MoRSE: Distilling EEG Foundation Models into a Single Montage-Robust Multi-Task Encoder

Anonymous authors
Paper under double-blind review

Abstract

Electroencephalography (EEG) foundation models are pre-trained across many electrode montages, and several accept any number of channels by construction. However, they are typically deployed after fine-tuning on a task recorded with a single montage, which largely removes this montage flexibility. Existing remedies, such as channel interpolation or montage-sampled fine-tuning, still leave one large model per task. To address this, we propose MoRSE (Montage-Robust Student for EEG), which distills fine-tuned EEG foundation models into a single small multi-task encoder that remains accurate on electrode subsets. In this asymmetric distillation, teachers observe the full montage while the student observes a random electrode subset resampled each step, and so learns to infer the full-montage representation from the electrodes available. Having found that fine-tuning typically compresses representations to an effective dimensionality that tracks the class count, we align the student with a never-fine-tuned backbone. We evaluate MoRSE on eight tasks spanning seizure detection, sleep staging, emotion recognition, motor imagery and dementia diagnosis, where a model fine-tuned without montage variation loses 0.142 balanced accuracy when electrodes are removed. Across 30 task-by-montage cells, a single 1.97M-parameter encoder retains 95% of the balanced accuracy of the per-cell specialists, with 89$\times$ fewer deployed parameters. Against eight montage-sampled specialists that are 24$\times$ larger in aggregate, it retains 98%. It also transfers zero-shot to unseen electrode layouts and to two external sleep-staging datasets. Thus, one encoder a third the size of a single specialist can replace multiple task-specific foundation models across tasks and layouts. Anonymous code is available via the Reproducibility Statement.

1 Introduction

Electroencephalography (EEG) records the electrical activity of the brain non-invasively from scalp electrodes. It is inexpensive, portable and temporally precise, and deep learning models now decode it across tasks such as seizure detection, sleep staging, emotion recognition and motor imagery (Jing et al., 2023; Supratak et al., 2017; Song et al., 2020; Lawhern et al., 2018). Most are trained on a single dataset, and a dataset arrives with a montage, the set of electrodes it records. The possible electrode positions are standardized. The international 10–20 system (Jasper, 1958) and its 10–10 and 10–5 extensions (Chatrian et al., 1985; Oostenveld & Praamstra, 2001) name the scalp positions that clinical guidelines prescribe (Seeck et al., 2017). Which of those positions a recording uses, and how many, is not standardized. A clinical review montage carries twenty-three (Obeid & Picone, 2016), a clinical sleep study four (Alvarez-Estevez & Rijsman, 2022) and a consumer headset fourteen (Duvinage et al., 2013). This variation is independent of the task, since emotion recognition alone is recorded at sixty-two (Liu et al., 2022), thirty-two (Koelstra et al., 2012) and fourteen electrodes (Katsigiannis & Ramzan, 2018). A model trained on one montage therefore does not transfer to a dataset recorded with another, because its input layer expects electrodes that dataset lacks. EEG datasets thus differ along two independent axes, the task and the montage, and the montage is set not by the deployer but by an existing recording or by the device worn by the user.

EEG foundation models are designed to break that tie. They pre-train on corpora spanning many datasets with different caps (Jiang et al., 2024; El Ouahidi et al., 2025), and their architectures

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

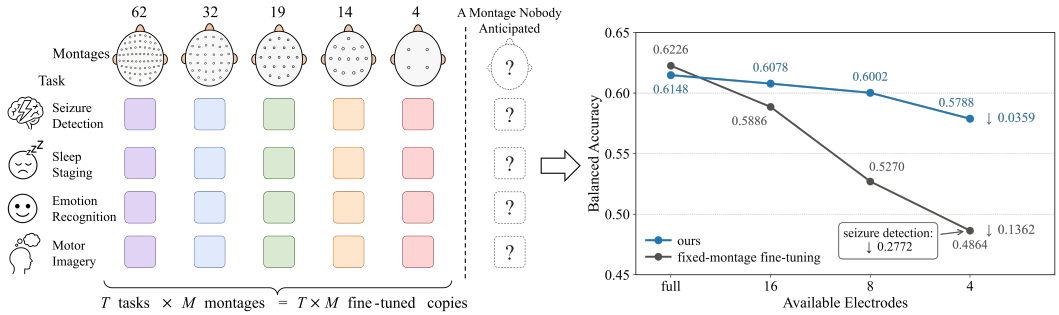

Figure 1: **A model fine-tuned on one (task, montage) cell is bound to that montage. Left:** covering T tasks on M montages by per-cell fine-tuning takes $T \times M$ copies, and a montage nobody anticipated, the dashed column, is covered only once labels for it have been collected. **Right:** balanced accuracy as electrodes are removed at test time, over the seven tasks with reduced budgets.

accept this variation. LaBraM represents a recording as per-channel patches and accepts an arbitrary number of channels (Jiang et al., 2024), and the conditional positional encoding of CBraMod adapts to diverse formats (Wang et al., 2025). Accepting a montage is therefore not the difficulty. A foundation model is deployed after task fine-tuning, which fits it to the one montage its task was recorded with, so the montage flexibility is a property of the backbone rather than of the deployed model. A copy fine-tuned on one (task, montage) cell is bound to that cell, so covering T tasks on M montages takes $T \times M$ copies (Figure 1, left). A student that inherits this architecture and pre-training but is fine-tuned without montage variation loses 0.142 balanced accuracy averaged over eight tasks when electrodes are dropped at test time (0.136 over the seven tasks reduced to four electrodes in Figure 1, right), and 0.277 on seizure detection. Each copy is also expensive to carry on wearable hardware, since the backbones we retrain have 4.9 to 69.2M encoder parameters and the task head of CBraMod alone runs 12 to 40M by task.

Three lines of prior work address the problem in part. Architectural approaches build montage invariance into the encoder (Yi et al., 2023; Döner et al., 2025; Foumani et al., 2025; Chen et al., 2025), but they pre-train a new foundation model and are evaluated after the per-task fine-tuning whose effect we measure in Figure 1. Input-side approaches project a reduced montage back to a canonical one before inference, and spherical-spline interpolation is one of four such adaptations benchmarked by Kokate et al. (2026). They still keep one full-montage specialist per task and pay its full-montage inference cost on every window. Distillation approaches compress EEG foundation models into small students (Wang et al., 2026), but only at a fixed montage. The two that grant the teacher more electrodes than the student (Kumaravel et al., 2023; Zhang et al., 2022) fix the reduced set of the student in advance and are thus robust to that set only. A simpler remedy, sampling the montage while fine-tuning the specialist itself, needs no new component. We adopt it and interpolation as baselines, and the former is the strongest in this paper (Section 3.3). None of these remedies reduces the number of models. Each leaves one fine-tuned backbone per task, so the matrix becomes a column of T large models, and this column is what our method compresses. To our knowledge, no method both resamples the electrode subset at every training step and evaluates the resulting model on layouts the sampler draws with negligible probability.

We propose **MoRSE** (Montage-Robust Student for EEG), a shared encoder with one linear head per task and 1.97M parameters in total, distilled from frozen teachers that always see the full montage while the student sees a random electrode subset, resampled at every step. The name denotes a scheme rather than a set of weights. MoRSE pre-trains no new foundation model, but compresses existing ones into a model deployable on any subset of the electrode vocabulary it knows. Logits come from the fine-tuned experts and per-patch tokens from the un-fine-tuned backbone, because we find that fine-tuning collapses the structure that representation distillation must transfer and leaves intact the decision boundary that logit distillation needs (Section 2.2). We evaluate on seven hardware layouts that the sampler draws with negligible probability, among them a frontal headband, a temporal seizure strip and a single-hemisphere setup. One of them is more clustered than 99.9% of same-size uniform draws. The contributions of this work are summarized as follows:

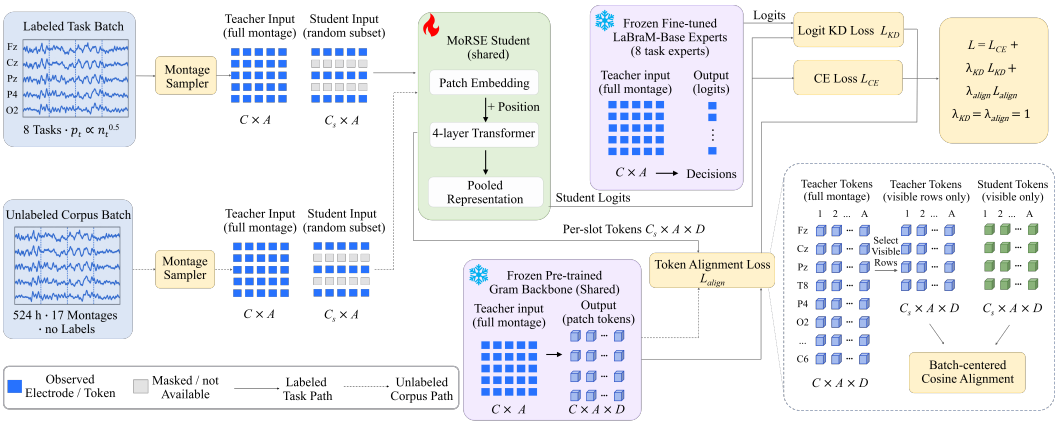

Figure 2: **Overview of MoRSE training.** The teachers see the full montage, whereas the student sees a random electrode subset resampled at every step. Fine-tuned LaBraM-Base experts supply logits and the never-fine-tuned Gram backbone supplies per-patch tokens. Solid arrows mark the labeled path, which carries all three losses, and dashed arrows the unlabeled corpus, which carries only $\mathcal{L}_{\text{align}}$ (Equation 12). **Inset:** the slot restriction of Equation 7.

- **Montage-robust distillation.** We propose MoRSE, which compresses fine-tuned EEG foundation models into one shared encoder with a linear head per task. The teachers always see the full montage while the student sees a random electrode subset resampled at every step, so distillation becomes inference of the full-montage representation from the electrodes available. MoRSE reuses the backbones the field already has rather than pre-training a new one (Section 2).
- **Teacher roles from a representation diagnosis.** We find that task fine-tuning keeps the decision a task needs and typically collapses the per-patch structure a representation target has to carry, and we turn that finding into the assignment of the two teachers. The logits come from the fine-tuned experts, the per-patch targets from a backbone that was never fine-tuned. To our knowledge this consequence of fine-tuning has not been drawn before (Section 2.2).
- **One encoder in place of 30 specialists.** We evaluate MoRSE on eight tasks, 30 task-budget cells, 42 held-out task-layout pairs and two external cohorts. One 1.97M encoder covers all of them, staying within 0.0291 balanced accuracy of the per-cell specialists it replaces while deploying 89× fewer parameters, and it gains +0.0528 over dropping channels on layouts never drawn during training. It trails eight montage-sampled specialists with 24× the parameters by only 0.0255, so one encoder carries the robustness of eight (Sections 3.2 and 3.3).

2 METHOD

One training step of MoRSE is illustrated in Figure 2. A recording window arrives as $\mathbb{R}^{C\times T}$ at 200 Hz, is scaled by $1/100$ and reshaped into one-second patches, giving $x \in \mathbb{R}^{C\times A\times 200}$ with C electrodes and $A = \lfloor T/200 \rfloor$ patches. Its electrode set is $\mathcal{C}_x = \{c_1, \ldots, c_C\} \subseteq \mathcal{V}$, where $\mathcal{V}$ is the electrode vocabulary of the encoder (the named scalp sites it holds an embedding for), and $\iota(c) \in \{1, \ldots, |\mathcal{V}|\}$ is the index of electrode c in it. An electrode subset is $S \subseteq [C]$ of size $C_s = |S|$, the input restricted to it is $x_S \in \mathbb{R}^{C_s\times A\times 200}$, the full montage is $x_F = x_{[C]}$, and $\mathcal{S}(S) = \{(c, a) : c \in S,\ a \in [A]\}$ is the set of (electrode, patch) slots that subset exposes. Task t has $C_y^{(t)}$ classes (written C_y where a single task is meant) and n_t training windows. Three models appear throughout, namely the student f_θ, the frozen fine-tuned expert E_t of task t, and a frozen backbone B that is never fine-tuned. Only f_θ is trained, and E_t and B supply its targets.

2.1 STUDENT ENCODER

The student is the NeuralTransformer of LaBraM (Jiang et al., 2024) at width $D = 200$, $L = 4$ blocks and ten heads, 1.97M parameters including one linear head per task. Each (electrode,

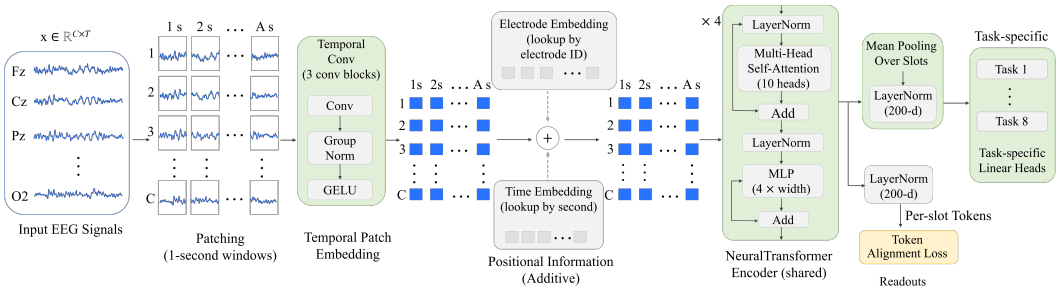

Figure 3: **The student encoder.** A shared encoder and a linear head per task, 1.97M parameters with 5.8k in the heads. Pooled tokens feed the heads and per-slot tokens the alignment loss.

patch) segment is embedded independently by a temporal convolution stack ϕ of three 1-D-convolution/GroupNorm/GELU blocks, and the result is summed with two learned lookup tables, an electrode embedding $p \in \mathbb{R}^{|\mathcal{V}|\times D}$ indexed by electrode identity and a time embedding $q \in \mathbb{R}^{A_{\max}\times D}$ indexed by position in time. The token at slot (c, a) is

$$h_{c,a} \;=\; \phi\big(x[c,a]\big) \;+\; p\big[\iota(c)\big] \;+\; q[a] \;\in\; \mathbb{R}^D, \qquad c \in S,\; a \in [A], \tag{1}$$

so a window seen on subset S yields $C_s A$ tokens. One encoder spans heterogeneous montages because the electrode embedding $p[\iota(c)]$ in Equation 1 is gathered by identity $\iota(c)$, not by the row of c in the input tensor. The encoder is therefore defined for every S with $\mathcal{C}_{x_S} \subseteq \mathcal{V}$ and every $A \le A_{\max}$ without architectural change (here $C \in [4, 66]$ and $A \in [4, 15]$), while an electrode outside $\mathcal{V}$ has to be mapped to the nearest site in $\mathcal{V}$ by 3-D distance.

The tokens pass through L pre-norm transformer blocks Enc with MLP ratio four. A `[CLS]` token is prepended and discarded at readout, so both readouts are taken over the $C_s A$ slot tokens of the same forward pass:

$$z_\theta \;=\; \mathrm{LN}\big(\mathrm{Enc}(h)\big) \;\in\; \mathbb{R}^{C_s A\times D}, \tag{2}$$

$$s_\theta \;=\; W_t\,\mathrm{LN}\Big(\tfrac{1}{C_s A}\textstyle\sum_{c\in S}\sum_{a\in[A]}\mathrm{Enc}(h)_{c,a}\Big) \;\in\; \mathbb{R}^{C_y^{(t)}}, \tag{3}$$

where W_t is the linear head of task t. Equation 2 is the per-slot representation the alignment term consumes and Equation 3 the logits the classification and distillation terms consume. They are not views of one vector, since layer normalization does not commute with averaging. Figure 3 shows both paths. Weights are initialized from LaBraM-Base, whose blocks 0, 4, 7 and 11, spread across its depth, become the four student blocks.

2.2 Teacher assignment

The two teacher slots are assigned by measuring what fine-tuning does to a representation. For a representation matrix $Z \in \mathbb{R}^{n\times D}$ over n windows, let $\sigma_1 \ge \cdots \ge \sigma_D$ be the singular values of the column-centered Z and

$$\mathrm{erank}(Z) \;=\; \exp\Big(-\sum_{i=1}^{D} p_i \log p_i\Big), \qquad p_i \;=\; \frac{\sigma_i^2}{\sum_{j=1}^{D}\sigma_j^2} \tag{4}$$

be its entropy-based effective rank (Roy & Vetterli, 2007), which equals D for an isotropic representation and 1 for a rank-one representation and needs no variance cutoff. Let ρ^{ft} denote the effective rank of the pooled representation of a backbone after fine-tuning on a C_y-way task, and $\kappa = \rho^{\mathrm{ft}}/(C_y - 1)$ its ratio to the $C_y - 1$ directions spanned by a linear head, so that κ near unity indicates a representation no larger than the task decision requires. On the seven tasks the backbone fits, fine-tuning compresses its representation to an effective dimensionality that tracks the class count, with $r = +0.94$ and κ within 0.9–2.5, while SEED-V, poorly fit, sits at 3.6. The per-patch effective rank falls further still (Appendix F). Such a representation still carries the task decision but is a poor source of features. The two distillation targets therefore come from different models, and alignment acts per slot because pooling discards most of the remaining rank.

Because Equation 10 is a slot-wise cosine, the alignment teacher B must emit one token per (electrode, patch) with the width D of the student. BENDR and EEGPT keep no per-electrode axis and are therefore excluded (Appendix D), whereas a width mismatch is resolved by widening the student rather than by changing the term (Appendix G.4). Among admissible models the choice matters, and B is a frozen Gram backbone (Li et al., 2026b). The logit teachers E_t only need to produce calibrated logits, and among such models the choice does not measurably matter, so their family is not a hyperparameter of the method. We use the eight full-montage LaBraM-Base specialists that the per-cell baseline of Table 1 already produces. Appendix G.1 measures both slots.

2.3 MONTAGE SAMPLING

At every step the electrode subset of the student is redrawn in two stages, size then membership:

$$k \;\sim\; \mathcal{U}\big\{\, k_{\min}(C),\, \ldots,\, C \,\big\}, \qquad S \mid k \;\sim\; \mathcal{U}\Big(\big\{\, S \subseteq [C] \;:\; |S| = k \,\big\}\Big), \tag{5}$$

where $k_{\min}(C) = \min\big(4,\, \max(2,\, \lfloor C/2 \rfloor)\big)$, so every admissible size is equally likely and membership is uniform over k-subsets without replacement. The student receives x_S, while both teachers keep x_F. This asymmetry is the method, and Figure 2 shows one step of it. It is an instance of learning under privileged information (Lopez-Paz et al., 2016; Xu et al., 2020) in which the privileged channel is the montage itself. At $S = [C]$ the alignment term of Equation 10 asks the student to reproduce a representation it already has the inputs for, whereas at $S \subsetneq [C]$ the same term asks it to infer the contribution of the missing electrodes, which is the axis the evaluation measures.

Because both models emit one token per slot, the alignment term reaches the fraction C_s/C of the token field of the teacher, 17.4% with four of twenty-three electrodes and 6.5% with four of the sixty-two of SEED-V. Two properties of Equation 5 were tested rather than assumed. The size is drawn as well as the membership so that one student covers every budget. Fixing it at $k = 8$ costs 0.0027 and is not superior even at a test-time budget of eight (Appendix I.2). The draw is also spatially unbiased, whereas deployment layouts are clustered, a mismatch quantified in Section 3.3 and examined in Appendix I. Over one run the student sees 10,019 distinct subsets across the eight tasks, where a per-cell specialist sees one. Tasks are drawn from a third distribution, flattened because the natural proportions are severely imbalanced:

$$\Pr[\,t\,] \;=\; \frac{n_t^{\alpha}}{\sum_{t'} n_{t'}^{\alpha}}. \tag{6}$$

We set $\alpha = 0.5$ without tuning, the midpoint between two unacceptable endpoints. HMC contributes 195,076 of the 307,103 training windows (Appendix B) and BCI IV 2a 2,880, so $\alpha = 1$ starves the small tasks of gradient while $\alpha = 0$ presents each BCI IV 2a window sixty-eight times as often as each HMC window.

2.4 TRAINING OBJECTIVE

Teacher inputs. B was pre-trained on average-referenced recordings, so it reads the referenced window and its tokens are then restricted to the slots the student saw:

$$\tilde{x}_F[c] \;=\; x_F[c] \;-\; \frac{1}{C}\sum_{c'=1}^{C} x_F[c'], \qquad z_B \;=\; B\big(\tilde{x}_F\big)\big|_{\mathcal{S}(S)} \;\in\; \mathbb{R}^{|\mathcal{S}(S)|\times D}. \tag{7}$$

The reference is taken over the full montage and never over S, because referencing the subset would make the target of a fixed window drift with the draw. The restriction is one-directional. z_B is the field of the teacher restricted to $\mathcal{S}(S)$, while z_θ is computed from x_S and is never selected out of that field. The expert reads the full montage unreferenced, $s_t = E_t(x_F)$.

Losses. Let $\bar{u}_n = u_n - \frac{1}{N}\sum_{m=1}^{N} u_m$ denote centering over the joint batch-and-slot axis of N tokens. With label smoothing ε and temperature T, the three terms are

$$\mathcal{L}_{\mathrm{CE}} = \mathrm{CE}_{\varepsilon}\big(y,\, s_\theta\big), \tag{8}$$

$$\mathcal{L}_{\mathrm{KD}} = T^2\,\mathrm{KL}\Big(\mathrm{softmax}(s_t/T)\,\Big\|\,\mathrm{softmax}(s_\theta/T)\Big), \tag{9}$$

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{|\mathcal{S}(S)|}\sum_{(c,a)\in\mathcal{S}(S)} \Big(1 - \cos\big(\bar{z}_\theta[c,a],\, \bar{z}_B[c,a]\big)\Big), \tag{10}$$

and the objective minimized is

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{CE}} \;+\; \lambda_{\mathrm{KD}}\,\mathcal{L}_{\mathrm{KD}} \;+\; \lambda_{\mathrm{align}}\,\mathcal{L}_{\mathrm{align}}. \tag{11}$$

Both weights are set to 1 without search, with $T = 2$ and $\varepsilon = 0.1$. The T^2 factor keeps the gradient scale of Equation 9 comparable to that of Equation 8 as T varies. The centering in Equation 10 is necessary. The per-slot tokens that the term compares have a cosine of 0.51 between unrelated windows and 0.26 once one shared mean direction is removed, so without $\bar{\cdot}$ the term is near-satisfied at initialization and carries almost no gradient. The failure is silent, since the loss still appears to decrease (Appendix I.2). Every labeled step is followed by one step on an unlabeled corpus of eleven other datasets (Appendix B.2). Such a window carries no label and has no expert, so $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{L}_{\mathrm{KD}}$ are undefined on it and the step minimizes

$$\mathcal{L}^{\mathrm{corpus}} \;=\; \lambda_{\mathrm{align}}\,\mathcal{L}_{\mathrm{align}}, \tag{12}$$

the same term of Equation 10 rather than a fourth loss, along the dashed path of Figure 2.

3 Experiments

3.1 Experimental setup

Tasks and budgets. We evaluate on eight cross-subject tasks: TUEV (Obeid & Picone, 2016) (event type), HMC (Alvarez-Estevez & Rijsman, 2022) (sleep staging), SEED-V (Liu et al., 2022) (emotion), BCI IV 2a (Tangermann et al., 2012) (BCI, motor imagery), CHB-MIT (Shoeb, 2009) (CHB, seizure), MentalArithmetic (Zyma et al., 2019) (MentalArith or Ment., cognitive load), Mumtaz2016 (Mumtaz et al., 2017) (Mum., depression) and ADFTD (Miltiadous et al., 2023) (dementia diagnosis). Electrode budgets are fixed subsets of full/16/8/4 for each task except HMC, which has four electrodes and is evaluated at full/2, giving 30 cells (Appendix B.1).

Preprocessing. For every model, signals are band-pass filtered at 0.1–75 Hz and resampled to 200 Hz as in LaBraM, with the notch filter placed at the mains frequency of the country where each dataset was recorded (Appendix D.5). Appendix B.2 lists the full training configuration.

Baselines and metrics. Table 1 compares the student against a LaBraM-Base specialist fine-tuned on each cell, whose full-montage instance is the *ceiling* of the task, and Appendix D.1 interpolates the reduced cells with spherical splines (Perrin et al., 1989) for that instance. Table 2 compares three families retrained under our splits and preprocessing (Appendix B.3). The non-foundation baselines are the classical pipelines and EEGNet (Lawhern et al., 2018) released by Yang et al. (2026), and five neural models run using the BIOT code (Yang et al., 2023a): SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2023b), CNN-Transformer (Peh et al., 2022), FFCL (Li et al., 2022) and ST-Transformer (Song et al., 2021). The foundation-model baselines are LaBraM-Base (Jiang et al., 2024), Gram (Li et al., 2026b), CBraMod (Wang et al., 2025) and REVE-Base (El Ouahidi et al., 2025), each fine-tuned per task, and the third family, MoRSE$\langle X \rangle$, builds all three models from one backbone X. Only CBraMod pre-trains on one of our tasks, TUEV, and that entry is marked † in Table 2 (Appendix B.4). All accuracies are balanced accuracy, the mean per-class recall.

Reported quantities. Two GPU runs with the same seed can differ by about 0.02 balanced accuracy on a task. Unless stated otherwise, comparisons therefore use the same five seeds, 0–4, and differences between arms are paired by seed (Equation 13), which removes the variance the arms share. A paired difference $\bar{\delta}$ is reported with its standard deviation over seeds, and its ratio to the standard error as t (Appendix C). In every table and figure, $\pm$ denotes the seed standard deviation, Δ a seed-paired difference between two arms, and **bold** the unique best entry.

3.2 One student versus 30 specialists

Comparison with the specialists. As Table 1 shows, a single 1.97M encoder covers all 30 cells with 89× fewer deployed parameters than the per-cell specialists, 175M across thirty copies of 5.82M. It trails them by 0.0291 balanced accuracy overall, and by less once BCI IV 2a is excluded (0.6380 against 0.6522 over 26 cells, $\Delta = -0.0142$), a task on which every specialist trails a 5.2k-parameter EEGNet (Appendix D.3). The gap narrows as electrodes are removed, from -0.040 at the full montage to -0.006 at four electrodes (Figure 4), so the single model is closest to the specialists at

Table 1: **One shared 1.97M student replaces 30 per-cell specialists.** Test-set balanced accuracy averaged over the budgets of each task. The mean covers all 30 cells.

Models	Mean	TUEV	HMC	SEED-V	BCI	CHB	Ment.	Mum.	ADFTD
MoRSE (ours)	0.6044	0.5765	0.6606	0.2758	0.3862	**0.9100**	0.6121	0.8960	0.5463
LaBraM specialists	0.6335	0.6290	**0.7010**	0.3065	0.5118	0.7870	0.6490	**0.9173**	**0.6000**
Gram specialists	**0.6395**	**0.6526**	0.6802	**0.3173**	**0.5126**	0.8641	**0.6524**	0.9055	0.5515
Δ vs LaBraM	-0.0291	-0.0525	-0.0404	-0.0308	-0.1256	**+0.1230**	-0.0370	-0.0214	-0.0537
Δ vs Gram	-0.0351	-0.0761	-0.0197	-0.0415	-0.1264	**+0.0459**	-0.0403	-0.0095	-0.0052

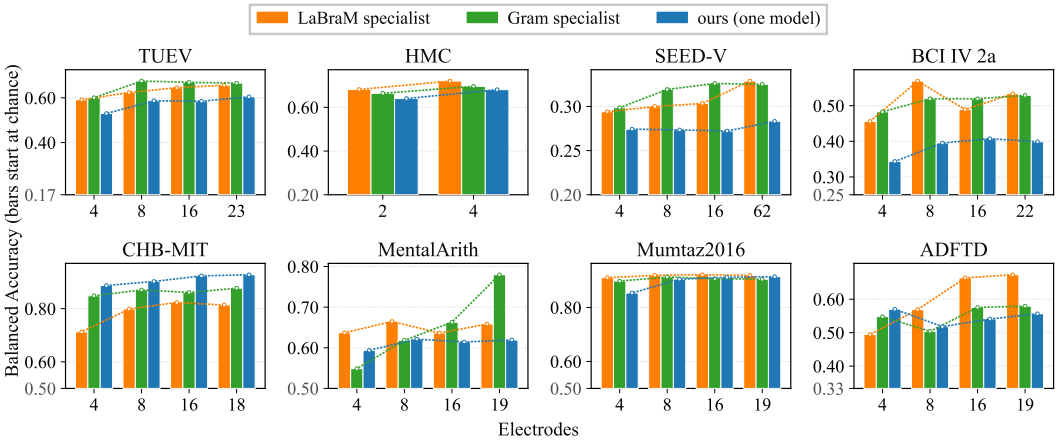

Figure 4: **Balanced accuracy across electrode budgets.** Blue denotes the shared student, and orange and green the LaBraM and Gram specialists. Panels start at task-specific chance.

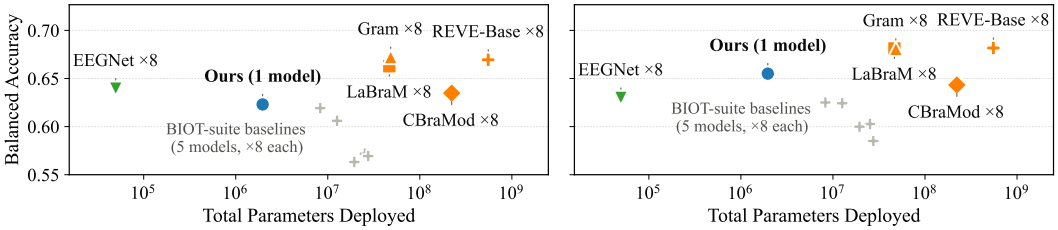

Figure 5: **Deployment cost versus full-montage balanced accuracy. Left:** eight tasks. **Right:** excluding BCI IV 2a. Gray crosses are BIOT-suite baselines, orange foundation-model specialists.

the smallest budget. It leads in 5 of the 30 cells, including all four CHB-MIT cells. CHB-MIT has only 3,006 training windows, and the benefit of sharing the encoder grows as training data shrink (Appendix G.5). A second specialist family, Gram fine-tuned at all 30 cells, gives a stronger column at 0.6395. Against it the gap widens to -0.0351 while the parameter reduction grows to $92\times$, yet the student leads on 10 cells rather than 5 (Appendix D.2). Fine-tuning one LaBraM specialist per task with the same sampler narrows the lead of the LaBraM specialists to 0.0100 (0.6144 against 0.6044), while the eight of them carry $24\times$ the parameters of the student (Appendix G.3).

Comparison with all baselines. We compare with every baseline at the full montage, the setting each baseline is built for and our most conservative test (Figure 4), and retrain each baseline under our protocol as one model per task (Table 2 and Figure 5). The student reaches 0.6230. It exceeds the classical pipelines and all five BIOT-suite networks, which deploy four to fourteen times its parameters, and trails EEGNet (0.6394) and the four foundation-model specialists (0.6348 to 0.6723). EEGNet leads only through BCI IV 2a. Without it the student exceeds EEGNet, 0.6551 against 0.6298, and only foundation-model specialists over twenty times larger remain ahead (Figure 5,

Table 2: **Full-montage comparison with all baselines.** Balanced accuracy under our protocol. Each baseline deploys one model per task, whereas each student is one shared encoder. The **mean** covers eight tasks, bold marks the column maximum, and † marks pre-training on that task.

Models	Mean	TUEV	HMC	SEED-V	BCI	CHB	Ment.	Mum.	ADFTD
Classical	0.5393	0.3382	0.5494	0.2532	0.5087	0.7698	0.5254	0.8534	0.5164
EEGNet	0.6394	0.5468	0.6602	0.3167	**0.7068**	0.8395	0.5686	0.9085	0.5685
SPaRCNet	0.6192	0.5131	0.6968	0.3229	0.5781	0.7972	0.5889	**0.9383**	0.5182
ContraWR	0.6060	0.4809	0.7058	0.3253	0.4785	0.7665	0.6225	0.9148	0.5538
CNN-Transf.	0.5719	0.4333	0.6966	0.3279	0.3559	0.7321	0.6343	0.9056	0.4890
FFCL	0.5632	0.4837	0.6984	0.2862	0.3061	0.7512	0.5933	0.9124	0.4743
ST-Transf.	0.5694	0.4308	0.6818	0.2709	0.4601	0.7833	0.5333	0.9198	0.4751
LaBraM-Base	0.6628	0.6561	**0.7202**	0.3288	0.5333	0.8135	0.6584	0.9187	**0.6730**
Gram	**0.6723**	**0.6844**	0.6796	**0.3327**	0.6127	0.8740	**0.7152**	0.9194	0.5605
CBraMod	0.6348	0.5845†	0.6875	0.2609	0.5753	0.8498	0.6194	0.9128	0.5882
REVE-Base	0.6695	0.6764	0.6856	0.3203	0.5839	0.8963	0.7102	0.9283	0.5549
MoRSE⟨LaBraM⟩	0.5993	0.5803	0.6897	0.3226	0.3306	0.8102	0.6095	0.9157	0.5360
MoRSE⟨Gram⟩	0.6140	0.6110	0.6722	0.2960	0.3797	0.8916	0.5698	0.9024	0.5895
MoRSE⟨REVE⟩	0.5810	0.5198	0.6592	0.2551	0.3497	0.8679	0.5943	0.9259	0.4763
MoRSE (ours)	0.6230	0.6044	0.6809	0.2832	0.3988	**0.9274**	0.6190	0.9140	0.5564

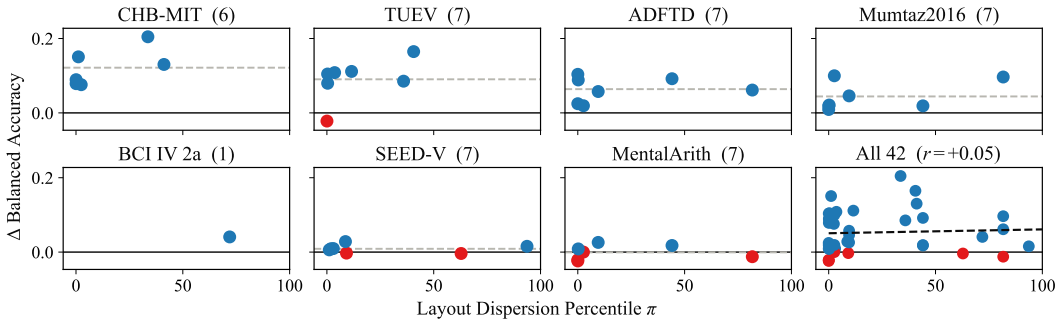

Figure 6: **Montage sampling minus no sampling on all 42 task–layout pairs.** Blue and red mark gains and losses, and dashed gray lines mark task means. HMC has no eligible layout.

right). Each single-backbone instance MoRSE$\langle X \rangle$ trails the specialists of X by 0.058 to 0.089, and MoRSE, whose teachers come from two backbones, outperforms all three instances. Every foundation-model specialist has at least 2.3× the latency and 1.9× the FLOPs of the student. Three compact baselines run faster, but each serves one cell and the student all 30 (Table 11).

3.3 Transfer to unseen layouts and cohorts

Unseen layouts. Deployment layouts are often clustered, as in a frontal headband or a temporal seizure strip, whereas most reduced budgets above are spread like the draws of the sampler. We therefore evaluate seven hardware layouts and keep a (task, layout) pair when the task montage contains at least three electrodes of the layout, which yields 42 pairs (Appendix B.1). The dispersion percentile π (Equation 16) measures how clustered a layout is relative to the subsets the sampler draws, and a frontal layout on a 23-electrode montage has $\pi = 0.1$, so it is more clustered than 99.9% of them. As Figure 6 shows, montage sampling improves the student by +0.0528 balanced accuracy over an otherwise identical student trained without it (0.5670 against 0.5143). The gain appears on 35 of the 42 pairs and on all five paired seeds ($t = 16.77$), and it is uncorrelated with π ($r = +0.05$). The same sampler raises the 5.82M LaBraM-Base specialists by +0.0454, so the robustness comes from the training distribution rather than from the student architecture. These specialists remain ahead of the student by 0.0255 but carry 24× its parameters (Appendix G.3).

External cohorts. We apply the sleep head unchanged to two cohorts from other sites, ISRUC (Khalighi et al., 2016) and the PhysioNet/CinC 2018 challenge set (P2018) (Ghassemi et al., 2018).

Table 3: **Zero-shot transfer to two external sleep cohorts.** Balanced accuracy for MoRSE and its no-sampling ablation. Bold marks the better arm in each row and the largest Δ per cohort.

	ISRUC (82 subj.)			P2018 (150 subj.)		
Electrodes	**MoRSE**	**No sampling**	Δ	**MoRSE**	**No sampling**	Δ
2	**0.6159** $\pm$0.0130	0.4891 $\pm$0.0129	**+0.1267**	**0.6300** $\pm$0.0103	0.5137 $\pm$0.0148	**+0.1163**
4	**0.6753** $\pm$0.0129	0.6736 $\pm$0.0030	+0.0017	**0.6809** $\pm$0.0085	0.6686 $\pm$0.0071	+0.0123
6	**0.6294** $\pm$0.0238	0.6229 $\pm$0.0193	+0.0065	**0.6614** $\pm$0.0212	0.6514 $\pm$0.0164	+0.0100

Table 4: **Distillation and montage sampling work best together in aggregate.** Balanced accuracy, averaged as in Table 1. **Dist.** adds both distillation terms and **Samp.** montage sampling to a cross-entropy student, and Δ_{int} is their interaction (Equation 14).

Dist.	**Samp.**	**Mean**	**TUEV**	**HMC**	**SEED-V**	**BCI**	**CHB**	**Ment.**	**Mum.**	**ADFTD**
		0.5359	0.4385	0.6204	0.2458	0.2882	0.7945	0.5661	0.8767	0.4994
	✓	0.5656	0.5144	0.6453	0.2549	0.3081	0.8815	0.5687	0.8933	0.4987
✓		0.5583	0.4833	0.5886	**0.2802**	**0.3994**	0.7473	0.6098	0.8626	0.5105
✓	✓	**0.6044**	**0.5765**	**0.6606**	0.2758	0.3862	**0.9100**	**0.6121**	**0.8960**	**0.5463**
Δ_{int}		+0.0164	+0.0173	+0.0471	-0.0136	-0.0332	+0.0757	-0.0003	+0.0168	+0.0365

Both use the five AASM stages of HMC, and neither appears in any pre-training list, our distillation corpus or our eight tasks (Appendix E). Each records the four HMC derivations (F4, C4, O2 and C3) plus F3 and O1, and we test all six, these four, and C4 and C3, the two-electrode budget of HMC. As Table 3 shows, the recipe reaches 0.675 on ISRUC and 0.681 on P2018 with the four HMC derivations, against chance at 0.200. With four or six electrodes, which keep the whole training montage of HMC, the two arms differ by at most 0.0123. With two, where F4 and O2 are missing, montage sampling adds $+0.1267$ on ISRUC ($t = 49.39$) and $+0.1163$ on P2018 ($t = 24.08$), positive on all five seeds in both cohorts. Its benefit thus arises where training electrodes are missing and survives a change of site.

3.4 Contribution of each component

To measure the contribution of each component, Table 4 compares four students that share the recipe, tasks and seeds and differ only in distillation, which brings the corpus steps with it, and in montage sampling. Sampling alone adds +0.0297 and distillation alone +0.0224, and the two together +0.0685. Distillation is therefore worth +0.0388 under the sampler rather than the +0.0224 it adds alone, and sampling is worth +0.0461 once distillation is present, giving a paired interaction of $\Delta_{\text{int}} = +0.0164$ ($t = 1.82$ over five seeds, four positive). The interaction is consistent with the asymmetry of Section 2.3, because at a fixed montage the alignment term asks for nothing the student cannot already compute. A loss-level ablation further locates the gain of distillation in representation alignment, worth $+0.0247$ over the supervised baseline without the unlabeled corpus and $+0.0360$ with it, while logit distillation is not separable from zero (Appendix G.2).

4 Conclusion

In this paper, we propose MoRSE, which replaces 30 per-cell EEG specialists with a single 1.97M encoder distilled from frozen EEG foundation models. The teachers see the full montage while the student sees a random electrode subset resampled at every step, so distillation becomes inference of the full-montage representation from the electrodes available. The roles of the teachers follow a representation diagnosis, with logits taken from fine-tuned experts and per-patch representations from a never-fine-tuned backbone. Distillation and montage sampling work best together. The resulting encoder trails the specialists it replaces by 0.0291 balanced accuracy over 30 cells with 89$\times$ fewer deployed parameters, and it gains +0.0528 over dropping channels at test time on hardware layouts it has never seen. We hope this work encourages EEG models to be designed and evaluated for robustness to electrode configurations, advancing deployment on real-world devices.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this manuscript, to support retrieval and discovery of relevant literature, and to assist in editing and debugging software code. All AI-assisted content, retrieved information and code were reviewed and verified by the authors, and the code was checked with the test suite of the released package and against the stored result files. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the implementation details, preprocessing procedures, training configuration, and additional evaluations required to reproduce our results in the main text and appendices. Appendix B gives the datasets, splits, baseline protocol and full training configuration (Table 8), Appendix D.5 the preprocessing, and Appendix C the definition of every reported quantity. Every training run archives its source tree and command line alongside its checkpoint, and all reported results are obtained through a common evaluation pipeline on the full validation and test sets. A provenance record identifies preprocessing generations and implementation revisions so that results from incompatible runs are never pooled. The core implementation of MoRSE, covering the student, the montage sampler, the distillation objective and the permutation-equivariance test of Appendix F.2, is provided in an anonymous repository at `https://anonymous.4open.science/r/MoRSE-CD7C`. All datasets used in this work are publicly available, several of them through PhysioNet (Pollard et al., 2026).

REFERENCES

Salar Abbaspourazad, Anshuman Mishra, Joseph Futoma, Andrew C. Miller, and Ian Shapiro. Wearable accelerometer foundation models for health via knowledge distillation. *arXiv preprint arXiv:2412.11276*, 2024.

Diego Alvarez-Estevez and Roselyne Rijsman. Haaglanden Medisch Centrum sleep staging database (version 1.1). PhysioNet, 2022. doi: 10.13026/t79q-fr32.

Bruno Aristimunha, Dung Truong, Pierre Guetschel, Seyed Yahya Shirazi, Isabelle Guyon, Alexandre R. Franco, Michael P. Milham, Aviv Dotan, Scott Makeig, Alexandre Gramfort, et al. EEG foundation challenge: From cross-task to cross-subject EEG decoding. In *NeurIPS 2025 Competition Track*, 2025.

Hubert Banville, Sean U. N. Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort. Robust learning from corrupted EEG with dynamic spatial filtering. *NeuroImage*, 251: 118994, 2022.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tidiane Camaret Ndir, Robin Tibor Schirrmeister, and Tonio Ball. EEG-CLIP: Learning EEG representations from natural language descriptions. *Frontiers in Robotics and AI*, 12:1625731, 2025.

Gian Emilio Chatrian, Ettore Lettich, and Paula L. Nelson. Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities. *American Journal of EEG Technology*, 25(2):83–92, 1985.

Zhige Chen, Chengxuan Qin, Wenlong You, Rui Liu, Congying Chu, Rui Yang, Kay Chen Tan, and Jibin Wu. HEAR: An EEG foundation model with heterogeneous electrode adaptive representation. *arXiv preprint arXiv:2510.12515*, 2025.

Nicholas Cooper, Lijun Chen, Sailesh Dwivedy, and Danna Gurari. Logit-based losses limit the effectiveness of feature knowledge distillation. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025.

Alexander Craik, Yongtian He, and Jose L. Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, 16(3):031001, 2019.

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 *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7):846, 2020.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems*, 2025.

Matthieu Duvinage, Thierry Castermans, Mathieu Petieau, Thomas Hoellinger, Guy Cheron, and Thierry Dutoit. Performance of the Emotiv Epoc headset for P300-based applications. *BioMedical Engineering OnLine*, 12(1):56, 2013. doi: 10.1186/1475-925X-12-56.

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*, 2025.

Matin Fallahi, Patricia Arias-Cabarcos, and Thorsten Strufe. NeuroShield: A device-agnostic foundation model for EEG authentication. *arXiv preprint arXiv:2606.20673*, 2026.

Navid Mohammadi Foumani, Soheila Ghane, Nam Nguyen, Mahsa Salehi, Geoffrey I. Webb, and Geoffrey Mackellar. EEG-X: Device-agnostic and noise-robust foundation model for EEG. *arXiv preprint arXiv:2511.08861*, 2025.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269, 2020.

Mohammad M. Ghassemi, Benjamin E. Moody, Li-wei H. Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G. Mark, M. Brandon Westover, and Gari D. Clifford. You snooze, you win: The PhysioNet/Computing in Cardiology Challenge 2018. In *Computing in Cardiology*, volume 45, 2018.

Anders Gjølbye, Lina Skerath, William Lehn-Schiøler, Nicolas Langer, and Lars Kai Hansen. SPEED: Scalable preprocessing of EEG data for self-supervised learning. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2024.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Hanning Guo, Hanwen Bi, Farah Abdellatif, Andrei Galbenus, Jon N. Shah, Abigail Morrison, and Jürgen Dammers. Brain-OF: An omnifunctional foundation model for fMRI, EEG and MEG. *arXiv preprint arXiv:2602.23410*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *Advances in Neural Information Processing Systems*, 2022.

Kangwook Jang, Sungnyun Kim, Se-Young Yun, and Hoirin Kim. Recycle-and-distill: Universal compression strategy for transformer-based speech SSL models with attention map reusing and masking distillation. In *Interspeech*, 2023.

Kangwook Jang, Sungnyun Kim, and Hoirin Kim. STAR: Distilling speech temporal relation for lightweight speech self-supervised learning models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Herbert H. Jasper. The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, 10:371–375, 1958.

Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. GraphSleepNet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *International Joint Conference on Artificial Intelligence*, pp. 1324–1330, 2020.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F. Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100 (17):e1750–e1762, 2023.

Stamos Katsigiannis and Naeem Ramzan. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016.

Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

Kuntal Kokate, Bruno Aristimunha, Dung Truong, and Arnaud Delorme. Channel adaptation for EEG foundation models: A systematic benchmark across architectures, tasks, and training regimes. *arXiv preprint arXiv:2604.23091*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Velu Prabhakar Kumaravel, Una Pale, Tomas Teijeiro, Elisabetta Farella, and David Atienza. Knowledge distillation-based channel reduction for wearable EEG applications. *TechRxiv preprint 22651156*, 2023. doi: 10.36227/techrxiv.22651156.v1.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems*, 2022.

Junhong Lai, Jiyu Wei, Lin Yao, and Yueming Wang. A simple review of EEG foundation models: Datasets, advancements and future perspectives. *arXiv preprint arXiv:2504.20069*, 2025.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

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, 2019.

Chenqi Li, Yu Liu, Shuo Zhang, Timothy Denison, and Tingting Zhu. Standing on the shoulders of giants: Rethinking EEG foundation model pretraining via multi-teacher distillation. *arXiv preprint arXiv:2603.04478*, 2026a.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022.

Ziyi Li, Wei-Long Zheng, Jiwen Xu, Yong Lu, and Bao-Liang Lu. Gram: A large general EEG model for raw data classification and restoration. *IEEE Transactions on Affective Computing*, 17 (1):841–855, 2026b.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG-FM-Compass: Progress, benchmarking, and future directions for EEG foundation models. *National Science Review*, 2026a. doi: 10.1093/nsr/nwag466.

Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, Yuheng Chen, Shaojie Hou, Xiaolian Zhu, and Dongrui Wu. MIRepNet: A pipeline and pre-trained model for EEG-based motor imagery classification. *Knowledge-Based Systems*, 343:115966, 2026b.

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*, 11:131, 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.

Yuan Liu, Zhuolan Gui, De Yan, Zhuang Wang, Ruisi Gao, Ningxin Han, Junying Chen, Jialing Wu, and Dong Ming. Lower limb motor imagery EEG dataset based on the multi-paradigm and longitudinal-training of stroke patients. *Scientific Data*, 12:314, 2025.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1–R13, 2007.

Jun Ma and Tuukka Ruotsalo. Adapting frozen foundation models for montage-agnostic high-resolution EEG event segmentation. *Journal of Neural Engineering*, 23(3):036020, 2026. doi: 10.1088/1741-2552/ae6142.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Evripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8 (6):95, 2023.

Wajid Mumtaz, Likun Xia, Mohd Azhar Mohd Yasin, Syed Saad Azhar Ali, and Aamir Saeed Malik. A wavelet-based technique to predict treatment outcome for major depressive disorder. *PLOS ONE*, 12(2):e0171409, 2017.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

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

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp EEG. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602, 2022.

F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2):184–187, 1989.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Arman Savran, Koray Çiftçi, Guillaume Chanel, Javier Cruz Mota, Luong Hong Viet, Bülent Sankur, Lale Akarun, Alice Caplier, and Michèle Rombaut. Emotion detection in the loop from brain signals and facial images. In *eNTERFACE 2006 Workshop*, 2006.

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

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

Margitta Seeck, Laurent Koessler, Thomas Bast, Frans Leijten, Christoph Michel, Christoph Baumgartner, Bin He, and Sándor Beniczky. The standardized EEG electrode array of the IFCN. *Clinical Neurophysiology*, 128(10):2070–2077, 2017.

Ali H. Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Urban Širca, Maryam Alimardani, Stefanos Zafeiriou, and Konstantinos Barmpas. Beyond accuracy: Robustness, interpretability and expressiveness of EEG foundation models. *arXiv preprint arXiv:2605.17562*, 2026.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2020.

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

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *International Conference on Learning Representations*, 2024.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740, 2024.

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

Toyotaro Suzumura, Hiroki Kanezashi, and Shotaro Akahori. Graph adapter of EEG foundation models for parameter efficient fine tuning. In *AAAI 2025 Workshop on Health Intelligence*, 2025.

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

Mario Giovanni Terzano, Liborio Parrino, Adriano Sherieri, Ronald Chervin, Sudhansu Chokroverty, Christian Guilleminault, Max Hirshkowitz, Mark Mahowald, Harvey Moldofsky, Agostino Rosa, et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep Medicine*, 2(6):537–553, 2001.

Rahul Thapa, Bryan He, Magnus Ruud Kjær, Hyatt Moore IV, Gauri Ganjoo, Emmanuel Mignot, and James Zou. SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. In *International Conference on Machine Learning*, pp. 48019–48037, 2024.

Logan Trujillo. Raw EEG data. Texas Data Repository, 2020. doi: 10.18738/T8/SS2NHB.

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*, 2024.

Jingyuan Wang, Zhihao Jia, Chenyu Liu, Xinliang Zhou, Haoran Luo, Ziyu Jia, Yong Li, Fang Li, Junfeng Yao, and Yi Ding. DLink: Distilling layer-wise and dominant knowledge from EEG foundation models. *arXiv preprint arXiv:2604.15016*, 2026.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998, 2022.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *Advances in Neural Information Processing Systems*, 2025.

Wei Xiong, Junming Lin, Jiangtong Li, Jie Li, and Changjun Jiang. ALFEE: Adaptive large foundation model for EEG representation. *arXiv preprint arXiv:2505.06291*, 2025.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. In *International Conference on Machine Learning*, 2026.

Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. Privileged features distillation at Taobao recommendations. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

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*, 2023a.

Chaoqi Yang, Cao Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: Model development and evaluation study. *JMIR AI*, 2:e46769, 2023b.

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

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European Conference on Computer Vision*, 2022.

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

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

Marzieh Zare. A negative-control protocol for clinical EEG foundation-model benchmarks: Dataset identity and external-cohort stress testing. *arXiv preprint arXiv:2607.24519*, 2026.

Chao Zhang, Yiqiao Liao, Siqi Han, Milin Zhang, Zhihua Wang, and Xiang Xie. Multichannel multidomain-based knowledge distillation algorithm for sleep staging with single-channel EEG. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(11):4608–4612, 2022.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems*, 2025.

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

A Related work

Task-specific EEG decoders. Before foundation models, EEG decoding progressed through architecture families trained separately for each dataset. Classical pipelines paired hand-designed spatial and spectral features with linear classifiers (Lotte et al., 2007). Convolutional networks subsequently learned these filters directly. Examples include the deep and shallow ConvNets of Schirrmeister et al. (2017), the compact EEGNet (Lawhern et al., 2018) for motor imagery and event-related potential classification, and recurrent-convolutional models such as DeepSleepNet (Supratak et al., 2017) for single-channel sleep staging. Craik et al. (2019) survey these deep-learning decoders. Graph networks then modeled inter-channel connectivity for emotion recognition (Song et al., 2020) and sleep staging (Jia et al., 2020), while EEG Conformer (Song et al., 2023) combined convolutional front ends with temporal self-attention. These models assume one dataset and one electrode layout. Foundation models seek to remove that assumption. The compact size of task-specific decoders nevertheless makes them an essential reference for any compression claim, and Yang et al. (2026) find that EEG foundation models often fail to significantly outperform compact neural networks in data-scarce settings. We independently reproduce that comparison with their released EEGNet implementation and hyperparameters. Appendix D.3 reports the tasks on which the conclusion applies to MoRSE.

EEG foundation models. BENDR (Kostas et al., 2021) adapts wav2vec-style contrastive pre-training to EEG, and BIOT (Yang et al., 2023a) tokenizes heterogeneous biosignals into a shared sequence. LaBraM (Jiang et al., 2024) represents one-second, single-channel patches with a learned vector-quantized codebook and performs masked prediction with a transformer whose positional vocabulary contains 128 electrode entries. CBraMod (Wang et al., 2025) replaces full attention with a criss-cross factorization over the channel-by-patch grid, while Gram (Li et al., 2026b) adds multi-scale processing and an explicit channel embedding. CSBrain (Zhou et al., 2025) tokenizes EEG across several temporal and anatomical scales. The family now includes Neuro-GPT (Cui et al., 2024), ALFEE (Xiong et al., 2025), MIRepNet (Liu et al., 2026b) for motor imagery, multimodal systems such as SleepFM (Thapa et al., 2024), and language-supervised models such as EEG-CLIP (Camaret Ndir et al., 2025). EEGPT (Wang et al., 2024) uses a masked dual self-supervised objective and aligns spatio-temporal representations rather than raw signals. It is excluded from our teacher slots for the architectural reason given in Appendix D.5. Self-supervised EEG models have also been applied to image decoding (Song et al., 2024), and a challenge and a benchmark provide standardized evaluation of foundation models (Aristimunha et al., 2025; Xiong et al., 2026). Broader surveys are provided by Lai et al. (2025) and Liu et al. (2026a), the latter covering 55 models. Recent models increase corpus scale and sensor heterogeneity simultaneously. BrainOmni (Xiao et al., 2025) quantizes EEG and MEG within one discrete vocabulary, and Brain-OF (Guo et al., 2026) extends this scope to fMRI. REVE (El Ouahidi et al., 2025) pre-trains on 25,000 subjects from 92 datasets and uses a 4D positional encoding for arbitrary electrode arrangements.

Montage heterogeneity. These models provide strong evidence that pre-training can accommodate montage heterogeneity. Downstream deployment nevertheless requires task adaptation and must satisfy device constraints. Most models are fully fine-tuned on a target montage, producing one adapted copy for every task-layout pair. Parameter-efficient alternatives (Suzumura et al., 2025) reduce the adaptation cost but retain this one-copy-per-pair structure. Architectural methods build invariance into the encoder. Yi et al. (2023) learn topology-agnostic representations through a unified topology and shared latent space. LUNA (Döner et al., 2025) projects a variable number of channels onto fixed learned queries through cross-attention, making channel processing permutation invariant and decoupling computation from electrode count. EEG-X (Foumani et al., 2025) embeds electrode locations rather than electrode indices, while HEAR (Chen et al., 2025) and NeuroShield (Fallahi et al., 2026) address heterogeneous electrode layouts. NeuroShield considers authentication rather than decoding. These methods obtain invariance from the architecture and then fine-tune per task. MoRSE obtains it from the training distribution and preserves it during adaptation. A direct comparison would require pre-training each architecture with its own corpus and objective, whereas our scope is post-hoc compression of existing backbones. Input-side methods reconstruct a canonical montage by interpolation or zero-filling before inference. Kokate et al. (2026) evaluate four such methods across five backbones and five tasks and find that the best one depends on the backbone. We include one of them, spherical-spline interpolation (Perrin et al., 1989), as a training-free baseline in

Appendix D.1. Ma & Ruotsalo (2026) interpolate channel embeddings from electrode coordinates so that a frozen foundation model accepts any montage, training only that layer and a small segmentation head per dataset. Their adapted system still carries the frozen backbone at inference and one head per dataset, whereas we compress the adapted models themselves. Dynamic spatial filtering (Banville et al., 2022) learns an attention-based filter that preserves a task-specific model under channel corruption or loss. MoRSE instead compresses multiple fine-tuned foundation models into one multi-task student. Širca et al. (2026) further argue that evaluation of EEG foundation models places excessive emphasis on clean in-distribution accuracy. The held-out layouts of Section 3.3 directly test this limitation.

Distillation and multi-task learning. Logit distillation (Hinton et al., 2015) and its refinements (Zhao et al., 2022; Huang et al., 2022; Sun et al., 2024; Beyer et al., 2022), together with feature-based methods (Yang et al., 2022; Cooper et al., 2025), are established techniques, and Gou et al. (2021) provide a survey. Giving the teachers the full montage while the student sees a subset is an instance of learning with privileged information (Lopez-Paz et al., 2016; Xu et al., 2020), applied to the montage rather than to auxiliary features. Within EEG, DLink (Wang et al., 2026) distills foundation models layer by layer into compact students, and Li et al. (2026a) combine several non-EEG teachers before distillation into an EEG student. In both, teacher and student see the same electrodes. Related work has also distilled biosignal students across modalities (Abbaspourazad et al., 2024) and compressed self-supervised speech encoders (Jang et al., 2023; 2024). The closest objectives are those of Kumaravel et al. (2023), who distill an all-channel teacher into a reduced-channel wearable EEG student, and Zhang et al. (2022), who distill a multi-channel sleep model into a single-channel model. Both methods fix the channel set of the student before training and neither uses a foundation model. They therefore provide robustness to one predefined montage. MoRSE resamples the subset at every step, and Section 3.4 measures the effect of this change. The student shares one encoder across eight task heads and therefore inherits the standard problem of conflicting task gradients (Yu et al., 2020; Liu et al., 2021). We sum the per-task losses because Kurin et al. (2022) find unitary scalarization competitive with specialized methods after matching regularization. Appendix G.5 measures the benefit of sharing the encoder relative to training the tasks separately.

B Datasets, splits, baselines and training configuration

B.1 Downstream tasks and held-out layouts

The eight tasks, summarized in Table 5, were chosen to span the conditions that a claim of montage robustness must cover rather than to form a convenient subset. They range from two to six classes, from four to sixty-two electrodes, from four- to fifteen-second windows, and from 1,317 training windows on MentalArithmetic to 195,076 on HMC. This factor of 148 in training data motivates flattening the task distribution (Equation 6) and is also the axis along which the value of a shared encoder varies (Appendix G.5). All splits are cross-subject, and subject and recording disjointness is verified for every split. The CHB-MIT cases chb01 and chb21 come from the same patient, recorded 1.5 years apart, and are therefore treated as one subject. Each ceiling is a LaBraM-Base model fine-tuned per task on the full montage, with a learning rate swept and selected by validation balanced accuracy, and is therefore the same quantity as the full-montage LaBraM specialist of Figure 4. Test-subject counts are derived from the splits, except for TUEV, whose preprocessing removes the patient code.

The seven held-out layouts, shown in Figure 7, are each named for the hardware they represent. The two frontal layouts, the posterior layout and the left-hemisphere layout have a dispersion percentile π (Section 3.3) below 10 on every task, whereas the temporal strip and Emotiv 14 are less clustered. The uniform draw shown for comparison is the one of 2000 eight-electrode draws whose dispersion is closest to their median, because a single arbitrary draw can appear clustered by chance. Table 6 reports, for every (task, layout) pair, the number of layout sites that the task montage records. A pair is evaluated only when at least three sites are present, which leaves 42 of the 56 pairs. With two sites, π depends on a single pairwise distance and cannot separate a clustered layout from a spread one, which is the axis against which transfer is reported. HMC, whose four electrodes (F4, C4, O2, C3) cover at most two sites of any layout, supports none, and the entirely sensorimotor montage of BCI IV 2a supports one. Both limits follow from the electrodes each dataset records.

Table 5: **The eight downstream tasks.** Ceilings are LaBraM-Base models fine-tuned per task on the full montage, and split sizes count windows.

Datasets	Classes	Electrodes	Window	Train/val/test	Ceiling	Test subjects
TUEV	6	23	5 s	68,270/15,662/28,305	0.6561	80
HMC	5	4	15 s	195,076/37,914/41,496	0.7202	22
SEED-V	5	62	4 s	18,230/5,469/5,469	0.3288	3
BCI IV 2a	4	22	4 s	2,880/1,152/1,152	0.5333	2
CHB-MIT	2	18	5 s	3,006/598/430	0.8135	4
MentalArithmetic	2	19	4 s	1,317/397/420	0.6584	7
Mumtaz2016	2	19	5 s	9,095/2,682/2,906	0.9187	12
ADFTD	3	19	5 s	9,229/2,581/2,264	0.6730	15

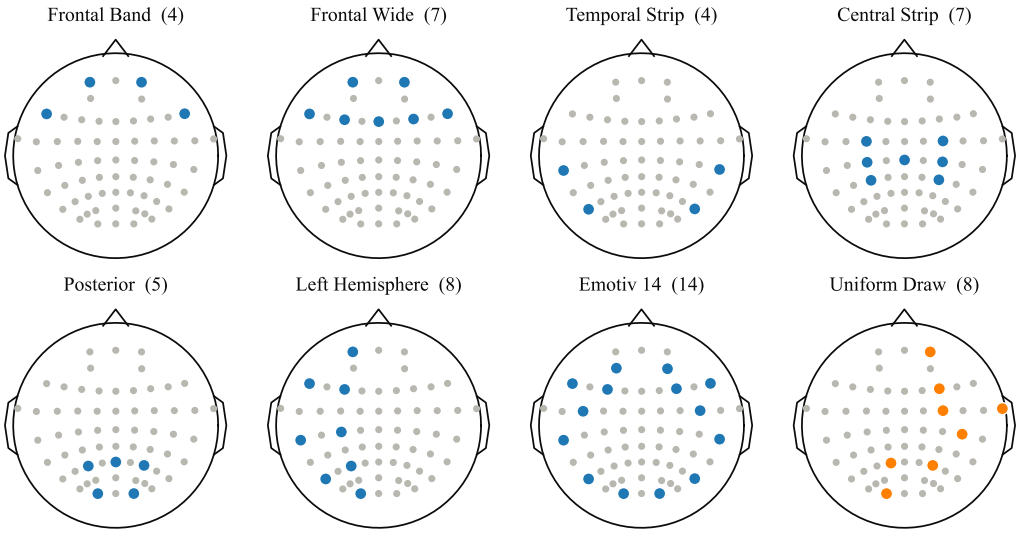

Figure 7: **The seven held-out layouts on the scalp.** Grey dots are the 10–20 sites recorded by any task montage, blue marks the layout, and the number beside each name is its electrode count. The orange panel is a uniform draw of eight electrodes, the kind of subset the sampler trains on.

Table 6: **Held-out layouts supported by each task montage.** Each cell gives the sites found over the sites the layout defines, and **bold** marks the 42 evaluated pairs, which have at least three sites.

Layouts	TUEV	HMC	SEED-V	BCI	CHB	Ment.	Mum.	ADFTD	Tasks
Electrodes	*23*	*4*	*62*	*22*	*18*	*19*	*19*	*19*	
frontal band	**4/4**	0/4	**4/4**	0/4	**4/4**	**4/4**	**4/4**	**4/4**	6
frontal wide	**7/7**	1/7	**7/7**	1/7	**6/7**	**7/7**	**7/7**	**7/7**	6
temporal strip	**4/4**	0/4	**4/4**	0/4	**4/4**	**4/4**	**4/4**	**4/4**	6
central strip	**3/7**	2/7	**7/7**	**7/7**	2/7	**3/7**	**3/7**	**3/7**	6
posterior	**5/5**	1/5	**5/5**	1/5	**4/5**	**5/5**	**5/5**	**5/5**	6
left hemisphere	**8/8**	1/8	**8/8**	1/8	**8/8**	**8/8**	**8/8**	**8/8**	6
Emotiv 14	**10/14**	2/14	**14/14**	0/14	**10/14**	**10/14**	**10/14**	**10/14**	6
Eligible layouts	**7**	**0**	**7**	**1**	**6**	**7**	**7**	**7**	**42**

B.2 Unlabeled corpus and training configuration

On corpus steps, the alignment term uses the 524 hours of unlabeled EEG listed in Table 7. The eleven datasets are Siena (Detti et al., 2020), CAP (Terzano et al., 2001), LowerLimbMI (Liu et al.,

Table 7: **The unlabeled distillation corpus.** Windows are 5 s at 200 Hz, and In pre-training marks the datasets in the pre-training list of LaBraM or Gram.

Datasets	Paradigm	Electrodes	Hours	In pre-training
Siena	seizure	29–31	141.0	yes
CAP	sleep	4–5	72.6	no
LowerLimbMI	lower-limb motor imagery	40–59	56.4	no
OpenBMI	upper-limb motor imagery	62	49.3	no
EEGMMIDB	motor execution & imagery	64	47.6	yes
Texas	cognitive / resting	66	43.0	yes
SEED	emotion	62	42.2	yes
DEAP	emotion	32	33.3	no
DREAMER	emotion (consumer)	14	30.5	no
AcuteStrokeMI	upper-limb motor imagery	30	4.4	no
eNTERFACE	emotion	64	3.7	yes
Total	11 datasets, 17 montages	4–66	524.0	5 of 11

2025), OpenBMI (Lee et al., 2019), EEGMMIDB (Schalk et al., 2004), Texas (Trujillo, 2020), SEED (Zheng & Lu, 2015), DEAP (Koelstra et al., 2012), DREAMER (Katsigiannis & Ramzan, 2018), AcuteStrokeMI (Liu et al., 2024) and eNTERFACE (Savran et al., 2006). The corpus carries no labels and contains none of the eight tasks, so it reaches the student only through Equation 12. An arm trained without the alignment term therefore cannot use it at all, and Appendix G.2 separates the contribution of the corpus from that of the alignment term. A corpus ratio of 1 specifies one unlabeled corpus batch for every labeled batch, so each epoch uses 600 batches of each type, as recorded in the training log. The corpus is deliberately not disjoint from the pre-training lists of LaBraM and Gram, because its role is to supply montage variety to the alignment term rather than new content. No dementia recording appears in it, so ADFTD remains a clean test of transfer.

Every reported number comes from the configuration in Table 8. Each row is read from the default arguments of the training script and then overridden by the command line archived with the checkpoint of the run, so a flag that the recipe changes from its default cannot appear with the default value. Neither loss weight was searched, and both are fixed at 1.

B.3 Baseline and specialist protocol

Every row of Table 2 was retrained under our protocol rather than quoted, because published numbers are not comparable to ours. CBraMod, for instance, splits SEED-V by trial rather than by subject and uses one-second windows, both of which favor a large model. The classical pipelines are deterministic, since the random-forest pipeline fixes its random state and CSP with LDA involves no randomness, so their standard deviation is zero by construction. The neural baselines, the LaBraM column and our own row are covered by the default seed protocol. The † in Table 2 marks the one cell whose backbone pre-trained on that task, CBraMod on TUEV (Section 3.1 and Appendix B.4). All generated tables are produced directly from the result files, and no number is transcribed by hand. Specialists use a batch size set per task, from 32 to 512, and two full-montage cells depart from the common recipe, BCI IV 2a with 200 epochs and SEED-V with drop path 0.3 and weight decay 0.1. Specialist cells are scored at the best validation epoch after warm-up. The schedule uses a five-epoch linear ramp, so validation scores within it are not comparable to later ones, and on the 598-window validation set of CHB-MIT the ramp often produces a spurious peak worth up to 0.09 of test balanced accuracy.

B.4 Pre-training overlap with the downstream tasks

We check overlap dataset by dataset against each published pre-training list rather than by corpus family, because the family is not the unit that leaks. LaBraM and Gram both draw on TUH, but on TUAR, TUEP, TUSZ and TUSL and on TUAB and TUSZ respectively, none of which is TUEV. Both include the SEED series (SEED, SEED-IV, SEED-GER and SEED-FRA), which does not contain SEED-V, and both include BCI Competition IV-1, which is not IV-2a. LaBraM states that its

Table 8: **Training configuration for MoRSE and its specialists.** Values are read from the archived configurations of the reported runs.

Block	Settings	Values
Model	student width / depth / heads	200 / 4 / 10
	parameters (deployed)	1.97M
Optimization	optimizer	AdamW
	learning rate	0.0005, cosine to 1e-06
	warmup epochs	3
	batch size	128
	epochs $\times$ steps	30×600
	label smoothing	0.1
	gradient clip	3.0
Objective	KD temperature T	2.0
	λ_{KD} / λ_{align}	1 / 1 (not searched)
Sampling	task sampling	$p_t \propto n_t^{0.5}$
	montage sampling	$k \sim \mathcal{U}\{k_{\min}, C\}$, then uniform k-subset
	corpus ratio	1
Execution	precision	bf16 autocast (training), fp16 autocast (evaluation)
	seeds	5 (all reported stochastic arms)
	checkpoint selection	best mean validation balanced accuracy over the 30 cells
Specialists	architecture	LaBraM-Base, 12 layers, 5.82M
	learning rate	0.0001–0.001 (swept)
	epochs	50–200
	batch size	32–512
	warmup epochs	5
	layer decay	0.65
	drop path	0.1–0.3
	weight decay	0.05–0.1
	checkpoint selection	best post-warmup validation balanced accuracy

Table 9: **Pre-training overlap between the compared backbones and the eight tasks.** Each cell states whether the published pre-training list of the backbone contains the task.

Tasks	LaBraM	Gram	CBraMod	REVE
TUEV	no	no	**yes**	no
HMC	no	no	no	no
SEED-V	no	no	no	no
BCI IV 2a	no	no	no	no
CHB-MIT	no	no	no	no
MentalArithmetic	no	no	no	no
Mumtaz2016	no	no	no	no
ADFTD	no	no	no	no

downstream datasets are excluded from pre-training, and REVE states that it removed the datasets used in its downstream evaluation from its 92-dataset corpus. The exception is CBraMod, which pre-trains on all of TUEG and therefore on TUEV. Its authors acknowledge this and report a variant of CBraMod that excludes TUEV for this reason. Table 9 gives the full check, taken from the dataset inventories in LaBraM (Jiang et al., 2024, Appendix D), Gram (Li et al., 2026b, Table I), CBraMod (Wang et al., 2025) and REVE (El Ouahidi et al., 2025, Appendix). The five non-foundation baselines of Table 2 are trained from scratch per task and cannot leak. The distillation corpus of Table 7 overlaps these lists, but it carries no labels and enters only the alignment term, so the overlap is a disclosure rather than contamination.

Zare (2026) argue that an overlap audit is insufficient on its own, because dataset identity can be decoded perfectly from every encoder they test, and they propose negative controls and external-

cohort stress tests. Our splits are cross-subject within each dataset, so the audit above does not address the confound they describe, and the two external cohorts of Appendix E implement the part of their recommendation that applies to our setting.

C DEFINITIONS OF THE REPORTED QUANTITIES

When comparing two arms, we first average over cells within each seed and then average the paired seed differences:

$$\delta_s = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left(a^{\mathrm{B}}_{g,s} - a^{\mathrm{A}}_{g,s} \right), \qquad \text{reported as } \bar{\delta} \pm \mathrm{sd}_s(\delta_s), \tag{13}$$

where $\mathcal{G}$ is the cell set and $a_{g,s}$ is the balanced accuracy of cell g under seed s. This order makes the error bar a spread across seeds rather than across cells. Pairing also removes variance shared by the two arms. We report the paired statistic $t = \bar{\delta} / (\mathrm{sd}_s(\delta_s)/\sqrt{n})$ over n seeds and the effect size $d = |\bar{\delta}|/\mathrm{sd}_s(\delta_s)$. The former carries the sign of $\bar{\delta}$, measures the resolution of the estimated difference and grows with n. The latter measures the difference in units of seed variability and does not depend on n. Resolving an effect at $|t| = 2$ requires approximately $(2/d)^2$ seeds. Where a comparison is paired over cells rather than seeds, t is computed in the same way over the n per-cell differences.

For a 2×2 design over distillation $i \in \{0, 1\}$ and montage sampling $j \in \{0, 1\}$, let μ_{ij} denote the corresponding cell mean. The two conditional effects and the interaction are

$$\begin{aligned} \text{distillation, sampling on:} \quad & \mu_{11} - \mu_{01}, \\ \text{sampling, distillation on:} \quad & \mu_{11} - \mu_{10}, \\ \text{interaction:} \quad & \Delta_{\mathrm{int}} = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00}), \end{aligned} \tag{14}$$

with each contrast computed per seed and paired according to Equation 13. A zero interaction indicates additive components, while a positive value indicates that each component increases the contribution of the other. The standard deviations of the four rows of Table 4 over their five seed-level 30-cell means are ± 0.0061, ± 0.0093, ± 0.0098 and ± 0.0040 for neither factor, sampling only, distillation only and both factors, respectively.

The interaction estimate illustrates why the seed count matters. On the first one, two, three and five seeds its value is $+0.0173$, $+0.0252$, $+0.0292$ and $+0.0164$. At two and three seeds it exceeds the effect of either component alone on the same seeds, whereas at five seeds it is smaller than both, $+0.0224$ for distillation and $+0.0297$ for sampling. At five seeds the paired statistic is $t = 1.82$, with four of five differences positive, so the estimate is only moderately resolved. This trajectory cautions against ablations on few seeds.

Two aggregates of the accuracy lost under electrode removal use different cell sets and consequently differ by as much as 0.006. The retention curve in Figure 1 covers the seven tasks with a four-electrode budget and compares the full montage with four electrodes. Its reductions are 0.1362 for the fixed-montage arm and 0.0359 for the full recipe. The value of 0.142 in the abstract and introduction, and the corresponding 0.037 for the full recipe, instead average all eight tasks and compare the full montage with the smallest supported budget of each task. The two sets differ only by HMC, which has four electrodes in total and hence a single reduced cell of two electrodes.

Balanced accuracy is not directly comparable across tasks with two and six classes. We therefore report the fraction of the available headroom of a task that the student recovers:

$$\eta_t = \frac{a_t^{\mathrm{student}} - a_t^{\mathrm{chance}}}{a_t^{\mathrm{ceiling}} - a_t^{\mathrm{chance}}}, \qquad a_t^{\mathrm{chance}} = 1/n_t^{\mathrm{classes}}, \tag{15}$$

where a_t^{student} is the full-montage balanced accuracy of the student on task t and a_t^{ceiling} that of the ceiling of the task. A value $\eta_t = 1$ indicates that the student matches the specialist ceiling, and $\eta_t > 1$ indicates that it exceeds the ceiling.

We measure how clustered a deployment layout is relative to the subsets drawn by the sampler of Equation 5. The spread $d(S)$ of a subset is the mean pairwise distance between its electrodes under

standard 10–20 coordinates $x_i \in \mathbb{R}^3$, and $\pi(S)$ locates this spread within the distribution of uniform subsets of equal size, from which the sampler draws once the size is fixed:

$$\begin{aligned} \mathcal{S}_m &= \big\{\, S' \subseteq [C] \;:\; |S'| = m \,\big\}, \\ d(S) &= \binom{|S|}{2}^{-1} \sum_{\{i,j\} \subseteq S} \|x_i - x_j\|_2, \\ \pi(S) &= 100 \cdot \widehat{\Pr}_{S' \sim \mathcal{U}(\mathcal{S}_{|S|})} \big[\, d(S') < d(S) \,\big], \end{aligned} \tag{16}$$

where the probability is estimated from 20,000 draws. A low value of π indicates a layout more spatially clustered than the subsets usually produced by uniform sampling. The reference distribution is the sampler rather than the data, so π describes a layout relative to the training distribution rather than electrode geometry alone.

We compare representations using linear CKA (Kornblith et al., 2019) on column-centered matrices $U \in \mathbb{R}^{n \times D_U}$ and $V \in \mathbb{R}^{n \times D_V}$ whose n rows are matched across the two models, per-patch tokens for the backbones of Appendix G.1 and pooled per-window vectors for the three students of Appendix H:

$$\mathrm{CKA}(U, V) = \frac{\big\|V^\top U\big\|_F^2}{\big\|U^\top U\big\|_F \, \big\|V^\top V\big\|_F}, \tag{17}$$

which is invariant to orthogonal transformations and isotropic scaling. It therefore does not reward two models for sharing only a coordinate basis.

Tables and figures follow three further conventions. Pale blue identifies the main recipe or the target reduced-montage condition, and bold differences mark the focal comparison rather than statistical significance. A per-task column or row averages over the budgets of that task.

D Boundaries of the approach

This appendix documents five boundaries of the approach, three of which qualify the headline comparison. First, spherical-spline interpolation feeding a full-montage specialist is not dominated by our student on a per-cell basis, and the two routes differ mainly in cost and coverage rather than accuracy. Second, Gram, fine-tuned as a second specialist family, is stronger than the LaBraM family on 16 of 30 cells, so the gap against a per-cell oracle over two backbones is -0.0527 rather than -0.0291 against the LaBraM family alone. Third, on small-sample motor imagery a 5.2k-parameter EEGNet outperforms every foundation-model specialist we trained, and this single task reverses the eight-task aggregate. The remaining two boundaries are that deployment cost does not follow parameter count and that the notch frequency has to follow the mains frequency of each recording. The appendix also records three exclusions, and Appendix J states the limitations that these boundaries imply.

D.1 Interpolation baseline

A training-free alternative is to reconstruct the absent channels with spherical splines (Perrin et al., 1989) and pass the result to an existing full-montage specialist. This is one of the four adaptation methods in a recent channel-adaptation benchmark (Kokate et al., 2026). It is also an exception to the five-seed protocol of Section 3.1. Each cell uses the seed-0 specialist of its task, so the student is evaluated on seed 0 rather than as a five-seed mean. Over the 22 reduced cells for which both routes exist, our student averages 0.5957 against 0.5712 for interpolation with a specialist, a margin of +0.0244, and wins 12 cells against 9, with one tie. We do not interpret this as a clear win. The difference between the student and interpolation is -0.1432 on BCI IV 2a at sixteen electrodes and +0.1702 on ADFTD at eight, so neither route dominates per cell. The two routes differ instead in cost and coverage. Interpolation requires a full-montage specialist for every task, so it removes the montage axis but retains the column of T large models that this paper compresses. It also incurs full-montage inference cost on every window. Because attention is quadratic in the number of tokens, interpolating four electrodes up to twenty-three costs roughly 33× the attention of processing the four directly.

Table 10: **Per-cell balanced accuracy of MoRSE and the LaBraM-Base specialists.** Bold marks the cells in which the student is ahead.

Tasks	Electrodes	MoRSE	Specialists	Δ
TUEV	full	0.6044 ±0.0183	0.6561 ±0.0133	-0.0516
TUEV	16	0.5849 ±0.0263	0.6452 ±0.0131	-0.0603
TUEV	8	0.5866 ±0.0255	0.6237 ±0.0200	-0.0371
TUEV	4	0.5299 ±0.0352	0.5910 ±0.0121	-0.0610
HMC	full	0.6809 ±0.0067	0.7202 ±0.0067	-0.0394
HMC	2	0.6403 ±0.0046	0.6818 ±0.0014	-0.0415
SEED-V	full	0.2832 ±0.0051	0.3288 ±0.0083	-0.0456
SEED-V	16	0.2722 ±0.0076	0.3034 ±0.0087	-0.0312
SEED-V	8	0.2735 ±0.0069	0.2999 ±0.0105	-0.0265
SEED-V	4	0.2743 ±0.0098	0.2941 ±0.0158	-0.0198
BCI IV 2a	full	0.3988 ±0.0254	0.5333 ±0.0401	-0.1345
BCI IV 2a	16	0.4073 ±0.0277	0.4891 ±0.0913	-0.0818
BCI IV 2a	8	0.3951 ±0.0320	0.5691 ±0.0115	-0.1740
BCI IV 2a	4	0.3436 ±0.0119	0.4557 ±0.0298	-0.1122
CHB-MIT	**full**	**0.9274** ±0.0095	**0.8135** ±0.0272	**+0.1140**
CHB-MIT	**16**	**0.9233** ±0.0055	**0.8233** ±0.0203	**+0.1000**
CHB-MIT	**8**	**0.9023** ±0.0214	**0.7986** ±0.0279	**+0.1037**
CHB-MIT	**4**	**0.8870** ±0.0133	**0.7126** ±0.0119	**+0.1744**
MentalArith	full	0.6190 ±0.0186	0.6584 ±0.0298	-0.0394
MentalArith	16	0.6143 ±0.0118	0.6362 ±0.0480	-0.0219
MentalArith	8	0.6213 ±0.0104	0.6648 ±0.0188	-0.0435
MentalArith	4	0.5937 ±0.0213	0.6368 ±0.0295	-0.0432
Mumtaz2016	full	0.9140 ±0.0058	0.9187 ±0.0062	-0.0046
Mumtaz2016	16	0.9116 ±0.0065	0.9212 ±0.0069	-0.0096
Mumtaz2016	8	0.9047 ±0.0054	0.9190 ±0.0053	-0.0143
Mumtaz2016	4	0.8535 ±0.0166	0.9104 ±0.0022	-0.0568
ADFTD	full	0.5564 ±0.0190	0.6730 ±0.0158	-0.1167
ADFTD	16	0.5408 ±0.0157	0.6637 ±0.0099	-0.1229
ADFTD	8	0.5182 ±0.0162	0.5689 ±0.0307	-0.0507
ADFTD	**4**	**0.5699** ±0.0135	**0.4945** ±0.0436	**+0.0754**

D.2 A SECOND SPECIALIST FAMILY

Because our alignment teacher is Gram and our logit teachers are LaBraM specialists, the specialist column should not come from one family only. We therefore also fine-tuned Gram at all 30 cells. The comparison requires matched input pipelines. The released data loader of Gram re-references each cell to the mean of its retained electrodes, whereas ours does not. This difference alone was worth +0.0225 to that column and produced reduced-electrode cells that outperformed their own full montage. With matched pipelines, the Gram column averages 0.6395 against 0.6335 for the LaBraM column and is the stronger of the two on 16 of 30 cells, at 181M parameters across its thirty copies. Our logit teachers therefore come from the weaker family on this cell set, although the family that supplies them has no measurable effect on the student (Appendix G.1). The ordering became visible only once both columns had five seeds. On seed 0 alone, the same comparison gave 0.6289 against 0.6345 and placed Gram behind. Taking the better specialist per cell gives 0.6571, so the gap of the student is -0.0527 against a per-cell oracle over two backbones and -0.0291 against the LaBraM family alone. The student and the LaBraM column are compared cell by cell in Table 10, where the student leads in 5 of the 30 cells. The seed spread of the specialist column is not uniform, from 0.09 on BCI IV 2a at sixteen electrodes to less than 0.007 on every cell of HMC and Mumtaz2016.

D.3 COMPACT AND CLASSICAL BASELINES

Yang et al. (2026) report that EEG foundation models often fail to significantly outperform compact neural networks in data-scarce settings, and we reproduce this comparison with their released EEG-Net implementation and hyperparameters. On BCI IV 2a, a 5.2k-parameter EEGNet reaches 0.7068

Table 11: **Deployment cost of each model.** MFLOPs are millions of floating-point operations per window, and cells served counts the task-by-montage cells that one loaded copy covers.

Models	Parameters	MFLOPs	Latency (ms)	Windows/s	Cells served
MoRSE (ours)	1.97M	491	2.48	18507	30
LaBraM-Base	5.82M	1468	5.72	7072	1
Gram	6.03M	1515	7.69	7411	1
CBraMod	28.09M	950	7.38	11642	1
REVE-Base	69.19M	16504	6.63	2084	1
EEGNet	0.01M	104	1.83	31630	1
SPaRCNet	0.99M	126	4.10	15435	1
ContraWR	1.58M	71	1.69	31486	1
CNN-Transformer	3.16M	100	8.73	6113	1
FFCL	2.42M	71	8.84	6641	1
ST-Transformer	3.43M	104	1.83	31175	1

balanced accuracy, compared with 0.5479 for a fine-tuned LaBraM-Base and 0.6127 for Gram, the strongest foundation-model specialist we trained on this task. Both values come from the shared protocol of Appendix G.1. The tuned LaBraM-Base of Table 2 reaches 0.5333 on this task, and EEGNet exceeds all three. The advantage is not caused by muscle artifact, because adding a 40 Hz low-pass filter, the band CBraMod uses for this dataset, changes EEGNet by 0.008. The subject split has a larger effect. Using the split of CBraMod, which tests on subjects 8 and 9 rather than 3 and 8, lowers EEGNet to 0.5964 under the same low-pass filter. We did not retrain the foundation-model specialists under that split, so we do not claim that the ordering on this task holds beyond our split.

This single task is sufficient to reverse the aggregate, which is why Figure 5 has two panels. Averaged over all eight full-montage cells, the EEGNets total 49.7k parameters and reach 0.6394, ahead of the 0.6230 of our student. Excluding BCI IV 2a reverses the ordering, with the student at 0.6551 against 0.6298 for EEGNet. Per task, the student is ahead on CHB-MIT ($+0.088$), TUEV ($+0.058$), MentalArithmetic ($+0.050$), HMC ($+0.021$) and Mumtaz2016 ($+0.006$), and behind on ADFTD (-0.012), SEED-V (-0.033) and BCI IV 2a (-0.308). Our compression argument therefore holds against foundation-model specialists but not against EEGNet. The eight task-specific instances of EEGNet are together more than an order of magnitude smaller than our student and, on this task mix, no less accurate on average. The argument does hold against the other five compact baselines of Table 2, each of which is larger than EEGNet by at least two orders of magnitude and none of which exceeds our student on the eight-task mean. Against EEGNet, the student is more accurate on five of the eight tasks, deploys one model rather than eight and needs no new training run for a new montage (Section 3.3), whereas EEGNet needs one.

The same benchmark also reports that foundation models often fail to outperform classical non-neural decoders. We tested this second claim with the released scikit-learn pipelines of the same authors, namely CSP with LDA for motor imagery and relative band power with a random forest elsewhere. Every classical pipeline lies above chance and below EEGNet on every task, for example 0.5087 against 0.7068 for EEGNet on BCI IV 2a, 0.3382 against 0.5468 on TUEV and 0.8534 against 0.9085 on Mumtaz2016. They also lie below every foundation-model specialist on every task and below our student everywhere except BCI IV 2a. On this task set and these cross-subject splits, EEGNet exceeds every foundation-model specialist only on BCI IV 2a, and no classical pipeline exceeds any of them. Table 2 gives every method on every task.

D.4 Deployment cost

We measure the cost of one forward pass on the window shape of TUEV, 23 electrodes and five one-second patches, in fp32 on an otherwise idle NVIDIA H100 80GB GPU (Table 11). FLOPs are counted per window at batch size 1, latency is the median of 50 timed forward passes at batch size 1 after 10 warm-up passes, and throughput is measured at batch size 64. Parameters, FLOPs and latency do not produce the same ordering, which is why all three are reported. CBraMod is the second-largest model in the comparison but not the slowest, because its parameters lie in

one flattened classifier rather than in twelve sequential blocks. CNN-Transformer has $1.6\times$ our parameters and $3.5\times$ our latency. Our student is at least $2.3\times$ faster and at least $1.9\times$ cheaper in FLOPs than every foundation model, and one loaded copy serves all thirty cells, whereas every other row requires one model per cell. The compact CNNs are a different case. ContraWR, EEGNet and ST-Transformer are all faster than our student, at 1.69, 1.83 and 1.83 ms against 2.48 ms, and EEGNet is smaller by more than two orders of magnitude.

D.5 Preprocessing and exclusions

Reusable preprocessing for EEG self-supervision is itself an active research topic (Gjølbye et al., 2024). The dataset scripts released with LaBraM apply a 50 Hz notch, which matches the mains frequency of most of its recordings. TUH is recorded in the United States, where the mains frequency is 60 Hz, and on TUEV we measure a 60 Hz peak 4.4 times the local median of the 40–75 Hz spectrum, left inside the pass band while genuine gamma activity is removed by the notch. Correcting the notch improves balanced accuracy by +0.046 for EEGNet on TUEV over three seeds and by +0.029 for a single fine-tuned LaBraM-Base. Cho2017, also recorded on 60 Hz mains and corrected in the same way, serves as a control, and the same LaBraM-Base fine-tuning shows an effect of 0.000 there, which we attribute to the acquisition system. The conclusion is therefore not that correcting the notch helps in general, but that the notch should follow the mains frequency of each recording and that its effect must be measured per dataset. Every model in our comparison uses the same pipeline, so the choice favors no arm.

We evaluated and rejected a ninth candidate task, upper-limb motor imagery in chronic stroke patients, because both a fine-tuned LaBraM-Base (0.5077) and an EEGNet (0.5022) remain at chance. Agreement at chance between two architectures three orders of magnitude apart is a property of the dataset under cross-subject splits rather than of the models, and a task at chance dilutes every aggregate that contains it. We also excluded two candidate alignment teachers, both for the same architectural reason and neither for a reason of quality. The first convolution of the BENDR encoder mixes channels into the feature dimension, so per-(channel, patch) tokens do not exist, and the number of input channels is fixed in the pre-trained weights. EEGPT (Wang et al., 2024) reaches the same situation by a different route. Its encoder appends a small fixed set of summary tokens to the channel-patch sequence and retains only those tokens after the transformer blocks, so the channel axis is discarded by construction. Because Equation 10 is defined slot-wise on that axis, neither model can occupy the B slot at any width. This case differs from the width mismatch that keeps a 512-wide backbone out of the alignment term of a 200-wide student (Section 2.2). That mismatch is resolved by giving the student the same width, as Appendix G.4 does for REVE, whereas a missing token axis cannot be resolved. We record these exclusions rather than report empty rows.

E The two external cohorts

This appendix gives the provenance, preprocessing and additional results behind Table 3.

E.1 Provenance and preprocessing

Neither cohort appears in any pre-training list we compare against. The check follows Appendix B.4 and is made dataset by dataset rather than by corpus family. The fifteen public datasets in the appendix of LaBraM and the fifteen-row table of Gram contain no sleep data, CBraMod pretrains on TUEG only, and the ninety-two-dataset list of REVE contains neither cohort, its PhysioNet group being Siena and ICARE. ISRUC does appear in the downstream evaluations of CBraMod and REVE, which is a different matter and is why we name it. Neither cohort is part of the 524-hour distillation corpus of Appendix B.2 or of the eight tasks.

ISRUC subgroup 1 contains one night per subject, and we use 82 of its 100 recordings. One recording lacks F3 and F4 entirely. The other seventeen come from a gain-anomalous acquisition stream. They run a whole night at a median window amplitude of 0.8–$2.1\,\mu$V against 6.7–$41.3\,\mu$V for the rest, and the largest multiplicative gap in the cohort separates the two groups. They are also exactly the subjects whose EDF header declares a physical range of $\pm 25\,\mu$V where every other subject declares ± 2490 or ± 5013. We exclude these recordings rather than rescale them, because the measured ratio matches neither $1\times$ nor the $99.6\times$ implied by the headers. The issue matters only because

our pipeline uses absolute microvolts, and a pipeline that normalized per recording would not have detected it. From the PhysioNet 2018 training set we take a fixed random sample of 150 of the 994 annotated records. Its sleep stages are annotated at stage transitions rather than per epoch, roughly a hundred marks for a seven-hour night, and we expand them to epochs. Read as per-epoch labels, they would mislabel almost every epoch without raising an error.

Preprocessing is identical to that of HMC except for the notch of P2018. Each scored $30\,\mathrm{s}$ epoch becomes two $15\,\mathrm{s}$ windows at $200\,\mathrm{Hz}$ that share its label and lie in the same split, after a 0.1–$75\,\mathrm{Hz}$ band-pass. The notch follows the per-country rule of Appendix D.5 and is therefore at $60\,\mathrm{Hz}$ for Boston rather than the $50\,\mathrm{Hz}$ of HMC and Coimbra. Both cohorts are split by subject at the proportions used throughout, but nothing is fitted on either. The sleep head is applied as trained on HMC and no target label enters anything but the metric, so the split carries no information here and Table 3 is scored on the whole assembled cohort, all 82 ISRUC and all 150 P2018 subjects. Scoring only the test split, about a seventh of the subjects, moves the two-electrode effect to $+0.1182$ and $+0.1017$ and leads to the same conclusion.

E.2 Consistency and null results

The two cohorts differ in country and acquisition system, ISRUC coming from Coimbra and P2018 from Massachusetts General Hospital in Boston, yet the two-electrode gains of Table 3 differ by only 0.010. The effect also appears in how much each arm loses as the montage shrinks. Reducing the montage from four electrodes to two costs the recipe 0.0594 on ISRUC and 0.0509 on P2018, against 0.1845 and 0.1549 without sampling, a difference of $+0.1250$ and $+0.1040$ that is positive on every seed in both cohorts. At two electrodes, the REM recall of the arm without sampling falls to 0.024 on ISRUC and 0.042 on P2018, against 0.406 and 0.433 with sampling.

The four- and six-electrode rows of Table 3 are null, as expected. HMC has four electrodes, so the four-electrode row is the training montage of the fixed-montage arm and the six-electrode row adds two electrodes, whereas a sampler is useful when electrodes are removed. Adding those two electrodes, which the sleep head never saw during training on HMC, reduces the accuracy of both arms by 0.046–0.051 on ISRUC and 0.017–0.019 on P2018. Published ISRUC results, 0.72–0.79 for the methods evaluated by Wang et al. (2025) and El Ouahidi et al. (2025), are not comparable to this table. They fine-tune on ISRUC and model sequences of twenty epochs, whereas our student reads a single $15\,\mathrm{s}$ window from a dataset it has never seen.

F Representation collapse under fine-tuning

This appendix gives the full representation diagnosis summarized in Section 2.2, together with an implementation hazard that silently corrupts the same alignment term. Table 12 and Figure 8 report the effective rank of Equation 4 for LaBraM-Base before and after task fine-tuning. Pooled values are computed on up to $4{,}000$ test windows per task, a fixed random subset where the test set is larger and all of it where it is smaller, and per-patch values on $4{,}000$ patch tokens.

F.1 Collapse to the class dimension

Ordering the tasks by class count reveals a regularity. Across the seven tasks the backbone fits, the fine-tuned pooled effective rank tracks the number of classes at $r = +0.94$, with κ between 0.9 and 2.5. This value has a simple explanation. Softmax is invariant to a constant shift of its logits, so the logits of a C_y-way head occupy a (C_y-1)-dimensional affine subspace, and a linear head needs only $C_y - 1$ directions of its 200-dimensional input to determine them. The argument does not imply, but Table 12 shows, that the remaining directions retain almost no variance after fine-tuning. The pooled rank falls from 15–30 before fine-tuning to 1.5–4.9 after, so the effective dimensionality falls to roughly what the task decision requires. The measurement does not establish which directions survive, since an effective rank counts dimensions without identifying them with the row space of the head. SEED-V, at $\kappa = 3.6$, supports the mechanism, because a model that never fit the task never underwent the collapse.

In the terminal phase of training the class means of a network form a $(C_y - 1)$-dimensional simplex and within-class variability vanishes (Papyan et al., 2020), and the effective rank measured here is

Table 12: **Effective rank of LaBraM-Base before and after task fine-tuning.** Values are out of 200, κ divides the fine-tuned pooled rank by $C_y - 1$, and patch reduction divides the pre-trained per-patch rank by the fine-tuned one.

Tasks	C_y	Pre-trained backbone		Fine-tuned expert		Compression	
		Pooled	Per-patch	Pooled	Per-patch	κ	Patch reduction
CHB-MIT	2	15.2	73.4	2.5	5.3	2.5	14×
MentalArithmetic	2	30.1	56.0	1.5	1.8	1.5	31×
Mumtaz2016	2	17.0	57.2	1.5	1.6	1.5	35×
ADFTD	3	20.8	68.3	2.4	3.1	1.2	22×
BCI IV 2a	4	26.2	47.1	3.2	3.1	1.1	15×
HMC	5	22.7	60.5	4.9	8.1	1.2	7×
TUEV	6	19.1	74.1	4.7	7.5	0.9	10×
SEED-V	5	27.6	70.0	14.4	38.4	3.6	2×

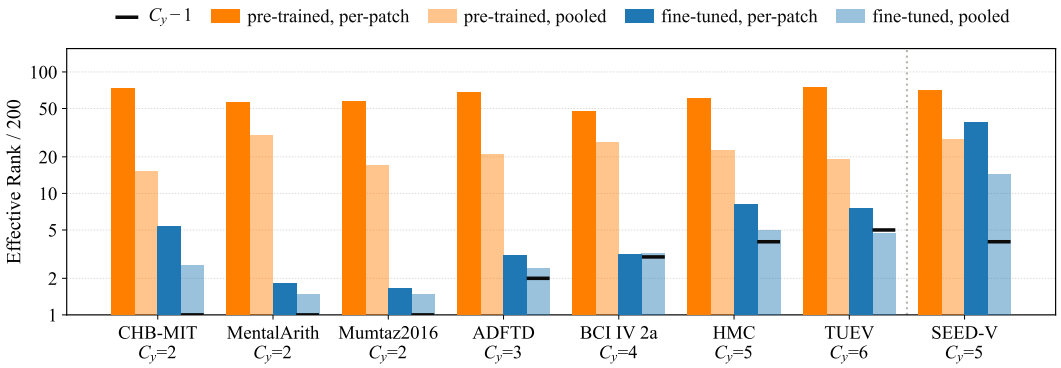

Figure 8: **Effective rank of LaBraM-Base before and after fine-tuning on all eight tasks.** Black marks show $C_y - 1$, the dotted rule separates SEED-V, and the y axis is logarithmic.

the dimensional signature of that geometry. A representation compressed to about $C_y - 1$ effective dimensions still carries the task decision but is an ineffective source of features, so the two distillation signals must be taken from different models. Two design choices follow. On the seven tasks the backbone fits, fine-tuning lowers the per-patch rank by a factor of 7 to 35, and distilling features from a fine-tuned expert would align the student to a subspace of one and a half to eight dimensions, so features must come from the backbone before fine-tuning. Pooling also lowers the rank of the pre-trained model by a factor of 1.8 to 4.8, so alignment must act per patch.

The measurement reads inputs only, never labels and never accuracy, and repeating it on the validation split of every task reproduces it. There the fine-tuned pooled rank tracks the class count at $r = +0.91$, against $+0.94$ on the test split. On the seven tasks the backbone fits, κ lies between 0.8 and 2.1, and SEED-V is again the only exception, at 2.7. Which model occupies which teacher slot therefore does not depend on test data.

F.2 Token order as an implementation hazard

Backbones disagree about the order of the flattened (channel, patch) token sequence. LaBraM and CBraMod are channel-major, whereas Gram is time-major. Because $C \cdot A = A \cdot C$, feeding one convention to a model that expects the other raises no error, and every patch simply receives the channel and time embedding of another slot. This error occurred in our pipeline and produced a large spurious interaction in an earlier version of Table 4. We now guard against it with a permutation-equivariance test, in which permuting the electrode order of the input together with the channel list must permute the output tokens and change nothing else. Under the wrong convention the deviation

Table 13: **Alignment teacher over eight tasks and 30 cells, with everything else identical.**

Alignment teachers	Balanced accuracy	Pre-training hours
Gram	**0.6044** $\pm$ 0.0040	7,290
Gram + LaBraM	0.5924 $\pm$ 0.0118	—
LaBraM	0.5797 $\pm$ 0.0045	2,534
CBraMod	0.5750 $\pm$ 0.0069	>9,000

Table 14: **Four pre-trained backbones fine-tuned as specialists on the full montage.** Headers give encoder parameters and pre-training hours.

Tasks	Gram 6.0 M, 7,290 h	LaBraM 5.8 M, 2,534 h	CBraMod 4.9 M, >9,000 h	REVE-Base 69.2 M, 61,415 h
TUEV	**0.6844**	0.6549	0.5845	0.6764
HMC	0.6796	**0.7075**	0.6875	0.6856
SEED-V	0.3327	**0.3386**	0.2609	0.3203
BCI IV 2a	**0.6127**	0.5479	0.5753	0.5839
CHB-MIT	0.8740	0.8042	0.8498	**0.8963**
MentalArith	**0.7152**	0.6473	0.6194	0.7102
Mumtaz2016	0.9194	0.9195	0.9128	**0.9283**
ADFTD	0.5605	**0.6839**	0.5882	0.5549

is 3.53 against a token magnitude of 7.12, and under the correct one it is 4.5×10^{-6}. We release the test.

G Additional results

This appendix collects further results on the teacher slots, the loss terms, the held-out layouts, the backbone instances and the per-task behavior of the shared encoder.

G.1 The teacher slots

Varying the alignment teacher alone (Table 13), Gram exceeds LaBraM by +0.0247 ($t = 12.8$) and CBraMod by +0.0294 ($t = 8.8$). Both differences are paired over the five seeds every arm has and are positive on every seed. Selecting on the validation split alone picks the same alignment teacher. REVE-Base cannot occupy this slot in a 200-wide student, because Equation 10 is a cosine over the token dimension and its tokens are 512-wide, so it enters only as a whole instance (Appendix G.4). CBraMod is the weakest teacher despite the largest pre-training corpus of the three. Scale does not order this slot. The eight-task means of the four backbones fall within 0.038 of each other, and Gram leads REVE-Base by 0.0028 with an eighth of its corpus and an eleventh of its encoder (Table 14). Gram and LaBraM each lead on three of the eight tasks, REVE-Base on two and CBraMod on none.

The four backbones are compared as specialists in Table 14 under one protocol, with our splits and preprocessing throughout and learning rates selected on validation. All four columns train for 50 epochs, 200 on BCI IV 2a, so no column is given a budget its neighbors did not have. Encoder parameters exclude the task head, which is not comparable across the four. The head of CBraMod flattens every token into a linear layer and ranges from 12 to 40 M parameters by task, two and a half to eight times its own 4.9 M encoder, so a total count would rank the backbones by readout rather than by backbone. The LaBraM column of this table follows the shared protocol rather than the tuned ceiling used elsewhere in this paper. It uses batch size 64 like the other three columns instead of the per-task ladder from 32 to 512, and drop path 0.1 with weight decay 0.05 on SEED-V instead of the tuned 0.3 and 0.1. The tuned runs serve as the reference for the student elsewhere, where the strongest attainable specialist is appropriate, whereas this comparison asks which backbone is better and therefore gives every column the same protocol. Our experiments use the 512-wide REVE-Base release. The other evaluated backbones are base-tier models with 4.9 to 6.0 M encoder parameters,

whereas REVE-Base has 69.2 M. REVE-Large changes the width from 512 to 1216 and retains the same 22 layers, the same 61,415-hour pre-training corpus and the same objective. Its number of attention heads and of positional frequencies both follow from this width. A REVE-Large column would therefore vary scale rather than backbone, so we do not include one. Two departures for REVE are deliberate. Its position bank has no entry for T1/T2 (TUEV) or CB1/CB2 (SEED-V), so these four are mapped to the nearest site it carries, at a measured distance of 0 mm. It also reads electrode coordinates rather than a name vocabulary, which is the property it is designed for.

The family of logit teachers does not matter. E_t is LaBraM only for the practical reason given in Section 2.2, so we built the missing family. Gram specialists were fine-tuned at all 30 cells with the same recipe and reused as logit teachers behind an adapter verified to reproduce the test accuracy of each specialist exactly. Substituting them into the main recipe changes it by +0.0007 over five paired seeds, 0.6051 $\pm$ 0.0051 against the same recipe with LaBraM logit teachers. This corresponds to $d = 0.09$, with three of the five paired differences positive, so the effect is absent rather than small and does not have a consistent sign. Two further families give the same answer. REVE-Base has the largest pre-training corpus of the four backbones, and its specialists form the second-strongest column of Table 14. Used as logit teachers in place of those of LaBraM, they change the main recipe by -0.0031 over five paired seeds (0.6013 $\pm$ 0.0052, $d = 0.45$, one of the five differences positive). CBraMod gives the sharpest test, because it is the weakest specialist column of Table 14, 0.0375 below Gram on the eight-task mean. Its specialists as logit teachers change the main recipe by -0.0034 (0.6010 $\pm$ 0.0054, $d = 0.47$, one of the five differences positive), and the TUEV disclosure attached to the CBraMod specialist column applies to this arm as well. The four families span 0.0041 as teachers, whereas the specialist columns they come from span 0.0375, nearly ten times as much. The same substitution under the variant recipe with the subset-consistency loss of Appendix I gives -0.0066 (0.5775 $\pm$ 0.0022, $d = 1.07$), also over five seeds, which is the same answer at a different operating point.

This result cannot be attributed to weaker Gram teachers. At five seeds per cell, the Gram specialist column averages 0.6395 against 0.6335 for the LaBraM column and is ahead on 16 of 30 cells, yet using the better specialist family as the source of logits still gives the student no benefit. This is consistent with Appendix F. What makes Gram the better alignment teacher is the geometry of its per-patch tokens, and fine-tuning collapses that geometry, by a factor of 7 to 35 per patch in the backbone measured there. A better backbone is thus a better source of representations but not of decisions, which agrees with the observation above that pre-training hours do not order teacher quality. Gram also does not win by carrying a richer representation in the obvious sense. Its per-patch tokens reach an effective rank of 66.5 against 75.6 for LaBraM, slightly lower, with a randomly initialized Gram at 43.7. Neither pre-training scale nor representational rank is therefore a proxy for the quality of an alignment target.

The fourth arm of Table 13 uses Gram and LaBraM as alignment teachers simultaneously. Aligning to Gram alone is better by +0.0120 at $t = 2.0$, paired seed by seed and positive on four of the five seeds. The gap did not shrink as seeds were added, reading $+0.0079$ over the first three seeds and $+0.0120$ over five, and the dual-teacher arm has the widest seed spread of the four, 0.0118 against 0.0040–0.0069. We regard this result as suggestive rather than established, because a standard deviation from five samples is not a firm basis for argument. The arm does establish a weaker claim, namely that the two representations are non-redundant. Linear CKA (Equation 17) between Gram and LaBraM is 0.169, against 0.101 for a randomly initialized Gram of the same architecture, a necessary control because unrelated high-dimensional representations do not score zero. Combining the two teachers nonetheless gives no benefit, so a difference of this kind is not sufficient for additive distillation.

The two slots have now been varied one at a time, and they behave differently. The alignment teacher matters and the logit teacher does not. The remaining objection is that the recipe mixes two families, so the mixture might have an effect that the measurements above do not isolate. A 2×2 design addresses it. The logit teachers are held at LaBraM, and the initialization of the student is crossed with the alignment teacher. With LaBraM initialization, alignment to LaBraM and to Gram gives 0.5797 and 0.6044. With Gram initialization, the corresponding means are 0.5758 and 0.5904. All statistics are paired seed by seed. The alignment teacher has an effect of the same sign in both rows, +0.0247 with a LaBraM student ($t = 12.8$) and +0.0146 with a Gram student ($t = 7.0$), so the preference is a property of the slot rather than of the student that measured it. The student initialization has a small effect in the opposite direction, with LaBraM ahead of Gram in both columns (-0.0039 and

Table 15: **Leave-one-out over the loss terms.** Montage sampling is on throughout, and every row is scored on the same 30 cells.

Configurations	$\mathcal{L}_{\mathrm{CE}}$	$\mathcal{L}_{\mathrm{KD}}$	$\mathcal{L}_{\mathrm{align}}$	Corpus	Balanced accuracy
Supervised baseline	✓				0.5656 ±0.0093
Logit-only	✓	✓			0.5603 ±0.0054
Feature-only, no corpus	✓		✓		0.5903 ±0.0047
Feature-only	✓		✓	✓	0.6016 ±0.0040
Full recipe	✓	✓	✓	✓	**0.6044** ±0.0040

-0.0141), and the interaction is -0.0101 ($t = -6.0$). Pairing a student with an alignment teacher from its own family therefore costs accuracy consistently across seeds. This interaction justifies the mixture. The initialization and the alignment slot have independent preferences and a same-family penalty rather than a bonus, so choosing the best occupant of each slot necessarily draws the recipe from two backbones, and this configuration follows from the measurements rather than being a confound. The result also answers the reverse concern, that a Gram student aligned to a Gram teacher is favored by the arrangement, which matters for Appendix G.4, where the generality claim rests on instances built entirely from one backbone.

G.2 Removing each distillation term

The comparisons above vary which model fills each teacher slot. Table 15 instead varies whether each slot is filled at all. The objective carries two distillation terms (Equation 11), logits from the fine-tuned experts and per-patch tokens from the backbone before fine-tuning, and Table 4 prices them together as one factor. Here each term is removed separately, with montage sampling on and the same 30 cells. Two of the five arms already existed, and each of the other three differs from one of them by a single flag.

The effect lies in one of the two terms. Representation alignment with the corpus is worth $+0.0360$ over the supervised baseline, $d = 4.4$ on the paired seed differences and positive on all five seeds. Logit distillation is not separable from zero. It gives -0.0054 alone ($d = 0.7$, positive on one seed of five) and $+0.0028$ on top of alignment ($d = 0.4$, positive on two of five). The $+0.0028$ is carried entirely by one seed, since dropping that seed makes the mean negative, and resolving an effect of this size at $t = 2$ would require about thirty seeds per arm. We therefore report it as absent. This agrees with the teacher-family result of Appendix G.1, and the presence of the logit term has almost no effect either. In this recipe the decisions supplied by the experts are already supplied by cross-entropy on the labels. We keep the term because it costs nothing at inference and because every other number in this paper was produced with it, not because it is what makes the recipe work.

The 524-hour unlabeled corpus carries $\mathcal{L}_{\mathrm{align}}$ and no other loss, so the arms without alignment are corpus-free by construction, whereas the feature-only arm gains 600 unlabeled steps per epoch in addition to the term itself. The feature-only arm without the corpus removes this confound. It differs from the feature-only arm only in the corpus ratio and takes the same labeled steps as the supervised baseline. The alignment term alone is worth $+0.0247$ over the supervised baseline ($t = 5.66$, positive on all five seeds), and the corpus adds $+0.0113$ on top of it ($t = 3.99$, positive on all five seeds). About two thirds of the alignment gain therefore comes from the term itself rather than from the additional unlabeled data. The contrast on which the logit null rests is unaffected, because the supervised and logit-only rows are both corpus-free and differ by one flag.

G.3 Held-out layouts

The zero-shot student is compared in Table 16 with a specialist fine-tuned on each of three (task, layout) pairs and selected by the checkpoint rule of Appendix B.3. The seed spread of these comparisons is large and uneven, and it is carried by the specialist rather than the student. On the temporal strip the specialist column has a seed standard deviation of 0.0466, and the oracle column 0.0214, against 0.0108 and 0.0089 on the posterior layout. The margin on the first row is therefore less than two seed standard deviations of the column it is compared with, and we do not treat any single row as decisive. The table is informative mainly through the student column, which required no training

Table 16: **The cost of zero-shot deployment on three held-out layouts.** The layout specialist is a LaBraM-Base fine-tuned on that layout, and the oracle column takes the best of three learning rates for each seed.

Tasks / layouts	MoRSE Zero-shot	Layout specialist	Specialist Oracle LR	MoRSE No sampling
CHB-MIT × temporal strip	**0.8312**	0.7572	0.7847	0.6265
CHB-MIT × posterior	**0.7940**	0.7721	0.7777	0.6433
TUEV × Emotiv-14	0.5589	**0.6218**	**0.6242**	0.3941

run for these layouts. The pooled result of Section 3.3 is paired seed by seed. Over the 42 pairs of Table 17, the five paired differences between the two students are all positive, and their mean of +0.0528 has a 95% interval of ± 0.0087 ($t = 16.77$, four degrees of freedom).

The comparison in Section 3.3 is between two students, which leaves a harder question open. Montage sampling is a property of the training distribution rather than of our architecture, so it can be applied to any model, including the 5.82M LaBraM-Base specialist from which the student is compressed. We therefore fine-tuned LaBraM per task with the same sampler, the same uniform size distribution and four-electrode floor, and evaluated it zero-shot on the same 42 pairs. Its epoch is selected by the rule used for the student, the best mean validation balanced accuracy across the four budgets, rather than by full-montage validation, which would have given the baseline a choice unavailable to the student.

The effect transfers with a similar magnitude. Sampling improves the LaBraM specialist by +0.0454 on these layouts, 0.5925 with it against 0.5472 without, on 36 of 42 cells ($t = 4.39$, paired over the 42 cells), compared with +0.0528 for the student measured the same way (0.5670 against 0.5143, 35 of 42, $t = 6.22$). The effect is therefore a property of the training distribution rather than of our student, since it survives a threefold change in capacity and a change from one shared encoder to eight per-task encoders. The two-student comparison alone could not support this stronger statement of the mechanism.

In absolute terms this baseline is ahead of the student on these layouts, 0.5925 against 0.5670, a gap of -0.0255 ($t = -3.12$ over the 42 cells), with the student ahead on 8 of 42 cells. On the 30 budget cells of Table 1 it reaches 0.6144 against 0.6044 for the student, a gap of 0.0100, which is two and a half times the seed spread of the student (± 0.0040) and cannot be attributed to noise. This bounds the contribution of the student. It does not offer more robustness to unseen layouts than a foundation model can be given, but it offers the same robustness at 1.97M parameters in one model rather than 46.6M in eight, at a cost of about 0.026 balanced accuracy. The comparison is like for like in the relevant sense, since neither model has seen a hardware layout before test. Table 18 gives the four arms per task and reveals two effects hidden by the aggregate. The deficit of the student is uneven. It is ahead on CHB-MIT by +0.0649 and behind on TUEV by -0.0801, the same split observed on the budget cells. Sampling is also heterogeneous for the specialist. BCI IV 2a is the only task that declines, the gains on SEED-V, MentalArithmetic and CHB-MIT are much smaller, and the +0.0454 aggregate is carried by ADFTD, TUEV and Mumtaz2016.

G.4 Backbone instances

A method evaluated over a single backbone would describe only LaBraM, so we built the whole chain twice more. An instance takes all three slots from one family X. The student is truncated from X, the alignment teacher is a frozen copy of X, and the logit teachers are the specialists of X. Each instance is scored against the specialists of its own backbone and never against another instance. The two 200-wide backbones give similar gaps, -0.0637 and -0.0583, but they share the same shape, so their agreement is the weaker form of the claim. REVE provides the stronger form. It is 512-wide, uses a gated feed-forward network, and locates electrodes by a 4D Fourier encoding of coordinates rather than by a name lookup. The chain still runs end to end, and this is the only configuration in which REVE can occupy the alignment slot, since a REVE-shaped student removes the width mismatch of Section 2.2 without a projection. The price is a larger gap of -0.0885. In Table 19, compression is against thirty per-cell specialists of the same backbone, and student sizes

Table 17: **Balanced accuracy on every held-out (task, layout) pair.** Dispersion is the percentile π of Equation 16, Sampling and None are two students that differ only in montage sampling, and Appendix B.1 records the tasks without an eligible layout.

Tasks	Layouts	Electrodes	Dispersion	Sampling	None
TUEV	frontal band	4	3.6	0.5093	0.4010
TUEV	frontal wide	7	0.0	0.5237	0.5457
TUEV	temporal strip	4	35.9	0.4008	0.3156
TUEV	central strip	3	11.5	0.3607	0.2492
TUEV	posterior	5	0.4	0.3517	0.2721
TUEV	left hemisphere	8	0.3	0.5411	0.4366
TUEV	emotiv 14	10	40.6	0.5589	0.3941
SEED-V	frontal band	4	9.2	0.2416	0.2443
SEED-V	frontal wide	7	1.1	0.2392	0.2332
SEED-V	temporal strip	4	62.8	0.2476	0.2514
SEED-V	central strip	7	3.0	0.2363	0.2268
SEED-V	posterior	5	1.9	0.2201	0.2111
SEED-V	left hemisphere	8	8.7	0.2553	0.2270
SEED-V	emotiv 14	14	93.7	0.2599	0.2444
BCI IV 2a	central strip	7	71.9	0.3894	0.3486
CHB-MIT	frontal band	4	2.4	0.7721	0.6963
CHB-MIT	frontal wide	6	0.1	0.8051	0.7158
CHB-MIT	temporal strip	4	33.7	0.8312	0.6265
CHB-MIT	posterior	4	1.1	0.7940	0.6433
CHB-MIT	left hemisphere	8	0.1	0.8805	0.8019
CHB-MIT	emotiv 14	10	41.2	0.9298	0.7995
MentalArith	frontal band	4	2.6	0.5362	0.5362
MentalArith	frontal wide	7	0.0	0.5400	0.5610
MentalArith	temporal strip	4	44.1	0.5235	0.5057
MentalArith	central strip	3	9.6	0.4930	0.4670
MentalArith	posterior	5	0.0	0.5305	0.5543
MentalArith	left hemisphere	8	0.2	0.5987	0.5902
MentalArith	emotiv 14	10	81.7	0.5337	0.5460
Mumtaz2016	frontal band	4	2.6	0.8535	0.7537
Mumtaz2016	frontal wide	7	0.0	0.8923	0.8832
Mumtaz2016	temporal strip	4	44.1	0.9118	0.8929
Mumtaz2016	central strip	3	9.6	0.8685	0.8227
Mumtaz2016	posterior	5	0.0	0.8808	0.8622
Mumtaz2016	left hemisphere	8	0.2	0.9111	0.8900
Mumtaz2016	emotiv 14	10	81.7	0.9063	0.8098
ADFTD	frontal band	4	2.6	0.4248	0.4056
ADFTD	frontal wide	7	0.0	0.4708	0.4461
ADFTD	temporal strip	4	44.1	0.4873	0.3955
ADFTD	central strip	3	9.6	0.4572	0.3998
ADFTD	posterior	5	0.0	0.5203	0.4167
ADFTD	left hemisphere	8	0.2	0.5411	0.4526
ADFTD	emotiv 14	10	81.7	0.5847	0.5233

include the eight task heads, as does the 1.97M quoted elsewhere in this paper. The heads have 5.8k parameters for a 200-wide student and 15.4k for the REVE-shaped one.

The larger gap of the REVE instance could not be explained away. It is not due to the learning rate. Sweeping $\{10^{-4}, 5{\cdot}10^{-4}, 10^{-3}\}$ on the student side, with the validation rule used for the specialists, selects $5{\cdot}10^{-4}$ for all three instances. It is not due to depth either. The student depth is itself selected on validation, and increasing the REVE student from two blocks to three, from 6.40 M to 9.54 M parameters, moves Δ to -0.0933, in the wrong direction, so the deficit is not simply the cost of a stronger compression. What remains is a property of the backbone. Two of 22 blocks form a different object from four of 12, and the representation of REVE may be harder to reproduce from a few electrodes. We report the boundary without claiming a mechanism.

Table 18: **Montage sampling given to the LaBraM-Base specialist on the held-out hardware layouts.** Every column is zero-shot and parentheses give the number of layouts per task. Effect is Sampling minus No sampling, and the last column is MoRSE minus the sampled specialist.

Tasks	LaBraM-Base 5.82M			MoRSE 1.97M			MoRSE − LaBraM
	No samp.	Sampling	Effect	No samp.	Sampling	Effect	
ADFTD (7)	0.4506	**0.5333**	+0.0827	0.4342	0.4981	+0.0638	-0.0353
Ment. (7)	0.5739	**0.5927**	+0.0188	0.5372	0.5365	-0.0007	-0.0562
Mum. (7)	0.8456	**0.9137**	+0.0681	0.8449	0.8892	+0.0442	-0.0245
SEED-V (7)	0.2370	**0.2506**	+0.0137	0.2340	0.2429	+0.0088	-0.0078
TUEV (7)	0.4564	**0.5438**	+0.0874	0.3735	0.4638	+0.0903	-0.0801
CHB (6)	0.7464	0.7705	+0.0241	0.7139	**0.8354**	+0.1216	+0.0649
BCI (1)	**0.5575**	0.4233	-0.1342	0.3486	0.3894	+0.0408	-0.0339
All 42 cells	0.5472	0.5925	+0.0454	0.5143	0.5670	+0.0528	-0.0255

Table 19: **One chain, three backbones.** Each instance is compared only with the specialists of its own backbone on the eight full-montage cells, and student sizes include the eight task heads.

Backbones	Width	Parameters	Compression	Δ vs own specialists	Seeds
LaBraM-Base	200	1.97 M	89×	-0.0637	5
Gram	200	2.05 M	88×	-0.0583	5
REVE-Base	512	6.40 M	324×	-0.0885	5

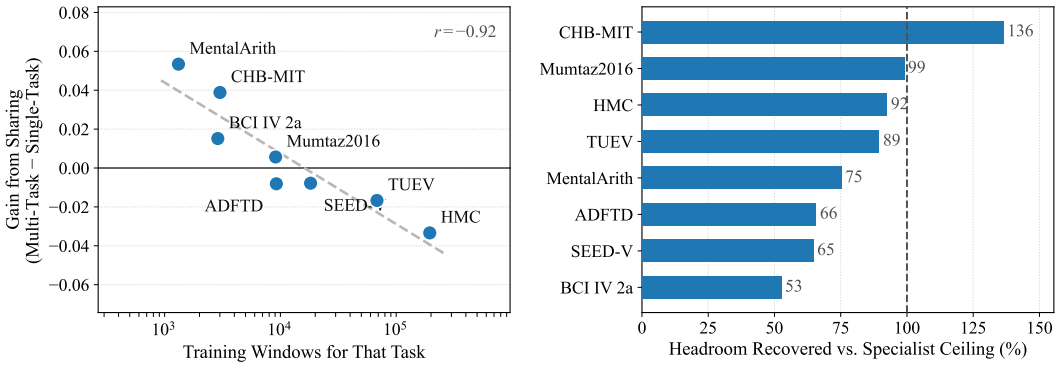

Figure 9: **Two per-task views of the shared encoder. Left:** the gain from sharing against the training-set size of each task. **Right:** the headroom recovered by the student on the full montage (Equation 15), with the dashed line at 100%.

G.5 Per-task behavior of the shared encoder

We additionally train eight single-task students, which use the same recipe and architecture as the shared encoder and differ only in being given one task. Each receives the full per-epoch step budget, so it sees eight times the gradient steps per task of the multi-task model and eight times its total compute. The baseline is deliberately advantaged so that the comparison cannot be attributed to under-training. Averaged over the eight tasks, sharing one encoder is worth +0.0059 against these single-task students. Figure 9 plots the gain from sharing, the mean over the cells of a task of the multi-task minus single-task balanced accuracy, against the training-set size of each task, and the two are almost perfectly anti-correlated ($r = -0.92$). Sharing is worth +0.0534 on MentalArithmetic, which has 1,317 training windows, and −0.0334 on HMC, which has 195,076. This relation also accounts for two results that the main text reports separately. CHB-MIT is the only task on which

the student beats its specialist on average over its budgets (Section 3.2) and the task on which zero-shot transfer to an unseen layout beats a layout-specific specialist (Table 16). Both follow from the small training set of CHB-MIT, 3,006 windows, rather than from seizure detection itself, so all three observations reflect the same mechanism. The right panel of Figure 9 rescales full-montage accuracy by the headroom above chance (Equation 15), which makes two-class and six-class tasks comparable. The student exceeds its ceiling only on CHB-MIT and recovers the least headroom on BCI IV 2a.

Aggregate scores report how much balanced accuracy a reduced budget costs, and Figure 10 shows what is lost. The matrices are row-normalized, so the diagonal gives per-class recall and remains readable under heavy class imbalance. The failure differs across tasks. TUEV, whose background class is 69% of its test set, drifts toward that class as electrodes are removed, which balanced accuracy registers but does not distinguish. ADFTD instead loses separation between Alzheimer's disease and controls, while frontotemporal dementia stays well separated. The distinction matters for deployment. A model that collapses to the majority class on a clinical screening task fails differently, and more dangerously, than one that confuses two adjacent sleep stages.

H STUDENT REPRESENTATIONS UNDER ELECTRODE REMOVAL

Section 3.3 and Appendix G.4 argue from scores. This appendix examines the representations themselves, using the pooled test-window vectors produced by our evaluation. No model is trained here, and every number comes from forward passes that had already been run. The analyses use the seed-0 checkpoint of each student, a model whose accuracy is already part of the five-seed means reported in the main text.

H.1 RETENTION ON HARDWARE LAYOUTS

Section 3.3 shows that the zero-shot student is accurate on layouts it never saw, but not why. Figure 11 addresses this question with representations instead of scores. Each window is encoded twice by the same student, once on the full montage and once on a hardware layout, and we measure how far the pooled representation moves as the cosine between the two batch-centered pooled vectors. Across the 42 (task, layout) pairs of Section 3.3, the student trained with montage sampling moves less on 38 of 42 pairs, with a mean cosine of 0.6228 against 0.5364, a gap of +0.0864. Pooled over the windows of all 42 pairs it retains more on 75% of windows, and on 85% for CHB-MIT with the temporal strip, the pair shown in the bottom panel, so the effect is not carried by a few windows.

This comparison requires a control, because a model could score well on it by ignoring its input. Cosines between different windows average 0.00 for both students and lie within 0.03 of zero on every layout, so both representations remain specific to the window and the gap reflects a real difference rather than a collapse. The asymmetric objective of Section 2.3 therefore teaches the student not only to score well without electrodes but also to keep more of its full-montage representation without them, which is the mechanism the method was designed around. The bottom panel shows a histogram rather than a t-SNE embedding, because t-SNE optimizes neighborhood probabilities rather than distances and cannot show that a representation moved less.

H.2 RETENTION ALONG THE BUDGET AXIS

The representation of a window can also be followed along the budget axis (Figure 12), which adds a measure with no free parameter. A test window is encoded at a reduced budget, and we ask whether its nearest neighbor among the full-montage vectors of that task, a pool of up to four thousand windows, is the same window. At eight electrodes the sampled student retrieves the window itself 81% of the time and the fixed-montage student 64%. At four electrodes the values are 45% and 35%, against a chance level of 0.098% averaged over the seven tasks, so a representation computed from four electrodes still identifies which window produced it nearly half the time. The batch-centered cosine gives the same result. Its control pairs each window with the next one in the test set, usually a neighboring window of the same recording, and lies between 0.20 and 0.37. It exceeds the 0.00 of Figure 11 because that control pairs windows at random, and with random pairing it is 0.00 here as well. In both cases it is far below the matched-window value. The failure at low budgets is not a rank collapse. The effective rank of Equation 4 does not fall as electrodes are removed but rises, from

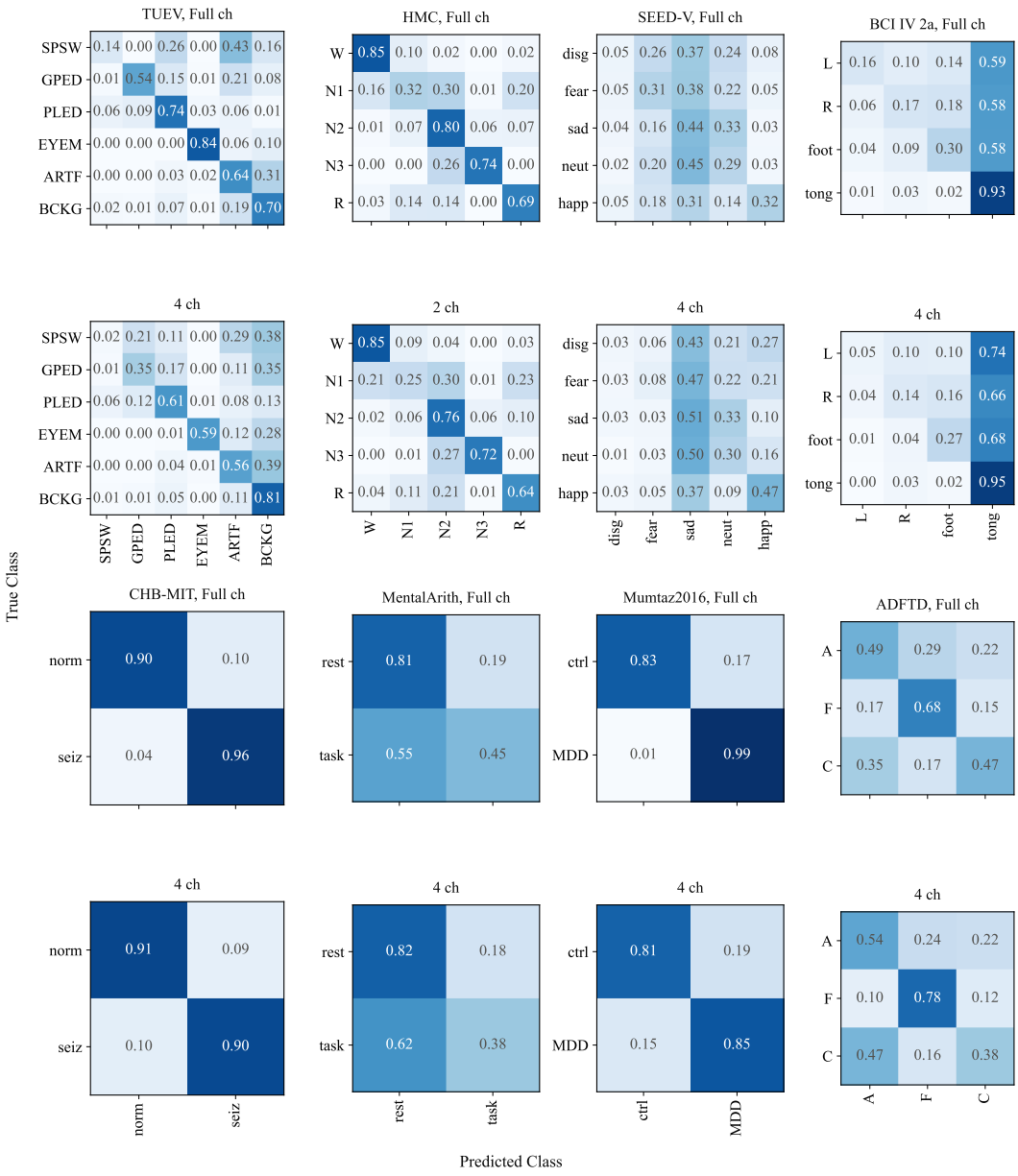

Figure 10: **Row-normalized confusion matrices for all eight tasks.** Rows are the true class and columns the predicted class, and each full-montage matrix sits directly above that of the smallest budget. The matrices are from the seed-0 student.

21.7 at the full montage to 30.4 at four electrodes over the same seven tasks, so what is lost is not dimensionality. Montage sampling also does not improve calibration. The expected calibration error is slightly worse for the sampled student at the full montage and at sixteen, eight and four electrodes, 0.15 against 0.13 at sixteen. Sampling thus improves accuracy and representational stability but not the agreement between confidence and accuracy, and a clinical deployment would require separate calibration.

H.3 Agreement between the three instances

The chain works for all three backbones (Appendix G.4), and Figure 13 asks whether the three students also compute similar representations, using linear CKA (Equation 17), which is defined

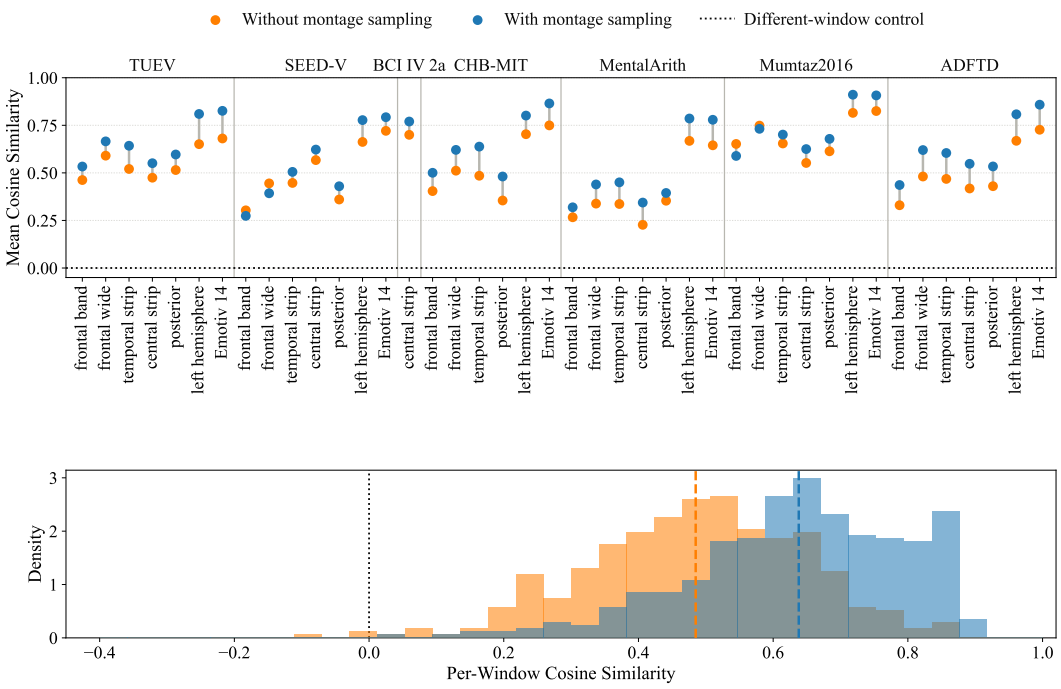

Figure 11: **The effect of montage sampling on the representation. Top:** mean retention for each (task, layout) pair, grouped by task, with the different-window control dotted. **Bottom:** per-window retention for CHB-MIT on the temporal strip, with dashed lines at the means.

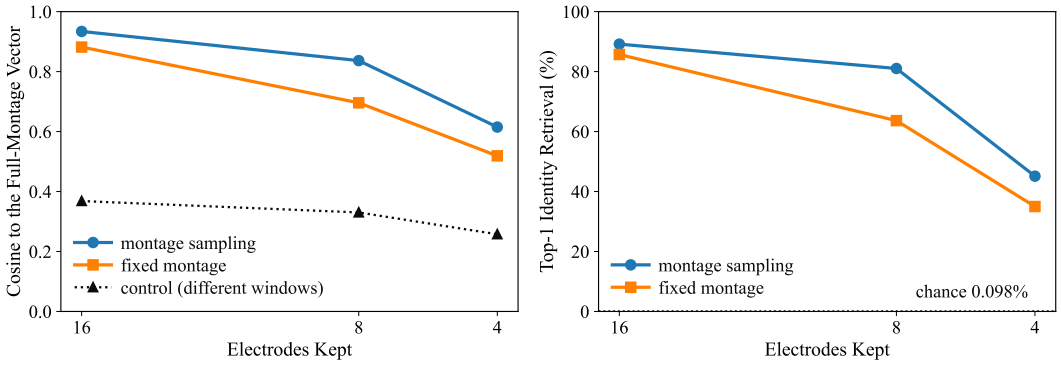

Figure 12: **Retention along the budget axis, averaged over the seven tasks with more than four electrodes. Left:** batch-centered cosine between the full-montage and reduced-budget vectors of a window, with a control between adjacent windows of the test set. **Right:** top-1 identity retrieval of a window from its reduced-budget vector.

across the 200- and 512-wide students. Their agreement depends on the task rather than on the pair of instances compared. Within a task the three pairwise values have a standard deviation of 0.039 on average, whereas across tasks their means span 0.156 to 0.829. Agreement tracks the performance of the student on the task ($r = +0.70$), and the lowest cell by a wide margin is BCI IV 2a, the one task on which the foundation-model approach as a whole fails (Appendix D.3). Three chains built from three different backbones therefore converge where they work and diverge only where none of them works.

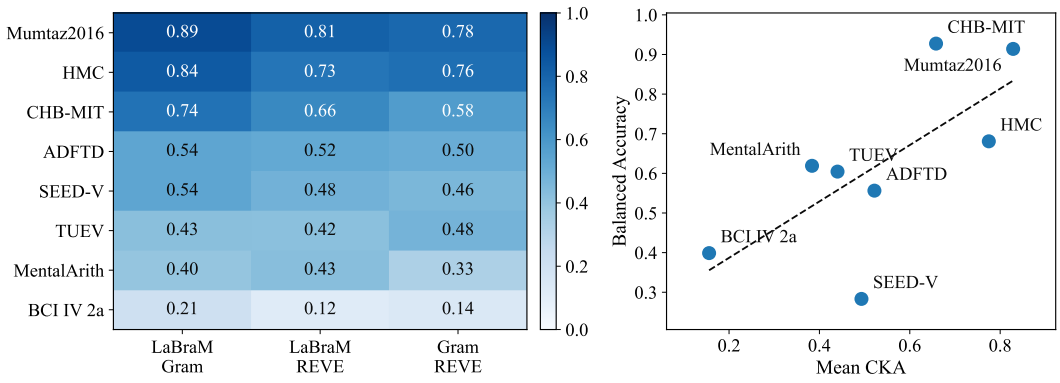

Figure 13: **Agreement between the three backbone instances. Left:** linear CKA between the pooled test representations of each pair of instances, per task. **Right:** the mean over the three pairs against the full-montage balanced accuracy of the student.

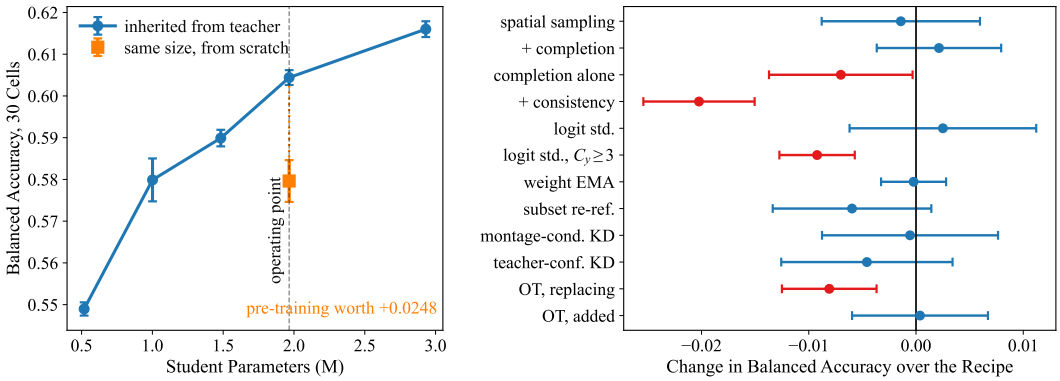

Figure 14: **Capacity against the eleven interventions. Left:** depth ladder at fixed width 200, every rung initialized from the same teacher by depth truncation, and the same 1.97M student trained from scratch. **Right:** paired-seed change of each intervention over the recipe on the 30 cells, with 95% t intervals, red where the interval excludes zero.

I ELEVEN INTERVENTIONS THAT DID NOT IMPROVE THE RECIPE

The recipe in Section 2 was fixed before any of the interventions below was run, and none of them was adopted, so these test-set comparisons did not shape it. None of eleven modifications to the sampler, logits, alignment target and optimizer improved on the full recipe beyond seed noise on the 30 budget cells. Five of them were harmful, and two predictions we had written down before the runs were contradicted by the outcome. All interventions are free at inference, so none was rejected because of deployment cost. The eleven comprise two sampler/objective variants, two forms of optimal-transport alignment, and seven other interventions. Stochastic comparisons are paired by seed against the recipe (Equation 13), except for weight averaging, which averages the five seed checkpoints of the recipe itself. Figure 14 places them beside a fixed-width depth ladder, and the last paragraph of this appendix shows that at this scale capacity changes the aggregate while targeted changes to the objective generally do not. The right panel of Figure 14 has twelve rows for the eleven interventions. The sampler and completion variants contribute a third row for their joint arm, montage-conditioned weighting contributes a second row for its teacher-confidence ablation, and weight averaging is absent because it is scored on one task rather than as an arm. Eight of the twelve intervals cross zero, and the four that do not are all negative.

I.1 THE ELEVEN INTERVENTIONS

Two interventions test spatial sampling and electrode completion. The clustered sampler draws grown spatial patches and hemisphere splits instead of uniform subsets. The completion head predicts the frozen teacher token for each electrode withheld from the student. On the 42 task-layout pairs, spatial sampling adds +0.0056 and completion alone +0.0005. Their joint arm adds +0.0065 ($t = 5.39$ over five paired seeds, positive on every seed). The factorial interaction, however, is only +0.0005 ($t = 0.08$), and completion on top of spatial sampling adds +0.0009 ($t = 0.18$). The joint gain on these layouts is therefore carried mainly by the clustered sampler, although its effect alone is not resolved at five seeds, and completion makes no resolved contribution. On the 30 budget cells, completion alone is harmful (Figure 14). The five-seed result thus overturns the coupling interpretation suggested by the first three seeds. The two individual variants remain counted among the eleven interventions, and their combined arm is the additional row in Figure 14.

Two further interventions modify the alignment term with optimal transport. The alignment term in Equation 10 pairs every observed student electrode with the same teacher electrode and discards the remaining teacher tokens. It therefore uses the fraction C_s/C of the teacher token field described in Section 2.3, which is 17.4% at a four-of-twenty-three budget and 6.5% on SEED-V. Because the student and teacher token sets differ in cardinality, we tested an entropic optimal-transport cost that asks the observed student tokens to represent the full teacher electrode distribution while the transport plan selects the correspondence. Unlike the completion head, the modification adds no parameters, and it reduces to Equation 10 when $S = [C]$. Before running it we predicted that its gain would increase as the student-to-teacher cardinality ratio decreased. Replacing slot-wise alignment with optimal transport changes the aggregate by -0.0081 ($d = 2.3$, negative on all five seeds) and changes performance on the hardware layouts by -0.0038. Adding the transport cost to the original alignment term gives +0.0004 ($d = 0.1$), with inconsistent signs across seeds. The prediction stated above also fails. The degradation grows as electrodes are removed, from -0.0020 at the full montage to -0.0069, -0.0101 and -0.0153 at sixteen, eight and four electrodes, the opposite of the predicted trend. These results show that a learned transport correspondence does not improve the objective. They do not show that the discarded teacher tokens are uninformative, because the experiment does not distinguish an uninformative target from one that is difficult to predict.

None of the remaining seven interventions improved on the recipe. Their stochastic comparisons are paired against the recipe of Section 2. A subset-consistency loss, added to the joint sampler and completion arm, maps two electrode subsets from the same window to the same pooled vector and changes the 30-cell mean by -0.0202 relative to the recipe ($d = 4.8$, negative on every seed). It imposes an inappropriate invariance because frontal and occipital subsets can legitimately contain different information. Averaging the weights of the five seed checkpoints, a technique that often helps elsewhere (Wortsman et al., 2022), reduces TUEV performance by 0.14. The seeds are not linearly mode-connected in the sense of Frankle et al. (2020), because each seed also changes both the sampler and the data order. Logit standardization (Sun et al., 2024) gives +0.0025, while restricting it to multiclass heads gives -0.0092 because the task-dependent transform also changes task weighting. An exponential moving average gives -0.0002. Weighting distillation by the support that the observed subset provides for the full-montage decision of the teacher gives -0.0006. This weighting is an instance of learning with privileged information (Lopez-Paz et al., 2016). Its confidence-only ablation gives -0.0046, so the montage-conditioned component contributes +0.0040 relative to that ablation. At $d = 0.9$, the difference is not resolved from seed noise. With three seeds the same contrast was +0.0073 at $d = 3.4$, which illustrates the instability of the smaller estimate. Neither weighting variant exceeds the full recipe. Finally, re-referencing each subset to the mean of its retained electrodes improves the Gram specialist column by +0.0225 but changes the student by -0.0060. It moves a backbone toward the reference used in its own pre-training and provides no corresponding benefit to the student.

Two of these failures have simple arithmetic explanations. Standardizing a C_y-class logit vector removes its mean and scale, leaving $C_y - 2$ degrees of freedom. For $C_y = 2$, every teacher logit vector therefore maps to ± 0.7071, so confident and uncertain predictions yield identical targets and distillation reduces to hard labels. This degeneracy is especially relevant to clinical EEG benchmarks, which are often binary. Re-referencing C channels to their retained-channel mean similarly

removes one degree of freedom. The loss is negligible with twenty-three channels but severe with two, and the change on the two-electrode cell of HMC is -0.1420.

I.2 Further design checks

Equation 5 samples both subset membership and subset size. Before running it we predicted that a fixed-k student would outperform the sampled student at its training value of k and underperform at other budgets, corresponding to the usual specialist-generalist tradeoff. The prediction does not hold. Fixing $k = 8$ over five seeds changes the 30-cell aggregate by -0.0027. The deficit is present at the full montage and at sixteen, eight and four electrodes, including -0.0028 on the eight-electrode cells used during training, compared with -0.0023 at the full montage and -0.0031 at four electrodes. The aggregate difference is not resolved at five seeds ($t = -1.94$, four of five seeds negative). Sampling the subset size is therefore not necessary for the observed gain, but fixing the size also provides no measurable specialization benefit at that budget.

Equation 10 computes cosine similarity after centering across the combined batch-and-slot axis. This operation is necessary because a shared mean direction dominates between-sample cosine similarity in these backbones. For the alignment teacher, pooled representations from unrelated windows have a cosine of 0.91 before centering and -0.02 afterward. The per-patch tokens used by the alignment term change from 0.51 to 0.26. The reduction after removing one mean direction indicates anisotropy rather than representation collapse. Without centering, the alignment objective is almost satisfied at initialization and provides little gradient even though its loss value appears to decrease.

A fixed-width depth ladder (Figure 14, left), initialized at every depth by truncating the same teacher, gives 0.5490 at 0.518M parameters (337× compression), 0.5799 at 1.001M, 0.5899 at 1.483M, 0.6044 at 1.97M, and 0.6160 at 2.931M. The curve has not saturated at its largest model, so the 1.97M operating point reflects a deployment budget rather than an optimum. A 1.97M student trained from scratch reaches 0.5796, making pre-trained initialization worth +0.0248, an improvement larger than that obtained from half a million additional depth parameters. Every objective modification above is small relative to this capacity trend. We also considered register tokens (Darcet et al., 2024) but did not train the intervention because the motivating pathology was absent. Across all budgets, the maximum residual-stream norm is 1.25 times the median and no token exceeds three times the median, whereas the reported pathology spans two orders of magnitude.

J Discussion

J.1 Pre-training scale and teacher quality

This paper assigns the same four backbones to four roles and measures each, as fine-tuned specialists, as frozen alignment teachers and as sources of logit teachers (Appendix G.1), and as the models being compressed (Appendix G.4). Two roles have three candidates, because REVE-Base cannot fill the alignment slot of a 200-wide student and CBraMod was not compressed. Across the four backbones the corpus spans a factor of twenty-four, from $2{,}534$ to $61{,}415$ hours, and the encoder spans 4.9 to 69.2 M parameters, yet the same mid-sized backbone ranks first in all four roles (Figure 15) and the largest ranks first in none, although in the logit role the four families differ by only 0.0041, within seed noise. We state this as an observation rather than a law, since four points do not support a slope, and Figure 15 therefore reports no correlation coefficient. It does support a negative claim strong enough to act on. Scale orders none of the four roles, so a teacher cannot be chosen from model cards and the slots must be filled by measurement, as Section 2.2 does, which is why the two slots of our objective hold different families rather than the largest available model.

J.2 Evidence from the eleven interventions

Appendix I reports eleven interventions with no inference cost that did not improve on the recipe beyond seed noise on the 30 budget cells, five of them harmful. Two of its results argue for reporting factorial effects over a common seed set rather than inferring mechanisms from one-at-a-time or low-seed ablations. On hardware layouts, the coupling between clustered sampling and completion suggested by the first three seeds does not survive five, with a factorial interaction of $+0.0005$ ($t = 0.08$), whereas the main 2×2 retains a positive descriptive interaction of $+0.0164$ ($t = 1.82$).

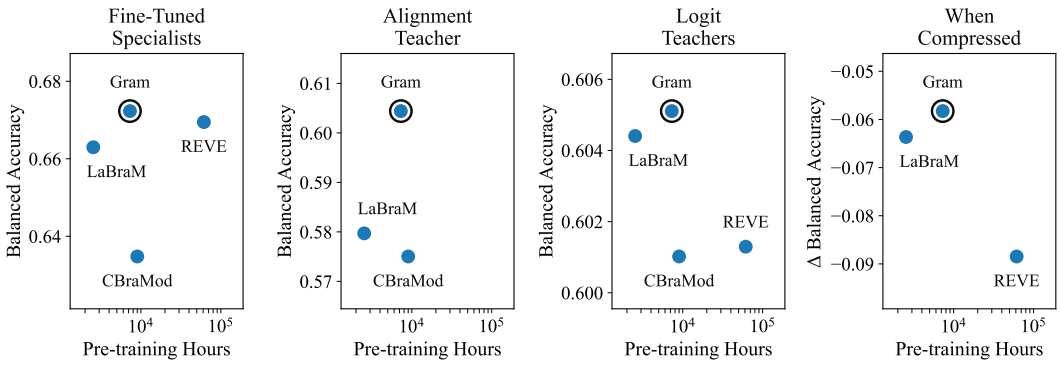

Figure 15: **Pre-training hours against four roles of a backbone.** From left to right, the balanced accuracy of the fine-tuned specialists over eight tasks, of the student over 30 cells for each alignment teacher and for each source of logit teachers, and the difference between each compressed instance and its own specialists. Rings mark the best entry in each panel.

The two predictions we had written down before the runs also failed. A fixed-k student did not beat the sampled one even at its own k, and the loss from optimal transport grew rather than shrank as electrodes were removed. We report these results as measured outcomes of this recipe, not as evidence that the tested components are intrinsically unhelpful, because a recipe assembled from interventions kept when they helped and forgotten when they did not cannot be reproduced.

J.3 Limitations

Our work has several limitations. First, the compression argument holds against foundation-model specialists but not against EEGNet. On small-sample motor imagery a 5.2k-parameter EEGNet outperforms every foundation-model specialist we trained and places the eight EEGNets ahead of our student on the eight-task mean, although the student leads on five of the eight tasks. Second, the gap is larger against stronger references, -0.0527 against a per-cell oracle over the LaBraM and Gram families rather than the -0.0291 of Table 1, and spherical-spline interpolation feeding a full-montage specialist is not dominated per cell, although it retains the T large models this paper compresses. Third, the student is not the most accurate model on held-out layouts. The 5.82M backbone trained with the same montage sampler is 0.0255 more accurate, so the contribution of the student is the same robustness in one 1.97M model rather than eight. Fourth, the student is robust to subsets of the electrodes it knows rather than to arbitrary electrodes. An electrode outside the vocabulary is mapped to the nearest known site, and we do not measure how performance degrades with that distance. Fifth, every result is obtained on eight tasks under one cross-subject protocol and one preprocessing pipeline, which is also why published numbers are not comparable to ours, and splitting by subject does not exclude the dataset-identity confound described by Zare (2026), which our two external cohorts address only in part. Finally, the teacher slots have to be filled by measurement, and every number in this paper is bounded by the quality of the backbones MoRSE distills from, since we neither pre-train a new foundation model nor claim a new state of the art.

J.4 Future work

These limitations suggest three directions. First, replacing the name vocabulary of the student with a coordinate-based electrode encoding, as REVE uses for its backbone, would extend robustness from subsets of known electrodes to arbitrary ones. Second, evaluating the student on further external cohorts and with dataset-identity negative controls would separate montage robustness from dataset identity. Third, montage sampling does not improve the calibration of the student (Appendix H), so calibration under electrode loss remains to be addressed before clinical deployment.