How Pretraining Objectives Shape EEG Representations: A Controlled Comparison

Anonymous authors
Paper under double-blind review

Abstract

EEG foundation models learn transferable representations from large unlabelled recordings, but differences in datasets, architectures, and evaluation protocols obscure how pretraining objectives shape what they learn. We address this gap through a controlled comparison of masked autoencoding (MAE), autoregressive prediction (AR), and subject-level contrastive learning (SLCL), using a shared frozen tokenizer and GPT-2 backbone with pretraining data drawn from 50,894 hours of four-channel consumer EEG. Across seven downstream tasks, MAE shows the clearest performance gains on sleep staging, age prediction, gender classification, and EEG event classification, while SLCL best learns subject-specific features. Increasing pretraining data generally improves downstream performance, while LoRA fine-tuning largely preserves the relative performance of the three objectives observed under frozen probing. To interpret these objective-specific strengths, we probe how established EEG features are encoded in the learned representations. MAE retains strong band-power information from the tokenizer and correlates most with cross-channel coherence features, as it reconstructs masked tokens from visible context across channels and time. AR most clearly exposes band power trends, since it predicts subsequent activity from preceding context. SLCL favors alpha-peak frequency and alpha asymmetry, suggesting that aligning recordings from the same participant across sessions emphasizes participant-related characteristics. Layer-wise probing further reveals that task-relevant EEG information is distributed differently across model depth depending on the pretraining objective, with all seven tasks reaching their highest probe scores before the final layer. Together, these findings connect pretraining formulations to distinct spectral, spatial, and temporal feature profiles, providing an interpretable basis for understanding their downstream strengths and trade-offs.

1 Introduction

Electroencephalography (EEG) non-invasively records brain activity with high temporal resolution and relatively low cost, supporting applications in brain-computer interfaces (BCI), clinical assessment, and mental state decoding. Yet, EEG signals are inherently difficult to decode due to low signal-to-noise ratio, non-stationarity, and pronounced inter-subject variability. Foundation models address these challenges by learning representations from large unlabelled EEG corpora that transfer through fine-tuning or linear probing and can outperform supervised learning from scratch when labels are limited (Banville et al., 2021; Jiang et al., 2024; El Ouahidi et al., 2025; Döner et al., 2025; Jiang et al., 2025; Niu et al., 2025).

Self-supervised pretraining learns from unlabelled data through objectives that shape the resulting representations. In natural language processing, autoregressive (AR) next-token prediction has been the dominant paradigm (Radford et al., 2018; Brown et al., 2020), while masked prediction remains central to representation learning (Devlin et al., 2019). In computer vision, the trajectory has been more varied: contrastive learning (CL) closed much of the gap with supervised features (Chen et al., 2020), masked autoencoders (MAE) showed that reconstructing heavily masked images yields strong representations (He et al., 2022), and autoregressive image modelling has re-emerged as a competitive alternative (El-Nouby et al., 2024; Fini et al., 2025). Together, prior work suggests distinct representational biases: masked reconstruction favours contextual integration, AR emphasizes predictive dependencies (Zhang et al., 2024), and CL promotes features shared across paired views or

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.

modalities (Radford et al., 2021). This motivates a controlled examination of how these objectives shape EEG representations and transfer across tasks.

These pretraining paradigms are increasingly being adapted to physiological time-series data. In photoplethysmography (PPG) and electrocardiography (ECG), Abbaspourazad et al. (2024) showed that subject-level contrastive learning captures participant-specific physiological information, while Narayanswamy et al. (2025) demonstrated predictable scaling of masked-input modelling across wearable sensor data. In electromyography (EMG), aligning EMG with complementary pose information through contrastive learning improved gesture generalization (Cui et al., 2025).

In EEG, foundation-model development has centred primarily on masked prediction. LaBraM (Jiang et al., 2024) tokenizes EEG patches into discrete neural codes and predicts masked codes from visible context, providing an MAE formulation compatible with a shared tokenizer. Subsequent models extend masked pretraining to heterogeneous recording configurations through Fourier-based positional encodings in REVE (El Ouahidi et al., 2025), topology-agnostic latent compression in LUNA (Döner et al., 2025), and a shared EEG–MEG sensor space in BrainOmni (Xiao et al., 2025). These models demonstrate the flexibility of masked pretraining but couple the objective with specialized mechanisms for sensor heterogeneity.

Contrastive and autoregressive objectives have also been explored in EEG, although less extensively than masked prediction. Banville et al. (2021) examined relative positioning, temporal shuffling, and contrastive predictive coding as self-supervised tasks for clinical EEG, illustrating several ways in which temporal structure can provide supervision without annotations. For consumer wearable EEG, Niu et al. (2025) used subject-level contrastive learning to capture inter-subject variability, following evidence that cross-session participant pairing can preserve important subject-dependent features in wearable EEG and other biosignals (Abbaspourazad et al., 2024). NeuroLM (Jiang et al., 2025) provides an EEG-specific formulation in which embeddings from a frozen vector-quantized encoder are processed using multi-channel next-token prediction.

Although all three objective families have been applied to EEG, differences in training data, channel layouts, architectures, and evaluation protocols prevent their effects from being isolated through cross-paper comparison. Recent surveys catalogue the available methods but leave this relationship unresolved (Weng et al., 2025). A controlled comparison combining downstream evaluation with probes of established EEG features is therefore needed to determine how the objectives shape learned representations and how these differences relate to performance across tasks.

Based on this literature, we compare three EEG pretraining formulations: NeuroLM-style AR (Jiang et al., 2025) for prediction across time, LaBraM-style MAE (Jiang et al., 2024) for predicting masked EEG tokens, and subject-level contrastive learning (SLCL) for aligning recordings from the same participant across sessions (Niu et al., 2025; Abbaspourazad et al., 2024). We retain NeuroLM's EEG autoregressive task without its language components, replace LaBraM's raw-patch inputs with frozen tokenizer features, and train SLCL on the same features. All three formulations share a frozen tokenizer and GPT-2 backbone architecture (Radford et al., 2019), while retaining their respective prediction, masking, and pairing procedures.

We train each formulation with a 306 M-parameter backbone on 25%, 50%, and 100% of a 50,894-hour four-channel meditation and sleep EEG corpus. Using matched frozen-representation evaluation protocols, we assess sleep staging, Hori staging, age regression, gender classification, and inter-subject variability (ISV), alongside the public clinical EEG benchmarks TUAB and TUEV. We further test whether objective-specific strengths persist after low-rank adaptation (LoRA) fine-tuning. To interpret these differences, we examine how well established EEG features can be recovered from the learned representations and how their downstream usefulness changes across model layers. Our contributions are as follows:

- **Controlled comparison of pretraining objectives.** We compare AR, MAE, and SLCL across seven tasks under shared conditions, including evaluation with varying pretraining data and LoRA fine-tuning.
- **Objective-specific EEG feature encoding.** We characterize how the three objectives shape spectral, spatial, and temporal feature profiles to interpret their downstream strengths.
- **Layer-wise representation analysis.** We examine how task and pretraining objective affect downstream performance across model depth.

2 METHODS

We use a two-stage framework to compare AR, MAE, and SLCL with a common input representation and backbone architecture. Stage 1 trains a vector-quantized tokenizer, whose encoder and codebook are then frozen. Stage 2 trains a GPT-2 backbone under each objective using the same encoder features. Figure 1 summarizes the framework.

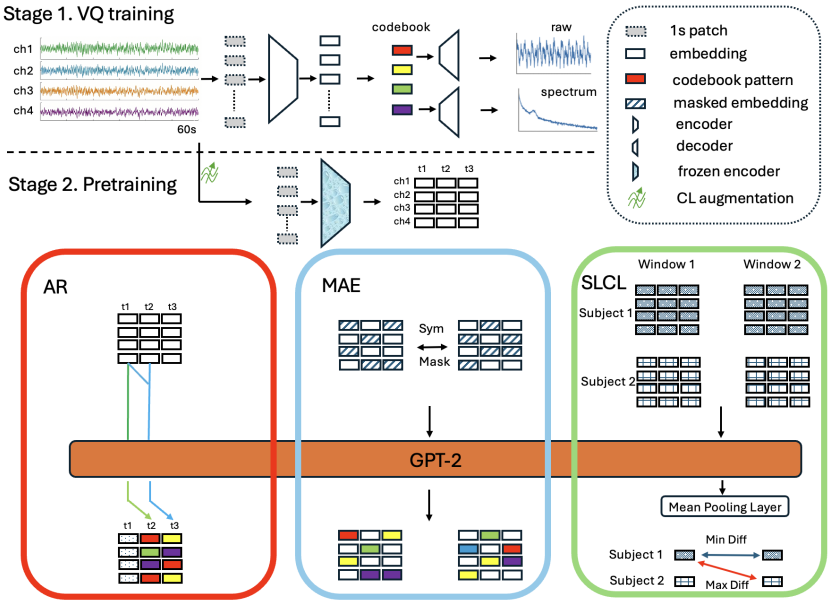

Figure 1: **Two-stage architecture and the three pretraining objectives.** Stage 1 segments 60 s four-channel EEG windows into one-second single-channel patches and trains a vector-quantized tokenizer to reconstruct each patch's waveform and Fourier magnitude. The encoder and codebook are then frozen. Stage 2 passes the encoder features to a shared GPT-2 backbone trained with autoregressive prediction (AR), masked autoencoding (MAE), or subject-level contrastive learning (SLCL), shown in the three lower panels.

Notation. Let $\mathbf{X} \in \mathbb{R}^{C\times T}$ denote a multi-channel EEG window with $C = 4$ channels and T time samples at sampling rate $f_s = 128$ Hz. We segment $\mathbf{X}$ into non-overlapping single-channel patches $\{x_i \in \mathbb{R}^P\}_{i=1}^N$ with patch length $P = 128$ samples (one second of signal), the total token count is $N = CT/P$. For the 60 s windows used throughout pretraining, $T = 7680$ and $N = 240$.

2.1 STAGE 1: VECTOR-QUANTIZED EEG TOKENIZER

The tokenizer follows NeuroLM and LaBraM (Jiang et al., 2024; 2025). A Transformer maps each patch to a 128-dimensional vector, assigns it to the nearest of 1,024 cosine-normalized codebook entries, and reconstructs its waveform and Fourier magnitude. Its VQ-VAE objective is

$$\mathcal{L}_{\mathrm{VQ}} = \sum_{i=1}^{N} \Big[\|o_i^t - x_i\|_2^2 + \|o_i^f - \phi_i\|_2^2 + \|\mathrm{sg}(\ell_2(q_i)) - \ell_2(u_{z_i})\|_2^2 + \|\ell_2(q_i) - \mathrm{sg}(\ell_2(u_{z_i}))\|_2^2 \Big], \quad (1)$$

Here, x_i and ϕ_i are the waveform and Fourier-magnitude targets for patch i, and o_i^t and o_i^f are the corresponding decoder predictions. The encoder output $h_i \in \mathbb{R}^{768}$ is projected to $q_i \in \mathbb{R}^{128}$ for quantization, with z_i denoting the selected codebook index and $u_{z_i} \in \mathbb{R}^{128}$ its embedding. The operator $\ell_2(\cdot)$ denotes unit-length normalization, $\mathrm{sg}(\cdot)$ stops gradients, and N is the number of patches. The codebook uses exponential moving-average updates (van den Oord et al., 2017). We omit NeuroLM's text-alignment loss. The same trained encoder and codebook are frozen and reused in every Stage-2 run.

2.2 Stage 2: Shared Backbone, Three Objectives

All three formulations use the same GPT-2 architecture with 305.7 M parameters and 24 blocks. Each backbone receives the continuous feature sequence $h_{1:N} \in \mathbb{R}^{N \times D_e}$ from the frozen encoder, where $D_e = 768$. AR predicts subsequent codebook indices, MAE predicts indices at masked positions, and SLCL contrasts pooled representations of augmented recordings from the same participant.

Each objective is trained on 25%, 50%, and 100% of the corpus for 50 epochs using AdamW, a batch size of 1,024, and a cosine learning-rate schedule, with learning rates selected for stable training. Within each data fraction, this design matches training exposure and schedule across objectives, without equalizing computational cost or requiring a common convergence criterion. Full hyperparameters are reported in Appendix A.

Autoregressive Pretraining AR follows NeuroLM's multi-channel autoregressive formulation, predicting the next time step's EEG codebook indices from preceding encoder features (Jiang et al., 2025). The frozen tokenizer produces a continuous feature and a discrete codebook index for each one-second channel patch. The backbone receives the continuous features, while the indices serve as prediction targets. A stair-stepping attention mask allows each backbone token at time t to attend to features from all channels up to and including t, then predict its own channel's codebook index at $t + 1$. Training minimizes the next-time-step cross-entropy:

$$\mathcal{L}_{\mathrm{AR}} = -\sum_{t=1}^{S-1} \sum_{c=1}^{C} \log p_\theta(z_{c,t+1} \mid h_{1:C,1:t}), \tag{2}$$

where $C = 4$ is the number of channels, S is the number of one-second time steps, $z_{c,t+1}$ is the target codebook index for channel c at time $t + 1$, and $h_{1:C,1:t}$ denotes the encoder features at positions up to time t across all channels. The predictor, parameterized by θ, assigns probabilities over the 1,024 codebook entries. Our AR variant retains only EEG next-token prediction, omitting NeuroLM's text branch, merged EEG and text vocabulary, and EEG–text alignment loss.

Masked Autoencoder Pretraining

MAE follows the masked EEG modelling formulation of LaBraM (Jiang et al., 2024), predicting codebook indices at masked positions from the visible context (Fig. 1, MAE panel), but operates on frozen Stage-1 encoder features rather than raw EEG patches, unlike the original LaBraM. We sample a mask $M \subset \{1, \ldots, N\}$ uniformly without replacement from the channel–time positions, with $|M| = \lfloor 0.5N \rfloor$, and replace the features at masked positions with a learnable embedding $e_{[\mathrm{MASK}]} \in \mathbb{R}^{D_e}$, yielding a corrupted sequence $\tilde{h}_{1:N}$. The backbone processes this corrupted sequence and predicts the tokenizer's codebook indices at the masked positions, where the cross-entropy loss is evaluated. Following LaBraM, a second pass uses the complementary mask $\bar{M}$, yielding $\tilde{h}'_{1:N}$ and making the previously visible positions prediction targets. Both passes reuse the same tokenizer targets, and their losses are summed:

$$\mathcal{L}_{\mathrm{MAE}} = -\sum_{i \in M} \log p_\theta\big(z_i \mid \tilde{h}_{1:N}\big) \; - \sum_{i \in \bar{M}} \log p_\theta\big(z_i \mid \tilde{h}'_{1:N}\big), \tag{3}$$

Subject-Level Contrastive Learning Pretraining We adopt subject-level contrastive learning following evidence that participant-level pairing improves age and sex prediction over segment-level pairing in wearable EEG (Niu et al., 2025), consistent with findings on demographic and physiological targets in PPG and ECG (Abbaspourazad et al., 2024). Positive pairs consist of 60 s windows from different sessions of the same participant, while windows from other participants serve as negatives. Each window randomly receives one raw-signal augmentation: temporal masking, amplitude scaling, additive noise, or a time shift. Backbone outputs are mean-pooled and ℓ_2-normalized to obtain a window embedding g. Let g_1^i and g_2^i denote the embeddings of the two augmented windows from participant i. Following Niu et al. (2025); Abbaspourazad et al. (2024), we combine a symmetric contrastive loss with KoLeo regularization (Sablayrolles et al., 2019):

$$\mathcal{L}_{\mathrm{CL}} = -\frac{1}{2B} \sum_{a \neq b} \sum_{i=1}^{B} \log \frac{\exp(\mathrm{sim}(g_a^i, g_b^i)/\tau)}{\sum_{j \neq i} \exp(\mathrm{sim}(g_a^i, g_b^j)/\tau)} - \frac{\lambda_{\mathrm{KL}}}{2B} \sum_{a=1}^{2} \sum_{i=1}^{B} \log \min_{j \neq i} \|g_a^i - g_a^j\|_2^2, \tag{4}$$

where $(a, b) \in \{(1, 2), (2, 1)\}$, $\tau = 0.05$, and $\lambda_{\mathrm{KL}} = 10^{-5}$. Exact augmentation ranges and implementation details are in Appendix A.

3 EXPERIMENTAL SETUP

Datasets. We use non-public consumer EEG datasets for pretraining and in-domain evaluation, and public clinical EEG datasets for external evaluation. All consumer EEG was recorded from the Muse 2 and Muse S headbands, which share the same four-electrode layout (TP9, AF7, AF8, TP10) and a native sampling rate of 256 Hz. Recordings were obtained under a data-use agreement with the device manufacturer and de-identified before analysis. Participants had provided informed consent for the use of their EEG data in scientific research at the time of collection. Pretraining used two large-scale unlabelled datasets of meditation and sleep sessions totaling 50,894 hours of EEG. Downstream evaluation used five supervised tasks covering sleep staging, Hori staging (Strauss et al., 2022), age regression, gender classification, and inter-subject variability (ISV), summarized in Table 1 and detailed in Appendix B. Participants in the consumer EEG evaluation sets are excluded from pretraining. To test transfer beyond the source corpus, we additionally evaluate on publicly available TUAB (binary abnormality detection) and TUEV (six-class event classification). For each external dataset, we construct a four-channel montage approximating the Muse channel layout using six available electrodes referenced to a synthetic frontal signal (exact mapping provided in Appendix B).

Table 1: Pretraining and consumer-EEG downstream datasets.

Dataset	Sessions	Win/sess	Window len.	Mean dur.	Total EEG	Type	Metric
Pretraining							
Meditation	149,429	10	60 s	11.7 min	29,138 h	unsupervised	—
Sleep	2,754	300	60 s	7.9 h	21,756 h	unsupervised	—
Downstream, consumer EEG							
sleep	355	250	30 s	7.0 h	2,485 h	5-class	$\kappa \uparrow$
hori	85	250	10 s	6.5 h	553 h	9-class	$\kappa \uparrow$
agereg	4,149	10	60 s	13.6 min	940 h	regression	MAE (yr) $\downarrow$
gender	4,188	10	60 s	13.6 min	949 h	binary	bal. acc. $\uparrow$
ISV	3,310	5	60 s	10.4 min	574 h	662-class	top-1 (KNN) $\uparrow$

Preprocessing and sampling. Pretraining and consumer-EEG downstream recordings undergo notch filtering at 50 and 60 Hz, bandpass filtering at 0.1–49 Hz, and resampling to 128 Hz. Pretraining samples 60 s windows uniformly from each session. Downstream window lengths follow the task annotations, with 30 s for sleep, 10 s for Hori, and 60 s for age, gender, and ISV. TUAB and TUEV recordings undergo montage conversion, line-noise notch filtering, 0.1–49 Hz bandpass filtering, and resampling to 128 Hz before window extraction. TUAB recordings are divided into non-overlapping 10 s windows from the start of each recording, discarding incomplete tails. TUEV supplies one 5 s window centered on each annotated event, allowing windows to overlap. Dataset construction and sampling details are provided in Appendix B.

Frozen-representation evaluation. We evaluate the pretrained representations against two baselines: VQ-only, which uses the frozen Stage-1 encoder without backbone pretraining, and bandpower, which uses 48 handcrafted spectral features (Appendix B). For VQ-only, each window embedding is obtained by mean-pooling the VQ encoder's final-layer outputs. For the three pretrained models, we instead pool the corresponding backbone's final-layer outputs.

For consumer-EEG classification tasks, we fit logistic regression probes, while age prediction uses ridge regression. Using the metrics in Table 1, we estimate performance with five-fold subject-disjoint cross-validation and nested validation. In each iteration, four outer folds are split 80/20 by participant to select regularization, and the remaining fold is held out for testing. ISV uses five sessions from each of 662 participants. Each session serves once as the query, with the other four forming the gallery, yielding five evaluation folds. Each query window is assigned to its nearest gallery participant by cosine similarity, giving chance accuracy of $1/662 \approx 0.15\%$. We report means and population standard deviations across folds, with full probe settings in Appendix B.

For TUAB and TUEV, we retain the dataset-provided test partitions, following the fixed-test evaluation used by REVE (El Ouahidi et al., 2025). The remaining participants are divided into training and validation sets in ratios of 80/20 for TUAB and 90/10 for TUEV. We train a single fully connected layer preceded by dropout on frozen, mean-pooled embeddings using cross-entropy loss, with validation balanced accuracy guiding learning-rate scheduling and early stopping. Performance is reported as mean balanced accuracy and population standard deviation over 10 probe seeds, evaluated

on individual TUAB windows or TUEV events. Full training settings and differences from published protocols are given in Appendix B.

LoRA fine-tuning. To evaluate transfer after adapting the backbone, we apply rank-8 LoRA to the attention projections in every block of the three models, keeping the Stage-1 tokenizer frozen. Learning rate, training duration, and adapter dropout are selected on validation data for each model and task. Sleep, Hori, age, and gender use participant-grouped five-fold cross-validation, while TUAB and TUEV use fixed splits and five seeds. ISV uses frozen nearest-neighbor evaluation and is not included in fine-tuning. Task-specific splits and training details are provided in Appendix D.

Feature probing. To characterize how EEG properties are encoded in the learned representations, we analyze 20,872 sixty-second windows from the labelled meditation corpus, covering 6,122 participants after artifact rejection. Windows are taken consecutively from session starts, with up to four sampled per participant. For every window, we compute EEG features directly from the signal and extract four mean-pooled embeddings, one from the frozen VQ encoder and one from each of the AR, MAE, and SLCL backbones. For each representation and scalar feature, we fit a separate ridge regression probe that predicts the measured feature value from the embedding. We use the same five subject-disjoint folds for all representations, training on four folds and testing on the held-out fifth. Held-out R^2 measures how closely the predictions recover the measured features, with higher scores indicating greater linear recoverability (Belinkov, 2022; Tang et al., 2026).

The 13 features in Figure 2 are organized into four families. Full feature definitions and evaluation details are provided in Appendix E.

- *Relative band power.* This measures the log proportion of summed band power in the delta (1–4 Hz), alpha (8–12 Hz) and beta (16–30 Hz) bands.
- *Spectral shape.* We estimate the spectral slope using Theil–Sen regression, which takes the median slope across pairs of observations (Abbaspoor et al., 2026), applied here to log-power and log-frequency. This family includes the spectral exponent, defined as the negative fitted slope, and fitted log-power at 10 Hz, which describes the spectrum's level at that frequency. It also includes alpha-peak frequency, the location of the strongest eligible peak in 7–13 Hz, and alpha-peak height above the fitted aperiodic background (Donoghue et al., 2020).
- *Cross-channel structure.* This family includes alpha/delta magnitude-squared coherence, which measures frequency-specific association between channels (Halliday et al., 1995). Frontal (AF8 versus AF7) and temporoparietal (TP10 versus TP9) alpha asymmetry measure right-minus-left log alpha power (Smith et al., 2017; Hashemi et al., 2016).
- *Power trends.* These quantify whether relative delta, alpha and beta power rises or falls across the window, using Kendall's τ rank correlation between time and the channel-averaged one-second log-relative-power series (Abbaspoor et al., 2026).

4 Results

4.1 Downstream Evaluation

Linear probing. The shared tokenizer already supports better downstream performance than hand-crafted bandpower features on all seven tasks (Table 2). Further backbone pretraining improves on this baseline in different ways: AR and MAE improve sleep, age, gender, and TUEV performance, while SLCL produces the largest ISV gain, increasing top-1 accuracy from 0.243 to 0.506. However, none of the objectives substantially improves Hori performance, and SLCL performs worse than VQ-only on TUEV.

MAE achieves the highest mean performance on six of seven tasks, with its clearest advantages on sleep, age, gender and TUEV, while SLCL produces a substantially higher observed ISV score. MAE also exceeds VQ-only on ISV, indicating that participant-linked information is encoded without explicit subject-level contrast. Hori and TUAB offer less separation between the leading objectives, with MAE exceeding AR by 0.004 on Hori and SLCL by 0.007 on TUAB.

LoRA fine-tuning. LoRA fine-tuning largely preserves the performance patterns observed under frozen probing (Table 2). The ordering remains MAE, AR, then SLCL on sleep, Hori, age, gender, and TUEV, while AR has a higher observed TUAB score than SLCL after LoRA fine-tuning.

Table 2: Downstream performance (mean ± population standard deviation). Bold and underline indicate the best and second-best scores per task within each evaluation block. A dash indicates an unevaluated setting.

Model	Sleep $\kappa \uparrow$	Hori $\kappa \uparrow$	Age MAE (yr) $\downarrow$	Gender bal. acc. $\uparrow$	ISV top-1 $\uparrow$	TUAB bal. acc. $\uparrow$	TUEV bal. acc. $\uparrow$
Bandpower	$0.497_{\pm 0.016}$	$0.272_{\pm 0.020}$	$9.22_{\pm 0.24}$	$0.631_{\pm 0.004}$	$0.196_{\pm 0.009}$	$0.739_{\pm 0.002}$	$0.384_{\pm 0.012}$
Linear probing							
VQ only	$0.600_{\pm 0.012}$	$0.304_{\pm 0.021}$	$8.80_{\pm 0.12}$	$0.642_{\pm 0.003}$	$0.243_{\pm 0.004}$	$0.747_{\pm 0.005}$	$0.439_{\pm 0.015}$
AR	$\underline{0.666}_{\pm 0.010}$	$\underline{0.308}_{\pm 0.022}$	$\underline{8.06}_{\pm 0.11}$	$\underline{0.680}_{\pm 0.008}$	$0.201_{\pm 0.005}$	$0.768_{\pm 0.003}$	$\underline{0.479}_{\pm 0.018}$
MAE	$\mathbf{0.692}_{\pm 0.008}$	$\mathbf{0.312}_{\pm 0.016}$	$\mathbf{7.61}_{\pm 0.13}$	$\mathbf{0.722}_{\pm 0.008}$	$\underline{0.305}_{\pm 0.010}$	$\mathbf{0.780}_{\pm 0.002}$	$\mathbf{0.537}_{\pm 0.009}$
SLCL	$0.628_{\pm 0.005}$	$0.283_{\pm 0.018}$	$8.54_{\pm 0.14}$	$0.642_{\pm 0.010}$	$\mathbf{0.506}_{\pm 0.008}$	$\underline{0.773}_{\pm 0.003}$	$0.398_{\pm 0.016}$
LoRA fine-tuning							
AR	$\underline{0.682}_{\pm 0.011}$	$\underline{0.307}_{\pm 0.022}$	$\underline{7.90}_{\pm 0.16}$	$\underline{0.693}_{\pm 0.007}$	–	$\underline{0.804}_{\pm 0.002}$	$\underline{0.503}_{\pm 0.010}$
MAE	$\mathbf{0.699}_{\pm 0.016}$	$\mathbf{0.320}_{\pm 0.013}$	$\mathbf{7.09}_{\pm 0.14}$	$\mathbf{0.744}_{\pm 0.007}$	–	$\mathbf{0.817}_{\pm 0.002}$	$\mathbf{0.546}_{\pm 0.019}$
SLCL	$0.608_{\pm 0.011}$	$0.295_{\pm 0.032}$	$8.52_{\pm 0.12}$	$0.636_{\pm 0.010}$	–	$0.791_{\pm 0.004}$	$0.426_{\pm 0.011}$

4.2 Pretraining Data Scale

Increasing pretraining data generally improves downstream performance across the three objectives. Comparing models trained on 25% and 100% of the session pool, all three improve in mean performance on age, gender, and TUEV (Table 3). MAE and SLCL also improve on sleep and ISV, while SLCL gains on TUAB. The benefits vary across tasks, with little improvement on Hori and performance degradation in AR's sleep, ISV, and TUAB scores. Complete results, including the intermediate 50% fraction, are provided in Appendix C.

Table 3: Downstream performance with 25% $\rightarrow$ 100% of the pretraining data. Bold marks an improved mean at 100%, without implying statistical significance.

Objective	Sleep $\kappa \uparrow$	Hori $\kappa \uparrow$	Age MAE (yr) $\downarrow$	Gender bal. acc. $\uparrow$	ISV top-1 $\uparrow$	TUAB bal. acc. $\uparrow$	TUEV bal. acc. $\uparrow$
AR	.670 $\rightarrow$ .666	.319 $\rightarrow$ .308	8.16 $\rightarrow$ **8.06**	.677 $\rightarrow$ **.680**	.227 $\rightarrow$ .201	.772 $\rightarrow$ .768	.447 $\rightarrow$ **.479**
MAE	.669 $\rightarrow$ **.692**	.311 $\rightarrow$ **.312**	8.15 $\rightarrow$ **7.61**	.694 $\rightarrow$ **.722**	.257 $\rightarrow$ **.305**	.778 $\rightarrow$ **.780**	.506 $\rightarrow$ **.537**
SLCL	.574 $\rightarrow$ **.628**	.286 $\rightarrow$ .283	8.76 $\rightarrow$ **8.54**	.633 $\rightarrow$ **.642**	.400 $\rightarrow$ **.506**	.745 $\rightarrow$ **.773**	.364 $\rightarrow$ **.398**

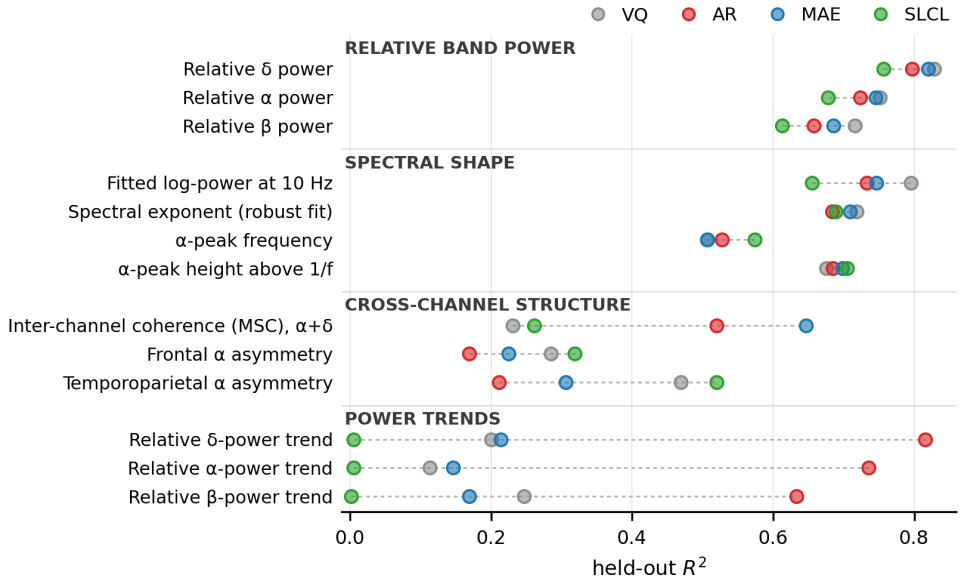

Figure 2: **EEG feature encoding.** Linear-probe prediction of 13 EEG features in the frozen VQ tokenizer and AR, MAE, and SLCL representations, measured by the participant-weighted held-out R^2 of ridge-regression probes.

4.3 EEG Feature Encoding

The feature probes show that band-power information is already encoded by the shared tokenizer. VQ-only reaches $R^2 = 0.829$, 0.752, and 0.716 for relative δ, α, and β power, respectively, exceeding all three backbones on these features (Figure 2).

Among the pretrained backbones, MAE achieves the highest band-power probe scores, reaching 0.820, 0.746, and 0.685 for relative δ, α, and β power, respectively. MAE also achieves the highest mean coherence score at 0.647, compared with 0.231 for VQ-only, 0.520 for AR, and 0.261 for SLCL. SLCL instead leads alpha-peak frequency at 0.574, compared with 0.506–0.528 for the other representations, and is the only backbone above VQ-only on both frontal and temporoparietal alpha asymmetry (0.318 versus 0.284 and 0.520 versus 0.469, respectively).

Temporal power trends distinguish AR from the other objectives, with AR achieving the highest probe scores across all three reported bands. Its scores reach 0.815, 0.736, and 0.633 for relative δ, α, and β power trends, compared with 0.200, 0.113, and 0.247 for VQ-only, while MAE reaches 0.146–0.214 and SLCL 0.002–0.005. These rank-based targets measure whether band power rises or falls within a window, identifying a temporal component of AR's feature profile. Full feature definitions and evaluation details are provided in Appendix E.

4.4 Layer-wise Probing

All tasks reach their highest observed probe scores at intermediate layers in each model. Across the three objectives and seven tasks, these peaks occur between layers 9 and 21 of the 24-layer backbone (Figure 3). For example, MAE achieves $\kappa = 0.712$ on sleep staging at layer 19, compared with 0.692 at the final layer, illustrating the benefit of using intermediate representations for downstream prediction.

The location of these peaks also differs across objectives, with AR and MAE generally reaching their highest scores later in the network than SLCL. Median relative peak depths are 0.79 for AR, 0.75 for MAE, and 0.46 for SLCL, showing that the objectives differ in where their representations best support the downstream tasks.

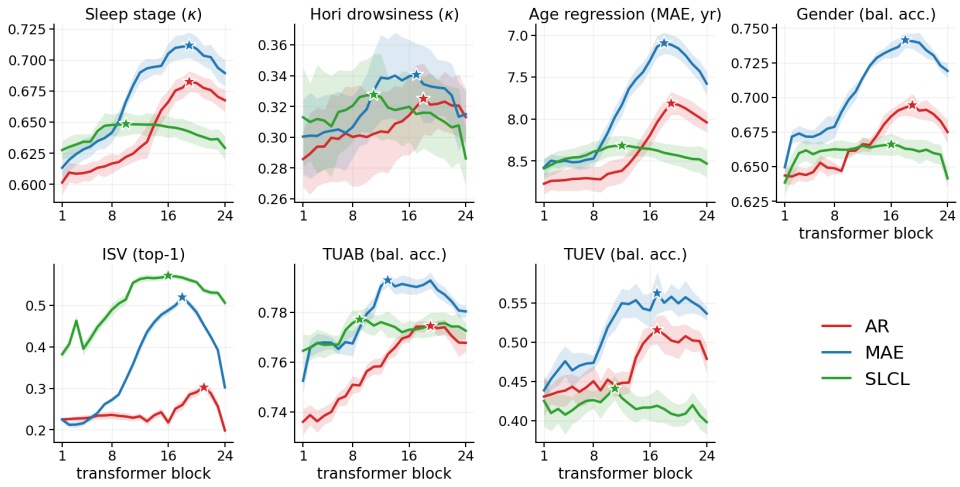

Figure 3: **Layer-wise probing on seven downstream tasks.** Curves show mean performance using representations from successive transformer layers. Stars mark the highest observed scores, and shaded bands show one population standard deviation above and below the mean. The age axis is inverted so that higher positions indicate better performance.

5 Discussion and Conclusion

The shared tokenizer provides the starting point for all three objectives. It learns EEG representations by reconstructing waveforms and spectra, encoding spectral features well and outperforming hand-

crafted bandpower features on all seven downstream tasks. Backbone pretraining further improves downstream prediction, with each objective emphasizing different EEG properties.

MAE builds on this spectral foundation by retaining band-power information while strengthening cross-channel feature encoding. Random masking distributes reconstruction targets across the window, allowing each masked token to be predicted from nearby observations and visible tokens elsewhere in time. When spectral patterns persist within a 60-second sleep or meditation window, they provide shared information for reconstructing multiple masked regions, which may encourage band-power retention. Visible tokens from other channels at the same time also support reconstruction, offering a plausible explanation for MAE's stronger coherence encoding. Spectral features are informative for sleep staging (Vallat & Walker, 2021), sex classification (Bučková et al., 2020), and abnormal EEG detection (Gemein et al., 2020), while channel relationships contribute to age prediction (Banville et al., 2024) and temporal and spatial context supports clinical event classification (Golmohammadi et al., 2019).

AR shows stronger encoding of temporal power trends. Unlike MAE, which reconstructs tokens using context from both sides in time, AR predicts future tokens from preceding observations alone. This training task may encourage learning how EEG activity changes over time. AR also improves coherence encoding, consistent with using preceding observations from all channels to predict each channel's next token. Despite these strengths, AR generally follows MAE in downstream performance, suggesting that the current tasks may not fully reward its temporal features. Longer-sequence tasks involving clear power or state changes could test their usefulness more directly. Similar band-power and spectral probe scores suggest overlapping information in AR and MAE. Their complementary temporal and cross-channel strengths motivate combining masked reconstruction with prediction of subsequent activity.

SLCL shows stronger encoding of alpha-peak frequency and alpha asymmetry. Its training objective aligns recordings from the same participant across sessions, rewarding properties that help distinguish individuals despite changes between recordings. Spectral profiles and alpha asymmetry show consistent individual differences across repeated recordings (Näpflin et al., 2007; Metzen et al., 2022). These findings connect SLCL's feature profile to its ISV performance, while MAE's improvement over VQ-only shows that participant-related information can also emerge without explicit pairing. In our experiments, same-participant contrast improves ISV performance, but has mixed effects across other tasks. This pattern is specific to the chosen pairing rule, as other contrastive designs may emphasize different information (Tian et al., 2020; Banville et al., 2021).

Increasing pretraining data improves downstream performance in most combinations of task and objective, while MAE retains the lead after LoRA fine-tuning on all evaluated tasks. Layer-wise probing further shows that all seven tasks achieve their highest probe scores before the final layer in each model. This agrees with evidence of stronger transfer from intermediate transformer layers in language models (Liu et al., 2019) and reduced transferability as higher layers specialize toward their training task in vision (Yosinski et al., 2014). SLCL generally peaks earlier than AR and MAE, possibly reflecting its focus on participant-related features that become useful at shallower layers.

Several limitations constrain the scope of these findings. We compare three specific formulations, and alternative prediction, masking, or pairing designs may produce different results within each objective family. The shared reconstruction-trained tokenizer controls the input representation but may interact differently with each formulation. Full-parameter fine-tuning was not evaluated and could change the relative performance.

Together, these findings connect EEG pretraining objectives to distinct spectral, spatial, and temporal feature profiles. Probing with established EEG features associates MAE with band-power and cross-channel structure, AR with temporal power trends, and SLCL with participant-related spectral and asymmetry features. Layer-wise evaluation further shows that their downstream usefulness changes across model depth. This controlled comparison provides an interpretable basis for selecting or combining objectives according to the signal properties and tasks of interest.

AI USE STATEMENT

Generative AI assisted manuscript organization and polishing, interpretation of existing results, and analysis code. The authors checked all code and claims against source artifacts and take responsibility for the paper.

REFERENCES

Saman Abbaspoor, Ayman Aljishi, and Kari L. Hoffman. Experience reorganizes content-specific memory traces in macaques. *Nature Neuroscience*, 29:1976–1986, 2026.

Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Hubert Banville, Maurice Abou Jaoude, Sean U.N. Wood, Chris Aimone, Sebastian C. Holst, Alexandre Gramfort, and Denis-Alexander Engemann. Do try this at home: Age prediction from sleep and meditation with large-scale low-cost mobile EEG. *Imaging Neuroscience*, 2, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Christopher S.Y. Benwell, Raquel E. London, Chiara F. Tagliabue, Domenica Veniero, Joachim Gross, Christian Keitel, and Gregor Thut. Frequency and power of human alpha oscillations drift systematically with time-on-task. *NeuroImage*, 192:101–114, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Barbora Bučková, Martin Brunovský, Martin Bareš, and Jaroslav Hlinka. Predicting sex from EEG: Validity and generalizability of deep-learning-based interpretable classifier. *Frontiers in Neuroscience*, 14:589303, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

Wenhui Cui, Christopher Sandino, Hadi Pouransar, Ran Liu, Juri Minxha, Ellen L. Zippi, Aman Verma, Anna Sedlackova, Erdrin Azemi, and Behrooz Mahasseni. CPEP: Contrastive pose-EMG pre-training enhances gesture generalization on EMG signals. In *NeurIPS Workshop on Foundation Models for the Brain and Body*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

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

Thomas Donoghue, Matar Haller, Erik J. Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Antonio H. Lara, Joni D. Wallis, Robert T. Knight, Avgusta Shestyuk, and Bradley Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, 23:1655–1665, 2020.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Vaishaal Shankar, Alexander Toshev, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *International Conference on Machine Learning (ICML)*, 2024.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G. Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Lukas A.W. Gemein, Robin T. Schirrmeister, Patryk Chrabąszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220:117021, 2020.

Meysam Golmohammadi, Amir Hossein Harati Nejad Torbati, Silvia Lopez de Diego, Iyad Obeid, and Joseph Picone. Automatic analysis of EEGs using big data and hybrid deep learning architectures. *Frontiers in Human Neuroscience*, 13:76, 2019.

D. M. Halliday, J. R. Rosenberg, A. M. Amjad, P. Breeze, B. A. Conway, and S. F. Farmer. A framework for the analysis of mixed time series/point process data: Theory and application to the study of physiological tremor, single motor unit discharges and electromyograms. *Progress in Biophysics and Molecular Biology*, 64(2–3):237–278, 1995.

Ali Hashemi, Lou J. Pino, Graeme Moffat, Karen J. Mathewson, Chris Aimone, Patrick J. Bennett, Louis A. Schmidt, and Allison B. Sekuler. Characterizing population EEG dynamics throughout adulthood. *eNeuro*, 3(6):ENEURO.0275–16.2016, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations (ICLR)*, 2025.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL-HLT*, pp. 1073–1094. Association for Computational Linguistics, 2019.

Dorothea Metzen, Erhan Genç, Stephan Getzmann, Mauro F. Larra, Edmund Wascher, and Sebastian Ocklenburg. Frontal and parietal EEG alpha asymmetry: A large-scale investigation of short-term reliability on distinct EEG systems. *Brain Structure and Function*, 227(2):725–740, 2022.

Aleksandra Miljevic, Neil W. Bailey, Fidel Vila-Rodriguez, Sally E. Herring, and Paul B. Fitzgerald. Electroencephalographic connectivity: A fundamental guide and checklist for optimal study design and evaluation. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 7(6):546–554, 2022.

Markus Näpflin, Marc Wildi, and Johannes Sarnthein. Test-retest reliability of resting EEG spectra validates a statistical signature of persons. *Clinical Neurophysiology*, 118(11):2519–2524, 2007.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *International Conference on Learning Representations (ICLR)*, 2025.

Jiansheng Niu, Aravind Ravi, Nicolas Florez Torres, David Baum, Taha Morshedzadeh, Christopher Aimone, and Maurice Abou Jaoude. Foundation models on wearable EEG using self-supervised learning. In *47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *International Conference on Learning Representations (ICLR)*, 2019.

Ezra E. Smith, Samantha J. Reznik, Jennifer L. Stewart, and John J. B. Allen. Assessing and conceptualizing frontal EEG asymmetry: An updated primer on recording, processing, analyzing, and interpreting frontal alpha asymmetry. *International Journal of Psychophysiology*, 111:98–114, 2017.

Mélanie Strauss, Jacobo D. Sitt, Lionel Naccache, and Federico Raimondo. Predicting the loss of responsiveness when falling asleep in humans. *NeuroImage*, 251:119003, 2022.

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

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Raphael Vallat and Matthew P Walker. An open-source, high-performance tool for automated sleep staging. *eLife*, 10:e70092, 2021.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Peter D. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhaohua Yang, Yuchen Liu, and Yiqiang Chen. Self-supervised learning for electroencephalogram: A systematic survey. *ACM Computing Surveys*, 57(12):1–38, 2025.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

Qi Zhang, Tianqi Du, Haotian Huang, Yifei Wang, and Yisen Wang. Look ahead or look around? A theoretical comparison between autoregressive and masked pretraining. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58819–58839, 2024.

APPENDICES

A PRETRAINING HYPERPARAMETERS

Stage-1: VQ Tokenizer Training. Table 4 lists the hyperparameters for training the vector-quantized neural tokenizer. The tokenizer is trained once and frozen before all Stage-2 runs.

Table 4: Hyperparameters for Stage-1 VQ tokenizer training.

Hyperparameters	**Values**
Encoder	
Transformer encoder layers	10
Attention heads	4
Hidden size	768
MLP size	3,072
Decoder (raw & frequency)	
Transformer decoder layers	4
Attention heads	4
Hidden size	768
Codebook	
Codebook size	$1{,}024 \times 128$
EMA decay	0.99
Commitment weight β	1.0
Training	
Batch size	400
Peak learning rate	2×10^{-5}
Minimal learning rate	1×10^{-6}
Learning rate scheduler	Cosine
Optimizer	AdamW
Adam β	$(0.9, 0.999)$
Weight decay	1×10^{-4}
Total epochs	60
Warmup epochs	5

Stage-2: Self-Supervised Pretraining. All training runs used NVIDIA H100 GPUs on 4-GPU or 8-GPU nodes. The Stage-2 models share the same optimizer schedule, data sampling, preprocessing, tokenizer, and architecture, while the objective and pretraining data fraction vary. Table 5 lists the shared backbone configuration and Table 6 the training settings.

Table 5: Shared Stage-2 backbone configuration. Parameter counts are reported for the trainable GPT-2 backbone alone and combined with the frozen Stage-1 VQ encoder.

L (blocks)	H (heads)	D (hidden)	Backbone	+ VQ encoder
24	16	1,024	305.7 M	377.0 M

SLCL augmentations and readout. Each window independently undergoes one augmentation at random, applied to the raw EEG before the frozen tokenizer. These augmentations include temporal masking of 10–30% of samples within a patch, amplitude scaling by a factor in $[0.5, 2]$, additive Gaussian noise, or a time shift of 10–100 samples. Gaussian noise is applied as $x' = x + 0.3\epsilon$, where $\epsilon_i \sim \mathcal{N}(0, (r\sigma_x)^2)$ independently, $r \sim \mathcal{U}(0.2, 0.8)$, and σ_x is computed across all elements of the input window, pooling channels and time. The model outputs are mean-pooled and ℓ_2-normalized. The temperature is $\tau = 0.05$ and the KoLeo weight is 10^{-5}, as in Eq. 4.

B DOWNSTREAM DATASETS

Consumer-EEG downstream. The consumer-EEG datasets use the preprocessing in Section 3. Their participants were selected from users excluded from pretraining.

Table 6: Hyperparameters for Stage-2 runs.

Parameter	Value
Input channels	4
Patch size	128 samples (1 s)
Block size (tokens)	240
Window length	60 s
Meditation windows / sess.	10
Sleep windows / sess.	300
Total epochs	50
Warmup epochs	3
LR schedule	cosine, peak 4–6 $\times 10^{-4}$, min 6×10^{-5}
Optimizer	AdamW, β_1=0.9, β_2=0.95, weight decay 0.1
Gradient clip	1.0
Batch size	1,024

- **Gender.** This is a binary classification task distinguishing Male from Female participants. Each user contributes one meditation session. The dataset contains 4,188 sessions in total (2,110 Male / 2,078 Female).
- **Age Regression.** This is a regression task predicting participant age in the range of 18 to 80 years. The dataset contains 4,149 sessions with a mean age of 48.2 years and a median of 47.0 years.
- **Sleep.** This is a 5-class classification task over the stages Wake, N1, N2, N3, and REM. Labels were assigned by one expert using 30-second epochs following AASM guidelines. The dataset contains 355 sessions, yielding 265,270 scored epochs (25,240 W / 10,468 N1 / 128,876 N2 / 42,608 N3 / 58,078 REM). An additional 31,215 unscored epochs are excluded.
- **Hori.** This is a 9-class classification task over the Hori wake-to-sleep transition stages H0 through H8 (Strauss et al., 2022). Labels were assigned by one expert scorer using 10-second epochs. The dataset contains 85 sessions, yielding 24,666 scored epochs (12,039 H0 / 154 H1 / 3,443 H2 / 2,672 H3 / 719 H4 / 430 H5 / 1,002 H6 / 1,571 H7 / 2,636 H8).
- **Inter-Subject Variability.** This is a 662-class nearest neighbor evaluation task. The dataset was constructed by selecting 662 subjects from the meditation corpus, each contributing exactly 5 meditation sessions (3,310 sessions total). It is evaluated via cosine-similarity 1-NN over ℓ_2-normalized pooled embeddings.

Mean-pooling produces one embedding per window, with 1,024 dimensions for each pretrained backbone and 768 dimensions for VQ-only. Participants are divided into five disjoint folds shared across models. In each iteration, one fold is held out for testing, while participants in the remaining four folds are split 80/20 into training and validation subsets for regularization selection. Logistic regression uses at most 5,000 iterations and $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Ridge regression uses $\alpha \in \{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$. Regularization is selected using validation Cohen's κ for sleep and Hori, balanced accuracy for gender, and mean absolute error for age. The probe is then refitted on all four non-test folds using the selected regularization before evaluation on the held-out test fold.

For ISV, each of the 662 participants contributes five sessions. In each of five evaluation folds, one session per participant provides query windows, while the other four provide gallery windows. Each query window is assigned the participant label of its nearest gallery window using cosine similarity between ℓ_2-normalized embeddings, without combining predictions across windows. Every session serves once as the query session, and we report mean top-1 accuracy and population standard deviation across the five folds.

The 48-dimensional bandpower baseline. The baseline concatenates 12 spectral features from each of TP9, AF7, AF8, and TP10, giving 48 features per window (Table 7). Power spectral density (PSD) is estimated using Welch's method with four-second Hann segments, 50% overlap, and constant detrending. Band powers are computed by trapezoidal integration, and relative powers are normalized by total power over 0.5–45 Hz.

TUAB and TUEV. TUAB is a binary normal/abnormal EEG classification task, while TUEV classifies events as spike and sharp wave, generalized periodic epileptiform discharge, periodic

Table 7: Spectral features computed for each channel in the bandpower baseline, giving 48 features across four channels. $S_c(f)$ denotes the Welch power spectral density and $P_{c,\mathrm{tot}}$ the total power over 0.5–45 Hz.

Feature group	Definition	Count
Log band powers	$\log_{10} P_{c,b}$ for δ [0.5,4), θ [4,8), α [8,13), β [13,30), and γ [30,45) Hz	5
Relative band powers	$P_{c,b}/P_{c,\mathrm{tot}}$ for the same five bands	5
Spectral slope	OLS slope of $\log_{10} S_c(f)$ against $\log_{10} f$	1
Log total power	$\log_{10} P_{c,\mathrm{tot}}$	1

lateralized epileptiform discharge, eye movement, artifact, or background. Both retain the dataset-provided test partition. The remaining participants are split into training and validation sets at 80/20 for TUAB and 90/10 for TUEV, with the same splits used across models and probe seeds.

Both datasets are converted to an approximate four-channel Muse montage from six source electrodes (FP1, FP2, F7, F8, T5, T6). With the synthetic frontal reference $r = (FP1 + FP2)/2$, the model inputs are

$$x_{\mathrm{TP9}} = T5 - r, \qquad x_{\mathrm{AF7}} = (FP1 + F7)/2 - r,$$
$$x_{\mathrm{AF8}} = (FP2 + F8)/2 - r, \quad x_{\mathrm{TP10}} = T6 - r.$$

Each constructed channel is assigned the positional embedding of its corresponding Muse channel (TP9, AF7, AF8, or TP10). The source electrodes approximate these locations but do not match them exactly. Recordings undergo 0.1–49 Hz bandpass filtering, line-noise notch filtering, and resampling to 128 Hz before window extraction. TUAB uses non-overlapping 10 s windows from the start of each recording, discarding incomplete tails and assigning the recording label to each window. For TUEV, we extract one 5 s window centered on the midpoint of each annotated event.

We train a single fully connected layer preceded by dropout on frozen, mean-pooled embeddings using cross-entropy loss and AdamW. Validation balanced accuracy guides learning-rate scheduling, early stopping, and checkpoint selection. We evaluate the selected checkpoint on individual TUAB windows or TUEV events and report mean balanced accuracy and population standard deviation across ten random seeds using the same data split.

C Pretraining Data Scale: Complete Results

Data fractions subsample pretraining sessions, while the tokenizer, model architecture, optimizer, 50-epoch schedule, and probing protocol remain fixed. Table 8 reports all three objectives at 25%, 50%, and 100% of the session pool.

Table 8: Downstream evaluation scores for AR, MAE, and SLCL at three pretraining data fractions, reported as mean ± population standard deviation over five folds (consumer EEG) or 10 probe seeds (TUAB/TUEV). Bold marks the best score per task and underline the second-best within each data fraction.

Objective	Sleep $\kappa\uparrow$	Hori $\kappa\uparrow$	Age MAE (yr) $\downarrow$	Gender bal. acc. $\uparrow$	ISV top-1 $\uparrow$	TUAB bal. acc. $\uparrow$	TUEV bal. acc. $\uparrow$
25% data							
AR	**0.670**$_{\pm 0.010}$	**0.319**$_{\pm 0.020}$	8.16$_{\pm 0.11}$	0.677$_{\pm 0.004}$	0.227$_{\pm 0.006}$	0.772$_{\pm 0.004}$	0.447$_{\pm 0.017}$
MAE	0.669$_{\pm 0.008}$	0.311$_{\pm 0.019}$	**8.15**$_{\pm 0.08}$	**0.694**$_{\pm 0.007}$	0.257$_{\pm 0.007}$	**0.778**$_{\pm 0.003}$	**0.506**$_{\pm 0.018}$
SLCL	0.574$_{\pm 0.009}$	0.286$_{\pm 0.027}$	8.76$_{\pm 0.12}$	0.633$_{\pm 0.006}$	**0.400**$_{\pm 0.008}$	0.745$_{\pm 0.005}$	0.364$_{\pm 0.019}$
50% data							
AR	**0.672**$_{\pm 0.012}$	**0.313**$_{\pm 0.020}$	**8.11**$_{\pm 0.12}$	0.682$_{\pm 0.004}$	0.200$_{\pm 0.010}$	0.768$_{\pm 0.005}$	0.467$_{\pm 0.017}$
MAE	0.664$_{\pm 0.009}$	0.310$_{\pm 0.027}$	8.17$_{\pm 0.14}$	**0.698**$_{\pm 0.006}$	0.285$_{\pm 0.006}$	**0.774**$_{\pm 0.005}$	**0.520**$_{\pm 0.019}$
SLCL	0.599$_{\pm 0.005}$	0.280$_{\pm 0.013}$	8.68$_{\pm 0.09}$	0.635$_{\pm 0.003}$	**0.439**$_{\pm 0.008}$	0.754$_{\pm 0.006}$	0.395$_{\pm 0.019}$
100% data							
AR	0.666$_{\pm 0.010}$	0.308$_{\pm 0.022}$	8.06$_{\pm 0.11}$	0.680$_{\pm 0.008}$	0.201$_{\pm 0.005}$	0.768$_{\pm 0.003}$	0.479$_{\pm 0.018}$
MAE	**0.692**$_{\pm 0.008}$	**0.312**$_{\pm 0.016}$	**7.61**$_{\pm 0.13}$	**0.722**$_{\pm 0.008}$	0.305$_{\pm 0.010}$	**0.780**$_{\pm 0.002}$	**0.537**$_{\pm 0.009}$
SLCL	0.628$_{\pm 0.005}$	0.283$_{\pm 0.018}$	8.54$_{\pm 0.14}$	0.642$_{\pm 0.010}$	**0.506**$_{\pm 0.008}$	0.773$_{\pm 0.003}$	0.398$_{\pm 0.016}$

D LoRA Fine-Tuning Protocol

We fine-tune the three pretrained models using rank-8 LoRA adapters in the attention projections of every backbone block, while keeping the Stage-1 tokenizer frozen. Learning rate, adapter dropout, and training duration are selected on validation data for each model and task.

For the consumer-EEG tasks of sleep, Hori, age, and gender, we use five participant-disjoint folds, holding out one fold for testing in each iteration. For TUAB and TUEV, we repeat fine-tuning five times with different random seeds, using the same training, validation, and official test partitions in every run. We report means and population standard deviations across the five folds or five runs, respectively. Table 2 compares fine-tuning performance with the frozen-probe results.

E Feature Probing: Protocol and Definitions

Sampling and representations. The analysis uses 20,872 sixty-second windows from 6,122 participants in the labelled meditation corpus after artifact rejection. We take consecutive, non-overlapping windows from the start of each recording, with up to four windows per participant. EEG features and embeddings are extracted from the same windows. For each window, we mean-pool the VQ encoder outputs and each backbone's final-layer outputs to obtain four representations for comparison.

Probe fitting and scoring. For each EEG feature, we train a separate ridge regression probe on each of the four representations to predict the feature value from the window embedding (Belinkov, 2022; Tang et al., 2026). Participants are divided into five disjoint folds shared across representations. In each iteration, one fold is held out for testing, while participants in the remaining four folds are used for training and validation to select regularization.

We combine the held-out predictions from all five folds and evaluate them against the measured feature values using R^2. Each representation is evaluated on the same windows with valid targets, with higher scores indicating more accurate recovery of the feature from the embedding. For features measured separately at multiple channels or channel-pair groups, we average their individual R^2 scores to obtain one value per representation in Figure 2.

Shared spectral estimation. For each channel, we estimate the power spectral density using Welch's method with four-second Hann windows, 50% overlap, and mean removal within each segment (Welch, 1967). The sampling rate is 128 Hz. We use six frequency bands: $\delta = [1, 4)$, $\theta = [4, 8)$, $\alpha = [8, 12)$, $\sigma = [12, 16)$, $\beta = [16, 30)$, and $\gamma = [30, 45]$ Hz. Band power is calculated by trapezoidal integration over the frequency bins within each band.

Relative delta, alpha, and beta power. For each channel and band b, we calculate $\log_{10}(P_b / \sum_k P_k)$, where P_b is the band power and the denominator sums power across all six bands. We fit separate probes for each channel's delta, alpha, and beta values. For each band, Figure 2 reports the mean R^2 across the four channel-specific probes.

Spectral exponent and fitted log-power at 10 Hz. For each channel, we fit log-power against log-frequency over 2–40 Hz using Theil–Sen regression, whose slope is the median of the slopes between pairs of observations (Abbaspoor et al., 2026). Writing the fitted spectrum as $\log_{10} P(f) = a + s \log_{10} f$, we define the spectral exponent as $-s$ and fitted log-power at 10 Hz as $a + s \log_{10}(10)$. The intercept is calculated as $a = \mathrm{median}(\log_{10} P) - s\,\mathrm{median}(\log_{10} f)$. These are two separate probe targets, each summarized by averaging R^2 across the four channels.

Alpha-peak frequency and height. For each channel, we fit the power spectrum over 2–40 Hz as an aperiodic background plus oscillatory peaks (Donoghue et al., 2020). Among peaks with center frequencies within 7–13 Hz, we select the peak with the greatest height above the fitted background. Its center frequency defines alpha-peak frequency, while its height above the background in log-power units defines alpha-peak height. Height therefore measures the prominence of the peak rather than total alpha-band power.

We retain only spectra for which the fitted model explains at least 90% of the variance in log power. Separate regression probes predict alpha-peak frequency at TP9 and TP10 and alpha-peak height

at all four channels. For each feature, Figure 2 reports the average R^2 across the corresponding channel-specific probes.

Inter-channel coherence. We calculate magnitude-squared coherence between channels x and y as $|S_{xy}(f)|^2/[S_{xx}(f)S_{yy}(f)]$, where S_{xy} is the cross-spectrum and S_{xx} and S_{yy} are the power spectra (Halliday et al., 1995). Estimation uses the native reference and the same Welch settings described above. We average coherence over frequency bins separately within delta [1,4) Hz and alpha [8,12) Hz. Each band provides three targets: AF7–AF8 coherence, TP9–TP10 coherence, and mean coherence across the four frontal–temporal channel pairs. We fit a separate probe for each of these six targets and average their R^2 scores for the coherence value in Figure 2. This sensor-space measure describes spectral association between channels rather than establishing neural communication (Miljevic et al., 2022).

Alpha asymmetry. Using the alpha-band power defined above, we calculate asymmetry as right-minus-left natural-log power (Smith et al., 2017; Hashemi et al., 2016), using AF8 and AF7 for frontal asymmetry and TP10 and TP9 for temporoparietal asymmetry. Positive values indicate greater right-channel alpha power, while negative values indicate greater left-channel alpha power. Each asymmetry measure provides one scalar target per window, with a separate probe fitted for each representation.

Relative delta-, alpha-, and beta-power trends. These features measure whether relative band power increases or decreases across a window, motivated by observations of temporal variation in EEG power (Benwell et al., 2019). For each one-second channel patch, we remove its mean and sum squared FFT magnitudes within each of the six bands, using a rectangular window. We divide each band's power by the sum across all six bands, apply a logarithmic transform, and average across channels at each second. For delta, alpha, and beta separately, Kendall's tie-corrected τ between time and the resulting 60-point series defines the trend target (Abbaspoor et al., 2026). Positive values indicate an increasing tendency and negative values a decreasing tendency. Each band produces one scalar target per window, so its displayed R^2 comes from a single probe without further channel averaging.