# DIVER-1: SCALING INTRACRANIAL EEG FOUNDATION MODELS FOR TRANSFERABLE REPRESENTATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Intracranial EEG (iEEG) provides direct, millisecond-scale recordings of human neural activity, but reusable representation learning is difficult because electrode layouts, anatomical coverage, referencing schemes, and recording conditions vary across patients and centers. We introduce DIVER-1, a self-supervised iEEG foundation model for variable-input recordings that combines any-variate electrode–time attention, spatio-temporal resampling, input-conditioned positional embeddings, and multi-domain masked reconstruction without assuming a fixed electrode montage. We pretrain two variants, DIVER-1-0.1s and DIVER-1-1s, on 5,310 hours of ECoG and SEEG spanning 352k channel-hours, roughly 54× the BrainTreeBank-based pretraining volume. We evaluate DIVER-1 on two held-out benchmarks: iMINDBench for naturalistic decoding and MAYO for seizure detection. DIVER-1-0.1s achieves the best results on both benchmarks—outperforming all prior foundation models and matching the strongest classical method on MAYO—despite being pretrained exclusively on task-free clinical monitoring recordings from external institutions. DIVER-1-1s achieves comparable performance with an order-of-magnitude smaller token budget, offering a practical option for compute-constrained settings. Finally, we present, to our knowledge, the first controlled, compute-aware scaling study of self-supervised iEEG pretraining, varying data scale, subject count, training duration, and model size up to 1.8B parameters. For masked reconstruction, we find that expanding unique recordings and training sufficiently long are more reliable scaling axes than increasing parameter count alone, highlighting the need for broader datasets and self-supervised objectives better aligned with downstream use. Code is available at link.

## 1 INTRODUCTION

Large-scale self-supervised pretraining has enabled foundation models across language, vision, and speech: models trained on broad unlabeled corpora to learn reusable representations for many downstream tasks (Bommasani et al., 2021). In neuroscience, the analogous goal is an encoder that learns from long unlabeled neural recordings and transfers across subjects, tasks, and clinical centers. Intracranial EEG (iEEG) is a compelling testbed for this goal: electrodes placed on or within the brain record field potentials close to their sources, preserving high-frequency, millisecond-scale neural dynamics that are attenuated and spatially blurred in non-invasive EEG (Buzsáki et al., 2012; Zahorodnii et al., 2025; Oganesian et al., 2025). These properties make iEEG valuable for cognitive decoding, brain–computer interfaces (Card et al., 2024; Wairagkar et al., 2025), and precision neuromedicine (Scangos et al., 2021; Khambhati et al., 2019).

Yet current iEEG foundation models have not fully realized this promise: under leakage-aware evaluation, simple spectrogram-based linear decoders often outperform pretrained iEEG models (Zahorodnii et al., 2025). The core difficulty is that iEEG is clinically implanted and patient-specific, so recordings vary in channel count, anatomical coverage, electrode type, reference scheme, and recording center. Thus, general iEEG foundation modeling requires more than standard Transformer pretraining; it requires addressing three bottlenecks in architecture, input interface, and scale.

**Bottleneck 1: restricted attention architectures for spatio-temporal neural dynamics.** Transformer-based foundation models model interactions through attention between patch tokens, i.e., local segments of the spatio-temporal neural recording. However, self-attention is order-agnostic



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.

by default: without additional structure, it does not know whether two tokens are adjacent in time or from the same electrode. This is important for iEEG, where behaviorally relevant activity can involve both local temporal dynamics within an electrode and coordination across distributed electrodes.

Existing iEEG models address this with different trade-offs. BrainBERT (Wang et al., 2023) models electrodes independently, precluding modeling direct cross-channel interaction. Brant (Zhang et al., 2023) and PopT (Chau et al., 2025) introduce spatial modeling but separate temporal and spatial processing. These designs are efficient, but cross-channel-time interactions remain indirect: activity at electrode $a$ and time $t$ can influence activity at electrode $b$ and time $t + \Delta$ only through staged or restricted attention routes. This is potentially limiting for modeling the brain's delayed inter-regional coordination, phase-lagged coupling, and distributed oscillatory communication (Deco et al., 2011; Stam et al., 2007; Fries, 2005; Varela et al., 2001).

Full spatio-temporal attention is therefore attractive because every electrode–time token can attend to every other token, but it still needs structure identifying each token's temporal and channel origin. BaRISTA (Oganesian et al., 2025) uses full spatio-temporal attention and injects temporal order directly into attention with RoPE, but adds spatial information as input-level metadata embeddings. This leaves same- versus cross-channel relations implicit in token features, despite Su et al. (2024) suggesting that relational structure can be more effective when encoded in attention rather than added to inputs. Moreover, many iEEG datasets also lack complete channel metadata (Bbrinkm & Cukierski, 2014; Carzaniga et al., 2025), so a general encoder should remain channel-relation aware even without spatial metadata embeddings.

**Bottleneck 2: fixed pretraining interfaces for variable iEEG inputs.** iEEG foundation models must also handle heterogeneity in the input interface itself. Clinical iEEG varies across patients in channel count, anatomical coverage, and electrode sub-modality (SEEG/ECoG). Downstream tasks also require different temporal contexts, from sub-second decoding of rapid perceptual or linguistic events to multi-second clinical windows in seizure detection. A general iEEG backbone should therefore not assume a single native channel count, electrode sub-modality, or task timescale.

Existing iEEG foundation models are trained through narrower interfaces. BrainBERT, Brant, PopT, and BaRISTA (Wang et al., 2023; Zhang et al., 2023; Chau et al., 2025; Oganesian et al., 2025) are pretrained on SEEG and use fixed pretraining input windows, creating a mismatch with ECoG recordings and with downstream tasks that spans shorter or longer windows. Furthermore, channel semantics depend not only on anatomical location, but also on the rest of the implant and the reference scheme. Under common-average referencing, for example, each contact is measured relative to the average of that patient's available contacts. Thus, a temporal-lobe contact in a temporal-focused implant is referenced against a different signal mixture than a comparable contact in a broader implant spanning temporal and extra-temporal regions. In this sense, a channel's meaning depends on the implant as a set, not only on its coordinate. EEG methods such as ACPE (Wang et al., 2024b) improve input adaptivity through conditional positional encodings, but channel-axis convolutions make the encoding order-sensitive: permuting the same electrodes changes the convolutional neighborhood even though the recording set is unchanged. For iEEG, where channel order is arbitrary across patients and centers, the input interface should preserve channel permutation equivariance.

**Bottleneck 3: under-scaled pretraining and missing iEEG scaling laws.** Current iEEG pretraining remains narrow and small. BrainBERT, PopT, and BaRISTA (Wang et al., 2023; Chau et al., 2025; Oganesian et al., 2025) train on BrainTreeBank (Wang et al., 2024a), and are evaluated on downstream tasks from the same dataset (Zahorodnii et al., 2025). This enables controlled benchmarking, but limits pretraining to pediatric movie-watching recordings at modest scale, roughly 40–43 hours total. It also differs from the intended use of foundation-models, where transfer should span subjects, centers, age groups, behavioral states, and recording protocols.Clinical monitoring is uniquely valuable because implanted electrodes can remain in place for days, yielding long recordings during naturalistic hospital behavior (Bbrinkm & Cukierski, 2014). Brant (Zhang et al., 2023) is an important scale exception, but its fixed 6 s patch interface limits use in sub-second contexts such as rapid perceptual and linguistic tasks.

Once pretraining data expand, compute allocation becomes the next bottleneck. Under a fixed compute budget, should iEEG foundation models scale data, parameters, training duration, or subject diversity? In language and vision, scaling-law studies have made such decisions less heuristic (Kaplan et al., 2020a; Hoffmann et al., 2022b; Muennighoff et al., 2023). For neural data, recent scaling studies

have focused mainly on supervised decoding or participant diversity in specific task settings (Banville et al., 2025; Bomatter & Gouk, 2025), while EEG/iEEG foundation models provide only coarse evidence that larger datasets or models improve performance (Zhang et al., 2023; Wang et al., 2024b; Jiang et al., 2024). The field therefore still lacks controlled, compute-aware sweeps over data, model size, training duration, and subject diversity for self-supervised iEEG foundation-model pretraining.

**DIVER-1: an iEEG-specific architecture and scaling study.** We address these three bottlenecks with DIVER-1, a self-supervised iEEG foundation model that combines any-variate electrode–time attention (Bottleneck 1), input-conditioned encoding and spatio-temporal resampling for variable layouts (Bottleneck 2), and large-scale heterogeneous pretraining paired with a controlled scaling-law study (Bottleneck 3). Our contributions are:

- We introduce **DIVER-1**, a self-supervised iEEG foundation model for variable electrode–time inputs. Its any-variate attention encodes temporal order and same-/cross-channel structure within attention itself, enabling direct cross-channel, cross-time modeling. STR (Spatio-Temporal Resampling) and STCPE (Spatio-Temporal Conditional Positional Embedding) make the input interface robust to variable electrode layouts and temporal contexts, while MDRO trains the model to reconstruct masked patches across complementary signal domains. We instantiate two patch-scale variants: DIVER-1-0.1s and DIVER-1-1s.
- We demonstrate **best-reported cross-dataset iEEG decoding from large-scale pretraining**. By pretraining on 5,310 hours of iEEG—over $54\times$ more channel-hours than BrainTreeBank-based pretraining—DIVER-1 transfers across 15 naturalistic decoding tasks spanning three institutions' datasets (iMINDBench) and clinical seizure detection (MAYO), all disjoint from pretraining. On iMINDBench, DIVER-1 outperforms all competing models despite no exposure to the benchmark's movie-watching datasets during pretraining, demonstrating transfer across subjects, institutions, and recording contexts.
- We provide, to our knowledge, the **first controlled compute-aware scaling-law study** for self-supervised iEEG foundation-model pretraining. Sweeping data scale, model size up to 1.82B parameters, training duration, compute, and subject count, we find that iEEG pretraining follows a data-constrained scaling regime, yielding a practical recipe for future iEEG foundation models: scale data first, train longer second, and scale parameters last.

## 2 METHODS

### 2.1 MODEL ARCHITECTURE

As illustrated in Figure 1, DIVER-1 patchifies each iEEG channel, encodes patches with a temporal CNN, and forms the token grid $\mathbf{X} = \mathbf{Y}_{\text{CNN}} + \mathbf{E}_{\text{STCPE}} + [\mathbf{E}_{\text{position}}, \mathbf{E}_{\text{modality}}] + \mathbf{E}_{\text{spectral}}$ where $[\cdot, \cdot]$ denotes concatenation and the embedding terms are defined below. A MOIRAI-based any-variate Transformer processes $\mathbf{X}$ across variable iEEG electrode layouts. We train two patch-size variants, DIVER-1-1s and DIVER-1-0.1s. Architectural hyperparameters are provided in Table 5.

**Patch encoding ($\mathbf{Y}_{\text{CNN}}$).** Given an input $\mathbf{Y} \in \mathbb{R}^{C\times T}$, we split each channel into non-overlapping temporal patches of length $P$, where $P = 500$ for DIVER-1-1s and $P = 50$ for DIVER-1-0.1s at 500 Hz. A three-layer CNN then extracts local features from each patch, yielding $\mathbf{Y}_{\text{CNN}} \in \mathbb{R}^{C\times N\times d_{\text{model}}}$, where $N = T/P$ is the number of patches and $d_{\text{model}}$ is the embedding dimension.

**Spatio-Temporal Conditional Positional Embedding (STCPE) ($\mathbf{E}_{\text{STCPE}}$).** STCPE injects **input-conditioned local spatio-temporal positional bias** without assuming a fixed iEEG montage. Because iEEG voltages are reference-dependent, a channel's signal reflects both its 3D location and its coupling to other channels through the reference scheme. ACPE (Wang et al., 2024b) captures such local channel–time context with convolutions, but its channel-axis convolution makes the encoding channel order-sensitive. STCPE replaces this with a sliding-window any-variate Transformer block (MOIRAI), yielding **temporally translation-equivariant** and **channel-permutation-equivariant** positional bias.

Let $\tilde{\mathbf{X}} = P_{\downarrow}(\mathbf{X})$, where $P_{\downarrow} : \mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{d_{\text{model}}/8}$, and let $m = (w-1)/2$ for a temporal window of width $w$. For each window center $\tau$, MOIRAI processes all channels in the local temporal window and STCPE aggregates the overlapping window outputs:

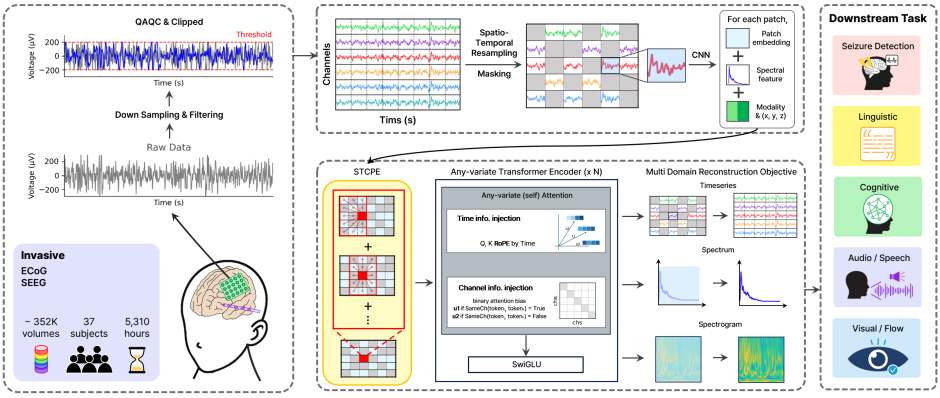


Figure 1: **DIVER-1 architecture, pretraining, and downstream evaluation pipeline.** After preprocessing, input patches are randomly masked and enhanced by adding modality, spectral, and CNN patch encodings, along with STCPE. The patches are then processed through any-variate Transformer encoders and trained to reconstruct missing patches across multiple signal domains. The pretrained model is then applied to diverse downstream tasks.

$$\begin{aligned}\mathbf{H}_\tau &= \mathrm{MOIRAI}\Big(\tilde{\mathbf{X}}_{[:,\,\tau-m:\tau+m,:]}\Big) \in \mathbb{R}^{C\times w\times(d_{\mathrm{model}}/8)},\\ \mathbf{E}_{\mathrm{STCPE}}[:,t,:] &= P_{\uparrow}\left(\sum_{\tau:\,|t-\tau|\le m} \mathbf{H}_\tau[:,\,t-\tau+m,:]\right) \in \mathbb{R}^{C\times d_{\mathrm{model}}}, \qquad t=1,\dots,N.\end{aligned} \tag{1}$$

Here, $t-\tau+m$ indexes the relative temporal offset within the window, and boundary handling is omitted for clarity. STCPE therefore preserves temporal translation equivariance through sliding windows and channel permutation equivariance through MOIRAI, allowing DIVER-1 to encode local spatio-temporal context without learning biases tied to a fixed channel order.

**Position, modality, and spectral embeddings.** We augment each patch with channel metadata and frequency information. Channel metadata is encoded as $[\mathbf{E}_{\text{position}}, \mathbf{E}_{\text{modality}}] = [PE^{(x)}(x), PE^{(y)}(y), PE^{(z)}(z), \mathbf{e}_{\text{type}}]$, where $(x, y, z)$ are MNI coordinates, $\mathbf{e}_{\text{type}}$ denotes electrode type (depth for SEEG; grid/strip for ECoG), and $PE$ is the sinusoidal coordinate encoding from PopT (Chau et al., 2025). Unavailable coordinates are encoded with $\mathbf{E}_{\text{position}} = \mathbf{0}$. We also add $\mathbf{E}_{\text{spectral}}$, obtained by FFT-transforming each patch and linearly projecting the features to $d_{\text{model}}$, following CBraMod (Wang et al., 2024b).

**Any-variate Transformer encoder (MOIRAI).** The enhanced token grid $\mathbf{X} \in \mathbb{R}^{C\times N\times d_{\text{model}}}$ is processed by any-variate Transformer blocks adapted from MOIRAI (Woo et al., 2024). Each token is indexed by temporal patch index $i$ and channel index $m$. **The key design choice is any-variate attention**: instead of restricting attention to temporal-only, channel-only, or axis-factorized interactions, DIVER-1 applies full self-attention over all $CN$ electrode–time tokens. This directly addresses the restricted-attention bottleneck while preserving compatibility with variable iEEG inputs, allowing each channel–time segment to interact with every other segment without assuming a fixed electrode layout.

To make full attention expressive rather than layout-agnostic, DIVER-1 directly embeds spatio-temporal structure into the attention computation itself. Specifically, any-variate attention combines RoPE for temporal offsets with a binary same-channel versus cross-channel attention bias. Omitting layer indices, head indices, and the usual scaling factor, the attention energy and output are

$$\begin{aligned}E_{ij,mn} &= (\mathbf{W}^Q\mathbf{x}_{i,m})^\top \mathbf{R}_{i-j}(\mathbf{W}^K\mathbf{x}_{j,n}) + u^{(1)}\mathbb{1}_{\{m=n\}} + u^{(2)}\mathbb{1}_{\{m\neq n\}},\\ \mathbf{z}_{i,m} &= \sum_{j=1}^{N}\sum_{n=1}^{C} \frac{\exp(E_{ij,mn})}{\sum_{k=1}^{N}\sum_{o=1}^{C}\exp(E_{ik,mo})}\mathbf{W}^V\mathbf{x}_{j,n}.\end{aligned} \tag{2}$$

Here, $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are the attention projections, $\mathbf{R}_{i-j}$ is the RoPE matrix for temporal offset $i-j$, and $u^{(1)}, u^{(2)} \in \mathbb{R}$ are learnable same-channel and cross-channel biases that may differ by layer and attention head. Because the binary bias depends only on whether two tokens originate from the same channel, rather than on the absolute channel identity, the encoder preserves channel permutation equivariance while still distinguishing within-electrode dynamics from cross-electrode interactions. Thus, even when electrode metadata are unavailable, DIVER-1 remains channel-relation aware.

### 2.2 PRETRAINING DIVER-1

DIVER-1 is pretrained with masked multi-domain reconstruction on randomly sampled spatio-temporal subviews of 30 s iEEG segment.

**Spatio-Temporal Resampling (STR).** To improve **robustness to variable electrode coverage and temporal context**, DIVER-1 is pretrained on random spatio-temporal subviews of each 30 s iEEG segment. At each step, we sample $C' \leq \min(C, 32)$ contacts and $N' \leq 30$ temporal patches from a scaled $\mathrm{Beta}(3, 1)$ distribution, favoring larger subviews. The temporal cap keeps the effective context budget comparable across DIVER-1-1s and DIVER-1-0.1s. Details of the sampling procedure are provided in Appendix B.5.

**Multi-Domain Reconstruction Objective (MDRO).** Electrophysiological signals are commonly analyzed as time-domain waveforms, frequency spectra, and time-frequency representations, each capturing transient responses, rhythmic activity, and evolving spectral structure (Luck, 2014; Cohen, 2014; Buzsáki et al., 2012). Following these standard views and recent electrophysiology foundation models (Wang et al., 2023; Zhang et al., 2023; Jiang et al., 2024; Wang et al., 2024b), DIVER-1 reconstructs multiple signal domains rather than only the raw waveform. Details in Appendix B.6.

## 3 EXPERIMENTAL SETUP

**QA/QC and preprocessing.** All iEEG recordings were high-pass filtered at 0.5 Hz, notch filtered at 60 Hz, resampled to 500 Hz, and segmented into 30-second windows. Clipping-based QA/QC removed electrodes when more than $3.33\%$ of samples exceeded the clipping threshold, and discarded segments when more than $50\%$ of channels were affected. After clipping, amplitudes were scaled from $[-200, 200]$ µV to $[-1, 1]$, following Wang et al. (2024b). Further details are provided in Appendix E.3.

**Pretraining dataset and setup.** DIVER-1 was pretrained on a multi-country, cross-center iEEG corpus spanning ECoG and sEEG, comprising $5{,}310$ hours, 352k channel-hours, and 37 subjects. This is $\mathbf{54}\times$ **more channel-hours than the BrainTreeBank corpus used to pretrain BrainBERT, PopT, and BaRISTA**. DIVER-1 also exceeds Brant's 281k channel-hour private SEEG corpus, with an even larger gap in effective training tokens: Brant's 6 s patches yield 169M channel-patch tokens, whereas DIVER-1 yields 1.27B tokens for DIVER-1-1s and 12.7B tokens for DIVER-1-0.1s. Thus, DIVER-1 exposes the model to $\mathbf{7.5}\times$ **and** $\mathbf{75}\times$ **more training tokens** at 1 s and 0.1 s resolution, respectively. This finer temporal granularity is important for subsecond cognitive decoding while preserving coverage of longer-timescale clinical dynamics. Dataset details are summarized in Table 3; comparisons with prior iEEG foundation models are provided in Table 11.

At this scale, DIVER-1-0.1s and DIVER-1-1s were trained across model widths, yielding variants from 13M up to 1.8B parameters (exact per-variant counts in Table 2). Since training the largest 1.8B-parameter variant required up to 128 A100 GPUs, naive hyperparameter sweeps at full scale are computationally prohibitive. We therefore adopt maximal update parameterization ($\mu$P) (Yang et al., 2022), which preserves update magnitudes across model widths and enables $\mu$Transfer: hyperparameters tuned on small proxy models transfer directly to larger ones, letting us scan DIVER-1 from 13M to 1.83B parameters under a single configuration. Details are in Appendix B.4 and B.8.

**Evaluation datasets and tasks.** We evaluate DIVER-1 on iMINDBench naturalistic decoding benchmark (Chau et al., 2026) and the MAYO seizure-detection dataset (Bbrinkm & Cukierski, 2014). iMINDBench harmonizes 15 binary language, auditory, and visual decoding tasks across three naturalistic movie-watching datasets; Neuroprobe/Brain Treebank (NP), Bang! You're Dead (BYD),

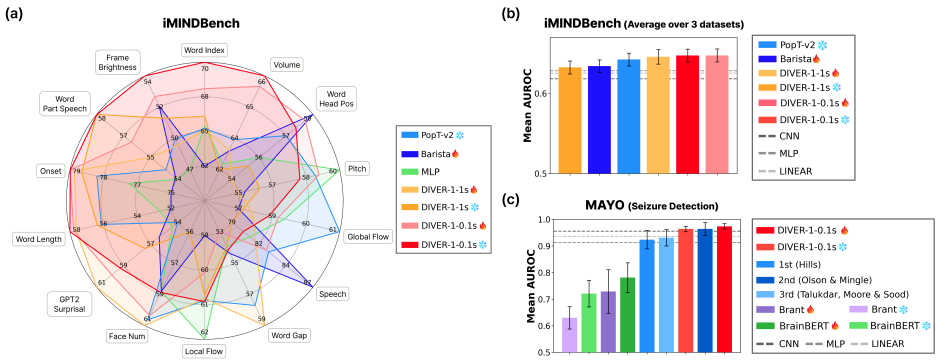


Figure 2: **DIVER-1 achieves strong iEEG decoding across naturalistic and clinical benchmarks.** (a) Per-task iMINDBench performance across the three constituent datasets (NP, BYD, and Pippi), comparing the four DIVER-1 variants (0.1s/1s, frozen and fully finetuned), PopT-v2 (Chau et al., 2026), and the best-performing non-foundation-model baseline (multi-STFT + MLP). (b) Overall iMINDBench performance; bars show foundation models, and dashed lines denote non-foundation-model baselines. (c) AUROC on MAYO seizure detection using 6 s windows, comparing the DIVER-1 variants with foundation-model baselines (BrainBERT and Brant); dashed lines denote non-foundation-model baselines (full comparison in Appendix Table 10). Error bars denote participant-level SEM. Snowflake denotes frozen-encoder linear probing, and fire icons denote full finetuning.

and Pippi, using fixed two-fold within-session splits and 1-second neural windows (Chau et al., 2026). Dataset details are provided in Appendix E.2.

**Baselines and finetuning.** We compare DIVER-1 against four self-supervised iEEG foundation models—BrainBERT (Wang et al., 2023), PopT-v2 (Chau et al., 2026), Brant (Zhang et al., 2023), and BaRISTA (Oganesian et al., 2025)—as well as benchmark-reported supervised baselines when available. These models are evaluated with linear probing or full finetuning using a linear classifier on flattened token representations. For iMINDBench, our frozen evaluation follows the benchmark's Logistic protocol, fitting a logistic-regression classifier on flattened aligned-channel representations with settings fixed across datasets and tasks (Chau et al., 2026). On MAYO, we additionally evaluate six non-foundation-model baselines: CNN, MLP, and linear decoders on STFT spectrograms, and reimplementations of the three top-ranked algorithms from the original challenge leaderboard (Baldassano et al., 2017). Details are provided in Appendix B.9.

## 4 Results and Discussion

### 4.1 DIVER-1 learns transferable iEEG representations

**DIVER-1 achieves state-of-the-art iEEG decoding under cross-dataset shift.** DIVER-1-0.1s achieves highest performance on iMINDBench and MAYO. On the iMINDBench Main decodability subset, both frozen and full-finetuning results of DIVER-1-0.1s outperform all competing models (Overall ROC-AUC $0.648$ vs. $0.645$ for PopT-v2; Figure 2(b), Table 9). Frozen features alone match full finetuning (Overall $0.648$ for both), indicating that the pretrained representation itself, rather than downstream adaptation, drives the gain. Notably, PopT-v2 (Chau et al., 2026), the strongest prior model, is pretrained on Brain Treebank—one of the benchmark's three constituent datasets—and thus on naturalistic movie-watching data matching the benchmark's behavioral context, whereas DIVER-1 is pretrained exclusively on task-free clinical monitoring recordings from external institutions. DIVER-1 nevertheless transfers simultaneously to all three institutions' datasets (NP, BYD, and Pippi), demonstrating the best reported decoding performance under genuine cross-dataset shift.

On MAYO seizure detection, fully finetuned DIVER-1-0.1s attains the highest mean AUROC among all evaluated configurations ($0.974 \pm 0.010$), outperforming prior iEEG foundation models and performing on par with the strongest hand-engineered classical pipelines ($0.964 \pm 0.024$; Figure 2c;

full comparison in Appendix Table 10). Despite operating on 6-second clinical windows—rather than iMINDBench's rapid sub-second events—the 0.1-second-patch model performs best on MAYO, indicating that DIVER-1 transfers across task timescales rather than being tied to its pretraining patch scale. The 1-second variant remains a practical alternative, using an order of magnitude fewer tokens while approaching the best MAYO AUROC ($0.959 \pm 0.020$), making it well suited to compute-constrained or long-recording settings where sub-second resolution is unnecessary.

Table 1: **Architecture and pretraining strategy ablation.** Single-component ablations of DIVER-1-0.1s on representative iMINDBench tasks. Rows remove components from three groups: encoder attention, input embeddings, and pretraining strategy. All models are $d_{\text{model}} = 256$ and pretrained for 8 epochs, and evaluated in logistic regression of frozen features. Overall is the unweighted mean of the NP, BYD, and Pippi dataset-level AUROC means. The best reported value in each column is bold.

| | | NP | BYD | Pippi | Overall |
|---|---|---|---|---|---|
| **Atten.** | w.o. RoPE | 72.40 | 62.53 | 57.25 | 64.06 |
| | w.o anyV attention | 71.96 | 62.25 | 56.48 | 63.56 |
| | w.o RoPE and anyV attention (Vanilla Attention) | 70.06 | 61.13 | 56.70 | 62.63 |
| **Embedding** | w.o. STCPE | 72.04 | 62.05 | 56.29 | 63.46 |
| | w.o Channel sub-modality emb. | 72.05 | 62.05 | 57.87 | 63.99 |
| | w.o Channel 3d position emb. | 72.97 | 62.70 | 56.72 | 64.13 |
| | w.o Spectral feature emb. | 72.15 | 62.02 | 56.19 | 63.46 |
| **Pretrain.** | w.o Multi-domain reconstruction (only raw) | 72.00 | 61.53 | 56.05 | 63.20 |
| | w.o channel and time resampling | 69.64 | 59.46 | 56.96 | 62.02 |
| | w.o channel resampling | 70.56 | 59.59 | 56.27 | 62.14 |
| | w.o time resampling | 70.67 | 60.18 | 56.36 | 62.40 |
| DIVER-1-0.1s | | **73.09** | **63.02** | **58.17** | **64.76** |

**Architecture and pretraining strategy ablation.** Table 1 reports component-wise ablations of DIVER-1-0.1s. The largest changes come from the resampling ablations. Replacing DIVER-1's attention with vanilla self-attention consistently reduces performance, while removing RoPE or the any-variate same/cross-channel bias alone has smaller, task-dependent effects, suggesting complementary benefits. Among embeddings and objectives, STCPE and spectral features provide modest gains, channel sub-modality has limited impact, and MDRO improves over raw-only reconstruction.

The resampling ablations show that stochastic resampling is critical for effective transfer. In particular, resampling substantially improves performance over pretraining on fixed windows, with clear benefits along both the temporal and channel dimensions.

### 4.2 Scaling laws reveal a data-constrained iEEG regime

We sweep training compute, data tokens, model size, training duration, and subject diversity to ask whether self-supervised iEEG pretraining scales predictably, and which axes matter most for downstream transfer.

**Scaling-law setup.** Because iEEG data are scarce and multi-pass training is often unavoidable, we fit the data-constrained scaling law of Muennighoff et al. (2023), which extends Kaplan-style power laws (Kaplan et al., 2020a) by replacing raw data and parameter counts with effective counts. Specifically, we model loss as $L(N, D) = A(N')^{-\alpha} + B(D')^{-\beta} + E$, with $D' = g(U_D, R_D, R_D^*)$, $N' = g(U_N, R_N, R_N^*)$, and $g(U, R, R^*) = U[1+R^*(1-e^{-R/R^*})]$. Here, $U_D$ and $R_D$ denote unique tokens and additional passes over them, while $R_D^*$ controls diminishing returns from repetition; $U_N, R_N, R_N^*$ play the analogous role for excess model capacity. This lets us test whether iEEG pretraining scales predictably across data, model size, and training duration. Details can be found in Appendix A.2.

**Fixed-duration sweeps show predictable pretraining scaling.** At fixed training duration, held-out loss decreases approximately linearly on log–log axes as compute, data size, and model size increase (Figure 3a), consistent with Kaplan-style scaling (Kaplan et al., 2020a) and suggesting that **masked iEEG reconstruction scales predictably**. Larger models, however, reduce loss only after sufficient training passes—at short durations they remain undertrained (Figure 3a, rightmost Epochs panel). Thus, **model capacity helps only when supported by sufficient training duration**, motivating the compute-optimal allocation analysis below.

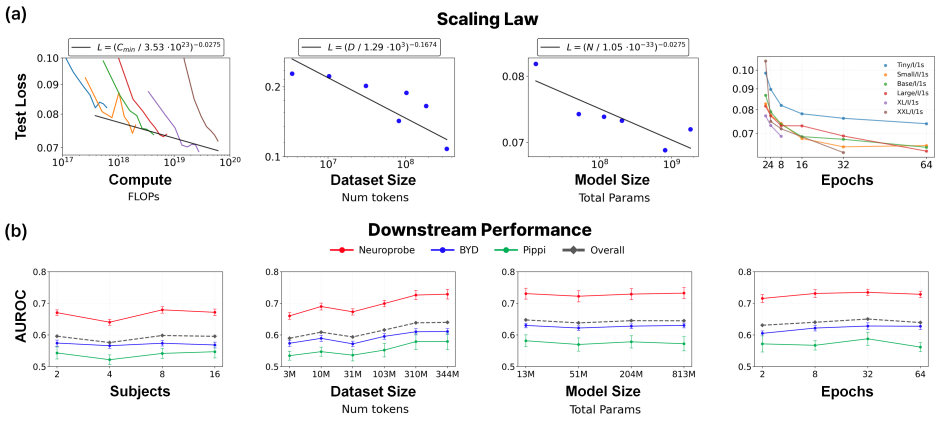


Figure 3: **Scaling laws and downstream scaling of DIVER-1.** (a) Held-out pretraining loss of DIVER-1-1s follows Kaplan et al. (2020a)-style power-law trends as training compute, data, and model parameters increase. Training-duration curves show that larger parameter-count variants require sufficient repeated exposure to the data before outperforming smaller variants, consistent with data-constrained scaling (Muennighoff et al., 2023). See Appendix C.2 for DIVER-1-0.1s and extended results. (b) Downstream dataset-level AUROC on the three iMINDBench datasets (NP, BYD, and Pippi) and their Overall mean, under controlled scaling of subject count, dataset size, model size, and epochs(training duration) on DIVER-1-0.1s. Dataset size and training duration yield the most consistent gains, whereas model size and subject diversity show no clear trend under fixed data budgets.

**Downstream scaling favors data and training duration over parameter count.** Figure 3(b) shows that increasing dataset size gives the most reliable downstream gains across tasks in iMINDBench, whereas downstream performance remains similar across model sizes, in line with Yang et al. (2026): although pretraining loss scales along the model-size axis, those gains do not carry over downstream—pretraining loss does not predict downstream performance. This pattern is also reflected in the pretraining scaling curves in Figure 3(a): the dataset-size axis produces a larger reduction in held-out reconstruction loss than the model-size axis. Together, these results suggest that DIVER-1 is currently more data-limited than parameter-limited. Notably, additional unique recordings appeared most useful when they broadened the population, center, and implant distributions, rather than through repeated exposure to the same data.

Training duration is another reliable scaling axis: AUROC generally improves up to 32 training passes, with mild decline at 64. Subject diversity shows no clear scaling trend under fixed total pretraining volume. Together, these results motivate the compute-optimal analysis below: **for iEEG foundation models, unique data and sufficient training duration should be prioritized before scaling parameter count**.

Although pretraining loss follows the Kaplan form for both variants, it does not reliably predict downstream decoding performance across our sweep. This result is specific to masked reconstruction, a common EEG foundation-model objective (Kuruppu et al., 2025), and motivates exploring alternative self-supervised objectives rather than questioning scaling in iEEG broadly. Achieving language-model-like scaling gains in iEEG may therefore require design choices tailored to neural data, beyond increases in parameters and compute. To our knowledge, this is the first empirical characterization of what additional data and compute can and cannot deliver for iEEG pretraining.

**Compute-optimal allocation.** The scaling-law fits give two practical compute-allocation rules for iEEG. Classical compute-optimal analyses ask how to split compute between data and parameters when fresh data can grow with compute (Kaplan et al., 2020a; Hoffmann et al., 2022a). In iEEG, however, data is usually fixed, so extra compute can only be spent on larger models or more training passes over the same recordings.

**First, scale data before parameters when new recordings are available.** Figure 4(a) shows that at an epoch of two, DIVER-1-1s favors a much more data heavy allocation than LLMs (Kaplan et al.,

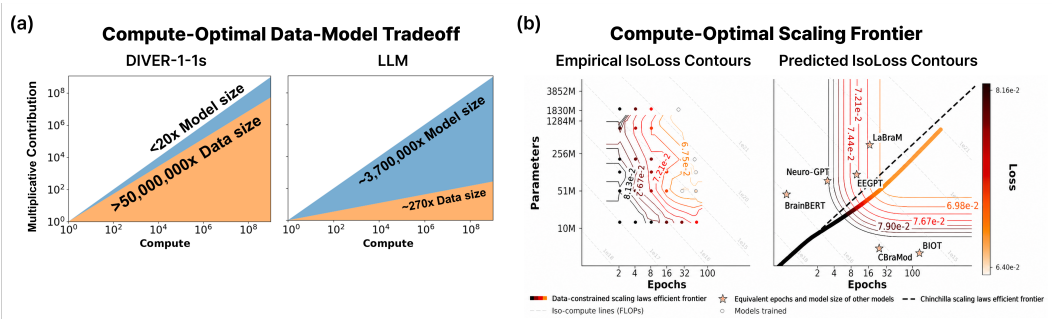


Figure 4: **Compute-optimal scaling of DIVER-1.** (a) **Optimal data–model allocation when training duration is fixed.** With training fixed to two passes, we use the fitted DIVER-1-1s scaling law to estimate how the compute-optimal dataset size and parameter count should grow as total compute increases. The y-axis shows the fold-increase assigned to each factor. Compared with LLM scaling of Kaplan et al. (2020a), DIVER-1-1s allocates much more of the compute increase to data growth and much less to parameter growth. (b) **Optimal parameter–training-duration allocation when the corpus is fixed.** Using fixed pretraining corpus, we plot held-out loss over model size and number of training passes. IsoCompute curves mark equal FLOPs, and IsoLoss contours show equal reconstruction loss. The efficient frontier favors small-to-mid-sized models trained longer over largest models trained briefly. See Appendix C.3 for result of DIVER-1-0.1s variant.

2020a). DIVER-1-0.1s shows the same trend (Appendix C.2), indicating that better iEEG foundation models depend primarily on expanding unique pretraining recordings, not simply adding parameters.

**Second, when the corpus is fixed, scale training duration before model size.** Figure 4(b) shows that realistic compute budgets favor small-to-mid-sized models trained for more passes, rather than very large models trained briefly. The predicted contours from Eq. A.2 reproduce this frontier, suggesting that the fitted law can guide parameter–duration choices before expensive pretraining. Overall, the compute-efficient order is: **increase unique data first, train sufficiently long second, and scale parameters last**. Fitted parameters and additional plots are provided in Appendix C.3.

## 5 Conclusion

We introduced DIVER-1, a self-supervised iEEG foundation model designed to learn transferable representations from heterogeneous clinical recordings. By combining an iEEG-specific architecture with large-scale external pretraining, DIVER-1 improves cross-dataset transfer on both naturalistic decoding and clinical seizure detection, outperforming prior iEEG foundation models and the strongest hand-engineered spectral baselines on iMINDBench, while matching the best classical pipelines on MAYO—an encouraging step forward for the field. Our scaling study further shows that iEEG masked-reconstruction pretraining remains data-constrained: pretraining data and longer training are more reliable drivers of performance than parameter growth alone. At the same time, pretraining loss does not consistently predict downstream decoding performance. This suggests that simply scaling parameters or compute may not be enough to close the remaining gap; instead, we expect more diverse recordings and better-aligned training objectives to matter most. These results suggest a practical path for future iEEG foundation models: scale data first, train longer second, and scale model size only after those bottlenecks are addressed. As public iEEG resources grow toward the scale studied here, we hope this work serves as a milestone, and its scaling analysis as a recipe and a basis for independent reproduction, for the next generation of iEEG foundation models.

### AI USE STATEMENT

In this work, we used generative AI tools for editing (grammar, spelling, and word choice), for drafting portions of the manuscript text under author direction, for retrieval and discovery of related work, for assistance in writing experiment code, executing experiments, and aggregating results during analysis, and for assistance in visualizing results for submission. We did not use generative AI tools for generating research ideas, designing experiments, or interpreting the results. All AI-assisted text, code, and analysis outputs were reviewed, verified against the underlying experimental records, and revised by the authors. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses intracranial EEG recorded from patients implanted for clinical reasons. The public pretraining and evaluation corpora (AJILE12, BrainTreebank/Neuroprobe, and the MAYO seizure-detection challenge data) were collected and released by their original custodians under their own consent and governance procedures, and we use them under their stated terms. Our self-collected clinical recordings were acquired with local institutional review board approval.

## REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902, 2019. URL http://arxiv.org/abs/1907.10902.

Steven N Baldassano, Benjamin H Brinkmann, Hoameng Ung, Tyler Blevins, Erin C Conrad, Kent Leyde, Mark J Cook, Ankit N Khambhati, Joost B Wagenaar, Gregory A Worrell, and Brian Litt. Crowdsourcing seizure detection: algorithm development and validation on human implanted device recordings. *Brain*, 140(6):1680–1691, 2017.

Hubert Banville, Yohann Benchetrit, Stéphane d'Ascoli, Jérémy Rapin, and Jean-Rémi King. Scaling laws for decoding images from brain activity, 2025. URL https://arxiv.org/abs/2501.15322.

S Bbrinkm and W Cukierski. Upenn and mayo clinic's seizure detection challenge, 2014.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

Philipp Bomatter and Henry Gouk. Is limited participant diversity impeding EEG-based machine learning?, 2025. URL https://arxiv.org/abs/2503.13497.

Rishi Bommasani, Drew A Hudson, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

Nicholas S. Card, Maitreyee Wairagkar, others, Sergey D. Stavisky, and David M. Brandman. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391 (7):609–618, 2024. doi: 10.1056/NEJMoa2314132.

Francesco Carzaniga, Michael Hersche, Abu Sebastian, Kaspar Schindler, and Abbas Rahimi. A foundation model with multi-variate parallel attention to generate neuronal activity. *arXiv preprint arXiv:2506.20354*, 2025.

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. *ArXiv*, pp. arXiv–2406, 2025.

Geeling Chau, Saba Hashemi, Yonghyeon Gwon, Eshani Patel, Jan DeWitt, Christopher Wang, Andrii Zahorodnii, Sabera J Talukder, Danny Dongyeop Han, Chun Kee Chung, Maryam M Shanechi, and Yisong Yue. imindbench: ieeg multi-institution neural decoding benchmark, 2026. URL https://arxiv.org/abs/2609.18104.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. *arXiv preprint arXiv:2401.10278*, 2024.

Mike X. Cohen. *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press, 2014. URL https://direct.mit.edu/books/monograph/4013/analyzing-neural-time-series-datatheory-and.

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

Gustavo Deco, Viktor K Jirsa, and Anthony R McIntosh. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature reviews neuroscience*, 12(1):43–56, 2011.

Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6):254–262, 2011.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011.

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

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022a. URL https://arxiv.org/abs/2203.15556.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *arXiv preprint*, 2022b.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

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

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

Ankit N Khambhati, Ari E Kahn, Julia Costantini, Youssef Ezzyat, Ethan A Solomon, Robert E Gross, Barbara C Jobst, Sameer A Sheth, Kareem A Zaghloul, Gregory Worrell, et al. Functional control of electrophysiological network architecture using direct neurostimulation in humans. *Network Neuroscience*, 3(3):848–877, 2019.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, Sandipan Pati, Gregory Worrell, and Yogatheesan Varatharajah. Eeg foundation models: A critical review of current progress and future directions, 2025. URL https://arxiv.org/abs/2507.11783.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG foundation models: Progresses, benchmarking, and open problems, 2026. URL https://arxiv.org/abs/2601.17883.

Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, 2 edition, 2014. URL https://mitpress.mit.edu/9780262525855/an-introduction-to-the-event-related-potential-technique/.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 2014.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Lucine L Oganesian, Saba Hashemi, and Maryam M Shanechi. Barista: Brain scale informed spatiotemporal representation of human intracranial neural activity. *arXiv preprint arXiv:2512.12135*, 2025.

Steven M Peterson, Satpreet H Singh, Benjamin Dichter, Michael Scheid, Rajesh PN Rao, and Bingni W Brunton. Ajile12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific data*, 9(1):184, 2022.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

Josef P Rauschecker and Sophie K Scott. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6):718–724, 2009.

Katherine W Scangos, Ankit N Khambhati, Patrick M Daly, Ghassan S Makhoul, Leo P Sugrue, Hashem Zamanian, Tony X Liu, Vikram R Rao, Kristin K Sellers, Heather E Dawes, et al. Closed-loop neuromodulation in an individual with treatment-resistant depression. *Nature medicine*, 27 (10):1696–1700, 2021.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022. URL http://jmlr.org/papers/v23/20-1111.html.

Fanqi Shen, Enhong Yang, Jiahe Li, Junru Hong, Xiaoran Pan, Zhizhang Yuan, Meng Li, and Yang Yang. Brain4FMs: A benchmark of foundation models for electrical brain signal, 2026. URL https://arxiv.org/abs/2602.11558.

Cornelis J Stam, Guido Nolte, and Andreas Daffertshofer. Phase lag index: assessment of functional connectivity from multi channel eeg and meg with diminished bias from common sources. *Human brain mapping*, 28(11):1178–1193, 2007.

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

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

Maitreyee Wairagkar, Nicholas S. Card, Tyler Singer-Clark, others, David M. Brandman, and Sergey D. Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, 644(8075):145–152, 2025. doi: 10.1038/s41586-025-09127-3.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023.

Christopher Wang, Adam Yaari, Aaditya Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph Madsen, Scellig Stone, Gabriel Kreiman, et al. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. *Advances in Neural Information Processing Systems*, 37:96505–96540, 2024a.

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. *arXiv preprint arXiv:2412.07236*, 2024b.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *International Conference on Machine Learning*, 2024.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

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

Andrii Zahorodnii, Christopher Wang, Bennett Stankovits, Charikleia Moraitaki, Geeling Chau, Andrei Barbu, Boris Katz, and Ila R Fiete. Neuroprobe: Evaluating intracranial brain responses to naturalistic stimuli. *arXiv preprint arXiv:2509.21671*, 2025.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36:26304–26321, 2023.

Hui Zheng, Hai-Teng Wang, Wei-Bang Jiang, Zhong-Tao Chen, Li He, Pei-Yang Lin, Peng-Hu Wei, Guo-Guang Zhao, and Yun-Zhe Liu. Du-in: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. *Advances in Neural Information Processing Systems*, 37: 79996–80033, 2024.

# A APPENDIX

## APPENDIX TABLE OF CONTENTS

**Section A:** Related Works
- A.1 Electrophysiology Data Foundation Models
- A.2 Neural Scaling Law

**Section B:** Experimental Setup Details
- B.1 Tested Models
- B.2 Pretraining Dataset
- B.3 Downstream Task and Dataset Overview
- B.4 Pretraining Setup and Model Scaling
- B.5 Spatio-Temporal Resampling
- B.6 DIVER Architecture and Pretext Task Hyperparameter Setting
- B.7 Pretraining Hyperparameter Search using $\mu$-Parameterization and $\mu$-Transfer
- B.8 $\mu$-parameterization ($\mu P$)
- B.9 Finetuning Setup

**Section C:** Scaling Law
- C.1 Scaling Law Experiment Details
- C.2 Extended Kaplan Scaling Law Results
- C.3 Data-Constrained Scaling Law Results

**Section D:** Extended Results
- D.1 Comprehensive DIVER Downstream Task Results
- D.2 Interpretation Results

**Section: E** Data Details
- E.1 Pretraining Dataset Description
- E.2 Finetuning Dataset Description
- E.3 QAQC and Preprocessing

**Section F:** Comparison With Existing EEG/iEEG Foundation Models

# A Related Works

## A.1 Electrophysiology Data Foundation Models

Ephys decoding has progressed from pipelines that coupled hand-crafted features with classical classifiers to end-to-end deep architectures that learn task-relevant representations directly from raw signals. Early EEG studies relied on feature engineering (e.g., band-power, CSP) paired with SVMs or LDA; subsequent work introduced convolutional backbones and sequence models that absorb spectral–temporal patterns with minimal preprocessing. Building on advances in self-supervised learning (SSL) and "foundation" paradigms, recent efforts pretrain large models on heterogeneous, weakly labeled or unlabeled EEG/iEEG corpora and adapt them to diverse downstream tasks (e.g., event detection, cognitive state decoding, BCI control). Representative lines include large-scale EEG transformers and montage-aware encoders (e.g., EEGFormer(Chen et al., 2024), NEURO-GPT(Cui et al., 2024), LaBraM(Jiang et al., 2024)), intracranial representation learners (e.g., BrainBERT(Wang et al., 2023), foundation models for intracranial neural signals), and broader neuro-sequence backbones emphasizing transfer and robustness (e.g., CBraMod(Wang et al., 2024b)).

## A.2 Neural Scaling Law

**Kaplan (Kaplan et al., 2020a) scaling law** provides a principled framework for predicting model performance as a function of model size and dataset size for large language models. They demonstrated that language model cross-entropy loss follows smooth power-law relationships with respect to model parameters ($N$) and training data ($D$). Concretely, they proposed a relation of the form

$$L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E \tag{3}$$

where $A, B, E$ are constants, and $\alpha, \beta > 0$ are scaling exponents. This formulation implies that increasing the number of parameters or training data yields a predictable reduction in loss, enabling systematic optimization of compute allocation across model size and training duration.

**Data-constrained scaling law.** In scientific domains, including Ephys, unique high-quality data are inherently limited. To address this, Muennighoff et al. (2023) extended the scaling framework to data-constrained regimes, where models must repeatedly train on the same corpus for multiple epochs. They proposed a modified law

$$L(N, D) = \frac{A}{(N')^{\alpha}} + \frac{B}{(D')^{\beta}} + E \tag{4}$$

where the effective number of parameters $N'$ and effective tokens $D'$ account for diminishing returns: repeated tokens are progressively less valuable, and excessively large models are less sample-efficient. They further define

$$D' = U_D + U_D R_D^* \left(1 - e^{-\frac{R_D}{R_D^*}}\right) \tag{5}$$

$$N' = U_N + U_N R_N^* \left(1 - e^{-\frac{R_N}{R_N^*}}\right) \tag{6}$$

where $U_D$ is the number of unique tokens, $R_D$ is the number of repetitions (epochs$-1$), and $R_D^*$ is a learned constant describing the "half-life" of repeated data. Analogously, $U_N$ is the compute-optimal parameter count for $U_D$, and $R_N^*$ governs diminishing returns beyond that point. Empirical results suggest that up to $\sim$4 epochs, repeated data is almost as useful as new data, but returns decrease sharply thereafter.

Compared to the Chinchilla law (Hoffmann et al., 2022a), which assumes abundant data and one training epoch, this formulation makes epoch count itself a central axis of scaling. This perspective is crucial for domains constrained by limited samples, such as iEEG, and it guides how compute should be allocated between model size and additional training passes.

Finally, the data-constrained scaling law also connects scaling analysis to more fundamental theory. Recent work has suggested that the exponents governing power-law scaling, such as $\alpha$ and $\beta$, are intimately connected to the *intrinsic dimension* of the underlying data manifold (Sharma & Kaplan,

2022). In particular, it has been argued that the data exponent $\beta$ can be interpreted in terms of an effective dimension $d$ via an approximate inverse relation of the form:

$$\beta \approx \frac{2}{d}. \tag{7}$$

From this view, estimating $\beta$ provides not only a measure of how data volume translates into performance but also an insight into the intrinsic dimensionality of iEEG signals, where scarcity and structural complexity are defining features.

# B Experimental Setup Details

## B.1 Tested Models

We adopt a systematic naming convention for all DIVER model variants: DIVER-1-Granularity$_{\text{Size}}$. The *Granularity* indicates temporal resolution (0.1s or 1s window). The *Size* component indicates model size (Tiny: 256, Small: 512, Base: 768, Large: 1024, XL: 2048, XXL: 3072 hidden dimensions). For example, DIVER-1-1S$_{\text{Base}}$ represents a base-sized model trained with 1s temporal patchsize. Unless otherwise specified, DIVER-1-1or0.1S refers to the Tiny variant, i.e., DIVER-1-1or0.1S$_{\text{Tiny}}$.

Table 2: **Model configurations with measured parameters and total FLOPs per epoch.**

| Models | # Parameters | Modality | Hidden Dimension ($d_{\text{model}}$) | total FLOPs / epoch |
|---|---|---|---|---|
| DIVER-1-0.1S $_{\text{Tiny}}$ | 12.72M | 0.1s | 256 | 76.34P |
| DIVER-1-0.1S $_{\text{Small}}$ | 50.75M | 0.1s | 512 | 253.96P |
| DIVER-1-0.1S $_{\text{Base}}$ | 114.07M | 0.1s | 768 | 532.77P |
| DIVER-1-0.1S $_{\text{Large}}$ | 202.70M | 0.1s | 1024 | 912.83P |
| DIVER-1-0.1S $_{\text{XL}}$ | 810.19M | 0.1s | 2048 | 3.40E |
| DIVER-1-0.1S $_{\text{XXL}}$ | 1.82B | 0.1s | 3072 | 7.50E |
| DIVER-1-1S $_{\text{Tiny}}$ | 13.03M | 1s | 256 | 77.52P |
| DIVER-1-1S $_{\text{Small}}$ | 51.36M | 1s | 512 | 256.44P |
| DIVER-1-1S $_{\text{Base}}$ | 115.00M | 1s | 768 | 536.82P |
| DIVER-1-1S $_{\text{Large}}$ | 203.95M | 1s | 1024 | 918.56P |
| DIVER-1-1S $_{\text{XL}}$ | 812.85M | 1s | 2048 | 3.46E |
| DIVER-1-1S $_{\text{XXL}}$ | 1.83B | 1s | 3072 | 7.64E |

## B.2 Pretraining Dataset

DIVER-1 was pretrained on the largest iEEG corpus to date, comprising 352k channel-hours from 37 subjects using iEEG data (ECoG/sEEG). The pretraining dataset description is given in Table 3. Private data were collected from patients who provided informed consent to participate in an IRB-approved study.

Table 3: **Summary of DIVER-1 pretraining datasets.**

| Datasets | Data Type | # Subj. | Volume (channel-hours) | Duration (hours) | Sampling Rate (Hz) |
|---|---|---|---|---|---|
| AJILE12 (Peterson et al., 2022) | ECoG | 12 | 124,423 | 1,282 | 1,000 |
| Self-collected iEEG (DIVER$_{\text{I}}$) | ECoG/sEEG | 25 | 227,612 | 4,028 | 2,000 |
| **Total** | — | 37 | 352,035 | 5,310 | — |

### B.3 Downstream Task and Dataset Overview

Table 4 provides a comprehensive overview of all downstream tasks and datasets used in our evaluation. Our evaluation covers diverse neural decoding objectives across visual, auditory, language domains and seizure detection.

**iMINDBench.** We evaluate naturalistic decoding on the three iMINDBench datasets—Neuroprobe/Brain Treebank (NP), Bang! You're Dead (BYD), and Pippi—using 15 binary targets spanning language, audition, and vision (Chau et al., 2026). Table 4 summarizes the dataset-level specifications; Section E.2 provides the detailed dataset and target definitions.

**MAYO.** We evaluate seizure detection on a custom human-only construction of the MAYO Kaggle challenge data (Bbrinkm & Cukierski, 2014). It uses 6 s segments formed from the original 1 s clips and contains the challenge's 8 human participants; Section E.2 details its construction and comparability limitations.

Table 4: **Overview of downstream tasks and datasets.** Sampling rates were adjusted to 500Hz across all datasets to match the pretraining configuration. Arrows (→) indicate resampling or channel selection from the original dataset. For iMINDBench, channel counts are the mean numbers of selected electrodes per subject, and subject/session counts follow the benchmark's evaluation units. MAYO contains both ECoG and SEEG recordings, but per-channel modality labels are not provided. For BYD, SEEG are recorded through Macroelectrodes on hybrid Behnke-Fried depth electrodes.

| Modality | Task Name | Datasets | Sampling Rate | Mean # Ch. | # Subj./sess. | Label |
|---|---|---|---|---|---|---|
| **sEEG** | NP | iMINDBench | 2048 → 500Hz | 118 | 5 | 15 binary tasks |
| **sEEG*** | BYD | iMINDBench | 1000 → 500Hz | 77 | 29 | 15 binary tasks |
| **sEEG** | Pippi | iMINDBench | 2048 → 500Hz | 110 | 5 | 15 binary tasks |
| **Mixed (ECoG+SEEG)** | Seizure Detection | MAYO | 5000 → 500Hz | Var.(16–72) | 8 | 2-class |

### B.4 Pretraining Setup and Model Scaling

Training experiments were conducted across two high-performance computing configurations. The primary server consisted of nodes each equipped with a single 2.8 GHz AMD EPYC Milan 7543P 32-core CPU and four NVIDIA A100 GPUs, which was more heavily utilized throughout the training process. For large model variants, we additionally employed a secondary server equipped with dual Intel Xeon Platinum 8480+ processors (112 cores total) and eight NVIDIA H200 GPUs with 144GB memory each. Training experiments were conducted using either 16,32,128 A100 GPUs or 32, 24, 16 H200 GPUs depending on the experimental configuration. We maintained a fixed global batch size of 192 across all training runs, with the per-GPU batch size adjusted dynamically based on the number of nodes employed.

We varied the model size by modifying the hidden dimension of the transformer, resulting in sizes of 13M, 51M, 115M, 203M, 813M, 1.83B parameters, while keeping the depth fixed at 12 layers. This capacity adjustment leverages the benefits of $\mu$ parameterization for stable training across different model sizes. DIVER-1 was implemented on the Python 3.12.3 and Pytorch 2.6.0 + cuda version 12.4. To enhance training efficiency, we employed DeepSpeed ZeRO Stage 2, BF16 precision. Optimization was performed using a custom implementation of the DeepSpeed's MuAdam optimizer (Yang et al., 2022) with utilizing DeepSpeed's FusedAdam backend (Rasley et al., 2020) for computational efficiency and learning rate calibration. A cosine annealing learning rate scheduler with warm-up restarts was applied, with cycle length matching the total training steps and minimum learning rate set to 0.01× the initial rate.

### B.5 Spatio-Temporal Resampling

STR is applied during self-supervised pretraining before masking and reconstruction. Each pretraining example is a 30 s iEEG segment sampled at 500 Hz. For DIVER-1-1s, the patch length is $P = 500$ samples, yielding $N_0 = 30$ temporal patches; for DIVER-1-0.1s, the patch length is $P = 50$ samples, yielding $N_0 = 300$ temporal patches.

Given a segment with $C$ contacts, STR first defines the maximum sampled channel and temporal-token budgets as

$$C_{\max} = \min(C, 32), \qquad N_{\max} = \min(N_0, B_{\max}),$$

where $B_{\max} = 30$ for DIVER-1-1s and $B_{\max} = 3$ for DIVER-1-0.1s.

We then independently sample

$$u_C, u_N \sim \mathrm{Beta}(3, 1),$$

and set

$$C' = \max(1, \lceil C_{\max} u_C \rceil), \qquad N' = \max(1, \lceil N_{\max} u_N \rceil).$$

The $\mathrm{Beta}(3, 1)$ distribution has density concentrated toward larger values, so the model usually observes large subviews while still being exposed to smaller channel subsets and shorter temporal contexts. We sample $C'$ contacts uniformly without replacement. For the temporal axis, we sample a contiguous crop of $N'$ patches by drawing a start index

$$\tau \sim \mathrm{Uniform}\{1, \ldots, N_0 - N' + 1\},$$

and retaining patches $\{\tau, \ldots, \tau + N' - 1\}$. The resulting subview has shape $C' \times N'$.

This resampling serves two purposes. First, random contact subsets expose the encoder to the variable electrode coverage typical of clinical iEEG, where patients differ in channel count, anatomical sampling, and implant geometry. Second, random temporal crops expose the encoder to variable context lengths.

After STR, masked reconstruction is performed only on the retained $C' \times N'$ token grid. STR is used only during pretraining; during downstream finetuning and evaluation, each task uses its native temporal window and available channel set.

### B.6 DIVER Architecture and Pretext Task Hyperparameter Setting

**Architecture setting** Table 5 lists the detailed architectural hyperparameter settings used for DIVER-1 pretraining.

**FFT, STFT setting** For the FFT, we used a window size of 500 time points with a sampling frequency of 500 Hz. A cutoff frequency of 200 Hz was applied, and the FFT amplitudes were converted to absolute values, normalized, and then compressed using a log(1 + x) transform. For the STFT, we employed a multi-resolution approach with window sizes of 200 and 100 time points, respectively. Each window was shifted with 50% overlap and tapered with a Hann window function. Consistent with the FFT settings, a cutoff frequency of 200 Hz was applied, and the STFT amplitudes were converted to absolute values, normalized, and compressed using the log(1 + x) transform.

**MDRO setting** The total objective is a weighted reconstruction loss over masked patches and domains, $\mathcal{L}_{\text{total}} = \sum_{(c,n)\in\mathcal{M}} \sum_{k\in\mathcal{K}} \lambda_k \mathcal{L}_{\text{MSE}}(\mathbf{y}_{c,n}^{(k)}, \hat{\mathbf{y}}_{c,n}^{(k)})$, where $\mathcal{K} = \{\mathrm{raw}, \mathrm{FFT}, \mathrm{STFT}\}$, $\lambda_k$ denotes the domain weight, and $\mathcal{M}$ denotes masked patch indices. We use $(1, 0.1, 1)$ for DIVER-1-1s and $(1, 1, 0)$ for DIVER-1-0.1s, choosing weights to keep losses on comparable numerical scales; for $P = 50$, the STFT term is omitted because the window is too short for a meaningful spectrogram target.

Table 5: **Hyperparameters for DIVER-1 pretraining.** Two model variants were trained for inputs with 1s and 0.1s patch size respectively. The 1s and 0.1s models share all settings except for patch size, patchwise CNN embedding settings, and SSL weights. Some hyperparameters are defined as a function of $d_{\text{model}}$, which we vary across {256, 512, 768, 1024, 2048, 3072}.

| | Hyperparameters | Settings |
|---|---|---|
| Input & Masking | Patch size | 500 (1s)<br>50 (0.1s) |
| | Mask ratio | 0.5 |
| | Masking type | Patch random |
| Patch Encoder (CNN) | Intermediate channel ($C_{\text{inter}}$) | $d_{\text{model}}/8$ (1s)<br>$d_{\text{model}}/16$ (0.1s) |
| | Input dimension | $\{1, C_{\text{inter}}, C_{\text{inter}}\}$ |
| | Output dimension | $\{C_{\text{inter}}, C_{\text{inter}}, C_{\text{inter}}\}$ |
| | Stride | {64, 3, 3} (1s)<br>{4, 3, 3} (0.1s) |
| | Kernel size | {63, 3, 3} |
| | Padding | {31, 1, 1} |
| | Depth | 3 |
| Patch Encoder (Spectral) | Spectral FFT size | $d_{\text{model}}/2 + 1$ |
| | Spectral dropout | 0.1 |
| STCPE | STCPE dimension | $d_{\text{model}}/8$ |
| | STCPE layers | 1 |
| | STCPE heads | $d_{\text{model}}/256$ |
| | STCPE $d_{ff}$ | $d_{\text{model}}/2$ |
| | Time window size | 7 |
| Positional Embedding | Channel type dimension | $d_{\text{model}}/4$ |
| | Embedding style | CPE (Learnable) |
| | Temperature | 2000 |
| | Scale | 1/256 |
| Transformer | Model dimension | $d_{\text{model}}$ |
| | Layers | 12 |
| | Heads | $d_{\text{model}}/32$ |
| | Feed-forward dimension | $4 * d_{\text{model}}$ |
| | Activation | SiLU |
| | Attention type | Flash attention |
| | Dropout | 0.1 |
| SSL Head | Domain | Time, FFT, STFT |
| | Loss weight ($\lambda_{\text{Time}}$) | 1.0 |
| | Loss weight ($\lambda_{\text{FFT}}$) | 0.1 (1s) / 1.0 (0.1s) |
| | Loss weight ($\lambda_{\text{STFT}}$) | 1.0 (1s) / 0.0 (0.1s) |
| Training | Parameterization | $\mu$P |

### B.7 Pretraining Hyperparameter Search using $\mu$-Parameterization ($\mu$P) and $\mu$Transfer

We employed a two-stage hyperparameter optimization approach to determine optimal learning rate (lr) and weight decay (wd) values; grid search followed by optuna optimization. The search was conducted using a 50M parameter model, a 12-layer architecture with 512-dimensional attention layers. Hyperparameter searches were performed over 2 epochs to balance computational efficiency with reliable performance estimation.

**Stage1: Grid search** We evaluated 30 distinct lr/wd combinataions across four stages of increasing specificity (Table 6), identifying lr=6.0e-03 and wd=1.0e-06 as the best configuration for DIVER-1-1S $_{\text{SMALL}}$. This configuration was also used as the starting point for the Stage 2 search of DIVER-1-0.1S $_{\text{SMALL}}$.

**Stage2: Optuna Optimization** We further refined the hyperparameters using Optuna (Akiba et al., 2019) or bayesian hyperparameter optimization. The search space was defined as $\pm 1$ order of

Table 6: **Stage 1: Grid search hyperparameter exploration.**

| Search Step | Hyperparameter Values |
|---|---|
| Initial Learning Rate Exploration | Learning rate: {1e-5, 1e-4, 1e-3, 1e-2} (wd=1e-2) |
| Weight Decay Exploration | Weight decay: {1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1} (lr=1e-3) |
| Refined Learning Rate Search | Learning rate: {2e-4, 3e-4, 5e-4, 8e-4, 2e-3, 3e-3, 5e-3, 6e-3, 8e-3} |
| Cross-combinations | lr=6e-3 with wd: {1e-6, 1e-5, 1e-4, 1e-3, 1e-1} |

magnitude around the best grid search configurations (range: $\times 0.1$ to $\times 10$), with 50 trials conducted to systematically explore this refined hyperparameter space. The optimal hyperparameter settings identified through Optuna optimization are presented in Table 7.

Table 7: **Optimal learning rate and weight decay**

| Models | Granularity | Learning Rate | Weight Decay |
|---|---|---|---|
| DIVER-1-1s $_{\text{Small}}$ | 1s | 2.30e-03 | 2.17e-07 |
| DIVER-1-0.1s $_{\text{Small}}$ | 0.1s | 4.91e-03 | 3.75e-06 |

### B.8 $\mu$-PARAMETERIZATION ($\mu P$)

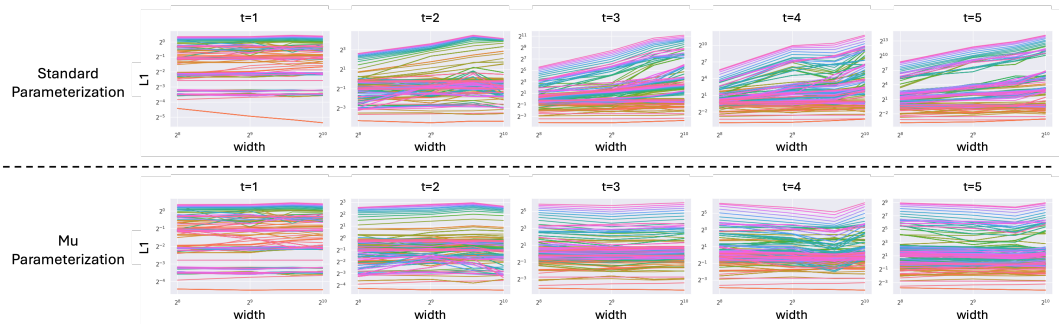


Figure 5: **Verification of the $\mu P$ implementation.** The L1 norm of activation vectors (y-axis) is plotted against model width (x-axis) for five training timesteps ($t$=1 to $t$=5) across four different widths $(256, 512, 768, 1024)$. **(Top Row)** With standard parameterization, activation norms are unstable and diverge as model width increases. **(Bottom Row)** In contrast, our $\mu$P implementation yields stable activation norms that are independent of model width. This confirms the model is correctly parameterized, a critical prerequisite for successful hyperparameter transfer via $\mu$Transfer.

The stable initialization shown in the **Figure 5** translates directly to stable training dynamics. We confirmed this by tracking the training loss while varying the model width across the same four configurations. With $\mu P$ enabled, the training loss remained low and stable for all model sizes. In stark contrast, training without $\mu P$ led to severe instability; the loss diverged rapidly as the model grew, with values exploding to over 100 for the largest width. This empirical result demonstrates that our use of $\mu P$ was essential for reliably training larger models in our scaling experiments.

### B.9 FINETUNING SETUP

Our frozen downstream evaluation corresponds to the Logistic decoder in iMINDBench (Chau et al., 2026). For each dataset–subject/session–task–fold evaluation unit, we freeze the DIVER-1 encoder and flatten the aligned-channel token representation of each 1-second example. We then perform z-normalization using statistics computed from the training set, apply the same transformation to the test set, and fit a separate scikit-learn logistic-regression classifier. We use a maximum of 10,000 iterations, tolerance $10^{-3}$, and random seed 42, keeping this configuration fixed across datasets and tasks. We separately evaluate representations from the penultimate (depth 11) and top (depth 12) encoder layers.

For full fine-tuning, we train the entire model with AdamW using a batch size of 32, evaluating learning rates of $2 \times 10^{-3}$ and $1 \times 10^{-2}$. The benchmark's fixed two-fold within-session splits are retained, and preprocessing statistics are estimated from the training split only. The iMINDBench comparison values are taken from its Main-unit scorecards (Chau et al., 2026).

For the MAYO task, frozen DIVER-1 also same configuration in sckit-learn logistic-regression above. For full fine-tuning, we used a learning rate of 3.86e-03, a weight decay of 4.87e-02, and a batch size of 32. The MAYO seizure dataset is an extended dataset that we constructed to match Brant's minimum input length, and therefore required training on the baseline models. BrainBERT was set at lr 1.08e-04 for classifier and lr 1.08e-05 for encoder, with AdamW optimizer, with batch size 128, which same batch size from original paper (Wang et al., 2023) The features in time [l-5+5] were first meaned, and concatenated along the channel dimension. For the frozen setting, only the classifier was trained using the classifier learning rate specified above. Brant was set at lr 9.13e-04 for the classifier and 9.13e-7 for encoder layers, with betas=(0.9, 0.999), eps=1e-8, batch size 4 and Adam. Except for the learning rates, all other settings were the same as in their publicly released code. For the frozen setting, only the classifier was trained using the classifier learning rate specified above.

For the MAYO dataset, we performed hyperparameter searches for DIVER-1, BrainBERT, and Brant, using 24 trials and selecting the best learning rate and weight decay based on the validation AUROC on subject 1, fold 1.

We could not evaluate PopT on MAYO because LPI coordinates are not available. Brant was pretrained with a 6 s patch and a total of 15 patches (90 s), so we were therefore unable to evaluate it on the neuroprobe (1 s), and there is a mismatch with the original pretrained context in MAYO (6 s).

For all models other than BaRISTA, we trained for 40 epochs using a CosineAnnealing scheduler, applied the same validation split to each training set, and used early stopping when the validation AUROC did not improve for 10 epochs. BaRISTA was trained for 30 epochs using the original codebase, with their default setting as learning-rate schedule consisting of a 5-epoch linear warmup followed by exponential decay.

# C Scaling Law

## C.1 Scaling Law Experiment Details

Pretraining loss curves for the trained models are presented in Figure 6 for the DIVER-1-1s model family and Figure 7 for the DIVER-1-0.1s model family. We trained separate instances for each epoch to obtain the corresponding loss curves. Importantly, we employed cosine annealing with warmup, where the decay schedule depends on the total number of training epochs. Therefore, extracting the loss at epoch 2 from a model trained for 32 epochs differs from the loss at epoch 2 of a model trained for only 2 epochs, as the learning rate trajectories diverge under these configurations.

**Unique number of tokens** First of all, the estimated total number of unique tokens is $U_D =$ number of data sample $\times$ number of tokens per sample. The number of tokens per sample can be expressed as number of channels $\times$ number of timestamps $\times$ proportion of unmasked patches.

The total number of data samples is 636,480. Since we randomly sampled from a 32 channel x 30 timestamp token grid using Beta(3,1) distribution for both axis, we estimate the number of tokens per sample as $32 \times 30 \times 0.75 \times 0.75 = 540$ tokens. Thus, we get $U_D = 636,480 \times 540$ for the data.

**Compute** To investigate compute scaling properties, we conducted systematic pretraining experiments across different epoch counts for the 1-second granularity models. For shorter training regimes (1, 2, 4, and 8 epochs), we pretrained all six model variants: DIVER-1-1S $_{\text{Tiny}}$, DIVER-1-1S $_{\text{Small}}$, DIVER-1-1S $_{\text{Base}}$, DIVER-1-1S $_{\text{Large}}$, DIVER-1-1S $_{\text{XL}}$, and DIVER-1-1S $_{\text{XXL}}$. For longer training regimes (16, 32, and 64 epochs), computational constraints limited our experiments to the four smaller model variants: DIVER-1-1S $_{\text{Tiny}}$, DIVER-1-1S $_{\text{Small}}$, DIVER-1-1S $_{\text{Base}}$, and DIVER-1-1S $_{\text{Large}}$.

For the DIVER-1-0.1s model family with 0.1-second granularity models, we followed a similar training protocol. We trained six model variant for epochs 2, 8 and five model variants for epochs 32, excluding DIVER-1-0.1S $_{\text{Base}}$ and three model variants for epochs 64, excluding DIVER-1-0.1S $_{\text{Base}}$, DIVER-1-0.1S $_{\text{Large}}$ and DIVER-1-0.1S $_{\text{XXL}}$.

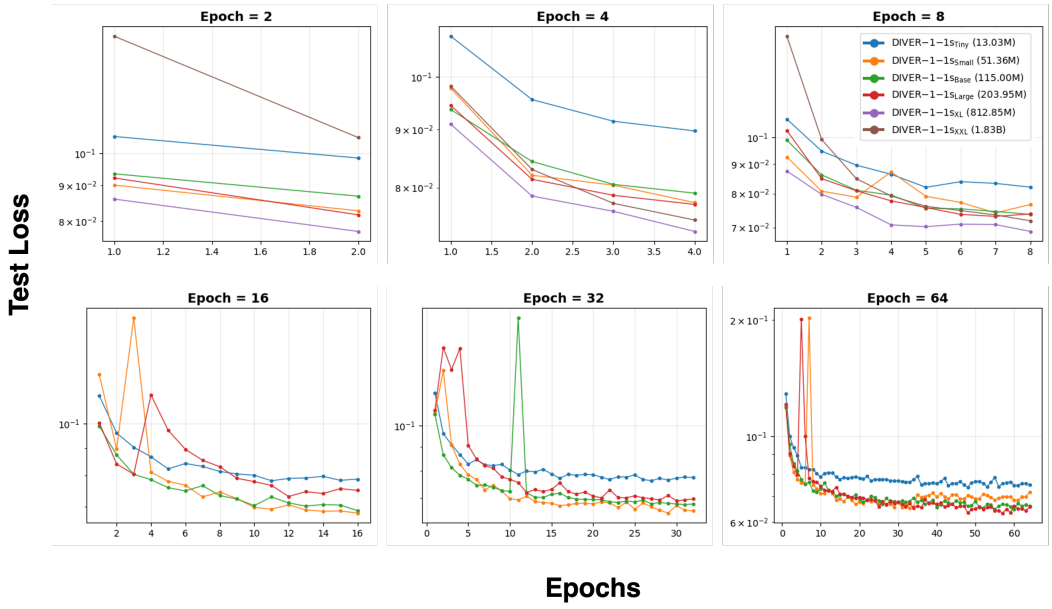


Figure 6: Loss curves of the DIVER-1-1s model family. Test loss across epochs is shown.

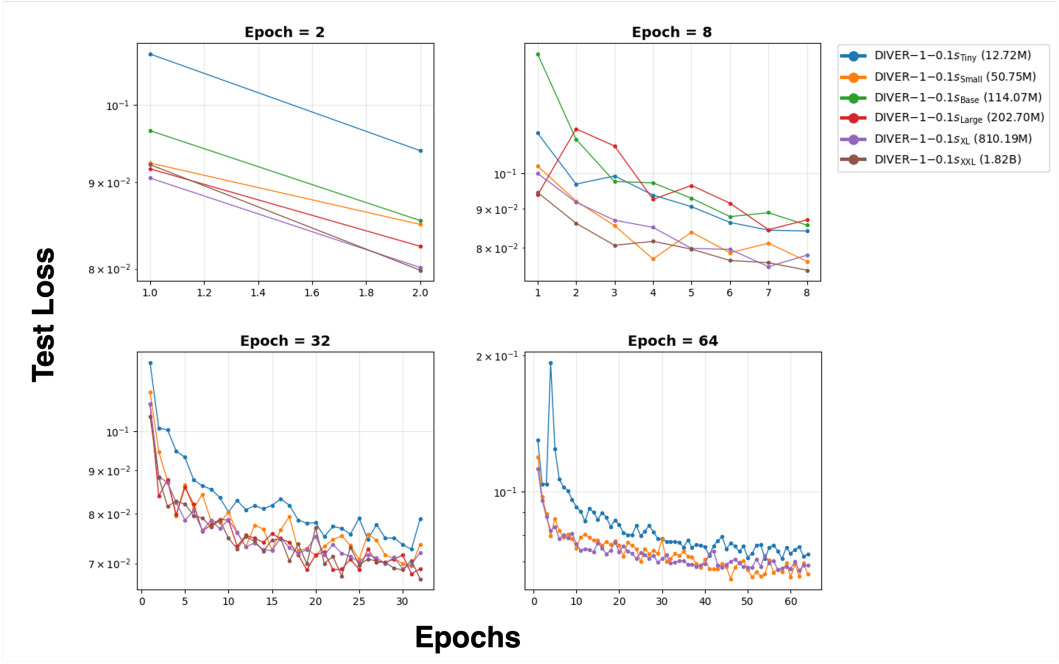


Figure 7: Loss curves of the DIVER-1-0.1s model family. Test loss across epochs is shown.

**Data Size** Data size scaling was done in 1, 3, 9, 24, 50, 90, 100% of data on DIVER-1-1s $_{\text{Small}}$ for 2 epochs with hyperparameters fixed as $lr = 6.0e-03, wd = 1.0e-06$.Number of token = number of data sample $\times$ number of tokens per sample, while the total number of data samples was 636,480. Number of tokens per sample estimated as 540, given that we randomly sampled from a 32×30 token grid using Beta(3,1) distribution. A detailed explanation can be found in the aforementioned **Unique number of tokens** section.

**Model Size** We fixed the number of epochs to 2. The models varied by their width, number of layers, and patch size. The detailed experiment conditions are on table 2. We observed that the models with different number of layers or patch size show different scaling behavior, so we fitted them separately.

**Number of Subjects** Subject scaling experiments were done with datasets containing 2, 4, 5, 8, 10, 15, 16 subjects respectively, while maintaining a constant dataset size.DIVER-1-0.1s $_{\text{Small}}$ trained for 2 epochs, due to compute constraint.

**Data-constrained Scaling Law** We trained a total of 31 DIVER-1-1s models with varying parameter counts and numbers of training epochs, while keeping the dataset fixed. The hidden dimension was fixed to 12 layers. Since different granularity led to different scaling behavior, we experimented on two granularity conditions and fitted them separately.

### C.2 Extended Kaplan (Kaplan et al., 2020a) Scaling Law Results

We tested models at the 1-second and 0.1-second granularity. For 1-second granularity models, all six models listed in Table 2 as DIVER-1-1s were tested for epochs 2, 4, 8. At epochs 16, 32, and 64, evaluation was conducted on the following four models: DIVER-1-1s $_{\text{Tiny}}$, DIVER-1-1s $_{\text{Small}}$, DIVER-1-1s $_{\text{Base}}$, and DIVER-1-1s $_{\text{Large}}$. At earlier epochs (2, 4, and 8), the general trend showed decreasing loss as model size increased. However, as observed in DIVER-1-1s $_{\text{XXL}}$, larger models exhibited substantially higher loss when trained with only a few epochs. This confirms what was also suggested by the data-constrained scaling isoplots: training very large models with insufficient updates is ineffective. Another possibility is that the aspect ratio of DIVER-1-1s $_{\text{XXL}}$ (256) places it outside the region where loss remains stable. Future work should therefore evaluate larger models within the aspect-ratio regime where stable loss behavior is maintained.

### C.3 Data-Constrained Scaling Law Results

Table 8: **Fitted data-constrained scaling law parameters for DIVER-1-1s and DIVER-1-0.1s model families.**

| | DIVER-1-1s | DIVER-1-0.1s |
|---|---|---|
| $A$ | 19.217 | 101.52 |
| $B$ | 57.065 | 1.1550 |
| $E$ | 0.0092 | 0.0030 |
| $\alpha$ | 0.3773 | 0.5248 |
| $\beta$ | 0.3504 | 0.1246 |
| $R_D^*$ | 9.5372 | 19.705 |
| $R_N^*$ | 3.3850 | 0.7191 |
| $R^2$ (linear) | 0.7858 | 0.7575 |
| $R^2$ (log) | 0.8152 | 0.7718 |

Table 8 reports the fitted data-constrained scaling-law parameters for both DIVER-1-1s and DIVER-1-0.1s. The exponents $\alpha$ and $\beta$ govern the marginal benefit of scaling parameters and data, respectively. For DIVER-1-1s, our values ($\alpha = 0.377$, $\beta = 0.350$) are similar in magnitude and comparable to language-domain results ($\alpha = 0.348$, $\beta = 0.366$; (Besiroglu et al., 2024)). The Kaplan-stye allocation in Figure 4(a) is computed at fixed two-epochs training where data is effectively fresh ($R_D \approx 1$); under this fresh-data regime the data component naturally exerts greater influence on the loss, whereas the data-constrained fit aggregates across multi-epoch runs and yields balanced exponents. DIVER-1-0.1s shows a more parameter-leaning exponent profile in isolation ($\alpha = 0.525$, $\beta = 0.125$), but its dataset-axis fit is less stable under our sampling density ($R^2 \approx 0.77$); the small $\beta$

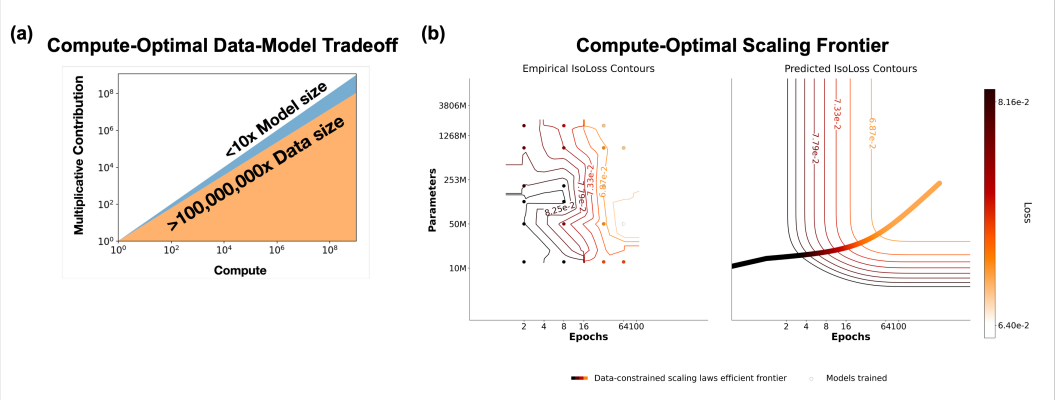


Figure 8: **Compute-optimal scaling of DIVER-1-0.1s.** (a) Optimal data–model allocation when training duration is fixed. With training fixed to two passes, we use the fitted DIVER-1-0.1s scaling law to estimate how the compute-optimal dataset size and parameter count should grow as total compute increases. (b) IsoLoss contours for the DIVER-1-0.1s model family. (Left) Twenty models with 0.1s patches were trained across varying epochs and parameter counts; iso-loss contours are obtained by linear interpolation between measured data points. (Right) Corresponding contours predicted by the fitted scaling law. The fading line denotes the minimum-loss configuration for each compute budget.

should therefore be read as reduced fit reliability rather than evidence that data scaling is unhelpful at 0.1s.

Importantly, the characteristic half-lives $R_D^* = 9.5$ and $R_N^* = 3.3$ quantify diminishing returns under repeated data and excessive parameters. The relatively larger $R_D^*$ implies that repeated data remains useful for many epochs before saturation, whereas the smaller $R_N^*$ suggests that the benefit of adding parameters decays more quickly. Together, these results suggest that gains are most effectively pursued by scaling dataset repetition while keeping model sizes small.

Figure 4(a) compares fixed-duration compute-optimal data–model allocation for DIVER-1-0.1s models against the Kaplan language-model fit (Kaplan et al., 2020b). Although iEEG and text tokens are not directly comparable, the relative allocation differs strikingly across domains. Both DIVER-1-1s(Figure 3(a) and DIVER-1-0.1s(Figure 8(a)) exhibit a strongly data-leaning regime, in which optimal scaling devotes the overwhelming majority of additional compute to dataset size ($> 5 \times 10^7\times$ for 1s model and $> 10^8\times$ for 0.1s model) while model size grows by less than two orders of magnitude ($< 20\times$ and $< 10\times$, respectively). In contrast, the Kaplan LLM fit is strongly parameter-dominated, allocating roughly $3.7 \times 10^6\times$ to model size against only $\sim$270$\times$ to data. The two iEEG patch granularities therefore behave qualitatively similarly to each other and qualitatively differently from the language-model regime, suggesting that compute-optimal scaling for iEEG signals favors data scaling over parameter scaling—consistent with the limited per-sample information density of neural recordings relative to text.

The empirical isoLoss contours in Figure 4 show less smoothness compared to the original scaling law paper(Muennighoff et al., 2023), primarily due to sampling density. While the original study used 93 model configurations with dense sampling across all loss ranges, we evaluated 30 configurations. Despite this visual difference, our empirical isoLoss contours (Figure 4(b) and Figure 8(b), left panel) align well with the predicted contours (Figure 4(b) and Figure 8(b), right panel), demonstrating that data-constrained scaling laws generalize to neural data. Although our scaling fits are estimated on DIVER-1, the setup is representative of current EEG/iEEG FMs: recent reviews report that 82% use Transformer backbones (Liu et al., 2026) and 70% use generation-based SSL(Shen et al., 2026). Other objectives, tokenizations, and corpora may change the exact exponents, but the intrinsic constraints of iEEG suggest that the qualitative recipe should persist: scale data first, train longer second, and scale parameters last.

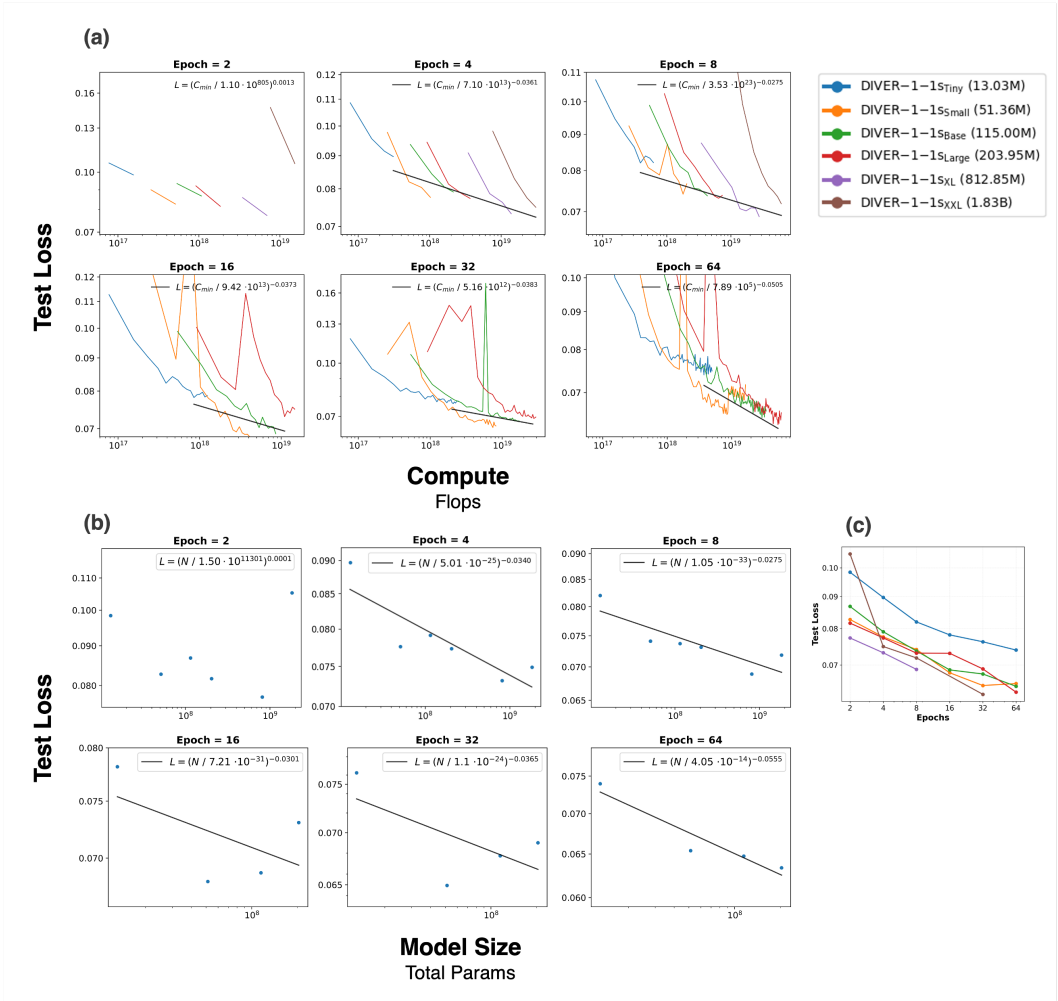


Figure 9: Scaling law extended results for the DIVER-1-1s family. Compute scaling and model size scaling plots are given for models trained for 2, 4, 8, 16, 32, and 64 epochs.(a) Compute scaling and (b) model size scaling plots are given for models trained for 2, 8, 32, and 64 epochs. (c) Epoch scaling plot of the models reported in Fig. 3 (d). The same training runs are reused, with losses re-plotted against parameters and dataset size in log-log scale.

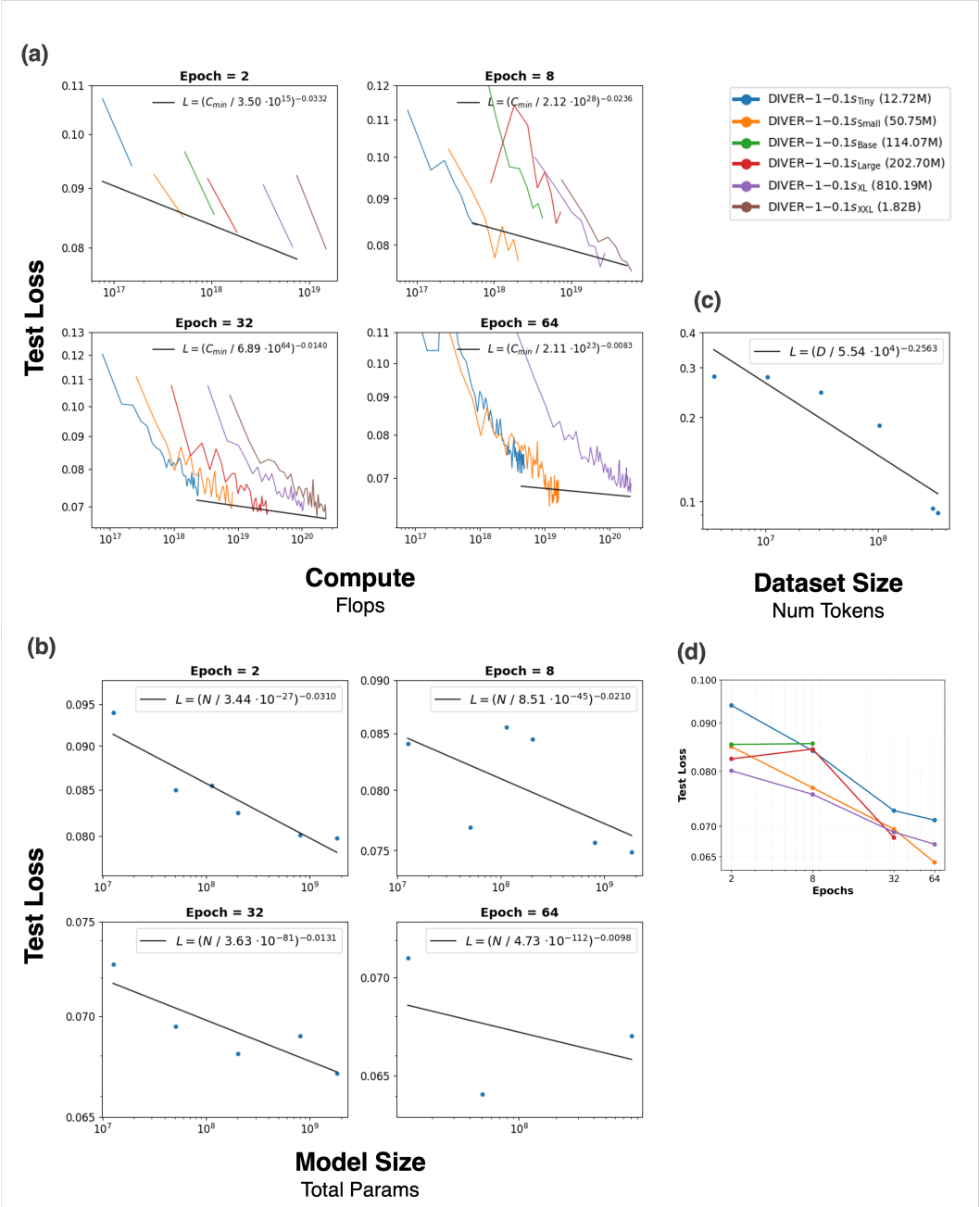


Figure 10: Scaling law extended results for the DIVER-1-0.1s family. (a) Compute scaling and (b) model size scaling plots are given for models trained for 2, 8, 32, and 64 epochs. (c) Dataset size scaling plots are given for models trained for 2 epochs. (d) Epoch scaling plot of the models reported in Fig. 8. The same training runs are reused, with losses re-plotted against parameters and dataset size in log-log scale.

Table 9: **Comparison with iMINDBench baselines across datasets.** Published Main-unit within-session ROC-AUC point estimates for NP, BYD, and Pippi (Chau et al., 2026). DIVER-1 results are obtained from penultimate-layer representations. Overall is the unweighted mean across the three datasets, and the best reported value in each column is bold.

| Models | NP | BYD | Pippi | Overall |
|---|---|---|---|---|
| Logistic regression (Multi-STFT) | 0.697 | 0.624 | 0.564 | 0.628 |
| MLP (Multi-STFT) | 0.702 | 0.625 | 0.567 | 0.631 |
| CNN (Multi-STFT) | 0.710 | 0.611 | 0.539 | 0.620 |
| PopT-v2 | 0.723 | **0.632** | 0.579 | 0.645 |
| BaRISTA | 0.716 | **0.632** | 0.558 | 0.636 |
| BrainBERT (frozen + linear head) | 0.664 | 0.619 | 0.575 | 0.619 |
| DIVER-1-0.1s (full-finetune) | 0.737 | 0.630 | 0.576 | **0.648** |
| DIVER-1-0.1s (frozen) | 0.731 | 0.630 | **0.582** | **0.648** |
| DIVER-1-1s (full-finetune) | **0.738** | 0.625 | 0.575 | 0.646 |
| DIVER-1-1s (frozen) | 0.704 | 0.616 | 0.579 | 0.633 |

# D Extended Results

## D.1 Comprehensive DIVER Downstream Task Results

This section reports the complete numerical results underlying Figure 2. Table 9 presents the iMIND-Bench Main-unit within-session ROC-AUC comparison on NP, BYD, and Pippi: baseline values are the point estimates published in the benchmark scorecard (Chau et al., 2026), and the DIVER-1 rows cover both patch scales (0.1 s and 1 s) under frozen (the benchmark's Logistic protocol) and full-finetuning evaluation, primarily with penultimate-layer readouts. Overall is the unweighted mean of the three dataset-level means. For MAYO, we evaluated DIVER-1-1s (with $d_{\text{model}} = 256$, pretrained for 32 epochs) in both frozen (linear probing) and full-finetuning settings, and DIVER-1-0.1s in both settings as well. We could not evaluate PopT because the dataset does not contain any coordinates. By contrast, although our model is trained with 3D positional embeddings, it can handle missing position information by replacing the positional embedding with a zero vector for electrodes with unknown location. Brant exhibited a substantial performance drop under full finetuning, likely because we deviated from its original pretraining configuration by using only a single patch, which can impair optimization when updating all parameters with original context windows (90s), whereas the frozen setting simply uses the fixed embedding.

Table 10 reports the consolidated MAYO comparison, including the three Neuroprobe-provided non-foundation-model decoders (CNN/MLP/linear on STFT spectrograms), our reimplementations of the three top-ranked classical algorithms from the original challenge leaderboard (Baldassano et al., 2017), and hyperparameter-optimized BrainBERT and Brant under both linear probing and full fine-tuning (Section B.9). DIVER-1 (0.1 s, full fine-tuning) has the best mean AUROC; however, it did not significantly outperform any classical or STFT-based baseline, so we describe the MAYO result as parity with the strongest classical methods. In contrast, all four prior foundation-model configurations—BrainBERT and Brant under both linear probing and full fine-tuning—yielded significantly lower AUROC than DIVER-1 (two-sided paired Wilcoxon signed-rank tests, BH-FDR-corrected $q = .029$ for each comparison).

Table 10: **MAYO seizure-detection comparison under the matched evaluation protocol.** Mean AUROC ± SEM across the 8 participants (participant-level SEM; values are means over 3 seeds with resampled validation splits on a fixed test set). Rows are grouped by model family and sorted by AUROC within each group; the best value overall is bold. "Leaderboard reimpl." rows are our faithful Python reimplementations of the top-3 classical algorithms of the original challenge (Baldassano et al., 2017); see Section E.2 for why leaderboard-reported AUCs are not comparable to ours.

| | Model | AUROC |
|---|---|---|
| DIVER-1 (ours) | **0.1 s, full fine-tuning, top layer** | $\mathbf{0.974 \pm 0.010}$ |
| | 0.1 s, frozen, penultimate layer | $0.964 \pm 0.013$ |
| | 0.1 s, frozen, top layer | $0.963 \pm 0.013$ |
| | 1 s, full fine-tuning, top layer | $0.959 \pm 0.020$ |
| | 1 s, frozen, penultimate layer | $0.936 \pm 0.023$ |
| | 1 s, frozen, top layer | $0.936 \pm 0.021$ |
| Leaderboard reimpl. | 2nd place (Olson & Mingle) | $0.964 \pm 0.024$ |
| | 3rd place (Talukdar, Moore & Sood) | $0.931 \pm 0.031$ |
| | 1st place (Hills) | $0.924 \pm 0.035$ |
| STFT decoders | CNN | $0.956 \pm 0.028$ |
| | Linear | $0.937 \pm 0.034$ |
| | MLP | $0.913 \pm 0.038$ |
| Prior iEEG FMs | BrainBERT, full fine-tuning (optimized) | $0.781 \pm 0.056$ |
| | Brant, full fine-tuning (optimized) | $0.729 \pm 0.082$ |
| | BrainBERT, linear probing (optimized) | $0.721 \pm 0.050$ |
| | Brant, linear probing (optimized) | $0.630 \pm 0.042$ |

### D.2 Interpretation Results

We performed the interpretation analyses on Neuroprobe, one of the three datasets included in the iMINDBench benchmark. Figure 11 shows a visualization of representation analysis on downstream datasets using UMAP (McInnes et al., 2018). We examine the embeddings obtained from the pre-trained DIVER-1-0.1s model on the test set of speech, onset, and volume tasks in neuroprobe dataset without any finetuning. The results indicate that DIVER learned meaningful iEEG representations from pretraining, thereby capturing label-relevant structure in downstream datasets even in the absence of finetuning.

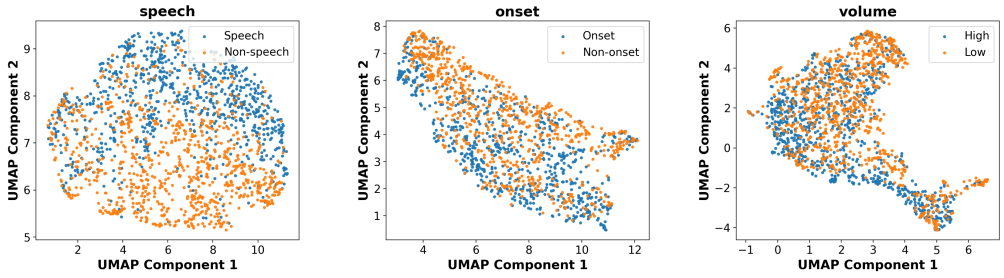


Figure 11: Visualizations of representations on neuroprobe downstream tasks. Each plot shows test set embeddings from one fold of a single trial from a single subject.

To obtain an interpretable salience measure of the model, we applied attention rollout (Abnar & Zuidema, 2020), which aggregates attention weights across all attention layers. Although DIVER's any-variate attention produces a high-dimensional matrix, the attention rollout yielded a rank-1 matrix across all subjects and sessions. We therefore used a single representative row, reflecting a uniform attention distribution across keys independent of query. Figure 12 shows task-dependent temporal and cortical attribution patterns derived from the attention rollout of DIVER-1-0.1s. Attention rollout weights were computed on test sets for four high AUROC scored Neuroprobe tasks (volume, pitch, GPT-2 surprisal, and word head position). We follow PopT's method of mapping attention weights onto the cortical surface, normalizing within each subject and scaling by the test AUROC (Chau et al., 2025). The resulting weights were averaged across subjects, trials, and folds.

Across both auditory (volume, pitch) and language (GPT-2 surprisal, word head position) tasks, attention rollout exhibits two temporal peaks: an early window (0–300 ms) and a late window (800–1000 ms). The early peak coincides with word onset of each sample and engages core speech–language regions, including inferior temporal gyrus and Broca's and Wernicke's areas (Mesgarani et al., 2014). The late peak shows task-specific spatial patterns. For auditory tasks, attribution remains concentrated in primary auditory cortex (Heschl's gyrus and superior temporal gyrus). This is consistent with the possible continued auditory processing of subsequent words within the 1s window. In contrast, language tasks show increased engagement of temporal-frontal regions in the later window. This aligns with the known post-auditory latency of higher language networks (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). Overall, these results indicate that DIVER attends to tokens in a temporally structured manner with task-relevant spatial attribution. Notably, the right middle occipital gyrus remains consistently salient for GPT-2 surprisal, suggesting a contribution of visual-context representations to surprisal estimation (Friederici, 2011; Dehaene & Cohen, 2011). Collectively, DIVER's task-dependent attributions are consistent with established neural models of auditory and language processing (Hickok & Poeppel, 2007).

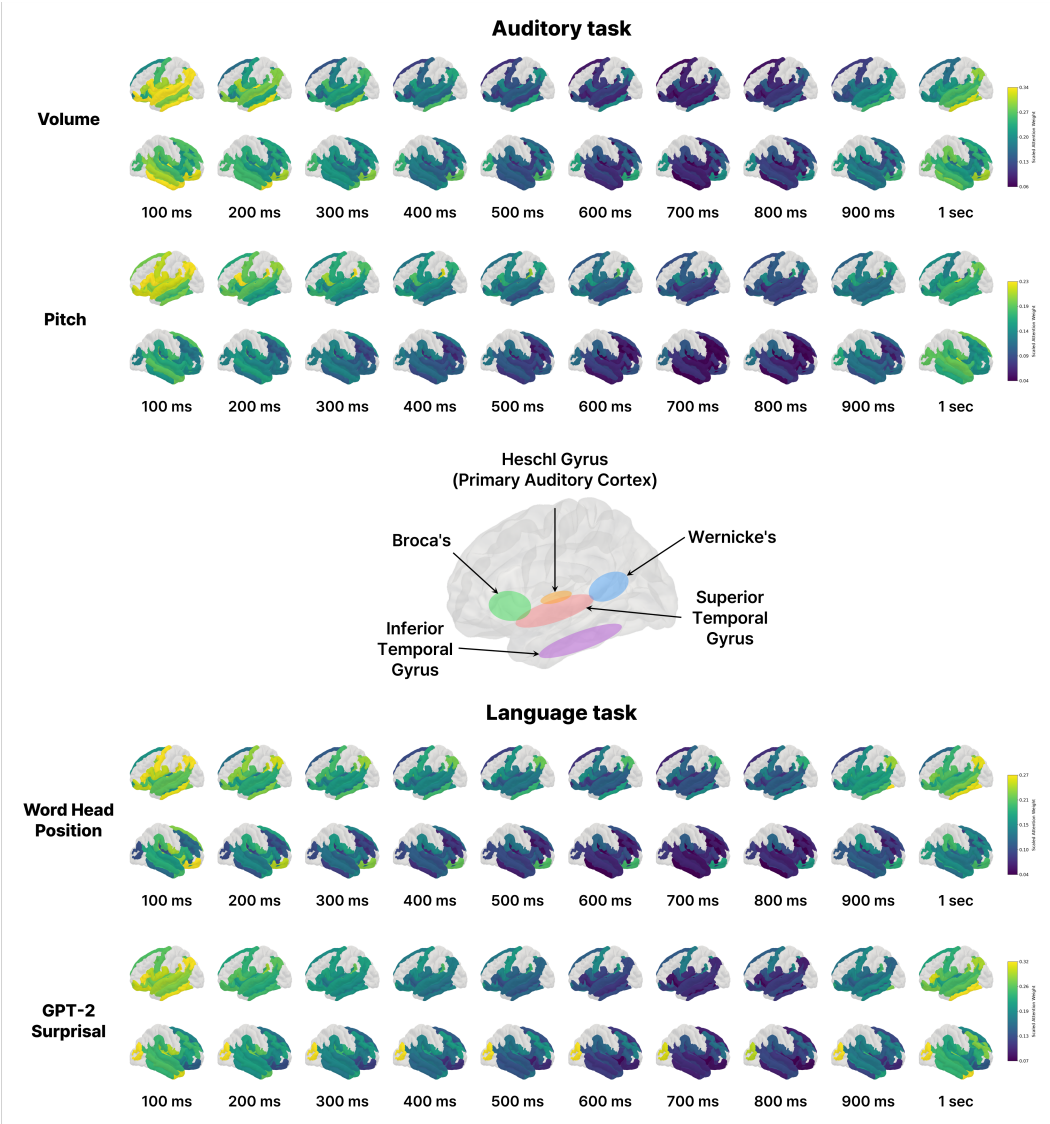


Figure 12: **Scaled attention rollout weights mapped onto the cortical surface for four Neuroprobe downstream tasks.** Each panel shows DIVER-1-0.1s scaled attention weights extracted via attention rollout and aggregated in 100 ms bins from 0 ms to 1 s relative to stimulus onset. For each task, the top row shows the left hemisphere and the bottom row shows the right hemisphere; brighter colors indicate larger scaled attention weights, which we interpret as greater model salience on the corresponding channel-temporal token for each task.

# E DATA DETAILS

## E.1 PRETRAINING DATASET DESCRIPTION

The following datasets were utilized for the pretraining of our DIVER models. The total pretraining time is 5,310 hours.

- **AJILE12 (Annotated Joints in Long-term Electrocorticography)** (Peterson et al., 2022): An ECoG dataset from 12 epilepsy patients, recorded semi-continuously over 55 days. Signals were collected from $\geq$ 64 electrodes at 1 kHz sampling rate and paired with synchronized video-based 3D human pose estimation and annotated wrist-movement events.

- **Self-collected iEEG dataset**: An intracranial EEG dataset from 25 drug-resistant epilepsy patients ($\sim$7 days, $\sim$168 h per subject) with long-term ECoG and sEEG recordings (mean $56.4 \pm 3.38$ channels, sampled at 2 kHz) during naturalistic hospital behaviors.

## E.2 FINETUNING DATASET DESCRIPTION

The following datasets were utilized for the downstream evaluation of our DIVER models. An overview of the dataset specifications and task definitions is provided in Table 4.

**iMINDBench.** iMINDBench is a multi-institution iEEG neural-decoding benchmark that harmonizes three independently collected naturalistic movie-watching datasets: Neuroprobe/Brain Treebank (NP), Bang! You're Dead (BYD), and Pippi (Chau et al., 2026). NP comprises held-out English feature-film recordings, BYD uses an 8-minute black-and-white Hitchcock excerpt with English audiovisual content, and Pippi uses a 6.5-minute Dutch-dubbed short film with interleaved speech and music. Their original sampling rates are 2048, 1000, and 2048 Hz, respectively. The benchmark release contains 5, 29, and 5 subject/session units, with averages of 118, 77, and 110 selected electrodes per subject for NP, BYD, and Pippi, respectively; these are benchmark subject/session counts rather than necessarily distinct-participant counts.

Each example pairs a 1-second neural window with one of 15 binary decoding targets spanning language (sentence onset, speech presence, word index, word gap, GPT-2 surprisal, dependency-head position, word length, and noun-versus-verb part of speech), audition (change in volume, volume, and pitch), and vision (global optical flow, local optical flow, frame brightness, and face presence). Continuous annotations are binarized within session using the bottom and top quartiles for NP and the bottom and top terciles for BYD and Pippi; categorical annotations are mapped to binary labels. Table 4 summarizes the corpus specifications, while Section B.9 describes the fixed splits, leakage-controlled preprocessing, Logistic evaluator, and ROC-AUC aggregation.

**MAYO.** Our seizure-detection evaluation uses a custom construction of the MAYO Kaggle challenge data (Bbrinkm & Cukierski, 2014; Baldassano et al., 2017) rather than the original challenge protocol, for two reasons. (i) Compatibility with Brant, whose 6 s patch cannot accommodate the challenge's native 1 s clips: we concatenate consecutive clips into 6 s segments and redefine the splits; because this procedure substantially reduces the size of the training set, we reverse the original training and test split. (ii) The challenge scores 12 subjects, of which 4 are canines recorded at 400 Hz on a different implanted device; we evaluate the 8 human participants only. AUROCs on our construction are therefore *not comparable* to the original challenge leaderboard: the winning entry's private-leaderboard AUC (0.96288, Hills (Baldassano et al., 2017)) and our numbers (e.g., DIVER-1 at 0.974 and our Hills reimplementation at 0.924) are computed on different evaluation sets. To provide classical references under matched conditions, we faithfully reimplemented the three top-ranked leaderboard algorithms as described in Baldassano et al. (2017) (the originals are Python 2): Hills (1st) — channel cross-correlation and its sorted eigenvalues, 1–47 Hz FFT magnitudes, 3000-tree random forest; Olson & Mingle (2nd) — automated selection from ten log-spaced bandpass filters (5–200 Hz), normalized per-clip covariance, ensemble of 100 two-hidden-layer MLPs; Talukdar, Moore & Sood (3rd) — 100 Hz downsampling, channel-wise and global amplitude/variance/FFT features, 1000 Extremely Randomized Trees.

### E.3 QAQC AND PREPROCESSING

All data underwent quality assessment and control (QAQC) and preprocessing with a philosophy of minimal intervention to retain as much original signal information as possible. For QAQC, we clipped amplitude values exceeding these normalization thresholds, only discarding electrodes when more than 3.33% of samples required clipping and removing whole segments when more than 50% of channels were compromised. This conservative strategy enabled us to preserve substantially more usable data. Then we normalized signals by dividing by 200 μV. While Jiang et al. (2024) (EEG model) applied normalization without QAQC and Wang et al. (2024b) (EEG model) removed entire segments if even one timepoint exceeded 100 μV, we adopted a more conservative clipping approach to prevent data loss.

For preprocessing, we applied minimal filtering: a high-pass filter (0.5 Hz) to remove low-frequency drift, 60 - 120 - 180 notch filters for power line noise suppression, and no low-pass filtering to preserve high-frequency components. All datasets were resampled to 500 Hz and segmented into 30-second non-overlapping windows.

## F COMPARISON WITH EXISTING IEEG FOUNDATION MODELS

We compare DIVER-1 to all prior iEEG foundation models on the axes most relevant for scaling: parameter count, subject diversity, channels, dataset volume (channel-hours), and effective compute, the last reported as Scaled Epochs $= \Phi/352{,}035$ where $\Phi$ is total channel-hours processed during pretraining and $352{,}035$ is the DIVER-1 corpus size.

Table 11: **Comparison with prior iEEG foundation models.**

| Model | Modality | Params | #Subj. (sites) | Cohort | Volume (ch-hr) | Train. Epochs | Sc. Eps. on Ours[a] |
|---|---|---|---|---|---|---|---|
| BrainBERT (Wang et al., 2023) | sEEG | 43.18M | 10 (1) | Brain Treebank[b] | 4,551 | 39* | 0.51 |
| Brant (Zhang et al., 2023) | sEEG | 505.69M | 9 (1) | private clinical | 281,860* | 426* | 340.88 |
| Du-IN (Zheng et al., 2024)[c] | sEEG | 4.38M | 12 (1) | private (Beijing)[c] | 150[c] | 400 | 0.17 |
| BaRISTA (Oganesian et al., 2025) | sEEG | 1M | 10 (1) | Brain Treebank[b] | 3,640* | 70 | 0.72 |
| PopT (Chau et al., 2025)[d] | sEEG | 20M[d] | 10 (1) | Brain Treebank[b] | 6,570* | n.r. | n.r. |
| **Ours** (DIVER-1) | ECoG+sEEG | 12.72M–1.83B | **37 (2)** | AJILE12 + private (multi-site) | **352,035** | 64 | 64 |

[a] Sc. Eps.(Scaled Epochs) $=$ (Volume $\times$ Train. Epochs)$/352{,}035$, normalising total channel-hours processed during pretraining by the DIVER-1 corpus size. This is a unit-consistent compute metric across models with different per-sample structures.
[b] BrainBERT, BaRISTA, and PopT all pretrain on the same 10-subject sEEG corpus (Boston Children's Hospital; movie-watching). BrainBERT used the pre-release version; the corpus was subsequently released as "Brain Treebank". Volume differences reflect channel/session selection only.
[c] Du-IN is a per-subject model; we report the SOTA configuration (62.70% acc.) pretrained on full 15 h recordings (12 h non-task + 3 h task). Multi-subject ("poms", 59.18%) and non-task-only (12 h, 60.02%) variants underperformed and are omitted.
[d] PopT is trained on top of a frozen BrainBERT; the 20 M count excludes the 43 M BrainBERT base.
n.r. Not reported. Specific information is missing from the original publication and cannot be accurately derived.
* Estimated; derivations in Table 12.

Table 12: **Estimation of model and training specifications.** We detail the assumptions and calculations used to derive data volume (channel-hours) and training epochs for prior models. Bold values represent the final estimates derived from the reported configurations.

| Model | Quantity | Est. Value | Justification / Derivation Method |
|---|---|---|---|
| **BrainBERT** | Train. Epochs | **39.06** | $U$=500k, $B$=256, $A$=1 (Wang et al., 2023). Each sample is a 5 s single-channel spectrogram ($h_s = 5/3600$ ch-hr).<br>$\Phi = U \cdot B \cdot A \cdot h_s = 1.78 \times 10^5$ ch-hr.<br>**Epochs** $= \Phi/4{,}551 = \mathbf{39.06}$. |
| **Brant** | Volume | **281,860** | Individual ch-hr summed across 9 subjects ($\sum c_i h_i$) using metadata from App. E Table 7 (totaling 894 channels and 2,528 recording hours). This precise count is highly consistent with the reported "$> 10^{12}$ timestamps" ($\geq 277,778$ ch-hr) at 1,000 Hz. |
| | Train. Epochs | **425.74** | $U$=750k, $B$=16, $A$=4 $\Rightarrow 48 \times 10^6$ sample-passes. Based on 1,500 patches/sample ($L$=15, $C \approx$100) and $\tau$=6 s, $h_s = 2.5$ ch-hr.<br>$\Phi = 48 \times 10^6 \cdot 2.5 = 1.20 \times 10^8$ ch-hr.<br>**Epochs** $= \Phi/281{,}860 = \mathbf{425.74}$. |
| **BaRISTA** | Volume | **3,640** | 35,089 segments of 3 s totaling 29.2 h (Oganesian et al., 2025). Mean channel count $\bar{C} = 1{,}247/10 = 124.7$.<br>**Volume** $= 29.2 \cdot 124.7 \approx \mathbf{3,640}$ ch-hr. |
| **PopT** | Volume | **6,570** | **Volume** = 19 sessions $\times$ 2.07 h/session $\times$ 167 mean ch/subj (Chau et al., 2025) $\approx \mathbf{6,570}$ ch-hr. |

*Notation.* $U$: optimizer updates, $B$: minibatch size, $A$: gradient accumulation steps, $h_s$: per-sample channel-hours, $\Phi$: total ch-hr processed ($\Phi = U \cdot B \cdot A \cdot h_s$). **Epochs:** Ratio of total throughput ($\Phi$) to dataset volume. **Scaled Epochs on DIVER$_1$:** $\Phi/352{,}035$, providing a unit-consistent comparison across heterogeneous sampling pipelines (e.g., single-channel spectrograms vs. multi-channel windows).