FILM: A Dataset of Flexible-Electrode Intracortical Recordings for Longitudinal Motor Decoding

Anonymous authors
Paper under double-blind review

Abstract

Reliable long-term use of intracortical brain–computer interfaces (iBCIs) depends on sustained recording quality and decoding performance across days. Flexible intracortical electrodes support chronic recordings, but the long-term stability of motor decoding from these signals remains insufficiently characterized. To address this gap, we introduce FILM (Flexible-Electrode Intracortical Recordings for Longitudinal Motor Decoding), a longitudinal dataset for evaluating motor decoding from chronic flexible-electrode recordings. FILM pairs neural features derived from 512-channel recordings with synchronized two-dimensional cursor kinematics from two non-human primates performing Center-Out and Random Target reaching tasks. FILM comprises 178 sessions organized into three subject–task subsets spanning up to 277 days. Using FILM, we characterize longitudinal behavioral and neural changes, evaluate within-day and cross-day decoding performance, and assess target-day recalibration. Mean velocity profiles for each movement direction show substantial similarity across days, while movement speed, trial duration, neural activity, and estimated velocity tuning vary. Within-day decoding performance remains generally high, whereas cross-day decoding performance without recalibration tends to decline as the interval between training and testing days increases. Recalibration with limited target-day data improves cross-day decoding. Compared with matched recalibration without historical weights, incorporating these weights improves mean performance at the smallest calibration budget, with gains diminishing as more target-day data become available. Retrospective analyses further associate larger source–target weight mismatch with smaller or negative gains. FILM provides a resource for studying longitudinal neural distribution shifts and data-efficient adaptation in chronic iBCIs.

1 Introduction

Intracortical brain–computer interfaces (iBCIs) translate neural activity related to intended movement into control signals for assistive devices and have enabled cursor and robotic-limb control, restoration of functional movement through electrical stimulation, and computer access and communication through cursor-based typing and attempted handwriting (Hochberg et al., 2006; Collinger et al., 2013; Bouton et al., 2016; Pandarinath et al., 2017; Willett et al., 2021; Card et al., 2026). However, reliable long-term iBCI deployment depends not only on decoder design, but also on recording quality and the stability of neural–kinematic relationships (Chestek et al., 2011; Pun et al., 2024; Wilson et al., 2026). Motor iBCI studies have traditionally relied on rigid penetrating microelectrode arrays, exemplified by the Utah array, for chronic intracortical recording (Maynard et al., 1997; Hochberg et al., 2006; Sponheim et al., 2021). Flexible intracortical electrodes, which reduce the mechanical mismatch between implanted devices and brain tissue, offer a promising alternative for chronic neural interfaces (Luan et al., 2017; Wang et al., 2023b).

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.

Flexible intracortical interfaces have progressed from months-scale, high-density recordings in rodents to chronic recording and motor-decoding experiments in non-human primates (NHPs). Early ultraflexible arrays enabled long-term high-density neural recording in rodents (Zhao et al., 2023b). Flexible multichannel electrodes subsequently established the feasibility of intracortical recording in behaving NHPs, while mechanically robust ultraflexible electrode arrays enabled months-scale chronic recording and online cursor control (Wang et al., 2023a; Tian et al., 2023). More recently, flexible-electrode technologies have been scaled to hundreds or thousands of recording sites and used for continuous upper-limb kinematic decoding (Fang et al., 2026).

Despite this progress, chronic recording capability does not by itself establish long-term decoding stability. Existing flexible-electrode studies have primarily characterized electrode design, recording longevity, neural signal quality, and the feasibility of motor decoding. Whether neural–kinematic mappings learned on one recording day remain usable on subsequent days has received considerably less attention. Because intracortical recordings and their relationships to behavior can change over time, a decoder trained on one day may progressively degrade when transferred across recording days, increasing the need for recalibration (Chestek et al., 2011; Pun et al., 2024). Systematic cross-day evaluation is therefore necessary to distinguish persistent recording capability from reliable long-term motor decoding.

Recent datasets and benchmarks have progressively extended standardized neural-decoding evaluation from within-session population modeling to cross-day generalization and longitudinal neural variability. The Neural Latents Benchmark (NLB) established standardized evaluation across intracortical datasets (Pei et al., 2022), while FALCON evaluated zero-shot cross-day generalization and few-shot recalibration (Karpowicz et al., 2024b). The LINK dataset further provided long-term NHP intracortical recordings and kinematics for investigating longitudinal changes in neural activity and decoding performance (Temmar et al., 2026). However, existing resources have not yet covered long-term cross-day motor decoding evaluation using chronic flexible-electrode intracortical recordings, leaving the decoding stability of flexible neural interfaces insufficiently characterized.

To address this gap, we introduce FILM (**F**lexible-Electrode **I**ntracortical Recordings for **L**ongitudinal **M**otor Decoding), a longitudinal dataset pairing neural features derived from 512-channel flexible-electrode recordings and synchronized two-dimensional cursor kinematics from two NHPs performing Center-Out and Random Target reaching tasks. FILM comprises 178 sessions across three subject–task subsets, the longest spanning 277 days, and provides two neural feature types: threshold-crossing rate (TCR) and spiking-band power (SBP). Using FILM, we characterize longitudinal changes in behavior, neural activity, and estimated velocity tuning and evaluate decoder transfer and recalibration across recording days. Decoders trained on one set of trials generally predict cursor velocity well on held-out trials from the same session; transfer to another day without recalibration is less reliable, particularly over longer intervals. Recalibration with limited target-day data improves cross-day decoding. In matched comparisons, incorporating historical decoder weights yields higher mean performance at the smallest calibration budget than recalibration without those weights, but this advantage narrows as more target-day data become available. Retrospective analyses further associate larger source–target weight mismatch with smaller or negative gains. FILM provides a benchmark for studying longitudinal neural changes and evaluating robust decoding, domain adaptation, and data-efficient recalibration methods under chronic neural non-stationarity.

2 Related Work

Standardized neural datasets and benchmarks have become important resources for neural representation learning and motor decoding research. The Neural Latents Benchmark (NLB) established a unified framework for neural population modeling and decoding evaluation across intracortical datasets (Pei et al., 2022). FALCON further focused on cross-day decoder robustness by benchmarking zero-shot generalization and few-shot recalibration (Karpowicz et al., 2024b). The LINK dataset provided long-term NHP intracortical recordings with kinematics for studying longitudinal changes in neural activity and decoding per-

formance (Temmar et al., 2026). However, existing longitudinal neural datasets have not specifically addressed long-term motor decoding evaluation under chronic flexible-electrode intracortical recordings.

Flexible intracortical interfaces have been developed to improve chronic neural recording capability, increase recording scale, and enable long-term neural recordings from rodents to NHPs. Early studies using ultraflexible electrode arrays achieved months-scale, high-density neural recordings in rodents (Zhao et al., 2023b). Subsequent flexible neural interfaces, including mesh electronics and ultra-flexible tentacle electrodes (UFTEs), demonstrated long-term tracking of individual neurons and neuronal ensembles (Zhao et al., 2023a; Yasar et al., 2024). Multidirectional ultraflexible probes further introduced self-assembled three-dimensional, ring-like electrode configurations for long-term neural recordings across distributed brain regions (Yang et al., 2023). With the extension of flexible electrode technologies to NHPs, mechanically robust ultraflexible electrode arrays (MERF) enabled months-scale chronic intracortical recording and online cursor control (Tian et al., 2023). More recently, kirigami-inspired microelectrode arrays (ki-MEAs) expanded flexible-electrode intracortical recordings to hundreds or thousands of recording sites and enabled continuous upper-limb kinematic decoding (Fang et al., 2026). However, existing studies have primarily focused on electrode design, chronic recording capability, and signal quality, while long-term motor decoding evaluation under flexible-electrode intracortical recordings remains limited.

Spike-sorted single-unit activity remains important for cell-resolved analyses of neuronal tuning and longitudinal stability, particularly when neuron-specific changes must be distinguished from changes in population-level dynamics (Gallego et al., 2020; Jensen et al., 2022). For iBCI decoding, however, stable isolation of individual units is often unnecessary: unsorted threshold-crossing activity can preserve task-relevant information and population structure while achieving decoding performance comparable to that of spike-sorted activity (Todorova et al., 2014; Perge et al., 2014; Christie et al., 2014; Perel et al., 2015; Dai et al., 2019; Trautmann et al., 2019). Accordingly, threshold-crossing activity has been widely used as a decoding feature in iBCIs, supporting cursor control, typing, attempted-handwriting communication, bimanual control, and long-term operation (Gilja et al., 2015; Jarosiewicz et al., 2015; Pandarinath et al., 2017; Brandman et al., 2018; Simeral et al., 2021; Willett et al., 2021; Fan et al., 2023; Karpowicz et al., 2025; Wilson et al., 2026). Spiking-band power (SBP) provides a continuous-valued representation of high-frequency spiking-related activity that complements discrete threshold-crossing features and has been used either alone or together with threshold crossings in real-time motor and communication iBCIs (Nason et al., 2020; Willsey et al., 2022; Willett et al., 2023; Ali et al., 2024; Karpowicz et al., 2024a; Willsey et al., 2025; Card et al., 2026). Reflecting the complementary roles of these sorting-free features, FILM provides both SBP and TCR as preprocessed neural features for longitudinal motor decoding.

3 The FILM Dataset

FILM provides longitudinal intracortical neural features and synchronized two-dimensional cursor kinematics from two non-human primates performing joystick-controlled reaching tasks. The dataset contains three subject–task subsets, MA_CO, MA_RT, and MB_CO, with 512 acquisition channels per subject and recording spans of up to 277 days. Fig. 1 summarizes the recording setup, neural feature extraction, an example aligned trial, data distributions, and temporal coverage.

3.1 Experimental Setup and Recording Configurations

Subjects and Recording Configuration. FILM contains intracortical recordings from two non-human primates, denoted MA and MB. Each subject was implanted with four 128-channel flexible arrays, providing 512 acquisition channels. The arrays used for MA and MB followed the designs of Yang et al. (2023) and Fang et al. (2026), respectively. All four MA arrays were implanted in dorsal premotor cortex (PMd), with the same configuration for CO and RT recordings. For MB_CO, two arrays were in PMd, one in ventral premotor cortex (PMv), and one in primary motor cortex (M1). Each array includes 124 electrode

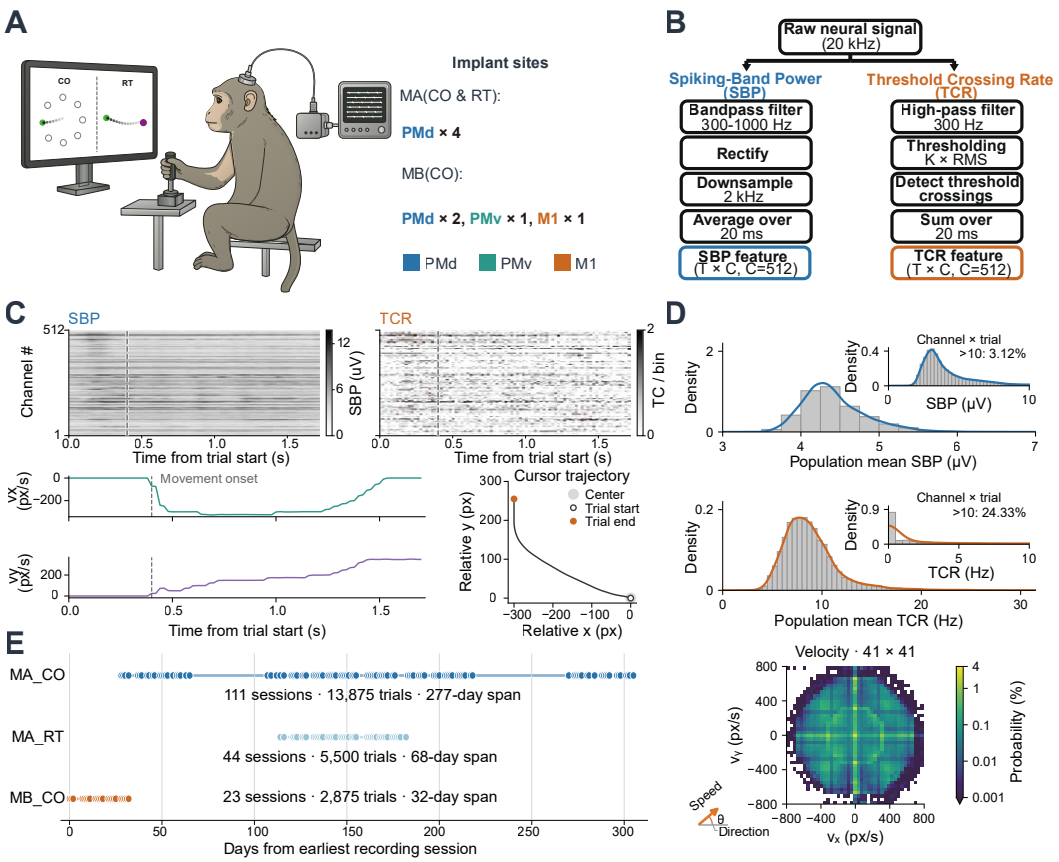

Figure 1: **Overview of the FILM dataset.** (A) Joystick-controlled Center-Out (CO) and Random Target (RT) tasks and flexible-array implantation configurations. (B) SBP and TCR extraction from 20-kHz recordings, with 20-ms output bins. TCR uses the signed threshold coefficient $K = -5$. (C) Example trial with 512-channel SBP (μV), threshold-crossing counts per 20-ms bin, cursor velocity, and cursor trajectory. Time is relative to the first valid trial sample; dashed lines mark movement onset. (D) Neural and velocity distributions over common post-onset trial windows in MA_CO. Main neural panels show trial-level population means; insets show channel-specific trial means, including the reference channels that pass quality control. Histograms and weighted Gaussian kernel density estimates give sessions equal weight, trials equal weight within sessions, and channels equal weight in the insets. The lower panel shows a 41×41 velocity histogram, normalized within each session and averaged equally across sessions. Color gives probability per cell (%) on a logarithmic scale; white cells have zero probability. The arrow illustrates speed and direction. (E) Session coverage on a common calendar axis, relative to the earliest recording in the collection. Annotations give session counts, trial counts, and elapsed spans for each subset.

sites and four reference channels. The 512-channel feature arrays preserve this acquisition layout, and analyses include all channels that meet their stated validity criteria. Fig. 1A and Appendix A.2 provide the implant configuration and channel maps.

Behavioral Paradigms. Subjects used a joystick to move a two-dimensional cursor toward visual targets. MA performed Center-Out (CO) and Random Target (RT) tasks, whereas MB contributed CO recordings. CO trials started at the workspace center and ended at one of eight peripheral targets. RT movements started at the cursor's current position and ended at a newly presented target. In both tasks, acquiring the target within the allotted time produced a successful trial and juice reward. Neural activity and cursor kinematics were recorded synchronously. Task procedures are described in Appendix A.

3.2 Neural and Behavioral Data Processing

Neural Feature Extraction. Neural signals were acquired at 20 kHz using a SpikeGadgets [1] system and converted to spiking-band power (SBP) and threshold-crossing rate (TCR) features at 20-ms resolution (Fig. 1B). TCR detects negative crossings of the channel-specific threshold $\theta_c = KV_{\mathrm{RMS},c}$, where $V_{\mathrm{RMS},c}$ is estimated separately for each session and the signed default coefficient is $K = -5$. Crossings are stored as integer counts per bin; rates in Hz are obtained by dividing by 0.02 s. SBP summarizes rectified 300–1000 Hz voltage within the same bin duration, following the feature construction of Nason et al. (2020). Both representations retain 512 acquisition channels. Appendix B specifies the preprocessing, longitudinal RMS analysis, and ten-condition TCR threshold comparison.

Behavioral Processing and Alignment. Within each trial, TCR counts, SBP amplitudes, and cursor kinematics are aligned on a common 20-ms grid with absolute Unix timestamps in seconds. TCR counts are summed within each bin from 1-ms intervals. Cursor positions are interpolated at 1-ms resolution, then positions and derived velocities are averaged within each 20-ms bin. SBP samples are assigned to the same bins using their acquisition timestamps and averaged when coverage is complete. The streams share timestamps and are paired by row index; bins without complete SBP coverage are excluded from SBP analyses. Fig. 1C illustrates the aligned neural features, cursor velocities, and trajectory for an example trial. Fig. 1D shows the distributions of SBP amplitude, TCR rate, and two-dimensional cursor velocity over common post-onset trial windows in MA_CO. Further details of temporal binning, synchronization, and trial segmentation are provided in Appendix B.

3.3 Dataset Composition and Organization

FILM contains 178 sessions and 22,250 trials, with 125 retained trials per session. Sessions were selected using within-day WF and GRU decoding criteria before the longitudinal analyses (Appendix C); within-day benchmark scores therefore describe this selected collection. MA_CO contributes 111 sessions and 13,875 trials over 277 days; MA_RT contributes 44 sessions and 5,500 trials over 68 days; and MB_CO contributes 23 sessions and 2,875 trials over 32 days. These subsets contain 111, 44, and 23 distinct recording days, respectively. A span is the elapsed calendar time between the first and last recording dates, without adding one day. Fig. 1E places all three subsets on the same calendar axis.

Each session is stored in a Neurodata Without Borders (NWB, Rübel et al., 2022) file containing SBP, TCR at $K = -5$, cursor position and velocity, trial annotations, and session metadata. Retained trial segments are concatenated in chronological order while preserving their timestamps and original trial identifiers. The decoding benchmark uses the first 100 retained trials for training and the final 25 for testing. Descriptive analyses use the trial sets specified for each figure. Appendix C describes session curation and NWB data organization. Data and code access arrangements, together with loading procedures and validity checks, are provided in Appendix C.3.

4 Longitudinal Characterization and Motor Decoding

We first characterize behavioral and neural changes across recording days, then evaluate frozen cross-day decoding and the value of historical decoder weights for limited-data recalibration. The main figures use SBP from MA_CO, which contains 111 recording days over 277 calendar days. Complementary analyses cover SBP and TCR in all three subsets (Appendices F and G). Except for the threshold sweep in Appendix B.4, TCR analyses use features extracted with the channel-specific detection threshold $\theta_c = K * V_{\mathrm{RMS},c}$ ($K = -5$).

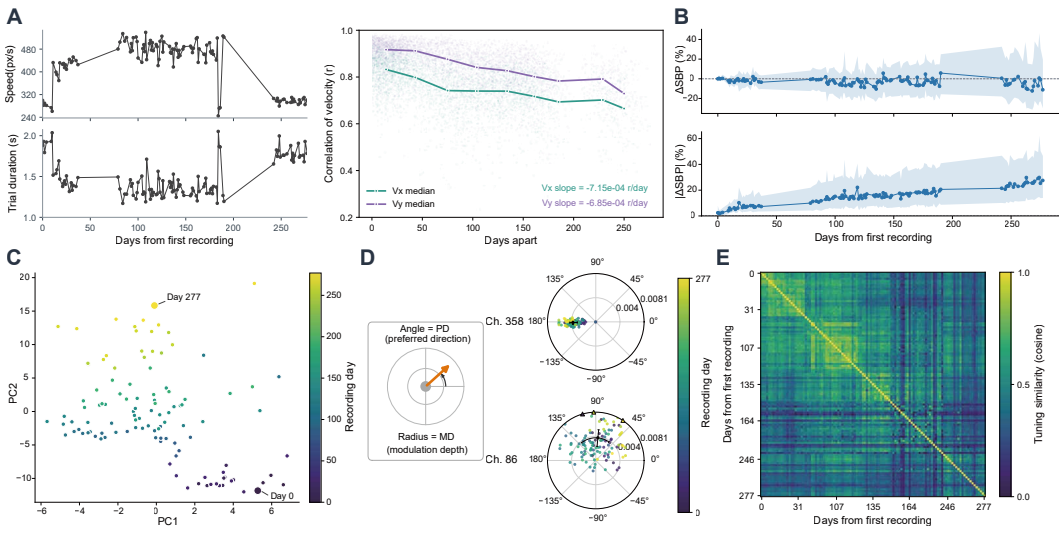

Figure 2: **Behavioral, neural, and velocity-tuning variation in** MA_CO**.** (A) Daily trial-averaged post-onset cursor speed and complete trial duration (left). Cross-day correlations of direction-conditioned, phase-normalized velocity profiles (right): points are day pairs, lines are 30-day-bin medians, and annotations give Theil–Sen slopes. (B) Signed and absolute channel-wise SBP changes relative to each channel's median daily amplitude over the first five eligible days. Lines and shading show medians and interquartile ranges (IQRs) across 512 channels. (C) Session centroids in a globally standardized, shared PCA space. (D) Preferred direction (PD; angle) and modulation depth (MD; radius) of the velocity encoding model for acquisition-channel indices 358 and 86. Colors denote recording day; black markers and bars summarize medians and IQRs, using circular summaries for PD. Triangles mark MD values outside the displayed radius. (E) Cosine similarity between population vectors of raw velocity coefficients on matched channels. Sessions are equally spaced chronologically, with elapsed-day tick labels. Methods are in Appendices D and F.

4.1 Behavioral and Neural Variability Across Recording Days

Behavioral Variability. We compare daily cursor speed and trial duration with direction-conditioned velocity-profile similarity (Fig. 2A). Movements in the middle recording period are generally faster and shorter than those in the late period. Horizontal and vertical velocity profiles remain correlated across days, although similarity tends to decrease with recording-day separation. Thus, similar velocity-profile shapes coexist with changes in movement speed and duration.

Neural Activity Variability. Relative to a fixed early SBP baseline, the median signed channel change stays near zero and is usually negative, while its spread and median absolute change grow across the recording span (Fig. 2B). A shared PCA projection of valid SBP samples places early and late session centroids in different regions (Fig. 2C). These views reveal channel and population variation alongside the behavioral changes.

4.2 Kinematic Tuning Across Recording Days

A linear model relates each channel's neural response to horizontal and vertical cursor velocity (Perel et al., 2015; Inoue et al., 2018). For SBP in MA_CO, trial-disjoint validation on the first 100 trials selects a lag of 200 ms (Appendix D.3; Fig. 8A). Preferred direction and modulation depth summarize the fitted velocity coefficients. The two example channels differ in the across-day dispersion of preferred direction, while both show changes in modulation depth (Fig. 2D).

[1] `https://spikegadgets.com`

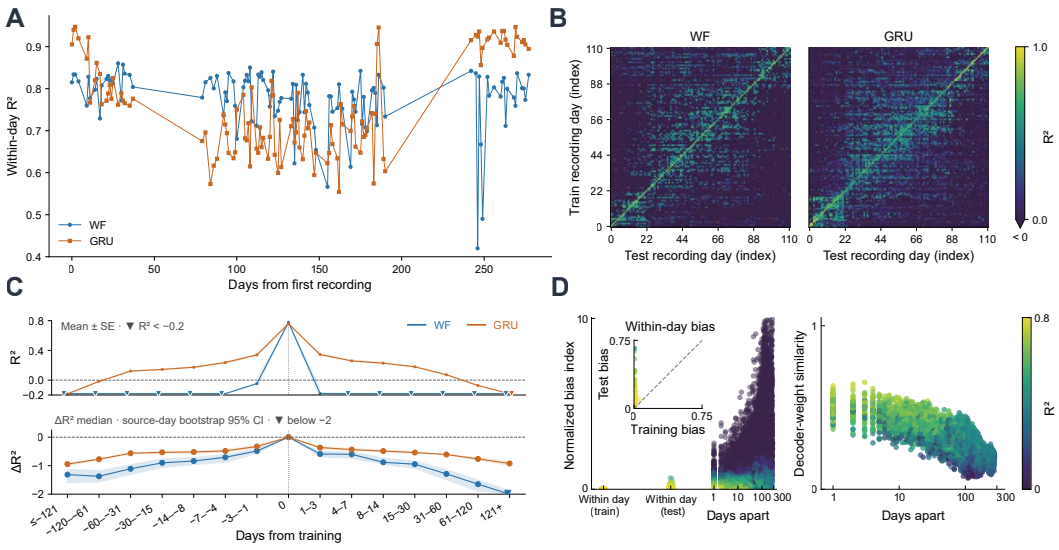

Figure 3: **Within-day decoding, frozen cross-day transfer, and decoder mismatch in MA_CO using SBP.** (A) Within-day held-out R^2 for Wiener filter (WF) and gated recurrent unit (GRU) decoders across 111 recording days. (B) Cross-day scores with source days in rows and target days in columns. Each source model and its preprocessing are fixed for its entire row, including the diagonal. Scores average horizontal and vertical velocity R^2; negative values saturate at the lower color limit. (C) Signed target-minus-source calendar-day separation. Upper: mean R^2 and descriptive pair-level SE. Lower: median cross-day R^2 minus the corresponding source-day diagonal score, with pointwise 95% source-day bootstrap intervals. Zero denotes within-day evaluation; boundary triangles mark values below the display limits. (D) Normalized output bias (left) and cosine similarity of WF weights fitted to normalized inputs (right). The left panel includes within-day controls; its inset compares training and test bias. Colors show frozen source-normalized R^2 on the left and target-normalized R^2 on the right. Methods are in Appendix E.

Cosine similarity of the raw velocity coefficients on the same valid channel indices is generally higher between nearby sessions than distant ones (Fig. 2E). Across-day similarity often falls below matched within-day split-half references, particularly in the MA subsets (Fig. 8E). Appendix D reports held-out encoding and permutation controls; Fig. 9D extends the tuning comparison to SBP and TCR in all subsets. The variation in estimated tuning motivates testing whether source-day decoders transfer across days.

4.3 Within-Day and Cross-Day Motor Decoding

Wiener filter (WF; Wiener, 1949) and gated recurrent unit (GRU; Chung et al., 2014) decoders predict two-dimensional cursor velocity from the first 100 trials of each session and are tested on its final 25 trials. Both retain substantial within-day performance across the recording span (Fig. 3A).

With source-day models and preprocessing held fixed, many cross-day transfers underperform the source-day diagonal (Fig. 3B). These losses occur in both temporal directions and are generally larger over longer intervals and for WF than GRU (Fig. 3C). Fig. 10 shows the corresponding results for SBP and TCR ($K = -5$) across all subsets; Appendix E gives the evaluation protocol.

4.4 Cross-Day Changes in Decoder Tuning and Output Bias

We use an output-bias index adapted from Wilson et al. (2026) and decoder-weight similarity to describe cross-day WF mismatch (Fig. 3D). The index measures systematic prediction offsets under frozen source-day z-scoring relative to directional output gain. Weight similarity

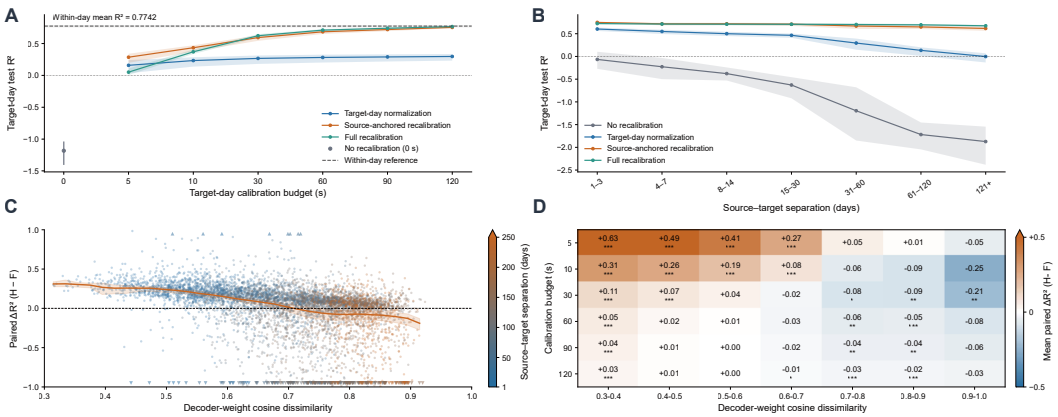

Figure 4: **Historical-weight benefits depend on calibration budget and decoder mismatch when using SBP features from the MA_CO subset.** (A) Target-day test R^2 for target-day normalization (N), source-anchored recalibration (H), and zero-centered full recalibration (F). The 0-s point denotes no recalibration with the frozen source model; the dashed line is the full target-day training reference (REF). (B) Performance at 60 s by source–target calendar-day separation. (C) Paired historical-weight benefit $G = R^2_H - R^2_F$ at 10 s versus weight dissimilarity $d = 1 - S$, colored by day separation. The curve summarizes target-day-weighted local means; triangles mark observations outside the display range. (D) Mean paired benefit by budget and dissimilarity bin. Entries show signed means; stars denote two-sided zero-centered target-day block-bootstrap tests with Holm correction across supported cells in this panel ($^*p_{\text{adj}} < 0.05$, $^{**}p_{\text{adj}} < 0.01$, $^{***}p_{\text{adj}} < 0.001$). All transfers use earlier sources and later targets. Means first average sources within each target day, then weight targets equally. Bands in A–C are pointwise 95% intervals from segmented target-day moving-block bootstrap. Weight dissimilarity uses full target-day training labels as a post-hoc diagnostic. Methods are in Appendix E.3.

compares independently fitted source-day and target-day coefficients at the same channel indices and lags. The within-day weight-fitting reference appears separately in Fig. 13A (Appendix E.2).

Output bias generally increases and weight similarity decreases with day separation. Bias is paired with fully frozen test R^2; weight similarity is paired with R^2 after substituting target-day training statistics for source-day input z-scoring while keeping source weights fixed. This substitution probes the contribution of input mean and scale shifts but need not remove all output bias. Poor transfer accompanies both output offsets and fitted-weight differences, motivating the limited-data recalibration comparison.

4.5 Cross-Day Recalibration and Historical-Weight Benefits

For earlier-source/later-target pairs, we compare frozen WF transfer with target-day normalization (N), source-anchored recalibration (H), and zero-centered recalibration (F). Nested sets of complete target-day training trials provide 5, 10, 30, 60, 90, or 120 s of calibration. H and F share labels, normalization, test trials, and the source-selected ridge parameter; only the penalty center differs, using source weights for H and zero for F. The source-centered penalty in H is analogous to L^2-SP regularization for transfer learning (LI et al., 2018). Their paired difference measures the incremental value of historical weights.

For SBP in MA_CO, all three updates improve mean target-day test R^2 over frozen transfer (Fig. 4A). Normalization alone provides partial recovery that levels off, while H and F improve with more labeled data and approach the full-training reference. H has higher mean performance at the shortest budgets, and the two curves converge as the budget grows. At 60 s, both supervised updates remain strong across the displayed day separations, whereas N declines with separation (Fig. 4B).

We relate paired benefit $G = R_H^2 - R_F^2$ to weight dissimilarity $d = 1 - \cos(W_s, W_{\mathrm{ref}(t|s)})$, where the reference fit uses full target-day training labels and the same λ_s. The 10-s view, selected after inspecting all budgets, shows smaller benefit at larger dissimilarity (Fig. 4C). Across budgets, benefits are largest with little calibration data and low dissimilarity, but can turn negative as either increases (Fig. 4D). This is a retrospective diagnostic, not a limited-budget selection rule. The recalibration weights use channel-by-lag standardization, distinct from the channel standardization in Fig. 3D.

Appendices G.3–G.5 extend the calibration curves, weight comparisons, and budget-by-mismatch summaries to all six subset–feature combinations. Appendix E specifies the paired evaluation and target-day statistical procedures.

5 Limitations, Discussion, and Conclusions

Limitations. FILM contains retained sessions from two non-human primates performing controlled two-dimensional reaching tasks, limiting generalization across subjects, behaviors, recording configurations, and follow-up durations. Session retention required both WF and GRU to pass within-session validation (Appendix C), so reported performance does not estimate overall recording yield. The observational cross-day comparisons cannot isolate behavioral, neural, and recording contributions. Releasing preprocessed SBP and TCR rather than raw signals restricts spike sorting and alternative feature extraction. Evaluations are offline and cover WF and GRU decoding and WF recalibration; other decoders and closed-loop performance remain untested. The weight-mismatch diagnostic uses full target-day training labels, so using it to decide whether to retain historical weights would require an estimate available within the calibration budget.

Discussion. FILM characterizes decoder transfer across days of chronic flexible-electrode recording. In the retained sessions, WF and GRU achieve substantial within-day decoding performance, while frozen transfer becomes less reliable over longer intervals (Fig. 3). Direction-conditioned velocity profiles remain similar across many days even as movement speed, trial duration, recorded activity, and estimated velocity tuning vary (Fig. 2). These findings motivate evaluating within-day accuracy and cross-day reliability separately when assessing chronic neural interfaces. The recalibration comparison examines how target-day data and historical weights contribute to recovery. For SBP in MA_CO, normalization based on unlabeled target neural data provides partial recovery, while supervised weight updates yield larger gains (Fig. 4). Under matched calibration conditions, anchoring to historical weights improves mean performance over zero-centered recalibration at the smallest budgets. This advantage diminishes as more target-day data become available or retrospective weight mismatch increases, and reverses in some budget–mismatch settings. This motivates methods that adjust how strongly they retain historical parameters as target data accumulate.

Conclusions. FILM provides neural features from chronic flexible-electrode recordings paired with cursor kinematics across 178 retained sessions from two non-human primates, with 512 acquisition channels and recording spans of up to 277 days. The baseline evaluations show substantial within-day decodability alongside less reliable frozen cross-day transfer. In the evaluated WF settings, target-day recalibration can improve transfer, while the incremental benefit of historical weights varies with calibration budget and retrospective mismatch. These data and protocols support reproducible comparisons of data-efficient cross-day decoder adaptation.

AI use statement

In this work, we used generative AI tools to discuss and refine research methodology, experimental plans, and plans for analysis; to assist with the implementation of computational methods; to assist with translation during manuscript editing; and to discuss possible interpretations of experimental results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses,

clean or reformat the dataset, or support qualitative or thematic data analysis. Formulating mathematical claims, providing essential elements of mathematical proofs, and writing proofs were not applicable to this work.

Additionally, we used generative AI tools to search for relevant literature, brainstorm research directions, suggest manuscript structure and titles, draft selected parts of the manuscript from author-provided outlines and methodological details, and improve the clarity and readability of the text throughout the manuscript.

The authors reviewed and revised AI-assisted material, checked the literature identified with AI assistance, and manually inspected the logic of AI-assisted code and experimental plans. The authors also checked AI-suggested interpretations against the experimental results and made the final methodological, analytical, and interpretive decisions. We take full responsibility for the manuscript, including its text, claims, analyses, code, and other research artifacts.

Ethics Statement

All experimental protocols in this study were approved by the responsible Animal Management Committee. All experiments were conducted in accordance with applicable ethical regulations. The institution name and identifying approval details are withheld for double-blind review. All experimental protocols in this study were approved by the responsible Animal Management Committee. All experiments were conducted in accordance with applicable ethical regulations. The institution name and identifying approval details are withheld for double-blind review.

Reproducibility statement

We provide code for data loading, analysis, decoder training and evaluation, recalibration, and figure generation as an anonymized ZIP archive in the supplementary materials, together with precomputed numerical results for figure regeneration. The experimental setup and preprocessing are described in Section 3, Appendix A, and Appendix B. Dataset curation, organization, access, and loading are documented in Appendix C. The analysis and evaluation protocols are described in Section 4, Appendix D, and Appendix E. An anonymized download link to the FILM dataset in NWB format will be shared with the reviewers and area chairs through a restricted-visibility OpenReview comment once the discussion forum opens.

References

Yahia H Ali, Kevin Bodkin, Mattia Rigotti-Thompson, Kushant Patel, Nicholas S Card, Bareesh Bhaduri, Samuel R Nason-Tomaszewski, Domenick M Mifsud, Xianda Hou, Claire Nicolas, Shane Allcroft, Leigh R Hochberg, Nicholas Au Yong, Sergey D Stavisky, Lee E Miller, David M Brandman, and Chethan Pandarinath. Brand: a platform for closed-loop experiments with deep network models. *Journal of Neural Engineering*, 21(2):026046, apr 2024. doi: 10.1088/1741-2552/ad3b3a. URL `https://doi.org/10.1088/1741-2552/ad3b3a`.

Chad E. Bouton, Ammar Shaikhouni, Nicholas V. Annetta, Marcia A. Bockbrader, David A. Friedenberg, Dylan M. Nielson, Gaurav Sharma, Per B. Sederberg, Bradley C. Glenn, W. Jerry Mysiw, Austin G. Morgan, Milind Deogaonkar, and Ali R. Rezai. Restoring cortical control of functional movement in a human with quadriplegia. *Nature*, 533(7602): 247–250, May 2016. ISSN 1476-4687. doi: 10.1038/nature17435. URL `https://doi.org/10.1038/nature17435`.

David M Brandman, Tommy Hosman, Jad Saab, Michael C Burkhart, Benjamin E Shanahan, John G Ciancibello, Anish A Sarma, Daniel J Milstein, Carlos E Vargas-Irwin, Brian Franco, Jessica Kelemen, Christine Blabe, Brian A Murphy, Daniel R Young, Francis R Willett, Chethan Pandarinath, Sergey D Stavisky, Robert F Kirsch, Benjamin L Walter,

A Bolu Ajiboye, Sydney S Cash, Emad N Eskandar, Jonathan P Miller, Jennifer A Sweet, Krishna V Shenoy, Jaimie M Henderson, Beata Jarosiewicz, Matthew T Harrison, John D Simeral, and Leigh R Hochberg. Rapid calibration of an intracortical brain–computer interface for people with tetraplegia. *Journal of Neural Engineering*, 15(2):026007, jan 2018. doi: 10.1088/1741-2552/aa9ee7. URL https://doi.org/10.1088/1741-2552/aa9ee7.

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, July 2026. ISSN 1546-170X. doi: 10.1038/s41591-026-04414-6. URL https://doi.org/10.1038/s41591-026-04414-6.

Cynthia A Chestek, Vikash Gilja, Paul Nuyujukian, Justin D Foster, Joline M Fan, Matthew T Kaufman, Mark M Churchland, Zuley Rivera-Alvidrez, John P Cunningham, Stephen I Ryu, and Krishna V Shenoy. Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. *Journal of Neural Engineering*, 8(4):045005, jul 2011. doi: 10.1088/1741-2560/8/4/045005. URL https://doi.org/10.1088/1741-2560/8/4/045005.

Breanne P Christie, Derek M Tat, Zachary T Irwin, Vikash Gilja, Paul Nuyujukian, Justin D Foster, Stephen I Ryu, Krishna V Shenoy, David E Thompson, and Cynthia A Chestek. Comparison of spike sorting and thresholding of voltage waveforms for intracortical brain–machine interface performance. *Journal of Neural Engineering*, 12(1):016009, dec 2014. doi: 10.1088/1741-2560/12/1/016009. URL https://doi.org/10.1088/1741-2560/12/1/016009.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. URL http://arxiv.org/abs/1412.3555.

Jennifer L Collinger, Brian Wodlinger, John E Downey, Wei Wang, Elizabeth C Tyler-Kabara, Douglas J Weber, Angus JC McMorland, Meel Velliste, Michael L Boninger, and Andrew B Schwartz. High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet*, 381(9866):557–564, February 2013. ISSN 0140-6736. doi: 10.1016/S0140-6736(12)61816-9. URL https://doi.org/10.1016/S0140-6736(12)61816-9.

Jun Dai, Peng Zhang, Hongji Sun, Xin Qiao, Yuwei Zhao, Jinxu Ma, Shaohua Li, Jin Zhou, and Changyong Wang. Reliability of motor and sensory neural decoding by threshold crossings for intracortical brain–machine interface. *Journal of Neural Engineering*, 16(3):036011, apr 2019. doi: 10.1088/1741-2552/ab0bfb. URL https://doi.org/10.1088/1741-2552/ab0bfb.

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 *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=STqaMqhtDi.

Runjiu Fang, Huihui Tian, Yan Du, Yan Zhao, Yinan Yang, Shouliang Guan, Shengguang Li, Mengcheng Liu, Ke Xu, and Ying Fang. Flexible kirigami microelectrode arrays for neuronal activity recordings in non-human primate brains. *Nature Electronics*, 9(3):266–278, March 2026. ISSN 2520-1131. doi: 10.1038/s41928-025-01560-6. URL https://doi.org/10.1038/s41928-025-01560-6.

Juan A. Gallego, Matthew G. Perich, Raeed H. Chowdhury, Sara A. Solla, and Lee E. Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23(2):260–270, February 2020. ISSN 1546-1726. doi: 10.1038/s41593-019-0555-4. URL https://doi.org/10.1038/s41593-019-0555-4.

Vikash Gilja, Chethan Pandarinath, Christine H Blabe, Paul Nuyujukian, John D Simeral, Anish A Sarma, Brittany L Sorice, János A Perge, Beata Jarosiewicz, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson. Clinical translation of a high-performance neural prosthesis. *Nature Medicine*, 21(10):1142–1145, October 2015. ISSN 1546-170X. doi: 10.1038/nm.3953. URL https://doi.org/10.1038/nm.3953.

Leigh R. Hochberg, Mijail D. Serruya, Gerhard M. Friehs, Jon A. Mukand, Maryam Saleh, Abraham H. Caplan, Almut Branner, David Chen, Richard D. Penn, and John P. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, July 2006. ISSN 1476-4687. doi: 10.1038/nature04970. URL https://doi.org/10.1038/nature04970.

Yoh Inoue, Hongwei Mao, Steven B. Suway, Josue Orellana, and Andrew B. Schwartz. Decoding arm speed during reaching. *Nature Communications*, 9(1):5243, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07647-3. URL https://doi.org/10.1038/s41467-018-07647-3.

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):313ra179–313ra179, 2015. doi: 10.1126/scitranslmed.aac7328. URL https://www.science.org/doi/abs/10.1126/scitranslmed.aac7328.

Kristopher T. Jensen, Naama Kadmon Harpaz, Ashesh K. Dhawale, Steffen B. E. Wolff, and Bence P. Ölveczky. Long-term stability of single neuron activity in the motor system. *Nature Neuroscience*, 25(12):1664–1674, December 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01194-3. URL https://doi.org/10.1038/s41593-022-01194-3.

Brianna M Karpowicz, Bareesh Bhaduri, Samuel R Nason-Tomaszewski, Brandon G Jacques, Yahia H Ali, Robert D Flint, Payton H Bechefsky, Leigh R Hochberg, Nicholas AuYong, Marc W Slutzky, and Chethan Pandarinath. Reducing power requirements for high-accuracy decoding in ibcis. *Journal of Neural Engineering*, 21(6):066001, nov 2024a. doi: 10.1088/1741-2552/ad88a4. URL https://doi.org/10.1088/1741-2552/ad88a4.

Brianna M. Karpowicz, Joel Ye, Chaofei Fan, Pablo Tostado-Marcos, Fabio Rizzoglio, Clayton B Washington, Thiago Scodeler, Diogo S de Lucena, Samuel R. Nason-Tomaszewski, Matthew Mender, Xuan Ma, Ezequiel Matias Arneodo, Leigh Hochberg, Cynthia Chestek, Jaimie M. Henderson, Timothy Q Gentner, Vikash Gilja, Lee E. Miller, Adam G. Rouse, Robert Gaunt, Jennifer L Collinger, and Chethan Pandarinath. Few-shot algorithms for consistent neural decoding (FALCON) benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=FN02v4nD8y.

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. Stabilizing brain-computer interfaces through alignment of latent dynamics. *Nature Communications*, 16(1):4662, May 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-59652-y. URL https://doi.org/10.1038/s41467-025-59652-y.

Xuhong LI, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2825–2834. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/li18a.html.

Lan Luan, Xiaoling Wei, Zhengtuo Zhao, Jennifer J. Siegel, Ojas Potnis, Catherine A Tuppen, Shengqing Lin, Shams Kazmi, Robert A. Fowler, Stewart Holloway, Andrew K. Dunn, Raymond A. Chitwood, and Chong Xie. Ultraflexible nanoelectronic probes form reliable, glial scar–free neural integration. *Science Advances*, 3(2):e1601966, 2017. doi:

10.1126/sciadv.1601966. URL https://www.science.org/doi/abs/10.1126/sciadv.1601966.

Edwin M. Maynard, Craig T. Nordhausen, and Richard A. Normann. The utah intracortical electrode array: A recording structure for potential brain-computer interfaces. *Electroencephalography and Clinical Neurophysiology*, 102(3):228–239, 1997. ISSN 0013-4694. doi: https://doi.org/10.1016/S0013-4694(96)95176-0. URL https://www.sciencedirect.com/science/article/pii/S0013469496951760.

Samuel R. Nason, Alex K. Vaskov, Matthew S. Willsey, Elissa J. Welle, Hyochan An, Philip P. Vu, Autumn J. Bullard, Chrono S. Nu, Jonathan C. Kao, Krishna V. Shenoy, Taekwang Jang, Hun-Seok Kim, David Blaauw, Parag G. Patil, and Cynthia A. Chestek. A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces. *Nature Biomedical Engineering*, 4(10):973–983, October 2020. ISSN 2157-846X. doi: 10.1038/s41551-020-0591-0. URL https://doi.org/10.1038/s41551-020-0591-0.

Chethan Pandarinath, Paul Nuyujukian, Christine H Blabe, Brittany L Sorice, Jad Saab, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson. High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife*, 6:e18554, feb 2017. ISSN 2050-084X. doi: 10.7554/eLife.18554. URL https://doi.org/10.7554/eLife.18554.

Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity, 2022. URL https://arxiv.org/abs/2109.04463.

Sagi Perel, Patrick T. Sadtler, Emily R. Oby, Stephen I. Ryu, Elizabeth C. Tyler-Kabara, Aaron P. Batista, and Steven M. Chase. Single-unit activity, threshold crossings, and local field potentials in motor cortex differentially encode reach kinematics. *Journal of Neurophysiology*, 114(3):1500–1512, 2015. doi: 10.1152/jn.00293.2014. URL https://doi.org/10.1152/jn.00293.2014. PMID: 26133797.

János A Perge, Shaomin Zhang, Wasim Q Malik, Mark L Homer, Sydney Cash, Gerhard Friehs, Emad N Eskandar, John P Donoghue, and Leigh R Hochberg. Reliability of directional information in unsorted spikes and local field potentials recorded in human motor cortex. *Journal of Neural Engineering*, 11(4):046007, jun 2014. doi: 10.1088/1741-2560/11/4/046007. URL https://doi.org/10.1088/1741-2560/11/4/046007.

Sean M Perkins, Elom A Amematsro, John Cunningham, Qi Wang, and Mark M Churchland. An emerging view of neural geometry in motor cortex supports high-performance decoding. *eLife*, 12:RP89421, feb 2025. ISSN 2050-084X. doi: 10.7554/eLife.89421. URL https://doi.org/10.7554/eLife.89421.

Tsam Kiu Pun, Mona Khoshnevis, Tommy Hosman, Guy H. Wilson, Anastasia Kapitonava, Foram Kamdar, Jaimie M. Henderson, John D. Simeral, Carlos E. Vargas-Irwin, Matthew T. Harrison, and Leigh R. Hochberg. Measuring instability in chronic human intracortical neural recordings towards stable, long-term brain-computer interfaces. *Communications Biology*, 7(1):1363, October 2024. ISSN 2399-3642. doi: 10.1038/s42003-024-06784-4. URL https://doi.org/10.1038/s42003-024-06784-4.

Oliver Rübel, Andrew Tritt, Ryan Ly, Benjamin K Dichter, Satrajit Ghosh, Lawrence Niu, Pamela Baker, Ivan Soltesz, Lydia Ng, Karel Svoboda, Loren Frank, and Kristofer E Bouchard. The neurodata without borders ecosystem for neurophysiological data science. *eLife*, 11:e78362, oct 2022. ISSN 2050-084X. doi: 10.7554/eLife.78362. URL https://doi.org/10.7554/eLife.78362.

John D. Simeral, Thomas Hosman, Jad Saab, Sharlene N. Flesher, Marco Vilela, Brian Franco, Jessica N. Kelemen, David M. Brandman, John G. Ciancibello, Paymon G. Rezaii, Emad N. Eskandar, David M. Rosler, Krishna V. Shenoy, Jaimie M. Henderson,

Arto V. Nurmikko, and Leigh R. Hochberg. Home use of a percutaneous wireless intracortical brain-computer interface by individuals with tetraplegia. *IEEE Transactions on Biomedical Engineering*, 68(7):2313–2325, 2021. doi: 10.1109/TBME.2021.3069119. URL https://doi.org/10.1109/tbme.2021.3069119.

Caleb Sponheim, Vasileios Papadourakis, Jennifer L Collinger, John Downey, Jeffrey Weiss, Lida Pentousi, Kaisa Elliott, and Nicholas G. Hatsopoulos. Longevity and reliability of chronic unit recordings using the utah, intracortical multi-electrode arrays. *Journal of Neural Engineering*, 18(6):066044, dec 2021. doi: 10.1088/1741-2552/ac3eaf. URL https://doi.org/10.1088/1741-2552/ac3eaf.

Hisham Temmar, Yixuan Wang, Nina Gill, Nicholas Mellon, Chang Liu, Luis Hernan Cubillos, Rio I. Parsons, Joseph T Costello, Matteo Ceradini, Madison M. Kelberman, Matthew Mender, Aren Hite, Dylan Michael Wallace, Samuel R. Nason-Tomaszewski, Parag Ganapati Patil, Matt Willsey, Anne Draelos, and Cynthia Chestek. Long-term intracortical neural activity and kinematics (LINK): An intracortical neural dataset for chronic brain-machine interfaces, neuroscience, and machine learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=TAdeh1dLzu.

Yixin Tian, Jiapeng Yin, Chengyao Wang, Zhenliang He, Jingyi Xie, Xiaoshan Feng, Yang Zhou, Tianyu Ma, Yang Xie, Xue Li, Tianming Yang, Chi Ren, Chengyu Li, and Zhengtuo Zhao. An ultraflexible electrode array for large-scale chronic recording in the non-human primate brain. *Advanced Science*, 10(33):2302333, 2023. doi: https://doi.org/10.1002/advs.202302333. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202302333.

Sonia Todorova, Patrick Sadtler, Aaron Batista, Steven Chase, and Valérie Ventura. To sort or not to sort: the impact of spike-sorting on neural decoding performance. *Journal of Neural Engineering*, 11(5):056005, aug 2014. doi: 10.1088/1741-2560/11/5/056005. URL https://doi.org/10.1088/1741-2560/11/5/056005.

Eric M. Trautmann, Sergey D. Stavisky, Subhaneil Lahiri, Katherine C. Ames, Matthew T. Kaufman, Daniel J. O' Shea, Saurabh Vyas, Xulu Sun, Stephen I. Ryu, Surya Ganguli, and Krishna V. Shenoy. Accurate Estimation of Neural Population Dynamics without Spike Sorting. *Neuron*, 103(2):292–308.e4, July 2019. ISSN 0896-6273. doi: 10.1016/j.neuron.2019.05.003. URL https://doi.org/10.1016/j.neuron.2019.05.003.

Yang Wang, Qifan Wang, Ruichen Zheng, Xinxiu Xu, Xinze Yang, Qiang Gui, Xiaowei Yang, Yijun Wang, He Cui, and Weihua Pei. Flexible multichannel electrodes for acute recording in nonhuman primates. *Microsystems & Nanoengineering*, 9(1):93, July 2023a. ISSN 2055-7434. doi: 10.1038/s41378-023-00550-y. URL https://doi.org/10.1038/s41378-023-00550-y.

Yang Wang, Xinze Yang, Xiwen Zhang, Yijun Wang, and Weihua Pei. Implantable intracortical microelectrodes: reviewing the present with a focus on the future. *Microsystems & Nanoengineering*, 9(1):7, January 2023b. ISSN 2055-7434. doi: 10.1038/s41378-022-00451-6. URL https://doi.org/10.1038/s41378-022-00451-6.

Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. The MIT Press, August 1949. ISBN 978-0-262-25719-0. doi: 10.7551/mitpress/2946.001.0001. URL https://doi.org/10.7551/mitpress/2946.001.0001.

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 https://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 https://doi.org/10.1038/s41586-023-06377-x.

Matthew S. Willsey, Samuel R. Nason-Tomaszewski, Scott R. Ensel, Hisham Temmar, Matthew J. Mender, Joseph T. Costello, Parag G. Patil, and Cynthia A. Chestek. Real-time brain-machine interface in non-human primates achieves high-velocity prosthetic finger movements using a shallow feedforward neural network decoder. *Nature Communications*, 13(1):6899, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34452-w. URL https://doi.org/10.1038/s41467-022-34452-w.

Matthew S. Willsey, Nishal P. Shah, Donald T. Avansino, Nick V. Hahn, Ryan M. Jamiolkowski, Foram B. Kamdar, Leigh R. Hochberg, Francis R. Willett, and Jaimie M. Henderson. A high-performance brain–computer interface for finger decoding and quadcopter game control in an individual with paralysis. *Nature Medicine*, 31(1):96–104, January 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03341-8. URL https://doi.org/10.1038/s41591-024-03341-8.

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, July 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01536-z. URL https://doi.org/10.1038/s41551-025-01536-z.

Yinan Yang, Ke Xu, Shouliang Guan, Jianfei Ding, Jinfen Wang, Ying Fang, and Huihui Tian. Ultraflexible Neural Probes for Multidirectional Neuronal Activity Recordings over Large Spatial and Temporal Scales. *Nano Letters*, 23(18):8568–8575, September 2023. ISSN 1530-6984. doi: 10.1021/acs.nanolett.3c02348. URL https://doi.org/10.1021/acs.nanolett.3c02348.

Tansel Baran Yasar, Peter Gombkoto, Alexei L. Vyssotski, Angeliki D. Vavladeli, Christopher M. Lewis, Bifeng Wu, Linus Meienberg, Valter Lundegardh, Fritjof Helmchen, Wolfger von der Behrens, and Mehmet Fatih Yanik. Months-long tracking of neuronal ensembles spanning multiple brain areas with Ultra-Flexible Tentacle Electrodes. *Nature Communications*, 15(1):4822, June 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49226-9. URL https://doi.org/10.1038/s41467-024-49226-9.

Joel Ye, Fabio Rizzoglio, Xuan Ma, Adam Smoulder, Hongwei Mao, Gary H Blumenthal, William Hockeimer, Nicolas Guazzelli Kunigk, Dalton D. Moore, Patrick J. Marino, Raeed H. Chowdhury, J. Patrick Mayo, Aaron Batista, Steven Chase, Michael L Boninger, Charles M. Greenspon, Andrew B. Schwartz, Nicholas G. Hatsopoulos, Lee E. Miller, Kristofer Bouchard, Jennifer L Collinger, Leila Wehbe, and Robert Gaunt. A generalist intracortical motor decoder. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=utXSSdD9mt.

Siyuan Zhao, Xin Tang, Weiwen Tian, Sebastian Partarrieu, Ren Liu, Hao Shen, Jaeyong Lee, Shiqi Guo, Zuwan Lin, and Jia Liu. Tracking neural activity from the same cells during the entire adult life of mice. *Nature Neuroscience*, 26(4):696–710, April 2023a. ISSN 1546-1726. doi: 10.1038/s41593-023-01267-x. URL https://doi.org/10.1038/s41593-023-01267-x.

Zhengtuo Zhao, Hanlin Zhu, Xue Li, Liuyang Sun, Fei He, Jason E. Chung, Daniel F. Liu, Loren Frank, Lan Luan, and Chong Xie. Ultraflexible electrode arrays for months-long high-density electrophysiological mapping of thousands of neurons in rodents. *Nature Biomedical Engineering*, 7(4):520–532, April 2023b. ISSN 2157-846X. doi: 10.1038/s41551-022-00941-y. URL https://doi.org/10.1038/s41551-022-00941-y.

A Experimental Setup Details

A.1 Behavioral Paradigms

FILM contains recordings from two non-human primate subjects, denoted MA and MB. Both subjects were trained to use a joystick to control a two-dimensional cursor on a screen and acquire visually presented targets. Recordings were collected under two trial-based behavioral paradigms: Center-Out (CO) and Random Target (RT).

Center-Out task. Each CO trial began with the cursor at the center of the workspace. One of eight peripheral targets was then presented, and the subject moved the cursor from the center to the target using the joystick. The eight target locations sampled multiple movement directions around the workspace. A trial was marked as successful when the cursor entered the target within the allotted time, at which point a juice reward was delivered. Trials in which the cursor did not reach the target within this interval were marked as unsuccessful. A trial terminated upon target acquisition; no additional target-hold period was required.

Random Target task. In the RT task, target locations varied across consecutive trials. Each movement began from the cursor's current position and ended at the newly presented target, producing reaches with varying start positions, directions, and distances rather than repeated movements from a fixed center. The success criterion and reward contingency matched those of the CO task: target acquisition within the allotted time produced a juice reward, whereas failure to acquire the target produced an unsuccessful trial. No additional target-hold period was imposed.

Neural activity and cursor kinematics were recorded synchronously throughout task performance. The released dataset retains task-related trial periods and their original temporal order; trial curation and NWB organization are described in Appendix C.

A.2 Implant Configurations

Both subjects were implanted with four flexible intracortical arrays of 128 acquisition channels each, providing 512 acquisition channels per subject. The MA arrays followed the design of Yang et al. (2023) and comprised four shanks; the MB arrays followed Fang et al. (2026) and comprised eight shanks. Fig. 5 maps each array's 124 electrode-site IDs (E001–E124) to acquisition-channel IDs (A000–A127), with four reference channels listed separately. Sites are ordered by electrode-site number. The diagram represents wiring and site order; it does not specify physical spacing or implant orientation.

The four reference channels within each 128-channel array are A031, A063, A095, and A127 for MA, and A003, A063, A064, and A124 for MB. For a global acquisition-channel ID $c \in \{0, \ldots, 511\}$, the corresponding within-array ID is $c \bmod 128$. The feature arrays retain all 512 acquisition channels per subject, including the 16 reference channels. The longitudinal RMS analysis uses all 512 channels; other analyses start from the same set and apply their stated validity criteria to individual observations.

For MA, all four arrays were implanted in the dorsal premotor cortex (PMd), with the same configuration for CO and RT recordings. For MB_CO, two arrays were located in PMd, one in ventral premotor cortex (PMv), and one in primary motor cortex (M1). These are the three subject–task subsets included in FILM. Recordings within each subset retain the same implant configuration and acquisition settings (Appendix C).

B Neural Signal Processing and Behavioral Alignment

This appendix specifies the neural features, their alignment to cursor kinematics, and the descriptive analyses in Fig. 1C–D. It then characterizes longitudinal variation in the channel-specific RMS voltage estimates used for TCR detection and evaluates within-day decoding across ten detection thresholds.

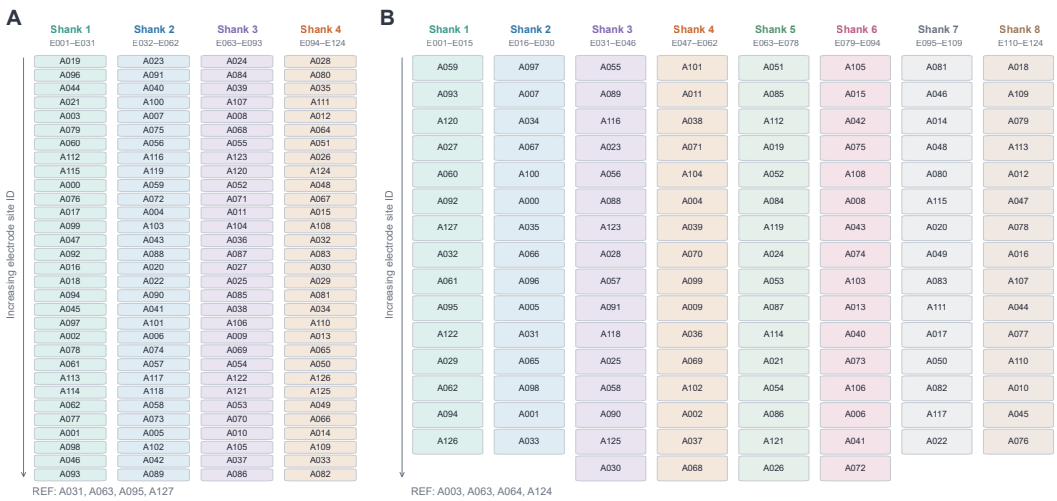

Figure 5: **Electrode-site to acquisition-channel mappings.** (A) MA: four shanks, each with 31 mapped sites. (B) MB: eight shanks, with 15 sites on shanks 1, 2, 7, and 8 and 16 sites on shanks 3–6. Colors identify shanks; E-prefixed labels give electrode-site IDs in increasing order from top to bottom, and A-prefixed labels give acquisition-channel IDs (A000–A127). Each array contains 124 mapped sites and four reference channels, listed under `REF`. Site order does not represent physical spacing or array orientation.

B.1 Neural Signal Processing

Raw electrophysiological signals were acquired using a SpikeGadgets system at a sampling rate of 20 kHz and processed independently for each recording channel. Separate processing pipelines were used to extract TCR and SBP features, as detailed below. Both feature representations had a temporal resolution of 20 ms, with their alignment to behavioral variables described in Section B.2.

Threshold-crossing rate (TCR). For each channel and recording session, we estimated the RMS voltage using a procedure adapted from Christie et al. (2014). We sampled six 5-s segments starting at 5%, 20%, 35%, 50%, 65%, and 80% of the recording duration. Each segment was high-pass filtered at 300 Hz using a fourth-order Butterworth filter applied forward and backward, then divided into non-overlapping 20-ms windows. We sorted the window-wise mean squared voltages, retained values between the 5th and 25th percentiles, and took the square root of their mean to obtain a segment-level RMS estimate. The median across valid segments defined $V_{\mathrm{RMS},c}$ for channel c. Longitudinal variation in these RMS estimates is shown in Fig. 6, and the sensitivity of within-day decoding to the TCR threshold multiplier is evaluated in Fig. 7.

Negative threshold-crossing events were detected in the high-pass-filtered signal at $\theta_c = KV_{\mathrm{RMS},c}$, with $K = -5$. Event channel identities and acquisition timestamps were retained during Trodes export. Events were counted independently for each channel in 1-ms bins, then summed into the 20-ms bins used for neural–behavioral alignment. The stored values are event counts per 20-ms bin; dividing each count by the bin width in seconds converts it to TCR in Hz.

Spiking-band power (SBP). Following Nason et al. (2020), continuous voltage recordings were band-pass filtered at 300–1000 Hz and full-wave rectified. The rectified signal was downsampled from 20 to 2 kHz by averaging consecutive, non-overlapping blocks of 10 samples, then smoothed with a Gaussian kernel with a standard deviation of 20 ms. Timestamps were retained for behavioral alignment, during which the smoothed samples were averaged into 20-ms bins. The resulting SBP values are rectified voltage amplitudes in μV. Subsequent analyses used the saved amplitudes without repeating the signal processing.

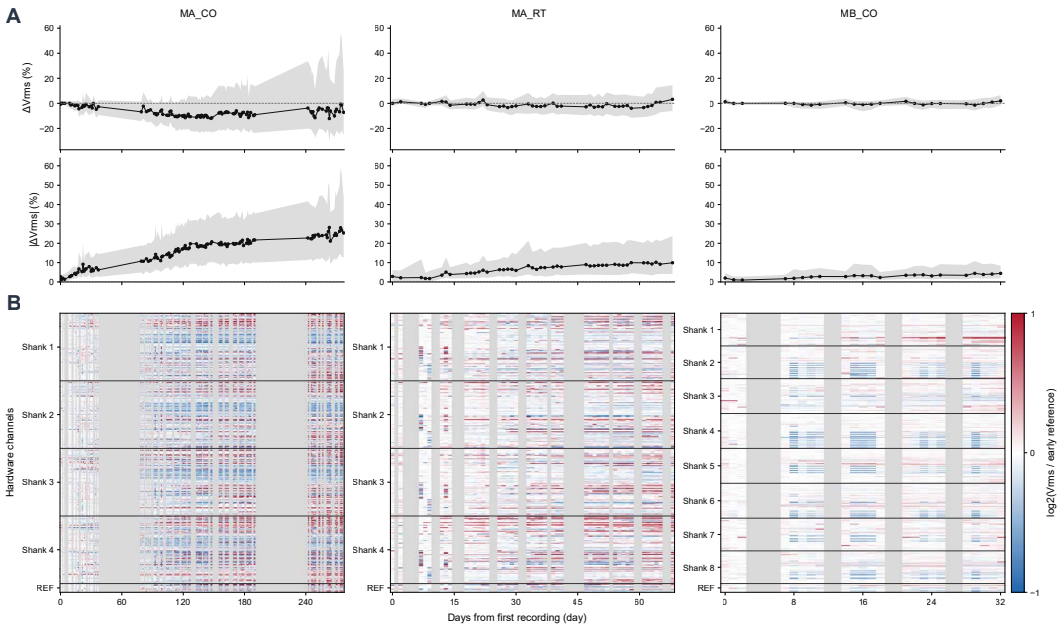

Figure 6: **Longitudinal RMS voltage changes relative to early baselines.** Columns show MA_CO (111 days), MA_RT (44 days), and MB_CO (23 days), with all 512 acquisition channels, including reference channels, in each subset. (A) Signed and absolute percentage changes from channel-specific baselines. Points and lines give daily medians across channels; shading gives the interquartile range. (B) Channel-level $\log_2$ RMS ratios for all 512 acquisition channels, including reference channels. Electrode-site channels are organized by physical shank according to the electrode-site-to-acquisition-channel mapping in Fig. 5, which also identifies the reference channels.

B.2 Behavioral Alignment

Timestamp synchronization. Neural and behavioral data were recorded as separate streams. Behavioral timestamps were recorded in Unix seconds. Neural acquisition times were recorded as 20-kHz sample ticks. These were converted to Unix seconds using the Unix time and corresponding tick at recording creation. This placed both streams on a common absolute time axis. Trials were identified from $t_{\text{TrialStart}}$ events and their associated cursor-coordinate records. Trial intervals retained their original identifiers and temporal order. Neural and behavioral samples were aligned within these intervals; inter-trial samples were excluded.

Temporal binning and kinematic processing. Within each retained trial, TCR counts and cursor kinematics were placed on a common 20-ms grid on this Unix time axis. The 1-ms TCR counts were summed within each bin. Cursor positions were interpolated at 1-ms resolution; positions and velocities derived from them were then averaged within each 20-ms bin. Smoothed 2-kHz SBP samples were located using their recorded Trodes timestamps, and 40 samples were averaged for each fully covered bin. The saved TCR, SBP, position, velocity, and `eval_mask` share absolute Unix timestamps in seconds for each bin and are paired by row index. The mask identifies bins with complete SBP coverage; other bins contain `NaN` SBP values and are excluded from SBP analyses. The temporal windows used for this alignment are checked for timestamp continuity and kept within trial boundaries.

B.3 Longitudinal Variation in RMS Voltage

Fig. 6 uses the original channel-specific RMS estimates in μV from the 111, 44, and 23 recording days in MA_CO, MA_RT, and MB_CO. Hardware channel IDs, rather than input row order, identify observations. Multiple estimates for the same channel and date

are combined by their median. Estimates must be finite, positive, and supported by at least one valid RMS segment. All 512 acquisition channels per subject, including the 16 reference channels, enter this analysis. RMS values are analyzed before applying K, with no fallback substitution.

Channel-specific early baseline. For each subset, let $\mathcal{D}_5$ contain its first five recording dates, and let $V_{c,d}$ denote the valid RMS estimate for acquisition channel c on date d. A channel is retained if it has valid estimates on at least three of these five dates. Its baseline is

$$b_c = \underset{\substack{d \in \mathcal{D}_5 \\ V_{c,d} \text{ valid}}}{\text{median}} V_{c,d}. \tag{1}$$

This channel- and subset-specific baseline remains fixed throughout the longitudinal analysis.

Population summaries and channel-level maps. For each channel with a valid baseline b_c and each recording day with a valid RMS estimate $V_{c,d}$, we compute

$$\Delta_{c,d} = 100 \left(\frac{V_{c,d}}{b_c} - 1 \right), \qquad H_{c,d} = \log_2 \left(\frac{V_{c,d}}{b_c} \right). \tag{2}$$

In Fig. 6, panel A shows daily medians and interquartile ranges of the signed change $\Delta_{c,d}$ and its magnitude $|\Delta_{c,d}|$ across valid channels, including reference channels. Absolute values are taken before aggregation so that increases and decreases do not cancel in the magnitude summary. Panel B shows $H_{c,d}$ for all 512 acquisition channel IDs, with a separate row for each hardware channel. Fig. 5 gives the shank and site mapping for electrode channels and identifies reference channels separately. Both panels use elapsed calendar days from the first recording date of each subset. Panel B retains one column per calendar day; unrecorded days and invalid observations remain missing. Lines in panel A connect recorded dates without imputing values in the gaps. The heatmap color scale saturates at ± 1 for display, while the underlying values are retained.

B.4 Sensitivity to the TCR Detection Threshold

We used the within-day WF and GRU velocity-decoding protocol in Appendix E to evaluate TCR at $K \in \{-6.5, -6, -5.5, -5, -4.5, -4, -3.75, -3.5, -3, -2\}$, where $\theta_c = K V_{\text{RMS},c}$. The TCR conditions use the same recording days and trial splits, with each session's RMS estimates held fixed across thresholds. SBP provides a single reference condition independent of the TCR detection threshold.

Fig. 7 shows that mean TCR scores vary modestly from $K = -6.5$ through -3.5 across the three subsets and both decoders. Scores decline as K approaches -2, with the lowest means at $K = -2$ in all six panels. SBP means are comparable to or higher than the TCR means. The current default $K = -5$ lies within the range of relatively stable within-day performance.

C Dataset Curation, NWB Organization, and Usage

C.1 Data Curation

FILM was curated by recording-configuration screening and within-session decoding criteria. Configuration and session selection precede the longitudinal analyses; the current figures use the selected session collection with TCR extracted at $K = -5$

Recording configuration consistency. Within each subject–task subset, we retained CO or RT sessions with the same electrode implantation configuration, cortical recording sites, and neural acquisition settings. All recordings were acquired using the SpikeGadgets system at 20 kHz. Spontaneous-activity recordings, sessions with configuration changes, and sessions lacking required neural or behavioral data were excluded. Fig. 1A and Appendix A summarize the implant configurations and recording sites of the retained subsets.

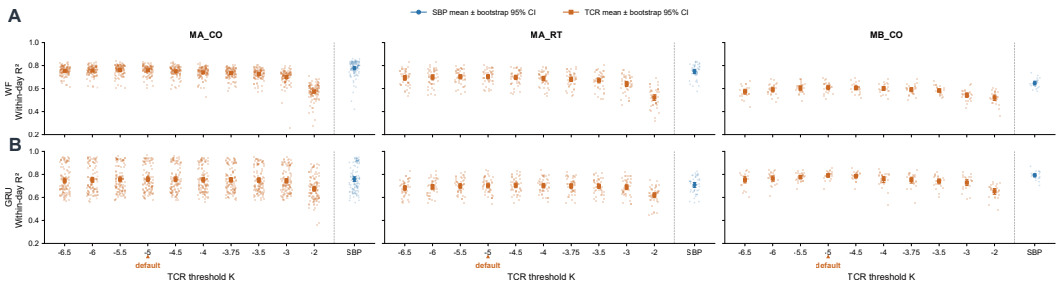

Figure 7: **Within-day decoding sensitivity to the TCR detection threshold.** (A) WF and (B) GRU scores across the three subsets (columns; 111, 44, and 23 sessions). Orange squares show ten signed TCR threshold coefficients K; blue circles show the single SBP reference condition. Faint points represent session scores. Large markers and error bars show equal-session means and pointwise 95% bootstrap confidence intervals (10,000 resamples; seed 42). Arrows mark the default $K = -5$. Test scores average R^2_x and R^2_y on the final 25 trials. Means and intervals retain scores outside the displayed 0.2–1.0 range

Automated session-level quality control. We selected sessions using only the first 100 retained trials of each candidate session. These trials were divided into five contiguous, trial-disjoint folds of 20 trials. For each outer fold, WF and GRU were trained on the other 80 trials and evaluated on the held-out fold. Normalization parameters were estimated from the outer-fold training trials; the WF ridge parameter was selected by trial-disjoint inner validation within those trials. GRU training settings were fixed before session screening. Training and validation histories did not cross an outer-fold boundary. For each model, we pooled out-of-fold predictions over the 100 trials and computed the mean of horizontal and vertical velocity R^2. A session was retained if both $R^2_{\mathrm{WF,CV}} \geq 0.3$ and $R^2_{\mathrm{GRU,CV}} \geq 0.3$. The final 25 trials were excluded from session screening and model selection and were used for the reported held-out evaluation. Selection did not use cross-day decoding results.

Trial sets and recording units. Each retained session contains 125 trials in their original chronological order, including nonconsecutive original trial IDs. A session is one NWB file, whereas a recording day is a calendar date. The MA_CO, MA_RT, and MB_CO subsets contain 111, 44, and 23 sessions, each recorded on a distinct date within its subset. For decoding, the first 100 retained trials form the training pool and the last 25 form the held-out pool. Descriptive analyses instead apply the trial sets and validity criteria specified for their figures. Recording spans use the date difference between the first and last sessions, independent of the number of recorded dates.

C.2 NWB Data Organization

File organization and metadata. Sessions are organized by subject, task, and recording date. A filename such as `MA-CO-20231225-01.nwb` encodes the subject, task, date, and session index within that date and task. Session metadata include the subject identifier, task description, session identifier, and session start time. The NWB `session_start_time` is derived from the Unix time of the first 20-ms TCR bin.

Neural, behavioral, and trial data. The `acquisition/binned_spikes` series stores TCR counts in 20-ms bins as a time-by-channel array with 512 acquisition channels. The `acquisition/sbp` series stores 20-ms SBP amplitudes. Cursor position and velocity are accessed through `acquisition/cursor_pos` and `acquisition/cursor_vel`; analyses read their x and y components. All four series share absolute Unix timestamps in seconds for each 20-ms bin within retained trials. The `intervals/trials` table records the `start_time`, `stop_time`, and original `trial_id` of each retained trial. These boundaries define the retained trial segments. For cross-session analyses, channels are aligned by acquisition-channel ID.

Cross-modal temporal alignment. Within each retained trial, TCR, SBP, cursor position, cursor velocity, and `eval_mask` have matching absolute Unix timestamps for each 20-ms bin and are paired by row index (Appendix B). The mask identifies bins with complete SBP coverage. SBP values in other bins are `NaN`, and those bins are excluded from analyses using SBP. Aligned rows are checked for timestamp continuity within each trial. Decoder input windows follow the rules for each model in Appendix E.

C.3 Data Access and Loading

Data access. The code and derived results needed to regenerate the figures are provided as an anonymized ZIP archive in the supplementary materials. When the OpenReview author–reviewer discussion period begins, we will make the FILM dataset available in NWB format through an anonymized data repository. We will share the repository link in a comment visible only to the reviewers and area chairs assigned to this submission. If the paper is accepted, we plan to release the dataset publicly through a data repository.

Loading NWB files. Sessions can be opened with PyNWB [2]. Before building features, a loader reads neural values and timestamps, both behavioral components, the trials table, channel identities, and temporal-validity information. It selects trials by their saved order, applies the stored SBP mapping, and rejects unmatched or invalid observations. TCR remains in counts for the decoding inputs and is divided by 0.02 s when an analysis requires rates in Hz. SBP analyses retain the saved amplitude scale. These steps preserve the trial split used in all decoding evaluations and align neural inputs by acquisition-channel ID for cross-day transfer and recalibration.

D Kinematic Tuning Analysis Details

We use velocity encoding to characterize channel tuning (Fig. 2D–E) and its cross-day variation (Fig. 9D). Fig. 8 reports lag selection, held-out encoding, and tuning reproducibility for SBP and TCR ($K = -5$) in all three subsets.

D.1 Velocity Model

For channel c and lag ℓ, we fit

$$x_c(t) = \beta_{0,c} + \beta_{x,c} v_x(t+\ell) + \beta_{y,c} v_y(t+\ell) + \epsilon_c(t). \tag{3}$$

Here $x_c(t)$ is SBP or TCR activity. Positive ℓ pairs neural activity with later velocity. Features and velocity are averaged in causal 100-ms windows, then within each nonempty trial–condition combination. Ordinary least squares (OLS) uses these means from the first 100 trials. The final 25 trials are held out. TCR counts are divided by 0.02 s to obtain Hz.

D.2 Movement Windows and Conditions

Movement starts at the first of three consecutive 20-ms bins exceeding 10% of the trial's peak speed and ends with the trial. We exclude trials already above threshold in their first bin or first crossing at peak speed. Only complete 100-ms neural and velocity windows within movement are used. Candidate lags range from −100 to 300 ms in 20-ms steps, at time points valid for every lag. Windows with mean velocity magnitude below 20 in stored units are removed. The remaining windows form 16 conditions: eight 45-degree directions by two speed groups, divided at the subset's median training speed.

D.3 Training-Side Lag Selection

Five trial-disjoint folds of the 100 training trials select one lag per subset and feature. At each lag, out-of-fold predictions and observations are averaged by condition and correlated across at least five supported conditions per channel. We take the median over channels

[2] `https://pynwb.readthedocs.io`

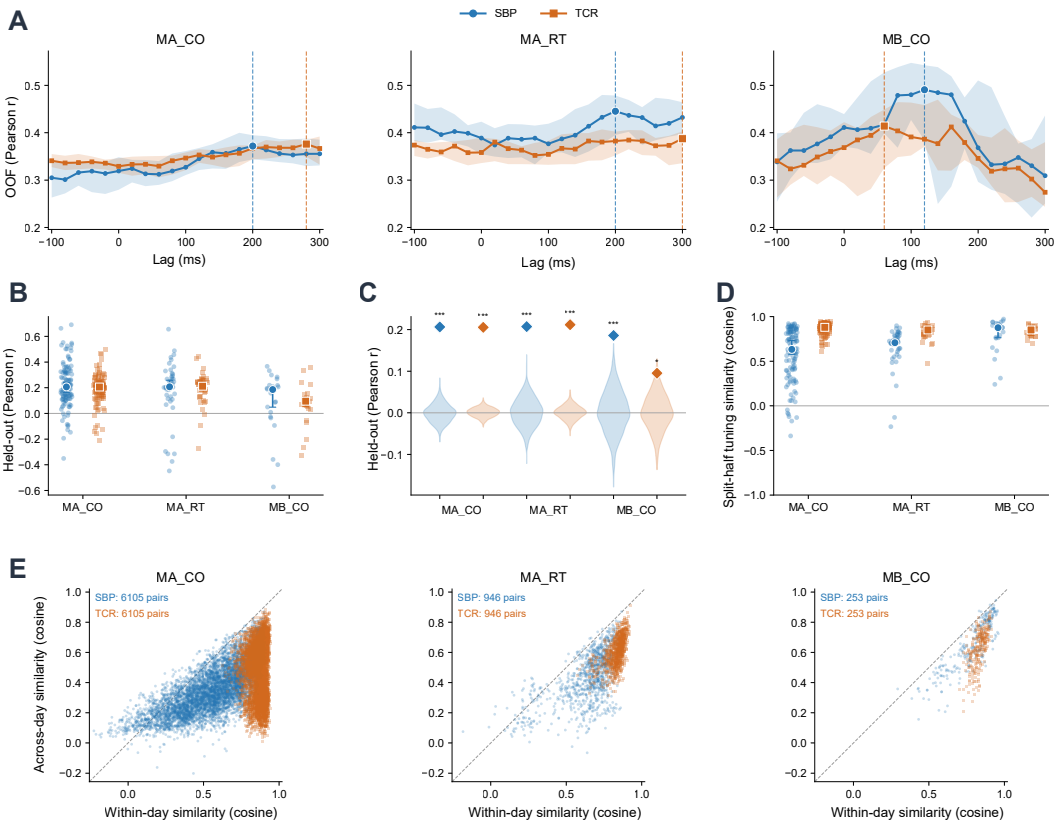

Figure 8: **Velocity encoding, lag selection, and coefficient reproducibility.** Blue circles denote SBP and orange squares denote TCR at $K = -5$. (A) Training-side out-of-fold condition-mean Pearson correlations versus lag, with cohort medians and pointwise session-bootstrap 95% intervals. Selected lags are highlighted. (B) Held-out encoding: faint points are session median channel correlations. Large points and bars give cohort medians and bootstrap 95% intervals. (C) Observed cohort scores (diamonds) and 1,000 full-procedure permutation null values. Stars denote one-sided empirical $p < 0.05$, 0.01, and 0.001. (D) Within-session coefficient cosine from repeated 50/50 training-trial splits, with session-bootstrap 95% intervals. (E) Matched within-day and across-day coefficient cosine for 6,105, 946, and 253 date pairs per feature. Each point summarizes common valid split repeats. The diagonal marks equal similarity. Date pairs sharing a session are dependent. Bootstrap intervals use 10,000 session resamples.

scoreable at every lag, then the median over sessions. Ties favor the smallest absolute lag, then the smaller signed lag.

Selected SBP/TCR lags are 200/280 ms for MA_CO, 200/300 ms for MA_RT, and 120/60 ms for MB_CO. The MA_RT TCR optimum lies at the search boundary, and the 60-ms MB_CO TCR windows overlap. These are alignment choices, not causal-delay estimates.

D.4 Held-Out Encoding and Permutation Reference

At the selected lag, each channel's OLS model is refitted on all 100 training trials. Its held-out score is the Pearson correlation between predicted and observed condition-mean responses on the final 25 trials. Fig. 8B summarizes scores by the median across channels within each session and across sessions within each subset.

To construct the null distribution, we repeat lag selection and held-out evaluation 1,000 times. In each repeat, multichannel trial–condition response rows are shuffled relative to

velocity within each training fold, with matched shuffles across candidate lags; validation responses remain unchanged. At the reselected lag, we draw a separate shuffle of response rows from all 100 training trials, refit the models, and evaluate them on the unchanged 25 held-out trials. Each repeat yields a cohort median held-out score. One-sided empirical p values use the plus-one correction.

D.5 Preferred Direction and Modulation Depth

Coefficients fitted on all 100 training trials give $\mathrm{PD}_c = \mathrm{atan2}(\beta_{y,c}, \beta_{x,c})$ and $\mathrm{MD}_c = (\beta_{x,c}^2 + \beta_{y,c}^2)^{1/2}$. Intercepts are excluded, and coefficients are not rescaled by day. In Fig. 2D, the black angle is the circular median minimizing absolute wrapped deviations, and its radius is the median MD. Error bars show the 25th–75th percentiles. Both examples use the pooled 99th-percentile MD as their radial limit. Boundary triangles mark larger values.

D.6 Population Tuning Similarity

For recordings i and j, $\mathbf{b}_i$ and $\mathbf{b}_j$ concatenate the raw β_x coefficients followed by β_y coefficients for the same valid channel indices in ascending order. Population similarity is the uncentered cosine

$$S_{ij} = \frac{\mathbf{b}_i^\top \mathbf{b}_j}{\|\mathbf{b}_i\|_2 \|\mathbf{b}_j\|_2}. \tag{4}$$

The vectors exclude intercepts and use no day-specific centering or standardization. Fig. 2E shows MA_CO SBP similarities. Negative values remain in the analysis despite the heatmap's zero color minimum. Fig. 9D extends the metric across day intervals and subsets.

D.7 Matched Split-Half Tuning Comparisons

We split each session's 100 training trials into two disjoint 50-trial sets 100 times and refit OLS at the selected lag. Fig. 8D reports each session's median split-half cosine over valid repeats, then the cohort median with a session-bootstrap interval. For each date pair and repeat, all four half-session vectors share valid channels. The within-day reference averages the two split-half cosines, and across-day similarity averages the four between-day half-pair cosines. Panel E takes their medians over matched repeats, requiring at least 80 valid repeats. Equal training counts make within- and across-day similarity comparable.

E Decoding and Recalibration Methods

E.1 Data, Models, and Normalization

Each day contributes one session of 125 retained chronological trials: the first 100 train the decoder and the last 25 test it. The MA_CO, MA_RT, and MB_CO subsets contain 111, 44, and 23 days. Inputs are saved, behavior-aligned 20-ms SBP or TCR features from 512 acquisition channels; TCR uses a threshold of $K * V_{\mathrm{rms}}$, where $K = -5$.

Wiener filter (WF). WF causally smooths features within each trial using a smoothing time constant of $\tau = 120$ ms (Ye et al., 2025) and predicts two-dimensional velocity from six consecutive bins (Perkins et al., 2025). The first five bins of each trial are excluded. A two-output ridge regression has an unpenalized intercept. Four ordered, row-based source-training folds select λ from 20 log-spaced values between 10^{-3} and 10^{5}; the final fit uses all source training rows. Histories can overlap across fold boundaries.

Gated recurrent unit (GRU). GRU predicts velocity at the last bin of each 100-bin input window constructed from the concatenated retained samples within each split. Each window contains 2 s of sampled data at 20-ms resolution. Windows may span trial boundaries and therefore need not cover a continuous 2-s interval on the original recording timeline. They never cross the training–test boundary. Two unidirectional GRU layers have 128 hidden units and 0.1 inter-layer dropout, followed by two 64-unit ReLU/LayerNorm layers and

a two-output layer. AdamW uses batch size 64, learning rate 5×10^{-4}, weight decay 10^{-3}, and 30 fixed epochs. Training loss controls the learning-rate schedule; no held-out data select epochs or models.

Input normalization. Both decoders z-score inputs and clip them to $[-5, 5]$ using source-training statistics. GRU standardizes unsmoothed channels before windowing. WF standardizes smoothed channels before forming histories for Fig. 3A–D, Fig. 10, the TCR threshold sweep, and Fig. 13A. For Fig. 4 and Figs. 11, 12, 13B–C, and 14, WF instead standardizes each channel–lag column after forming histories. The first WF procedure uses source-training statistics throughout cross-validation; the second refits them within each training fold. Weight comparisons use the same normalization convention and the same channel, lag, and output order.

E.2 Cross-Day Evaluation and Diagnostics

Frozen transfer. Each source model is evaluated on its own held-out trials and on every target day's held-out trials, including the source-day diagonal, with weights, intercept, channel order, and source normalization fixed. Valid predictions are pooled to compute R_x^2 and R_y^2 against each test set's mean velocity; the score is $(R_x^2 + R_y^2)/2$. WF and GRU may have different valid window counts. Fig. 3C groups directed pairs by signed calendar separation (positive means an earlier source), showing mean pair scores with descriptive standard errors and median $\Delta R^2_{s\to t} = R^2_{s\to t} - R^2_{s\to s}$. Pointwise intervals for the latter resample source days with their target pairs. Fig. 10 extends this analysis across subsets and features.

Output bias and weight similarity. Fig. 3D analyzes all directed recording-day pairs in the MA_CO subset using SBP features and WF with per-channel z-scoring before history construction. Following Wilson et al. (2026), we use a velocity-adapted normalized bias index (NBI) to quantify systematic offsets in decoded velocity relative to the directional response amplitude. Larger NBI values indicate stronger output bias relative to the directional signal. NBI is computed with source-day z-scoring, decoder weights, and intercept held fixed. Weight similarity is measured as the intercept-free cosine between source- and target-day WF coefficients in matched coordinates and is paired with target-normalized held-out R^2. Target normalization replaces source-day input statistics with target-day training statistics while keeping the source weights and intercept fixed. This approximately compensates for shifts in input means and scales, although residual output bias may remain. Fig. 11 uses earlier-to-later day pairs and applies z-scoring to each channel–lag feature after history construction.

Fig. 13A compares a fixed fit on the target day's 25 held-out trials with 100 random 25-trial training fits from the target or source day. A separate within-day reference splits the 100 training trials into 50/50 fits. Each auxiliary fit uses its day's training pool for channel statistics. Held-out-trial fits serve only this weight diagnostic and never enter predictions, model selection, or recalibration.

E.3 Target-Day Recalibration

For each nominal budget b, we select the longest suffix of complete trials from the target-day training pool whose cumulative duration does not exceed b. The realized calibration duration can therefore be smaller than the nominal budget. The last 25 trials stay held out. No recalibration keeps the source scaler and model; normalization-only recalibration (N) estimates target channel–lag statistics from calibration neural data without velocity labels while retaining source weights. A WF trained on all 100 target training trials provides a separate within-day reference (REF).

Historical-weight recalibration (H) and zero-centered recalibration (F) use identical target trials, target normalization, source-selected penalty λ_s, and a free intercept. For target design $\mathbf{Z}_{t,b}$ and velocities $\mathbf{Y}_{t,b}$, they minimize

$$(\widehat{\mathbf{W}}_C, \widehat{\mathbf{c}}_C) = \arg\min_{\mathbf{W},\mathbf{c}} \|\mathbf{Y}_{t,b} - \mathbf{Z}_{t,b}\mathbf{W} - \mathbf{1}\mathbf{c}\|_F^2 + \lambda_s \|\mathbf{W} - \mathbf{C}\|_F^2, \qquad \mathbf{C} = \begin{cases} \mathbf{W}_s & \text{for H,} \\ \mathbf{0} & \text{for F.} \end{cases} \tag{5}$$

All strategies use the same valid target test bins. The paired benefit is $G_{s,t,b} = R^2_{H,s,t,b} - R^2_{F,s,t,b}$. Weight dissimilarity is $d_{s,t} = 1 - \cos(\mathbf{W}_s, \mathbf{W}_{\mathrm{ref}}(t \mid s))$, where the target reference uses all 100 training trials, penalty λ_s, and matched coordinates. Because this reference uses target velocity labels, $d_{s,t}$ is a retrospective diagnostic. Analyses across budgets use the same eligible source–target pairs.

Aggregation and statistical comparisons. Sources are averaged within each target day, then target days receive equal weight; REF contributes once per target. Each H–F contrast uses matched held-out trials, and each subset–feature combination retains the same eligible pairs across budgets. Pointwise 95% intervals use 2,000 moving-block bootstrap resamples of target days. Day sequences are split at gaps greater than 14 calendar days, and blocks contain five consecutive recorded days within a segment. Each draw retains a target's source pairs and is shared across strategies and budgets.

Fig. 4C and Fig. 13C summarize 10-s benefit in overlapping dissimilarity windows of width 0.1 and step 0.02; this budget was selected after inspecting the results. Figs. 4D and 14 use seven fixed dissimilarity bins from 0.3–0.4 to 0.9–1.0. Windows and cells average sources within target and then targets equally. A cell requires at least ten target days and 95% valid bootstrap draws. Map-cell significance uses 100,000 two-sided, zero-centered target-day block-bootstrap resamples with a plus-one p correction; Holm correction applies within each subset–feature panel. These intervals and tests account for local target-day dependence while treating shared source models as fixed.

F Additional Longitudinal Characterization Results

This appendix specifies the analyses behind Fig. 2A–C and extends the behavioral, neural-feature, and velocity-tuning summaries to all three subsets in Fig. 9. Across subsets, behavior and neural signals vary over recording days while movement profiles and estimated tuning retain some cross-day similarity

F.1 Behavioral Summaries and Cross-Day Similarity

We use saved, aligned position and velocity. Trial direction is the angle of maximum displacement from the initial position, assigned to one of eight 45-degree bins before temporal cropping; displacement must exceed 10^{-9}. Invalid trials, undefined directions, nonfinite samples, and trials with fewer than three valid samples are excluded. The descriptive window starts at the first saved sample with speed above 10^{-8}, or the first sample if movement has already begun, and ends at trial stop. Stationary trials and windows shorter than three samples are excluded; later zero-speed samples remain. This onset differs from the sustained 10%-of-peak onset used for velocity encoding (Appendix D). Fig. 2A reports mean post-onset speed and full stop-minus-start duration, pooling same-day sessions with equal trial weights.

For velocity profile similarity, post-onset velocity components are linearly interpolated over normalized phase to the subset's rounded median window length; amplitudes are retained. Dates require at least five trials in each direction. We average trials equally within direction, correlate each direction between dates, and combine the eight correlations separately for each velocity component using $\tanh(\frac{1}{8}\sum_{k=1}^{8}\operatorname{arctanh}(r_k))$. All directions and both components must be valid. Fig. 2A uses 30-day bins with at least 20 pairs and a Theil–Sen trend; Fig. 9A uses the calendar-day bins shown on its axis. Pair summaries are descriptive because dates recur across pairs.

F.2 Longitudinal Neural Activity Variability

Within trials, neural and behavioral features are paired by their common 20-ms row index. SBP summaries exclude bins without complete coverage under `eval_mask` or with nonfinite values. Temporal windows are checked for timestamp continuity and kept within trial boundaries; each valid feature row is counted once. Daily channel means weight finite, non-

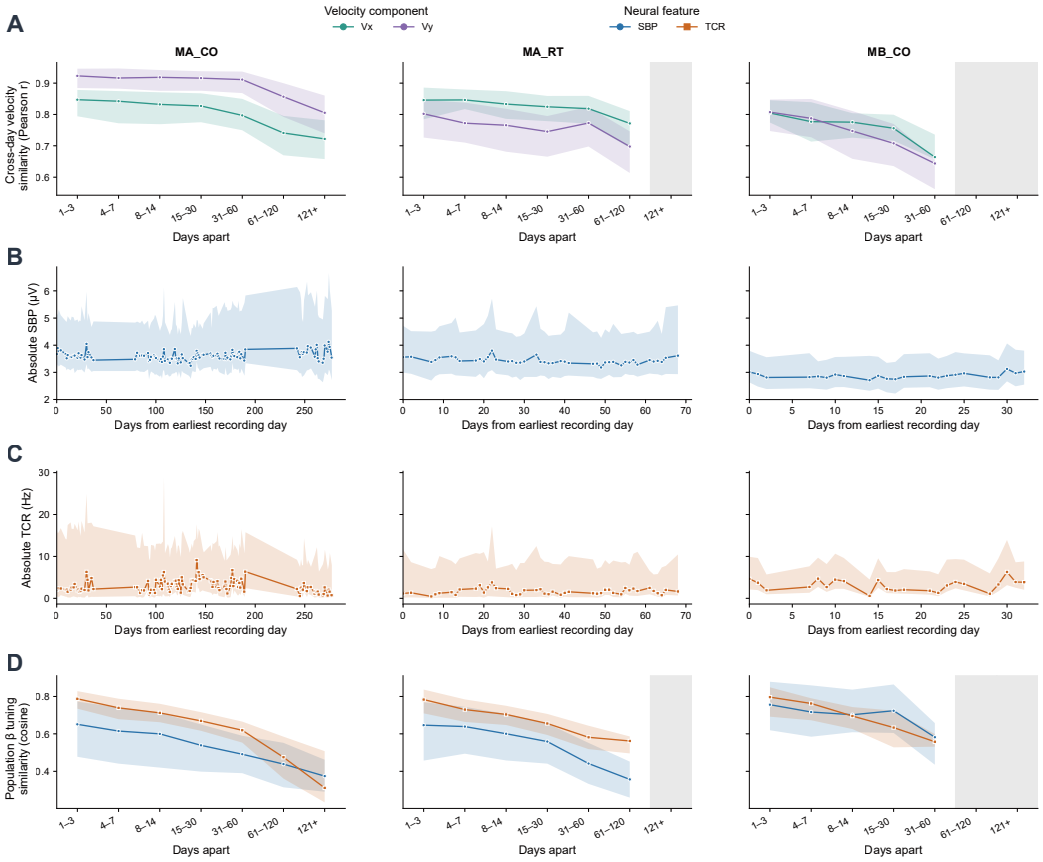

Figure 9: **Longitudinal behavior, neural activity, and velocity tuning.** Columns show MA_CO, MA_RT, and MB_CO. (A) Cross-day Pearson correlations of direction-averaged horizontal (v_x, green) and vertical (v_y, purple) post-onset velocity profiles. (B) Raw daily SBP amplitude in μV and (C) raw daily TCR rate in Hz at $K = -5$. (D) Cross-day cosine similarity of raw population velocity-encoding coefficients for SBP (blue) and TCR (orange). Lines and shaded bands show medians and interquartile ranges (IQRs), respectively, across date pairs in A and D and across channels in B and C. In A and D, date pairs are binned by their separation in calendar days. In B and C, the horizontal axis shows elapsed time since the first recording in each subset. Gray regions indicate bins with no eligible date pairs.

negative samples equally and retain channel indices across dates. For Fig. 2B, each channel's baseline b_c is the median positive daily SBP mean over the first five eligible dates. Signed and absolute percentage changes for daily mean $x_{d,c}$ are

$$L_{d,c} = 100\left(\frac{x_{d,c}}{b_c} - 1\right), \qquad A_{d,c} = |L_{d,c}|. \tag{6}$$

Daily medians and IQRs use the same 512 channels valid throughout the compared period, with absolute values taken before aggregation. No pseudocount, imputation, or daily z-scoring is applied. Fig. 9B shows raw SBP means for channels valid on every recorded date. Panel C uses the same aligned rows and reports TCR ($K = -5$) counts per 20-ms bin as Hz, retaining channels with positive, finite rates on every recorded date. Both panels summarize channels without filling recording gaps.

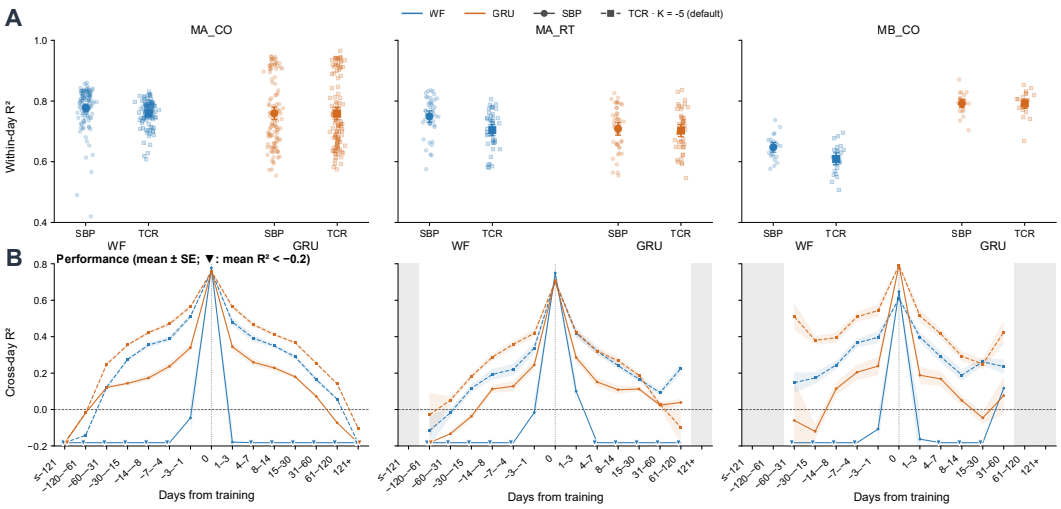

Figure 10: **Within-day and frozen cross-day velocity decoding.** Columns show the three subsets; colors distinguish WF and GRU, and markers distinguish SBP and TCR. (A) Recording-day held-out R^2, equal-day means, and day-bootstrap 95% intervals. (B) Mean raw test R^2 by signed target-minus-source calendar-day separation, with descriptive pair-level SE; zero denotes within-day evaluation. Solid and dashed lines distinguish the two features. Boundary triangles indicate scores beyond the display range, and gray regions have no observations; all scores remain in the summaries.

F.3 Shared Population Activity Projection

Fig. 2C fits one PCA space to all eligible MA_CO SBP rows finite on all 512 channels, weighting samples equally without balancing days or filtering by direction or behavior. After excluding constant channels, we standardize by pooled channel means and population standard deviations and eigendecompose $Z^\top Z$. Each component's sign makes its largest absolute loading positive. Session centroids are mean feature vectors projected with the same pooled statistics and loadings; each recorded day contributes one point. This descriptive projection is separate from decoding evaluation.

F.4 Population Tuning Similarity Across Recording Days

Fig. 9D applies the raw-coefficient cosine defined in Appendix D to the same valid channel indices for each date pair. It excludes intercepts, retains negative cosines, and reports medians and IQRs by calendar-day separation. Decoder-weight similarity is examined in Appendix G.2, and recalibration results are presented in Appendices G.3–G.5.

G Additional Decoding and Recalibration Results

Figs. 10–14 extend the decoding, decoder-mismatch, and recalibration analyses to SBP and TCR($K = -5$) across all three subsets. They show the main qualitative patterns while revealing variation with subset, feature, day separation, and weight mismatch.

G.1 Within-Day and Cross-Day Decoding

Fig. 10 applies the within-day and frozen cross-day protocol in Appendix E to every subset–feature combination. Within-day decoding generally exceeds frozen transfer, but the transfer curves vary with temporal direction and separation; the shorter subsets do not cover the longest intervals. These curves report raw R^2, whereas Fig. 3C also shows source-day-relative losses.

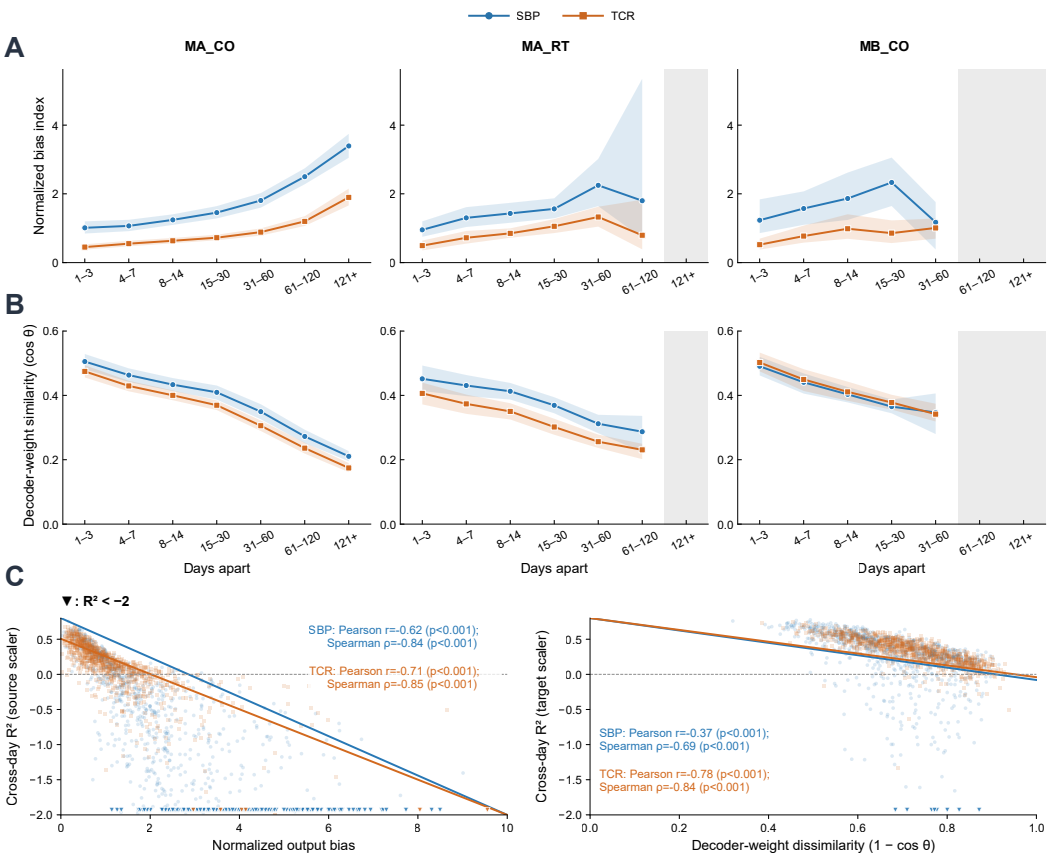

Figure 11: **WF output bias, weight mismatch, and transfer performance.** (A) NBI and (B) weight cosine similarity versus day separation, with mean curves and recording-day pigeonhole-bootstrap 95% intervals. (C) Frozen-transfer R^2 versus NBI and target-normalized R^2 versus decoder-weight cosine dissimilarity $d = 1 - \cos\theta$. Pearson and Spearman correlations and trends use all eligible pairs; points are subsampled for display. Colors and markers distinguish SBP and TCR. Boundary triangles indicate scores beyond the display range; pair-level p values are descriptive because dates recur across pairs.

G.2 Decoder Output Bias and Weight Similarity

Fig. 11 examines earlier-to-later WF transfer using feature-column standardization after history construction, unlike the channel standardization in Fig. 3D (Appendix E.1). Normalized bias index (NBI) is paired with frozen-transfer R^2; intercept-free weight cosine S is paired with target-normalized R^2 (Appendix E.2). Bias and weight differences accompany poorer transfer across the additional combinations, with varying strength.

Temporal summaries use pairs valid for both features, with finite NBI in $[0, 10]$ or S in $[0, 1]$ as appropriate. Seven calendar-day bins report equal-pair means with recording-day pigeonhole-bootstrap intervals. Pooled Pearson and Spearman associations screen features separately; trends fit equal-frequency-bin medians, and at most 900 points per feature are shown. All eligible pairs enter the associations. Shared recording days make pair-level p values descriptive.

G.3 Recalibration Across Subsets and Features

Fig. 12 extends Fig. 4A to all six subset–feature combinations under the matched-budget protocol in Appendix E.3. At the shortest budget, source-anchored recalibration (H) has higher mean R^2 than zero-centered full recalibration (F) in every combination; their curves

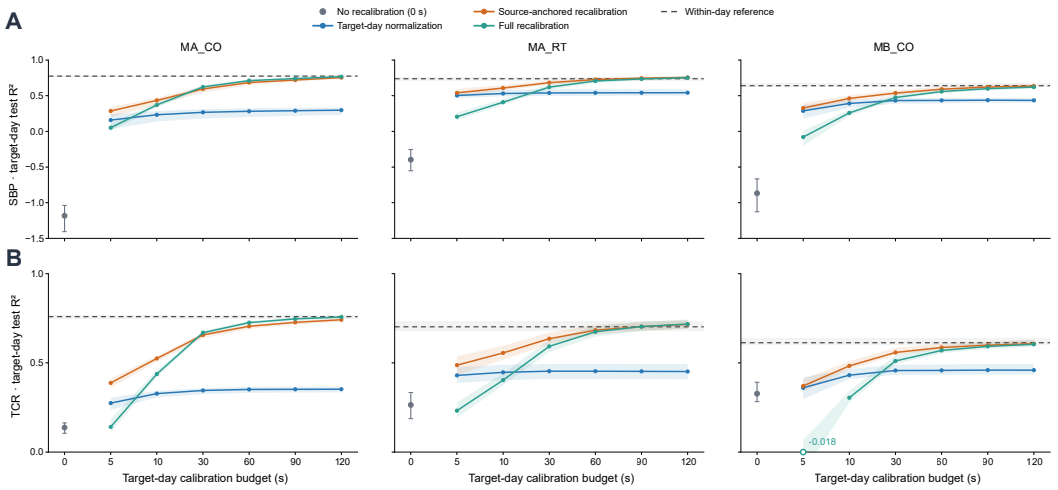

Figure 12: **Recalibration across subsets and neural features.** Columns show the three subsets; rows show SBP and TCR. Gray points denote frozen transfer without calibration; curves show normalization-only (N), source-anchored (H), and zero-centered full (F) recalibration. Points are equal-target-day mean held-out R^2 with pointwise target-day block-bootstrap 95% intervals. Dashed lines and gray bands show the full-training within-day reference (REF). An open marker indicates a mean below the display range.

converge as target-day data increase. At the shortest calibration budget, the mean R^2 for F using TCR features in MA_CO falls below the plotted range, as indicated by an open marker. Statistical methods are described in Appendix E.3.

G.4 Matched Weight Comparisons and 10-s Benefit

Fig. 13A compares within-day and across-day WF weight cosines at matched fitting sizes. The smaller within-day split lowers the reference similarity, yet the displayed across-day similarities remain lower. Rows B and C compare source weights with full-training target reference weights and relate their dissimilarity to ten-second paired benefit $R^2_H - R^2_F$. Greater dissimilarity accompanies smaller benefit, sometimes negative in MA_CO. This is a retrospective diagnostic because the distance uses full target-training labels; the ten-second view was chosen after inspecting the budget results.

Row A standardizes channels before history construction, while B and C standardize feature columns afterward; cosines are compared only within each coordinate system. Row C's local curve uses Gaussian weights with scale 0.5. The matched-fit controls and target-day aggregation are specified in Appendix E.2 and Appendix E.3.

G.5 Benefit Across Calibration Budgets and Weight Mismatch

Fig. 14 extends Fig. 4D across all six combinations, using paired benefit $R^2_H - R^2_F$ and the full-training weight-distance diagnostic. Short-budget, low-mismatch cells often favor historical weights, while negative benefits also occur as mismatch or calibration budget grows. Thus, the direction and size of the benefit depend on both factors. Eligible pairs, support criteria, intervals, and tests follow Appendix E.3.

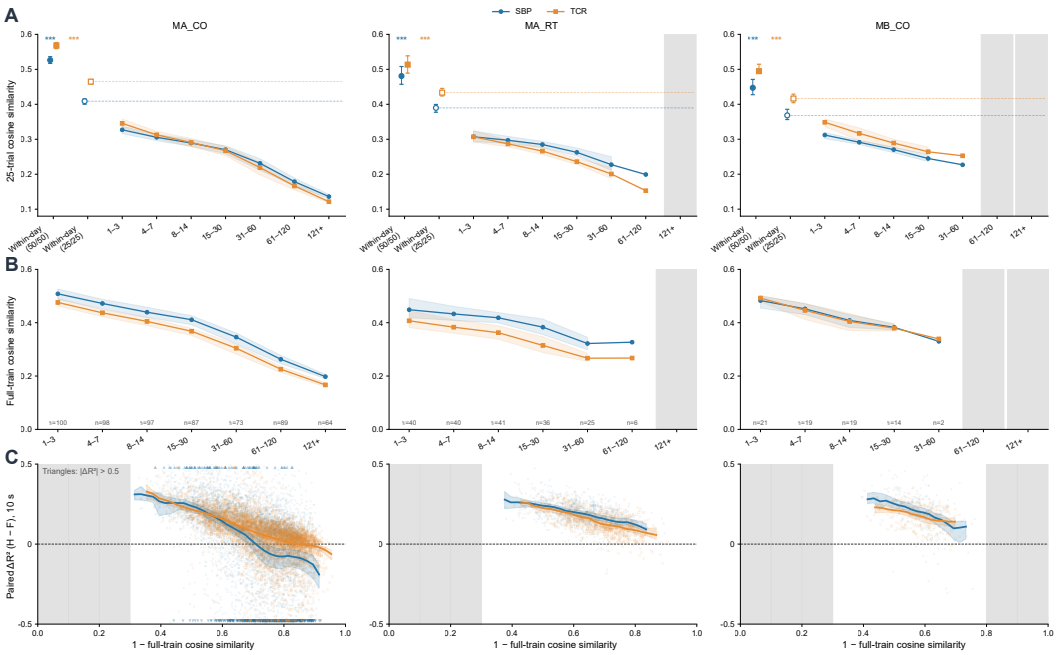

Figure 13: **Weight similarity and historical-weight benefit.** Columns show the three subsets; colors and markers distinguish SBP and TCR. (A) Within-day 50/50 training-split cosines and 25/25 cosines comparing a fixed target-day held-out fit with repeated training fits from the target or source day; dashed lines mark the within-day 25/25 reference. Stars compare the within-day 50/50 and 25/25 values by paired Wilcoxon tests with Benjamini–Hochberg adjustment ($^*q < 0.05$, $^{**}q < 0.01$, $^{***}q < 0.001$). (B) Full-training source–target weight cosine at the source-selected penalty. (C) Ten-second paired benefit versus full-training weight dissimilarity. Faint points are date pairs; curves are locally weighted equal-target-day means with pointwise target-day block-bootstrap 95% intervals. Row A standardizes channels before history construction, whereas rows B and C standardize feature columns after history construction. Triangles at the plot boundaries indicate benefit values outside the displayed range. Gray regions indicate insufficient data support for estimating local trends.

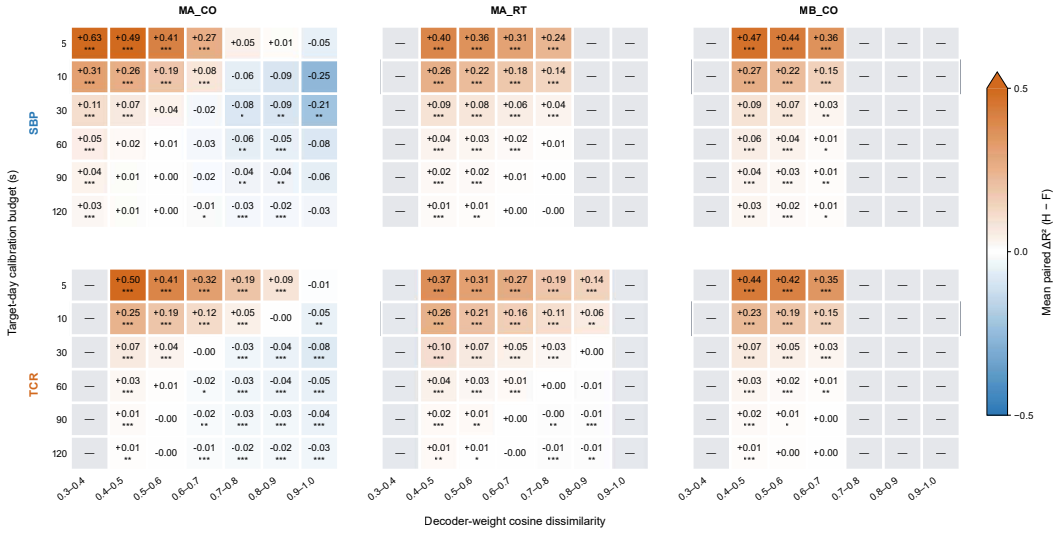

Figure 14: **Historical-weight benefit across calibration budgets and weight dissimilarity.** Columns show the three subsets; upper and lower panels show SBP and TCR. Rows give calibration budgets, and columns within each panel bin full-training weight dissimilarity. Cells show equal-target-day mean paired $R^2_H - R^2_F$; pointwise target-day block-bootstrap 95% intervals are estimated but not displayed. Stars mark two-sided zero-centered target-day block-bootstrap tests with panel-wise Holm correction ($^*p_{\text{adj}} < 0.05$, $^{**}p_{\text{adj}} < 0.01$, $^{***}p_{\text{adj}} < 0.001$). Gray cells lack support; boxes highlight the ten-second view. The top-left panel repeats Fig. 4D.