A CONTROLLED COMPARISON OF SELF-SUPERVISED PRETRAINING OBJECTIVES AND CORPORA FOR INTRACORTICAL NEURAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain-computer interfaces (BCIs) have recently made impressive progress in restoring communication by decoding intended behavior directly from intracortical neural activity. Inspired by the success of foundation models in computer vision and natural language processing, recent work has begun to explore large-scale self-supervised pretraining for neural decoding. However, it is unclear whether these approaches already provide consistent advantages at the scale of data currently available, which is still limited and highly fragmented. In this work we present the first controlled comparison of three self-supervised objective families for intracortical sequence decoding under a unified experimental framework. Using a common transformer backbone, we compare masked autoencoding (MAE), Joint-Embedding Predictive Architectures (JEPA), autoregressive (AR) pretraining, and training from scratch across corpora spanning multiple subjects, behaviors and species: up to 345 hours of intracortical recordings, combining 208 hours from 6 human participants performing attempted and imagined speech, handwriting, typing and cursor control with 137 hours of motor tasks from 19 non-human primates. Our results show that the gains pretraining provides on downstream tasks strongly depend on the amount and diversity of available data, and the resulting performance is often comparable to that of carefully tuned task-specific training. These findings establish a controlled benchmark for future research on neural foundation models.

1 INTRODUCTION

Communication is the channel through which inner states reach the outside world. Amyotrophic lateral sclerosis (ALS), brainstem stroke, spinal cord injury and other neurological conditions affecting the motor pathways that enable communication (e.g. speaking and writing) can have devastating consequences on quality of life. Recent advances in brain-computer interfaces (BCIs) have shown that communication can be partially restored by decoding intended behavior directly from neural activity Silva et al. (2024). Over the past decade the field has entered a new phase, driven by safer intracortical recording technologies, higher-quality neural signals, the availability of public datasets through initiatives such as the BrainGate consortium (BrainGate2), and rapid progress in deep learning. This has produced remarkable results (Willett et al., 2024) across a broad range of decoding tasks, most prominently attempted speech Boccato et al. (2026a), but also imagined handwriting Willett et al. (2021), typing Jude et al. (2026b), and cursor control (Wilson et al., 2025; Singer-Clark et al., 2025), with several systems reaching real-time operation and communication rates that are clinically meaningful for patients (Jude et al., 2026a; Card et al., 2024).

Despite this progress, the field remains in its infancy. Data are scarce: intracortical recordings require neurosurgical implantation, an inherently invasive procedure that restricts the number of available participants and makes data acquisition slow and expensive. Existing datasets are also fragmented, with substantial heterogeneity in recording hardware, implant location, channel count, preprocessing pipelines, session duration, and recorded behaviour. Moreover, intracortical signals are complex, high-dimensional, non-stationary multivariate time series, and there is as yet no consensus on which architectures and training strategies are best suited to them. As a result, most

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.

decoding systems are still trained independently for each subject and each task, which limits generalisation and leaves the field without a shared model that accumulates knowledge across studies.

Similar bottlenecks have been encountered elsewhere in machine learning. Computer vision, natural language processing, and speech recognition all eventually transitioned from task-specific supervised models to large-scale foundation models trained with self-supervised learning (SSL). Rather than learning a single task, such models learn general-purpose representations from large corpora of unlabelled data, which can then be adapted to downstream applications with comparatively little supervision. This naturally suggests that scaling both models and data could benefit neural decoding as well, and the community has indeed begun to explore neural foundation models trained on collections of recordings spanning multiple subjects, tasks, species, and acquisition modalities (Azabou et al., 2023; 2025; Ye & Pandarinath, 2021; Le et al., 2025; Willeke et al., 2026). Because human intracortical data are hard to obtain at scale, these efforts often supplement them with the more numerous non-human primate datasets available, which include generally simpler behavioural tasks.

Whether the benefits observed in other domains transfer to this setting, however, is not obvious. Intracortical recordings are non-stationary, and both the input space and the neural representations themselves vary across sessions and participants; cross-session and cross-subject generalisation remain only partially solved. The field is gradually moving away from the linear decoders traditionally favoured in neuroscience for their interpretability towards deep models, whose expressive power becomes advantageous only when sufficient training data are available. Whether the current data regime is already sufficient for foundation-model behaviour to emerge is not known. The central question of this paper is therefore: *at the scale of data currently available for intracortical BCIs, are neural foundation models an effective strategy?* Current findings remain mixed, reflecting the diverse assumptions across studies.

Ye et al. (2025) pretrain an autoregressive transformer on 2000 hours of spiking activity from over 30 monkeys and humans, predicting neural activity and behaviour jointly in an autoregressive formulation. They conclude that the advantage of pretrained models over from-scratch ones vanishes once roughly 1.5 hours of task-specific data are available, and scaling the corpus from 200 to 2000 hours degrades performance unless model capacity is increased at the same time. On the single sequence-to-sequence task they consider, CTC-based handwriting decoding (Willett et al., 2021; Fan et al., 2023), fine-tuning from pretrained weights degrades rather than improves performance. Zhang et al. (2025) reach a different conclusion with a different method. Pretraining a transformer encoder with purely self-supervised masked modelling on human and monkey recordings, and fine-tuning per human subject, they find that a larger and more diverse corpus consistently improves attempted and imagined speech decoding. These two studies differ simultaneously in pretraining objective, in supervision, in architecture and model scale, in corpus composition, and in downstream task. This motivates the questions: *do we actually need neural foundation models; if so, how should they be trained; and under which data regimes do they pay off?*

Two gaps in particular motivate this work. First, the dominant self-supervised paradigms have never been systematically compared under a common experimental framework, despite resting on markedly different assumptions about the structure of neural activity: masked autoencoders assume that missing observations can be inferred from their surrounding context, autoregressive models exploit the temporal predictability of neural dynamics, and JEPA-style methods learn predictive representations directly in latent space without reconstructing the signal at all. Whether any of these is intrinsically better suited to invasive recordings is unknown. Second, little is understood about how the composition of the pretraining corpus affects downstream performance. Existing studies rely on substantially different data, ranging from single-task human recordings to collections spanning multiple motor behaviors and non-human primate experiments. Larger and more diverse corpora are generally assumed to produce better representations, but it remains unclear how much diversity is actually beneficial, whether knowledge transfers across behaviours and whether non-human primate recordings consistently help on human decoding tasks.

We address both gaps with, to our knowledge, the first systematic study of self-supervised pretraining for invasive neural decoding under a unified experimental setting. We adopt the recently proposed BraIn-to-Text (BIT) encoder (Feghhi et al., 2025) as a common backbone and compare objective-specific pretraining pipelines across corpus compositions using a shared transformer backbone, common optimiser settings, and the same downstream evaluation. We compare masked au-

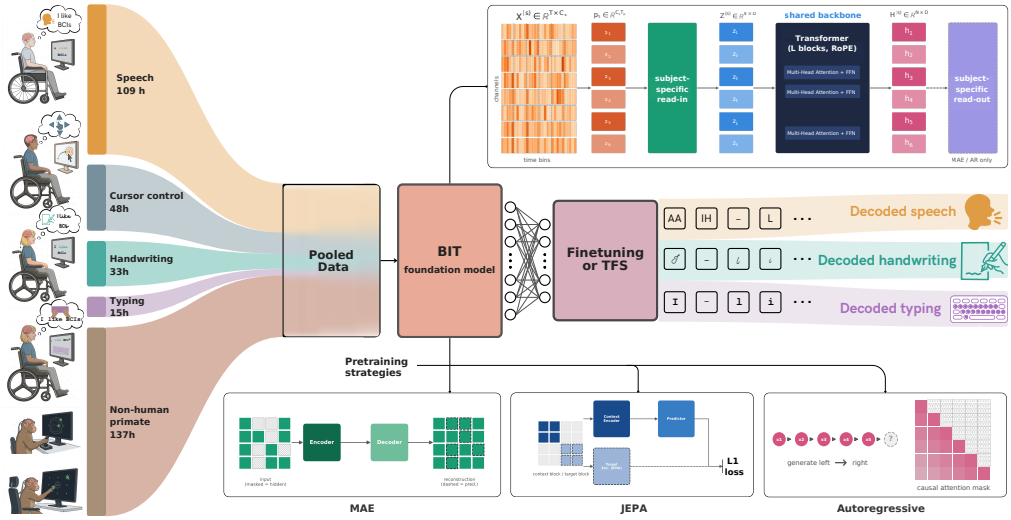

Figure 1: Overview of the experimental framework. The BIT encoder (top) is shared across the pretraining corpus (left) and the three SSL objectives (bottom); after pretraining it is fine-tuned with a CTC head, per subject or jointly, and evaluated on speech, handwriting and typing (right). Training from scratch follows the same fine-tuning protocol from random initialisation.

toencoding, Joint-Embedding Predictive Architectures, autoregressive pretraining, and training from scratch across the largest collection of publicly available human intracortical datasets assembled to date, spanning attempted speech, handwriting, typing, and cursor control, together with non-human primate motor recordings. Figure 1 provides an overview of the data and methods.

Our results challenge the assumption that neural foundation models are already the method of choice for invasive BCIs. The benefits of self-supervised pretraining prove highly dependent on the amount and composition of the available data, and are often more modest than expected. We argue that this reflects two complementary factors: modern neural decoders remain highly sensitive to architectural and optimisation choices, which keeps well-tuned task-specific models competitive whenever sufficient subject-specific data are available; and the scale at which foundation-model behaviour emerges in other domains may still be substantially larger than the amount of intracortical data currently accessible. Our code is publicly available.[1]

2 Methods

2.1 Data and downstream tasks

The human datasets comprise approximately 208 hours of recordings from participants implanted with intracortical Utah arrays, spanning attempted and imagined speech, handwriting, typing, and cursor-control tasks collected in previously published studies. The largest group consists of attempted and imagined speech recordings collected from Card et al. (2024), Willett et al. (2023), Wairagkar et al. (2025), Kunz et al. (2025), and Jude et al. (2026a). Handwriting data comprise recordings from Fan et al. (2023), Willett et al. (2021) and Karpowicz et al. (2024). Typing data comes from Jude et al. (2026b), while cursor-control recordings were collected from Wilson et al. (2025) and Singer-Clark et al. (2025). Neural activity in these datasets is represented using threshold crossing counts and spike band power features, when available; dataset-specific preprocessing details are provided in Appendix 4.

[1]Code: `https://github.com/anonymousrepo456123/neural-ssl`

In addition, approximately 165 hours of 19 non-human primate Utah-array recordings from seven previously published reaching and grasping datasets (Perich et al., 2018; Churchland et al., 2012; O'Doherty et al., 2017; Ma et al., 2023; Temmar et al., 2025; Even-Chen et al., 2019; Chowdhury et al., 2020) were included. Further details on the composition of each dataset and preprocessing are given in Appendices A.1 and A.2, with a visual overview (Figure 4).

2.2 EXPERIMENTAL DESIGN

Our experiments vary two factors while holding everything else fixed. The first is the *composition of the pretraining corpus*, with two levels: HUMAN, comprising the 208 hours of human intracortical recordings described above, and HUMAN+NHP, which adds the 165 hours of non-human primate recordings, for a total of 373 hours. The second is the *pretraining objective*, with four levels: MAE, JEPA, autoregressive prediction, and training from scratch (no pretraining). Training from scratch is by construction independent of the corpus, giving $3_{pretrained} \times 2_{corpora} + 1_{scratch} = 7$ configurations in total. Each configuration is evaluated on attempted and imagined speech, imagined handwriting and imagined typing. We also investigate how the composition of the fine-tuning corpus changes the downstream performance, comparing two alternatives: *joint-task* and *single* fine-tuning.

The from-scratch baseline uses the same encoder architecture and the same hyperparameter search budget as the pretrained models (Section 2.5); the only difference is that the encoder weights are randomly initialised rather than loaded from a pretraining checkpoint.

2.3 BIT ENCODER

We adopt the BraIn-to-Text (BIT) transformer of Feghhi et al. (2025), in the variant of Zhang et al. (2025), as the common backbone for all experiments. The binned neural activity of each subject is partitioned into non-overlapping temporal patches, and each patch is mapped into a shared latent space of dimension D by a subject-specific read-in module, which makes it possible to share the entire transformer backbone across the whole pretraining corpus despite the number of channels varying from subject to subject. The resulting token sequence is processed by a stack of transformer blocks with multi-head attention and rotary positional embeddings (RoPE) (Su et al., 2021), with bidirectional attention for MAE and JEPA and causally masked attention for autoregressive pretraining (Section 2.4.3); apart from this choice of attention mask, the backbone architecture is identical across objectives. Objectives that require reconstruction in signal space project the encoder output back to the original feature space through subject-specific read-out modules, whereas JEPA operates entirely in latent space and employs no reconstruction module. After pretraining, all objective-specific modules are discarded and the encoder is equipped with a task-specific prediction head for downstream decoding. Full architectural details and model hyperparameters are given in AppendixB.1 and B.2.1. Throughout this work the encoder architecture is held fixed; only the self-supervised objective and the composition of the pretraining corpus are varied.

2.4 SSL METHODS

Three self-supervised learning objectives were investigated: MAE (He et al., 2022), JEPA (Assran et al., 2023), and autoregressive modeling Vaswani et al. (2017). These objectives represent three complementary paradigms in modern self-supervised learning: reconstruction-based learning, latent predictive learning, and autoregressive sequence modeling, and capture different assumptions about how meaningful representations can be learned from neural dynamics.

2.4.1 MASKED AUTOENCODER

Following Zhang et al. (2025), the first objective is a masked reconstruction task in the spirit of masked autoencoders.

We sample a binary mask $m \in \{0,1\}^N$ over temporal patches by drawing contiguous spans of random length $\ell \sim \mathcal{U}\{1, \ldots, \ell_{\max}\}$ at random starting positions, until a target masking ratio r is reached. Span-based masking, rather than i.i.d. patch-wise masking, prevents the model from solving the task by trivial local interpolation between neighbouring patches and forces it to exploit longer-range temporal context. The embeddings of masked patches are replaced by a learnable mask

token $e_{\text{mask}} \in \mathbb{R}^D$: $z_n = (1 - m_n)\, f_s(p_n^{(s)}) + m_n\, e_{\text{mask}}, \qquad n = 1, \ldots, N$. The full sequence $Z = (z_1, \ldots, z_N)$ is processed by the encoder with bidirectional attention, and reconstruction is performed by a lightweight linear head rather than by a separate transformer decoder.

The BIT encoder g_θ produces contextual representations $H = g_\theta(Z) \in \mathbb{R}^{N \times D}$, and a subject-specific read-out $h_s : \mathbb{R}^D \to \mathbb{R}^{C_s T_p}$ projects each latent token back to the original feature space. Since both thresholded spike counts and spike-band power are z-scored within each recording day, we optimise a mean squared error on the masked positions,

$$\mathcal{L}_{\text{MAE}} = \frac{1}{C_s T_p \sum_n m_n} \sum_{n=1}^{N} m_n \left\| h_s(H_n) - p_n^{(s)} \right\|_2^2, \tag{1}$$

so that the gradient signal comes exclusively from patches the encoder could not observe. Only the read-in and read-out layers are subject-specific; the mask token, the transformer blocks and the positional encoding are shared across all subjects and species, which is what allows a single encoder to be pretrained on the pooled corpus.

2.4.2 JEPA

We adopt the Joint-Embedding Predictive Architecture (JEPA) of Assran et al. (2023). Unlike reconstruction-based methods, JEPA operates entirely in latent space: the model is trained to predict the representations of masked patches, not the original neural signals. The architecture maintains two copies of the BIT encoder: an *online* (student) encoder g_θ whose parameters are updated by gradient descent, and a *target* (teacher) encoder $g_{\bar{\theta}}$ whose parameters are a exponential moving average (EMA) of the online encoder: $\bar{\theta}_t \leftarrow m_t\, \bar{\theta}_{t-1} + (1 - m_t)\, \theta_t$ where the momentum coefficient m_t follows a cosine schedule from 0.996 to 1.0 over training. The target encoder is kept in evaluation mode and receives no gradient signal.

A random subset of patches is selected as prediction targets and replaced with a learned mask token in the input to the online encoder, which then processes the full sequence (context and masked positions alike) to produce contextualised representations $H^{(s)}_{\text{online}} \in \mathbb{R}^{N \times D}$. The target encoder processes the *unmasked* sequence and produces $H^{(s)}_{\text{target}} \in \mathbb{R}^{N \times D}$, which are layer-normalised.

A lightweight transformer p_ϕ, shallower than the main encoder (6 vs. 7 layers) but with the same hidden dimension, takes $H^{(s)}_{\text{online}}$ as input and maps it to a predicted target representation $\hat{H}^{(s)} = p_\phi(H^{(s)}_{\text{online}})$. At masked positions, the predictor input is replaced with a dedicated *predictor mask token* (separate from the encoder mask token) so that positional information is preserved while the identity of the missing patch is not leaked. The loss is a smooth-ℓ_1 loss computed only at masked, non-padded positions:

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{|\mathcal{M}|} \sum_{n \in \mathcal{M}} \text{SmoothL1}\Big(\hat{H}^{(s)}_n,\ \text{LN}(H^{(s)}_{\text{target},n}) \Big), \tag{2}$$

where $\mathcal{M}$ denotes the set of masked patch indices. No signal reconstruction is performed. After pretraining, the target (EMA) encoder weights are used to initialise the downstream decoder, following Assran et al. (2023).

2.4.3 Autoregressive Prediction

As third pretraining objective we adopt a next-patch prediction scheme analogous to GPT (Radford et al., 2018), applied to the patched neural token sequence. The autoregressive objective trains the encoder to predict future neural activity from past observations. Given a sequence of temporal patches, the transformer operates causally, ensuring that each patch attends only to previous time steps. The latent representation of each patch is projected through a subject-specific reconstruction head to predict the neural activity of subsequent patches.

The only structural change from the MAE encoder is the attention mask. Rather than bidirectional self-attention, each patch token is restricted to attend only to itself and all preceding patches:

$$H^{(s)} = g_\theta^{\text{AR}}\big(Z^{(s)}\big), \qquad \text{where } \big[g_\theta^{\text{AR}}\big]_{nt} = 0 \ \text{ if } \ t > n. \tag{3}$$

To avoid leaking future information at the input level, the temporal smoothing applied during data augmentation is also made causal when this mode is active. No masking is applied to the input sequence: with causal attention, position n already has access only to patches $1, \ldots, n$, so additional masking would merely remove signal the model is meant to condition on. The encoder output at patch n is projected back to signal space by the subject-specific read-out and trained to predict patch $n+1$:

$$\mathcal{L}_{\mathrm{AR}} = \frac{1}{C_s T_p |\mathcal{V}|} \sum_{(n,n+1)\in\mathcal{V}} \left\| h_s\big(H_n^{(s)}\big) - p_{n+1}^{(s)} \right\|_2^2, \tag{4}$$

where $\mathcal{V}$ is the set of valid consecutive-patch pairs (i.e. pairs of patches which are both non-padded).

Training Details: All objectives are pretrained with identical optimisation settings, so that downstream differences are not confounded by unequal tuning budgets (full hyperparameters in Appendix B.2.1). The pretraining checkpoint transferred to downstream fine-tuning is selected on a held-out validation split of the pretraining corpus (see Appendix A.3). For MAE and autoregressive pretraining we select the epoch maximising the R^2 between predicted and ground-truth neural features. This criterion is not applicable to JEPA, whose predictions live in a latent space; we select the epoch minimising the validation loss. Alongside the reconstruction metrics, we monitor RankMe (Garrido et al., 2022), an unsupervised estimate for the effective rank of the learned representations, suited for Joint Embedding Self Supervised Learning. A fully collapsed representation, with all variance along a single direction, gives 1, while a spectrum spread uniformly over R directions gives exactly R. Because it requires neither labels nor reconstruction in signal space, RankMe applies to all three SSL objectives, and in particular provides a collapse diagnostic for JEPA, whose training loss can decrease while the representations degenerate. We use this measure to detect failed pretraining runs (see Appendix A.2 and E).

2.5 FINE-TUNING

After pretraining, the encoder is equipped with a linear projection layer mapping token embeddings to logits over the decoding vocabulary. This consists of 41 tokens for speech (39 English phonemes, silence, and the CTC blank), 32 tokens for handwriting and 31 tokens for typing. The encoder and the linear head are jointly optimised on the downstream decoding task with the CTC loss (Graves et al., 2006). During fine-tuning, all models are equipped with bidirectional attention.

Single vs Joint Fine-Tuning: We perform the fine-tuning in two settings: *single* and *joint*. Single fine-tuning is the standard procedure in which the weights of the encoder and subject-specific head are optimised for the single subject. On the other hand, joint fine-tuning is performed according to the task performed by the subjects. We discern three tasks: speech, handwriting and typing; therefore we run three joint fine-tunings for each pretrained configuration. The training data for the joint fine-tuning consists of the union of all neural recordings from all subjects performing the given task. For example, for the joint fine-tuning over speech we compose a training set that is the union of recordings from Wairagkar, Card, Willett, Kunz, and Jude Speech Anarthria. During joint fine-tuning, the validation step is done separately on each dataset under examination, and the best performing epoch for each subject is selected for the final evaluation; model weights are updated according to the overall performance on the selected task. Single and joint fine-tuning are always implemented as described, with one exception. Given the small size of the training corpus of Kunz subjects T12 and T15, and that these are the same subjects respectively of Willett T12 and Card T15, implanted in the same area and acquired in a short period of time, we always perform a *pseudo-joint* fine-tuning for subjects T12 and T15: all single fine-tunings on these two subjects include in their training corpus also the data coming from the twin dataset. This differs from the pure joint fine-tuning because the latter includes also neural recordings from other datasets of speech, i.e. Jude Speech Anarthria and Wairagkar.

Evaluation: Decoding accuracy is measured by phoneme error rate (PER) for speech and character error rate (CER) for handwriting and typing. Predicted sequences are obtained by greedy CTC decoding, collapsing repeated emissions and removing blank tokens. The number of errors is then computed as the Levenshtein edit distance (Levenshtein, 1966) between the predicted and ground-truth sequences, counting substitutions, insertions, and deletions, and is normalised by the length of the ground-truth sequence. Reporting PER/CER, rather than word error rate, isolates the contribu-

tion of the neural encoder from that of any downstream language model, which is what we want to investigate when comparing pretraining objectives.

Optimisation: Encoder and linear head are optimised jointly with AdamW for 800 epochs. No layers are frozen: the full encoder is updated during fine-tuning. For each configuration we train 30 independent models with hyperparameters drawn at random from fixed ranges and report the run achieving the best validation performance. In the single fine-tunings regime this was done in practice for $6 + 1 = 7$ variants (pretrained and from scratch), and 8 unique subjects (T15 from Card and Kunz were always aggregated, as well as T12 from Willett and Kunz), giving $X = 7 \times 8 \times 30$ trainings. In the joint fine-tuning case we aggregated the subjects under three tasks (speech, handwriting, typing): this resulted in $Y = 7 \times 3 \times 30$ trainings. The total number of fine-tuning runs was $X + Y = 2310$. Search ranges and sampling details are given in Appendix C.

3 Results

Figure 2 reports downstream decoding performances for all pretraining configurations, including different data corpora and single as well as joint fine-tuning strategies, with the best-performing configuration highlighted for each subject. Full numerical results are given in Appendix G.

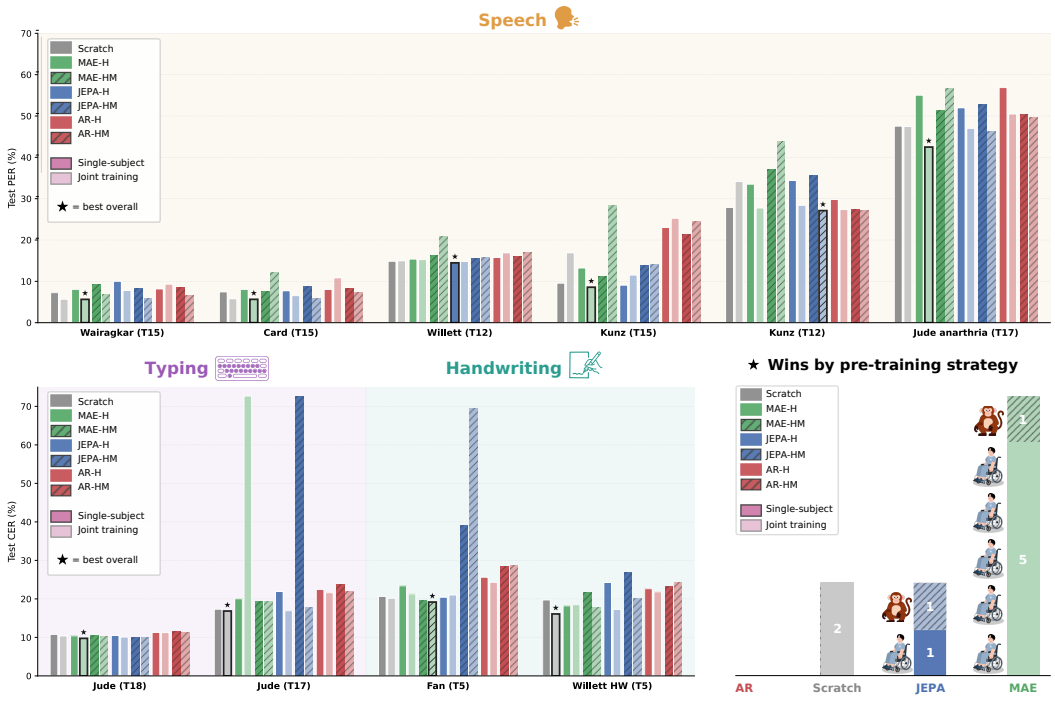

Figure 2: Downstream decoding performances on all subjects. Bars report test PER for speech (top) and CER for typing and handwriting (bottom left); lower is better. Each subject is shown with all 14 configurations. Hatching indicates the HUMAN+NHP corpus, solid bars the HUMAN corpus. Best result for each subject marked by a star. Bottom right: number of subjects for which each pretraining objective provides the best result, split by corpus composition.

Advantages of Pretraining and NHP Data: Across all 10 subjects, the best configuration is obtained from a pretrained encoder in 8 cases and from the from-scratch baseline in 2. Among the pretrained winners, 6 come from MAE and 2 from JEPA, with the HUMAN+NHP corpus accounting for only 2 of them. No configuration wins consistently across subjects or tasks, and the margins are small: on the two largest speech datasets the best pretrained model improves over the from-scratch baseline by 0.03 and 0.14 absolute percentage points of PER.

Only 2 of the 8 pretrained wins use the HUMAN+NHP corpus. One of these is on handwriting (Fan et al., 2023), the human task whose motor demands are closest to the reaching and grasping

behaviours that make up the primate corpus, which would be consistent with transfer being restricted to shared motor structure. We note, however, that the second HUMAN+NHP win is on imagined speech (Kunz et al., 2025), and that the remaining handwriting subject is best decoded by the from-scratch baseline, so the pattern is not clearly consistent with the idea of a shared motor structure between human and non-human primates. What we find is a weaker benefit than the one reported by Zhang et al. (2025) when incorporating NHP recordings in their corpus. However, their NHP corpus consists of 269 hours of recordings, against our 137, while their HUMAN corpus is roughly half of ours (98 vs 208 hours), making NHP a much larger fraction of their HUMAN+NHP corpus ($\sim 70\%$ vs $\sim 40\%$). Their scaling analysis reports that including NHP data gives smaller improvements than human data, with a real gain being detectable only in the upper part of the range they explore, when the NHP corpus surpasses the magnitude of the HUMAN corpus. Our results suggest that adding non-human primate recordings at this scale does not bring notable advantages. Taken together with Zhang et al. (2025), they are in line with the claim that consistent advantages from NHP data may be detectable only when the NHP corpus surpasses the Human one.

Joint fine-tuning: Independently of the pretraining objective, pooling subjects that perform the same task is markedly effective: 9 of the 10 best per-subject results in Figure 2 are obtained under the joint regime, the sole exception being Willett T12 on attempted speech.

Variability across the search and statistic robustness of differences: The spread of outcomes within a single sweep is large, and in a few cells the best configuration found within the 30-run budget is still unsatisfactory (e.g. 72.70% CER for MAE-H on typing subject T17 under joint fine-tuning). We report these results as obtained: every configuration receives the same fixed search budget, and the failure of a given initialisation to yield a usable model within that budget is part of what we are measuring. Given the high variability induced by the sampled hyperparameter configurations, we investigate whether pre-training followed by fine-tuning results in significantly different error rates compared to training from scratch. We carry out the following stability analysis. For each combination of pretraining strategy, subject, and finetuning paradigm (single vs joint), we select the best configuration found by the hyperparameter sweep, and re-train it under 5 fixed seeds, to obtain a distribution of outcomes at each combination's best setting. First, we observe high variability across finetuning outcomes, even in this 'best-scenario' setting, especially for some subjects (Kunz T15 and Jude Speech Anarthria for speech, all subjects for typing and handwriting). A visualization of the stability runs for all configurations is in Appendix F. In particular, we observe that a non-negligible fraction of seeds, rather than converging to a useful solution, settle at error rates substantially higher than the rest of the seed distribution for the same configuration. These "collapsed" runs appear across multiple model-paradigm combinations, more frequently for typing and handwriting subjects. To make the comparison as fair as possible we exclude collapsed runs as follows. First, Silverman's unimodality test (Silverman, 1981) ($\alpha = 10\%$) is performed on all 5-seed groups, to identify which groups do not follow a unimodal distribution (i.e., what groups actually contain collapsed runs). When unimodality is rejected, we partition the seeds into two clusters via optimal one-dimensional k-means and discard the upper cluster. For each subject we then apply a two-sided Mann-Whitney U test comparing the stability seeds of every pre-trained initialisation against the corresponding scratch baseline (Figure 3). Importantly, no pretrained configuration shows an uncorrected $p < 0.05$ improvement over scratch. In contrast, in several cases the pretrained models obtain significantly worse performances with respect to the baseline, an effect more concentrated in speech decoding (e.g. Wairagkar, Card, Kunz T15, Willett T12, Jude anarthria). Despite the absence of statistically significant differences, Figure 3 also reveals a complementary pattern: joint training from scratch outperforms single-subject training from scratch across several subjects (Wairagkar, Card, Fan Handwriting, Willett Handwriting, Jude Typing T17), consistent with the broader observation that pooling together same-task subjects may be beneficial.

4 DISCUSSION AND CONCLUSIONS

We compared three of the most widely used self-supervised pretraining objectives for intracortical neural decoding on a common transformer backbone while varying the composition of the pre-training corpus and the downstream fine-tuning strategy, and comparing them against trained-from-scratch baselines across multiple subjects and tasks. Our main finding is that, at this scale and for the investigated objectives, self-supervised pretraining does not provide a consistent advantage over carefully tuned training from scratch. This is in contrast with the scaling paradigm that has brought

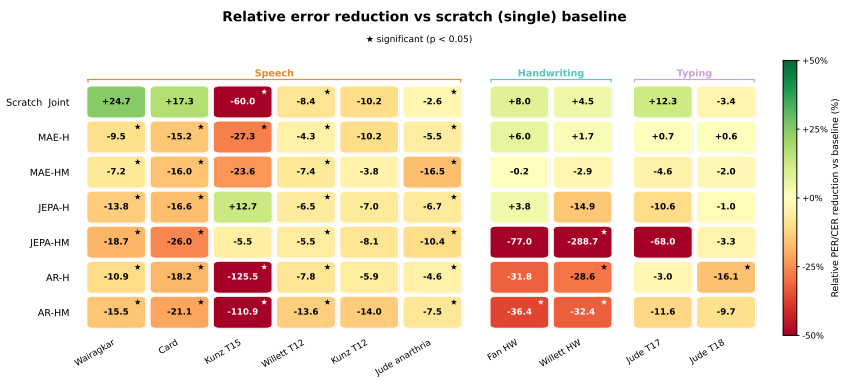

Figure 3: Relative PER/CER reduction of all configurations over scratch (single) baseline, over 5 stability seeds. Collapsed runs (Silverman's bimodality test, $\alpha = 10\%$, mean-ratio guard ≥ 1.25) are excluded. Cells show the median percentage reduction in PER/CER compared to the scratch single-subject baseline (positive = lower error = better). Cells are green for improvements and red for degradations. Stars indicate statistically significant differences (two-sided Mann-Whitney U test, uncorrected $p < 0.05$). Top row compares joint training from scratch vs the single-subject baseline.

progress in other fields of AI, such as NLP, automatic speech recognition and computer vision, where increasing the amount and diversity of training data has enabled more capable general-purpose models, although on corpora far larger than anything available here: Whisper (Radford et al., 2022) was pretrained on hundreds of thousands of hours of audio, against our 345.

In neural decoding, however, the benefits of scaling may depend not only on the amount of data, but also on its relevance to the downstream task. Differently from images, audio or language, intracortical recordings are highly heterogeneous across subjects, tasks, species, and recording configurations. Simply aggregating the heterogeneous recordings available to date may not provide the same benefits as scaling data within a more homogeneous domain. A result in this direction is the success obtained by joint fine-tuning across subjects performing the same task, in line with the results of Boccato et al. (2026b). This suggests that effectively exploiting labelled data across subjects may be at least as important as learning general-purpose representations through unsupervised pretraining. The main challenge for neural foundation models may be in understanding how to productively combine data from different subjects, tasks, and recording settings. Furthermore, our results suggest that learning a useful model of neural dynamics is not necessarily equivalent to learning a representation useful for decoding behavior. This highlights the importance of designing pretraining objectives that capture properties of neural activity relevant to the intended downstream task, rather than relying only on the modeling of neural dynamics to produce useful representations.

In conclusion, we argue that the benefits of neural foundation models may depend on reaching a sufficient scale of data, but particular care should be taken when designing appropriate pretraining objectives and handling neural heterogeneity.

This study has several limitations. First, although the pretraining corpus is large relative to existing intracortical datasets, it remains heterogeneous and limited compared with the datasets used to train foundation models in other domains. We therefore cannot determine whether the limited benefits of pretraining reflect an intrinsic limitation of self-supervised learning for neural data or simply the current scale and composition of the available corpora. Second, our comparison is performed using a single transformer backbone and a fixed training architecture across all pretraining objectives. This allows us to isolate the effect of the pretraining objective and corpus composition, but limits the generality of our conclusions to other architectures which may interact differently with the various pretraining objectives. Third, the finite hyperparameter search budget does not guarantee that each objective was evaluated at its optimal training configuration. A broader exploration which includes hyperparameter tuning in pretraining and explores more fine-tuning configurations could give different results. Finally, whether pretraining confers an advantage in the low-label regime remains only indirectly addressed here.

AI USE STATEMENT

In this work, we have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

The rest of the required disclosure tasks (formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs) are not applicable to this work.

Additionally, we used generative AI tools to create or edit parts of the software code, edit specific sections or sentences of the research paper to improve readability and improve the quality of some scientific figures. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

In this work we used publicly available datasets which contain human recordings. A full list of the datasets is in the Appendix A.1, including direct links to the official sources. We do not acquire any additional data, from humans or animals.

Neural recordings of course represent sensitive forms of personal data, and decoding models raise legitimate concerns about privacy. The properties of the present setting limit those concerns. First, the recordings require surgically implanted intracortical arrays, so acquisition cannot occur without the participant's deliberate enrolment in a clinical study. Second, decoding remains strongly subject-specific: models rely on subject-specific input layers and on labelled data from the participant themselves. No model studied here could be applied to an unconsenting individual.

REPRODUCIBILITY STATEMENT

We have made our best effort to make this work reproducible. To this end, we provide in the Appendix: links to the official data repositories for every included dataset; a full description of all preprocessing steps that were carried out; a full description of how the validation and test splits are obtained; the hyperparameters of all the considered models, as well as those defining the sweep-space for fine-tuning; the values we obtained during pretraining of useful diagnostic measures, for reference; extended results on both validation and test set. Furthermore, we provide a link to the full code related to this work.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding, 2023. URL https://arxiv.org/abs/2310.16046.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IuU0wcO0mo.

Tommaso Boccato, Michal Olak, and Matteo Ferrante. BrainWhisperer: Leveraging large-scale ASR models for neural speech decoding. March 2026a. URL https://arxiv.org/abs/2603.13321.

Tommaso Boccato, Michal Olak, and Matteo Ferrante. Cross-subject decoding of human neural data for speech brain computer interfaces. *Journal of Neural Engineering*, 23(4):046018, 2026b. ISSN 1741-2552. doi: 10.1088/1741-2552/ae8576. URL http://dx.doi.org/10.1088/1741-2552/ae8576.

BrainGate2. Braingate2: Feasibility study of an intracortical neural interface system for persons with tetraplegia (braingate2). URL https://clinicaltrials.gov/study/NCT00912041. NCT00912041.

Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, August 2024. ISSN 1533-4406. doi: 10.1056/nejmoa2314132. URL http://dx.doi.org/10.1056/NEJMoa2314132.

Raeed H Chowdhury, Joshua I Glaser, and Lee E Miller. Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. *eLife*, 9, January 2020. ISSN 2050-084X. doi: 10.7554/elife.48198. URL http://dx.doi.org/10.7554/eLife.48198.

Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012. ISSN 1476-4687. doi: 10.1038/nature11129. URL http://dx.doi.org/10.1038/nature11129.

Nir Even-Chen, Blue Sheffer, Saurabh Vyas, Stephen I. Ryu, and Krishna V. Shenoy. Structure and variability of delay activity in premotor cortex. *PLOS Computational Biology*, 15(2):e1006808, February 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006808. URL http://dx.doi.org/10.1371/journal.pcbi.1006808.

Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy Wilson, Leigh Hochberg, Krishna V. Shenoy, Jaimie Henderson, and Francis Willett. Plug-and-play stability for intracortical brain-computer interfaces: A one-year demonstration of seamless brain-to-text communication. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42258–42270, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023.

Ebrahim Feghhi, Shreyas Kaasyap, Nima Hadidi, and Jonathan C. Kao. Time-masked transformers with lightweight test-time adaptation for neural speech decoding, 2025. URL https://arxiv.org/abs/2507.02800.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. 2022. doi: 10.48550/ARXIV.2210.02885. URL https://arxiv.org/abs/2210.02885.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, ICML '06, pp. 369–376. ACM Press, 2006. doi: 10.1145/1143844.1143891. URL http://dx.doi.org/10.1145/1143844.1143891.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01553. URL http://dx.doi.org/10.1109/CVPR52688.2022.01553.

Justin J. Jude, Stephanie Haro, Hadar Levi-Aharoni, Hiroaki Hashimoto, Alexander J. Acosta, Nicholas S. Card, Maitreyee Wairagkar, David M. Brandman, Sergey D. Stavisky, Ziv M. Williams, Sydney S. Cash, John D. Simeral, Leigh R. Hochberg, and Daniel B. Rubin. Decoding intended speech with an intracortical brain-computer interface in a person with long-standing anarthria and locked-in syndrome. *Cell Reports*, 45(4):117162, April 2026a. ISSN 2211-1247. doi: 10.1016/j.celrep.2026.117162. URL http://dx.doi.org/10.1016/j.celrep.2026.117162.

Justin J. Jude, Hadar Levi-Aharoni, Alexander J. Acosta, Shane B. Allcroft, Claire Nicolas, Bayardo E. Lacayo, Nicholas S. Card, Maitreyee Wairagkar, Alisa D. Levin, David M. Brandman, Sergey D. Stavisky, Francis R. Willett, Ziv M. Williams, John D. Simeral, Leigh R. Hochberg, and

Daniel B. Rubin. Restoring rapid natural bimanual typing with a neuroprosthesis after paralysis. *Nature Neuroscience*, March 2026b. ISSN 1546-1726. doi: 10.1038/s41593-026-02218-y. URL http://dx.doi.org/10.1038/s41593-026-02218-y.

Brianna M. Karpowicz, Joel Ye, Chaofei Fan, Pablo Tostado-Marcos, Fabio Rizzoglio, Clay Washington, Thiago Scodeler, Diogo de Lucena, Samuel R. Nason-Tomaszewski, Matthew J. Mender, Xuan Ma, Ezequiel Matias Arneodo, Leigh R. Hochberg, Cynthia A. Chestek, Jaimie M. Henderson, Timothy Q. Gentner, Vikash Gilja, Lee E. Miller, Adam G. Rouse, Robert A. Gaunt, Jennifer L. Collinger, and Chethan Pandarinath. Few-shot algorithms for consistent neural decoding (falcon) benchmark. 2024. doi: 10.1101/2024.09.15.613126. URL http://dx.doi.org/10.1101/2024.09.15.613126.

Erin M. Kunz, Benyamin Abramovich Krasa, Foram Kamdar, Donald T. Avansino, Nick Hahn, Seonghyun Yoon, Akansha Singh, Samuel R. Nason-Tomaszewski, Nicholas S. Card, Justin J. Jude, Brandon G. Jacques, Payton H. Bechefsky, Carrina Iacobacci, Leigh R. Hochberg, Daniel B. Rubin, Ziv M. Williams, David M. Brandman, Sergey D. Stavisky, Nicholas AuYong, Chethan Pandarinath, Shaul Druckmann, Jaimie M. Henderson, and Francis R. Willett. Inner speech in motor cortex and implications for speech neuroprostheses. *Cell*, 188(17):4658–4673.e17, August 2025. ISSN 0092-8674. doi: 10.1016/j.cell.2025.06.015. URL http://dx.doi.org/10.1016/j.cell.2025.06.015.

Trung Le, Hao Fang, Jingyuan Li, Tung Nguyen, Lu Mi, Amy Orsborn, Uygar Sümbül, and Eli Shlizerman. Spint: Spatial permutation-invariant neural transformer for consistent intracortical motor decoding, 2025. URL https://arxiv.org/abs/2507.08402.

Vladimir I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8), 1966. URL https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training, 2018. URL https://arxiv.org/abs/1807.05118.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

Xuan Ma, Fabio Rizzoglio, Stephanie Thacker, Christian Ethier, Matthew Perich, and Lee Miller. Data from: Using adversarial networks to extend brain computer interface decoding accuracy over time, 2023. URL https://datadryad.org/dataset/doi:10.5061/dryad.cvdncjt7n.

Joseph E. O'Doherty, Mariana M. B. Cardoso, Joseph G. Makin, and Philip N. Sabes. Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology, 2017. URL https://zenodo.org/record/583331.

Matthew G. Perich, Juan A. Gallego, and Lee E. Miller. A neural population mechanism for rapid learning. *Neuron*, 100(4):964–976.e7, November 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.09.030. URL http://dx.doi.org/10.1016/j.neuron.2018.09.030.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Alexander B. Silva, Kaylo T. Littlejohn, Jessie R. Liu, David A. Moses, and Edward F. Chang. The speech neuroprosthesis. *Nature Reviews Neuroscience*, 25(7):473–492, May 2024. ISSN 1471-0048. doi: 10.1038/s41583-024-00819-9. URL http://dx.doi.org/10.1038/s41583-024-00819-9.

B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 43(1):97–99, 1981. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1981.tb01155.x. URL http://dx.doi.org/10.1111/j.2517-6161.1981.tb01155.x.

Tyler Singer-Clark, Xianda Hou, Nicholas S Card, Maitreyee Wairagkar, Carrina Iacobacci, Hamza Peracha, Leigh R Hochberg, Sergey D Stavisky, and David M Brandman. Speech motor cortex enables bci cursor control and click. *Journal of Neural Engineering*, 22(3):036015, May 2025. ISSN 1741-2552. doi: 10.1088/1741-2552/add0e5. URL http://dx.doi.org/10.1088/1741-2552/add0e5.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864.

Hisham Temmar, Yixuan Wang, Nina Gill, Nicholas Mellon, Chang Liu, Luis Hernan Cubillos, Rio I. Parsons, Joseph T. Costello, Matteo Ceradini, Madison M. Kelberman, and et al. Long-term intracortical neural activity and kinematics (link): An intracortical neural dataset for chronic brain–machine interfaces, neuroscience, and machine learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC13322217/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. June 2017. URL https://arxiv.org/abs/1706.03762.

Maitreyee Wairagkar, Nicholas S. Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M. Miller, Leigh R. Hochberg, David M. Brandman, and Sergey D. Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, 644(8075):145–152, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09127-3. URL http://dx.doi.org/10.1038/s41586-025-09127-3.

Konstantin F. Willeke, Polina Turishcheva, Alex Gilbert, Goirik Chakrabarty, Hasan A. Bedel, Paul G. Fahey, Yongrong Qiu, Marissa A. Weis, Michaela Vystrčilová, Taliah Muhammad, Lydia Ntanavara, Rachel E. Froebe, Kayla Ponder, Zheng Huan Tan, Emin Orhan, Erick Cobos, Sophia Sanborn, Katrin Franke, Fabian H. Sinz, Alexander S. Ecker, and Andreas S. Tolias. Omnimouse: Scaling properties of multi-modal, multi-task brain models on 150b neural tokens, 2026. URL https://arxiv.org/abs/2604.18827.

Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03506-2. URL http://dx.doi.org/10.1038/s41586-021-03506-2.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06377-x. URL http://dx.doi.org/10.1038/s41586-023-06377-x.

Francis R Willett, Jingyuan Li, Trung Le, Chaofei Fan, Mingfei Chen, Eli Shlizerman, Yue Chen, Xin Zheng, Tatsuo S Okubo, Tyler Benster, Hyun Dong Lee, Maxwell Kounga, E Kelly Buchanan, David Zoltowski, Scott W Linderman, and Jaimie M Henderson. Brain-to-Text benchmark '24: Lessons learned. December 2024. URL https://arxiv.org/abs/2412.17227.

Guy H. Wilson, Elias A. Stein, Foram Kamdar, Donald T. Avansino, Tsam Kiu Pun, Ronnie Gross, Tommy Hosman, Tyler Singer-Clark, Anastasia Kapitonava, Leigh R. Hochberg, John D. Simeral, Krishna V. Shenoy, Shaul Druckmann, Jaimie M. Henderson, and Francis R. Willett. Long-term unsupervised recalibration of cursor-based intracortical brain–computer interfaces using a hidden markov model. *Nature Biomedical Engineering*, 10(7):1466–1484, December 2025. ISSN 2157-846X. doi: 10.1038/s41551-025-01536-z. URL http://dx.doi.org/10.1038/s41551-025-01536-z.

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with neural data transformers. 2021. doi: 10.48550/ARXIV.2108.01210. URL https://arxiv.org/abs/2108.01210.

Joel Ye, Fabio Rizzoglio, Adam Smoulder, Hongwei Mao, Xuan Ma, Patrick Marino, Raeed H Chowdhury, Dalton D Moore, Gary Blumenthal, Will Hockeimer, Nicolas G. Kunigk, J Patrick Mayo, Aaron P Batista, Steven M Chase, Adam G Rouse, Michael L. Boninger, Charles Greenspon, Andrew B. Schwartz, Nicholas Hatsopoulos, Lee E Miller, Kristofer Bouchard, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. A generalist intracortical motor decoder. February 2025. doi: 10.1101/2025.02.02.634313. URL http://dx.doi.org/10.1101/2025.02.02.634313.

Yizi Zhang, Linyang He, Chaofei Fan, Tingkai Liu, Han Yu, Trung Le, Jingyuan Li, Scott Linderman, Lea Duncker, Francis R Willett, Nima Mesgarani, and Liam Paninski. A cross-species neural foundation model for end-to-end speech decoding, 2025. URL https://arxiv.org/abs/2511.21740.

A DATA

A.1 DATASETS

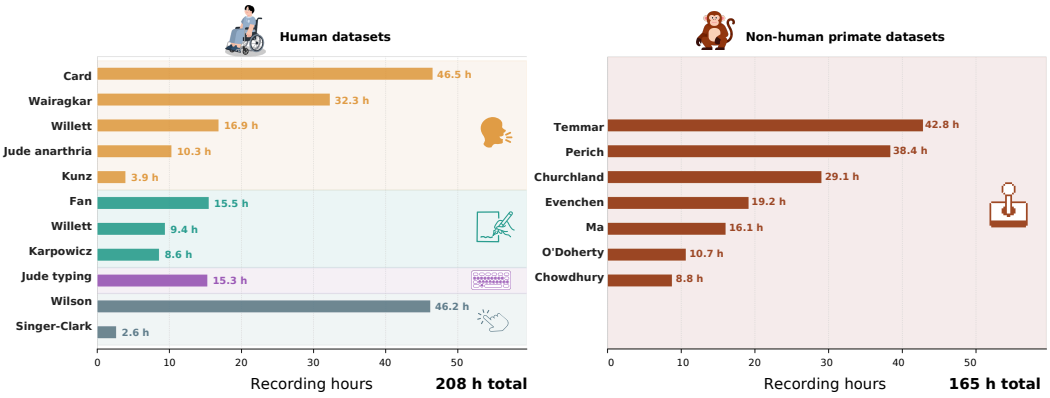

Figure 4: Left: composition of the HUMAN corpus sorted by available recoding hours and tasks performed by the participants (attempted or imagined speech, imagined handwriting, imagined typing, cursor control). Right: composition of the NHP corpus sorted by recording hours. NHP subjects performed various motor tasks, e.g. reaching and grasping. After preprocessing, the Perich dataset comes down to $\sim$ 10 hours (Appendix A.2), bringing the total NHP hours to $\sim$ 137.

Table 1: Human datasets. Tasks: AS = attempted speech, IS = imagined speech, HW = handwriting, TY = typing, CC = cursor control. *Elec.* is the number of recording channels and *Feat.* the signal types extracted from each (TX: threshold crossings; SBP: spike-band power), so that the input dimensionality per 20 ms bin is their product. The last column marks datasets also used as downstream decoding tasks.

Dataset	Subj.	Task	Elec.	Feat.	Hours	Downstream
Willett et al. (2023)	T12	AS	128	TX+SBP	16.9	✓
Card et al. (2024)	T15	AS	256	TX+SBP	46.5	✓
Wairagkar et al. (2025)	T15	AS	256	TX+SBP	32.3	✓
Jude et al. (2026a)	T17	AS	128	TX+SBP	10.3	✓
Kunz et al. (2025)	T12, T15	IS	128, 256	TX+SBP	3.4	✓
Kunz et al. (2025)	T16, T17	IS	64, 128	TX+SBP	0.5	
Willett et al. (2021)	T5	HW	192	TX	9.4	✓
Fan et al. (2023)	T5	HW	192	TX	15.5	✓
Karpowicz et al. (2024)	T5	HW	192	TX	8.6	
Jude et al. (2026b)	T17	TY	384	TX+SBP	7.5	✓
Jude et al. (2026b)	T18	TY	384	TX+SBP	7.8	✓
Wilson et al. (2025)	T5	CC	192	TX	46.2	
Singer-Clark et al. (2025)	T15	CC	256	TX+SBP	2.7	
Total					**207.6**	

A.2 PREPROCESSING

Neural data was binned into 20 ms time bins where not already provided at that resolution. For each dataset, spike counts were z-scored per channel; normalization statistics were computed across all available sessions of the same subject where possible, and within individual sessions otherwise.

Two datasets required specific handling to avoid pretraining failure. The documented failures were consistent across three indicators: the validation R^2 for neural reconstruction settled at approximately -0.1 within a few epochs instead of increasing steadily; the training loss stopped decreasing early; and RankMe (Garrido et al., 2022), an unsupervised measure of the effective rank of the learned representations, collapsed and/or oscillated around low values rather than growing over training.

Table 2: Non-human primate datasets. Tasks: RE = reaching, RG = reach/grasp, FM = finger movement. *# Subj.* counts biological subjects; the number of subject-specific read-in modules is larger, since a separate module is instantiated whenever channel identity is not comparable across recordings. All datasets provide threshold crossings only. None is used as a downstream decoding task.

Dataset	# Subj.	Task	Elec.	Feat.	Hours
Temmar et al. (2025)	1	FM	96	TX	42.8
Perich et al. (2018)	4	RE		TX	38.4
Churchland et al. (2012)	2	RE	192	TX	29.1
Even-Chen et al. (2019)	2	RE	192	TX	19.2
Ma et al. (2023)	5	RG	73–96	TX	16.1
O'Doherty et al. (2017)	2	RE	96, 192	TX	10.7
Chowdhury et al. (2020)	3	RE	42–167	TX	8.8
Total					**165.0**

Details about the Wilson dataset The cursor-control recordings of Wilson et al. (2025) in the training set contribute with more than $50,000$ trials, an order of magnitude more than any other human dataset. Including this dataset in full destabilised pretraining, with the shared backbone effectively dominated by a single subject and behaviour. We therefore cap the number of trials contributed by each subject to $10{,}000$ per epoch. The retained subset is resampled it at every epoch, so that over the course of training the model is exposed to all available trials. The cap is applied uniformly to all subjects, although in practice only Wilson et al. (2025) exceeds it.

Details about the Perich dataset Throughout this work, a *subject* is operationally defined as a unique (animal, channel-count) pair rather than as a biological individual: whenever recordings from the same animal are available with a different number of channels, we treat them as a distinct subject and instantiate a separate read-in module. Perich et al. (2018) requires special treatment. Unlike the other datasets, it is distributed after spike sorting, a procedure in which the extracellular waveforms recorded on each electrode are clustered into putative single neurons, so that the data dimensions correspond to isolated units rather than to physical recording channels. Because the set of units that can be isolated depends on the recording conditions of each session, the number of dimensions varies substantially across sessions and unit identity is not comparable across sessions even within the same animal. Under our definition above, every session constitutes a separate subject and requires its own read-in module. In practice we found that including the dataset in full would inflate the parameter count making this configuration untrainable: pretraining on the full corpus including all Perich sessions consistently failed to converge, which we attribute to the large number of read-in modules. To mitigate this, we retain only the 10 longest sessions of Perich et al. (2018) and exclude the remainder from the pretraining corpus. After this step, the total number of hours retained is 10.

A.3 Data splits

Downstream datasets are split into training, validation, and test sets. All datasets are distributed with pre-existing splits. For datasets which have no predefined validation or test split, or in cases where the test split does not include labels, we randomly split the validation set in half to create a validation and a test split both with labels. Model selection during fine-tuning uses the validation split, and we report metrics on the validation and the held-out test split.

B Models

B.1 BIT

Let $X^{(s)} \in \mathbb{R}^{T \times C_s}$ denote the binned neural activity of a recording from subject s, where T is the number of $20\,\text{ms}$ bins and C_s the number of recording channels, which varies across subjects.

Temporal patching The sequence is partitioned into $N = \lceil T/T_p \rceil$ non-overlapping temporal patches of T_p consecutive bins, so that the n-th patch $p_n^{(s)} \in \mathbb{R}^{C_s T_p}$ is obtained. Patching serves two purposes: it aligns the fine temporal resolution of intracortical recordings with the slower timescale of the behaviour being decoded, and it shortens the context length seen by the transformer.

Read-in Each patch is mapped to the shared latent dimension D by a subject-specific patch embedding module $f_s : \mathbb{R}^{C_s T_p} \to \mathbb{R}^D$, implemented as (LayerNorm, Linear, LayerNorm). These read-in layers make it possible for all transformer parameters to be shared across the entire pretraining corpus. Applying f_s independently to each patch yields the token sequence

$$Z^{(s)} = \left[f_s(p_1^{(s)});\ \ldots;\ f_s(p_N^{(s)}) \right] \in \mathbb{R}^{N \times D}, \tag{5}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes row-wise stacking.

Encoder The sequence is then processed by a stack of transformer blocks with multi-head self-attention and rotary positional embeddings (RoPE) (Su et al., 2021), producing contextualised representations

$$H^{(s)} = g_\theta\big(Z^{(s)}\big) \in \mathbb{R}^{N \times D}. \tag{6}$$

Attention is bidirectional for the MAE and JEPA objectives, and causally masked for autoregressive pretraining (Section 2.4.3); apart from this choice of attention mask, the backbone architecture is identical across objectives.

Heads Objectives that require reconstruction in signal space (MAE and autoregressive prediction) project the encoder outputs back to the original neural feature space through subject-specific read-out layers $h_s : \mathbb{R}^D \to \mathbb{R}^{C_s T_p}$, whereas JEPA operates entirely in latent space and employs no reconstruction head. After pretraining, all pretraining-specific modules are discarded and the encoder is equipped with a task-specific prediction head for downstream decoding.

B.2 Model architecture details

The BIT encoder consists of approximately 7 M parameters. Including subject specific read-in and read-out layers the total number of parameters is 26 M for the human corpus and 41 M for the humans + NHP corpus; these are the parameter counts for MAE and autoregressive pretraining. During JEPA pretraining, instead, a predictor network ($\approx$ 6 M parameters) is additionally used but no read-out layers are employed, amounting to 22 M and 30 M total trainable parameters for the human and human + NHP corpora respectively.

B.2.1 ENCODER ARCHITECTURE

For every objective, the architecture is pretrained with the AdamW optimiser (Loshchilov & Hutter, 2017) for 400 epochs, using a peak learning rate of 5×10^{-4}, weight decay 10^{-5}, batch size 64, a cosine learning-rate schedule with 5% linear warm-up, and mixed-precision (`bfloat16`) training.

Table 3: Encoder architecture and input augmentation, shared across all objectives and corpora. Values follow Zhang et al. (2025).

Hyperparameter	**Value**
Embedding dimension	384
Head dimension	512
Number of heads	6
Depth	7
Patch size T_p	5
Dropout rate	0.2
Attention dropout rate	0.4
White noise standard deviation	0.2
Constant offset standard deviation	0.05
Gaussian smoothing width	2.0

B.2.2 OBJECTIVE-SPECIFIC HYPERPARAMETERS

Table 4: Hyperparameters specific to each pretraining objective. MAE and JEPA use the same span-masking module and the same masking parameters, so that the space in which prediction is performed (signal vs. latent) is the only difference between them. Optimiser settings are shared across all objectives (AdamW, peak learning rate 5×10^{-4}, weight decay 10^{-5}, batch size 64, 400 epochs, cosine schedule with 5% linear warm-up), as are the architectural parameters (Table B.2.1).

Hyperparameter	**MAE**	**JEPA**	**AR**
Attention mask	bidirectional	bidirectional	causal
Masking ratio r	0.5	0.5	—
Max mask span $\ell_{\max}$	15	15	—
Predictor depth	—	6	—
Predictor hidden dimension	—	384	—
EMA momentum schedule	—	$0.996 \rightarrow 1.0$ (cosine)	—
Loss	MSE	smooth-ℓ_1	MSE
Read-out layers	yes	no	yes

C HYPERPARAMETER SEARCH FOR FINE-TUNING

The selection of the optimal hyperparameters for fine-tuning was performed with *Ray tune* (Liaw et al., 2018). We evaluated 30 configurations per experiment, stratified by masking ratio: the input masking ratio applied during fine-tuning was fixed to one of $\{0.0,\ 0.25,\ 0.5\}$, with 10 configurations allocated to each value, while the remaining hyperparameters were sampled at random from the following ranges: batch size $\sim \mathcal{U}\{8, 16, 32, 64\}$, weight decay $\sim \mathrm{LogUniform}(5 \times 10^{-5}, 10^{-1})$, and learning rate $\sim \mathrm{LogUniform}(5 \times 10^{-5}, 10^{-3})$.

D COMPUTE

All pretraining runs were performed on a single NVIDIA H100 GPU. On the HUMAN corpus, 400 epochs required approximately 15, 15 and 14 h for MAE, JEPA and AR; on HUMAN+NHP, approximately 30 h, 34 and 18 respectively.

Hyperparameter search for downstream fine-tunings was performed on NVIDIA A100 GPUs, for a total of roughly 2600 GPU-hours across all configurations, subjects and fine-tuning regimes. This number includes the stability analysis.

E PRETRAINING DIAGNOSTICS

We provide in Table E the values of RankMe and R^2 score obtained on the pretrained models. Since JEPA does not reconstruct the input signal directly, R^2 is not applicable in that case. We point out that in this domain, RankMe alone does not provide a reliable proxy for downstream performance: the AR model has higher RankMe values with respect to MAE and JEPA, yet its performances are notably scarcer.

Table 5: Properties of the pretrained representations. RankMe is computed on a balanced pool of 400 trials from each of the 10 downstream subjects, drawn from the training split ($n = 4000 \gg D = 384$). R^2 is the reconstruction score of each objective on the same pool, and is undefined for JEPA.

	RankMe		$\mathbf{R^2}$	
Objective	HUMAN	HUMAN+NHP	HUMAN	HUMAN+NHP
MAE	216.0	183.0	0.093	0.021
JEPA	274.1	210.6	—	—
AR	270.7	264.7	0.057	0.060

F STABILITY ANALYSIS

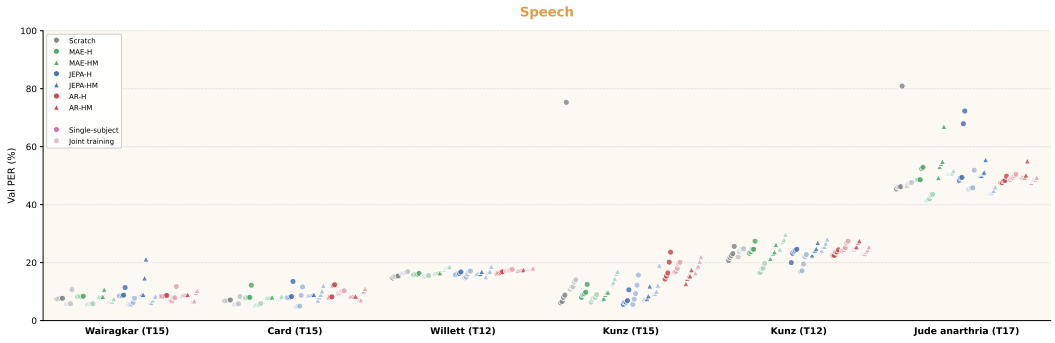

Figure 5: Distribution of validation Phoneme Error Rates across the considered speech decoding datasets and pre-training objectives. For each combination of model, subject, and fine-tuning regime, the best hyperparameter configuration found by sweep is retrained 5 times with different random seeds; each point in the plot represents one run. Colour encodes the pre-training objective (grey: Scratch, green: MAE, blue: JEPA, red: AR); circle markers indicate models pre-trained on human neural data only (H), triangles indicate models pre-trained on human and monkey data (HM). Dark and light shades distinguish single-subject and joint fine-tuning, respectively. Vertical spread within a group reflects seed-to-seed instability.

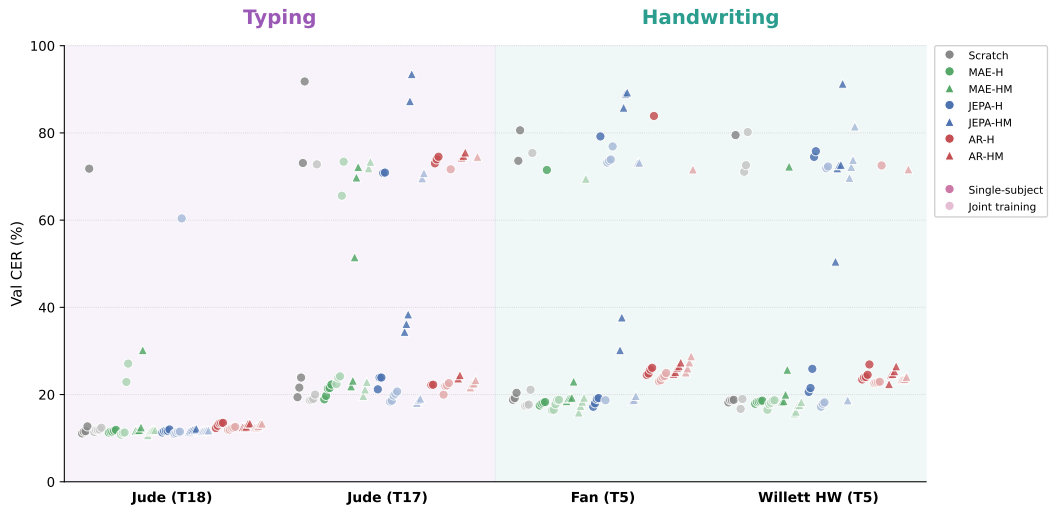

Figure 6: Distribution of validation Character Error Rates across the considered typing and handwriting decoding datasets and pre-training objectives. For each combination of model, subject, and fine-tuning regime, the best hyperparameter configuration found by sweep is retrained 5 times with different random seeds; each point in the plot represents one run. Colour encodes the pre-training objective (grey: Scratch, green: MAE, blue: JEPA, red: AR); circle markers indicate models pre-trained on human neural data only (H), triangles indicate models pre-trained on human and monkey data (HM). Dark and light shades distinguish single-subject and joint fine-tuning, respectively. Vertical spread within a group reflects seed-to-seed instability.

G EXTENDED RESULTS

Table 6: Speech PER (%), subject-specific training, on val/test splits. Card T15 and Kunz T15 are trained as a T15 pair; Willett T12 and Kunz T12 as a T12 pair.

Model	Wairagkar (T15)	Card (T15)	Willett (T12)	Kunz (T15)	Kunz (T12)	Jude anar. (T17)
Scratch	7.57 / **7.27**	7.43 / **7.46**	15.86 / 14.85	6.37 / 9.56	18.27 / **27.88**	46.13 / **47.55**
MAE-H	8.09 / 8.06	8.21 / 8.05	16.47 / 15.39	6.90 / 13.23	19.23 / 33.50	53.41 / 55.03
MAE-HM	9.65 / 9.22	7.92 / 7.58	17.16 / 16.36	8.49 / 11.28	22.36 / 36.96	48.87 / 51.34
JEPA-H	10.24 / 10.01	7.73 / 7.72	15.82 / **14.48**	5.57 / **9.07**	19.95 / 34.40	50.73 / 51.98
JEPA-HM	8.39 / 8.27	8.49 / 8.75	16.30 / 15.52	8.75 / 13.72	22.60 / 35.68	50.21 / 52.77
AR-H	8.38 / 8.17	8.02 / 8.04	16.09 / 15.77	13.00 / 23.04	20.43 / 29.80	54.28 / 56.88
AR-HM	8.73 / 8.48	8.43 / 8.35	16.84 / 16.11	16.18 / 21.32	22.12 / 27.49	49.32 / 50.25

Table 7: Speech PER (%) on val/test splits, joint training across all six speech subjects.

Model	Wairagkar (T15)	Card (T15)	Willett (T12)	Kunz (T15)	Kunz (T12)	Jude anar. (T17)
Scratch	5.79 / 5.67	5.56 / 5.79	16.49 / 14.97	9.28 / 16.91	26.20 / 34.14	46.43 / 47.46
MAE-H	5.69 / **5.64**	5.65 / **5.65**	15.82 / 15.26	7.16 / **8.58**	16.35 / 27.75	42.21 / **42.49**
MAE-HM	6.93 / 6.83	11.97 / 12.04	21.48 / 20.87	27.32 / 28.43	30.53 / 43.86	55.40 / 56.64
JEPA-H	7.59 / 7.78	6.49 / 6.55	15.72 / **14.82**	6.63 / 11.52	22.24 / 28.39	46.11 / 47.00
JEPA-HM	5.84 / 5.90	5.74 / 5.92	16.63 / 15.87	8.49 / 13.97	24.28 / **27.11**	44.37 / 46.20
AR-H	9.43 / 9.33	10.13 / 10.86	17.82 / 16.90	16.45 / 25.25	23.44 / 27.37	48.31 / 50.49
AR-HM	6.83 / 6.55	7.08 / 7.35	17.89 / 17.05	16.45 / 24.51	22.71 / 27.24	48.75 / 49.61

Table 8: Typing CER (%) on val/test splits.

	Jude (T18)		Jude (T17)	
Model	Subject-specific	Joint	Subject-specific	Joint
Scratch	11.59 / 10.78	11.28 / 10.38	19.53 / **17.32**	18.47 / **16.85**
MAE-H	11.36 / 10.42	10.85 / **9.75**	20.29 / 20.05	72.57 / 72.70
MAE-HM	11.47 / 10.50	10.97 / 10.30	20.21 / 19.42	20.67 / 19.27
JEPA-H	11.82 / 10.50	11.55 / 10.10	22.80 / 21.92	18.92 / 16.93
JEPA-HM	11.39 / **10.10**	10.85 / 9.94	73.02 / 72.62	20.14 / 17.78
AR-H	12.86 / 11.29	12.32 / 11.25	22.26 / 22.47	21.66 / 21.61
AR-HM	12.90 / 11.57	12.43 / 11.25	24.16 / 23.71	21.35 / 22.00

Table 9: Handwriting CER (%) on val/test splits.

	Fan		**Willett**	
Model	Subject-specific	Joint	Subject-specific	Joint
Scratch	18.21 / 20.64	18.74 / 20.17	18.43 / 19.74	16.39 / **16.09**
MAE-H	21.30 / 23.40	18.40 / 21.31	18.08 / **18.21**	17.79 / 18.49
MAE-HM	17.30 / **19.46**	16.87 / **19.18**	22.21 / 21.66	17.74 / 17.73
JEPA-H	18.84 / 20.43	18.78 / 21.02	23.16 / 24.26	18.14 / 17.26
JEPA-HM	36.60 / 39.14	69.33 / 69.38	26.77 / 26.75	19.77 / 20.23
AR-H	23.78 / 25.62	23.29 / 24.31	21.23 / 22.66	20.81 / 21.85
AR-HM	26.47 / 28.43	26.79 / 28.72	22.88 / 23.33	24.08 / 24.15