question of what kind of supervision a mask provides for EEG data, and how that supervision varies with the mask's spatial and temporal structure.

We approach this question with a controlled grid sweep. We first introduce a formal framework that unifies the masking strategies of the EEG self-supervised literature under a single formalism with three parameters: a spatial radius r, a temporal length L, and a target masking ratio ρ. Tuning r and L continuously interpolates between random patch masking, whole-channel masking, contiguous time-window masking, and spatio-temporal block masking. We sweep a parameter grid of 29 (L, r) configurations covering 6 lengths and 5 radii (excluding the degenerate full-mask cell). Following the results of REVE (Ouahidi et al., 2025b), we keep the mask ratio fixed at $\rho = 0.55$. To assess whether the conclusions generalise beyond a single pretext, we replicate the entire sweep for two SSL frameworks: input-space reconstruction à la MAE (He et al., 2022) and latent-space prediction à la JEPA (Assran et al., 2023). This yields 58 pre-trained foundation models, all sharing the REVE pipeline except for the two axes under study. Each model is evaluated on OpenEEGBench (OEB) (Guetschel et al., 2026a) using a linear probe on frozen features.

Our findings illuminate four points. First, MAE and JEPA agree on which mask geometries help and which do not: a shared optimum at moderate spatial radius and short temporal length, and three shared failure modes (single-channel blocks, full-channel blocks, and most long temporal blocks). Second, within the working regime, downstream performance is largely insensitive to the precise mask choice — many configurations are statistically indistinguishable from the best. Third, JEPA admits a multichannel-specific failure mode that we name *bias-inflation collapse* and which standard collapse detectors miss. Fourth, our pipeline is compute-efficient: at the optimum, both frameworks reach REVE-level downstream performance for a fraction of the original pre-training cost.

We contribute (i) a formal three-parameter framework that unifies the masking strategies of the EEG self-supervised literature; (ii) 58 open-source pre-trained checkpoints spanning the full (L, r) grid under both MAE and JEPA pretexts; (iii) an actionable default mask for practitioners pre-training EEG foundation models; and (iv) the characterisation of *bias-inflation collapse*, a JEPA-specific failure mode on multichannel signals that we do not find documented in the prior literature. In the tradition of design-axis ablations such as Tian et al. (2020)'s "What makes for good views for contrastive learning?", we isolate one component of the pipeline rather than introducing a new model. The code to reproduce the pre-training and evaluation pipelines, as well as the pre-trained checkpoints, is available at `https://anonymous.4open.science/r/eeg-fm-masking-6944/`.

2 Related work

EEG foundation models. Self-supervised foundation models for EEG and broader biosignals (Lee et al., 2025; Zhou et al., 2025a) increasingly rely on masked-prediction pretexts, but their gains are usually reported at the level of a full pipeline. Recent models change tokenisation (Jiang et al., 2024; Yang et al., 2023), attention or channel aggregation (Wang et al., 2025; Zhou et al., 2025b; Döner et al., 2025), the pretext objective (Kostas et al., 2021; Cui et al., 2024; Wang et al., 2024), and even the recording regime (Wang et al., 2023). This diversity makes it difficult to tell which design choices are responsible for downstream improvements, and motivates controlled studies that isolate one axis at a time.

Closest to our setting, REVE (Ouahidi et al., 2025b) pairs a vanilla transformer with an MAE objective and a mixed spatio-temporal block mask at ρ=0.55; ST-EEGFormer (Yang & Hulle, 2024) similarly shows that a plain MAE on raw patches can compete with more elaborate EEG architectures. Whereas these works propose complete pipelines with a fixed mask, we keep the REVE pipeline as a reference point and ablate only the mask geometry: a controlled sweep of 29 (L, r) configurations, replicated under both MAE and JEPA, yielding 58 checkpoints that map shared optima, shared failure modes, and a JEPA-specific collapse.

Masking strategies for sequential and multichannel signals. Block-masked prediction has a clear lineage from BERT (Devlin et al., 2019) for language, through MAE (He et al., 2022) for images, to ST-MAE (Feichtenhofer et al., 2022) for video, where the choice of mask shape is recognised as a primary design axis. EEG masked models also vary their masks, but usually as a fixed design choice rather than as an object of study: Signal-JEPA (Guetschel et al., 2024) uses channel blocks, EEG-VJEPA (Hojjati et al., 2026) adapts video-JEPA spatio-temporal tubelets, EEGPT (Wang et al.,

2024) uses a high-radius mixed mask, CBraMod (Wang et al., 2025) masks patches independently at $\rho \approx 0.5$, EEG2Rep (Mohammadi Foumani et al., 2024) preserves informative subsequences while predicting latent targets, and MAEEG (Chien et al., 2022) reports that concentrated temporal masks improve downstream sleep staging. These choices cover single-channel, full-channel, temporal, and spatio-temporal regimes, but the EEG literature lacks a common parameterization that shows how these regimes relate to one another or how performance changes between them. Our (L, r) parameterization recovers these shapes as boundary cases and indexes a continuous interpolation between them under a fixed pipeline.

MAE vs. JEPA beyond EEG. The contrast between input-space reconstruction and latent-space prediction has been studied at length outside EEG, in vision (He et al., 2022; Assran et al., 2023), video (Bardes et al., 2024; Feichtenhofer et al., 2022), and general-purpose SSL (Baevski et al., 2022b;a). Closest to our work, Van Assel et al. (2025) give a closed-form characterization of when latent-space prediction outperforms pixel-space reconstruction under high-variance observation noise, a regime that arguably matches EEG. While their analysis is theoretical and based on synthetic Gaussian noise, we measure the MAE–JEPA contrast empirically across 12 real EEG datasets and 58 pre-trained checkpoints under a fixed pipeline.

Ablation-centric studies. Several ablation-centric studies have shaped SSL and transformer practice without introducing a new architecture: Tian et al. (2020) characterized what makes a good view for contrastive learning, Steiner et al. (2021) disentangled pipeline choices for vision transformers, and Zhai et al. (2022) mapped scaling behavior for vision transformers. Our work follows this tradition in EEG: rather than proposing another foundation model, it isolates masking geometry as a critical pre-training axis and provides the controlled ablation that is missing from the current EEG literature.

3 Controlled evaluation of masking geometry and SSL framework

We propose a controlled experimental design in which 58 EEG foundation models are pre-trained and evaluated under a single fixed recipe, varying only two axes: the SSL framework (MAE vs. JEPA) and the spatio-temporal masking geometry (L, r). Any difference in downstream performance can therefore be attributed to one of these two axes or their interaction.

3.1 A shared pre-training pipeline

A shared training pipeline. First, we fix the backbone shared across all our foundation models (FMs) to the *REVE-Small* variant of REVE (Ouahidi et al., 2025b). We choose REVE because three benchmarks, OpenEEGBench (Guetschel et al., 2026a), NeuralBench (Banville et al., 2026) and NeuroAtlas (Kontras et al., 2026), independently identified it as the strongest published EEG foundation model. We nevertheless test the robustness of our results to the backbone in Appendix H.1. The REVE architecture consists of a transformer with 4 encoder layers, 2 decoder layers, 8 attention heads, and a token dimension of 512, for a total of $\sim$12.7 M parameters, excluding the decoder which is unused for downstream evaluations. This architecture uses GEGLU feed-forward blocks (Shazeer, 2020), RMSNorm (Zhang & Sennrich, 2019), the LLaMA-style FFN arrangement (Touvron et al., 2023), and a fixed 4D sinusoidal positional encoding (see Appendix A.1 for a complete description). Then, we fix the pre-training dataset for all our FM trainings to use the open part of the corpus used to train REVE. REVE is trained on a corpus of approximately 6 TB, which bundles a mix of open-licence datasets and access-restricted datasets. We restrict ourselves to the open-licence subset consisting of 4.4 TB so that we can freely distribute the resulting pre-trained checkpoints and training scripts. We also fix the optimization procedure for all our training, using AdamW (Loshchilov & Hutter, 2017) with a one-epoch warmup, a one-cycle cosine learning-rate schedule, and bf16 mixed precision (see Appendix A.2 for all training hyperparameters). The input of the network is fixed to 30 s windows with 32 channels. The training input windows are non-overlapping windows from the whole open corpus, rescaled with a robust scaler whose statistics are estimated only on the current window, and clipped to avoid values exceeding 15 standard deviations. We randomly sample 32 channels, padding with zeros when fewer channels are present. Windows are tokenised identically across configurations, with tokens of 1 s length and 0.1 s overlap. Each configuration is pre-trained

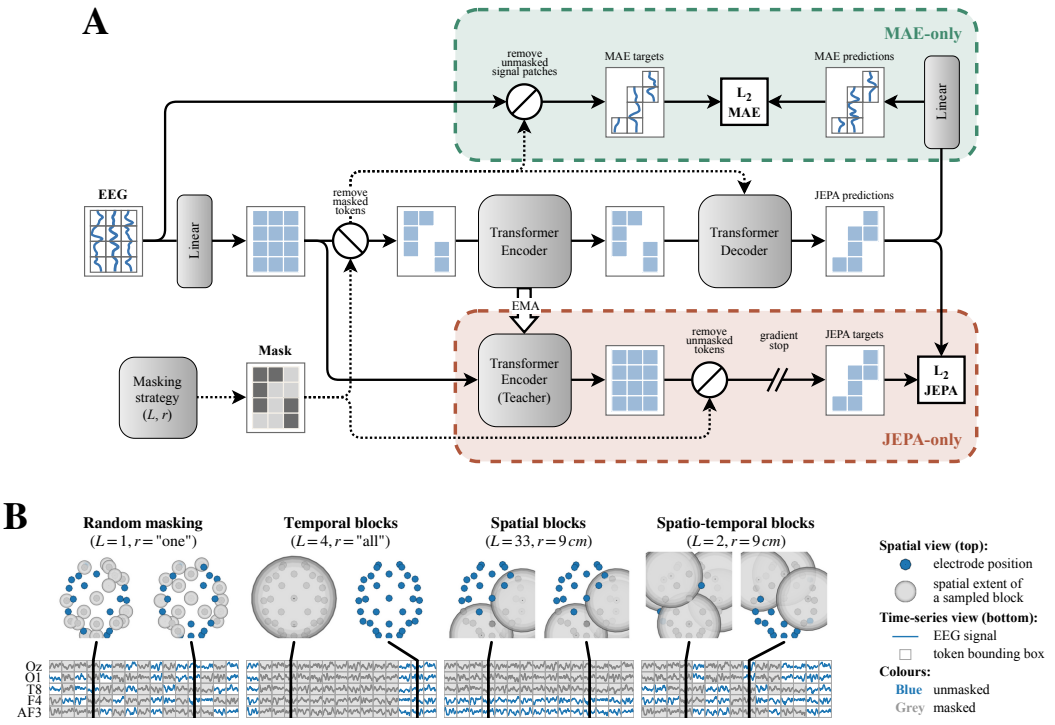


Figure 1: Controlled cross of two design axes for EEG masked-prediction pre-training. **(A) Framework axis.** The MAE and JEPA pretexts are superimposed on a shared pipeline; components outside the coloured dashed boxes (linear tokeniser, masking, encoder, decoder) are shared between the two, while each pretext adds its own branch (green: MAE input-space reconstruction; orange: JEPA latent prediction against an EMA teacher). **(B) Masking axis.** Block-masks parameterised by spatial radius r, temporal length L, and target ratio ρ. This formalism unifies the masking strategies used in the EEG self-supervised literature, recovering four canonical regimes at extreme (L, r) values: random patches, pure temporal blocks, pure spatial blocks, and spatio-temporal blocks. We sweep 29 (L, r) configurations at fixed $\rho = 0.55$. Holding every other component fixed (backbone, optimiser, corpus, tokenisation, downstream protocol) yields the $58 = 2 \times 29$ pre-trained foundation models that we evaluate on OpenEEGBench under a linear probe on frozen features.

on the same fixed number of seen tokens, corresponding to 10 epochs over the corpus. Additional details on this shared pipeline are given in Appendix A. We verified in preliminary experiments that our pipeline reproduces results close to those of REVE before launching the full ablation study.

The Framework axis. For reconstruction-based self-supervised learning, two main frameworks are competing. On one side, Masked Autoencoders (*MAE*) reconstruct masked input patches in the input space from the visible (unmasked) tokens, under an L_2 reconstruction loss. Typically, a final linear layer projects the decoded output tokens back to the input space. This framework is relatively straightforward and therefore uses neither a secondary loss nor a teacher network. On the other side, Joint-Embedding Predictive Architectures (*JEPA*) replace input-space reconstruction with a latent-space prediction: both the context and the target are first embedded into a latent space, and the target embeddings are then predicted from the context embeddings. To provide stable targets during training, the target embeddings are produced by an Exponential Moving Average (EMA) teacher that receives the full (unmasked) input and outputs the embeddings of the masked tokens. These target embeddings are predicted from the embeddings of the unmasked tokens under an L_2 loss. To stay in the simplest setting and vary as little as possible across configurations, we deliberately add no secondary loss to stabilise training. Although JEPA variants often add explicit anti-collapse terms such as VICReg-style variance–covariance penalties (Bardes et al., 2021), C-JEPA's contrastive correction (Mo & Tong, 2024), or SIGReg in LeJEPA (Balestriero & LeCun, 2025), we treat these as separate pipeline choices rather than part of the framework axis. This keeps JEPA at strict parity

with MAE on the "no auxiliary loss" axis. Preliminary runs showed no representational collapse at the REVE scale under this setting, so we avoided adding anti-collapse regularisation to keep the comparison as controlled as possible. These two frameworks are illustrated in Figure 1 (A), with additional details given in Appendices A.6 and A.7.

The Masking axis. Here, we unify the various masking strategies proposed in the literature into a single parametric family. As EEG signals are spatio-temporal signals, with auto-correlation in both space and time, the definition of masking strategies needs to account for these two directions. Let $x \in \mathbb{R}^{N_c \times N_t}$ denote a tokenized EEG segment with N_c (non-padding) channels at scalp positions $\{p_c\}_{c=1}^{N_c} \subset \mathbb{R}^3$ and N_t temporal patches. A binary mask $M \in \{0,1\}^{N_c \times N_t}$ is defined based on a set of K blocks $(B_1, \dots, B_K)$ parameterized by a spatial center and temporal start (c_i, t_i), associated with a spatial radius r and a temporal length L. This parameterization controls the span of the mask in both directions through the dimensions of the block:

$$B_i = \left\{ (c,t) \ : \ \|p_c - p_{c_i}\|_2 \leq r \ \text{ and } \ t_i \leq t \leq t_i + L - 1 \right\}. \tag{1}$$

The final mask is defined by $M_{c,t} = 1$ when $(c,t) \in \bigcup_{i=1}^{K} B_i$. The mask is then applied only to non-padding channels (at most 32).

To select the blocks B_i, each block is sampled independently from the others by sampling (c_i, t_i) uniformly, with c_i drawn uniformly over channels and $t_i \sim \text{Unif}\{1, \dots, N_t - L + 1\}$. The number of blocks K is chosen such that the expected mask ratio equals a fixed target $\rho^\star$, while accounting for overlaps between blocks. Since the K blocks are sampled i.i.d., a fixed cell (c,t) is left unmasked with probability $(1-f)^K$, where $f \in [0,1]$ denotes the expected fraction of cells covered by a single block. Setting $\mathbb{E}[\rho] = 1 - (1-f)^K$ equal to $\rho^\star$ and rounding up gives:

$$K \ = \ \left\lceil \frac{\log(1-\rho^\star)}{\log(1-f)} \right\rceil, \qquad f \ = \ f_{\text{sp}}(r) \cdot \frac{L}{N_t}, \tag{2}$$

where $f_{\text{sp}}(r)$ is the expected fraction of channels within a distance r of a uniformly drawn center channel. We use a uniform-scalp approximation $f_{\text{sp}}(r) = \pi r^2 / A_{\text{scalp}}$, with A_{scalp} the scalp surface area. The full sampling procedure is given in Algorithm 1 in Appendix A.4. In the following, we fix $\rho^\star = 0.55$ based on the pre-training performance obtained on initial training of the REVE architecture (Ouahidi et al., 2025b); Appendix H.2 explains why ρ cannot be varied jointly with (L, r) and shows that our conclusions hold across ratios.

The two extreme values of r are defined as special cases with exact closed forms: $f_{\text{sp}} = 1/N_c$ when r is below the smallest inter-channel distance (each block covers a single channel), and $f_{\text{sp}} = 1$ when r exceeds the largest inter-channel distance (each block covers every channel). We denote these two special cases as $r =$ `"one"` and $r =$ `"all"`, respectively.

This masking strategy yields spatio-temporal masks, similar to the block-masking strategy used in REVE (Ouahidi et al., 2025b) and Signal-JEPA (Guetschel et al., 2024). Particular values of r and L recover other strategies: using $L = 1, r =$ `"one"` yields the random masking strategy (Wang et al., 2025), while $r =$ `"all"` with short blocks yields the temporal block masking, a strategy that forces the model to learn to predict a full time-segment. Finally, $L = 33$ yields spatial blocks, where the goal of the model is to reconstruct a complete channel from the information of other channels.

3.2 Downstream evaluation protocol

Downstream protocol. To evaluate the obtained foundation models, we consider their performance on a set of diverse downstream tasks using only linear probing. A core promise of FMs is to produce rich embeddings of EEG recordings, which is best measured with a simple linear model on top of them, as it shows how well the model disentangles complex information. This protocol also maximizes the number of downstream datasets that can be evaluated under a fixed compute budget while still providing a strong signal of representation quality. To this aim, our evaluation relies on OpenEEGBench (Guetschel et al., 2026a), an open-source, community-driven benchmark to evaluate EEG-FM, originally introduced in Guetschel et al. (2026b).All pre-trained checkpoints are evaluated with a ridge-regression linear probe fitted on frozen representations, on the 12 OpenEEGBench datasets. We strictly follow the train/validation/test splits provided by the benchmark. Final models (epoch 10) are evaluated across 5 seeds that vary the random projection (the only stochastic component

of the foundation-model downstream pipeline). The per-dataset feature dimensions before the probe's random projection are reported in Appendix C.2, and per-dataset class distributions in Appendix C.1. Additional details on the downstream protocol are given in Appendix C.

Score aggregation and statistical testing. To aggregate the results for each model, raw scores are first normalised within the dataset (see Appendix C.4 for the exact per-dataset normalisation) and then aggregated as the mean of the normalised scores over the datasets and the seeds. Raw per-dataset scores are reported in Appendix F. Using the repeated evaluation across seeds and datasets, we also perform statistical testing using a two-level bootstrap that propagates the variance at both levels. This assesses if a configuration is better than the others, in the spirit of Demšar (2006) and Agarwal et al. (2021). At each of the $T = 10\,000$ bootstrap replicates, the 12 OpenEEGBench datasets are resampled with replacement (treating the benchmark as a sample of plausible EEG datasets), and within each resampled dataset, the 5 seeds of every configuration are themselves resampled with replacement, independently across configurations. Because raw scores are not comparable across datasets, we rank configurations within each dataset before averaging the 12 within-dataset ranks into a single mean rank $\bar{R}_c^{(t)} \in [1, 29]$ per configuration c. We summarize the resulting joint distribution by the pairwise dominance probability $\mathbb{P}(c \succ c') = \frac{1}{T}\sum_t \mathbf{1}[\bar{R}_c^{(t)} < \bar{R}_{c'}^{(t)}]$. We choose $T = 10\,000$ so that the Monte Carlo standard error of $\mathbb{P}(c \succ c')$ stays below 0.005 for any underlying value, comfortably above the $T \geq 2\,000$ recommended by Agarwal et al. (2021). We call configuration c *strictly better than* c' when $\mathbb{P}(c \succ c') > 0.975$; pairs with $\mathbb{P} \in [0.025, 0.975]$ are reported as equally ranked.

Baselines. The performance of our pre-trained models is compared against two families of baselines, both evaluated within the OpenEEGBench framework. First, we compare our models to the *REVE baseline*, which corresponds to the REVE architecture with pre-trained weights from the REVE-Base checkpoint (69 M parameters, pre-trained on the full 6 TB REVE corpus) released on HuggingFace (Ouahidi et al., 2025b). This model is evaluated under the same ridge-regression linear probe on frozen features as our models. A comparison with other published EEG foundation models is given in Appendix G. Additionally, we report the performances of three non-foundation baselines, chosen to span state-of-the-art models for the various tasks included in OpenEEGBench. We include EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2023), and ShallowFBCSPNet (Schirrmeister et al., 2017b), using their Braindecode (Aristimunha et al., 2026) reference implementations. These models are also trained and evaluated within OpenEEGBench under a full per-task training protocol. Here, our evaluation departs from the evaluation of FMs due to the nature of these supervised baselines: our goal is not to investigate the quality of the learned representations, but to provide a reference point for each task.

3.3 Grid sweep design

We pre-train multiple EEG foundation models that span the Cartesian product of the two reconstruction-based SSL frameworks (MAE and JEPA) and a grid of spatio-temporal masking configurations parameterised by a temporal length L and a spatial radius r. The grid spans $r \in \{\texttt{"one"}, 6, 9, 12, \texttt{"all"}\}$ cm and $L \in \{1, 2, 4, 8, 16, 33\}$ patches, yielding 30 configurations. We remove the degenerate $(L = 33, r = \texttt{"all"})$ where all the signal is masked, resulting in a total of 29 configurations per framework. The per-cell number of blocks K is listed in Table 4 of Appendix A.5, where we also verify empirically that the actual mask ratio matches $\rho^\star = 0.55$ across the 29 cells.

4 Masking geometry drives downstream performance

We investigate the downstream performance of the 58 pre-trained models of the core grid and uncover the central role of spatio-temporal masking geometry. We first report headline results at the optimal mask configuration (Figure 2), establishing that, with the right mask, our pipeline learns embeddings on par with the REVE baseline under both frameworks, at a fraction of its pre-training cost. We then open the box: Figure 3 maps the full (L, r) grid and reveals which configurations consistently win or fail, and investigate why, anchored by the spatial-vs.-temporal redundancy structure of EEG. Per-dataset breakdowns are in Appendix F.

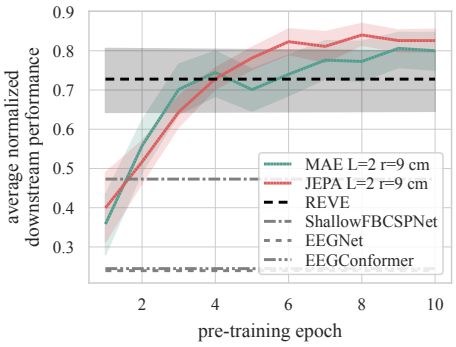


Figure 2: Mean normalised downstream performance on OpenEEGBench across the 12 datasets and 5 probe seeds, as a function of the pre-training epoch, for both frameworks at the optimal mask $(L, r) = (2\,\text{s}, 9\,\text{cm})$. Shaded bands are the 95% confidence interval across seeds and datasets. The dashed black line locates the REVE baseline; horizontal dashed grey lines locate the three supervised baselines (EEGNet, EEGConformer, ShallowFBCSPNet) trained end-to-end per task. Both frameworks reach the REVE baseline level by epoch $\sim$ 4 and remain at or above it.

Headline at the optimal mask. Figure 2 previews the main outcome of this section. At the optimal mask geometry $(L, r) = (2\,\text{s}, 9\,\text{cm})$, both embeddings learned with MAE and JEPA reach REVE-baseline level within a few pre-training epochs and stabilize at or above it throughout the rest of the training. These performances are obtained with fewer parameters and a fraction of the compute used to pre-train the REVE baseline. Choosing a suitable mask is therefore sufficient to match REVE-Base-level performance under both pretexts; the rest of this section unpacks how we arrive at this optimum, why this mask works, and where the approach breaks. The supervised baselines (EEGNet, EEGConformer, ShallowFBCSPNet) are trained end-to-end on each downstream task and are shown only as a per-task reference for what task-specific supervised training achieves; they are not directly comparable to the foundation-model curves, which use a frozen linear probe.

Spatial vs. temporal redundancy. The two axes along which we mask reveal very different structures. The within-channel autocorrelation in the EEG drops sharply with increased time-lag, falling below $R^2 \approx 0.2$ within $\sim$2 s, whereas between-channel correlation decays much more slowly with electrode distance and remains non-negligible across the entire scalp due to substantial volume conduction in EEG. This asymmetry anchors the qualitative reading of the masking grid: increasing L depletes the same-channel temporal context of a masked patch, forcing the predictor to look across channels; increasing r depletes the cross-channel spatial context, forcing it to look across time. The full characterisation is given in Appendix D. Figure 3 then asks how this redundancy structure shapes what the model learns: does masking along the temporal axis produce worse representations than masking across the spatially redundant channel axis? Four findings answer this question.

Finding 1: shared optimum and shared failure modes. Both frameworks converge on the same optimal mask, $(L, r) = (2\,\text{s}, 9\,\text{cm})$, and on the same failure modes (Figure 3 (A,B)). The redundancy argument predicts both: at $r =$ `"one"`, the masked patch is trivially predictable from its same-channel temporal neighbors, so the pretext provides too little learning signal; at $r =$ `"all"`, spatial context is entirely removed, leaving only the weakly predictive temporal axis — a degenerate regime that heavily impacts JEPA (Finding 3). Most long temporal blocks ($L \in \{8, 16\}$ at moderate r) similarly underperform, consistent with temporal redundancy decaying faster than spatial redundancy. A notable exception is $L = 33$, where the block spans the full window and the mask reduces to masking entire channels: pure cross-channel prediction, hard but tractable given the strong spatial redundancy of EEG. This regime is not the end of the L continuum but a different pretext task: a fully masked channel can only be inferred from its scalp position, whereas at $L \leq 16$ the masked channel remains partly visible and its relation to the neighbouring channels can be read off the visible part. MAE handles this well (three $L = 33$ cells in the ex-aequo cluster); JEPA less so (one cell).

Finding 2: performance is robust outside the failure modes. Once the failure modes are excluded, the choice of (L, r) matters surprisingly little: 11 MAE and 9 JEPA configurations are statistically indistinguishable from their respective framework's best under the hierarchical bootstrap (white circles in Figure 3 (B)). The *ex-aequo* cluster spans moderate radii ($r \in \{6, 9, 12\}$ cm) and short lengths ($L \in \{1, 2, 4\}$), with a secondary lobe at $L = 33$, consistent with the asymmetry identified in Finding 1, where full-channel masking is hard but tractable. This robustness is not symmetric across frameworks: the within-framework spread (best minus worst normalised score) is 0.32 for MAE and 0.77 for JEPA, meaning a poor mask choice is 2.4$\times$ more costly under JEPA than under MAE.

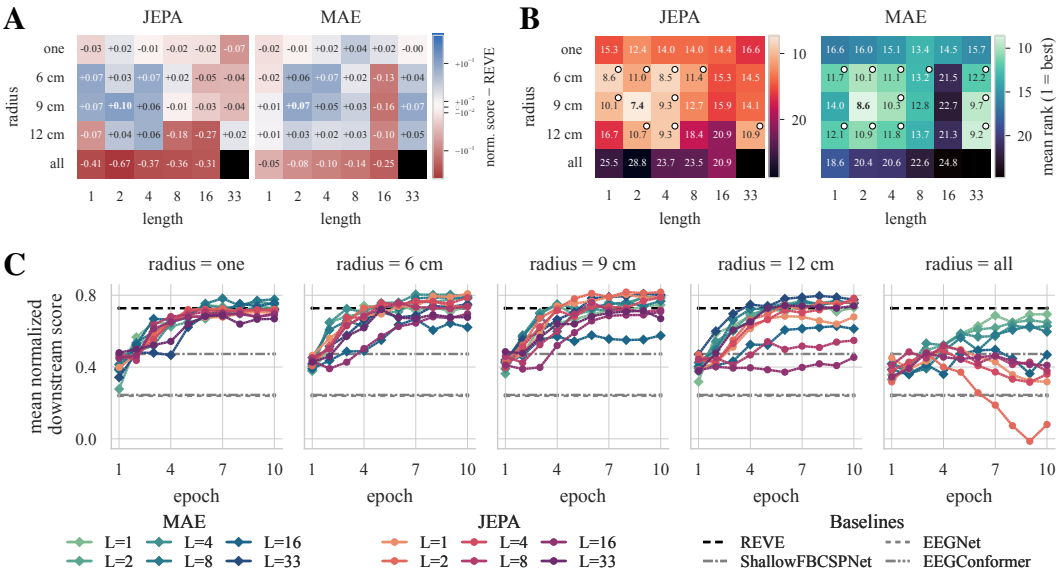


Figure 3: Effect of the masking geometry (L, r) on downstream performances. **(A) Mean normalised score** above the REVE baseline (0 = the public REVE-Base checkpoint) across the 12 OpenEEG-Bench datasets $\times$ 5 probe seeds. **(B) Mean rank** (1 = best) per cell under the hierarchical bootstrap ($T = 10{,}000$ replicates over datasets and seeds). The bold cell in each framework is the bootstrap winner; white circles mark cells not statistically distinguishable from it at $\mathbb{P}(\text{winner} \succ c) \leq 0.975$ (ex-aequo cluster). **(C) Training trajectories:** score of (A) as a function of the pre-training epoch, faceted by spatial radius r, for both frameworks and all six values of L. Dashed lines locate the REVE baseline (black) and the three supervised baselines (grey). All foundation-model curves use a frozen-feature ridge probe; supervised baselines are trained end-to-end per task.

Finding 3: bias-inflation collapse of JEPA at $r =$"all". Figure 3 (C) reveals two patterns. First, the across-dataset normalised score increases with pre-training epochs for nearly all configurations: pre-training leads to downstream performance improvements under most masks. Second, the $r =$ `"all"` panel exposes a clear exception that affects only JEPA: the MAE bundle plateaus below the REVE baseline but does not deteriorate, whereas the JEPA bundle descends systematically from epoch $\sim$3 onwards, so more pre-training makes the downstream score worse. We name this regime *bias-inflation collapse* and discuss its mechanism in Bias-inflation collapse; a complete analysis is given in Appendix E.

Finding 4: REVE-Base-level performance for a fraction of its compute. At the optimal mask $(L, r) = (2\,\text{s}, 9\,\text{cm})$, both frameworks reach the level of the REVE baseline (Figure 2), while each of our pre-training runs costs around 14 H100-hours (Appendix A.3), compared to the 260 A100-hours reported for REVE-Base (Ouahidi et al., 2025b). This efficiency stems from the whole pipeline — our models are smaller (12.7M vs. 69M parameters), pre-trained on less data (4.4 vs. 6 TB) on a shorter schedule — not from mask geometry alone: under the same pipeline and budget, the failure modes of Finding 1 fall well short of this level. Beyond this study, the fact that a competitive model can be pre-trained in 14 H100-hours lowers the barrier for future work; our open codebase is designed to let researchers rapidly prototype and evaluate new EEG foundation model ideas, and calls for systematic scaling-law studies in this domain.

5 Discussion and Limitations

Under the shared pipeline adopted in this work, the masking strategy is the dominant lever of downstream performance: the within-framework spread (0.32 normalised score for MAE, 0.77 for JEPA) exceeds the cross-framework gap at the optimum, where MAE and JEPA perform comparably. With well-chosen mask parameters $(L, r) = (2\,\text{s}, 9\,\text{cm})$, both frameworks reach REVE-Base-level downstream performance at a fraction of its pre-training compute.

Why does the mask matter? Beyond reporting which masks work, the pattern of failures and successes reveals something deeper: the masking geometry acts as a probe of the signal's redundancy structure, and downstream performance reflects how well the pretext task forces the network to exploit non-trivial statistical dependencies in the data. At $r =$ `"one"`, the high spatial redundancy of EEG makes the pretext trivially solvable from same-time context alone. The network never needs to integrate information across times, and the representations it learns carry little structure beyond spatial smoothness. At $r =$ `"all"`, the extreme opposite, the spatial context is entirely removed: the only surviving signal is the weakly predictive temporal axis, and the pretext degenerates into a task the network solves by ignoring the EEG content altogether. The successful regimes sit between these extremes, when the mask removes enough spatial redundancy to make the pretext non-trivial, yet leaves enough context for the network to learn genuinely informative representations. This suggests a general principle for SSL on multichannel biosignals: a good mask should be calibrated to the signal's redundancy structure, targeting the axis along which information is shared but not trivially accessible.

Bias-inflation collapse: a multichannel-specific failure mode of JEPA. At $r =$ `"all"`, JEPA admits a degenerate solution that MAE structurally cannot: MAE has no teacher network and its target is the per-channel raw input patch, so any failure to reproduce channel-specific signals is penalized at training time; under JEPA, the target is a contextual latent feature produced by the EMA teacher, and the predictor can, in principle, match this target by reproducing a per-channel constant from positional embeddings alone, without ever needing to resolve the EEG content. This shortcut is unlocked specifically at $r =$ `"all"`: when every channel of the same time-patch is masked together, no same-time spatial neighbor survives in the predictor's context, so the easiest student-teacher agreement collapses onto the position-only lookup table; at smaller r, spatial neighbors preserve a non-trivial input-driven path that the optimizer preferentially follows. We confirm in Appendix E that this is the regime the optimizer converges to: the encoder learns a trivial channel-position lookup table, the per-channel feature mean inflates by a factor of 2 to 3 while the input-driven residual shrinks by a factor of 5 to 10, and standard variance-preserving regularizers (VICReg, C-JEPA) do not rescue the regime because they are themselves satisfied by the inflated lookup-table representation. The collapse is spectrally selective, keeping broadband amplitude but losing the slow structure of the signal (Appendix E.3), and it persists with 0.5 s patches (Appendix H.3), so it is not an artefact of the tokeniser. The methodological implication is direct: the in-domain collapse metrics computed at training time (variance, EMA divergence, loss thresholds) are insufficient to validate JEPA pre-training in the multichannel setting, and building a reliable in-training detector is itself non-trivial on multi-montage corpora (Appendix E.6). Until such a detector exists, a downstream probe is required to expose the failure, and the practical mitigation is simply to avoid the masking regime that triggers it ($r =$ `"all"`) — consistent with the recommendations of Finding 2. We insist, however, that the deeper concern is not that $r =$ `"all"` fails, but that the failure is *invisible* to every standard diagnostic.

What we recommend to practitioners. For practitioners pre-training EEG foundation models, we recommend $(L, r) = (2\,\text{s}, 9\,\text{cm})$ as a robust default under both pretexts; within the ex-aequo cluster ($r \in \{6, 9, 12\}$ cm, $L \in \{1, 2, 4\}$) further tuning is unlikely to be productive. The recommendation holds on two other backbones (Appendix H.1) and across mask ratios (Appendix H.2). Avoid $r =$ `"one"` (pretext too easy, and its best ρ depends on the montage; Appendix H.2) and $r =$ `"all"` (bias-inflation collapse under JEPA, underperformance under MAE). If JEPA is the chosen framework, mask choice requires more care than under MAE, whose stability across the grid makes it more forgiving. Finally, optimise the mask before scaling compute: at a fixed budget, a poor mask costs up to 0.77 normalised score.

Limitations. The core grid uses a single backbone and a single pre-training corpus (dominated by clinical EEG and BCI paradigms); transfer to other corpora, hyperparameters, and consumer-grade EEG remains to be established. The grid fixes the mask ratio at $\rho = 0.55$; ρ and (L, r) interact (Appendix H.2), and a full joint sweep remains to be done. One of the 12 downstream datasets, PhysioNet-MI, is part of our pre-training corpus (48.5 h out of $\sim$34,000 h); the other eleven are unseen. The REVE-Base checkpoint, our main baseline, additionally saw TUAB and TUEV, and scores above both our $(2\,\text{s}, 9\,\text{cm})$ models on all three, so this exposure does not inflate our results. More broadly, OpenEEGBench overlaps with the pre-training corpora of other recent EEG foundation models, so collective benchmark overfitting cannot be ruled out.

AI USE STATEMENT

In this work, we used generative AI tools for one task with required disclosure: implementing methods. This covers code autocompletion while programming, as well as some larger implementation tasks carried out more autonomously from detailed specifications written by the authors; the architecture of the code base and all its design principles are the authors' own. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to propose or refine hypotheses, to design or provide feedback on the research methodology or experiments, to assist with translation, or to interpret results: the design of all experiments and the interpretation of all results are the authors' own. Generating synthetic data, proving mathematical claims or writing proofs, cleaning or reformatting datasets, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: editing the paper for grammar, spelling and word choice, drafting parts of the paper, searching for relevant literature, and facilitating the running of experiments (monitoring cluster jobs and running smoke tests). We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, code and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only previously collected, publicly available EEG recordings of human participants, released by their original authors: the open-licence subset of the REVE pre-training corpus, whose sources are cited in Appendix B, and the 12 downstream datasets of OpenEEGBench, described in Appendix C. No new data were recorded and no interaction with human participants took place for this study. We restricted pre-training to the open-licence part of the corpus so that the resulting checkpoints and code can be redistributed. Beyond the general dual-use concerns of brain-signal decoding, such as inferring cognitive or clinical states without informed consent, which apply to the field as a whole, we do not foresee harmful applications specific to this work.

REPRODUCIBILITY STATEMENT

All 58 models share the single recipe of section 3, so reproducing the study amounts to reproducing one pipeline under 29 mask configurations and two objectives. The code for pre-training and downstream evaluation, together with the cached results of the main grid, is available at the anonymous link given in the introduction. The backbone is specified in Appendix A.1, the training hyperparameters in Appendix A.2, the compute cost in Appendix A.3, and the two objectives in Appendices A.6 and A.7. The mask sampler is fully determined by (2) and Algorithm 1 (Appendix A.4), and its realised mask ratio is checked in Appendix A.5. The pre-training corpus is listed in Appendix B; the downstream datasets, splits, probe and score normalisation are given in Appendix C, and the statistical testing procedure in subsection 3.2.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. Curran Associates, Inc., 2021. doi: 10.48550/arXiv.2108.13264. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Blanca Aguado-Lopez, Ana F. Palenciano, Jose M. G. Penalver, Paloma Diaz-Gutierrez, David Lopez-Garcia, Chiara Avancini, Luis F. Ciria, and Maria Ruz. Competition_eeg, 2024. URL https://openneuro.org/datasets/ds005089/versions/1.0.1.

Edilberto Amorim, Wei-Long Zheng, Jong Woo Lee, Susan Herman, Mohammad Ghassemi, Adithya Sivaraju, Nicolas Gaspard, Jeannette Hofmeijer, Michel J A M van Putten, Matthew Reyna, Gari Clifford, and Brandon Westover. I-CARE: International Cardiac Arrest REsearch consortium Database. *PhysioNet*, June 2023. doi: 10.13026/avek-0p97. URL https://doi.org/10.13026/avek-0p97. Version 2.0.

Carlos Valle Araya, Carolina Mendez-Orellana, and Maria Rodriguez-Fernandez. Large spanish eeg, 2024. URL https://openneuro.org/datasets/ds004279/versions/1.1.2.

Bruno Aristimunha, Pierre Guetschel, Martin Wimpff, Lukas Gemein, Cedric Rommel, Hubert Banville, Maciej Sliwowski, Daniel Wilson, Simon Brandt, Théo Gnassounou, Joseph Paillard, Aman Srivastava, Bruna Junqueira Lopes, Sara Sedlar, Sarthak Tayal, Young Truong, Thomas Moreau, Sylvain Chevallier, Alexandre Gramfort, and Robin Tibor Schirrmeister. Braindecode: Toolbox for decoding raw electrophysiological brain data with deep learning models, 2026. URL https://zenodo.org/doi/10.5281/zenodo.19421349. Version v1.4.0.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01499. URL https://ieeexplore.ieee.org/document/10205476/.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech and Language, 2022a. URL https://arxiv.org/abs/2212.07525.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language, 2022b. URL https://arxiv.org/abs/2202.03555.

Imad J. Bajwa1, Andre S. Nilsen1, René Skukies1,3, Arnfinn Aamodt1, Gernot Ernst2, Johan F. Storm1, and Bjørn E. Juel1,2. A repeated awakening study exploring the capacity of complexity measures to capture dreaming during propofol sedation, 2024. URL https://openneuro.org/datasets/ds005620/versions/1.0.0.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, 2025. URL http://arxiv.org/abs/2511.08544.

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Mingfang, Zhang, Elisa Cascardi, Katelyn Begany, Teon Brooks, and Jean-Rémi King. NeuralBench: A unifying framework to benchmark NeuroAI models, 2026. URL https://arxiv.org/abs/2605.08495.

Amir Bar, Florian Bordes, Assaf Shocher, Mahmoud Assran, Pascal Vincent, Nicolas Ballas, Trevor Darrell, Amir Globerson, and Yann LeCun. Stochastic positional embeddings improve masked image modeling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. PMLR, 2024. doi: 10.48550/arXiv.2308.00566. URL https://arxiv.org/abs/2308.00566.

Alexandre Barachant. *Commande robuste d'un effecteur par une interface cerveau machine EEG asynchrone*. PhD thesis, Université de Grenoble, 2012.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, 2021. URL https://arxiv.org/abs/2105.04906.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video, 2024. URL http://arxiv.org/abs/2404.08471.

Cemre Baykan and Alexander C. Schütz. Eeg responses to the number of objects in partially occluded and uncovered scenes, 2024. URL https://openneuro.org/datasets/ds005586/versions/2.0.0.

Benjamin Lowe (Ben.Lowe@Mq.Edu.Au), Jonathan Robinson (Jonathan.Robinson@Monash.Edu), Naohide Yamamoto (Naohide.Yamamoto@Qut.Edu.Au), Hinze Hogendoorn (Hinze.Hogendoorn@Qut.Edu.Au), and Patrick Johnston (Dr.Pat.Johnston@Icloud.Com). Visual attribute-specific contextual trajectory paradigm, 2023. URL https://openneuro.org/datasets/ds004603/versions/1.1.0.

Ole Bialas, Emily Teoh, Andrew Anderson, and Edmund Lalor. Invariant encoding of phonemes in neural responses to continuous speech, 2023. URL https://openneuro.org/datasets/ds004408/versions/1.0.8.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. BCI Competition 2008 – Graz data set A. Technical Report 16, Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, 2008. URL https://www.bbci.de/competition/iv/desc_2a.pdf.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021. doi: 10.48550/arXiv.2104.14294. URL https://arxiv.org/abs/2104.14294.

James F Cavanagh and Trevor C J Jackson. Mood manipulation and pst, experiment 1, 2022. URL https://openneuro.org/datasets/ds004315/versions/1.0.0.

Luis Alberto Barradas Chacón and Selina C. Wriessnegger. Toonfaces, 2023. URL https://openneuro.org/datasets/ds004324/versions/1.0.0.

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. doi: 10.1038/s41597-023-02650-w.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. MAEEG: Masked Auto-encoder for EEG Representation Learning, 2022. URL http://arxiv.org/abs/2211.02625.

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.

Chuqin Xiang, Xinrui Fan, Duo Bai, Ke Lv, and Xu Lei. A resting-state eeg dataset for sleep deprivation, 2025. URL https://openneuro.org/datasets/ds004902/versions/1.0.8.

Jose Cordoba-Silva, Rafael Maya, Mario Valderrama, Luis Felipe Giraldo, William Betancourt-Zapata, Andrés Salgado-Vascob, Juliana Marín-Sánchez, Viviana Gómez-Ortega, and Mark Ettenberger. Dataset of electrophysiological signals (eeg, ecg, emg) during music therapy with adult burn patients in the intensive care unit., 2023. URL https://openneuro.org/datasets/ds004840/versions/1.0.1.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-GPT: Towards A Foundation Model For EEG. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024. ISBN 979-8-3503-1333-8. doi: 10.1109/ISBI56570.2024.10635453. URL https://ieeexplore.ieee.org/document/10635453/.

Ian Daly, Nicoletta Nicolaou, Duncan Williams, Faustina Hwang, Alexis Kirke, Eduardo Miranda, and Slawomir J. Nasuto. An eeg dataset recorded during affective music listening, 2024. URL https://openneuro.org/datasets/ds002721/versions/1.0.3.

Arnaud Delorme. Go-nogo categorization and detection task, 2021. URL https://openneuro.org/datasets/ds002680/versions/1.2.0.

Arnaud Delorme and Claire Braboszcz. Meditation vs thinking task, 2021. URL https://openneuro.org/datasets/ds003969/versions/1.0.0.

Arnaud Delorme and Tracy Brandmeyer. Eeg meditation study, 2024. URL https://openneuro.org/datasets/ds001787/versions/1.1.1.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. 7(1):1–30, 2006. URL http://jmlr.org/papers/v7/demsar06a.html.

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*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL http://aclweb.org/anthology/N19-1423.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and Topology-Agnostic Foundation Model for EEG Signal Analysis, 2025. URL http://arxiv.org/abs/2510.22257.

Pablo Rodríguez-San Esteban, Ana B. Chica, and José A. González-López. Neural representation of consciously seen and unseen information, 2024. URL https://openneuro.org/datasets/ds005273/versions/1.0.0.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners, 2022. URL https://arxiv.org/abs/2205.09113.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. PMLR, 2023. doi: 10.48550/arXiv.2210.02885. URL https://arxiv.org/abs/2210.02885.

Lukas Gehrke, Sezen Akman, Albert Chen, Pedro Lopes, and Klaus Gramann. Prediction error, 2024. URL https://openneuro.org/datasets/ds003846/versions/2.0.2.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Tijl Grootswagers, Ivy Zhou, Amanda Robinson, Martin Hebart, and Thomas Carlson. Human electroencephalography recordings from 50 subjects for 22,248 images from 1,854 object concepts, 2022. URL https://openneuro.org/datasets/ds003825/versions/1.2.0.

Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. Eeg-attention-rsvp-exp1, 2023a. URL https://openneuro.org/datasets/ds004816/versions/1.0.1.

Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. Eeg-attention-rsvp-exp2, 2023b. URL https://openneuro.org/datasets/ds004817/versions/1.0.1.

Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. Features-eeg, 2024. URL https://openneuro.org/datasets/ds004357/versions/1.0.1.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention. 2024. doi: 10.48550/ARXIV.2403.11772. URL https://arxiv.org/abs/2403.11772.

Pierre Guetschel, Bruno Aristimunha, Dung Truong, Kuntal Kokate, Thomas Moreau, Michael Tangermann, and Arnaud Delorme. Open EEG Bench: Benchmarking parameter-efficient fine-tuning of EEG foundation models, 2026a. URL https://doi.org/10.5281/zenodo.20060513. Version v0.4.0.

Pierre Guetschel, Bruno Aristimunha, Dung Truong, Kuntal Kokate, Michael Tangermann, and Arnaud Delorme. Toward OpenEEG-Bench: A live community-driven benchmark for EEG foundation models. In *Proceedings of the 34th European Signal Processing Conference (EUSIPCO 2026)*, pp. 1–5, Bruges, Belgium, August 2026b. EURASIP.

Cameron D. Hassall, Yan Yan, and Laurence T. Hunt. Drum trainer, 2022a. URL https://openneuro.org/datasets/ds004152/versions/1.1.2.

Cameron D. Hassall, Yan Yan, and Laurence T. Hunt. Continuous feedback processing, 2022b. URL https://openneuro.org/datasets/ds004262/versions/1.0.0.

Cameron D. Hassall, Yan Yan, and Laurence T. Hunt. Steer the ship, 2022c. URL https://openneuro.org/datasets/ds004264/versions/1.1.0.

Cameron D. Hassall, Laurence T. Hunt, and Clay B. Holroyd. Average task value, 2024. URL https://openneuro.org/datasets/ds004147/versions/1.0.2.

Christoffer Hatlestad-Hall, Trine Waage Rygvold, and Stein Andersson. Srm resting-state eeg, 2022. URL https://openneuro.org/datasets/ds003775/versions/1.2.1.

Marleen Haupt, Monika Graumann, Santani Teng, Carina Kaltenbach, and Radoslaw M. Cichy. Braille letters - eeg, 2024. URL https://openneuro.org/datasets/ds004951/versions/1.0.0.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, 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. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01553. URL https://ieeexplore.ieee.org/document/9879206/.

Jason Helbing, Dejan Draschkow, and Melissa L.-H. Võ. Search superiority recollection familiarity, 2024. URL https://openneuro.org/datasets/ds005189/versions/1.0.1.

Amirabbas Hojjati, Lu Li, Ibrahim Hameed, Anis Yazidi, Pedro G. Lind, and Rabindra Khadka. From Video to EEG: Adapting Joint Embedding Predictive Architecture to Uncover Saptiotemporal Dynamics in Brain Signal Analysis, 2026. URL http://arxiv.org/abs/2507.03633.

Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R. Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 International brain–computer interface competition: A review. 16, 2022. ISSN 1662-5161. doi: 10.3389/fnhum.2022.898300. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2022.898300/full.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. 2024. doi: 10.48550/ARXIV.2405.18765. URL https://arxiv.org/abs/2405.18765.

Michael J. Kahana and Joseph H. Rudoler. Penn electrophysiology of encoding and retrieval study (peers), 2023. URL https://openneuro.org/datasets/ds004395/versions/2.0.0.

Emmanuel K Kalunga, Sylvain Chevallier, and Quentin Barthélemy. Using riemannian geometry for ssvep-based brain computer interface. *arXiv preprint arXiv:1501.03227*, 2015.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. 124:180–192, 2016. ISSN 0169-2607. doi: 10.1016/j.cmpb.2015.10.013. URL https://www.sciencedirect.com/science/article/pii/S0169260715002734.

Konstantinos Kontras, Trui Osselaer, Stylianos G. Mouslech, Angeliki-Ilektra Karaiskou, Guido Gagliardi, Thomas Strypsteen, Mohammad Hossein Badiei, Anku Rani, Maarten Vanmarcke, Miguel Bhagubai, Chanakya Ekbote, Jaedong Hwang, Christos Chatzichristos, Paul Pu Liang, and Maarten De Vos. Neuroatlas: Benchmarking foundation models for clinical eeg and brain-computer interfaces, 2026. URL https://arxiv.org/abs/2605.14698.

Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders calibration-less P300-based BCI with modulation of flash duration Dataset (bi2015a)*. PhD thesis, GIPSA-lab, 2019a.

Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain invaders cooperative versus competitive: multi-user P300-based brain-computer interface dataset (BI2015b)*. PhD thesis, GIPSA-lab, 2019b.

Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders calibration-less P300-based BCI using dry EEG electrodes Dataset (bi2014a)*. PhD thesis, GIPSA-lab, 2019c.

Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain invaders solo versus collaboration: Multi-user P300-based brain-computer interface dataset (bi2014b)*. PhD thesis, GIPSA-lab, 2019d.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. 15:653659, 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.653659. URL https://www.frontiersin.org/articles/10.3389/fnhum.2021.653659/full.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces. 15(5):056013, 2018. ISSN 1741-2560, 1741-2552. doi: 10.1088/1741-2552/aace8c. URL http://arxiv.org/abs/1611.08024.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019a.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019b.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019c.

Na Lee, Konstantinos Barmpas, Alexandros Koliousis, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A Comprehensive Review of Biosignal Foundation Models, 2025. URL https://www.techrxiv.org/doi/full/10.36227/techrxiv.176369849.97173246/v1.

Alexander C. Li, Alexei A. Efros, and Deepak Pathak. Understanding collapse in non-contrastive Siamese representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022. doi: 10.48550/arXiv.2209.15007. URL https://arxiv.org/abs/2209.15007.

Haijie Liu, Penghu Wei, Haochong Wang, Xiaodong Lv, Wei Duan, Meijie Li, Yan Zhao, Qingmei Wang, Xinyuan Chen, Gaige Shi, et al. An eeg motor imagery dataset for brain computer interface in acute stroke patients. *Scientific Data*, 11(1):131, 2024.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022. doi: 10.1109/TCDS.2021.3071170.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2017. URL https://arxiv.org/abs/1711.05101.

Dominique Makowski, An-Shu Te, Stephanie Kirk, and Zi Liang Ngoi. Fakefaceemo_data, 2023. URL https://openneuro.org/datasets/ds004582/versions/1.0.0.

Donia Metwalli, Eslam Ahmed, Antony Emil, Yousef A. Radwan, Mariam Barakat, Anas Ahmed, Amro Omar, and Sahar Selim. Areeg: Arabic inner speech eeg dataset, 2025. URL https://openneuro.org/datasets/ds005262/versions/1.0.1.

Shentong Mo and Shengbang Tong. Connecting Joint-Embedding Predictive Architecture with Contrastive Self-supervised Learning, 2024. URL https://arxiv.org/abs/2410.19560.

Denise Moerel, Tijl Grootswagers, Amanda Robinson, Sophia Shatek, Alexandra Woolgar, Thomas Carlson, and Anina Rich. The time-course of feature-based attention effects dissociated from temporal expectation and target-related processes, 2022. URL https://openneuro.org/datasets/ds004043/versions/1.1.0.

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*, KDD '24, pp. 5544–5555. Association for Computing Machinery, 2024. ISBN 979-8-4007-0490-1. doi: 10.1145/3637528.3671600. URL https://dl.acm.org/doi/10.1145/3637528.3671600.

Wajid Mumtaz. MDD Patients and Healthy Controls EEG Data (New), 2016. URL https://figshare.com/articles/dataset/EEG_Data_New/4244171.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196.

Patrick Ofner, Andreas Schwarz, Joana Pereira, and Gernot R Müller-Putz. Upper limb movements can be decoded from the time-domain of low-frequency eeg. *PloS one*, 12(8):e0182578, 2017.

Julie Onton and Scott Makeig. Imagined emotion study, 2022. URL https://openneuro.org/datasets/ds003004/versions/1.1.1.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=ZeFMtRBy4Z.

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 38 (NeurIPS 2025)*, 2025b. doi: 10.48550/arXiv.2510.21585. URL https://arxiv.org/abs/2510.21585.

Maria J. Ribeiro and Miguel Castelo-Branco. Eeg, ecg and pupil data from young and older adults: rest and auditory cued reaction time tasks, 2021. URL https://openneuro.org/datasets/ds003690/versions/1.0.0.

Alexander P. Rockhill, Nicko Jackson, Jobi George, Adam Aron, and Nicole C. Swann. Uc san diego resting state eeg data from patients with parkinson's disease, 2022. URL https://openneuro.org/datasets/ds002778/versions/1.0.5.

Joseph H. Rudoler, Matthew R. Dougherty, Brandon S. Katerman, James P. Bruska, Woohyeuk Chang, David J. Halpern, Nicholas B. Diamond, and Michael J. Kahana. Spatial memory and non-invasive closed-loop stimulus timing, 2023. URL https://openneuro.org/datasets/ds004706/versions/1.0.0.

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, 2017a.

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. 38(11):5391–5420, 2017b. ISSN 1065-9471, 1097-0193. doi: 10.1002/hbm.23730. URL https://onlinelibrary.wiley.com/doi/10.1002/hbm.23730.

Tong Shan, Madeline S. Cappelloni, and Ross K. Maddox. Music and speech elicit similar subcortical responses in human listeners, 2024. URL https://openneuro.org/datasets/ds004356/versions/2.2.1.

Sophia M. Shatek, Amanda K. Robinson, Tijl Grootswagers, and Thomas A. Carlson. Capacity for movement is an organisational principle in object representations: Eeg data from experiment 1, 2023a. URL https://openneuro.org/datasets/ds003885/versions/1.0.8.

Sophia M. Shatek, Amanda K. Robinson, Tijl Grootswagers, and Thomas A. Carlson. Capacity for movement is an organisational principle in object representations: Eeg data from experiment 2, 2023b. URL https://openneuro.org/datasets/ds003887/versions/1.2.3.

Noam Shazeer. GLU Variants Improve Transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 1, 2025a. URL https://openneuro.org/datasets/ds005505/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 2, 2025b. URL https://openneuro.org/datasets/ds005506/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 3, 2025c. URL https://openneuro.org/datasets/ds005507/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 4, 2025d. URL https://openneuro.org/datasets/ds005508/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 5, 2025e. URL https://openneuro.org/datasets/ds005509/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 6, 2025f. URL https://openneuro.org/datasets/ds005510/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 7, 2025g. URL https://openneuro.org/datasets/ds005511/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 8, 2025h. URL https://openneuro.org/datasets/ds005512/versions/1.0.1.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. Healthy brain network (hbn) eeg - release 9, 2025i. URL https://openneuro.org/datasets/ds005514/versions/1.0.1.

Ali Hossam Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, 2009. URL https://hdl.handle.net/1721.1/54669.

Elizabeth M. Siefert, Sindhuja Uppuluri, Jianing Mu, Marlie C. Tandoc, James W. Antony, and Anna C. Schapiro. Siefert2024, 2024. URL https://openneuro.org/datasets/ds005121/versions/1.0.2.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. 31:710–719, 2023. ISSN 1534-4320, 1558-0210. doi: 10.1109/TNSRE.2022.3230250. URL https://ieeexplore.ieee.org/document/9991178/.

Jan Sosulski and Michael Tangermann. Spatial filters for auditory evoked potentials transfer between different experimental conditions. In *GBCIC*, 2019.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. 2021. doi: 10.48550/ARXIV.2106.10270. URL https://arxiv.org/abs/2106.10270.

Tasos Papastylianou, Rodrigo Ramele, Luca Citi, Caterina Cinel, and Riccardo Poli. Pes - pandemic emergency scenario, 2023. URL https://openneuro.org/datasets/ds004477/versions/1.0.2.

Jack E. Taylor, Rasmus Sinn, Cosimo Iaia, and Christian J. Fiebach. Alphabetic decision task, 2024. URL https://openneuro.org/datasets/ds005594/versions/1.0.3.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning?, 2020. URL https://arxiv.org/abs/2005.10243.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. URL https://arxiv.org/abs/2302.13971.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint Embedding vs Reconstruction: Provable Benefits of Latent Space Prediction for Self Supervised Learning, 2025. URL https://arxiv.org/abs/2505.12477.

J. Veillette, S. Heald, B. Wittenbrink, and H. Nusbaum. eeg-neuroforecasting, 2022. URL https://openneuro.org/datasets/ds004284/versions/1.0.0.

John Veillette, Pedro Lopes, and Howard Nusbaum. Illusion of agency over electrically-actuated movements, 2023. URL https://openneuro.org/datasets/ds004561/versions/1.0.0.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings, 2023. URL https://arxiv.org/abs/2302.14367.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 39249–39280. Neural Information Processing Systems Foundation, Inc., 2024. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding, 2025. URL http://arxiv.org/abs/2412.07236.

Weilong Li and Jiaxin Zhao. Perceiveimagine, 2024. URL https://openneuro.org/datasets/ds005697/versions/1.0.2.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 78240–78260. Neural Information Processing Systems Foundation, Inc., 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html.

Liuyin Yang and Marc M. Van Hulle. Learning Robust EEG Representations with a Large Spatiotemporal Transformer as a Foundation Model. 2024. URL https://openreview.net/forum?id=V5Zn0VVvBE.

Weibo Yi, Shuang Qiu, Kun Wang, Hongzhi Qi, Lixin Zhang, Peng Zhou, Feng He, and Dong Ming. Evaluation of eeg oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PloS one*, 9(12):e114853, 2014.

Yulin Wang, Wei Duan, Lihong Ding, Debo Dong, and Xu Lei. A test-retest resting and cognitive state eeg dataset, 2022. URL https://openneuro.org/datasets/ds004148/versions/1.0.1.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1204–1213. IEEE, 2022. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01179. URL https://ieeexplore.ieee.org/document/9880094/.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization, 2019. URL https://arxiv.org/abs/1910.07467.

Xiangdong Zhang, Shaofeng Zhang, and Junchi Yan. PCP-MAE: Learning to predict centers for point masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.48550/arXiv.2408.08753. URL https://arxiv.org/abs/2408.08753. Spotlight.

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

Bangyan Zhou, Xiaopei Wu, Zhao Lv, Lei Zhang, and Xiaojin Guo. A fully automated trial selection method for optimization of motor imagery based brain-computer interface. *PloS one*, 11(9): e0162657, 2016.

Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain Foundation Models: A Survey on Advancements in Neural Signal Processing and Brain Discovery, 2025a. URL http://arxiv.org/abs/2503.00580.

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, 2025b. URL http://arxiv.org/abs/2506.23075.

Natalia Zhozhikashvili, Maria Protopova, Tatiana Shkurenko, Marie Arsalidou, Ilya Zakharov, Boris Kotchoubey, Sergey Malykh, and Yuri Pavlov. Sternberg, 2024. URL https://openneuro.org/datasets/ds005095/versions/1.0.1.

Zoltan Kekecs, Kyra Girán, Vanda Vizkievicz, Anna Lutoskin, and Yeganeh Farahzadi. The effects of sham hypnosis techniques, 2026. URL https://openneuro.org/datasets/ds004572/versions/1.3.2.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during Mental Arithmetic Task Performance. 4(1):14, 2019. ISSN 2306-5729. doi: 10.3390/data4010014. URL https://www.mdpi.com/2306-5729/4/1/14.

A FRAMEWORKS DETAILS

A.1 DETAILED REVE-SMALL ARCHITECTURE

The REVE-Small backbone follows the REVE-Small architecture of Ouahidi et al. (2025b) with minor simplifications that we keep across all pre-training configurations. The pipeline goes as follows: a shared linear patch embedding turns each EEG channel into a sequence of overlapping temporal patches of length 200 samples (1 s at 200 Hz) with 20-sample (0.1 s) overlap, each linearly projected to $d = 512$ dimensions; the same embedding is reused by the student encoder, the EMA teacher (JEPA only) and the predictor. A non-learnable additive 4D sinusoidal positional encoder encodes the channel position (x, y, z) on $d_{\text{spat}} = 384$ dimensions (128 per coordinate) and the patch index on $d_{\text{time}} = 128$ dimensions, computed at the post-embedding sampling rate $\nu_{\text{feat}} = 200/180 = 10/9$ Hz; its parameters are buffers, not weights, so it does not appear in the parameter count below. The contextual encoder is a 4-layer transformer with 8 heads, $d_{\text{model}} = 512$, GEGLU (Shazeer, 2020) feed-forward blocks at the LLaMA $8/3$ expansion ratio ($d_{\text{ff}} = \lfloor 8 \cdot 512/3 \rfloor = 1365$), pre-norm RMSNorm (Zhang & Sennrich, 2019), no bias on linear layers, and dropout 0. The MAE decoder and the JEPA predictor share a single class: a 2-layer transformer decoder with the same width, head count, FFN geometry, and norm; each layer applies pre-norm self-attention over the mask tokens, cross-attention to the encoder output (memory), and a GEGLU feed-forward block. The decoder query is a single learned d-dimensional mask token (no separate positional embedding for the mask target: the additive PE from the encoder is reused). A final $d_{\text{model}} \to p$ linear layer maps each output token to the prediction target, with $p = 200$ for MAE (raw patch in input space) and $p = 512$ for JEPA (latent embedding of the same dimension as the encoder).

Table 1 reports the per-component parameter count, derived directly from the layer specifications above and from the `torch.nn` primitives we use (notably: in our GEGLU FFN, the input projection maps $d_{\text{model}} \to 2\, d_{\text{ff}}$ so that the gate-and-multiply activation can halve its output back to d_{ff}; RMSNorm contributes one learnable scale per channel; `bias=False` on every linear layer). The MAE student totals ~ 21.2M parameters and the JEPA student ~ 21.3M parameters; the JEPA EMA teacher is a frozen copy of the encoder only (~ 12.6M parameters held in fp32 memory but excluded from the optimiser, see Appendix A.7); the patch embedding is shared with the student rather than duplicated.

Table 1: Per-component parameter count of the REVE-Small backbone, derived from the configuration files. "†" marks components that exist only for one of the two frameworks; "‡" marks the JEPA-only EMA teacher, which is held in memory but not optimised. Numbers assume `bias=False` on every linear layer and the GEGLU expansion described in the prose.

Component	# params
Linear patch embedding ($200 \to 512$, with bias)	102,912
Encoder layer (single, $\times 4$):	3,146,240
Self-attention W_{QKV}, W_O	1,048,576
GEGLU FFN (gate + value, output projections)	2,096,640
$2\times$ RMSNorm scales	1,024
Encoder transformer total (4 layers)	12,584,960
Decoder layer (single, $\times 2$):	4,195,328
Self-attention W_{QKV}, W_O	1,048,576
Cross-attention W_{QKV}, W_O	1,048,576
GEGLU FFN (gate + value, output projections)	2,096,640
$3\times$ RMSNorm scales	1,536
Decoder transformer total (2 layers)	8,390,656
Learned mask token ($d = 512$)	512
Output projection †, MAE ($512 \to 200$)	102,400
Output projection †, JEPA ($512 \to 512$)	262,144
MAE student total	**21,181,440**
JEPA student total	**21,341,184**
JEPA EMA teacher‡ (encoder only; patch embedding shared with student)	12,584,960

A.2 TRAINING HYPERPARAMETERS

The same optimizer, learning-rate schedule, batch size, precision, and gradient clipping are used for every one of the 58 pre-training configurations of the core grid; only the framework loss and the masking strategy change between configurations. The values in Table 2 are copied verbatim from the launch scripts and traced to the corresponding default in the codebase when the launch script does not override them.

Table 2: Training hyperparameters shared by all 58 pre-training configurations of the core grid.

Hyperparameter	Value
Optimizer	AdamW
(β_1, β_2)	(0.9, 0.999) (PyTorch default)
weight decay	0.01
Peak learning rate	2.4×10^{-4}
Final learning rate η_{min}	10^{-6}
Warmup schedule	linear, factor $10^{-6} \to 1$
Warmup steps	3,080 (= 1 epoch)
LR schedule (after warmup)	cosine, $T_{\text{max}} = T_0 -$ warmup
T_0	30,800 (= total optimizer steps)
Total epochs	10
Steps per epoch	3,080 ($= \lceil 3{,}695{,}009/1200 \rceil$)
Total optimizer steps	30,800
Per-GPU batch size	600
Number of GPUs (DDP)	2 (1 H100 node)
Effective batch size	1,200 windows / step
Mixed-precision dtype	`bf16-mixed`
Gradient clipping	$\|g\|_2 \leq 3.0$
DDP strategy	`ddp_find_unused_parameters_true`
Random seed	42

The cosine schedule is wrapped in a `SequentialLR` that switches from a 1-epoch linear warm-up to `torch.optim.lr_scheduler.CosineAnnealingLR` with $T_{\text{max}} = T_0 -$ warmup_steps $=$ 27,720 at the warm-up milestone. We deliberately set T_0 *exactly* equal to the total optimizer step count: a smaller T_0 would let `CosineAnnealingLR` pass its minimum and the learning rate would climb back up over the last steps, which we observed to wreck the final weights in earlier sweeps.

A.3 PRE-TRAINING COMPUTE COST

All 58 grid runs share the same 10-epoch budget and the same hardware, two H100 GPUs on a single node (Table 2). We report the cost of each run in H100-hours, *i.e.* its measured wall-clock duration multiplied by the number of GPUs. Table 3 summarises these costs per framework: JEPA runs cost about 2 H100-hours more than MAE runs, and the spread within each framework is small. For reference, Ouahidi et al. (2025b) report an estimated 260 A100-hours for the original REVE-Base pre-training. On our cluster, an H100-hour is billed $1.5\times$ an A100-hour; at this rate, the original REVE-Base pre-training costs about $12\times$ one of our runs. That comparison is not scale-matched, since REVE-Base has 69M parameters and was trained on the full 6 TB corpus whereas our models have 12.7M parameters and use the 4.4 TB open subset, and the convergence trajectory of REVE-Base is not public; the gap in compute therefore cannot be attributed to the mask geometry alone.

Table 3: Pre-training cost of the 58 grid runs, in H100-hours (two GPUs per run; n runs per group).

Framework	n	Mean	Median	SD	Min–Max
JEPA	29	14.91	15.28	0.78	14.00–16.29
MAE	29	12.70	12.65	0.25	12.62–13.98
Overall	58	13.81	13.99	1.26	12.62–16.29

Algorithm 1 Spatio-temporal block-mask sampling. Inputs are the (L, r) axis values, the channel positions $\{p_c\} \subset \mathbb{R}^3$, the optional channel-padding mask, the target mask ratio $\rho^\star$, the scalp surface $A_{\text{scalp}} = 4\pi(0.1)^2 \cdot 3/4$ ($3/4$ of a 10 cm sphere), and the patch-grid size (N_c, N_t).

function $f_{\text{SP}}(r, N_c)$
  if $r = 0$ **then**
    return $1/N_c$ ▷ single-channel block (special case)
  else if $r = \infty$ **then**
    return 1 ▷ full-spatial block (special case)
  else
    return $\pi r^2 / A_{\text{scalp}}$ ▷ small-cap surface fraction
  end if
end function
$f \leftarrow f_{\text{sp}}(r, N_c) \cdot L/N_t$ ▷ per-block cell coverage
if $f \geq 1$ **then**
  $K \leftarrow 1$
else
  $K \leftarrow \max\left(1, \left\lceil \frac{\log(1-\rho^\star)}{\log(1-f)} \right\rceil\right)$ ▷ inclusion-exclusion, see Eq. (2)
end if
$M \leftarrow$ all-zeros boolean tensor of shape (N_c, N_t)
for $i = 1, \ldots, K$ **do**
  $c_i \leftarrow$ sample uniformly from the non-padding channels
  $t_i \leftarrow$ sample uniformly from $\{1, \ldots, \max(1, N_t - L + 1)\}$
  for $c = 1, \ldots, N_c$ **do**
    if $r = 0$ **then**
      in $\leftarrow [c = c_i]$ ▷ distance ≤ 0 matches only c_i
    else if $\|p_c - p_{c_i}\|_2 \leq r$ **then**
      in $\leftarrow$ True
    else
      in $\leftarrow$ False
    end if
    if in **then**
      for $t = t_i, \ldots, t_i + L - 1$ **do**
        $M_{c,t} \leftarrow$ True
      end for
    end if
  end for
end for
return M ▷ $M_{c,t} = 1$ iff cell (c, t) is masked

A.4 Mask sampling algorithm

For each pre-training configuration the masking strategy is fully specified by the pair (L, r) where $L \in \mathbb{N}_{>0}$ is the temporal block length in patch units and $r \in [0, \infty]$ is the spatial cap radius, together with the target mask ratio $\rho^\star = 0.55$ shared across all 58 configurations of the core grid. The mask sampler computes the number of independently sampled blocks K from the inclusion-exclusion formula and then samples the K blocks; Algorithm 1 states the procedure as it appears in the codebase.

A.5 Empirical validation of the block count K and mask fraction $\rho^\star$

We sanity-check the inclusion-exclusion formula for K (Equation 2) by counting, in simulation, the realised cell-level mask fraction it produces under our pre-training sampling pipeline. For each (L, r) pair we draw 1,000 batches of 64 examples and aggregate over the resulting 64,000 samples.

The number of blocks K given by (2) for each of the 29 cells is listed in Table 4; it spans more than two orders of magnitude, from $K = 2$ for the largest blocks to $K = 843$ for single-patch blocks.

Table 5 reports two empirical quantities per (L, r) cell: the per-block channel coverage (*ch/sphere*, the mean and standard deviation, over individual blocks, of the number of non-padding channels falling inside the spherical cap of one randomly drawn block) and the realised cell-level mask fraction (*total mask*, the mean and standard deviation, over the 64,000 samples, of the fraction of (c, t) tokens

Table 4: Number of blocks K sampled for an input x, for each combination radius-length (r, L), derived from the inclusion-exclusion formula (2) with the mask ratio $\rho^\star = 0.55$.

	Temporal block length L (patches)					
Spatial radius r	1	2	4	8	16	33
0.0 ($r =$ "one")	843	422	211	106	53	26
0.06 m	220	110	55	28	14	7
0.09 m	98	49	24	12	6	3
0.12 m	55	28	14	7	4	2
∞ ($r =$ "all")	26	13	7	3	2	–

Table 5: Empirical per-block channel coverage (*ch/sphere*) and realised total mask fraction (*total mask*) for every (L, r) cell, with K as in Table 4; mean $\pm$ std over $1{,}000 \times 64 = 64{,}000$ samples. Target mask fraction is $\rho^\star = 55\,\%$.

		Temporal block length L (patches)					
Spatial radius r	metric	1	2	4	8	16	33
0.0 ($r =$ "one")	ch / sphere	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	tot %	55.0 ± 1.1	54.8 ± 1.3	54.2 ± 1.8	53.1 ± 2.5	50.9 ± 3.4	55.8 ± 6.3
0.06 m	ch / sphere	3.08 ± 1.60	3.08 ± 1.60	3.09 ± 1.61	3.08 ± 1.60	3.09 ± 1.60	3.09 ± 1.60
	tot %	56.2 ± 4.1	56.0 ± 4.2	55.5 ± 4.5	54.7 ± 5.1	53.0 ± 5.8	60.3 ± 10.0
0.09 m	ch / sphere	6.09 ± 2.54	6.08 ± 2.52	6.10 ± 2.54	6.10 ± 2.54	6.09 ± 2.53	6.10 ± 2.54
	tot %	53.5 ± 3.4	53.4 ± 3.6	52.3 ± 4.2	51.6 ± 5.2	50.8 ± 6.3	59.0 ± 11.0
0.12 m	ch / sphere	10.53 ± 3.62	10.53 ± 3.63	10.52 ± 3.62	10.54 ± 3.63	10.51 ± 3.61	10.52 ± 3.63
	tot %	53.2 ± 3.3	53.8 ± 3.8	53.6 ± 4.7	53.4 ± 6.2	57.5 ± 7.5	70.0 ± 13.3
∞ ($r =$ "all")	ch / sphere	23.01 ± 6.01	23.01 ± 6.01	23.05 ± 6.03	23.04 ± 6.02	23.02 ± 6.02	–
	tot %	55.1 ± 5.1	55.4 ± 6.1	58.5 ± 8.2	54.6 ± 10.7	66.6 ± 12.8	–

flagged as masked after the K blocks are merged). The *ch/sphere* value is independent of L for fixed r, as expected. The realised mask fraction matches the $\rho^\star = 55\,\%$ target within a few points across most cells, with the largest deviations observed in the extreme regimes $r =$ "all" and $L = 33$.

Mask geometry: empirical vs. theoretical token distances. Figure 4 reports the distance from each masked token to its nearest unmasked token on the same channel (top row, varying L at $r =$ "one") and at the same time on a different channel (bottom row, varying r at $L = 1$), comparing the empirically realised distribution under our pre-training pipeline to the analytical single-block, infinite-density reference. The empirical and theoretical distributions track each other closely across the entire (L, r) grid, confirming that the inclusion-exclusion formula of Equation 2 produces the intended geometry on the actual REVE channel layouts: longer L pushes a masked patch further from the nearest unmasked patch *on the same channel* (so the network must look across channels to fill it in), and larger r pushes it further from the nearest unmasked patch *at the same time* (so the network must look across time). The cell $r =$ "all" is degenerate by construction: every channel is masked at masked time-points, so the spatial distance is undefined and only the temporal axis carries information about the masked tokens.

A.6 MAE DETAILS

The MAE framework reconstructs the *raw input patches* at masked positions. Concretely, the shared linear patch embedding returns both the d-dimensional embeddings (used as the encoder's local features) and the unfolded raw patches $\mathbf{P} \in \mathbb{R}^{B \times C \times T \times p}$ with $p = 200$ samples per patch; the predictor output $\hat{\mathbf{P}} \in \mathbb{R}^{B \times C \times T \times p}$ comes from the same 2-layer transformer decoder as in JEPA but with output dimension $p = 200$ (the input patch length) rather than $d = 512$. The objective is a masked MSE

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|\, p} \sum_{(c,t) \in \mathcal{M}} \|\hat{\mathbf{P}}_{c,t} - \mathbf{P}_{c,t}\|_2^2,$$

identical in form to the JEPA loss but with raw-patch targets; the target tensor is detached and zero-filled outside the target mask before the elementwise MSE. No per-patch target normalisation is

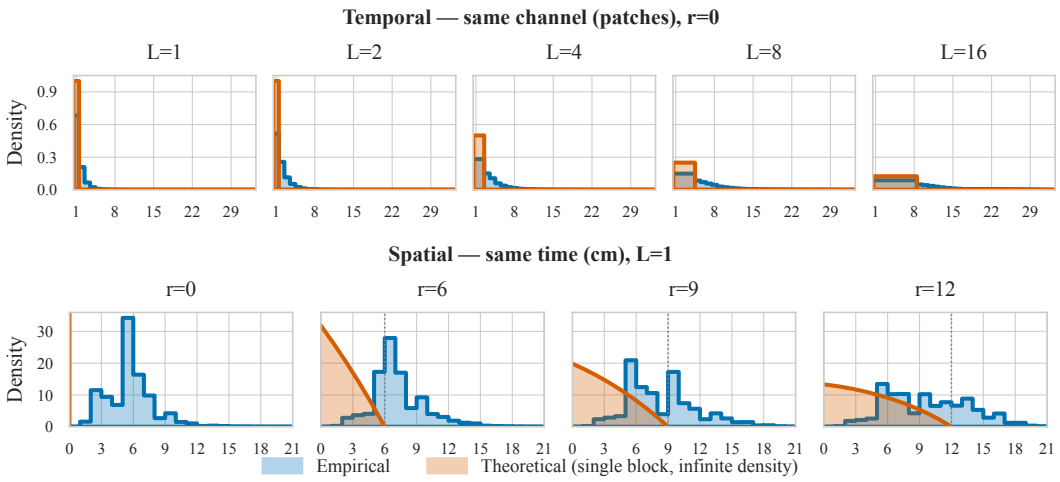


Figure 4: Distance from each masked token to its nearest unmasked token, empirical (blue) vs. single-block infinite-density theoretical (orange). *Top row:* temporal distance (in patches) on the same channel, sweeping L at $r =$ `"one"`. *Bottom row:* spatial distance (in centimeters) at the same time, sweeping r at $L = 1$. Empirical distributions are aggregated over 1,000 batches of 64 samples drawn from the REVE channel layouts.

applied at the loss level: the input window has already been scaled by the per-window robust scaler before patching, and we found in preliminary runs that the additional per-patch normalisation used in some computer-vision MAE pipelines was unnecessary at the REVE-Small scale. The variance and covariance regularisers exposed by the framework are unused in the 58 configurations of the core grid (their loss weights are set to 0), so the only term optimised is $\mathcal{L}_{\text{MAE}}$, preserving strict parity with the JEPA-no-reg configuration on the "no auxiliary loss" axis.

A.7 JEPA DETAILS

The JEPA framework regresses the predictor's output at masked positions onto the contextual features of an exponential-moving-average (EMA) teacher under a masked L_2 loss; in our *no-reg* configuration the variance and covariance regularisers of Bardes et al. (2021) are explicitly disabled by setting their weights to 0, so the only term optimised is the main prediction loss $\mathcal{L}_{\text{JEPA}} = \frac{1}{|\mathcal{M}|\,d} \sum_{(c,t)\in\mathcal{M}} \|\hat{z}_{c,t} - \bar{z}_{c,t}\|_2^2$, where $\mathcal{M} = \{(c,t) : M_{c,t} = 1 \text{ and } (c,t) \text{ is not padding}\}$, $\hat{z}_{c,t} \in \mathbb{R}^d$ is the predictor output, and $\bar{z}_{c,t} \in \mathbb{R}^d$ is the EMA-teacher contextual feature at the same position; the loss weights at masked, non-padded cells contribute equally regardless of the per-block density. The EMA teacher is built once at framework construction time as a `deepcopy` of the encoder reduced to its maskless form — it owns its own copy of the additive positional encoder and the 4-layer transformer but holds no masker, and the patch embedding is shared with the student (its output is cached during the student's forward and reused as the teacher's input) rather than duplicated. The teacher is kept in `eval()` mode throughout training; it receives the *unmasked* input, so its forward pass costs an additional $\sim$12.6M-parameter encoder evaluation per step (Table 1). Targets are taken from the last encoder layer only — no multi-layer averaging of teacher features.

The EMA decay schedule interpolates linearly from $\tau_{\text{start}} = 0.999$ at step 0 to $\tau_{\text{end}} = 1.0$ at step 30,000, after which it is held at $\tau_{\text{end}} = 1.0$ (i.e. the teacher is frozen for the last $\sim 2.6\,\%$ of training):

$$\tau(s) = \begin{cases} \tau_{\text{end}} - (\tau_{\text{end}} - \tau_{\text{start}}) \left(1 - \dfrac{s}{S_{\text{anneal}}}\right), & s < S_{\text{anneal}} = 30{,}000, \\ \tau_{\text{end}} = 1.0, & s \geq S_{\text{anneal}}, \end{cases}$$

with $\tau_{\text{start}} = 0.999$ and $\tau_{\text{end}} = 1.0$, where τ is the teacher's keep weight (so the per-step update on every trainable teacher parameter is $\theta_{\text{T}} \leftarrow \tau\,\theta_{\text{T}} + (1-\tau)\,\theta_{\text{S}}$, performed in `fp32` regardless of the training precision; non-trainable teacher buffers are copied directly from the student each step). The teacher update is performed *per training step*, immediately after the optimiser step; it is not called during validation or test.

Collapse diagnostics for JEPA. We monitor JEPA runs for representational collapse throughout pre-training, but do not add any corrective loss. The three failure modes we guard against are documented in the self-supervised learning literature: dimensional/variance collapse, addressed by the variance term of VICReg (Bardes et al., 2021); trivial-constant "entire collapse", which Mo & Tong (2024) show is not reliably prevented by the EMA teacher of I-JEPA (Assran et al., 2023) alone and motivates their addition of VICReg-style regularisation; and the post-collapse loss divergence that follows once the encoder outputs degenerate. A Lightning callback halts the run as soon as any of three detection conditions fires. First, the embedding standard deviation σ_z logged by the framework at every optimiser step, the actual standard deviation of the encoder outputs across the batch, must remain above $\epsilon = 10^{-2}$; if it falls below ϵ for 200 consecutive steps the run is flagged for variance collapse. This threshold is well below the unit-variance target $\gamma = 1$ used by the VICReg variance hinge (Bardes et al., 2021), so the rule fires only when the encoder has effectively lost its representational variance. Second, the main prediction loss exceeds 10^4, indicating post-collapse divergence of the kind reported by Mo & Tong (2024) when the EMA teacher fails to stabilise training. Third, the main prediction loss drops below 10^{-3} within the first 500 optimiser steps, which is the signature of the predictor latching onto a trivial constant target – the "entire collapse" mode of Mo & Tong (2024). Across all JEPA configurations reported in this paper, none of these conditions was ever triggered.

B Pre-training datasets

The open subset of the REVE pre-training corpus comprises the following datasets: (Ouahidi et al., 2025a; Rudoler et al., 2023; Makowski et al., 2023; Shan et al., 2024; Grootswagers et al., 2023b; Helbing et al., 2024; Shatek et al., 2023b; Moerel et al., 2022; Shatek et al., 2023a; Grootswagers et al., 2024; 2022; 2023a; Cordoba-Silva et al., 2023; Metwalli et al., 2025; Tasos Papastylianou et al., 2023; Esteban et al., 2024; Veillette et al., 2023; Haupt et al., 2024; Chacón & Wriessnegger, 2023; Zhozhikashvili et al., 2024; Delorme & Brandmeyer, 2024; Ribeiro & Castelo-Branco, 2021; Benjamin Lowe (Ben.Lowe@Mq.Edu.Au) et al., 2023; Delorme & Braboszcz, 2021; Hassall et al., 2024; Onton & Makeig, 2022; Daly et al., 2024; Hassall et al., 2022a; Aguado-Lopez et al., 2024; Hassall et al., 2022b;c; Cavanagh & Jackson, 2022; Bialas et al., 2023; Siefert et al., 2024; Hatlestad-Hall et al., 2022; Zoltan Kekecs et al., 2026; Rockhill et al., 2022; Gehrke et al., 2024; Araya et al., 2024; Yulin Wang et al., 2022; Chuqin Xiang et al., 2025; Delorme, 2021; Veillette et al., 2022; Kahana & Rudoler, 2023; Shirazi et al., 2025a;b;c;e;f;g;h;i;d; Weilong Li & Jiaxin Zhao, 2024; Bajwa1 et al., 2024; Taylor et al., 2024; Baykan & Schütz, 2024; Barachant, 2012; Cho et al., 2017; Lee et al., 2019a; Liu et al., 2024; Ofner et al., 2017; Schalk et al., 2004; Goldberger et al., 2000; Yi et al., 2014; Zhou et al., 2016; Schirrmeister et al., 2017a; Kalunga et al., 2015; Lee et al., 2019c; Korczowski et al., 2019c;d;a;b; Sosulski & Tangermann, 2019; Lee et al., 2019b; Amorim et al., 2023).

C Downstream evaluation details

We evaluate every model on the 12 OpenEEGBench datasets (Guetschel et al., 2026a). Foundation-model checkpoints are evaluated under a closed-form ridge linear probe on frozen features (Appendix C.2); the three non-foundation baselines (EEGNet, EEGConformer, ShallowFBCSPNet) are trained from scratch (Appendix C.3).

C.1 Downstream datasets

Datasets. The 12 OpenEEGBench downstream datasets are Arithmetic (Zyma et al., 2019; Goldberger et al., 2000), BCIC-IV-2a (Brunner et al., 2008), BCIC2020-3 (Jeong et al., 2022), PhysioNet-MI (Schalk et al., 2004; Goldberger et al., 2000), CHB-MIT (Shoeb, 2009; Goldberger et al., 2000), FACED (Chen et al., 2023), ISRUC-Sleep (Khalighi et al., 2016), MDD (Mumtaz, 2016), SEED-V (Liu et al., 2022), SEED-VIG (Zheng & Lu, 2017), TUAB (Obeid & Picone, 2016), and TUEV (Obeid & Picone, 2016).

Per-dataset class distributions. Table 6 reports the per-class window counts and percentages for each of the 12 OpenEEGBench downstream datasets.

Table 6: Per-class window counts and balance for each of the 12 OpenEEGBench downstream datasets, computed from the per-recording metadata of the Hugging Face Zarr mirror.

Dataset	n_{windows}	classes	Distribution (%)
`arithmetic_zyma2019`	1,707	2	25.3 / 74.7
`bcic2020-3`	6,000	5	20 / 20 / 20 / 20 / 20
`bcic2a`	5,184	4	25 / 25 / 25 / 25
`chbmit`	37,046	2	**2.5** / **97.5**
`faced`	19,217	8	13 / 14 / 13 / 12 / — / 12 / 17 / 7.7 / 12
`isruc-sleep`	98,480	5	21 / 14 / 33 / 19 / 14
`mdd_mumtaz2016`	14,743	2	47 / 53
`physionet`	9,688	4	25 / 25 / 25 / 25
`seed-v`	117,744	5	17 / 26 / 20 / 17 / 20
`seed-vig`	20,355	**regression**	range [0.022, 1.0], mean 0.467
`tuab`	42,136	2	49.7 / 50.3
`tuev`	3,330	6	**2.4** / 19 / 23 / 8.1 / 16 / **31**

Signal scaling. For every foundation-model checkpoint, the per-window scaling applied at downstream evaluation time matches the one used during its own pre-training. For the 58 checkpoints of the core grid, a per-window robust scaler divides every channel by the median, taken across channels, of the per-channel standard deviations; values exceeding ± 15 standard deviations are finally clipped. For the public REVE baseline, we apply per-window the standardisation described in the REVE paper (Ouahidi et al., 2025b): z-score followed by clipping at ± 15 standard deviations. The three non-foundation baselines (EEGNet, EEGConformer, ShallowFBCSPNet) are trained from scratch on each downstream task and receive the raw signals in microvolts, with no per-window normalisation.

C.2 Downstream pipeline of foundation models

For every (checkpoint, dataset) pair, the downstream pipeline of a foundation model operates in three stages: frozen feature extraction, Gaussian random projection, and a closed-form ridge solve. There is no stochastic gradient descent.

Frozen feature extractor. For every input window, the encoder is operated on full, unmasked inputs as a frozen feature extractor. It outputs the post-context tokens $\mathbf{Z} \in \mathbb{R}^{B \times C \times T \times d}$ produced by its last encoder layer, where B is the batch size, C the number of channels of the downstream dataset, T the number of temporal patches per window, and $d = 512$ the latent dimension. No trainable parameter is updated by the downstream protocol.

Per-dataset feature dimensionality before the random projection. The number of features D_{orig} that the ridge probe sees per example, before the Gaussian random projection described next, is determined entirely by the encoder's pre-head tensor shape: both the REVE baseline and our encoder emit a $(B, n_{\text{chans}}, n_{\text{patches}}, 512)$ tensor that is flattened to a single vector per example. With the shared tokenisation parameters — patch length of 200 samples, 20-sample overlap (so a stride of 180 samples), and embedding dimension 512 — the number of temporal patches is $n_{\text{patches}} = \lfloor (n_{\text{times}} - 200)/180 \rfloor + 1$ and the feature count is

$$D_{\text{orig}} \;=\; n_{\text{chans}} \cdot n_{\text{patches}} \cdot 512. \tag{3}$$

Table 7 reports D_{orig} on each of the 12 OpenEEGBench datasets, computed from the per-dataset window shapes $(n_{\text{chans}}, n_{\text{times}})$.

Random projection and seeds. The flattened output tensor $\mathbf{Z}$ of dimension D_{orig} is then projected into a lower-dimensional space by Gaussian random projection $\mathbb{R}^{D_{\text{orig}}} \to \mathbb{R}^{D_{\text{proj}}}$ with $D_{\text{proj}} = 5{,}000$; its entries are drawn from $\mathcal{N}(0,\, 1/D_{\text{proj}})$. This random projection is the only stochastic component of the foundation-model downstream pipeline (the closed-form ridge solve being deterministic given the projection); each seed therefore corresponds to an independent draw of the projection. We use 5 seeds for the final-epoch (epoch 10) checkpoint of every (framework, mask) configuration, so that the headline numbers reported in the main paper are summarised as a mean over 5 seeds. We use only 3

Table 7: Per-dataset feature dimension D_{orig} before the Gaussian random projection, computed from Eq. (3). Columns n_{chans} and n_{times} are read from the Hugging Face Zarr metadata of each dataset; n_{patches} follows from the shared tokenisation. The last column flags whether the projection step activates (always true here, since $D_{\text{orig}} > 5{,}000$ on every cell).

Dataset	n_{chans}	n_{times}	n_{patches}	D_{orig}	projected?
`arithmetic_zyma2019`	19	1,000	5	48,640	yes
`bcic2a`	22	800	4	45,056	yes
`bcic2020-3`	64	600	3	98,304	yes
`physionet`	64	600	3	98,304	yes
`chbmit`	17	2,000	11	95,744	yes
`faced`	26	2,000	11	146,432	yes
`isruc-sleep`	6	6,000	33	101,376	yes
`mdd_mumtaz2016`	19	1,000	5	48,640	yes
`seed-v`	62	200	1	31,744	yes
`seed-vig`	17	1,600	8	69,632	yes
`tuab`	16	2,000	11	90,112	yes
`tuev`	21	1,000	5	53,760	yes

seeds for the intermediate-epoch checkpoints used in the training-trajectory plots to save compute resources. This choice is reasonable because only the final-epoch checkpoints (5 seeds) are used in statistical comparisons; the training-trajectory curves (3 seeds) are meant to be used for qualitative comparison only.

Closed-form ridge solve. The ridge solver accumulates sufficient statistics ($\mathbf{Z}_{\text{proj}}^\top \mathbf{Z}_{\text{proj}}$, $\mathbf{Z}_{\text{proj}}^\top Y$, mean and second-moment sums) in double precision over the training split. The centred $D_{\text{proj}} \times D_{\text{proj}}$ correlation matrix is eigendecomposed once, and a fixed log-spaced grid of 17 values $\lambda \in \{10^{-8}, 10^{-7}, \ldots, 10^{8}\}$ is swept in the eigenbasis to pick the regulariser that maximises the validation metric (balanced accuracy for classification datasets, R^2 for regression datasets); ties are broken in favour of the largest λ.

C.3 Downstream pipeline of non-foundation models

The three non-foundation baselines (EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2023), and ShallowFBCSPNet (Schirrmeister et al., 2017b)) are trained from scratch on each downstream dataset using the OpenEEGBench per-task training protocol.

End-to-end training. The three non-foundation baselines are optimised with AdamW at a peak learning rate of 10^{-3} and a batch size of 64, under a cosine learning-rate schedule decaying from the peak rate down to $\eta_{\min} = 10^{-6}$ over the full training duration. Each model is trained for at most 30 epochs with early stopping on the validation loss (patience 10), and gradients are clipped to a norm of 1.0. The same hyperparameters are used across all 12 downstream datasets and the three architectures, with no per-dataset overrides.

Seeds. Each seed independently re-initialises the backbone weights and the linear-head weights, re-shuffles the training batches, so the reported variances reflect both initialisation and training-time stochasticity.

C.4 Cross-dataset score normalisation

The 12 OpenEEGBench datasets are scored with different metrics (balanced accuracy for the 11 classification tasks, R^2 for the `seed-vig` regression task) and on different absolute scales, so the raw scores cannot be averaged directly across datasets. To report a single cross-dataset summary (the *Average (norm.)* row of Figure 11 and the y-axis of Figure 12), we linearly map the raw score s on dataset d to a dataset-normalised score

$$\tilde{s}^d = \frac{s - s^d_{\min}}{s^d_{\max} - s^d_{\min}}, \tag{4}$$

Table 8: Per-dataset normalisation constants $s^d_{\min}$ and $s^d_{\max}$ used in Equation 4. They are the minimum and maximum, across the 58 v9 sweep configurations, of the seed-averaged score on each dataset (balanced accuracy or R^2). The negative values for `seed-vig` are an instance of the negative-R^2 phenomenon discussed in Appendix C.5.

Dataset	min	max
arithmetic_zyma2019	0.595	0.778
bcic2020-3	0.231	0.294
bcic2a	0.263	0.475
chbmit	0.771	0.922
faced	0.158	0.332
isruc-sleep	0.497	0.720
mdd_mumtaz2016	0.698	0.864
physionet	0.326	0.585
seed-v	0.263	0.318
seed-vig	-0.848	-0.082
tuab	0.706	0.824
tuev	0.792	0.949

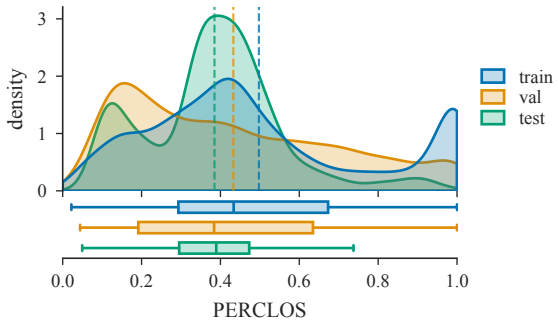


Figure 5: Per-window PERCLOS distribution on the OpenEEGBench `seed-vig` split (subject-based: train=$\{1, 2, 3, 10, \ldots, 21\}$, val=$\{4, 5\}$, test=$\{6, 7, 8, 9\}$). *Top:* Kernel-density estimate per split; the dashed verticals mark the per-split means. *Bottom:* matching horizontal box plots (medians, IQR, whiskers at $1.5\times$IQR; outliers hidden). The three splits differ both in location (means drift monotonically from train to test) and in shape: train and val are broad and have a heavy right tail near PERCLOS≈ 1, whereas test is unimodal, sharply concentrated around the central range, and lacks the high-PERCLOS tail entirely. This is what the closed-form ridge probe cannot recover from.

where $s^d_{\min}$ and $s^d_{\max}$ are the minimum and maximum, across the 58 sweep configurations (29 mask cells $\times$ 2 frameworks), of the seed-averaged score on dataset d. By construction, the worst sweep configuration of each dataset maps to $\tilde{s}^d = 0$ and the best to $\tilde{s}^d = 1$, so the resulting quantity is comparable across datasets and the cross-dataset average weights every dataset uniformly. Table 8 reports the constants $s^d_{\min}$ and $s^d_{\max}$ used throughout the paper.

C.5 SEED-VIG: NEGATIVE R^2 UNDER RIDGE PROBING

`seed-vig` is the only regression dataset in OpenEEGBench: the target is the per-window PERCLOS vigilance score in $[0, 1]$, and the test metric is `sklearn.metrics.r2_score`. Across all our checkpoints, ridge probing on `seed-vig` returns a *negative* R^2 around -0.35. This is not a bug: it is a structural consequence of the predefined subject split combined with the closed-form ridge solver, and the score still ranks models meaningfully above a constant baseline.

The distribution shift across the OpenEEGBench split is shown in Figure 5 and quantified in Table 9.The test distribution is much narrower than either train or val and is missing the heavy right tail near 1 that both training-side splits carry.

Table 9: Per-window PERCLOS statistics for the OpenEEGBench `seed-vig` split. The four test subjects happen to be both less drowsy on average and more uniform than the training subjects, which creates the negative-R^2 floor described in the text.

Split	n_{windows}	mean	std	range
train	13,275	0.498	0.283	[0.022, 1.000]
val	3,540	0.433	0.269	[0.044, 1.000]
test	3,540	0.385	0.175	[0.050, 1.000]

Table 10: Test R^2 on `seed-vig` for the train-mean dummy predictor and for the frozen REVE baseline under the OpenEEGBench ridge probe (5 random-projection seeds). The dummy score is the structural floor implied by the subject split visible in Figure 5; the ridge probe sits above it, indicating that the frozen features carry information about PERCLOS even though the metric stays negative.

Predictor	test R^2
`DummyRegressor` (predicts $\bar{y}_{\text{train}} = 0.498$)	-0.413
REVE baseline, ridge probe (mean over 5 seeds)	-0.355 ± 0.016
Gap above the dummy floor	$+0.058$

Both pathologies hurt the ridge probe in the same direction. The closed-form solver fixes its bias to $b = \bar{y}_{\text{train}} - W^\top \bar{h}_{\text{train}}$, so for any feature mapping with limited predictive power on the new subjects the predictions concentrate near $\bar{y}_{\text{train}} = 0.498$, i.e. around the right-most dashed line in Figure 5, while the test distribution is centred on the left-most one. The mean mismatch alone contributes

$$\frac{(\bar{y}_{\text{train}} - \bar{y}_{\text{test}})^2}{\text{Var}(y_{\text{test}})} = \frac{0.113^2}{0.175^2} \approx 0.413$$

to $1 - R^2$, setting a hard floor of $R^2 \approx -0.413$ for any near-constant predictor. Selecting λ on val cannot rescue this: the val box in Figure 5 has roughly the same width as train and only a small mean offset, so the val-side penalty is mild and the probe receives no signal that the test box has collapsed to roughly a third of its width.

To make this floor concrete we report the score of two reference predictors in Table 10: a `DummyRegressor(strategy='mean')` that always returns $\bar{y}_{\text{train}}$, and the REVE ridge-probe baseline (frozen, ridge probe with 5 projection seeds).

The frozen REVE features therefore lift the score by ≈ 0.06 R^2 above the constant-prediction floor, which is the relevant quantity when comparing checkpoints on this dataset. We report the raw R^2 throughout the paper for consistency with the rest of OpenEEGBench, but the reader should keep in mind that on `seed-vig` the informative axis is the gap to the -0.413 floor rather than the sign of the score itself.

D Spatial vs. temporal redundancy of the EEG signal

The two axes along which we mask, channels and time, are not interchangeable in the underlying signal. Figure 6 reports the squared Pearson correlation R^2 of the broadband EEG between (*left*) the same channel at two different time points and (*right*) two different channels at the same time point, measured on a random subsample of REVE windows. Within-channel autocorrelation drops sharply with the time-lag, falling below $R^2 \approx 0.2$ within ~ 2 s; between-channel correlation, in contrast, decays much more slowly with electrode distance and remains non-negligible across the entire scalp. A 1 s temporal patch of a given channel can therefore be predicted reasonably well from its own neighbouring patches alone, whereas channel-wise spatial structure is genuinely shared between channels and is harder to recover from local temporal context. This anchors the qualitative interpretation of the masking grid used in Section 4.

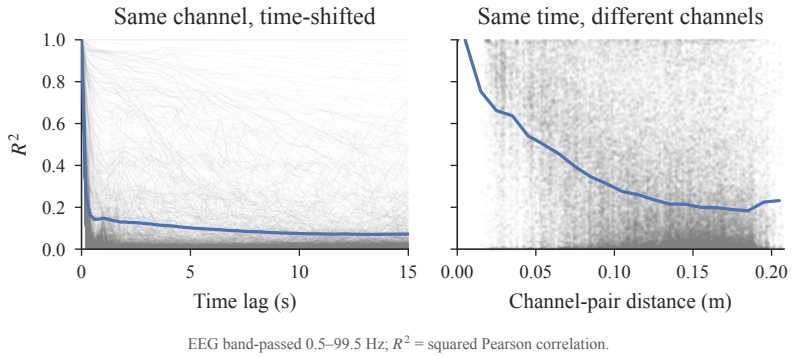


Figure 6: Squared Pearson correlation R^2 of the broadband EEG on a random subsample of REVE windows. *Left:* within-channel autocorrelation as a function of the time-lag (grey lines: per-window curves; coloured line: average across windows). *Right:* between-channel correlation as a function of the Euclidean inter-electrode distance (grey points: per-pair R^2; coloured line: distance-binned average). Temporal redundancy decays much faster than spatial redundancy.

E BIAS-INFLATION COLLAPSE

This appendix gives the full mechanistic analysis of *bias-inflation collapse*, the JEPA-only failure mode at $r = \texttt{"all"}$ described in Finding 3 and which, to our knowledge, has not been documented in the prior literature. We characterise the phenomenon geometrically (Appendix E.1), document its empirical signature (Appendix E.2), explain why it is the optimum of the JEPA loss at $r = \texttt{"all"}$ but not at smaller radii nor under MAE (Appendix E.4), show why none of the standard collapse detectors fire on it (Appendix E.5), discuss why an in-training detector for this regime is non-trivial in the multi-montage pre-training setting (Appendix E.6), list mechanistic hypotheses we ruled out during the investigation (Appendix E.7), and situate the phenomenon within the broader literature on self-supervised collapse (Appendix E.8).

E.1 PHENOMENON AND NOTATION

Let the encoder produce, at the end of a forward pass, a feature tensor $z(n, c, t) \in \mathbb{R}^d$ for every window n, non-padded channel c, and time-patch t. We decompose every token into its window-mean component and its input-driven residual:

$$z(n, c, t) \;=\; \mu(c, t) \;+\; \delta(n, c, t), \qquad \mu(c, t) \;:=\; \mathbb{E}_n\, z(n, c, t). \tag{5}$$

We track two scalars summarising this decomposition: the bias magnitude $\|\mu\|$, defined as the average of $\|\mu(c, t)\|$ over (c, t), and the input-driven magnitude $\sigma_{\text{token}} := \mathbb{E}_{n,c,t}\, \|\delta(n, c, t)\|$. The main empirical claim is that, across pre-training of JEPA at $r = \texttt{"all"}$, $\|\mu\|$ inflates by a factor of 2–3 while σ_{token} collapses by a factor of 5–10, yielding tokens dominated by a fixed per-channel-position bias and carrying vanishing input-driven information. Healthy ($r \leq 12$ cm) JEPA runs and the MAE control at $r = \texttt{"all"}$ both keep $\|\mu\|$ approximately stable and σ_{token} growing.

E.2 EMPIRICAL SIGNATURE

We measure the components of Equation 5 on the data received by the probe at downstream evaluation (a single subject of physionet and bcic2a, 84 and 576 windows respectively, processed through 10 representative pre-trained checkpoints — the 5 JEPA r-sweep cells at $L = 2$, the 4 JEPA L-sweep cells at $r = \texttt{"all"}$, and the MAE $r = \texttt{"all"}$ control) at every pre-training epoch (epochs $1, \ldots, 10$). Four scalars summarise each (run, epoch) cell: $\|\mu\|$, σ_{token}, the dimensionless ratio $\sigma_{\text{token}}/\sigma_{\text{chan}}$ where $\sigma_{\text{chan}} := \mathbb{E}_{n,t}\, \|z(n, c, t) - \mathbb{E}_c z(n, c, t)\|$ measures the spread of z across channels at fixed (n, t), and the mean intra-channel cosine similarity $\overline{\cos_{\text{intra}}}$ averaged over (c, t) of unit-normalised tokens across windows.

The four trajectories tell a coherent story: only JEPA at $r = \texttt{"all"}$ exhibits the simultaneous bias inflation, residual collapse, ratio collapse, and asymptote of the cosine similarity at 1.0. Every

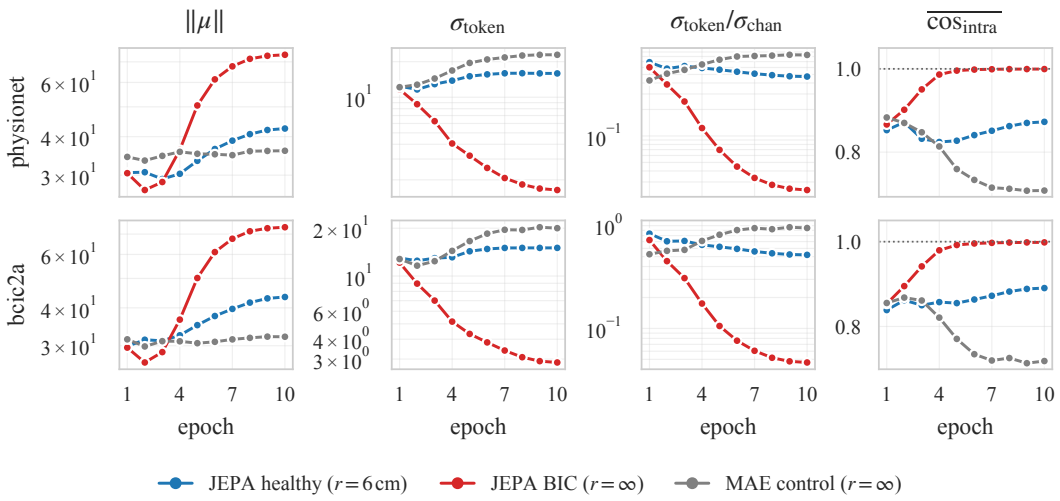


Figure 7: Bias-inflation collapse signature on OEB probe data, cross-window metrics computed on a single subject of physionet (top) and bcic2a (bottom) processed through each frozen checkpoint at every pre-training epoch. Only JEPA at $r =$ `"all"` shows the simultaneous bias inflation ($\|\mu\| \uparrow$), residual collapse ($\sigma_{\text{token}} \downarrow$), ratio drop and direction collapse ($\overline{\cos_{\text{intra}}} \to 1$) predicted by the lookup-table mechanism (Appendix E.4). Healthy JEPA ($r = 6$ cm) and the MAE control ($r =$ `"all"`) remain stable on all four metrics.

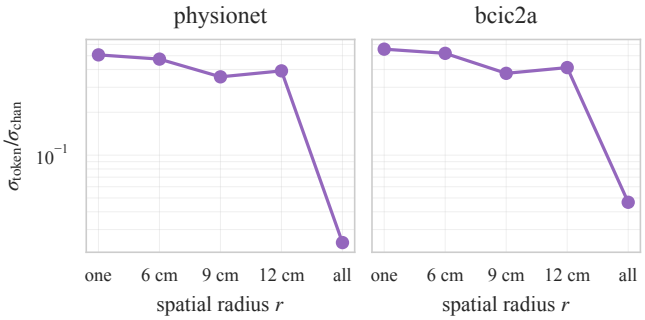


Figure 8: Phase transition in the spatial radius r at $L = 2$ and the final epoch. The ratio $\sigma_{\text{token}}/\sigma_{\text{chan}}$ stays clustered around 0.4 across the four finite-radius cells and drops by an order of magnitude at $r =$ `"all"` on both physionet and bcic2a — a sharp transition, not a smooth trend.

window of a given channel produces essentially the same direction in feature space. JEPA at $r = 6$ cm and MAE at $r =$ `"all"` remain stable on all four metrics across pre-training.

Phase transition in r. The dependence on the spatial radius is sharp rather than smooth. Sweeping the JEPA grid at $L = 2$ and the final epoch across $r \in \{\texttt{"one"}, 6, 9, 12\,\text{cm}, \texttt{"all"}\}$, the ratio $\sigma_{\text{token}}/\sigma_{\text{chan}}$ stays in $[0.36, 0.56]$ across the four finite-radius cells and then drops to ≈ 0.024 on physionet and ≈ 0.047 on bcic2a at $r =$ `"all"` — a phase transition of one order of magnitude in both datasets. The dependence on the mask ratio ρ, which the core grid holds fixed, is examined in Appendix H.2: ρ influences the collapsed cell more than any other, and the cell improves at higher ρ without becoming competitive.

Predictivity across the full grid. Aggregating across 10 configurations (full r-sweep at $L = 2$, full L-sweep at $r =$ `"all"`, and the MAE $r =$ `"all"` control) on both physionet and bcic2a, the log of $\sigma_{\text{token}}/\sigma_{\text{chan}}$ at the final epoch predicts the final downstream score with Pearson $r = +0.95$ on physionet and $+0.91$ on bcic2a. The bias-inflation signature is therefore not a quirk of one configuration but a quantitative predictor of downstream collapse across the full (L, r) grid.

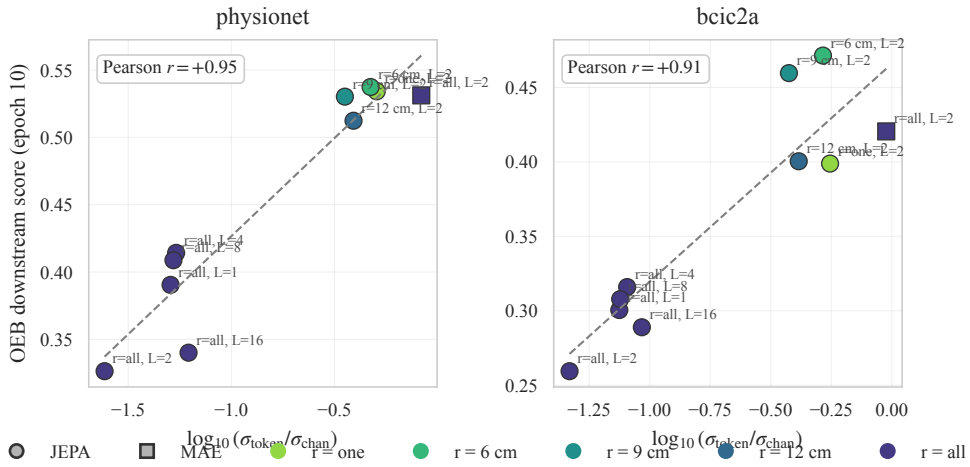


Figure 9: Predictivity of $\log_{10}(\sigma_{\text{token}}/\sigma_{\text{chan}})$ for the downstream OEB score at the final epoch across the 10 (framework, r, L) configurations. Each point is one frozen checkpoint at epoch 10. Markers: framework (○ JEPA, □ MAE). Colours: spatial radius (light → dark for small → large r). Pearson $r = +0.95$ on physionet and $+0.91$ on bcic2a.

Multi-subject robustness. We replicate the four metrics on five different subjects of each downstream dataset. Healthy and MAE trajectories are essentially subject-invariant; the bias-inflation trajectory is also robust, with subject-to-subject variation contained within 5% of the run-level mean.

E.3 Spectral selectivity of the collapsed representation

To characterise what the collapsed representation retains, we fit ridge probes on the frozen features of all 58 configurations of the core grid, across all twelve OpenEEGBench datasets, to predict the per-channel log band-power in the δ, θ, α, β and γ bands. Table 11 reports the decodability of each band (mean over channels), per dataset, for the two optima of the grid, the collapsed cell and an untrained encoder used as a floor. Bias-inflation collapse is spectrally selective, and this replicates on 12 of 12 datasets: the collapsed JEPA cell ($r =$ `"all"`, $L = 2$) retains the high bands far better than δ, θ and α, and its low-band decodability is the lowest of the grid on most datasets. The collapsed model therefore does not decode nothing: it keeps broadband amplitude and loses the slow structure of the signal.

E.4 The lookup-table mechanism

The structure of the JEPA loss admits, at $r =$ `"all"`, a particularly simple trivial fixed point. Recall that the JEPA loss

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{|\mathcal{M}|\, d} \sum_{(c,t)\in\mathcal{M}} \left\| \hat{z}_{c,t} - \bar{z}_{c,t} \right\|_2^2$$

is minimised whenever the predictor's output $\hat{z}_{c,t}$ matches the EMA teacher's output $\bar{z}_{c,t}$ at every masked position. Any function that depends only on the position (c, t) is a valid joint solution, since both student and teacher pass through the same positional encoder and the same architecture (modulo the EMA delay). We call such a solution a *lookup table*: a deterministic mapping $(c, t) \mapsto M(c, t) \in \mathbb{R}^d$ that the encoder learns to emit identically across windows, regardless of the input EEG. If the encoder converges to this solution then $z(n, c, t) \approx M(c, t)$ for every n, the EMA teacher does likewise, and the predictor (a small transformer that receives the positional embeddings of the masked tokens together with the encoder's contextual output) can reproduce $M(c, t)$ from positions alone. The JEPA loss reaches a near-zero floor without the encoder ever encoding any input-driven information.

Why $r =$ `"all"` unlocks the shortcut. The competing non-trivial solution is for the encoder to exploit the redundancy of EEG signals across spatial neighbours: when channel c at time-patch t

Table 11: Band-power decodability from frozen final-epoch features (R^2, mean over channels), per frequency band and per OpenEEGBench dataset, for the two optima of the core grid, the collapsed JEPA cell and an untrained encoder.

Band / dataset	MAE (9 cm, L=2)	JEPA (9 cm, L=2)	JEPA ("all", L=2), collapsed	Untrained
δ band				
arithmetic_zyma2019	0.72	0.50	−0.02	0.49
bcic2020-3	0.62	0.40	0.11	0.30
bcic2a	0.70	0.47	−0.08	0.29
chbmit	0.84	0.75	0.04	0.67
faced	0.67	0.51	0.03	0.42
isruc-sleep	0.82	0.64	−0.05	0.60
mdd_mumtaz2016	0.75	0.63	0.04	0.52
physionet	0.75	0.57	0.02	0.50
seed-v	0.68	0.49	0.13	0.33
seed-vig	0.66	0.50	0.03	0.46
tuab	0.77	0.63	0.05	0.48
tuev	0.80	0.68	0.14	0.57
θ band				
arithmetic_zyma2019	0.54	0.44	−0.06	0.17
bcic2020-3	0.63	0.55	0.24	0.20
bcic2a	0.59	0.45	0.10	0.08
chbmit	0.73	0.68	−0.01	0.40
faced	0.60	0.53	0.09	0.25
isruc-sleep	0.69	0.65	−0.12	0.19
mdd_mumtaz2016	0.77	0.70	−0.01	0.42
physionet	0.66	0.59	0.05	0.25
seed-v	0.54	0.48	0.11	0.06
seed-vig	0.64	0.59	0.13	0.35
tuab	0.73	0.68	0.07	0.37
tuev	0.75	0.71	0.10	0.39
α band				
arithmetic_zyma2019	0.71	0.74	0.17	0.40
bcic2020-3	0.68	0.64	0.24	0.24
bcic2a	0.73	0.73	0.07	0.16
chbmit	0.61	0.65	0.08	0.25
faced	0.58	0.60	0.07	0.15
isruc-sleep	0.72	0.75	−0.14	0.34
mdd_mumtaz2016	0.67	0.71	−0.04	0.34
physionet	0.65	0.66	0.08	0.20
seed-v	0.52	0.56	0.03	0.05
seed-vig	0.69	0.67	0.12	0.33
tuab	0.74	0.73	0.12	0.41
tuev	0.68	0.70	0.15	0.28
β band				
arithmetic_zyma2019	0.30	0.43	0.23	0.02
bcic2020-3	0.62	0.56	0.36	0.22
bcic2a	0.36	0.45	0.32	0.05
chbmit	0.71	0.76	0.49	0.41
faced	0.61	0.63	0.38	0.31
isruc-sleep	0.66	0.69	−0.19	0.36
mdd_mumtaz2016	0.66	0.73	0.54	0.32
physionet	0.49	0.56	0.34	0.10
seed-v	0.55	0.60	0.37	0.16
seed-vig	0.57	0.54	0.26	0.28
tuab	0.70	0.69	0.26	0.40
tuev	0.63	0.68	0.36	0.31
γ band				
arithmetic_zyma2019	0.15	0.28	0.32	0.02
bcic2020-3	0.51	0.47	0.42	0.16
bcic2a	0.51	0.67	0.69	0.07
chbmit	0.68	0.74	0.62	0.46
faced	0.59	0.68	0.60	0.35
isruc-sleep	0.59	0.64	0.20	0.29
mdd_mumtaz2016	0.42	0.52	0.24	0.15
physionet	0.36	0.50	0.53	0.05
seed-v	0.56	0.66	0.62	0.13
seed-vig	0.51	0.56	0.36	0.23
tuab	0.64	0.66	0.37	0.37
tuev	0.55	0.63	0.49	0.27

is masked but a nearby channel c' at the same t is not, volume-conducted activity at c' is strongly informative of the content at c. At $r \leq 12$ cm this same-time spatial neighbour is preserved in the predictor's context for nearly every masked position, and the spatial-redundancy path is at least as attractive as the lookup table. At $r =$ "all", every channel of the masked time-patch is masked together, so no same-time spatial neighbour ever survives in the context. The only signal left is the temporal axis, which is poorly predictive of EEG content at the timescale of our window (Spatial vs. temporal redundancy). The lookup table becomes the strictly easier path, and the optimiser converges to it.

Why MAE forbids the shortcut structurally. Under MAE, the target at masked position (c, t) is the raw input patch $x_{c,t} \in \mathbb{R}^p$, and the readout is a fixed linear map $W \in \mathbb{R}^{p \times d}$ followed by an L_2 loss. If the encoder converged to the same lookup-table solution $z(n, c, t) \approx M(c, t)$, the readout $W \cdot M(c, t)$ would be a per-position constant, identical across windows; the loss would then equal the cross-window variance of the raw EEG patch, which is of order one and far from zero. The reconstruction head therefore forbids the lookup-table solution and forces the encoder to keep multiple input-dependent latent directions alive — even at $r =$ `"all"`, where MAE merely under-performs but does not collapse.

E.5 Why standard detectors fail to fire

Three classes of collapse detector documented in the JEPA literature fail to register bias-inflation collapse, including the three triggers of our collapse-detection callback (Appendix A.7).

- **Per-dimension variance check.** The variance of z_d across the batch remains well above the $\epsilon = 10^{-2}$ threshold throughout pre-training, because the inflated bias $\mu(c, t)$ varies across (c, t) cells and contributes $\Theta(\|\mu\|^2)$ to the per-dimension variance.
- **Loss divergence.** The main JEPA loss does not blow up: in the lookup-table regime, student and teacher converge *together* to the same $M(c, t)$, and the loss drives smoothly toward zero rather than past 10^4.
- **Trivial-constant collapse.** The loss does not collapse to $< 10^{-3}$ within the first 500 steps either, because the lookup table is a non-constant function of position and is learnt over many epochs (the bias inflation occurs gradually between epochs 4 and 10, see Appendix E.2).

VICReg-style regularisers are unlikely to rescue the regime. The VICReg variance hinge $\mathcal{L}_{\text{var}} = \frac{1}{d} \sum_{i=1}^{d} \max(0, \gamma - \sqrt{\text{Var}(z_i)})$ is satisfied because $\text{Var}(z_i)$ is dominated by the bias term and remains above any reasonable target γ. The covariance term $\mathcal{L}_{\text{cov}} = \frac{1}{d} \sum_{i \neq j} \text{Cov}(z_i, z_j)^2$ is satisfied whenever the lookup table $\{M(c, t)\}$ spreads isotropically across the d feature dimensions, and we observe empirically that it does: the effective rank of $\{M(c, t)\}$ grows from ~ 5 at epoch 1 to ~ 25 at epoch 10 on physionet, indicating that the encoder actively distributes the lookup directions across feature dimensions. The combined objective $\mathcal{L}_{\text{JEPA}} + \mathcal{L}_{\text{var}} + \mathcal{L}_{\text{cov}}$ therefore appears to be satisfied by the lookup table at no cost, so VICReg (Bardes et al., 2021), C-JEPA (Mo & Tong, 2024), and similar variance-preserving regularisers are unlikely to detect or prevent bias-inflation collapse in the multichannel setting. We do not test this prediction empirically: our entire grid is run in the no-auxiliary-loss configuration to keep the framework axis at strict parity with MAE, so adding VICReg or C-JEPA at $r =$ `"all"` is left to future work. Distributional regularisers that constrain the full embedding distribution rather than its second moments — such as the SIGReg term of LeJEPA (Balestriero & LeCun, 2025) — might in principle rule out the lookup-table fixed point, but we have not verified this either.

E.6 Why detection during pre-training is non-trivial

The cross-window signature of Appendix E.2 ($\sigma_{\text{token}}/\sigma_{\text{chan}}$, $\overline{\text{cos}_{\text{intra}}}$) is a quantitative predictor of bias-inflation collapse, but it relies on a fixed channel set across windows. This assumption holds on a downstream evaluation dataset but breaks in pre-training: REVE windows come from recordings with different channel layouts, so there is no canonical correspondence of channels across windows in a batch. A practical in-training detector would therefore have to operate *within* a single window (or a single recording) without channel-set assumptions, and would have to discriminate the early-pre-training plateau of a collapsed run from the legitimately slow take-off of a healthy run. We could not construct a detector that meets both requirements reliably; this paragraph documents the natural candidates we explored and why each fell short, so that future work can build on the attempt.

Within-window candidate quantities. Given the encoder output $z \in \mathbb{R}^{W \times C \times T \times d}$ on a batch of W windows (with per-window non-padded channels indicated by a binary mask), the natural

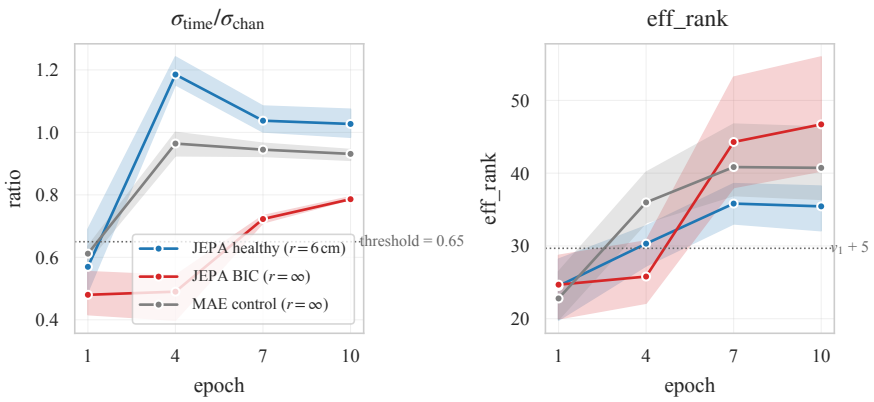


Figure 10: Candidate within-window proxies $\sigma_{\text{time}}/\sigma_{\text{chan}}$ (left) and $\mathrm{eff_rank}$ (right) on real REVE pre-training windows, at epochs $1, 4, 7, 10$ for the three core runs. Lines: mean across the 32 windows of the evaluation shard; bands: 25–75 percentile. By epoch 4 the BIC run is still at its epoch-1 values on both metrics, while healthy and MAE have already taken off — but neither metric discriminates the runs at initialisation, and both can plausibly look low for legitimate reasons in early training.

within-window analogues of the cross-window signature are

$$\sigma_{\text{time}} := \operatorname*{mean}_{w} \operatorname*{mean}_{c \text{ valid in } w} \operatorname*{mean}_{t} \left\| z_w(c,t) - \operatorname*{mean}_{t'} z_w(c,t') \right\|,$$
$$\sigma_{\text{chan}} := \operatorname*{mean}_{w} \operatorname*{mean}_{t} \operatorname*{mean}_{c \text{ valid in } w} \left\| z_w(c,t) - \operatorname*{mean}_{c' \text{ valid in } w} z_w(c',t) \right\|,$$

together with the effective rank $\mathrm{eff_rank} := \exp\big(-\sum_i p_i \log p_i\big)$ of the centred token matrix flattened to $(N_{\text{valid}}\, T, d)$, computed per window and averaged across the batch.

What we observed on real REVE windows. On a single 1 GB shard of the REVE training cache (32 windows per evaluation), we measured $\sigma_{\text{time}}/\sigma_{\text{chan}}$ and $\mathrm{eff_rank}$ at four epochs $(1, 4, 7, 10)$. At epoch 1 all three runs (healthy at $r = 6$ cm, BIC JEPA at $r =$ `"all"`, MAE control at $r =$ `"all"`) sit below the eventual healthy plateau on both metrics — no static threshold would discriminate the BIC run from a healthy run at initialisation. By epoch 4 ($\sim$30% of the budget) the healthy and MAE runs have already taken off ($\sigma_{\text{time}}/\sigma_{\text{chan}} \approx 1.19$ and 0.96 respectively, $\mathrm{eff_rank}$ around 30 and 36), whereas the BIC run is still at its epoch-1 values ($\sigma_{\text{time}}/\sigma_{\text{chan}} \approx 0.49$, $\mathrm{eff_rank} \approx 26$).

Why we do not turn this into a detector. Although a conjunctive rule of the form "$\sigma_{\text{time}}/\sigma_{\text{chan}} < 0.65$ AND $\mathrm{eff_rank} - \mathrm{eff_rank}^{(1)} < 5$ after a warm-up of $\sim$30% of the budget" separates the three runs in our shard test, we caution against using it as a stopping criterion in practice: (i) its thresholds are calibrated on the same three runs that motivate the rule, with no independent validation set; (ii) the warm-up window is itself a hyperparameter that depends on the planned epoch budget; (iii) the underlying proxies are confounded with the legitimate slow take-off of healthy runs in the early epochs; (iv) we have not tested the rule across other pre-training corpora, montage distributions, or backbones — all of which could shift the healthy baseline. A reliable in-training detector for bias-inflation collapse on multi-montage corpora most likely requires either a small held-out probe dataset evaluated periodically, or a distributional regulariser (such as the SIGReg term of LeJEPA (Balestriero & LeCun, 2025)) that structurally rules out the lookup-table fixed point; we leave both directions to future work.

Practical mitigation. Until a reliable in-training detector exists, the simplest and most effective mitigation is to *avoid the masking regime that triggers the collapse*. Across the 58-cell grid the failure is confined to JEPA at $r =$ `"all"` (Figure 3 (C)), so the actionable rule for practitioners pre-training EEG foundation models with a JEPA-style pretext is to never mask every channel of the same time-patch jointly, and to prefer a moderate spatial radius ($r \in \{6, 9, 12\}$ cm) as recommended in Finding 2. Under this rule, no run in our grid exhibits bias-inflation collapse, and the question of in-training detection is sidestepped.

E.7 HYPOTHESES WE RULED OUT

For completeness, the following plausible mechanistic stories were ruled out by the data:

- *Rank-1 magnitude modulation,* $z = \alpha(n, c, t) \cdot u(c, t) + \varepsilon$: ruled out by measuring the parallel fraction of δ along μ, which stays below 2% across all runs — δ is essentially orthogonal to μ rather than modulating its magnitude.
- *Amplification of the random-init direction:* ruled out by $\cos\big(\mu(c,t)^{(10)}, \mu(c,t)^{(1)}\big) \approx 0.28$ for both healthy and collapsed runs — the direction rotates as much during bias-inflation collapse as during healthy training, only the magnitude inflates differently.
- *Topographic readout:* ruled out by the cosine-vs-distance Pearson correlation across channel pairs, which is *lower* for collapsed runs (~ 0.40) than for healthy (~ 0.55) or MAE (~ 0.62) — the lookup table is essentially arbitrary across channels rather than smoothly topographic.
- *Bias-rescue by feature standardisation:* ruled out by a probe-with-scaler vs. probe-without-scaler comparison whose accuracy gap is below 1% across runs — the input information is structurally destroyed rather than merely masked by the bias term.
- *Monotonicity in* L *at* $r =$ `"all"`: ruled out by the L-sweep at the final epoch, where the most collapsed configuration is $L = 2$ and configurations with $L = 1$ or $L \geq 4$ all show milder bias inflation.
- *Artefact of the* $1\,s$ *tokeniser:* ruled out by two runs with $0.5\,\text{s}$ patches at $r =$ `"all"` (Appendix H.3), which display the full signature and score far below the healthy geometry downstream.

E.8 RELATION TO PRIOR WORK

We are not aware of a prior description of bias-inflation collapse in its specific form — a JEPA-only failure mode, on multichannel signals, triggered by a mask geometry that hides every channel of the same time-patch jointly, in which the encoder converges to a per-position lookup table $M(c, t)$ with an inflating bias μ and a vanishing input-driven residual δ. Several prior works, however, anticipate parts of the picture and we situate our contribution against them here.

Variance-fooling collapse. Li et al. (2022) describe a partial dimensional collapse of non-contrastive Siamese SSL in which the per-dimension variance is preserved while the embedding distribution concentrates on a low-dimensional subspace; they show that PCA singular values detect the failure where per-dimension variance does not. Bias-inflation collapse falls in the same broad family of "high-variance, degenerate-representation" regimes, but the failure mode is different: the encoder output is high-rank rather than low-rank (the effective rank actively grows from ~ 5 to ~ 25 during collapse, see Appendix E.5), and its degeneracy is position-indexed rather than subspace-restricted.

Mean drift in EMA-teacher SSL. The DINO "centering" trick (Caron et al., 2021) subtracts a running mean of the teacher output to prevent one dimension from dominating the softmax, and is one of the earliest acknowledgements that EMA teachers can drift along their own mean. Mo & Tong (2024), motivating C-JEPA, observe that I-JEPA "inadequately learns the mean of patch representations" and propose VICReg-style regularisation as a fix. Bias-inflation collapse is consistent with this family of mean-drift concerns, but is sharper in two ways: the drift is position-indexed (different for each (c, t) cell) rather than a single global mean vector, and the proposed VICReg fix is, by our analytical argument (Appendix E.5), structurally satisfied by the lookup-table solution.

Collapse via aggressive masking on highly correlated modalities. Baevski et al. (2022b) report that EMA-teacher latent prediction is prone to collapse for "modalities where adjacent targets are very correlated and where longer spans need to be masked, e.g. speech", and add target normalisation over the sequence (instance-norm for speech) to mitigate it. Our $r =$ `"all"` regime is the multichannel-spatial analogue of their long-temporal-span observation: every channel of the same time-patch is masked jointly, so the only surviving context is on the temporal axis, which is itself highly autocorrelated for EEG. data2vec's instance-norm fix targets a single sequence axis and does not generalise straightforwardly to a per-position $\mu(c, t)$ over a 2D channel-time grid.

Positional shortcuts in masked-prediction SSL. Zhang et al. (2024) (PCP-MAE) document a "center leakage" phenomenon in point-cloud MAE: the decoder reconstructs plausibly even when fed only the positional embeddings of masked patches, without the encoder's output, indicating that positional metadata alone supplies the prediction signal. Bar et al. (2024) show that deterministic positional embeddings let MIM models latch onto absolute location and propose stochastic positions to break the shortcut. Bias-inflation collapse is a member of the same broader family "the model exploits positional information to bypass the input", but with a distinct mechanism: it is the encoder (not the decoder) that converges to a position-only function, and the JEPA student-teacher loss collapses to zero rather than the MAE reconstruction succeeding from positions alone.

JEPA loss as a poor proxy for representation quality. A growing body of recent work argues that JEPA pre-training loss decreases independently of downstream representation quality, motivating non-loss probes such as RankMe (Garrido et al., 2023) or SIGReg (Balestriero & LeCun, 2025). Bias-inflation collapse is a concrete, mechanistically characterised instance of this broader concern, with one extra wrinkle: rank-based probes such as RankMe are themselves fooled by the lookup-table solution, since the per-position bias spreads isotropically across feature dimensions and yields a high effective rank (Appendix E.5). Distributional regularisers that constrain the full embedding distribution, such as SIGReg, would in principle rule out the lookup-table fixed point, but we have not verified this empirically.

F PER-DATASET BREAKDOWN OF THE DOWNSTREAM PERFORMANCE

This appendix reports per-dataset versions of the aggregated panels of section 4. Figure 11 expands the average heatmap of Figure 3 (A) into one row per OpenEEGBench dataset, on the raw score scale (each row uses its own colour-bar so the per-dataset diff is interpretable in the dataset-native metric, balanced accuracy or R^2). The bottom row *Average (norm.)* reproduces the across-dataset summary of Figure 3 (A) on the dataset-normalised scale and is included for direct comparison with the main paper. Figure 12 likewise expands the across-epoch summary of Figure 3 (C) into one sub-panel per dataset, with each 58 pre-training run drawn as a single transparent line so that within-framework variance over the 10 checkpoints is visible.

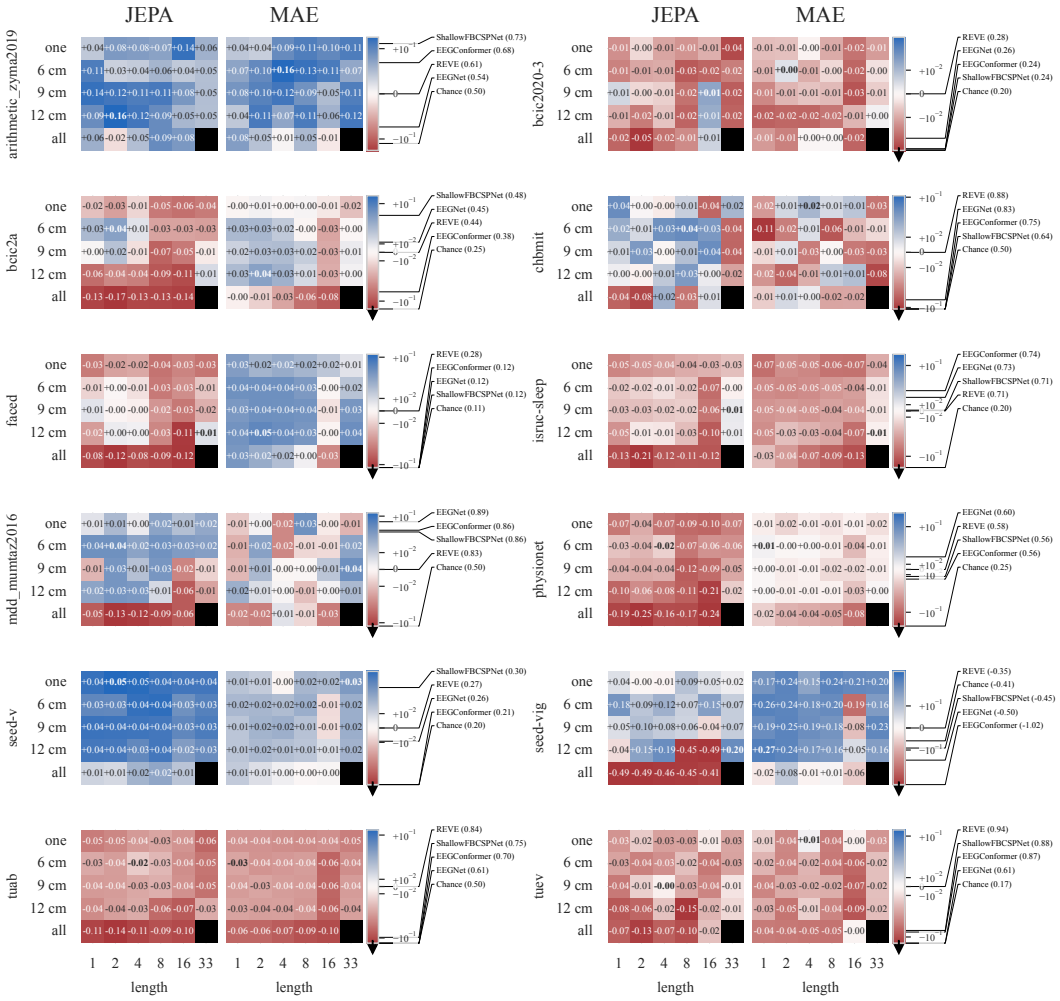


Figure 11: Per-dataset, per-(framework, L, r) score difference to the REVE ridge-probe baseline on OpenEEGBench. One row per dataset, one column per framework; each row uses its own diverging colour-bar so the diffs are interpretable in the dataset-native metric (balanced accuracy or R^2). Sidebars on the right of each row mark the absolute REVE level, the chance level, and the three supervised baselines (EEGNet, EEGConformer, ShallowFBCSPNet) trained end-to-end per task. The bottom row *Average (norm.)* reproduces Figure 3 (A) on the cross-dataset normalised scale. The trends summarised in Section 4 are visible dataset-by-dataset: the moderate-r, short-L cluster wins on most datasets, $r =$ `"all"` underperforms for JEPA on most datasets (with the exceptions of `arithmetic_zyma2019` and `seed-v`), and $(L, r) = (2, 9\,\text{cm})$ is the aggregate JEPA optimum, although the local per-dataset peak sometimes lies elsewhere.

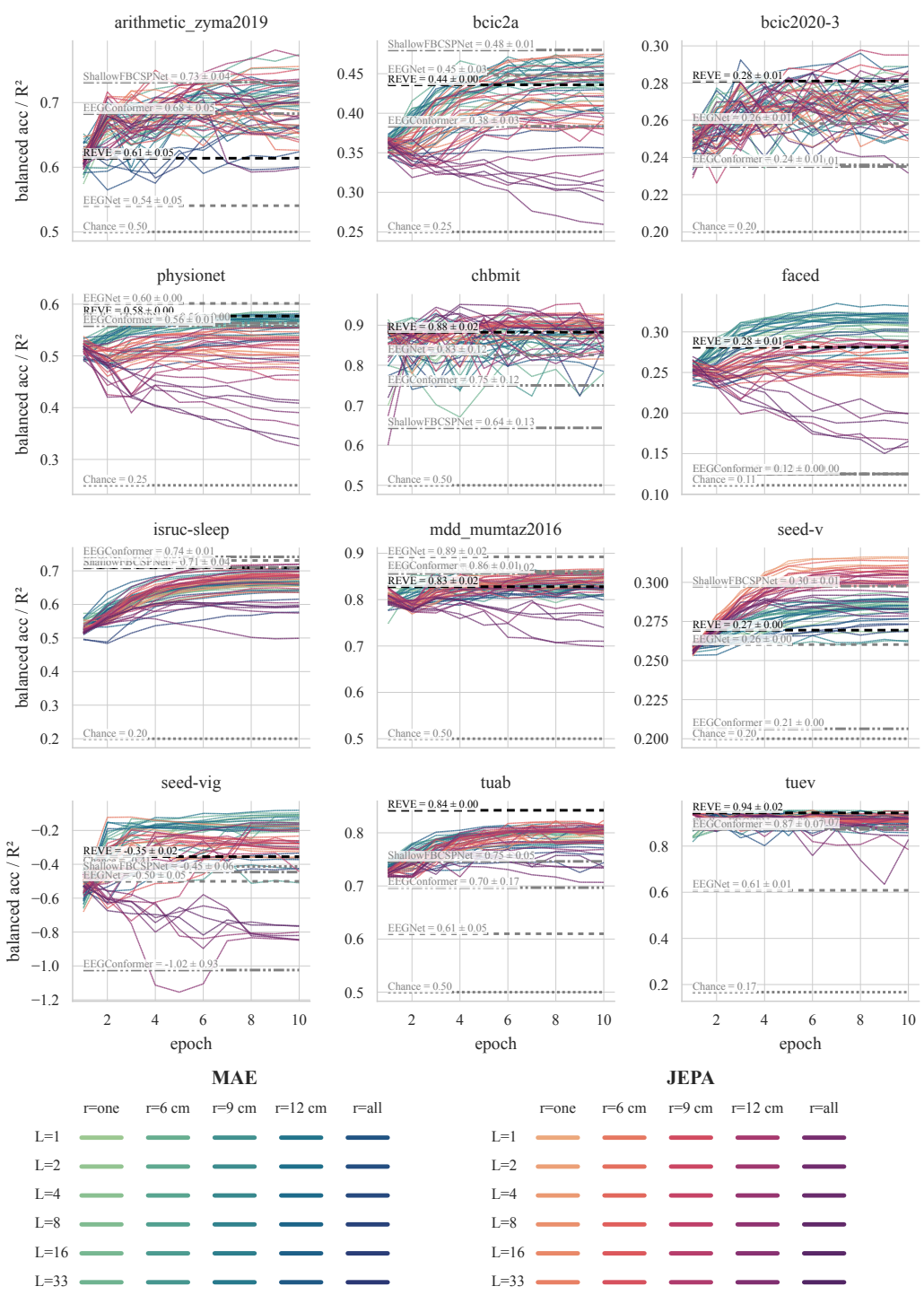


Figure 12: Per-dataset training trajectory of the downstream score across the 10 pre-training epochs. Each line is one of the 58 pre-training configurations (hue: framework), averaged across 3 probe seeds. Per-dataset chance (dotted grey), the REVE baseline mean ± std (dashed black), and the three supervised baselines (EEGNet, EEG Conformer, ShallowFBCSPNet, dashed coloured) are overlaid for reference; the supervised baselines are trained end-to-end per task, while all FM curves use a frozen-feature ridge probe. Across datasets, both frameworks reach a near-final ranking by epoch ~ 5; pathological JEPA runs at $r =$ `"all"` are visible as the bundle that drops below the REVE band early and never recovers.

G Comparison with published EEG foundation models

Table 12 compares our recommended geometry $(L, r) = (2\,\text{s}, 9\,\text{cm})$ with the public checkpoints of CBraMod (Wang et al., 2025), EEGPT (Wang et al., 2024), LaBraM (Jiang et al., 2024) and BIOT (Yang et al., 2023) on the 12 OpenEEGBench datasets. All models are evaluated by linear probing on a frozen backbone, with one of two ways of fitting the linear layer. The *ridge* probe is the analytical fit used throughout this paper (Appendix C.2); it was added to the benchmark after its original publication and requires the backbone to be entirely frozen. The *SGD* probe is the one of the original OpenEEGBench paper (Guetschel et al., 2026b), which fits the linear layer by iterative gradient descent; we report the scores published there, which cover ten of the twelve current datasets. The ridge probe does not apply to EEGPT, BIOT and LaBraM, because they contain modules that must be retrained on each dataset and therefore cannot be fully frozen; these three models appear under the SGD probe only.

Table 12: Downstream scores of the recommended geometry against public EEG foundation models on the 12 OpenEEGBench datasets (balanced accuracy; R^2 for `seed-vig`). Ridge: the analytical linear probe of this paper on fully frozen backbones. SGD: the linear probe of the original OpenEEGBench paper, whose published scores are reported (dashes: datasets it did not cover). Best score per row in bold.

	Ours $(2\,\text{s}, 9\,\text{cm})$, ridge		Public, ridge		Public, SGD (published)		
Dataset	MAE	JEPA	REVE-Base	CBraMod	EEGPT	LaBraM	BIOT
`arithmetic_zyma2019`	0.716	**0.737**	0.614	0.562	0.551	0.499	0.553
`bcic2020-3`	0.273	0.277	**0.281**	0.233	0.207	0.200	0.216
`bcic2a`	**0.462**	0.460	0.436	0.315	0.296	0.258	0.264
`chbmit`	0.888	0.908	0.882	**0.941**	—	—	—
`faced`	**0.321**	0.280	0.281	0.212	0.174	0.125	0.155
`isruc-sleep`	0.665	0.682	**0.709**	0.647	0.487	0.466	0.702
`mdd_mumtaz2016`	0.842	**0.858**	0.828	0.776	0.778	0.783	0.825
`physionet`	0.568	0.533	**0.577**	0.474	0.312	0.294	0.263
`seed-v`	0.288	**0.307**	0.269	0.255	0.212	0.228	0.239
`seed-vig` (R^2)	**−0.108**	−0.258	−0.355	−0.716	—	—	—
`tuab`	0.808	0.805	**0.842**	0.746	0.715	0.602	0.573
`tuev`	0.919	0.930	**0.944**	0.895	0.338	0.221	0.357

Under the ridge probe, our recommended geometry is ahead of CBraMod, the strongest published model after REVE-Base, on 11 of the 12 datasets, and none of the SGD-probed models match REVE-Base. The models trained in this study are therefore not toy models: they are competitive with the published state of the art.

H Ablations

This section tests how far the conclusions of the core grid reach, along three axes: whether the recommended geometry transfers to other backbones (Appendix H.1), whether it survives changes of the mask ratio (Appendix H.2), and whether bias-inflation collapse depends on the patch length (Appendix H.3). Throughout, downstream scores are mean normalised OpenEEGBench scores over the 12 datasets and 5 probe seeds, computed with the normalisation constants of the core grid (Appendix C.4) so that they can be read against the grid.

H.1 Transfer of the recommended geometry to other backbones

Does the recommended geometry help architectures other than REVE-Small? To find out, we retrained the CBraMod (Wang et al., 2025) and LaBraM (Jiang et al., 2024) architectures under MAE on our corpus, each with two masking strategies: the random masking they were originally trained with, $(r = \texttt{"one"}, L = 1)$ at $\rho = 0.5$, and our recommendation $(r = 9\,\text{cm}, L = 2)$ at the same $\rho = 0.5$, for fairness. One deviation from CBraMod deserves mention. CBraMod encodes channel positions with 1D convolutions along the channel axis, which depend on the order in which the channels are stored; since that order is arbitrary, the encoding is flawed by design, and rather than

re-implement it we used the sinusoidal encoding of the rest of the paper. On both backbones, the recommended geometry improves the downstream score over the original strategy (Table 13), which supports the generalisability of our recommendation across architectures.

Table 13: Mean normalised OpenEEGBench score (12 datasets $\times$ 5 probe seeds) of the CBraMod and LaBraM architectures re-trained under MAE on our corpus, with their published mask ($r =$ `"one"`, $L = 1$) and with the recommended geometry ($r = 9\,\text{cm}$, $L = 2$), both at $\rho = 0.5$. The REVE-Small row is the same contrast in the core grid, at $\rho^\star = 0.55$.

Backbone	Published mask	Recommended geometry	Recommended better on
REVE-Small (core grid)	0.710	0.800	9 of 12
CBraMod	0.597	0.640	10 of 12
LaBraM	0.605	0.642	7 of 12

H.2 Sensitivity to the mask ratio

The core grid fixes the mask ratio at $\rho^\star = 0.55$; this section asks whether its conclusions hold at other ratios. Sweeping ρ over the entire grid is out of reach ($58 \times 5 = 290$ pre-training runs), so we sweep $\rho \in \{0.25, 0.40, 0.55, 0.70, 0.85\}$ on the four geometries about which the paper makes its main claims, under both frameworks: the recommended cell ($r = 9\,\text{cm}, L = 2$), random masking ($r =$ `"one"`, $L = 1$), a large mask ($r = 12\,\text{cm}, L = 16$) and the bias-inflation cell ($r =$ `"all"`, $L = 2$). This amounts to 30 new pre-training runs; Table 14 reports their scores next to the corresponding cells of the core grid.

Table 14: Mean normalised OpenEEGBench score (12 datasets $\times$ 5 probe seeds) as a function of the target mask ratio ρ, for four geometries under both frameworks. The $\rho = 0.55$ column is the core grid.

Framework	Geometry (r, L)	$\rho = 0.25$	0.40	0.55 (grid)	0.70	0.85
MAE	(`"one"`, 1)	0.59	0.66	0.71	0.83	0.78
MAE	(9 cm, 2)	0.74	0.82	0.80	0.81	0.78
MAE	(12 cm, 16)	—	0.57	0.63	0.59	0.63
MAE	(`"all"`, 2)	0.61	0.65	0.65	0.66	0.56
JEPA	(`"one"`, 1)	0.59	0.64	0.70	0.72	0.72
JEPA	(9 cm, 2)	0.79	0.86	0.83	0.77	0.73
JEPA	(12 cm, 16)	—	0.45	0.46	0.57	0.59
JEPA	(`"all"`, 2)	−0.34	−0.16	0.05	0.32	0.37

The first lesson of Table 14 is that ρ and geometry are not independent knobs: the same change of ρ acts differently on different geometries. The spatially large masks ($12\,\text{cm}, L = 16$) and (`"all"`, $L = 2$) perform poorly at every ρ under both frameworks, which confirms that they are failure modes: their low scores come from the geometry, not from the ratio. The bias-inflation cell (JEPA, `"all"`, $L = 2$) is the cell that ρ influences most: it gets worse at lower ρ and better at higher ρ, while staying non-competitive throughout. We have no explanation for this dependence at this point; it could be a lead for understanding the mechanism in a follow-up study. The recommended geometry ($9\,\text{cm}, L = 2$) is the least sensitive to ρ. A slightly lower ratio, $\rho = 0.4$, gives a small gain over the default ($+0.023$ under MAE, $+0.030$ under JEPA), which suggests that this ratio might be optimal; we report it without leaning on it. Random masking (`"one"`, $L = 1$), finally, degrades sharply at low ρ but improves at $\rho = 0.7$ under both frameworks, to the point of matching the recommended geometry under MAE.

This last observation has a geometric explanation. Raising ρ changes the effective geometry of the mask: the number of blocks K grows, more blocks overlap, and the merged blocks become larger than the nominal geometry, all the more so when the nominal blocks are small.

The sweep also illustrates why the core grid holds ρ fixed. Changing ρ changes the learning task itself: it sets how many tokens the network has to predict, hence over how many tokens the learning signal is spread, and how many unmasked tokens remain available to make the prediction; both factors affect

the difficulty of the pretext task. Were ρ and the geometry (L, r) to change together, these influences would combine, the task difficulty would move in an uncontrolled way, and the two effects could not be separated. Published models use ρ close to 0.55, which we therefore took as a prior and held fixed. The overall conclusion of the sweep is that the geometry recommendation $(9\,\text{cm}, L = 2)$ holds.

H.3 Patch length and bias-inflation collapse

At $r =$ `"all"`, the only context left to the predictor is the temporal axis, and the patch length sets how predictive that axis is: with shorter patches, within-channel temporal correlations are stronger and the lookup-table shortcut of Appendix E.4 less attractive. The collapse could therefore be an artefact of our 1 s patches, which this ablation tests. The test cannot be made free of confounds. The window duration is fixed at 30 s, so halving the patch length doubles the token count, and one of two things must give: (A) keep the dimension of each token, so that the model size is unchanged but the signal is represented over twice as many features per second; or (B) keep the total dimensionality, so that each token becomes smaller, which requires adapting the architecture to the new token size. Either way the token count grows, and since attention is quadratic in it, a 0.5 s tokeniser is more expensive to pre-train than a 1 s one. We nevertheless ran both options under JEPA at $(r =$ `"all"`$, L = 1)$ with 0.5 s patches; each run needed 4 GPUs to complete in about the same time as the 1 s runs on 2. We then benchmarked the two runs downstream and re-ran the diagnostics of Appendix E.2 on them (Table 15). Both stay well below the healthy geometry on the downstream benchmark, and both display the complete collapse signature: $\|\mu\|$ inflates about twofold over pre-training, σ_{token} stays flat or shrinks, the ratio $\sigma_{\text{token}}/\sigma_{\text{chan}}$ falls monotonically, and the intra-channel cosine reaches 0.94–0.98. Bias-inflation collapse is therefore not an artefact of the 1 s patch length: it persists at 0.5 s, and the shorter patch only makes pre-training more expensive. The reference rows of Table 15 are cells of the core grid; for the reasons above, they are not directly comparable to the new runs.

Table 15: Patch-length ablation: diagnostics of Appendix E.2 at the final epoch, inflation of $\|\mu\|$ between epochs 1 and 10, parameters of the encoder / of the whole model, pre-training cost, and mean normalised OpenEEGBench score. Reference rows are cells of the core grid.

Run	Patch	Framework	Token dim.	Mask (r, L)	$\sigma_{\text{token}}/\sigma_{\text{chan}}$	$\overline{\cos}_{\text{intra}}$	$\|\mu\|$, ep. $1 \to 10$	Params enc. / tot.	GPU-h	OEB score
(A) same model size	0.5 s	JEPA	512	(`"all"`, 1)	0.18 / 0.17	0.98 / 0.98	×2.0 / ×2.2	12.6 M / 21.3 M	39.5	0.598
(B) same features per second	0.5 s	JEPA	256	(`"all"`, 1)	0.27 / 0.24	0.94 / 0.95	×2.1 / ×2.2	3.2 M / 5.3 M	30.6	0.417
Core grid, collapsed	1.0 s	JEPA	512	(`"all"`, 1)	0.05 / 0.08	1.00 / 1.00	×2.0 / ×2.0	12.7 M / 21.3 M	16.3	0.320
Core grid, collapsed	1.0 s	JEPA	512	(`"all"`, 2)	0.02 / 0.05	1.00 / 1.00	×2.4 / ×2.5	12.7 M / 21.3 M	14.0	0.054
Core grid, healthy	1.0 s	JEPA	512	(6 cm, 2)	0.47 / 0.52	0.87 / 0.89	×1.4 / ×1.5	12.7 M / 21.3 M	14.0	0.757
Core grid, MAE control	1.0 s	MAE	512	(`"all"`, 2)	0.83 / 0.95	0.71 / 0.72	×1.0 / ×1.0	12.7 M / 21.2 M	12.7	0.652