CORTICALL: A UNIVERSAL INTRACORTICAL DECODER FOR COMMUNICATION BCIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Intracortical brain–computer interfaces (BCIs) for communication remain difficult to scale across participants because their input is heterogeneous: individuals differ in the number and placement of implanted arrays, and each channel samples a distinct neural population. Decoders are therefore typically trained one participant at a time, forgoing the benefits of pooled data. We introduce **CorticAll**, a multimodal intracortical decoder that accepts heterogeneous channel sets and supports speech, handwriting, and typing decoding alongside cursor control. A shared Channel Tokeniser processes each channel, Channel Identity embeddings preserve its identity, and a cross-attention Channel Mixer maps any number of channels to a fixed-size representation, enabling end-to-end training across datasets. We evaluate CorticAll on 11 datasets from 7 participants, spanning 4 cortical regions and 4 modalities (≈173 h). We further introduce *supervised-on-scale*, joint supervised training across datasets and modalities, and study it for pooled decoding, fine-tuning, and transfer. Fine-tuning the pooled model reaches state-of-the-art error rates on 7 of the 8 datasets with published baselines and decodes cursor velocity at up to $R^2 = 0.93$. Finally, latent-space analyses reveal substantial shared structure across modalities and participants, suggesting that the model reuses common functional primitives across both.

1 INTRODUCTION

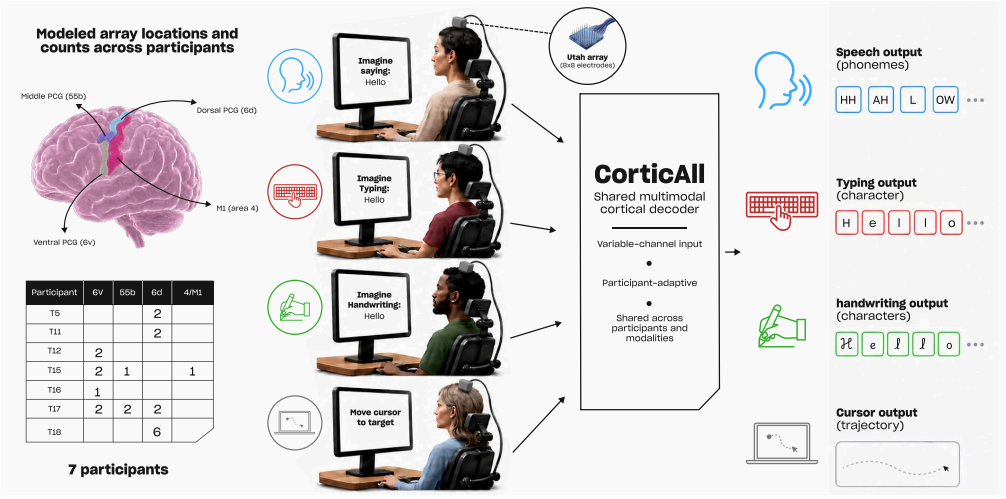

Figure 1: **CorticAll: a single shared decoder for heterogeneous intracortical inputs across four communication modalities.** Seven participants, implanted with Utah microelectrode arrays spanning four cortical regions, attempt or imagine one of four behaviours: vocalisation, typing, handwriting, or cursor control. CorticAll ingests participant-specific neural input with different array count and placement through a shared encoder and decodes it with per-modality heads into phonemes (speech), characters (typing and handwriting), or a 2-D cursor trajectory.

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.

Communication is fundamental to human life. We are a deeply social species, and the ability to convey thoughts and intentions underpins almost every form of participation in society; to lose it is to be cut off from that society. Yet neurodegenerative disease, such as amyotrophic lateral sclerosis (ALS), and acute injury, such as brainstem stroke or high spinal cord injury, can sever the pathway from the intact cognition to the muscles that would express it, in the extreme producing locked-in syndrome: full awareness with no reliable means of communication (Smith & Delargy, 2005). Intracortical brain–computer interfaces (BCIs) aim to reopen this channel by decoding intended behaviour directly from cortical activity. Successive studies have steadily raised the achievable throughput—from two-dimensional cursor control (Kennedy & Bakay, 1998), through imagined handwriting (Willett et al., 2021b) and self-paced keyboard typing (Jude et al., 2026d), to attempted speech within limited (Moses et al., 2021) and open vocabulary (Willett et al., 2023b)—each step conveying more information per unit time and further loosening the constraints of the locked-in state. Beyond demonstrating feasibility, these works mapped the cortical regions amenable to functional decoding and advanced the decoder itself from linear readouts to deep sequence models. In doing so, however, they inherited machine learning's defining constraint: performance is tightly coupled to the quantity of training data (Kaplan et al., 2020), of which a single participant recorded over a handful of sessions supplies relatively little.

A natural place to look for a way past this bottleneck is a neighbouring field that has already overcome it. Decoding text from intracortical signals is, in structure, an instance of automatic speech recognition (ASR): a continuous input must be transcribed into a sequence of discrete tokens with no supervised frame-level alignment between the two. It is unsurprising, then, that the two architectural families dominating ASR recur throughout neural speech decoding: autoregressive encoder–decoder models such as Whisper (Radford et al., 2022), and encoder-only models trained with a connectionist temporal classification (CTC) objective like wav2vec2 (Baevski et al., 2020). The decisive ingredient behind both, however, was scale: Whisper is trained on 680k hours of transcribed speech, and wav2vec2 is pretrained on roughly 50k hours of unlabelled audio before CTC fine-tuning. Such scaling is feasible because the input is homogeneous (a one-dimensional waveform), which lets arbitrarily many uncoordinated sources be pooled into a single corpus.

Intracortical BCIs afford no such homogeneity. Their input is multivariate: each dimension corresponds to a distinct recording channel or feature. Every channel reports activity from a distinct neural population, yet channels are also mutually correlated, so the signal available to a decoder is jointly determined by each channel's individual dynamics and by the co-activations among groups of channels (Gadonneix et al., 2026). The input is therefore **permutation-sensitive**: channel order carries meaning. This heterogeneity compounds across participants, who differ in the number, type, and placement of their arrays. It is a fundamental barrier to the scaling that homogeneous inputs permit, and it underlies the prevailing **one-participant, one-model** design, in which each decoder is trained on a single participant's data. That design sidesteps variable dimensionality and permutation sensitivity by construction, but forfeits the data-scaling benefits documented for neural decoding (Olak et al., 2026). A handful of works pool data across participants, and even species. NDT3 (Ye et al., 2025) trains across humans and monkeys, but does not explicitly preserve channel identity across datasets and remains sensitive to input ordering and sensor variability; its demonstrations are confined to relatively low-dimensional motor tasks, where coarse population-level structure may be sufficient. POYO (Azabou et al., 2023) explicitly addresses this heterogeneity through unit-specific representations, allowing a shared model to operate across recordings with variable neural populations. Yet both approaches decode behaviour at known timepoints rather than alignment-free, variable-length symbol sequences as required for communication tasks such as speech. BIT (Zhang et al., 2025) extends multi-dataset modelling to this harder setting, but preserves subject-specific channel structure through dedicated read-in and read-out layers while sharing only the central temporal Transformer; these dedicated front-ends each see only a fraction of the corpus, plausibly limiting the attainable scaling benefit.

The **training paradigm** is another open question. Prior multi-dataset efforts have relied largely on self-supervised pretraining: NDT2 (Ye et al., 2023) uses masked neural reconstruction and reports pretraining gains disappearing by 3,000 target trials ($\sim$ 50 min), while its successor NDT3 scales pretraining pool to 1750 h yet sees its advantage over training from scratch largely disappear by $\sim$1.5 h of target data. BIT similarly uses masked reconstruction, requiring 367 h of pretraining data to achieve up to $\sim$20% relative reductions in attempted-speech phoneme and word error rate (PER, WER). In contrast, recent supervised multi-dataset training (Boccato et al., 2026b;a) attains relative

reductions of a similar magnitude to BIT with roughly an order of magnitude less data (~ 25 h). To date, this strategy has been confined to a single modality; the effect of joint *multimodal* training is unexplored, and even long-horizon deployments default to per-modality models (Card et al., 2026).

We present **CorticAll**, a multimodal cortical decoder designed to scale across heterogeneous intracortical recordings (Figure 1). CorticAll accepts neural inputs with different channel counts and identities while preserving channel-specific information. Unlike prior systems that use a separate front-end for each dataset, its neural feature extractor is almost entirely shared across the pooled corpus, enabling end-to-end training that makes more efficient use of limited neural data. CorticAll further supports alignment-free, variable-length sequence decoding for speech, handwriting, and typing, alongside continuous cursor decoding within the same model. We evaluate it on 11 publicly released intracortical datasets from 7 human participants, spanning 4 cortical regions and 4 modalities, comprising $\approx$ 103k trials and $\approx 173h$ of recordings.

Toward a data-efficient scaling recipe, we introduce **supervised-on-scale training**: joint supervised training across datasets and modalities, driving the model toward a shared functional representation of neural activity. Whereas self-supervised pretraining seeks general structure in the neural signal itself, supervised-on-scale uses behaviourally relevant targets to shape a representation around the functions the decoder must ultimately recover. We propose this as a potentially more data-efficient alternative to self-supervised pretraining, and as a plausible path toward functional-decoding neural foundation models. Finally, we characterise CorticAll's global participant$\times$modality latent space and find substantial shared structure across speech, handwriting, and typing, with cursor control largely reusing subsets of the same representation.

Contributions.

1. **CorticAll**, a decoder that accepts intracortical input of arbitrary channel dimensionality while retaining persistent channel identity, using a Channel Tokeniser, Channel Identity embeddings, and a cross-attention Channel Mixer that collapses any number of channels to a fixed-size representation.
2. To our knowledge, the first **end-to-end trained, multi-participant, multimodal** intracortical decoder, jointly supporting alignment-free speech, handwriting, and typing decoding alongside continuous cursor control in a single model.
3. **Supervised-on-scale training**—joint supervised training across datasets and modalities to learn a shared functional representation, used both for joint decoding and as pretraining for target-specific fine-tuning and transfer.
4. **State-of-the-art performance** on 7 of the 8 datasets with published baselines, together with latent-space analyses revealing substantial shared structure across modalities.

2 METHODS

2.1 DATA

To test the limits of what a single model can absorb, we aggregate the major publicly available human intracortical datasets recorded with microelectrode (Utah) arrays. The resulting corpus spans **11 datasets from 7 participants, 4 cortical regions, and 4 modalities** (speech, handwriting, typing, cursor control), totalling $\approx$103,000 trials and $\approx$173 h of recording (Table 2). Each source dataset was released with a comparable pipeline of temporal binning and feature extraction, yielding up to two features per channel: **spike count**, the number of voltage threshold crossings per bin; and **spike-band power**, the mean squared voltage on the channel after high-pass filtering at 250 Hz—the spectral band associated with spiking activity. To reconcile the datasets, we interpolate all recordings to a common bin width of 20ms and z-score each feature per session. See section A.1 for detailed description of subjects and datasets, including definition of train/val splits.

Decoding targets. The ultimate goal is to restore communication, whose natural unit is the word. The signals recorded from implanted cortex, however, reflect predominantly *sub-lexical* structure. This is clearest in speech motor cortex, where population activity is organised around articulatory–gestural rather than lexical–semantic features (Bouchard et al., 2013) and encodes measured artic-

ulator kinematics more strongly than higher-level identity (Conant et al., 2018). Phonemes—not words—are therefore the sub-lexical unit most directly tied to this articulatory code, and the appropriate target for speech. For handwriting and typing the behaviour is character-level by construction: participants imagined drawing individual letters, or typed character by character. With those targets we therefore optimize for the best direct neural readout as opposed to highest accuracy of the final word-level communication output. Lexical targets are derived from sentence labels by lowercasing, then: for **speech**, stripping punctuation and mapping to the 39-symbol ARPAbet phoneme inventory plus an inter-word `SIL` token; for **typing**, the 26 alphabetic characters plus an inter-word token and the punctuation `,` `?` `.`; and for **handwriting**, the typing set augmented with the apostrophe. For **cursor** control the direct behavioural ground truth is unavailable, so we instead use each source study's decoded trajectories as targets—the closest available proxy, since those trajectories were confirmed by participants as their intended movement—giving a per-timestep target of 2-D velocity.

2.2 Model

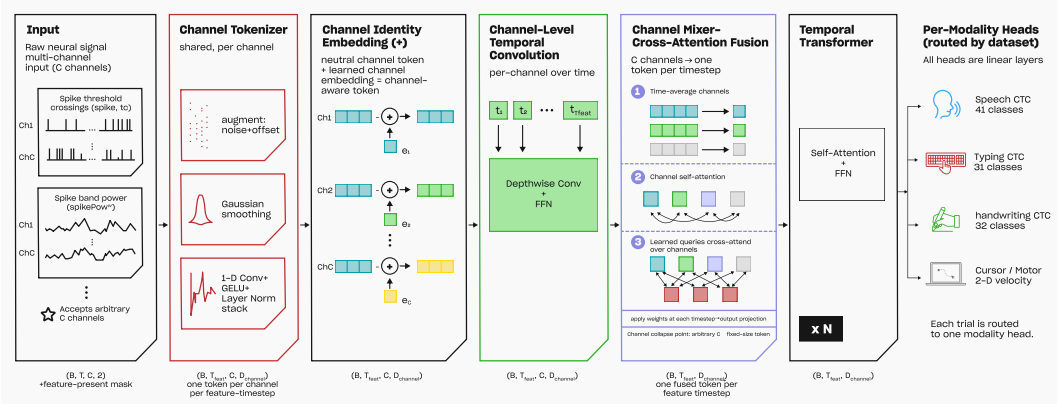

Figure 2: **Overview of the CorticAll architecture.** Per-channel representations are combined into a fixed-size sequence, temporally modeled, and routed to modality-specific decoding heads.

CorticAll is designed as a *universal* intracortical decoder that accepts arbitrary channel counts and configurations and decodes common BCI modalities. It is useful to view neural decoding as two coupled problems: a **spatial** problem, concerning what each channel means and how channels interact, and a **temporal** problem, concerning how neural activity over time encodes behaviour. CorticAll addresses the spatial problem with a Channel Tokeniser, Channel Identity Embeddings, and a cross-attention Channel Mixer, and the temporal problem with a Temporal Transformer and modality-specific decoding heads. Fig. 2 provides an overview, the modules below follow the forward pass, and App. A.5 outlines hyperparameters and parameter budget.

Channel Tokeniser. The Channel Tokeniser operates independently on each channel, using the time-binned, z-scored spike-count and spike-band-power features common to public Utah-array datasets (§2.1). Missing features are replaced by a learned mask token, allowing the same front-end to handle datasets with one or both features. Inputs are Gaussian-smoothed and, during training, augmented with additive noise and a random per-trial offset. A stack of strided 1-D convolutions, GELU nonlinearities, and normalisation layers then emits one channel token every four 20 ms bins (80 ms). By operating identically on every channel, the Tokeniser is encouraged to learn a general representation of per-channel neural features rather than channel-specific transformations. Its sharing across datasets further allows it to learn from the full training pool, unlike prior multi-dataset systems (Feghhi et al., 2025; Boccato et al., 2026b) whose feature extractors are dataset-specific.

Channel Identity Embedding. Each channel token receives a learnable Channel Identity embedding, added element-wise to provide a persistent representation of that channel whenever it appears. This allows the same architecture to operate across participants with distinct channel sets while retaining the identity of each recording channel.

Channel Mixer. Each channel is first contextualised independently over time by a depthwise temporal convolution and feed-forward block. The resulting representation is then time-averaged into one channel vector per channel, and these vectors are contextualised across channels by a three-layer Transformer encoder. Eight learnable queries cross-attend over the channel vectors, each defining a softmax weighting over channels; the resulting per-query pooled representations are concatenated and projected into a single token. The channel-mixing weights are computed once per trial and reused across all timesteps, making the output dimensionality independent of the input channel count. Because the channel vectors contain both Channel Identity and trial-specific activity, the mixing adapts to which channels are present and how active they are.

Temporal Transformer and decoding heads. Neural recordings are multivariate time series, and decoding them requires modelling how evidence distributed across time relates to the underlying behaviour. For this we adopt the Temporal Transformer of Feghhi et al. (2025), a transformer encoder with relative learnable positional encodings, aggressive temporal masking, and a patch embedder. This architecture outperforms the previously dominant RNN decoders (Willett et al., 2023b; Card et al., 2024) and underpins current state-of-the-art speech decoding (Zhang et al., 2025). We retain its stack and hyperparameters, replacing only its feature extractor (patch embedder) with the Channel Tokeniser and Channel Mixer. The resulting representation is mapped to behaviour by modality-specific linear heads.

2.3 Training and evaluation

Objective and optimisation. CorticAll is trained end-to-end with a supervised objective that sums the active modality losses, using CTC (Graves et al., 2006a) for speech, handwriting, and typing and masked Huber loss on 2-D cursor velocity. During mixed-modality training, each head's loss is normalised by an exponential moving average of its recent magnitude before applying its fixed weight, preventing differences in loss scale from dominating training. We optimise with AdamW using linear warmup followed by cosine decay. Full optimisation and training details are given in App.A.6.

Training regimes. We study five regimes defined primarily by their training pool and initialisation:

1. **Single-dataset (from scratch).** The prevalent one-participant one-model baseline
2. **Within-modality supervised-on-scale (from scratch).** Joint training on all datasets of a single modality. Testing the effect of leveraging cross-participant data in supervised training
3. **Full supervised-on-scale (from scratch).** Joint training on all datasets across all four modalities simultaneously. Extension of the cross-participant testing to a multimodal setup.
4. **Target fine-tuning.** Initialise from regime 2 or 3 checkpoint and fine-tune on a target dataset in that pool (columns *Fine-tuned Within-mod* and *Fine-tuned All-mod* respectively). Testing supervised-on-scale as a pretraining formulation
5. **Leave-one-dataset-out transfer.** Train regime 2 on every dataset of a modality *except* a held-out one, then fine-tune on the held-out dataset. Testing foundational-model transfer of supervised pretraining to unseen dataset. Limited to speech-only due to compute demand

Metrics. We report pooled error rate, computed as total Levenshtein edits divided by total reference length, giving phoneme error rate (PER) for speech and character error rate (CER) for handwriting and typing. For cursor control, we report velocity R^2. All metrics are computed on each dataset's validation split. For single-dataset and fine-tuned runs, we evaluate the best target-dataset checkpoint; for multi-dataset from-scratch runs, one checkpoint is selected per modality using its weighted validation metric and shared across all datasets of that modality. For fair comparison we decode greedily and compare against baselines without downstream language-model rescoring.

3 Results

Scaling across datasets improves decoding. Table 1 reports per-dataset decoding performance across the introduced training regimes, alongside prior published results. Training jointly on

Table 1: Decoding accuracy across datasets and training regimes. Cells report the mean ± standard deviation (SD) across three seeds; the leave-one-out fine-tuned (*Unseen*) column is a single seed and is reported without SD. Speech is scored by phoneme error rate (PER, %), handwriting and typing by character error rate (CER, %); cursor is scored by velocity R^2. Columns follow the regimes of Section 2.3. Italic rows report the per-modality weighted mean (metrics averaged proportionally to validation trial count); dashes are configurations not run.

		CorticAll: From scratch			**CorticAll: Fine-tuned**			
Dataset	**Subj.**	Single	Within-mod.	All-mod.	Unseen	Within-mod.	All-mod.	**SOTA**
Speech — PER (%)								
Card	T15	8.5 ± 0.5	7.9 ± 0.1	9.0 ± 1.1	7.7	**6.9** ± 0.3	7.5 ± 0.6	7.1[1]
Willett	T12	14.5 ± 0.4	14.6 ± 0.5	16.5 ± 1.6	14.8	**13.7** ± 0.3	14.8 ± 0.4	14.4[1]
Kunz	T12	24.8 ± 1.4	16.6 ± 1.1	16.4 ± 1.1	21.3	14.6 ± 1.2	**11.8** ± 0.6	18.1[1]
Kunz	T15	22.5 ± 4.7	17.7 ± 1.3	16.6 ± 2.4	15.9	**11.4** ± 0.8	12.4 ± 3.5	17.9[1]
Kunz	T16	23.0 ± 1.8	32.6 ± 3.3	23.6 ± 2.2	40.4	24.0 ± 1.4	**20.1** ± 3.3	–
Kunz	T17	67.4 ± 5.9	32.4 ± 0.5	33.2 ± 1.0	31.1	**27.8** ± 1.3	28.5 ± 0.7	–
Jude	T17	33.9 ± 1.3	25.4 ± 1.4	26.3 ± 4.2	27.6	**23.7** ± 0.6	23.9 ± 2.0	34.0[2]
Wairagkar	T15	11.2 ± 0.8	3.8 ± 0.1	4.2 ± 1.2	4.1	3.1 ± 0.4	**3.0** ± 0.6	7.0[3]
Weighted mean		*16.1* ± 0.7	*12.8* ± 0.3	*13.8* ± 1.2	*12.9*	***11.3*** ± 0.1	*11.8* ± 0.6	–
Handwriting — CER (%)								
Willett	T5	8.1 ± 0.4	7.1 ± 0.6	7.8 ± 0.3	–	7.1 ± 0.4	6.9 ± 0.1	**5.4**[4]
Fan	T5	17.2 ± 2.2	15.7 ± 0.9	17.5 ± 2.1	–	15.2 ± 0.7	**13.7** ± 1.2	19.2[5]
Jude	T18	8.5 ± 1.5	2.4 ± 0.7	3.7 ± 0.7	–	1.9 ± 0.1	**1.8** ± 0.2	–
Weighted mean		*12.2* ± 1.1	*10.7* ± 0.7	*11.8* ± 0.9	–	*10.4* ± 0.5	***9.6*** ± 0.5	–
Typing — CER (%)								
Jude	T17	29.7 ± 0.8	25.6 ± 1.2	23.0 ± 3.7	–	24.8 ± 0.5	**21.7** ± 2.8	27.9[6]
Jude	T18	6.4 ± 0.3	6.2 ± 0.2	5.4 ± 0.5	–	5.7 ± 0.1	**4.5** ± 0.4	8.1[6]
Weighted mean		*17.1* ± 0.6	*15.1* ± 0.5	*13.5* ± 1.9	–	*14.5* ± 0.3	***12.4*** ± 1.5	–
Cursor — velocity R^2 (higher is better)								
Singer-Clark	T15	0.80 ± 0.098	0.90 ± 0.006	0.89 ± 0.008	–	0.92 ± 0.004	**0.93** ± 0.010	–
Wilson	T5	**0.66** ± 0.009	0.62 ± 0.022	0.62 ± 0.005	–	0.64 ± 0.007	0.64 ± 0.003	–
Karpowicz	T11	**0.83** ± 0.005	0.73 ± 0.023	0.64 ± 0.102	–	0.78 ± 0.021	0.76 ± 0.027	–
Weighted mean		***0.66*** ± 0.009	*0.63* ± 0.021	*0.63* ± 0.005	–	*0.65* ± 0.007	*0.65* ± 0.003	–

[1] BIT-All variant from (Zhang et al., 2025); [2] (Jude et al., 2026c);[3] (Wairagkar et al., 2026b); [4] (Willett et al., 2021b); [5] (Fan et al., 2023b); [6] (Jude et al., 2026d). Cursor uses each source study's decoded trajectories as ground truth, so no external baseline applies.

datasets within a modality generally improves over the single-dataset baseline of the prevailing one-participant, one-model paradigm, with particularly large gains on several of the smaller datasets: within-modality training reduces Kunz T17 speech PER from 67.4 to 32.4, Wairagkar speech from 11.2 to 3.8, and Jude T18 handwriting CER from 8.5 to 2.4. At the modality level, it improves the from-scratch weighted mean of all three discrete modalities: speech $16.1 \rightarrow 12.8$, handwriting $12.2 \rightarrow 10.7$, and typing $17.1 \rightarrow 15.1$. Cursor is the exception: Wilson and Karpowicz are already strong when trained individually and do not benefit from pooling, while Singer-Clark improves substantially (R^2 $0.80 \rightarrow 0.90$). Pooling therefore allows the model to learn functional representations from data unavailable to any per-participant decoder, while the shared front-end extends this benefit to the neural feature extractor itself, realising the central premise of supervised-on-scale.

One shared model spans four modalities. Extending the training pool from one modality to all four retains similar aggregate accuracy across a shared multimodal model. Relative to within-modality training, weighted-mean speech PER changes from 12.8 to 13.8, handwriting CER from 10.7 to 11.8, typing CER improves from 15.1 to 13.5, and cursor R^2 remains unchanged at 0.63. Multimodal training can therefore support four behaviourally distinct modalities within the same shared parameterisation, with only modest changes in aggregate decoding performance.

Fine-tuning yields the strongest target performance. Target-specific fine-tuning broadly yields the strongest per-dataset performance and further improves the weighted mean for every modality. The preferred initialisation differs by modality: within-modality pretraining is strongest for speech, whereas all-modality pretraining is strongest for handwriting and typing, showing that multimodal pretraining can provide a useful initialisation even for a single target modality. Across source datasets with a published comparator, CorticAll achieves a lower mean error than the prior reported result on 7 out of 8. The improvements span datasets of different scales: from small collections such as Kunz T15 speech (11.4 vs. 17.9 PER), through medium-scale datasets such as Jude T17 typing (21.7 vs. 27.9 CER), to larger datasets such as Willett T12 speech (13.7 vs. 14.4 PER). The exception is Willett T5 handwriting (6.9 vs. 5.4 CER), whose baseline relied on substantial character-level synthetic data augmentation.

The shared representation transfers to unseen datasets. Holding each speech dataset out of pretraining and introducing it only during fine-tuning improves over single-dataset training on six out of eight datasets, reducing weighted-mean PER from 16.1 to 12.9. This nearly matches within-modality joint training that includes the target (12.8), although it remains behind subsequent target-specific fine-tuning from that pooled model (11.3). The shared representation thus captures structure that transfers to datasets and recording sessions absent from pretraining, rather than merely memorising the training pool, although the current data scale may not yet be sufficient for a fully target-agnostic, foundation-model-style regime.

Additional results. Architecture ablations in Appendix A.2 examine the shared front-end, channel identity, and per-channel temporal modelling. Qualitative examples in Appendix A.4 show discrete predictions and decoded cursor trajectories, including simultaneous speech and cursor decoding.

3.1 Shared latent structure across modalities

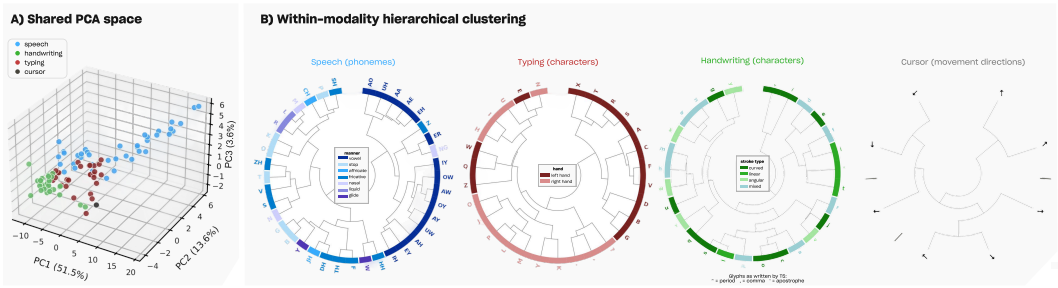

Figure 3: **Shared latent geometry and hierarchical structure of decoding primitives.** (a) Mean latent representation of each primitive, projected onto the top three principal components of CorticAll's multimodal latent activity. (b) Per-modality radial dendrograms (Ward clustering in the shared PC space), each computed on that modality's primitives and coloured by a category the model never sees during training: manner of articulation (speech), typing hand, stroke type (handwriting), and movement direction (cursor).

Latent geometry reflects how multi-modal primitives are produced. To understand how a single model represents four modalities, we analyse the variant trained from scratch on the multimodal corpus, using the 384-dimensional output of its final Temporal Transformer layer before the modality-specific heads. All analyses use the validation split and one seed. Neural activity is segmented into modality *primitives*: phonemes for speech, characters for handwriting and typing, and movement directions for cursor. Primitive onsets are obtained by forced alignment for the discrete modalities (Viterbi, 1967; Graves et al., 2006b; Hira); cursor trials are segmented at local minima in movement speed and each segment is labelled by its mean-velocity direction. PCA is fitted jointly across modalities, with the first three components capturing 68.7% of the variance. Their projection shows substantial overlap among the discrete modalities, while cursor forms a separate, compact cluster (Figure 3a). Within each modality, the latent geometry aligns with attributes of how its primitives are produced (Figure 3b): speech phonemes organise by manner of articulation,

typed characters by the hand used to produce them, handwritten characters by glyph shape, and cursor movements by direction. Although these production attributes are not provided as explicit supervision, they are reflected in CorticAll's learned latent organisation.

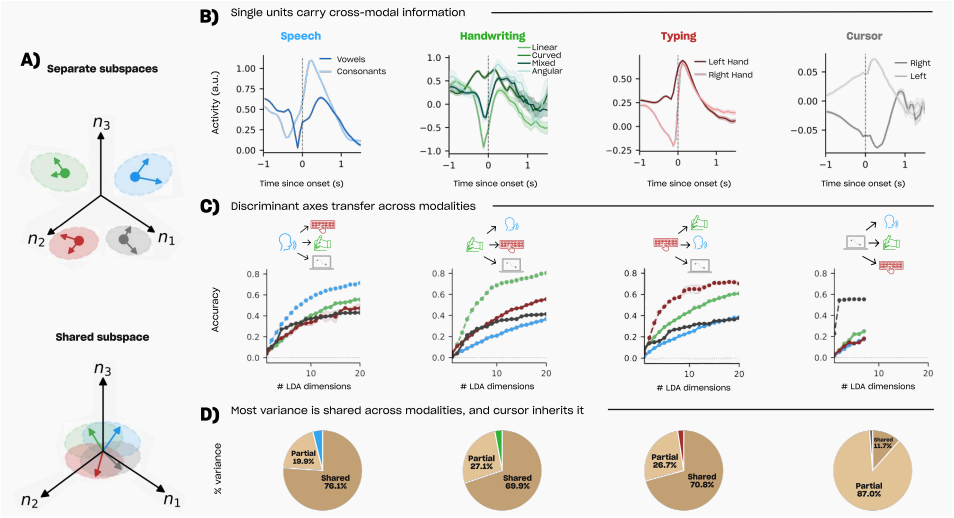

Figure 4: **Speech, handwriting, and typing share a latent subspace that cursor largely inherits.** (a) Two hypotheses: modality-private subspaces versus a single shared subspace. (b) An example dimension from temporal transformer; trial-averaged activity (mean $\pm$ s.e.m.) aligned to primitive onset. (c) LDA axes fit on one modality; population activity from every other modality projected onto the same axes. In each plot, the chance-corrected accuracy is shown as a function of number of linear discriminants. (d) DySO variance decomposed in shared / partial / unique as pie charts.

Latent subspaces are shared across modalities and participants. We ask whether CorticAll represents each modality in separate or shared subspaces (Figure 4a). As a qualitative check, we inspect individual latent dimensions of the Temporal Transformer output. An example unit (Figure 4b) is modulated across all four modalities, separating speech vowels from consonants, curved from linear handwriting, left- from right-hand typing, and leftward from rightward cursor movement. We quantify this by fitting linear discriminants (LDs) to distinguish primitive classes in one modality, then projecting the other modalities onto the same subspace (Figure 4c). Transfer is strong among speech, handwriting, and typing, weaker to cursor, and weakest from cursor-derived axes. Finally, Dynamic Subspace Overlap (DySO), generalising the two-condition analysis of Dekleva et al. (2024), decomposes each modality's variance into dimensions *shared* across all modalities, *partial* across a subset, or *unique* to that modality (Figure 4d). Most variance in speech, handwriting, and typing is fully shared (76.1%, 69.9%, and 70.8%, respectively), whereas cursor is mostly partial (87.0%), with only 11.7% fully shared and almost no unique variance. A complementary within-modality DySO analysis across participants shows that the same behaviour is represented largely in shared rather than participant-specific dimensions (App. A.3, Figure 6). Together, these analyses suggest that CorticAll maps heterogeneous neural inputs into a largely shared latent basis across participants and communication modalities, while cursor inherits a more restricted subset of that representation - possibly reflecting its lower-dimensional motor repertoire.

4 Discussion

Across machine-learning domains, scale has been a dependable route to better performance. Our results, together with growing evidence from multi-participant neural decoding (Boccato et al., 2026b; Zhang et al., 2025; Ye et al., 2025; Fogg et al., 2026; Azabou et al., 2025), suggest that intracortical BCIs can benefit from the same principle. Their central obstacle is heterogeneity: participants differ in channel count, placement, and identity, which has historically kept datasets siloed and decoders participant-specific. CorticAll addresses this with a shared Channel Tokeniser, Channel

Identity embeddings, and Channel Mixer, allowing one end-to-end trained model to pool data across participants, cortical regions, and four communication and control modalities. This pooling is beneficial, not merely possible: fine-tuning the pooled model achieves state-of-the-art performance on the majority of datasets with published baselines and achieves cursor-velocity R^2 of 0.64–0.93. A single model spanning speech, typing, handwriting, and cursor control could also simplify clinical deployment by replacing separate decoders for each function.

Beyond aggregate accuracy, CorticAll develops a largely shared latent representation across participants and modalities. This is consistent with evidence that neural population activity can reuse low-dimensional structure across different motor behaviours (Gallego et al., 2018), across related tasks (Tafazoli et al., 2025), and across individuals performing the same behaviour (Spalding et al., 2026). Our results extend this picture at the level of a learned decoder: speech, handwriting, and typing occupy largely shared dimensions, while cursor predominantly reuses subsets of this representation. These are, however, representations learned by CorticAll rather than direct measurements of a shared biological subspace, so the observed overlap may reflect common neural organisation, regularities of the behaviours, or structure induced by the labels and training objective. Regardless of its source, this shared organisation illustrates the principle behind supervised-on-scale: rather than learning general statistical structure from neural activity alone, as in self-supervised pretraining, behavioural supervision shapes the pooled representation around functional structure useful for decoding. Our results therefore establish supervised-on-scale as a viable way to exploit heterogeneous neural data, both for joint decoding and as pretraining for subsequent fine-tuning, although we do not compare it with self-supervised pretraining under matched conditions. Establishing whether the learned shared axes also correspond to known anatomical or physiological organisation is an important complementary direction.

A key question is how well such a pretrained representation transfers to an unseen dataset since joint training clearly benefits participants already represented in the training pool. In the leave-one-dataset-out setting, however, transfer is useful but incomplete: pretraining on the remaining datasets improves over training from scratch on most speech datasets, yet remains behind models that included the target during pretraining. Whether this gap reflects limits to cross-participant sharing or simply the current scale and diversity of available data remains open.

Scaling further is constrained not by data collection alone: intracortical data capture is increasing across academic and industry efforts, alongside a growing number of implanted participants (Patrick-Krueger et al., 2025), yet much of it cannot be openly released. Naturalistic recordings are particularly valuable because they capture the conditions in which BCIs are ultimately meant to operate: free-flowing communication with realistic linguistic context, prosody, paralinguistic cues, and communicative intent that constrained laboratory protocols cannot fully reproduce. Yet these same recordings are especially sensitive: Card et al. (2026) collected more than 3,800 hours of at-home BCI use but note that the neural data and transcripts cannot be released because they capture private daily conversation. Progress will therefore require coordination across labs to develop practical mechanisms for learning from such data without compromising participant privacy, including controlled access, privacy-preserving training, and foundation models trained on shareable data and adapted locally to participant data while keeping personal information under participant control.

Several limitations remain. Adaptation to a new participant requires initialising new Channel Identity embeddings and fine-tuning the model on labelled target data, limiting few- and zero-shot deployment. Inferring functionally informed channel representations from minimal calibration data is a natural next step. We also evaluate communication at the phoneme and character level, which isolates neural decoding performance but stops short of deployable word-level communication. Integrating word prediction directly into the architecture could enable end-to-end neural-to-word decoding without relying on a separate language-model rescoring stage. Our cross-paper comparisons may still reflect differences in data splits and evaluation protocols. Data augmentation offers another complementary route to reducing labelled-data requirements by synthesising plausible neural activity for desired outputs, and deserves systematic study in this multi-dataset setting. Despite these limitations, the central result is consistent across datasets: pooling heterogeneous neural data across participants and modalities is both feasible and beneficial, providing a path toward increasingly general intracortical decoders.

AI USE STATEMENT

In this work, we used generative AI tools to obtain feedback on research methodology or experiments, implement some of the methods and partially to clean and reformat datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, or to support qualitative and thematic data analysis or interpret results, while the following support is not applicable to this work: generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, creation of artifacts, summarize or analyse existing literature, edit the paper to improve readability, identify relevant literature. We have reviewed all AI-assisted work. The LLM-generated code was reviewed and tested by the main author including the code for processing the data, the summarized literature was subsequently reviewed, the corrected manuscript text was reviewed manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses the publicly available intracortical datasets referenced in Table 2; no new human data were collected. The datasets were gathered under respective IRB oversight with informed consent, and all analyses were performed on de-identified recordings. We complied with each dataset's license and did not attempt re-identification or linkage with external sources. Our models are intended to advance scientific understanding of neural speech decoding and are not suitable for clinical, legal, or surveillance use; any future deployment of neuro-prosthetic BCIs should include explicit consent, human oversight, and protections against non-consensual decoding.

REPRODUCIBILITY STATEMENT

All experiments use publicly released datasets, with no new or private neural recordings required to reproduce the results. Section 2.1 and Appendix A.1 document the datasets, neural features, preprocessing, decoding targets, and dataset-specific train/validation splits. The model architecture is described in Section 2.2 and Appendix A.5, including complete architecture hyperparameters and parameter counts. The training and evaluation regimes are defined in Section 2.3, while Appendix A.6 reports the optimisation, loss balancing, augmentation, checkpoint-selection procedure, random seeds, hardware, and computational cost used for the reported experiments. We additionally provide the code used for model training and evaluation as supplementary material.

REFERENCES

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh B Nachimuthu, Michael Jacob Mendelson, Blake Aaron Richards, Matthew G Perich, Guillaume Lajoie, and Eva L Dyer. A unified, scalable framework for neural population decoding. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sw2Y0sirtM.

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.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

Tommaso Boccato, Michal Olak, and Matteo Ferrante. Brainwhisperer: Leveraging large-scale asr models for neural speech decoding, 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. March 2026b. doi: 10.64898/2026.02.27.708564. URL http://dx.doi.org/10.64898/2026.02.27.708564.

Kristofer E. Bouchard, Nima Mesgarani, Keith Johnson, and Edward F. Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, February 2013. ISSN 1476-4687. doi: 10.1038/nature11911. URL http://dx.doi.org/10.1038/nature11911.

Nicholas Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis Willett, Erin Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrell Deo, Aparna Srinivasan, Eun Young Choi, Matthew Glasser, Leigh Hochberg, Jaimie Henderson, Kiarash Shahlaie, Sergey Stavisky, and David Brandman. Data for: An accurate and rapidly calibrating speech neuroprosthesis, 2025. URL https://datadryad.org/dataset/doi:10.5061/dryad.dncjsxm85.

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.

Nicholas S. Card, Tyler Singer-Clark, Hamza Peracha, Carrina Iacobacci, Xianda Hou, Maitreyee Wairagkar, Zachery Fogg, Elena C. Offenberg, Leigh R. Hochberg, Sergey D. Stavisky, and David M. Brandman. Long-term independent use of an intracortical brain–computer interface for speech and cursor control. *Nature Medicine*, 32(7):2504–2510, 2026. ISSN 1546-170X. doi: 10.1038/s41591-026-04414-6. URL http://dx.doi.org/10.1038/s41591-026-04414-6.

David F. Conant, Kristofer E. Bouchard, Matthew K. Leonard, and Edward F. Chang. Human sensorimotor cortex control of directly measured vocal tract movements during vowel production. *The Journal of Neuroscience*, 38(12):2955–2966, February 2018. ISSN 1529-2401. doi: 10.1523/jneurosci.2382-17.2018. URL http://dx.doi.org/10.1523/JNEUROSCI.2382-17.2018.

Brian M Dekleva, Raeed H Chowdhury, Aaron P Batista, Steven M Chase, Byron M Yu, Michael L Boninger, and Jennifer L Collinger. Motor cortex retains and reorients neural dynamics during motor imagery. *Nature human behaviour*, 8(4):729–742, 2024.

Darrel R. Deo, Elizaveta V. Okorokova, Anna L. Pritchard, Nick V. Hahn, Nicholas S. Card, Samuel R. Nason-Tomaszewski, Justin Jude, Thomas Hosman, Eun Young Choi, Deqiang Qiu, Yuguang Meng, Maitreyee Wairagkar, Claire Nicolas, Foram B. Kamdar, Carrina Iacobacci, Alexander Acosta, Leigh R. Hochberg, Sydney S. Cash, Ziv M. Williams, Daniel B. Rubin, David M. Brandman, Sergey D. Stavisky, Nicholas AuYong, Chethan Pandarinath, John E. Downey, Sliman J. Bensmaia, Jaimie M. Henderson, and Francis R. Willett. A mosaic of whole-body representations on the human precentral gyrus. *Nature*, 2026. doi: 10.1038/s41586-026-10653-x.

Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy Wilson, Leigh Hochberg, Krishna 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, 2023a. URL https://datadryad.org/dataset/doi:10.5061/dryad.hqbzkh1p6.

Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy H. Wilson, Leigh Hochberg, Krishna V. Shenoy, Jaimie M. Henderson, and Francis R. Willett. Plug-and-play stability for intracortical brain-computer interfaces: a one-year demonstration of seamless brain-to-text communication. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023b. Curran Associates Inc.

Ebrahim Feghhi, Shreyas Kaasyap, Nima Ryan Hadidi, and Jonathan Kao. Time-masked transformers with lightweight test-time adaptation for neural speech decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=0U7D9AFiZ0.

Zachery M. Fogg, Nicholas S. Card, Maitreyee Wairagkar, Aparna Srinivasan, Tyler Singer-Clark, Xianda Hou, Elizaveta Okorokova, Hamza Peracha, Carrina Iacobacci, Tiffany Brailow, Justin Jude, Hadar Levi-Aharoni, Trung Le, Domenick Mifsud, Pranav I. Deevi, Samuel R. Nason-Tomaszewski, Anna L. Pritchard, Yunnuo Zhang, Brice Richards, Payton Bechefsky, Leigh R. Hochberg, Ziv Williams, Kiarash Shahlaie, Nicholas AuYong, Daniel B. Rubin, Chethan Pandarinath, David M. Brandman, and Sergey D. Stavisky. A generalizable speech neuroprosthesis. 2026. doi: 10.64898/2026.07.23.739430. URL http://dx.doi.org/10.64898/2026.07.23.739430.

Julien Gadonneix, Mingfang Zhang, Jérémy Rapin, Linnea Evanson, Pierre Bourdillon, and Jean-Rémi King. Temporal structure of the language hierarchy within small cortical patches, 2026. URL https://arxiv.org/abs/2604.03021.

Juan A. Gallego, Matthew G. Perich, Stephanie N. Naufel, Christian Ethier, Sara A. Solla, and Lee E. Miller. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature Communications*, 9(1), October 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06560-z. URL http://dx.doi.org/10.1038/s41467-018-06560-z.

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, pp. 369–376, New York, NY, USA, 2006a. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143891. URL https://doi.org/10.1145/1143844.1143891.

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*, pp. 369–376, 2006b.

Moto Hira. Forced alignment with wav2vec2. URL https://docs.pytorch.org/audio/stable/tutorials/forced_alignment_tutorial.html.

Beata Jarosiewicz, Anish A. Sarma, Daniel Bacher, Nicolas Y. Masse, John D. Simeral, Brittany Sorice, Erin M. Oakley, Christine Blabe, Chethan Pandarinath, Vikash Gilja, Sydney S. Cash, Emad N. Eskandar, Gerhard Friehs, Jaimie M. Henderson, Krishna V. Shenoy, John P. Donoghue, and Leigh R. Hochberg. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science Translational Medicine*, 7(313), November 2015. ISSN 1946-6242. doi: 10.1126/scitranslmed.aac7328. URL http://dx.doi.org/10.1126/scitranslmed.aac7328.

Justin 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. Data from: Decoding intended speech with an intracortical brain-computer interface in a person with longstanding anarthria and locked-in syndrome, 2026a. URL https://datadryad.org/dataset/doi:10.5061/dryad.vq83bk481.

Justin Jude, Hadar Levi Aharoni, Alexander Acosta, Shane Allcroft, Claire Nicolas, Bayardo Lacayo, Nicholas Card, Maitreyee Wairagkar, Alisa Levin, David Brandman, Sergey Stavisky, Francis Willett, Ziv Williams, John Simeral, Leigh Hochberg, and Daniel Rubin. Restoring rapid natural bimanual typing with a neuroprosthesis after paralysis, 2026b. URL https://datadryad.org/dataset/doi:10.5061/dryad.cz8w9gjjk.

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 longstanding anarthria and locked-in syndrome. *Cell Reports*, 45(4):117162, April 2026c. 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 2026d. ISSN 1546-1726. doi: 10.1038/s41593-026-02218-y. URL http://dx.doi.org/10.1038/s41593-026-02218-y.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Brianna M. Karpowicz, Yahia H. Ali, Lahiru N. Wimalasena, Andrew R. Sedler, Mohammad Reza Keshtkaran, Kevin Bodkin, Xuan Ma, Daniel B. Rubin, Ziv M. Williams, Sydney S. Cash, Leigh R. Hochberg, Lee E. Miller, and Chethan Pandarinath. Data from: Stabilizing brain-computer interfaces through alignment of latent dynamics, 2025. URL https://datadryad.org/dataset/doi:10.5061/dryad.q83bk3jtp.

P R. Kennedy and R A. E. Bakay. Restoration of neural output from a paralyzed patient by a direct brain connection. *NeuroReport*, 9(8):1707–1711, 1998. ISSN 0959-4965. doi: 10.1097/00001756-199806010-00007. URL http://dx.doi.org/10.1097/00001756-199806010-00007.

Erin Kunz, Benyamin Abramovich Krasa, Foram Kamdar, Donald Avansino, Nick Hahn, Seonghyun Yoon, Akansha Singh, Samuel Nason-Tomaszewski, Nicholas Card, Justin Jude, Brandon Jacques, Payton Bechefsky, Carrina Iacobacci, Leigh Hochberg, Daniel Rubin, Ziv Williams, David Brandman, Sergey Stavisky, Nicholas AuYong, Chethan Pandarinath, Shaul Druckmann, Jaimie Henderson, and Francis Willett. Inner speech in motor cortex and implications for speech neuroprostheses, 2025. URL https://datadryad.org/dataset/doi:10.5061/dryad.gf1vhhn1j.

David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021. ISSN 1533-4406. doi: 10.1056/nejmoa2027540. URL http://dx.doi.org/10.1056/NEJMoa2027540.

Michal Olak, Tommaso Boccato, and Matteo Ferrante. Decoding the decoder: Contextual sequence-to-sequence modeling for intracortical speech decoding, 2026. URL https://arxiv.org/abs/2603.20246.

K. Michelle Patrick-Krueger, Ian Burkhart, and Jose L. Contreras-Vidal. The state of clinical trials of implantable brain–computer interfaces. *Nature Reviews Bioengineering*, 3(1):50–67, 2025.

János A Perge, Mark L Homer, Wasim Q Malik, Sydney Cash, Emad Eskandar, Gerhard Friehs, John P Donoghue, and Leigh R Hochberg. Intra-day signal instabilities affect decoding performance in an intracortical neural interface system. *Journal of Neural Engineering*, 10(3): 036004, April 2013. ISSN 1741-2552. doi: 10.1088/1741-2560/10/3/036004. URL http://dx.doi.org/10.1088/1741-2560/10/3/036004.

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.

Tyler Singer-Clark, Xianda Hou, Nicholas Card, Maitreyee Wairagkar, Carrina Iacobacci, Hamza Peracha, Leigh Hochberg, Sergey Stavisky, and David Brandman. Data from: Speech motor cortex enables bci cursor control and click, 2026. URL https://datadryad.org/dataset/doi:10.5061/dryad.prr4xgxzq.

Eimear Smith and Mark Delargy. Locked-in syndrome. *BMJ*, 330(7488):406–409, February 2005. ISSN 1468-5833. doi: 10.1136/bmj.330.7488.406. URL http://dx.doi.org/10.1136/bmj.330.7488.406.

Z. Spalding, S. Duraivel, S. Rahimpour, C. Wang, K. Barth, C. Schmitz, S. P. Lad, A. H. Friedman, D. G. Southwell, J. Viventi, and G. B. Cogan. Shared latent representations of speech production for cross-patient speech decoding. *Nature Communications*, 17(1), 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-75455-1. URL http://dx.doi.org/10.1038/s41467-026-75455-1.

Sina Tafazoli, Flora M. Bouchacourt, Adel Ardalan, Nikola T. Markov, Motoaki Uchimura, Marcelo G. Mattar, Nathaniel D. Daw, and Timothy J. Buschman. Building compositional tasks with shared neural subspaces. *Nature*, 650(8100):164–172, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09805-2. URL http://dx.doi.org/10.1038/s41586-025-09805-2.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

Maitreyee Wairagkar, Nicholas Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee Miller, Leigh Hochberg, David Brandman, and Sergey Stavisky. Data for: An instantaneous voice synthesis neuroprosthesis, 2026a. URL https://datadryad.org/dataset/doi:10.5061/dryad.2280gb64f.

Maitreyee Wairagkar, Aparna Srinivasan, Nicholas S. Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M. Miller, Leigh R. Hochberg, David M. Brandman, and Sergey D. Stavisky. Brain2voice 2.0: High-performance voice synthesis brain-computer interface. 2026b. doi: 10.64898/2026.06.30.735633. URL http://dx.doi.org/10.64898/2026.06.30.735633.

Francis Willett, Donald Avansino, Leigh Hochberg, Jaimie Henderson, and Krishna Shenoy. Data from: High-performance brain-to-text communication via handwriting, 2021a. URL https://datadryad.org/dataset/doi:10.5061/dryad.wh70rxwmv.

Francis Willett, Erin Kunz, Chaofei Fan, Donald Avansino, Guy Wilson, Eun Young Choi, Foram Kamdar, Matthew Glasser, Leigh Hochberg, Shaul Druckmann, Krishna Shenoy, and Jaimie Henderson. Data for: A high-performance speech neuroprosthesis, 2023a. URL https://datadryad.org/dataset/doi:10.5061/dryad.x69p8czpq.

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 2021b. 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 2023b. ISSN 1476-4687. doi: 10.1038/s41586-023-06377-x. URL http://dx.doi.org/10.1038/s41586-023-06377-x.

Guy Wilson, Elias Stein, Foram Kamdar, Donald Avansino, Tsam Pun, Ronnie Gross, Tommy Hosman, Tyler Singer-Clark, Anastasia Kapitonava, Leigh Hochberg, John Simeral, Krishna Shenoy, Shaul Druckmann, Jaimie Henderson, and Francis Willett. Data from: Long-term unsupervised recalibration of cursor bcis, 2025. URL https://datadryad.org/dataset/doi:10.5061/dryad.1jwstqk6g.

Joel Ye, Jennifer L Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=CBBtMnlTGq.

Joel Ye, Fabio Rizzoglio, Xuan Ma, Adam Smoulder, Hongwei Mao, Gary Blumenthal, William Hockeimer, Nicolas Kunigk, Dalton Moore, Patrick Marino, Raeed Chowdhury, J. Patrick Mayo, Aaron Batista, Steven Chase, Michael Boninger, Charles Greenspon, Andrew B Schwartz, Nicholas Hatsopoulos, Lee Miller, Kristofer Bouchard, Jennifer Collinger, Leila Wehbe, and

Robert Gaunt. A generalist intracortical motor decoder. In *Advances in Neural Information Processing Systems 38*, NeurIPS 2025, pp. 122532–122572. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-3692. URL http://dx.doi.org/10.52202/085713-3692.

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. doi: 10.48550/ARXIV.2511.21740. URL https://arxiv.org/abs/2511.21740.

A Appendix

A.1 Data details

The aggregated corpus comprises 11 publicly released source datasets, spanning 7 participants, 4 modalities, and $\approx$ 173 hours of recording. Table 2 summarises the dataset-specific recordings and inputs used for modelling, while Table 3 provides participant, implant, and clinical context. Below we describe shared feature processing and the dataset-specific preprocessing and train/validation splits.

Table 2: Aggregated intracortical datasets, grouped by source dataset. All recordings use 64- or 96-channel Utah arrays. "Count" and "power" denote spike count and spike-band power; Ch. gives channels used/total. In the implant column, † marks implanted arrays excluded from modelling, and $\times n$ denotes the number of arrays in the corresponding area. Trials gives the number of examples used for modelling. Bin denotes the temporal resolution of the released features; all inputs are resampled to 20 ms before modelling.

Dataset ref.	Modality	Subj.	Implant loc.	Ch. (u/t)	Trials	Hours	Span	Sess.	Feat.	Bin
Card et al. (2025)	Speech	T15	i6v, s6v, 55b, M1	256/256	9,495	46.5	611d	45	count, power	20 ms
Willett et al. (2023a)	Speech	T12	i6v, s6v, 44 $\times 2^{\dagger}$	128/256	9,680	16.9	119d	24	count, power	20 ms
Jude et al. (2026a)	Speech	T17	s6v $\times 2$, 55b $\times 2^{\dagger}$	128/256	4,124	10.3	140d	13	count, power	20 ms
Wairagkar et al. (2026a)	Speech	T15	i6v, s6v, 55b, M1	256/256	1,675	12.1	128d	13	count, power	10 ms
Kunz et al. (2025)	Speech	T12	i6v, s6v	128/128	500	0.8	1d	1	count, power	20 ms
	Speech	T15	i6v, s6v, 55b, M1	256/256	665	3.0	308d	3	count, power	10 ms
	Speech	T16	i6v, PEF/55b†, 6d $\times 2^{\dagger}$	64/256	415	0.9	233d	2	count, power	20 ms
	Speech	T17	s6v $\times 2$, 55b $\times 2^{\dagger}$	128/256	980	0.6	1d	1	count, power	10 ms
Jude et al. (2026b)	Typing	T17	s6v $\times 2$, 6d $\times 2$, 55b $\times 2^{\dagger}$	256/384	664	9.5	88d	8	count, power	10 ms
	Typing	T18	6d $\times 6$	384/384	881	8.6	63d	8	count, power	10 ms
	Handwriting	T18	6d $\times 6$	384/384	122	1.3	1d	1	count, power	10 ms
Willett et al. (2021a)	Handwriting	T5	6d $\times 2$	192/192	730	9.4	252d	10	count	20 ms
Fan et al. (2023a)	Handwriting	T5	6d $\times 2$	192/192	938	11.7	286d	21	count	20 ms
Wilson et al. (2025)	Cursor	T5	6d $\times 2$	192/192	69,141	38.5	1764d	103	count	20 ms
Karpowicz et al. (2025)	Cursor	T11	6d $\times 2$	192/192	119	0.02	n/a	2	count	20 ms
Singer-Clark et al. (2026)	Cursor	T15	i6v, s6v, 55b, M1	256/256	2,176	2.2	429d	3	count, power	10 ms
	Speech	T15	i6v, s6v, 55b, M1	256/256	554	1.0	429d	3	count, power	10 ms

Neural features and binning. Each array channel contributes threshold-crossing spike counts ("count") and, where available, spike-band power ("power"). The T5 and T11 handwriting and cursor datasets provide spike counts only; all remaining datasets provide both features. Regardless of the preprocessing supplied with each release, we re-normalise all features by z-scoring within each recording block or session and clip the resulting values at ± 10. All datasets are resampled to a common 20 ms temporal resolution before modelling. Count and power features are concatenated count-first into a per-participant vector, with channels ordered according to each release's native channel map; the Willett speech dataset additionally applies the fixed spatial permutation supplied with the release. Arrays retained for decoding and the resulting channel counts are listed in Table 2.

Speech datasets. Willett 2023 (T12). We use the attempted-speech dataset released for the Brain-to-Text 2024 competition, retaining only the two ventral area-6v arrays as in the original work.

Table 3: Participants contributing to the aggregated corpus. All arrays are Blackrock Utah microelectrode arrays with 1.5 mm electrodes; implants are left-hemisphere unless otherwise indicated. Array counts give the full implant, including arrays excluded from decoding (Table 2). Area targets follow the HCP multimodal parcellation (6v: ventral premotor; 6d: dorsal premotor / hand knob; 55b: middle precentral; 4/M1: primary motor; 44: inferior frontal; PEF: premotor eye field). Clinical and anatomical details for T5–T17 are reported by Deo et al. (2026); T18 details follow Jude et al. (2026b).

Subj.	Aetiology	Arrays (implanted)	Array targets	Task-relevant status
T5	Spinal cord injury (C4 AIS-C)	2×96 ch	hand-knob 6d ($\times 2$)	tetraplegia; full face/head movement
T11	Spinal cord injury (C4 AIS-B)	2×96 ch	hand-knob 6d ($\times 2$)	tetraplegia; very limited arm movement
T12	ALS (bulbar-onset)	4×64 ch	6v ($\times 2$); 44 ($\times 2$, excluded)	anarthric; functionally independent (limb strength 3–4/5)
T15	ALS	4×64 ch	6v ($\times 2$), 55b, 4/M1	anarthric (can vocalise); very limited body movement
T16	Pontine stroke	4×64 ch	6d ($\times 2$), 6v, 55b (PEF border)	dysarthric; slow, quiet speech; limited upper-limb movement
T17	ALS (rapidly progressive)	6×64 ch	6d ($\times 2$), 6v ($\times 2$), 55b ($\times 2$)	anarthric; quadriplegic; ventilator-dependent; eye movement only
T18	Spinal cord injury (C4)	6×64 ch	6d ($\times 4$ left, $\times 2$ right)	tetraplegia; no movement below C4

The official train/validation split is used (8,800/880 trials), while the 1,200-trial blinded competition hold-out is omitted. **Card 2025 (T15).** We use the attempted-speech dataset released for the Brain-to-Text 2025 competition and retain the full four-array T15 implant. We follow the official train/validation split (8,069/1,426 trials); the 1,450-trial official test set is omitted. **Jude speech (T17).** We retain the two ventral-6v arrays used in the original intended-speech study. The release does not provide a standard train/validation partition, so we construct one following the evaluation setup in the official code: 54 trials from recording day 68 form the validation set and the remaining 4,070 trials are used for training. The original study also reports PER on day 68, allowing direct comparison on the same evaluation data. **Wairagkar 2024 (T15).** We use the full four-array T15 layout. To avoid overlap with the Card dataset, we restrict training to 1,585 trials from recording days not represented in Card. Validation set comprises 90 trials denoted as "benchmark set" in original study, which was also used to provide the baseline performance. **Kunz 2025 (T12, T15, T16, T17).** This multi-participant attempted/imagined-speech dataset contains word-only and full-sentence recordings and does not provide a single evaluation protocol across participants. For T12, T15, and T16, we follow the evaluation convention of Zhang et al. (2025), utilizing only the full-sentence trials, using the 50-word-vocabulary trials for validation and the remaining trials for training. This yields train/validation counts of 400/100, 615/50, and 365/50, respectively. T17 contains only single-word trials, so we instead construct an 80/20 split stratified by word (784/196). **Singer-Clark speech (T15).** The speech-production blocks recorded during simultaneous speech and cursor control are routed to the speech head. They use the same split as the corresponding cursor dataset described below (554 trials total).

Handwriting datasets. **Willett 2021 (T5).** We use attempted handwriting recorded from the two hand-knob 6d arrays, with spike counts as the only neural feature and the official train/validation split (502/228 trials). The 31-symbol target set contains the 26 lowercase letters, comma, apostrophe, question mark, period, and space; T5 produced period and space as '~' and '>'. Characters were written in print, each on top of the previous one. We did not identify subsequent work reporting directly comparable CER on the same released split. **Fan 2023 (T5).** This later T5 handwriting collection uses the same two 6d arrays and spike-count-only. We use the release's official session-hold-out split, which matches the evaluation split used in the original study, giving train 739 / val 199 splits. The dataset is also included in the FALCON benchmark, but FALCON defines its own train/test partition and its reported scores are therefore not directly comparable to ours. We did not identify other published results using the split evaluated in this study, other than the original work. **Jude handwriting (T18).** We use attempted handwriting recorded from all six 6d arrays, with the same protocol and targets as the Willett handwriting dataset. We reserve 20% of trials for validation, prioritising globally unique sentences so that no sentence occurs in both splits (98/24 trials).

Typing datasets. **Jude typing (T17, T18).** We use bimanual attempted typing with gesture-encoded character targets. For T17, we retain the two s6v and two 6d arrays; for T18, we use all six 6d arrays. We follow the block-level train/validation assignments provided with the released data and evaluate on these validation blocks. The original study reports CER separately for each validation day rather than as a single aggregate value; for the benchmark in Table 1, we therefore

construct one overall CER by averaging the reported day-level values weighted by the number of validation trials on each day.

Cursor datasets. **Wilson 2025 (T5).** We split the cursor-control data 80/20 under a fixed seed, stratified by task (55,312/13,829 trials), and compute velocity normalisation on training data only before applying it to validation. With 103 sessions spanning $\approx$ 4.8 years, this is the largest and longest-running dataset in the corpus. **Karpowicz 2025 (T11).** For the memory-guided reaching task, successful non-calibration trials are segmented into short ($\leq$ 600 ms) outward reaches using a strict recentre-jump detector. This yields 119 reaches from 1,116 original trials, which are split 80/20 under a fixed seed and stratified by day and target (95/24). **Singer-Clark cursor (T15).** For centre-out and grid cursor control, centre-out trials are divided into outward and return sub-reaches. We then apply an 80/20 fixed-seed split stratified by calibration flag and recording day (1,740/436 trials).

A.2 Architecture ablations and frontend comparison

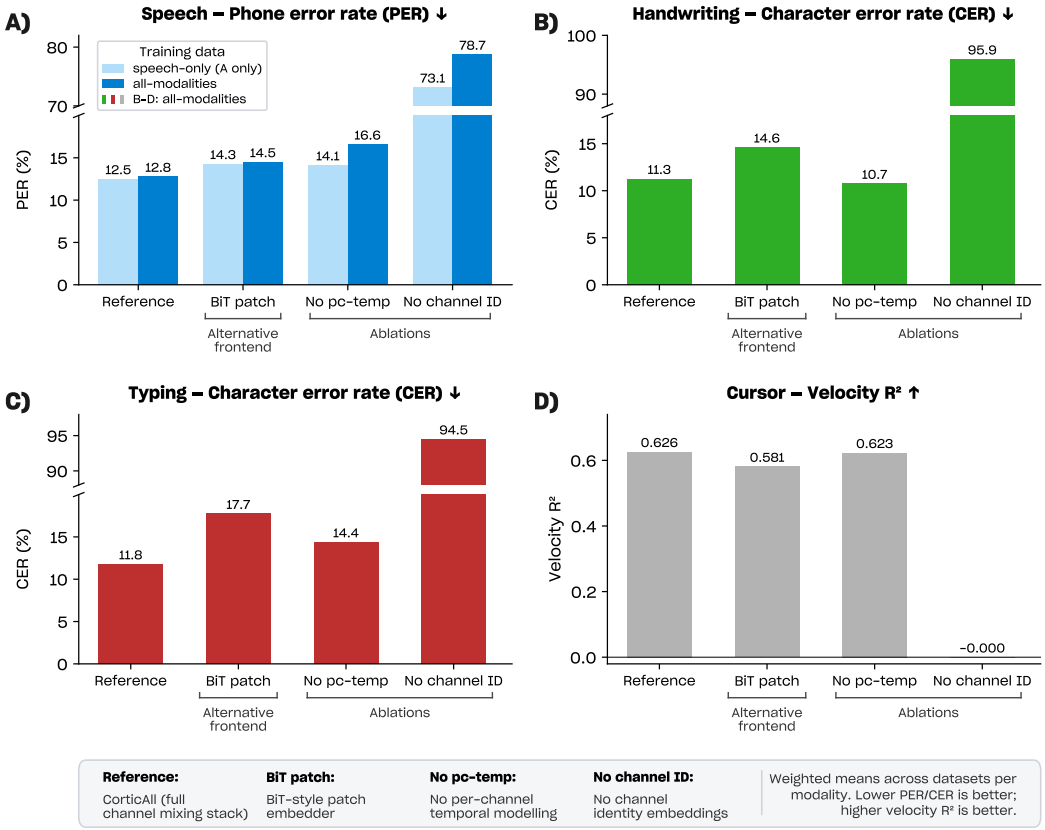

Figure 5: **CorticAll architecture ablations and frontend comparison.** We compare the reference architecture with a BiT-style dataset-specific linear patch embedder and two ablations: removal of per-channel temporal modelling (**No pc-temp**) and removal of channel identity embeddings (**No channel ID**). Bars report weighted mean performance across datasets within each modality from a single training seed; speech additionally includes a speech-only training condition. Lower PER and CER indicate better performance, whereas higher velocity R^2 is better.

CorticAll is designed to operate on heterogeneous intracortical recordings while sharing as much of the model as possible across datasets, participants, and modalities. Its channel-mixing stack generalizes the dataset-specific feature extractors used in prior work into a shared module reused across datasets. Persistent channel identity is central to this formulation, while processing channels separately before mixing also enables temporal modelling within each channel, in addition to the

temporal modelling applied after channels are combined. We therefore ablate both channel identity and this per-channel temporal modelling stage, and compare the shared frontend with the dataset-specific linear patch embedder used by the BiT family of models (Zhang et al., 2025; Feghhi et al., 2025). Figure 5 reports weighted mean performance across datasets within each modality from single-seed models trained from-scratch, jointly across all modalities; for speech, we additionally report speech-only training.

Removing **channel identity embeddings** causes decoding performance to collapse across all four modalities. Although CorticAll supports variable channel configurations, the channels therefore cannot be treated as exchangeable: the model requires persistent information about which channel generated each neural feature. For the discrete decoding tasks, the residual non-zero performance reflects degenerate predictions dominated by blanks and a small number of high-frequency tokens (speech: SIL/AH/DH; typing: I; handwriting: E/T), rather than meaningful decoding. The same behaviour is observed for speech under both speech-only and all-modality training.

Removing **per-channel temporal modelling** has a more modality-dependent effect. It substantially degrades speech decoding, particularly under all-modality training, and also worsens typing performance, while the changes for handwriting and cursor decoding are comparatively small in this single-seed comparison. This suggests that modelling temporal structure within individual channels provides an important additional degree of freedom for some communication behaviours, beyond temporal modelling performed only after channels have been combined.

Finally, replacing the shared channel-mixing stack with the **BiT-style linear patch embedder** consistently degrades performance across all four modalities, for speech under both speech-only and all-modality training. This choice of the front-end also governs how the models scale: each BiT projection costs about $3{,}000$ parameters per electrode, totalling 10.5M (46% of the 23.0M all-modality model) and growing with every dataset added, while CorticAll encodes dataset identity through additive channel embeddings at 128 parameters per electrode ($\sim 24\times$ fewer), keeping its dataset-specific footprint essentially flat (0.33M, $\sim 2\%$ of its 14.3M). At the current 16 dataset-participant frontend instantiations the two are already 23.0M versus 14.3M, and the gap widens with every participant added. Sharing the frontend thus improves accuracy and parameter efficiency at once.

A.3 Cross-participant subspace sharing

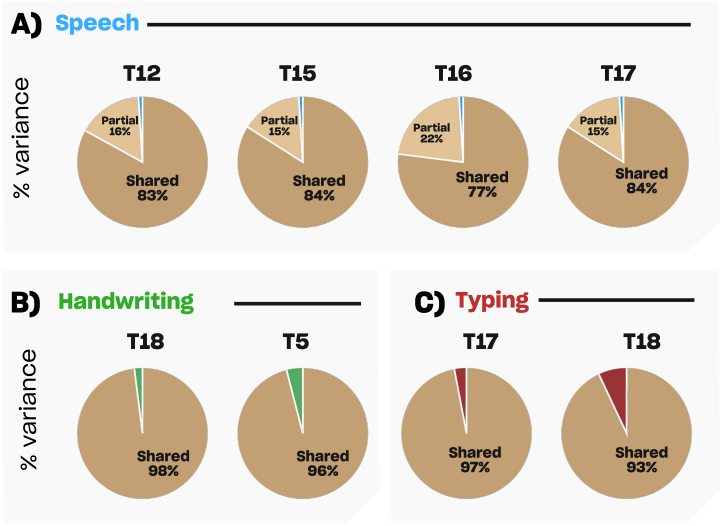

Figure 6: **Latent subspaces are largely shared across participants within each modality.** DySO is applied across participants separately for speech, handwriting, and typing. Pie charts show the fraction of each participant's latent variance lying in dimensions shared across all participants, shared with only a subset (*partial*), or unique to that participant.

We further apply DySO across participants within each modality to ask whether the shared model forms participant-specific or common latent subspaces. Speech, which contains four participants, shows predominantly shared variance across all participants (77–84%), with most of the remainder lying in dimensions shared by only a subset of participants. Handwriting and typing, each represented by two participants, show even stronger overlap, with 96–98% and 93–97% of variance shared, respectively (Figure 6). Thus, within a fixed behaviour, CorticAll's latent representation is largely shared across participants rather than partitioned into participant-specific subspaces.

A.4 Qualitative decoding examples

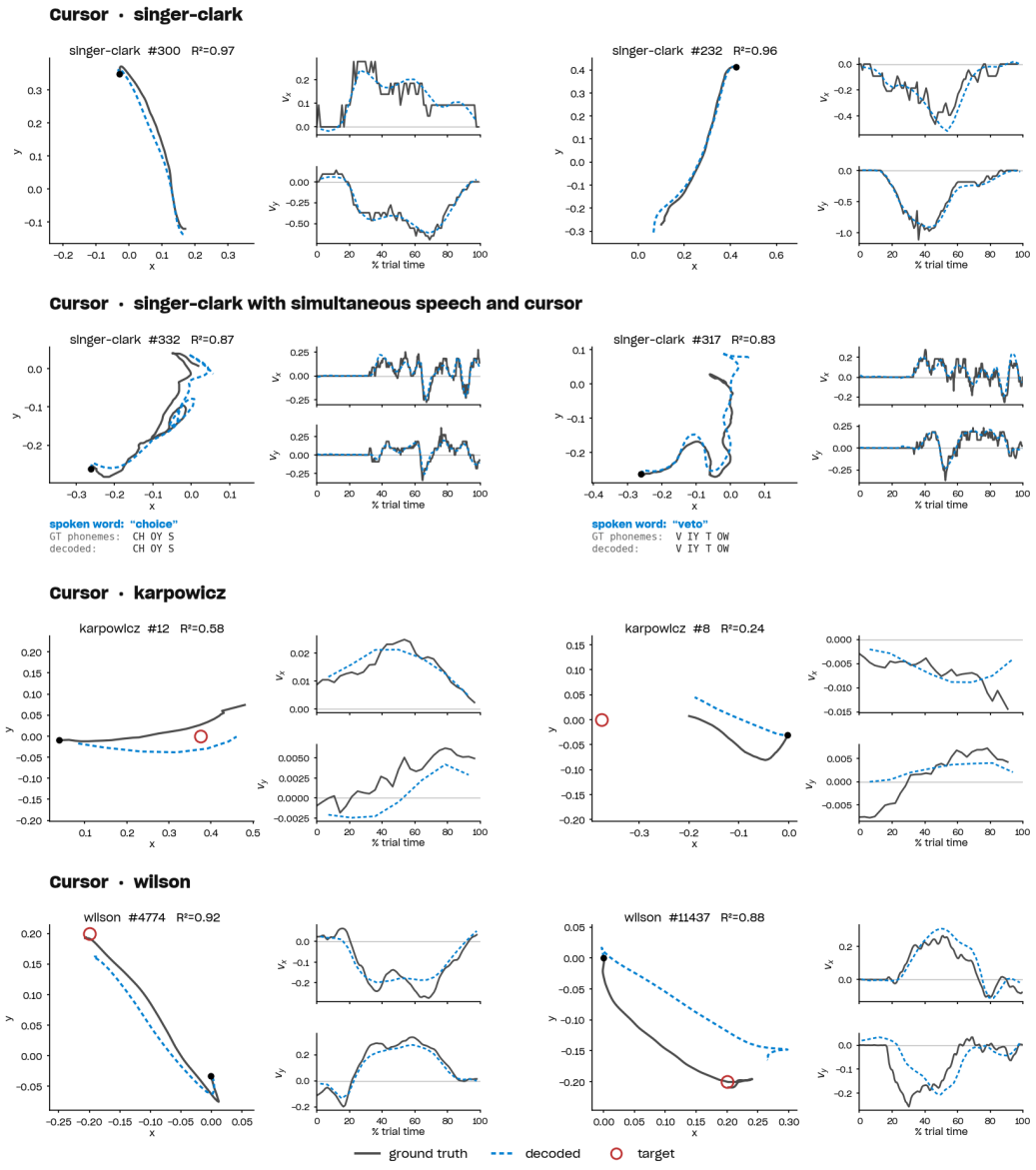

Figure 7: **Qualitative cursor-decoding examples.** Each trial shows the 2-D trajectory (left; position integrated from decoded velocity) and the decoded v_x, v_y traces over normalised trial time (right), annotated with velocity R^2. Ground truth is solid grey, decoded output dashed blue, target an open red circle. The second row shows singer-clark trials where the same model decodes cursor and speech simultaneously (phonemes shown beneath).

Speech

card (n=1426)

best	PER 0.00	**He does the yard.**
	GT	HH IY D AH Z DH AH Y AA R D
	dec	HH IY D AH Z DH AH Y AA R D
median	PER 0.04	**Were quite vocal about it.**
	GT	W ER K W AY T V OW K AH L AH B AW T IH T
	dec	W ER K W AY T F OW K AH L AH B AW T IH T
worst	PER 0.63	**Zoology department.**
	GT	Z OW AA L AH JH IY D IH P AA R T M AH N T
	dec	DH AH AA L AH T IY T AH P AY D

wairagkar (random-test) (n=90)

best	PER 0.00	**What is under this house?**
	GT	W AH T IH Z AH N D ER DH IH S HH AW S
	dec	W AH T IH Z AH N D ER DH IH S HH AW S
median	PER 0.00	**I don't mind it at all.**
	GT	AY D OW N T M AY N D IH T AE T AO L
	dec	AY D OW N T M AY N D IH T AE T AO L
worst	PER 0.18	**We will all miss you.**
	GT	W IY W IH L AO L M IH S Y UW
	dec	W IY W IH L AO L B EY K S Y UW

kunz T12 (50-word) (n=100)

best	PER 0.00	**Clean up.**
	GT	K L IY N AH P
	dec	K L IY N AH P
median	PER 0.17	**I have it up.**
	GT	AY HH AE V IH T AH P
	dec	AY HH AE V IH T HH OW P
worst	PER 0.33	**It is very bad.**
	GT	IH T IH Z V EH R IY B AE D
	dec	IH T IH Z V EH R IY K AH M IH NG

kunz T16 (50-word) (n=50)

best	PER 0.00	**It is okay**
	GT	IH T IH Z OW K EY
	dec	IH T IH Z OW K EY
median	PER 0.22	**I am okay**
	GT	AY AE M OW K EY
	dec	AY AE M AH K IY
worst	PER 0.80	**Bring my glasses please**
	GT	B R IH NG M AY G L AE S AH Z P L IY Z
	dec	D UW HH OW P DH AE HH AA

willett (n=880)

best	PER 0.00	**Just way in the back.**
	GT	JH AH S T W EY IH N DH AH B AE K
	dec	JH AH S T W EY IH N DH AH B AE K
median	PER 0.12	**Have a good evening.**
	GT	HH AE V AH G UH D IY V N IH NG
	dec	HH AE V AH G UH D L AE V N IH NG
worst	PER 0.75	**Were thoroughbreds.**
	GT	W ER TH ER OW B R EH D Z
	dec	W IY R S AW T AH B AE D

jude (speech) (n=54)

best	PER 0.00	**They're so small.**
	GT	DH EH R S OW S M AO L
	dec	DH EH R S OW S M AO L
median	PER 0.22	**Prisons all my life.**
	GT	P R IH Z AH N Z AO L M AY L AY F
	dec	P R ER S AH AO L M AY L AY F
worst	PER 0.57	**It's such a trouble here.**
	GT	IH T S S AH CH AH T R AH B AH L HH IY R
	dec	IH T S T AA N AA JH AH N T R AE B IH R

kunz T15 (50-word) (n=50)

best	PER 0.00	**Where is it?**
	GT	W EH R IH Z IH T
	dec	W EH R IH Z IH T
median	PER 0.08	**How do you feel?**
	GT	HH AW D UW Y UW F IY L
	dec	HH AW D UW Y UW HH IY L
worst	PER 0.53	**It is comfortable**
	GT	IH T IH Z K AH M F ER T AH B AH L
	dec	IH T IH Z HH AH NG G R IY

kunz T17 (n=196)

best	PER 0.00	**though**
	GT	DH OW
	dec	DH OW
median	PER 0.00	**though**
	GT	DH OW
	dec	DH OW
worst	PER 1.00	**were**
	GT	W ER
	dec	CH OY S

Handwriting

handwriting (n=228)

best	CER 0.00	**it's quite a wide brief, actually**
	dec	it's quite a wide brief, actually
median	CER 0.06	**that's when i threw up on the carpet**
	dec	that's whan i thew up on the carpet
worst	CER 0.28	**it's probably a road accident christine**
	dec	it's probably a road acsident

jude T18 (handwriting) (n=24)

best	CER 0.00	**they'll trample it down.**
	dec	they'll trample it down.
median	CER 0.04	**her optimism gave me heart.**
	dec	her aptimism gave me heart.
worst	CER 0.22	**they were pursuing him.**
	dec	they luvere pursuing bi.

fan (handwriting) (n=199)

best	CER 0.00	**revulsion in the desert**
	dec	revulsion in the desert
median	CER 0.16	**camp meals are no great problem**
	dec	anp meals are no shea problem
worst	CER 0.41	**he had become king at fifteen**
	dec	he a blome kins dififak

Typing

jude T17 (typing) (n=150)

best	CER 0.00	**It's show time.**
	dec	Its show time.
median	CER 0.18	**What do you mean?**
	dec	What doo you oman?
worst	CER 0.91	**It's just a.**
	dec	Hukisgit

jude T18 (typing) (n=176)

best	CER 0.00	**An environmental engineer.**
	dec	An environmental engineer.
median	CER 0.03	**I heard there was some stuff.**
	dec	I heard there was some stiff.
worst	CER 0.70	**It is not.**
	dec	Nonno it is not

best / median / worst: trials from the best 20%, the median, and the worst 20% of each dataset's per-trial error (PER for speech, CER for handwriting / typing)
GT reference (speech phonemes) **dec** decoded output red mismatch vs reference

Figure 8: **Qualitative decoding examples for the discrete modalities.** Best-20%, median, and worst-20% trials per dataset by per-trial error (PER speech, CER handwriting/typing; n = val trials). GT reference, dec greedy output (no LM), red = mismatch.

A.5 Architecture details

Table 4 lists the full hyperparameters of CorticAll and Table 5 its parameter budget; here we record only the details not already given by the figure, the main text, or those tables.

Training-time details. Datasets that provide only one of the two per-channel features have the missing feature replaced by a learned mask token, so a single front-end consumes both single- and dual-feature recordings unchanged. During training the tokeniser inputs are augmented with additive Gaussian noise and a random baseline shift (one constant per feature, added across all channels and timesteps of the trial). The per-channel temporal block is applied as a gated residual with LayerScale, initialised near-identity so it is engaged gradually over training.

Table 4: Architecture hyperparameters of CorticAll.

Component / hyperparameter	**Value**
Input & Channel Tokeniser	
Input features per channel	spike count, spike-band power (2); + binary presence flag ($\Rightarrow$ 4 conv channels)
Gaussian smoothing σ	2.0 bins, 20-tap kernel (fixed)
Conv stack	2 $\times$ Conv1d (stride 2, kernel 5, 64 hidden channels) + GELU; GroupNorm(1) then output LayerNorm
Total temporal stride	4 (one token / 4 bins, 80 ms)
Per-channel width d_{chan}	128
Channel Identity & per-channel temporal contextualiser	
Channel Identity embedding	per (dataset, channel); dim 128 ($= d_{\text{chan}}$)
Temporal contextualiser	depthwise temporal conv (ConvNeXt-style): pointwise $\times 2$ + depthwise k=127 + FFN ($\times 4$), gated residual + LayerScale
Dropout	0.1
Channel Mixer (global–local cross-attention)	
Channel encoder	3 layers, 4 heads
Learned queries	8
Readout attention	4 heads, dim/head 32
Mixer dropout	0.2
Fused token width d_{model}	384
Temporal Transformer	
Layers / heads / dim-per-head	7 / 6 / 64
MLP ratio	4
Dropout	0.3
Positional encoding	relative, learnable (per-layer bias)
Input dropout	0.2
Decoding heads (linear)	
Speech / handwriting / typing heads	41 / 32 / 31 classes (incl. CTC blank)
Cursor	2 (2-D velocity)

Parameter budget. Table 5 breaks down the $14{,}315{,}837$ parameters of the model. The Temporal Transformer dominates ($\approx 87\%$), while the entire spatial front-end (tokeniser, identity embeddings, per-channel temporal block and channel mixer) accounts for only $\approx 13\%$ and the task heads for the remainder. The backbone is fixed regardless of how many datasets are used: the *only* component that grows with the data is the channel-identity embedding, which adds exactly 128 parameters per electrode (from $8{,}192$ for a 64-channel array to $49{,}152$ for a 384-channel array, ≈ 0.33M across the 12 current arrays). Adding a new dataset therefore introduces just its own identity block and shares everything else.

A.6 Training and optimisation details

This appendix expands on Section 2.3. Shared optimisation, loss, and augmentation settings are in Table 6; the per-regime pool, initialisation, and the few settings that differ across regimes are in Table 7. The five regimes themselves are defined in Section 2.3.

Component	Params	%
Temporal transformer (7 layers, d=384, 6 heads, MLP×4)	12.42 M	86.7
Channel mixer / neural tokenizer	1.81 M	12.7
Global–local aggregator	1.24 M	8.7
Channel-identity embeddings (all datasets)	0.33 M	2.3
PC-temporal conv + FFN	0.23 M	1.6
Input projection / norm / mask tokens	0.02 M	0.1
Conv encoder (feature extractor)	0.04 M	0.3
Task heads (CTC speech/hw/typing + motor)	0.04 M	0.3
Total	**14.32 M**	**100**

Table 5: CorticAll parameter budget: 14,315,837 parameters in total, with percentages taken of the total.

Optimisation, loss balancing, and compute. All models are optimised with AdamW under linear warmup and cosine decay (Table 6). When several modalities are trained jointly, per-batch gradient magnitudes swing by more than an order of magnitude depending on the modality mix, which destabilises training in two ways that we address together. First, each head's loss is divided by an exponential moving average of its own recent magnitude (decay 0.99, roughly a 100-step horizon) before its fixed weight is applied, so no single modality dominates the total purely through loss scale. Second, we shorten AdamW's second-moment horizon by lowering β_2 from 0.999 to 0.95 for high-variance cursor runs, so a single gradient spike no longer inflates the update for the following $\sim$1000 steps. All runs use bf16 mixed precision and train on a single NVIDIA H100, with no multi-GPU parallelism. Across the three seeds, the reported training runs total $\approx$3,730 H100-hours ($\approx$156 H100-days). The leave-one-out regime is by far the most expensive per experiment: for a single seed its speech pretraining runs alone cost $\approx$400 H100-hours, plus $\approx$77 H100-hours for the held-out fine-tunes. For this reason we ran leave-one-out for speech datasets and a single seed only, while the other four regimes were repeated over all three seeds.

Augmentation. Beyond the additive noise and per-trial offset listed in Table 6, we apply time masking of (Feghhi et al., 2025) during training: up to ten contiguous spans are masked until 15% of a trial's valid frames are covered, and the masked positions (at the input to the Temporal Transformer) are replaced by a learned mask token, so the model must infer them from temporal context.

Cursor velocity normalisation. Because the cursor datasets are recorded in different native units and span very different velocity scales, we z-score each cursor dataset's 2-D velocity targets before computing the Huber loss, so that the regression objective is comparable across datasets and no single dataset dominates the motor head.

Reuse of channel-identity embeddings. When several datasets originate from the same participant, we can optionally tie their channel-identity embeddings, signalling to the model that these recordings share a physical substrate even though they are separate datasets. We found this tying to be beneficial only when a set of conditions hold jointly, and detrimental otherwise: (i) the datasets are collected relatively close in time, since chronic intracortical recordings drift over days and weeks (Perge et al., 2013; Jarosiewicz et al., 2015), so a shared identity becomes inconsistent as the signal-to-behaviour relationship shifts; (ii) they engage the same cortical territory, since speech tends to recruit ventral premotor cortex (area 6v) whereas handwriting and typing typically recruit the dorsal hand-knob (area 6d), so tying identities across these modalities would force the channel mixer to find a compromise between the importance of a given channel/region when determining the collapsing formula; and (iii) the electrode-to-electrode correspondence across datasets is known, i.e. channel i refers to the same electrode in both. Under these criteria we tie embeddings for T15 in speech (*card* and *wairagkar*; Kunz T15 is excluded due to separation in time), T17 in speech (*jude* and *kunz_t17*), T5 in handwriting (*willet* and *fan*), and T18 across typing and handwriting (both collected in the same period and from the same cortical region). We stress that these are observations gathered during model development rather than rigorously ablated findings. For the leave-one-out pretraining and fine-tuning we apply *no* embedding sharing: even when the same participant appears

in the pretraining pool (as several speech datasets do), the held-out dataset is given a distinct identity, which partially isolates the transfer effect from mere participant re-identification.

Checkpoint selection. Fine-tuning and transfer runs are initialised from the pretraining checkpoint with the best per-modality weighted-mean validation metric. For evaluation, single-dataset and target-specific fine-tuning runs use the best checkpoint for the target dataset, whereas multi-dataset from-scratch runs use one checkpoint per modality, selected by the modality-level weighted validation metric and shared across all datasets of that modality (Section 2.3).

Hyperparameter optimisation. We performed no systematic hyperparameter search: there were no sweeps, and only minimal manual tuning during development. The settings in Tables 6 and 7 are therefore lightly-tuned defaults rather than an optimised configuration, and we expect headroom from more careful tuning.

Table 6: Training, optimisation, and augmentation hyperparameters shared across regimes. Per-regime settings (epochs, β_2, gradient clip, loss normalisation, λ_{speech}) are given in Table 7.

Hyperparameter	**Value**
Optimisation	
Optimiser	AdamW
β_1	0.9
Weight decay	10^{-4}
Base learning rate	3×10^{-4}
Warmup	500 steps, linear
LR schedule	cosine decay to 1% floor
Precision	bf16 mixed
Parallelism	single H100 per run
Evaluated seeds	42, 2, 3
Loss	
Speech / handwriting / typing	CTC (blank $= 0$, zero-infinity)
Cursor	Huber ($\delta = 1.0$), masked to valid bins
Loss normalisation	EMA, decay 0.99 (full-multimodal only)
$\lambda_{\text{hw/typing/cursor}}$	1.0
Augmentation (training only)	
Additive noise std	0.2
Per-trial offset std	0.05
Time masking	15% of frames

Table 7: Per-regime training configuration. Dashes denote settings that do not apply (loss normalisation is only needed when modalities are mixed). †Cursor from-scratch uses β_2=0.99. ‡Cursor within-modality uses β_2=0.95 and gradient clip 0.5.

Regime	**Training pool**	**Init**	**Epochs**	β_2	**Gradient clip**	**Loss norm**	λ_{speech}
1. Single-dataset	one dataset	scratch	800	0.999†	0.7	–	1.0
2. Within-modality	all datasets, one modality	scratch	800	0.999‡	0.7‡	–	1.0
3. Full multimodal	all datasets, all modalities	scratch	800	0.999	0.5	✓	2.0
4. Target fine-tune	one target (present in pool)	ckpt (2 or 3)	400–800	0.999	0.7	–	1.0
5. Leave-one-out transfer	modality $\setminus$ target, then target	scratch $\rightarrow$ ckpt	800 $\rightarrow$ 400	0.999	0.7	–	1.0