NEURAL DATA NEEDS SEMANTIC TOKENIZATION: BEHAVIORAL EVENTS AS BOUNDARIES OF SESSION-TRANSFERABLE TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Extracellular electrophysiology records a different set of neurons in every session. Neural foundation models embed each neuron and each session into their tokens, so every new session is an input they have never seen, and they fail to generalize to it. A tokenizer for new sessions needs a unit that every session shares and that carries behavior. Population activity offers such a unit. It evolves on a low-dimensional manifold that persists across neuronal turnover and across animals once sessions are aligned. This manifold changes regime at task events such as stimulus onset and movement onset. Within each regime the population occupies a state, the part of the manifold it spans between two events, and each state carries its own behavioral meaning. We propose **Tokenization with States** (TWS), which segments each trial, one repetition of the task, at these events and converts every state into tokens of population geometry, with no neuron or session embedding. On held-out International Brain Laboratory (IBL) sessions, TWS decodes movement even from regime boundaries that carry no information about the target, while a per-neuron foundation model pretrained on those sessions decodes at chance. Frozen after training on mice alone, TWS transfers to macaques and Utah arrays with only a linear probe. On reach direction, a target that its boundaries do not define, it achieves a Matthews correlation of 0.23, where the event time alone achieves 0.01. For cross-session generalization, the token matters more than the model on top.

1 INTRODUCTION

Every extracellular electrophysiology session records a different set of neurons. Electrode positions shift between Neuropixels insertions, impedance changes alter unit isolation with Utah arrays, and spike sorting yields different populations even from identical trajectories (Jun et al., 2017; Dhawale et al., 2017; Steinmetz et al., 2021). Every new session is therefore *out-of-distribution* by default. Neural foundation models are exposed to this shift, because every token they build carries the identity of one neuron and of the session it came from. Neuron overlap across sessions is typically zero (Appendix A): no neuron label recurs, so per-neuron tokens form a vocabulary that is entirely session-specific and collapses on unseen sessions.

A tokenizer for new sessions therefore needs a unit with two properties: (1) a single decoder can use it in every session, and (2) it carries behavioral meaning. We look for such a unit in population activity, the joint activity of the recorded neurons. It evolves on a low-dimensional manifold (Cunningham & Yu, 2014; Gallego et al., 2017). This manifold separates into distinct regimes at behavioral events such as movement onset (Vyas et al., 2020; Churchland et al., 2012).[1] Elsayed et al. (2016) showed that at movement onset population activity moves into a subspace orthogonal to the one it occupied during preparation. (1) The manifold is shared across sessions. The same latent dynamics persist across months of neuronal turnover (Gallego et al., 2020) and across animals (Safaie et al., 2023). These studies align each session to the shared structure with a transform such as CCA, and a unit that describes the manifold without neuron identity needs no such alignment. (2) The manifold carries behavior. Within the window between two behavioral events, it tracks what the an-

[1] These regimes differ from the events of event segmentation theory (Appendix B).

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.

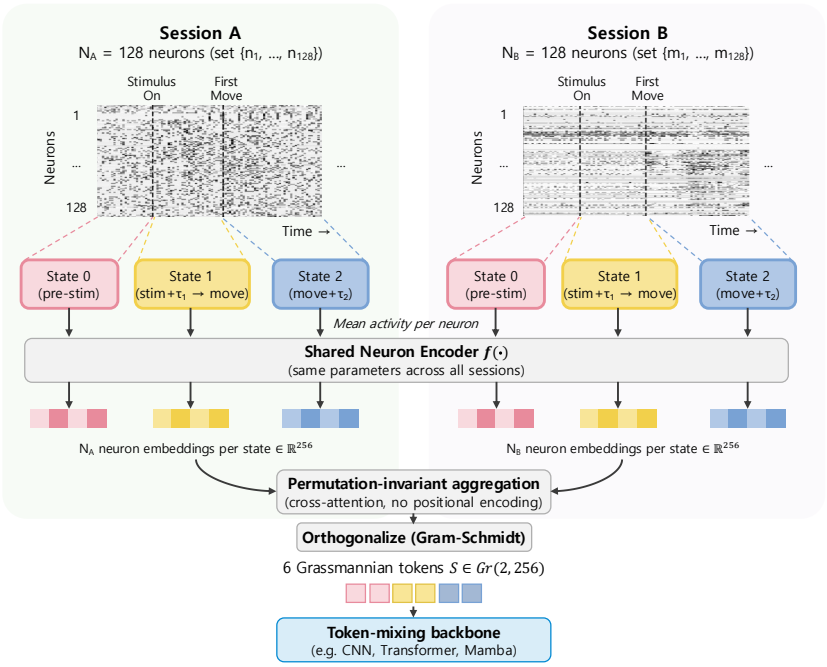

Figure 1: **Overview of Tokenization with States (TWS).** Each session records a different set of neurons. TWS segments every trial into behavioral states at task events and converts each state into tokens without neuron identity, so all sessions share one token space. States S0 to S2 are the states k=1, 2, 3 of Algorithm 1, and τ_1 and τ_2 are the event times shifted by two trainable offsets shared by all sessions.

imal does (Safaie et al., 2023). The manifold of population activity between two behavioral events therefore meets both requirements, and we use it as the unit of a token (Appendix C).

Language modeling has shown that the definition of a token matters. Whole words form a closed vocabulary that cannot represent unseen words (Sennrich et al., 2016). Single characters carry little meaning on their own and produce long sequences. Subword tokenizers such as BPE (Sennrich et al., 2016) and SentencePiece (Kudo & Richardson, 2018) occupy an intermediate level of granularity. They learn reusable, semantic units below the level of whole words from the frequency of character sequences. Neural data show both failure modes. Per-neuron tokens form a closed vocabulary, and single time bins, like single characters, carry little meaning on their own.

We therefore propose a new tokenizer that converts each behavioral regime of a trial, one repetition of the task from stimulus to response, into a token of the population manifold: **Tokenization with States** (TWS; Figure 1), where a state is the population activity of one behavioral regime. Many controlled electrophysiology experiments record the time of each task event of a trial, such as stimulus onset and the first movement. TWS segments the trial at these times and converts the activity of each regime into a few tokens: the subspace that this activity spans, built from all recorded neurons without their order, so the tokens need no neuron or session embedding. A backbone then mixes these tokens into a prediction of behavior. What the tokens carry therefore bounds what any backbone can decode, and for cross-session generalization the tokenizer decides more than the model on top.

Contributions.

- **Tokens built per behavioral regime rather than per neuron prevent decoding from collapsing on sessions with new neurons.** On unseen IBL sessions TWS decodes movement even when its regime boundaries carry no information about the target, while a per-neuron foundation model decodes at chance even after pretraining on those sessions and their movement labels.
- **A frozen TWS trained on mice transfers to other laboratories, to macaques and to Utah arrays.** On choice and reach direction, targets that the token boundaries do not

define, a linear probe on it outperforms every baseline of Table 4, including a per-neuron foundation model trained on the target datasets themselves.

- **A token space shared by every session lets a decoder fit on a few sessions generalize to the others.** On reaction time, a probe fit on five sessions outperforms every baseline of Table 1 fit on all sessions.

2 Related Work

Neural foundation models. Existing multi-session models retain neuron or session identity: POYO (Azabou et al., 2023), POYO+ (Azabou et al., 2025) and NDT2 (Ye et al., 2023) learn an embedding for every neuron, session or context, NEDS (Zhang et al., 2025) finetunes on each new session, and CEBRA (Schneider et al., 2023) needs a model or an input layer for every session; FALCON (Karpowicz et al., 2024) benchmarks the recalibration this requires. Alignment methods (Gallego et al., 2020; Degenhart et al., 2020; Karpowicz et al., 2022) fit a transform to every new session. SPINT (Le et al., 2025) is invariant to the order of the units but infers an identity for each unit from unlabeled trials of the new session. TWS infers no unit identity and fits no parameter to a new session, and its tokens are the states between task events (Appendix D). Appendix C covers further related work on neural manifolds and their behavioral meaning as well as subword tokenization and data-driven segmentation of neural states.

3 Method

3.1 Problem Setup

Session s records N_s neurons over T time bins per trial and gives trial activity $\boldsymbol{X}^{(s)} \in \mathbb{R}^{T \times N_s}$ where both N_s and the identity of the neurons change from session to session. We seek a representation $\boldsymbol{h} \in \mathbb{R}^d$ of a trial that captures task structure whichever neurons the session records.

Which variables such a representation can carry. Some behavioral variables such as locomotion and arousal occupy a low-dimensional subspace shared across the population (Stringer et al., 2019; Musall et al., 2019) and any large subset of neurons carries them. Others such as choice (Steinmetz et al., 2019) are decoded along directions that differ from session to session. We call the first kind Type A and the second kind Type B. A representation without neuron identity is designed to preserve Type A variables, while Type B variables need not survive the removal of neuron identity. TWS therefore targets Type A variables across sessions (Appendix E).

3.2 TWS: Tokenization with States

TWS segments a variable-neuron trial $\boldsymbol{X} \in \mathbb{R}^{T \times N_s}$ along time into $K{=}3$ behavioral regimes and converts each one into $k_{\mathrm{dim}}{=}2$ tokens of dimension $d{=}256$, six tokens for the trial (Figure 1, Algorithm 1). Soft sigmoid gates place the boundaries at known event times with learnable offsets:

$$w_1(t) = \sigma\left(\frac{b_{\mathrm{stim}} - t}{\gamma}\right), \quad w_2(t) = \sigma\left(\frac{t - \tau_1}{\gamma}\right) \cdot \sigma\left(\frac{b_{\mathrm{move}} - t}{\gamma}\right), \quad w_3(t) = \sigma\left(\frac{t - \tau_2}{\gamma}\right) \quad (1)$$

where $\tau_1{=}b_{\mathrm{stim}}{+}\Delta_1$ and $\tau_2{=}b_{\mathrm{move}}{+}\Delta_2$, Δ_1 and Δ_2 are two trainable scalars shared by all trials and sessions and bounded by a sigmoid (learned values in Appendix F), and γ is a temperature. Population activity does not switch regime at the event itself but settles into the new one after a delay, so each boundary carries a trainable offset instead of a fixed one. The two transitions, $[b_{\mathrm{stim}}, \tau_1)$ and $[b_{\mathrm{move}}, \tau_2)$, fall outside the three states. Before gating each neuron's spike counts go through $\log(1 + x)$ and a z-score over the trials of its session in the batch. This step uses only the unlabeled activity of that session. One encoder lifts the mean activity of every neuron to $\mathbb{R}^d$, and one cross-attention aggregates the N_s vectors with no positional encoding over them. Both are shared by every state, neuron and session; only the k_{dim} queries that extract a state differ between states. Each state gives $k_{\mathrm{dim}}{=}2$ tokens on $\mathrm{Gr}(k_{\mathrm{dim}}, d)$ (other values in Appendix G) and four blocks of a kernel-3 convolution mix the six tokens (Appendix H). Each token set is an orthonormal basis, a point on a Grassmannian manifold (Edelman et al., 1998), which bounds what one state can carry. The bound is what matters, not the geometry: tokens that skip the Gram–Schmidt step decode within 0.01 of these (Appendix G.2).

Algorithm 1 Tokenization with States (TWS): forward pass for one trial

Require: Trial activity $\boldsymbol{X} \in \mathbb{R}^{T \times N_s}$, event times $b_{\text{stim}}, b_{\text{move}}$ ($b_{\text{move}}{=}T$ if no first movement falls inside the window)
Ensure: Tokens $\{\boldsymbol{e}_{k,j}\} \in \mathbb{R}^{6 \times d}$, $d{=}256$; trial embedding $\boldsymbol{h}$ ▷ N_s neurons → 6 tokens of dim 256
$\boldsymbol{x}(t) \leftarrow$ z-score of $\log(1 + \boldsymbol{x}(t))$ over the trials of its session in the batch ▷ per neuron
Step 1: Regime boundary segmentation (cf. Fig. 1, state boxes)
for state $k = 1, \ldots, K$ **with** $K{=}3$ **do**
  Compute soft weights $w_k(t)$ via sigmoid gates at $b_{\text{stim}}, b_{\text{move}}$ ▷ Eq. 1
  $\bar{w}_k(t) \leftarrow w_k(t) \,/\, \sum_{t'} w_k(t')$
end for
Step 2: State-weighted population summary (cf. Fig. 1, shared encoder)
for state $k = 1, \ldots, K$ **do**
  $\bar{\boldsymbol{x}}_k \leftarrow \sum_{t=1}^{T} \bar{w}_k(t) \cdot \boldsymbol{x}(t) \in \mathbb{R}^{N_s}$ ▷ mean activity per neuron per state
end for
Step 3: Manifold commitment (cf. Fig. 1, aggregation and Gram–Schmidt)
for state $k = 1, \ldots, K$ **do**
  for neuron $n = 1, \ldots, N_s$ **do**
    $\boldsymbol{u}_n \leftarrow f_{\text{coord}}(\boldsymbol{p}_n) + f_{\text{act}}(\bar{x}_{k,n})$ ▷ scalar → $\mathbb{R}^d$; f_{act} shared by all states and sessions, f_{coord} a shared bias without coordinates (Appendix I)
  end for
  for $j = 1, \ldots, k_{\text{dim}}$ **do** ▷ $k_{\text{dim}}{=}2$
    $\boldsymbol{r}_{k,j} \leftarrow \mathrm{LN}\Big(\mathrm{CrossAttn}(\boldsymbol{q}_{k,j},\ \{\boldsymbol{u}_n\}_{n=1}^{N_s})\Big)$ ▷ N_s neurons → 1 vector $\in \mathbb{R}^d$; one CrossAttn for all states, query $\boldsymbol{q}_{k,j}$ per state
  end for
  $\{\boldsymbol{e}_{k,j}\}_{j=1}^{k_{\text{dim}}} \leftarrow \mathrm{GramSchmidt}(\{\boldsymbol{r}_{k,j}\})$ ▷ orthonormal basis $\in \mathrm{Gr}(k_{\text{dim}}, d)$
end for
▷ *N_s neurons → $K{\times}k_{\text{dim}}{=}6$ tokens of dim $d{=}256$. Invariant to the order of the neurons.*
Step 4: Backbone + behavioral decode
$\boldsymbol{h} \leftarrow \mathrm{Backbone}([\boldsymbol{e}_{1,1}, \boldsymbol{e}_{1,2}, \boldsymbol{e}_{2,1}, \boldsymbol{e}_{2,2}, \boldsymbol{e}_{3,1}, \boldsymbol{e}_{3,2}])$ ▷ $\in \mathbb{R}^{6 \times d}$
$\hat{y} \leftarrow \mathrm{LinearProbe}(\mathrm{mean}(\boldsymbol{h}))$ ▷ evaluation; training uses one linear head per state

Property (permutation invariance). The claim that TWS tokens carry no neuron identity rests on one property of this aggregation. Permuting the N_s recorded neurons together with their coordinates leaves every TWS token unchanged because each step before the tokens either acts on each neuron alone or sums over the neurons with a softmax that has no positional encoding over them (proof and numerical check in Appendix I). The tokens do depend on how many units a session records: with a quarter of the units TWS retains 83% of its movement decoding (Appendix J).

4 Experiments

4.1 Datasets and Setup

Data. IBL Brain Wide Map (Angelaki et al., 2025): 243 Neuropixels sessions from 109 mice in a visual decision task. **Steinmetz 2019** (Steinmetz et al., 2019): 39 Neuropixels sessions from 10 mice. **Gallego et al.** (Gallego et al., 2020): 111 Utah array sessions from 4 reaching macaques; reach direction is defined in the 84 center-out sessions and we score it there. **NLB MC_Maze** (Pei et al., 2021): 23 Utah array pseudo-sessions from 1 macaque. Each session enters with at most 128 units chosen without labels (most spikes for IBL, highest variance elsewhere, zero-padded when fewer; Appendix F), in 20 ms bins from 0.5 s before stimulus onset or go cue to 1.5 s after it ($T{=}100$; for NLB 1.2 s, $T{=}85$, so that every bin stays inside its trial). The IBL movement label covers 2 s after stimulus onset (Appendix K), so a first movement in the last half second makes the label positive while TWS gets no movement boundary.

Model versions. Four versions of one architecture separate what the architecture, pretraining and behavioral training contribute: **TWS (random init)**, a forward pass with random weights; **TWS pretrained**, after self-supervised training that reconstructs masked state means and pushes consecutive state subspaces toward orthogonality (Appendix F); **TWS from scratch**, trained on behavior

Table 1: Cross-session decoding on IBL over the 53 held-out sessions (Section 4.1). ±: SD over 5 model seeds. PCA and Population Statistics have no trained model; CEBRA and NEDS train one model per session and are run once.

	Type A		Type B	
Method	Movement	RT (R^2)	Choice	Block
Cross-session (leave-one-session-out)				
TWS pretrained + finetuned	0.553 ± 0.006	0.398 ± 0.004	$\mathbf{0.040 \pm 0.006}$	$\mathbf{0.031 \pm 0.003}$
TWS from scratch	$\mathbf{0.565 \pm 0.006}$	$\mathbf{0.402 \pm 0.003}$	0.038 ± 0.006	0.026 ± 0.004
TWS pretrained	0.428 ± 0.043	0.247 ± 0.036	0.022 ± 0.001	0.019 ± 0.002
TWS (random init)	0.274 ± 0.018	0.083 ± 0.011	-0.000 ± 0.004	0.005 ± 0.005
PerceiverIO	0.303 ± 0.009	0.169 ± 0.003	0.014 ± 0.009	0.011 ± 0.006
PCA (aligned)	0.349	0.226	0.000	−0.005
PCA (unaligned)	0.002	0.003	0.027	0.007
POYO	0.000 ± 0.001	-0.001 ± 0.001	-0.000 ± 0.002	0.000 ± 0.002
NEDS	0.000	−0.010	−0.022	0.076
CEBRA	0.000	0.000	0.005	−0.012
Population Statistics	0.000	0.009	0.005	−0.002

alone; and **TWS pretrained + finetuned**. Every IBL model is scored on 53 held-out sessions (49 drawn at random and the 4 test sessions of NEDS (Zhang et al., 2025)). TWS and PerceiverIO train without them. The baselines that fit a model to each session fit one to every held-out session by design.

Evaluation. We decode seven IBL variables (Appendix K, L); the main text reports movement, reaction time (RT), choice and block, and Appendix M all seven and the within-session scores. Cross-session scores are leave-one-session-out: a Ridge probe on frozen embeddings, fit on the other sessions, with no labels and no fitted parameters from the test session. Its unlabeled trials set the per-neuron z-score, and continuous targets are standardized within each session, which uses the label statistics of the test session. We report the Matthews correlation coefficient (MCC; 0 at chance) for discrete variables, R^2 for continuous ones, and *session identifiability* $\mathrm{MI}(\boldsymbol{h}; \mathrm{session})/H(\mathrm{session})$.

Baselines. POYO (Azabou et al., 2023) stands for per-neuron models; it pretrained on all IBL sessions, the held-out ones included, with an auxiliary loss on movement, wheel velocity and licking. Supervised CEBRA (Schneider et al., 2023) and NEDS (Zhang et al., 2025) fit one model per session, NEDS with its own per-neuron and per-session embeddings. PerceiverIO (Jaegle et al., 2021) is the POYO tokenizer without neuron or session embeddings on the TWS CNN backbone, trained from scratch. Aligned PCA rotates each session's PCA latents onto a shared template with that session's unlabeled trials (Appendix D), and PCA without alignment serves as a floor. Population Statistics summarizes each trial by the mean and spread of its activity, and Event-time information is the same probe on the first-movement bin alone, with no neural data.

Cross-dataset evaluation. On Steinmetz, Gallego et al. and NLB, TWS and PerceiverIO retain their IBL weights and only a Ridge classifier is fit, on all sessions of the dataset but the test one. CEBRA fits one model per session and POYO trains on each dataset (Appendix N). NLB pseudo-sessions share neurons, so NLB tests transfer across trials only.

4.2 TWS Tokens Decode Unseen Sessions Where Per-Neuron Tokens Collapse

TWS decodes a session it has never seen. TWS leads every baseline on movement and on reaction time (Table 1), and four baselines enter with an advantage. POYO pretrained on these very sessions and still predicts the same class for every trial. Finetuning POYO end-to-end on IBL behavior does not change this (Appendix D). CEBRA trains one model per session and also predicts a constant for every trial, on the classes and on the continuous targets alike (Appendix O). Aligned PCA rotates each new session onto a shared template and achieves 62% of the movement decoding of TWS and 56% of its reaction-time decoding. TWS fits nothing per session. With regime boundaries that carry no information about the target, TWS still decodes movement, where POYO, NEDS and CEBRA decode nothing (Section 4.3). On the five held-out sessions that also record video, TWS decodes whisker motion, a variable that training never uses, twice as well as PerceiverIO, and no model decodes pupil diameter across sessions (Appendix P).

Table 2: Tokenization boundary ablation (IBL, 53 held-out sessions). Every row uses the same architecture and only the boundaries change. **(a)** Number of boundaries. **(b)** Boundary placement at fixed $K{=}3$. Jittered: true boundary plus Gaussian noise with standard deviation σ bins. Fixed: boundary at a constant bin for all trials. Random: uniform draw per trial. Shuffled: real boundaries reassigned across (or within) sessions. Jitter and shuffling move only the boundaries that exist: a trial whose first movement falls outside the window keeps none, as in the true condition. Event-time information: the same probe on the one-hot first-movement bin that each condition gives, with no neural data; it measures how much of the label that input holds (mean over 5 draws of the perturbation). Cross-session leave-one-session-out, as in Table 1. $\pm$: SD over 5 model seeds, each with its own boundary randomization.

	TWS		Event-time information	
Boundary condition	Movement	RT R^2	Movement	RT R^2
(a) How many boundaries (K)				
$K{=}2$: stimulus onset only	0.214 ± 0.007	0.096 ± 0.007	–	–
$K{=}3$: stimulus + first movement	0.565 ± 0.006	0.402 ± 0.003	0.908	0.443
$K{=}3$ + transitions between states	0.563 ± 0.004	0.403 ± 0.004	0.908	0.443
(b) Where to place the boundary ($K{=}3$ fixed)				
True behavioral	$\mathbf{0.565 \pm 0.006}$	$\mathbf{0.402 \pm 0.003}$	0.908	0.443
Jittered ($\sigma{=}5$ bins)	0.523 ± 0.008	0.399 ± 0.003	0.712	0.441
Jittered ($\sigma{=}10$ bins)	0.515 ± 0.011	0.392 ± 0.003	0.577	0.436
Fixed center (bin 50)	0.215 ± 0.005	0.128 ± 0.003	0.000	0.000
Fixed late (bin 75)	0.188 ± 0.003	0.097 ± 0.006	0.000	0.000
Random uniform	0.171 ± 0.011	0.095 ± 0.001	0.000	0.000
Shuffled across sessions	0.199 ± 0.007	0.115 ± 0.004	0.000	0.000
Shuffled within session	0.196 ± 0.009	0.118 ± 0.003	0.000	0.000

Choice and block stay near chance across sessions. Every method, TWS included, stays near chance on choice and block across sessions (Table 1), as Section 3 predicts for Type B variables (Appendix E). Models that report these variables above chance on IBL, such as NEDS (Zhang et al., 2025), finetune on each test session; our cross-session probe fits nothing to the test session. The NEDS block score is the largest number in the two Type B columns and is still chance: its median over the 53 sessions is 0.000 and 11% of sessions score above zero ($p{=}0.08$, one-sided Wilcoxon signed-rank test).

4.3 Regime Boundaries Decide What the Tokens Can Hold

Table 2 shows that what the tokens hold depends on the number of regime boundaries and on whether each boundary is aligned with the recorded event time. Moving every boundary by 100 ms reduces movement decoding by 7% and moving it by 200 ms by 9%, while boundaries that ignore the trial reduce it by 62% when every trial is segmented at one constant bin and by 70% when each trial is segmented at a random bin. Training does not repair a wrong segmentation: with those boundaries 40 epochs of behavioral training stay below TWS (random init) with the regime boundaries (Table 1). Transition tokens between consecutive states change neither number, since the three state subspaces are already nearly orthogonal (Appendix Q); Appendix R explains why the trial is not segmented more finely.

Neural activity and the event time. TWS decodes movement from neural activity, not only from the event time it receives. With boundaries that carry no information about the target (a fixed, random or shuffled bin), the event time decodes nothing, and TWS retains a third of its movement decoding. The event time also defines movement and constrains reaction time, and a probe on it alone achieves movement MCC 0.91 and RT R^2 0.44 (Table 2, Event-time information). A large part of this movement score comes from the 11% of held-out trials in which the mouse moves shortly before the stimulus. The label counts these trials as trials without a movement, and the event time identifies them because it falls before the stimulus. Without these trials, TWS achieves 93% of the movement information in the event time, and PerceiverIO, which receives no event time, achieves less than half (Appendix R).

Table 3: Tokenizer × backbone on IBL (53 held-out sessions, cross-session protocol of Table 1); every model learns from scratch. ±: SD over 5 model seeds. The CNN row repeats Table 1.

		TWS		PerceiverIO	
	Backbone	Movement MCC	RT R^2	Movement MCC	RT R^2
Token mixing	CNN	$\mathbf{0.565 \pm 0.006}$	$\mathbf{0.402 \pm 0.003}$	0.303 ± 0.009	0.169 ± 0.003
	Mamba-1	0.536 ± 0.032	0.385 ± 0.035	0.287 ± 0.015	0.158 ± 0.005
	Mamba-2	0.555 ± 0.009	0.400 ± 0.005	0.294 ± 0.012	0.162 ± 0.012
	Transformer	0.552 ± 0.006	0.398 ± 0.002	$\mathbf{0.313 \pm 0.008}$	0.168 ± 0.004
No mixing	MLP	0.459 ± 0.016	0.305 ± 0.013	0.290 ± 0.010	0.166 ± 0.005
	Pooling	0.465 ± 0.007	0.288 ± 0.011	0.291 ± 0.004	0.167 ± 0.002

4.4 The Tokenizer, Not the Backbone, Determines Cross-Session Decoding on IBL

We separate the effect of the tokenizer from the effect of the backbone. Two tokenizers, TWS and PerceiverIO (Jaegle et al., 2021) (the POYO tokenizer without neuron or session embeddings), meet six backbones: pooling, MLP, CNN, Transformer (Vaswani et al., 2017), Mamba-1 (Gu & Dao, 2024) and Mamba-2 (Dao & Gu, 2024). Every combination trains from scratch, because PerceiverIO has no self-supervised objective (Table 3).

Backbones that mix the tokens differ from one another by at most 5% in movement decoding, while removing the mixing reduces it by 18%; their transfer to the other datasets is in Appendix S. Swapping TWS for PerceiverIO halves decoding with the same backbone. The embeddings of the two tokenizers also differ in what they spend their variance on. Session identity takes 31% of the PerceiverIO embedding and movement 1%, against 23% and 11% for TWS (Figure 2B; Appendix T). The backbone is interchangeable; the tokenizer is not.

4.5 Tokens Carry Neural State across Species and Recording Technologies

Table 4: Transfer of the frozen representation with a linear probe; POYO trains on each dataset (protocol in Section 4.1). MCC, mean ± SD over 5 seeds. Rows without ± are deterministic features. Event-time information uses the first-movement bin and no neural data.

	Steinmetz 2019			Gallego et al.	NLB MC_Maze
Method	Choice	Feedback	Movement	Reach direction	Movement
TWS from scratch	$\mathbf{0.151 \pm 0.015}$	$\mathbf{0.188 \pm 0.008}$	$\mathbf{0.581 \pm 0.003}$	$\mathbf{0.232 \pm 0.010}$	$\mathbf{0.392 \pm 0.010}$
TWS (random init)	0.036 ± 0.016	0.077 ± 0.017	0.550 ± 0.007	0.128 ± 0.002	0.247 ± 0.010
CEBRA	-0.000 ± 0.000	0.000 ± 0.000	0.005 ± 0.050	0.000 ± 0.010	-0.003 ± 0.018
POYO	0.008 ± 0.010	0.045 ± 0.021	0.111 ± 0.013	0.013 ± 0.009	0.162 ± 0.047
Population Statistics	-0.007	0.000	0.242	0.046	0.177
PCA (unaligned)	0.011 ± 0.010	0.012 ± 0.009	0.390 ± 0.003	0.090 ± 0.004	0.144 ± 0.011
Event-time information	-0.009	0.149	0.319	0.010	0.260

A mouse-Neuropixels-trained TWS transfers to macaques and Utah arrays. We freeze the TWS trained on IBL and fit only a linear probe on each new dataset. It decodes every target better than every baseline of Table 4, including POYO, which trains on the target dataset itself. Training on IBL raises choice, reach direction and NLB movement above TWS (random init), also when the untrained encoder receives biases of the trained size. A PerceiverIO trained on IBL, which also has no neuron or session embedding, transfers too (Appendix U). The transfer of TWS comes from neural activity: the first-movement time alone decodes neither choice nor reach direction (row Event-time information), and choice holds within each movement group and on go trials alone (Appendix V).

What the tokens carry. Each token is a learned summary of how activity is distributed across the neurons of one regime, and this learned summary is what transfers. Two fixed summaries of the same distribution test this: its mean, and its mean with the spread and quantiles over neurons (Appendix U). Both approach TWS on movement. On choice and reach direction the better of the two achieves a fifth and a third of TWS. The learned summary needs no time course inside a regime. TWS averages that time course away, and pretraining on data whose temporal order is destroyed transfers at least as well (Appendix U).

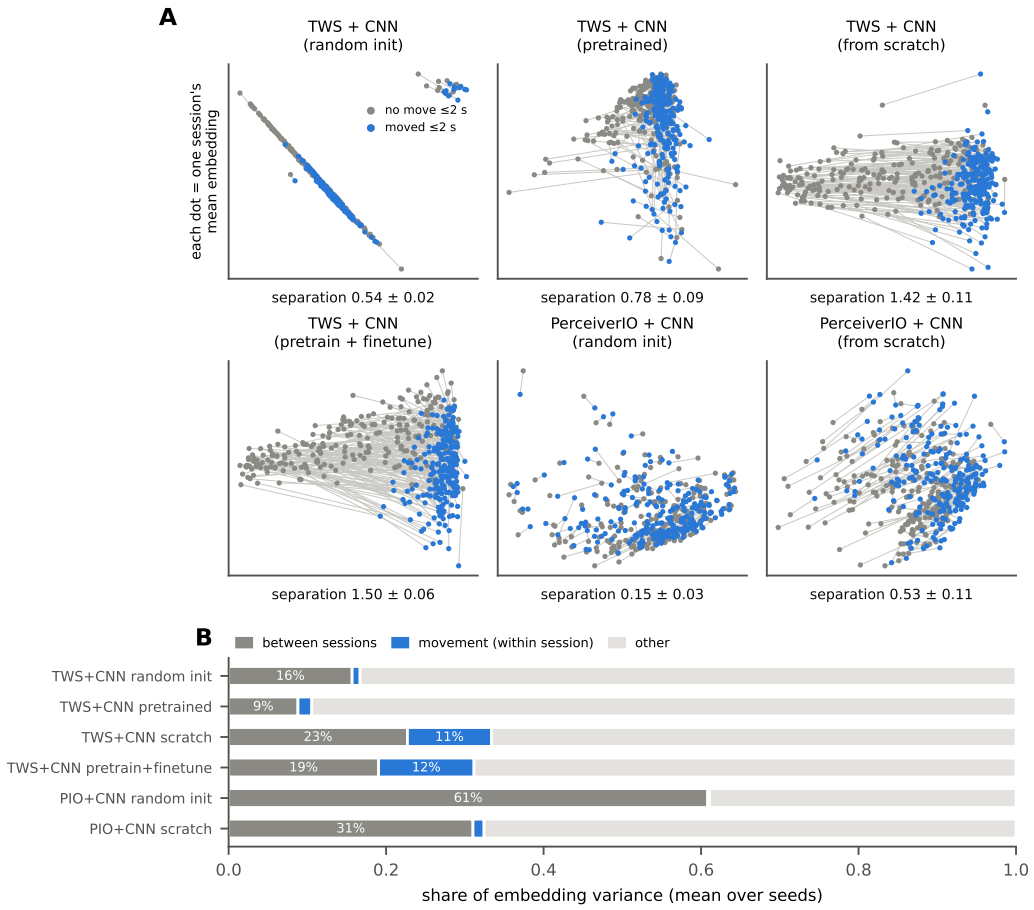

Figure 2: **Training organizes the TWS embedding space around behavior; the tokenizer decides how far. (A)** Each IBL session contributes two points, the mean embedding of its trials with (blue) and without (gray) a first movement within 2 s, and a line joins the two; PCA of these 486 points (first seed). Separation: distance between the two behavior centres over the spread of sessions around them, mean ± SD over 5 seeds. **(B)** Share of trial-embedding variance explained by session identity and, after removing each session's mean, by movement (mean over seeds). Models trained without the 53 held-out sessions; embeddings of all 243 sessions.

4.6 The Architecture Sets the Token Space, Behavioral Training Regroups It, Pretraining Lowers Session Identity

We compare the four versions of TWS on IBL, defined in Section 4.1; Appendix J adds two controls, one on the head that decodes the tokens and one on the number of units per session.

As shown in Table 1, TWS (random init) decodes movement close to a PerceiverIO trained from scratch. The architecture alone already makes movement decodable across sessions. Pretraining lowers the session identity of the embedding, from 23% of the variance after behavioral training alone to 19% when pretraining comes first (Figure 2B), and leaves the final score unchanged (Table 1). An earlier backbone shows the same for every combination of its losses (Appendix W).

In TWS (random init) the trials of each session cluster together whatever their movement state (Figure 2A). Pretraining separates the two movement states only slightly, from 0.54 to 0.78 in the separation of Figure 2A, and behavioral training pushes it to 1.50: the sessions then fall clearly onto two clouds, one per movement state. Subtracting the mean embedding of each session leaves movement decoding intact, trained or not (Appendix X).

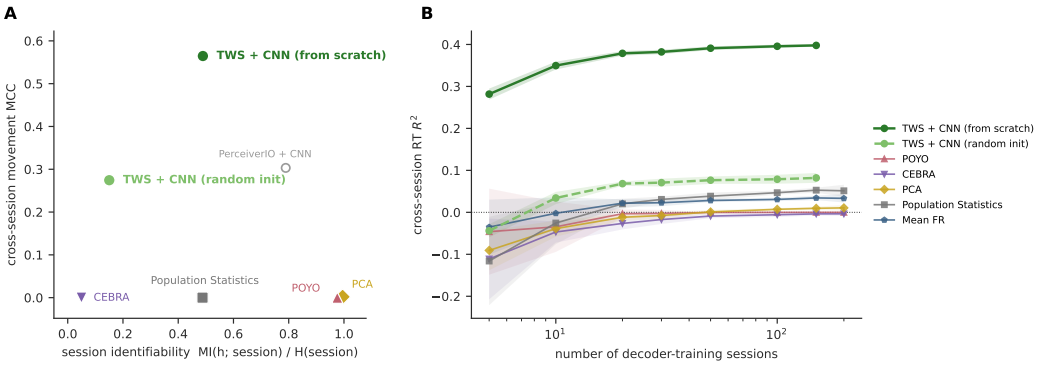

Figure 3: **Session invariance and data efficiency. (A)** Session identifiability (normalized MI of a 243-class probe, x-axis) versus cross-session movement MCC (y-axis, the values of Table 1); methods in the upper-left corner achieve invariance without sacrificing decoding. We draw no error bars because the SD over 5 model seeds is smaller than the markers (values in Appendix Y). **(B)** Reaction time R^2 versus number of decoder-training sessions (log scale). Shaded bands: ± 1 SD over 5 model seeds for TWS and over 20 draws of decoder-training sessions for the baselines. Session identifiability over all 243 IBL sessions; TWS and PerceiverIO decoding over the 53 held-out sessions, with decoder-training sessions in (B) drawn from the other 190.

4.7 Behavior and Session Identity Occupy Different Directions

In TWS, session identity and behavior occupy different directions of the embedding. Deleting the directions that separate sessions leaves movement decoding almost unchanged (0.56 against 0.55). Identifying the session from the two behavior decoding directions works no better than from two random directions (Appendix Z). The session information that remains therefore does not interfere with decoding. The other methods fail in two ways (Figure 3A). POYO and PCA identify the session of almost every trial and decode movement at chance, and CEBRA reduces session identity and loses behavior with it. TWS retains behavior with half the session information of POYO and less than PerceiverIO with the same backbone, at the laboratory level as well as the session level (Appendices AA, AB and AC). On Steinmetz, a dataset it never saw, the frozen TWS retains a third of the session information of per-neuron features (Appendix AD).

4.8 With Labels from Five Sessions, TWS Outperforms Baselines Trained on All of Them

A decoder for a new laboratory has to be fit on sessions with behavioral labels, and TWS needs few of them. Figure 3B shows reaction-time decoding against the number of sessions the probe is fit on. With five sessions, TWS from scratch decodes reaction time five times better than the best baseline of Figure 3B fit on every session. It also exceeds aligned PCA fit on every session by 25% (Table 1), and PerceiverIO on the same five sessions fourfold (Appendix AE). Across six models, this score follows how consistently the probe direction points across sessions, not session information (Appendix AE).

5 Discussion

We propose TWS, a tokenizer that turns the population activity between task events into tokens without neuron or session embeddings. On unseen sessions TWS decodes where per-neuron and per-session models decode at chance. Frozen after training on mice, TWS transfers to macaques and Utah arrays with a linear probe fitted on each dataset, including reach direction, a target that its boundaries do not define. The design sets the limits, and they are the price of the invariance. A tokenizer without neuron identity need not preserve a variable whose decoding direction differs from session to session. Averaging inside a regime gives up the time course within it, and TWS needs the event times that a task records. Free behavior therefore requires a tokenizer that finds the regimes in the data itself. TWS is a practical model for the electrophysiology datasets most laboratories hold. It fits nothing per session, it asks only for the event times a task already records, and a probe fit on a few sessions with behavioral labels generalizes to the remaining sessions.

Reproducibility Statement

All datasets are public and listed with their preprocessing in Section 4. The tokenizer is specified in Section 3 and Algorithm 1, the backbone in Appendix H, and remaining implementation details in Appendix I. Training procedures and all hyperparameters are in Appendix F, and the new-dataset protocol in Appendix N. Evaluation protocols, baselines and seeds are in Section 4, and the behavioral variables are defined in Appendix K. Code will be released upon acceptance.

AI Use Statement

Generative AI tools (Claude) were used for code assistance during implementation and for drafting portions of this manuscript. All AI-generated content was reviewed, verified, and edited by the authors. We take full responsibility for the final content.

References

Dora Angelaki, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, Kcénia Bougrova, Sebastian A Bruijns, Matteo Carandini, Joana A Catarino, et al. A brain-wide map of neural activity during complex behaviour. *Nature*, 645(8079):177–191, 2025.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36:44937–44956, 2023.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L. Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sangyoon Bae, Mehdi Azabou, Blake Aaron Richards, and Jiook Cha. Decoding dynamic visual experience from calcium imaging via cell-pattern-aware pretraining. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=z9kAjjRejs`.

Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487 (7405):51–56, 2012.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Alan D Degenhart, William E Bishop, Emily R Oby, Elizabeth C Tyler-Kabara, Steven M Chase, Aaron P Batista, and Byron M Yu. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nature biomedical engineering*, 4(7):672–685, 2020.

Ashesh K Dhawale, Rajesh Poddar, Steffen BE Wolff, Valentin A Normand, Evi Kopelowitz, and Bence P Ölveczky. Automated long-term recording and analysis of neural activity in behaving animals. *Elife*, 6:e27702, 2017.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Gamaleldin F Elsayed, Antonio H Lara, Matthew T Kaufman, Mark M Churchland, and John P Cunningham. Reorganization between preparatory and movement population responses in motor cortex. *Nature communications*, 7(1):13239, 2016.

Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

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, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021.

James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydın, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

Brianna Karpowicz, Joel Ye, Chaofei Fan, Pablo Tostado-Marcos, Fabio Rizzoglio, Clayton Washington, Thiago Scodeler, Diogo de Lucena, Samuel Nason-Tomaszewski, Matthew Mender, Xuan Ma, Ezequiel Arneodo, Leigh Hochberg, Cynthia Chestek, Jaimie Henderson, Timothy Gentner, Vikash Gilja, Lee Miller, Adam Rouse, Robert Gaunt, Jennifer Collinger, and Chethan Pandarinath. Few-shot algorithms for consistent neural decoding (FALCON) benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Brianna M. Karpowicz, Yahia H. Ali, Lahiru N. Wimalasena, Andrew R. Sedler, Mohammad Reza Keshtkaran, Kevin Bodkin, Xuan Ma, Lee E. Miller, and Chethan Pandarinath. Stabilizing brain-computer interfaces through alignment of latent dynamics. *bioRxiv*, 2022.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations*, pp. 66–71, 2018.

Trung Le, Hao Fang, Jingyuan Li, Tung Nguyen, Lu Mi, Amy Orsborn, Uygar Sümbül, and Eli Shlizerman. SPINT: Spatial permutation-invariant neural transformer for consistent intracortical motor decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22 (10):1677–1686, 2019.

Marius Pachitariu, Shashwat Sridhar, Jake Pennington, and Carsen Stringer. Solving the spike sorting problem with Kilosort. *Nature Methods*, 2024.

Felix Pei, Joel Ye, David M. 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. In *Advances in Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2109.04463.

Mostafa Safaie, Joanna C Chang, Junchol Park, Lee E Miller, Joshua T Dudman, Matthew G Perich, and Juan A Gallego. Preserved neural dynamics across animals performing similar behaviour. *Nature*, 623(7988):765–771, 2023.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 1715–1725, 2016.

Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.

Nicholas A Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588, 2021.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019.

The International Brain Laboratory, Valeria Aguillon-Rodriguez, Dora Angelaki, Hannah Bayer, Niccolò Bonacchi, Matteo Carandini, Fanny Cazettes, Gaelle Chapuis, Anne K. Churchland, Yang Dan, Eric Dewitt, Mayo Faulkner, Hamish Forber, Laura Haetzel, Michael Hausser, Sonja B. Hofer, Fei Hu, Anup Khanal, Christopher Krasniak, Ines Laranjeira, Zachary F. Mainen, Guido Meijer, Nathaniel J. Miska, Thomas D. Mrsic-Flogel, Masayoshi Murakami, Jean-Paul Noel, Alejandro Pan-Vazquez, Cyrille Rossant, Joshua Sanders, Karolina Socha, Rebecca Terry, Anne E. Urai, Hernando Vergara, Miles Wells, Christian J. Wilson, Ilana B. Witten, Lauren E. Wool, and Anthony M. Zador. Reproducibility of in-vivo electrophysiological measurements in mice. *bioRxiv*, 2022. doi: 10.1101/2022.05.09.491042. doi: 10.1101/2022.05.09.491042.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual review of neuroscience*, 43(1):249–275, 2020.

Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36:80352–80374, 2023.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: A mind–brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007.

Yizi Zhang, Yanchen Wang, Mehdi Azabou, Alexandre Andre, Zixuan Wang, Hanrui Lyu, The International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Neural encoding and decoding at scale. *arXiv preprint arXiv:2504.08201*, 2025.

A Neuron Identity Is Session Identity

Neuron identity is session identity (Section 1) because of how extracellular electrophysiology works. Each session records its own set of neurons, and the labels that name them never carry over to another session. We verify this on the IBL Brain Wide Map and on the Gallego et al. data.

No neuron label recurs. A per-neuron tokenizer keys its vocabulary on neuron labels. In IBL these labels never recur. Each session is an acute recording with its own probe insertion, so most sessions also record different cells. We compare the probe tracks of the 361 pairs of sessions from the same subject, using the Allen CCF coordinates of every channel from the aligned histology. Of these, 342 pairs (94.7%) never come within 100 μm of each other, with a median closest approach of 1.2 mm, and 19 pairs do. The labels do not match even for those 19 pairs. Spike sorting (Kilosort (Pachitariu et al., 2024)) runs on each session separately and assigns session-local cluster indices, and no registry links clusters across sessions. Across all $\binom{243}{2} = 29{,}403$ session pairs (109 subjects, median 1 session per subject) no label is shared. Of those pairs, 361 (1.2%) come from the same subject and 29,042 (98.8%) from different subjects.

Chronic recordings. With chronically implanted arrays, such as the Utah arrays of the Gallego et al. and NLB data, some cells are recorded again on later days. The data do not track them. All 111 sessions of the Gallego et al. data number their units from 0 within the session and carry no field that links a unit to another session. Dedicated tracking can follow part of the cells (Dhawale et al., 2017; Steinmetz et al., 2021), but it is rarely available and the tracked set shrinks as the recorded population turns over (Gallego et al., 2020). For the data these models receive, the per-neuron vocabulary is therefore session-specific. The population manifold persists through the turnover (Gallego et al., 2020), and that is what TWS encodes.

Neuron count. Every one of the 243 sessions has at least 128 units that pass quality control, and we keep 128 per session (`max_units`), so in IBL the dimensionality of the raw neural space is the same for every session while its axes are unrelated neurons. The count does vary on the Gallego et al. data (18 to 128 units per session, median 64, after the same cap).

Implication. A model that represents a trial as a vector in neuron-indexed space ($\mathbb{R}^{N_s}$) cannot compare two sessions. Two sessions with $N_1 \neq N_2$ live in spaces of different dimension. Even when $N_1 = N_2$, the axes correspond to unrelated neurons. Per-neuron tokenization, as in POYO and NEDS, answers this with learned neuron embeddings and per-session adaptation, which write neuron identity into the representation. TWS summarizes each state over an unordered set of neurons instead, which removes neuron identity by construction. Session identity is a different matter. TWS reduces it and does not remove it: trained TWS embeddings still carry 0.49 of the session entropy (Figure 3A).

B Regime Boundaries vs. Event Segmentation Theory

We use "regime boundary" rather than "event boundary" to distinguish TWS's boundaries from those in Event Segmentation Theory (Zacks et al., 2007). EST boundaries are observer-internal, driven by prediction error, and hierarchically organized; TWS boundaries are task-defined transitions (stimulus onset set by the task, movement onset measured from behavior) that demarcate distinct computational regimes. The two concepts overlap (both mark qualitative changes in processing) but differ in origin: exogenous task structure versus endogenous perceptual inference.

C Extended Related Work

Subword tokenization as an analogy. In NLP, subword methods such as BPE (Sennrich et al., 2016) and SentencePiece (Kudo & Richardson, 2018) solve the open-vocabulary problem by learning reusable units below the level of whole words, from how often character sequences occur. TWS draws an analogy. Regime boundaries define units between a single time bin and a whole trial. Every such unit marks the same stage of the task in every session, whichever neurons were recorded. BPE merges character pairs by corpus frequency; TWS segments at task events, stimulus onset and movement onset. The boundaries of a TWS token therefore carry domain knowledge rather than a statistical regularity. The Grassmannian projection that follows is the neural analogue of a shared

embedding table: it maps variable-length and variable-neuron states into a fixed geometry that carries no neuron identity and less session identity than per-neuron representations.

Behavioral meaning of the manifold. Population dynamics change regime at task events such as movement onset (Vyas et al., 2020; Churchland et al., 2012) and move between distinct subspaces such as the orthogonal preparatory and movement subspaces of motor cortex (Elsayed et al., 2016). The manifold follows the task and not the animal. One monkey performing two related tasks produces less similar latent dynamics than two monkeys performing the same task (Safaie et al., 2023). The manifold carries this meaning before the animal moves: while a monkey only plans a reach, its dynamics already predict which target it will reach for (Safaie et al., 2023).

Grassmannian and subspace methods. Grassmannian manifolds (spaces of linear subspaces) have a long history in computer vision for subspace tracking and video-based recognition, where each video clip is represented as a point on $\mathrm{Gr}(k, d)$ and geodesic distances provide an appearance-invariant similarity measure. In neuroscience, subspace overlap metrics (principal angles, canonical correlations) are standard tools for quantifying representational similarity across conditions or sessions (Elsayed et al., 2016). TWS uses the Grassmannian not as an analysis tool but as an architectural bottleneck: projecting each state onto $\mathrm{Gr}(k_{\mathrm{dim}}, d)$ with $k_{\mathrm{dim}}{=}2$ forces the representation through a two-dimensional subspace per state. The unordered aggregation removes neuron identity and the narrow bottleneck limits how much session-specific detail passes through. The Grassmannian is one lightweight way to impose this bottleneck.

Neural state segmentation. Hidden Markov models (HMMs) and switching linear dynamical systems have been used to discover latent neural states from continuous recordings (Linderman et al., 2017). These methods learn state boundaries from data. That serves exploratory analysis, but it does not serve cross-session transfer: boundaries learned on one neuron set need not align with those learned on another. TWS takes task-defined regime boundaries instead, stimulus onset and movement onset, which mark the same stage of the task in every session. The price is that TWS needs trial-structured data with known event times.

Appendix B contrasts TWS with Event Segmentation Theory.

D Alignment-Based Baselines versus Tokenization

Contrastive embeddings such as CEBRA (Schneider et al., 2023) are trained on each session and compared afterwards. Manifold-alignment methods (Gallego et al., 2020; Degenhart et al., 2020; Karpowicz et al., 2022) map a new session's latent trajectories onto a reference with a transform fitted to that session's data. Both buy invariance by fitting something to every new session. TWS has it by construction. Each token is a basis vector of a subspace that attention over the unordered neurons computes from the state's population activity, and no permutation of the neurons changes it. The task-defined boundaries then make the k-th token stand for the same computational regime in every session. A new session needs no fitting and no alignment data. The same forward pass encodes it, even at random initialization (Table 1). The price is that TWS needs the event times that define the boundaries, which alignment methods do not.

The baselines in Table 1 differ in what they fit to each session. CEBRA trains one supervised and one self-supervised encoder per session, 486 fits of 5,000 iterations for 243 sessions. POYO learns an embedding for every neuron and every session during pretraining, over all IBL sessions, so its test-session embeddings were fitted to those sessions' data. TWS fits nothing per session. Under the same leave-one-session-out probe it decodes cross-session movement at MCC 0.565 and RT at $R^2{=}0.402$, and at random initialization still at 0.274 and 0.083, while CEBRA and POYO stay at chance on both targets (Table 1). Per-session fitting makes the embedding space of one session internally consistent and leaves the axes of two sessions unrelated, which is what a shared linear probe cannot use. TWS makes the axes comparable before any probe runs. Finetuning POYO end-to-end on IBL choice and RT does not change this. Its embeddings still decode cross-session movement at the majority rate (accuracy 0.748, MCC 0.018; RT $R^2{=}0.036$; one model, scored on all 243 sessions).

A label-free alignment baseline. We implement alignment in the one form that runs on an unseen session. Per session, the log-transformed, z-scored spike counts are reduced to 10 principal components. Each session's trial-averaged latent trajectory, over all trials and aligned to the stimulus,

is rotated onto a common template by orthogonal Procrustes, and generalized Procrustes over all sessions builds that template. Single trials are then projected, rotated and averaged in ten 200 ms windows. The procedure uses no labels, and the held-out session enters through its own unlabeled trials alone. Under the probe of Table 1, aligned PCA decodes movement at MCC 0.349 and RT at R^2=0.226, and choice and block at chance. Without the rotation the same features decode neither target (0.002 and 0.003), so the rotation is what makes PCA transfer. TWS exceeds aligned PCA by 0.22 in movement MCC and 0.18 in RT R^2 and fits nothing to the new session (Table 27).

E Empirical Validation of the Type A / Type B Distinction

A variable is Type A when it survives a change of the recorded neurons. We measure that inside a session, before any model enters. The neurons of a session go into two random halves. Every trial of a half is summarized by statistics that do not depend on which neuron is which: the mean, the standard deviation and the 10, 25, 50, 75 and 90% quantiles over neurons, in five windows of the trial. A decoder is fit on the summaries of one half and scored on the summaries of the other, with the trials cross-validated, so training and test differ in both trials and neurons. The index divides this cross-half score by the score of a decoder that sees the full neuron vector of the same trials. It therefore reports the share of the decodable signal that survives the loss of neuron identity, not the strength of that signal. Every neuron is z-scored within its session first, as in TWS, and the summaries of each half are standardized over the trials of that half, without labels. Ten neuron splits and five folds per session, over 242 IBL sessions, 39 Steinmetz sessions, 81 Gallego et al. sessions and 23 NLB pseudo-sessions. The index needs 60 trials for its folds and 10 trials in the smallest class, which drops one IBL session of 50 trials. Reach direction exists only in the 84 center-out sessions of Gallego et al., since the random-target sessions draw a new target on every trial; three of those 84 fall below the trial or neuron minimum.

Results (Table 5). The index separates the movement targets from IBL choice and block. Over the eight pairs it follows the score of TWS from scratch only loosely (Spearman 0.55, p=0.16). Movement keeps more of its signal than choice or block on IBL and Steinmetz, and NLB movement keeps about a third of it. Choice on IBL keeps a quarter of it. Steinmetz choice keeps about 1.7 times as much as IBL choice, which is why the same target name transfers on one dataset and not on the other. The Steinmetz label contrasts a left turn with a right turn or no turn (Table 22), so part of it follows population-wide movement. The cross-half column and Table 4 share a metric: on Gallego et al. the identity-free decoder achieves 0.112 within a session, and TWS achieves 0.232 across sessions, species and recording technologies. Reaction time has no index, because the full-neuron decoder gives a negative R^2 within sessions and leaves the ratio undefined.

Table 5: How much of each target survives a change of the recorded neurons. MCC, averaged over sessions.

Dataset	Target	Full neurons	Cross-half summary	Index
IBL	movement	0.524	0.308	0.59
	choice	0.624	0.155	0.25
	block	0.560	0.143	0.25
Steinmetz	movement	0.804	0.611	**0.76**
	feedback	0.477	0.287	0.60
	choice	0.807	0.345	0.43
NLB Gallego et al.	movement	0.663	0.234	0.35
	reach direction	0.952	0.112	**0.12**

F Training Details and Hyperparameters

Geometry-only pretraining. Pretraining sets up the geometry of the token space before behavior is learned. Preparatory and movement-epoch activity occupy orthogonal subspaces (Elsayed et al., 2016), so we push the subspaces of consecutive states apart and let behavioral training fill a space that already has this structure; Section 4.6 shows that it adds session invariance rather than accuracy.

We train with three self-supervised losses, no behavioral labels:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\mathrm{MSE}}}_{\text{reconstruct epoch means}} + \lambda_{\mathrm{div}} \underbrace{\mathcal{L}_{\mathrm{div}}}_{\text{state token diversity}} + \lambda_{\mathrm{geo}} \underbrace{\mathcal{L}_{\mathrm{geo}}}_{\text{geodesic distance}} \tag{2}$$

$\mathcal{L}_{\mathrm{MSE}}$: Mean squared error between decoded state tokens and per-state neuron mean activities. $\mathcal{L}_{\mathrm{div}}$: Chordal distance (squared sine of principal angles between consecutive subspaces) plus pairwise cosine repulsion across all state token pairs, jointly encouraging manifold separation. $\mathcal{L}_{\mathrm{geo}}$: Negative geodesic distance on the Grassmannian between adjacent state pairs (S_k, S_{k+1}), encouraging near-orthogonality between temporally neighboring regimes, the prior stated above.

State masking pretext. During pretraining, we randomly mask one of K states (replacing its tokens with a learnable mask embedding) in 50% of trials. The model must reconstruct the masked state from the remaining states, learning inter-state relationships.

Behavioral finetuning. After pretraining, we add one linear head per token for each classification task, with the logits averaged over tokens, and regression heads on the mean token $\bar{\boldsymbol{h}} = \mathrm{mean}(\boldsymbol{h}) \in \mathbb{R}^d$, and finetune on behavioral labels. Models are evaluated both within-session (same session, held-out trials) and cross-session (test sessions never seen during training).

Table 6 lists the settings of every training stage. All stages use AdamW; IBL pretraining and behavioral training clip gradients at norm 1. Each session enters with at most 128 units. In IBL we keep the units that pass the IBL quality label and take the 128 with the most spikes over the session; every IBL session has at least 128 such units. In the other datasets we keep the 128 units with the largest variance and fill the remaining slots with zeros when a session has fewer. Pretraining further keeps IBL units that pass firing-rate, presence-ratio and contamination thresholds, cuts at the knee of their spike counts and keeps the 40% whose binned counts are closest to Gaussian, a selection inspired by Bae et al. (2026). No unit selection uses behavioral labels.

Learned offsets. The offsets of Eq. 1 are trained with the other weights and bounded by a sigmoid to $\Delta_1 \in [5, 25]$ and $\Delta_2 \in [0, 20]$ bins. In the five TWS from scratch models of Table 1 they converge to $\Delta_1 = 14.6$–15.6 bins (292–311 ms after the stimulus onset) and $\Delta_2 = 1.3$–3.0 bins (26–60 ms after the first movement). Δ_1 lands after the stimulus-evoked transient and Δ_2 almost on the movement itself; we read these as the delays at which each regime settles, not as a measurement of them. The architecture is the same in every stage: d=256, K=3 states with k_{dim}=2, gate temperature γ=2 bins, 4 cross-attention heads, at most 128 neurons and T=100 bins of 20 ms; the backbone has 4 blocks, each a layer normalization, a kernel-3 convolution over the 6 tokens, GELU and dropout 0.1, followed by a feed-forward layer.

Table 6: Training hyperparameters. The six IBL behavioral heads are choice, feedback, licking, movement, wheel velocity and RT. New-dataset training follows Appendix N.

	IBL pretraining	IBL behavioral training	New-dataset training
Objective	Eq. 2, no labels	six behavioral heads	class-weighted cross-entropy
Loss weights	λ_{div}=0.01, λ_{geo}=0.005	–	–
State masking	50% of trials	–	–
Learning rate	8×10^{-4}, constant	10^{-3}, cosine to 10^{-5}	encoder 10^{-4}, head 10^{-3}
Weight decay	0.01	0.05	0.01
Batch	256 trials from 16 sessions	128 trials from 8 sessions	4 sessions $\times$ 16 trials
Epochs	30 (patience 15)	40 (no early stopping)	at most 15 (patience 5)
Validation	5% of trials	15% of sessions	15% of training sessions

G Additional Ablations

Backbone note. The k_{dim} table and the first per-state angles below come from an earlier backbone on the same tokenizer (a multi-scale state-space model) and report accuracy; the other results in this section come from the held-out models of Table 1.

G.1 Manifold Dimensionality (k_{dim})

With 243 sessions of decoder training data, movement accuracy holds across k_{dim} (0.864–0.867). All three values decode within 0.003 of one another, and we use k_{dim}=2.

Table 7: Manifold dimensionality ablation (IBL).

k_{dim}	Movement	Interpretation
1	0.864	Slightly narrow: marginal loss of task-relevant variance
2	**0.865**	Selected: balances expressiveness and compression
4	0.867	Similar: extra dimensions add little at this scale

G.2 Tokens without Gram–Schmidt

We remove the Gram–Schmidt step of Algorithm 1 (Step 3). Each state then keeps the two LayerNorm outputs of its cross-attention as its tokens, and the tokens no longer form an orthonormal basis on $\mathrm{Gr}(k_{\mathrm{dim}}, d)$. Everything else stays as in TWS from scratch. Cross-session decoding changes by less than 0.01 on every target (Table 8). The Grassmannian form is therefore a convenient realization of the tokens and not a requirement.

Table 8: Cross-session decoding on the 53 held-out IBL sessions (protocol of Table 1); mean $\pm$ SD over 5 model seeds.

Tokens	Movement MCC	RT R^2	Choice MCC	Block MCC
Orthonormal basis (TWS from scratch)	0.565 ± 0.006	0.402 ± 0.003	0.038 ± 0.006	0.026 ± 0.004
Without Gram–Schmidt	0.562 ± 0.003	0.399 ± 0.003	0.033 ± 0.005	0.022 ± 0.005

G.3 Per-state Manifold Orthogonality

In an earlier geometry-only pretraining run the first principal angle between consecutive state subspaces was 88.8° for S0↔S1 and 88.8° for S1↔S2, against 1.5° and 5.4° without pretraining. The pretrained subspaces are nearly orthogonal; the others overlap heavily. The pretrained model analyzed in Appendix Q (with state masking, before finetuning) gives 87.0° and 87.3° for the first principal angle of the two consecutive pairs, over the trials of all 243 sessions. Geometric pretraining separates the state manifolds.

H Backbone Architecture

The backbone receives the six state tokens and nothing else. TWS has already averaged activity within each state (Algorithm 1, Step 2), so the backbone mixes a short sequence of state summaries rather than time bins. Any sequence model can fill that role (Section 4.4). We use a convolutional mixer by default: four blocks of a kernel-3 convolution over the six state tokens, with d=256. It is the simplest backbone that mixes the tokens at all, and it leads the Mamba and Transformer backbones by at most 0.03 in movement MCC (Table 3). Backbones that mix the tokens decode within 0.03 MCC of each other. They differ in how well they transfer to other datasets (Appendix S).

I Implementation Details

Permutation invariance: proof and check. Let Π permute the N_s recorded neurons and apply it to the activity $\boldsymbol{X}$ and to the coordinates $\{\boldsymbol{p}_n\}$. Then every TWS token is unchanged. *Proof.* Normalization, f_{act} and f_{coord} act on each neuron alone and so commute with Π. The cross-attention returns a softmax-weighted sum over the neurons with no positional encoding over them and so ignores their order. Gram-Schmidt then acts on the k_{dim} attention outputs in a fixed order. □ The backbone sees the six tokens in a fixed order so predictions and distances between tokens inherit the invariance. On five held-out sessions and 20 random permutations each the tokens of a trained model agree to within 4×10^{-6} of their scale.

Scope of the permutation invariance. Each token summarizes the mean activity of every neuron within the state. The invariance holds for the order of the neurons. Recording different neurons changes the activity that enters the aggregation. The aggregation has no padding mask, so in a session with fewer than 128 units the zero-filled slots enter the attention and the tokens depend on the number of recorded units; every IBL session has 128. The backbone receives the orthonormal basis that Gram-Schmidt returns, a point on the Stiefel manifold whose span is the Grassmannian point. The basis therefore enters the representation alongside its span.

Atlas coordinates. In the neuron feature computation $\boldsymbol{u}_n = f_{\text{coord}}(\boldsymbol{p}_n) + f_{\text{act}}(\bar{x}_{k,n})$, atlas coordinates ($\boldsymbol{p}_n \in \mathbb{R}^3$) are missing for 230 of the 243 IBL sessions, so f_{coord} reduces to one learned bias shared by all neurons. In the other 230 sessions every IBL result therefore comes from f_{act} alone: the tokens line up across sessions from per-state mean firing rates, with no anatomical identity. When atlas coordinates are available, f_{coord} provides an additional anatomical prior.

J Two Controls on the Head and on the Neurons

One head over all tokens. Behavioral training fits one linear head per state (Algorithm 1). Pooling the six tokens under a single head leaves cross-session decoding unchanged: movement 0.562 ± 0.004 against 0.565 ± 0.006, reaction time 0.402 ± 0.001 against 0.402 ± 0.003, over the 53 held-out sessions and 5 model seeds. What the tokens hold does not depend on how the head aggregates them.

Fewer neurons per session. Every session enters with at most 128 units (Section 4.1). We keep a random subset of them, zero the rest and leave the model of Table 1 unchanged. Decoding then falls gradually rather than abruptly (Table 9): 32 units keep 83% of the movement decoding of 128 units, and 64 units keep 93%.

Table 9: TWS from scratch on the 53 held-out sessions with fewer units per session, the rest zero-filled. Mean $\pm$ SD over 5 model seeds.

Units per session	Movement MCC	RT R^2	Choice MCC	Block MCC
32	0.469 ± 0.011	0.346 ± 0.015	0.018 ± 0.006	0.031 ± 0.004
64	0.523 ± 0.005	0.380 ± 0.004	0.025 ± 0.005	0.022 ± 0.005
128 (Table 1)	$\mathbf{0.565 \pm 0.006}$	$\mathbf{0.402 \pm 0.003}$	$\mathbf{0.038 \pm 0.006}$	0.026 ± 0.004

K Glossary of Behavioral Variables

Table 10 defines the seven behavioral variables used in this study. All variables are derived from the IBL biased-blocks visual decision-making task (The International Brain Laboratory et al., 2022; Angelaki et al., 2025), in which head-fixed mice turn a steering wheel to indicate the perceived location of a drifting visual grating.

L Behavioral Variable Class Distributions

Five of the seven behavioral variables separate no method from another (Table 12). Table 11 gives the reason. Each of the five is at ceiling, dominated by its base rate, or unpredictable across sessions by construction.

Engagement (P=0.938): predicting "engaged" for every trial yields 93.8% accuracy, and no method decodes it across sessions (MCC ≈ 0). **Feedback** (P=0.813) follows movement (P(fb=1 | move=1)=0.845), yet no method decodes it above MCC 0.11 across sessions. **Choice and block** ($P \approx 0.52$) are Type B. The Grassmannian bottleneck drops the neuron-specific patterns that encode them, so every cross-session method lands at chance, while per-neuron methods within a session confirm the information exists in the raw data (firing rates and Ridge: choice MCC 0.537, block 0.574; Table 12). **Wheel velocity** gives $R^2 \approx 0$ across sessions for every method. Its per-session means shift widely (σ_{cross}=1.62).

M Full Results Across All Behavioral Variables

Table 12 reports cross-session and within-session decoding of all 7 IBL behavioral variables (53 held-out sessions, 5 seeds), with the cross-session protocol of Table 1. Within-session: a random 60/20/20 split of the trials so that training and test trials share sessions. TWS and the other shared representations use one probe on the trials of the 53 held-out sessions together. The per-neuron references (PCA, firing rates and CEBRA) use one probe per session because their features differ between sessions. Per-variable class distributions and chance levels are detailed in Appendix L.

Table 10: Behavioral variable definitions. Type A: encoded across the population, recoverable from any sufficiently large neuronal subset. Type B: depends on which specific neurons are recorded. Scale: C = classification, R = regression. *Feedback is structurally Type A but its cross-session decodability is dominated by its correlation with movement.

Variable	Scale	Type	Base rate	Definition
Movement	C	A	P=0.76	Whether a first wheel movement (IBL `firstMovement_times`) occurs between stimulus onset and 2 s after it. This is the event that also places the first-movement regime boundary, so the boundary alone largely determines the label (Appendix R). Movement-related neural signals appear "almost everywhere in the brain" (Musall et al., 2019; Stringer et al., 2019).
Reaction time	R	A	—	Time (s) from stimulus onset to the IBL response time, which follows the first movement by a median of 0.19 s. Trials outside 50 ms–10 s are excluded. Reflects sensorimotor transformation latency; shorter for higher-contrast stimuli.
Choice	C	B	$P\approx$0.52	Direction of wheel turn: leftward (CW) or rightward (CCW). Correct when the grating is moved to center (0° azimuth). Requires lateralized selectivity in specific neurons (Steinmetz et al., 2019).
Feedback	C	A*	P=0.81	Trial outcome: correct ($+1$, water reward) or incorrect (-1, white-noise burst + 2 s timeout). Mechanistically linked to movement: P(fb=1 | move=1)=0.845.
Block	C	B	$P\approx$0.52	Dominant stimulus side in the current probability block. After 90 unbiased trials, prior probability alternates between 80:20 and 20:80 (left:right) in blocks of 20–100 trials. A latent context variable.
Engagement	C	A	P=0.94	Whether the mouse is actively performing: recent accuracy $\geq$0.6 over a 10-trial rolling window. Near-ceiling base rate reduces discriminative value.
Wheel velocity	R	A	—	Signed inverse latency of the first movement (1/s, sign of the choice), a measure of movement vigor. Trials without a first movement 0.05–2 s after stimulus onset are excluded. Per-session means shift widely (σ_{cross}=1.62).

Table 11: Class distributions and base rates for the 7 IBL behavioral variables (243 sessions, 152,794 trials). P(positive): proportion of positive class. $\sigma_{\text{cross}}(\mu)$: std of per-session means.

Variable	P(positive)	Majority baseline	Explanation
Classification variables			
Movement	0.760	0.760	Session range $[0.08, 0.91]$; **discriminative**
Engagement	0.938	0.938	93% positive; MCC $\approx$ 0 for all methods
Feedback	0.813	0.813	P(fb=1|move=1)=0.845; dominated by movement
Choice	0.521	0.521	69% of sessions in $[0.4, 0.6]$; Type B, $\approx$chance
Block	0.518	0.518	Session std = 0.070; Type B, $\approx$chance
Regression variables			
RT	$\sigma_{\text{global}} = 1.49$, $\sigma_{\text{cross}}(\mu) = 0.57$		Meaningful cross-session variance; **discriminative**
Wheel vel.	$\sigma_{\text{global}} = 7.63$, $\sigma_{\text{cross}}(\mu) = 1.62$		large per-session shifts; $R^2 \approx 0$ across sessions

N Training on New Datasets: Initialization and Frozen Encoders

For each dataset the sessions are split into 5 folds; each fold's sessions are held out for testing, 15% of the remaining sessions are used for early stopping, and the rest for training (hyperparameters in Table 6). The head is a layer normalization followed by a linear classifier on the mean of the

Table 12: Full results on all 7 IBL behavioral variables. MCC for classification; R^2 for wheel velocity and RT, with RT standardized per session across sessions and raw within sessions. Cross-session: leave-one-session-out Ridge probe, scored on the 53 held-out sessions. Within-session: TWS and PerceiverIO are probed on the trials of the 53 held-out sessions pooled (random 60/20/20 split, 5 repeats), which none of them trained on; POYO and Population Statistics on the trials of all sessions pooled; PCA, firing rates and CEBRA, whose feature spaces differ between sessions, per session (random 60/20/20 split). $\pm$: SD over 5 model seeds; rows without $\pm$ are a single run. PerceiverIO uses the same CNN backbone as TWS. Aligned and unaligned PCA are scored on four variables only (Appendix D). Per-variable class distributions in Appendix L.

Method	Movement	Feedback	Choice	Block	Engagement	Wheel (R^2)	RT (R^2)
Cross-session (leave-one-session-out)							
TWS pretrained + finetuned	0.553 ± 0.006	0.070 ± 0.012	$\mathbf{0.040 \pm 0.006}$	$\mathbf{0.031 \pm 0.003}$	0.000 ± 0.000	-0.000 ± 0.001	0.398 ± 0.004
TWS from scratch	$\mathbf{0.565 \pm 0.006}$	0.065 ± 0.011	0.038 ± 0.006	0.026 ± 0.004	0.000 ± 0.000	-0.001 ± 0.003	$\mathbf{0.402 \pm 0.003}$
TWS pretrained	0.428 ± 0.043	0.019 ± 0.019	0.022 ± 0.001	0.019 ± 0.002	-0.000 ± 0.000	-0.001 ± 0.001	0.247 ± 0.036
TWS (random init)	0.274 ± 0.018	0.001 ± 0.002	-0.000 ± 0.004	0.005 ± 0.005	0.002 ± 0.003	-0.002 ± 0.001	0.083 ± 0.011
PerceiverIO	0.303 ± 0.009	$\mathbf{0.102 \pm 0.012}$	0.014 ± 0.009	0.011 ± 0.006	0.000 ± 0.000	-0.005 ± 0.003	0.169 ± 0.003
PCA (aligned)	0.349	–	0.000	-0.005	–	–	0.226
PCA (unaligned)	0.002	–	0.027	0.007	–	–	0.003
POYO	-0.000 ± 0.001	-0.000 ± 0.000	-0.000 ± 0.002	0.000 ± 0.002	0.000 ± 0.000	-0.000 ± 0.001	-0.001 ± 0.001
CEBRA	0.000	0.000	0.005	-0.012	0.000	-0.000	0.000
Population Statistics	0.000	0.000	0.005	-0.002	0.000	-0.000	0.009
Event-time information	*0.908*	*0.000*	*0.031*	*0.020*	*0.000*	*0.003*	*0.443*
Within-session							
TWS pretrained + finetuned	0.585 ± 0.005	0.092 ± 0.034	0.106 ± 0.005	0.095 ± 0.005	0.000 ± 0.000	0.014 ± 0.001	0.543 ± 0.003
TWS from scratch	$\mathbf{0.606 \pm 0.004}$	0.107 ± 0.021	0.109 ± 0.006	0.096 ± 0.007	0.000 ± 0.000	0.013 ± 0.002	$\mathbf{0.547 \pm 0.004}$
TWS pretrained	0.448 ± 0.040	0.030 ± 0.023	0.081 ± 0.006	0.067 ± 0.010	-0.000 ± 0.000	0.009 ± 0.001	0.354 ± 0.044
TWS (random init)	0.360 ± 0.021	0.067 ± 0.005	0.050 ± 0.004	0.047 ± 0.001	$\mathbf{0.031 \pm 0.012}$	0.005 ± 0.001	0.250 ± 0.018
PerceiverIO	0.353 ± 0.008	$\mathbf{0.150 \pm 0.014}$	0.102 ± 0.008	0.069 ± 0.008	0.000 ± 0.000	$\mathbf{0.018 \pm 0.002}$	0.254 ± 0.006
POYO	0.013 ± 0.012	0.001 ± 0.002	0.041 ± 0.011	0.026 ± 0.007	-0.000 ± 0.000	0.005 ± 0.002	0.013 ± 0.003
CEBRA	0.330	0.090	**0.132**	**0.192**	0.011	0.017	0.147
Population Statistics	0.007	0.006	0.007	0.004	-0.001	0.000	0.015
Reference: within-session, per-neuron features							
PCA + Ridge	**0.469**	0.382	0.512	0.440	0.043	**0.344**	**0.378**
Firing Rate + Ridge	0.370	**0.498**	**0.537**	**0.574**	**0.203**	0.284	0.354

output tokens. *Scratch*: all weights randomly initialized. *IBL*: initialized from the self-supervised IBL checkpoint (TWS) or from the IBL PerceiverIO trained on IBL behavior (lr 10^{-4}). *Frozen*: IBL encoder fixed, only the head trained. The POYO row of Table 4 uses this protocol: POYO trains from scratch on the labels of each dataset, with unit and session embeddings learned on the training sessions, and is scored on the held-out sessions of each fold, whose units it has never seen. The CEBRA row fits one supervised model per session on the trial-mean firing rates and scores its embeddings with the leave-one-session-out Ridge probe of the other rows. *Frozen, behavior*: the same, with the encoders of Tables 4 and 20 on both sides, trained on IBL behavior from scratch with the CNN backbone and without the 53 held-out sessions, so the two tokenizers are matched. MCC is computed per test session and averaged over sessions and folds; Table 13 gives mean $\pm$ SD over 3 seeds.

Table 13: Held-out-session MCC after training on each dataset. In the IBL and Frozen columns the initializations are not matched: TWS starts from the self-supervised checkpoint, PerceiverIO from a model trained on IBL behavior with a Transformer backbone. Frozen, behavior: both encoders trained on IBL behavior with the CNN backbone.

		TWS				PerceiverIO			
Dataset	Target	Scratch	IBL	Frozen	Frozen, behavior	Scratch	IBL	Frozen	Frozen, behavior
Steinmetz	Choice	$\mathbf{0.261 \pm 0.005}$	0.238 ± 0.002	0.221 ± 0.004	0.230 ± 0.018	0.231 ± 0.007	0.252 ± 0.015	0.232 ± 0.019	0.218 ± 0.011
Gallego et al.	Reach direction	0.234 ± 0.003	0.225 ± 0.015	0.166 ± 0.023	0.200 ± 0.002	0.281 ± 0.007	$\mathbf{0.300 \pm 0.008}$	0.175 ± 0.007	0.183 ± 0.005
NLB	Movement	0.311 ± 0.013	0.354 ± 0.010	0.333 ± 0.019	0.333 ± 0.011	0.299 ± 0.043	$\mathbf{0.421 \pm 0.013}$	0.379 ± 0.008	0.399 ± 0.008

Within-state time and session identity come together. Trained on each dataset (Table 13), TWS leads on Steinmetz choice from scratch, and PerceiverIO leads on reach direction (Gallego et al.) and, from IBL behavior, on NLB; under the Ridge probe of Table 20 the order on choice and reach is reversed. With both encoders trained on IBL behavior and frozen, TWS leads on choice and reach direction and PerceiverIO on NLB movement (Frozen, behavior). One reading follows from the design. The trajectory of population activity during the movement carries a reach. TWS averages that trajectory within the post-movement state, and PerceiverIO keeps it in time patches. Session identity lives in the same time course: session identity explains 31% of the variance of trained PerceiverIO embeddings against 23% for TWS, and movement, after each session's mean is removed, 1% against 11% (Figure 2B). In the tokenizers we compared, a representation that kept within-state time also kept the session. TWS gives up within-state time and keeps less session identity (23% against 31%), not none. The IBL checkpoint adds to training from scratch on NLB ($+0.043$), the dataset farthest

from IBL. It costs a little on Steinmetz (-0.023), the mouse task IBL already resembles, and on Gallego et al. (-0.009).

O CEBRA Regression Performance

Without per-session standardization of RT, CEBRA gives a negative cross-session R^2 on wheel velocity (-0.057) and on RT (-0.191). Self-supervised CEBRA gives the same picture (-0.056 and -0.188). Per-session contrastive embeddings carry no continuous variable across sessions. With RT standardized per session, as in Table 1, the saved embeddings of one set of CEBRA models give $R^2{=}-0.001$; the same embeddings without standardization give -0.195, against -0.191 for the 5-seed mean above.

P Physiological Variable Decoding

Five IBL sessions in our set have video of the whisker pad and the pupil. All five are among the 53 held-out sessions. We decode whisker motion energy and pupil diameter per trial from the frozen embeddings of the held-out models, with the Ridge probe of Table 1. Cross-session leaves one of the five sessions out; within-session splits the trials of the five at random. No model trains on either variable (Table 14).

Table 14: Whisker motion energy and pupil diameter (R^2) on the 5 held-out IBL sessions with video. Mean $\pm$ SD over 5 model seeds.

	Cross-session		Within-session	
Method	Whisker	Pupil	Whisker	Pupil
TWS from scratch	$\mathbf{0.462 \pm 0.015}$	-0.033 ± 0.008	$\mathbf{0.636 \pm 0.012}$	0.549 ± 0.013
TWS pretrained + finetuned	0.459 ± 0.018	-0.035 ± 0.009	0.607 ± 0.018	0.521 ± 0.045
TWS pretrained	0.090 ± 0.036	-0.035 ± 0.002	0.499 ± 0.026	0.329 ± 0.097
TWS (random init)	0.091 ± 0.016	$\mathbf{0.008 \pm 0.001}$	0.306 ± 0.066	0.116 ± 0.145
PerceiverIO	0.227 ± 0.034	-0.072 ± 0.081	0.619 ± 0.019	0.751 ± 0.012
POYO	-0.081 ± 0.148	-0.039 ± 0.054	0.452 ± 0.057	$\mathbf{0.835 \pm 0.080}$

Across sessions TWS decodes whisker motion twice as well as PerceiverIO, and POYO decodes none of it. Behavioral training produces the result: TWS (random init) and TWS pretrained both stay near 0.09. No model decodes pupil diameter across sessions, and POYO and PerceiverIO decode it well within sessions. Five sessions is a small sample, so we read these numbers as a direction only.

Q State Manifold Characteristics

Do the learned TWS states stand for distinct neural regimes, or are they arbitrary partitions of the trial? We extract the state representations of all 243 IBL sessions and analyze their geometry (Figure 4).

Orthogonal subspaces. Session-demeaned PCA of state centroids reveals that the three states occupy near-orthogonal directions in representational space (Figure 4A). The two principal angles between the per-trial Grassmannian bases of the pretrained model (before behavioral finetuning) confirm this: consecutive pairs are nearly orthogonal (S0↔S1: $87.0°$ and $89.3°$, S1↔S2: $87.3°$ and $89.6°$), while the non-adjacent pair shares structure (S0↔S2: $9.0°$ and $30.6°$). Near-orthogonality between consecutive pairs is encouraged by $\mathcal{L}_{\text{geo}}$ (Section 3), but the partial structure between non-adjacent states and the per-state dimensionality differences emerge without direct supervision.

Dimensionality differs across states. The participation ratio (PR, effective dimensionality of each state's representation) is 1.3 before the stimulus, 2.5 between the stimulus and the movement and 1.3 after the movement (Figure 4B), so the state that the two boundaries enclose is the one that spreads over more directions. Reaction time follows the same order. Within a session, no principal direction of the pre-stimulus state correlates with reaction time ($|r| \leq 0.03$). The third direction of the stimulus-to-movement state achieves $|r|{=}0.26$ and the third direction of the post-movement state $|r|{=}0.49$. Movement follows the same order over the three states: the largest correlation any direction of a state achieves is 0.08, 0.26 and 0.32. The states that the movement boundary delimits therefore carry the timing, which is what TWS tokens keep across sessions (Section 4.2).

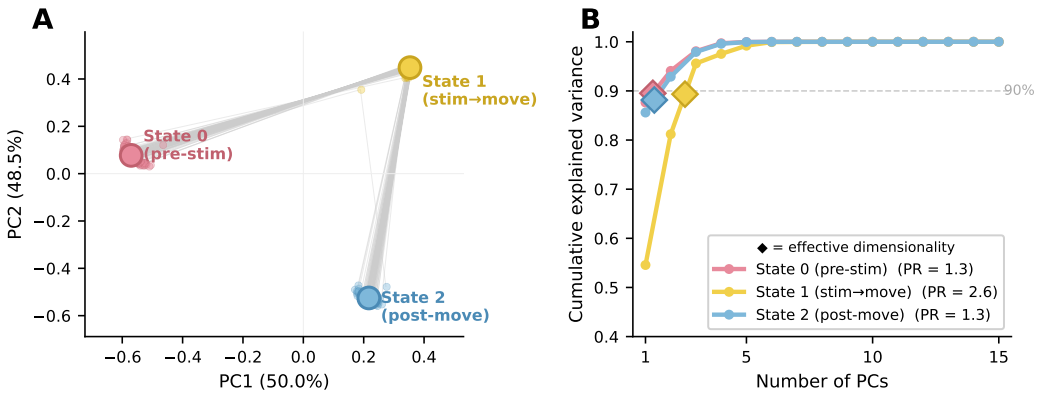

Figure 4: **TWS states occupy orthogonal subspaces with distinct dimensionality. (A)** Session-demeaned PCA of state centroids (243 sessions; pretrained model, first seed). Small dots: per-session centroids; large circles: grand means; gray lines connect the three states within individual sessions. **(B)** Cumulative explained variance per state. Diamonds mark each state's participation ratio (PR, effective dimensionality). Colors and state labels match Figure 1.

R Regime Boundary Information Pathway Analysis

Why two boundaries. The IBL protocol records five event times per trial, but only stimulus onset and first movement fall reliably inside the analysis window and mark a clear shift in population dynamics (Churchland et al., 2012). Cutting the trial more finely than these two events would also leak the labels: a narrow window placed after the first movement stays empty whenever the mouse does not move, and its emptiness states the movement label.

Regime boundary tokenization raises one concern: the state tokens $\boldsymbol{h}$ could pass the boundary time b_{move} straight to the decoder and carry no neural content. Reaction time satisfies $\text{RT} \propto b_{\text{move}} - b_{\text{stim}}$ with b_{stim} constant, so a model that recovers b_{move} from $\boldsymbol{h}$ predicts RT without decoding anything neural.

We decompose the information in TWS state tokens into two pathways: (i) an *architectural timing prior* arising from the interaction between the gate window shape and generic neural temporal structure, and (ii) *trial-specific neural content* within each epoch.

Null result: constant activity carries zero boundary information. The normalized weights $\bar{w}_k(t) = w_k(t) / \sum_{t'} w_k(t')$ guarantee $\sum_t \bar{w}_k(t) = 1$. When neural activity is temporally constant ($\boldsymbol{x}(t) = \boldsymbol{c}$ for all t), the state mean $\bar{\boldsymbol{x}}_k = \sum_t \bar{w}_k(t)\, \boldsymbol{x}(t)$ reduces to $\bar{\boldsymbol{x}}_k = \boldsymbol{c}$ regardless of b_{move}. The gate cannot transmit boundary information without temporal modulation in the input signal. This rules out direct metadata passthrough.

Controlled decomposition. Real neural activity has temporal structure, such as peri-stimulus modulation and movement-onset transients, and that structure interacts with the position of the gate window. We measure the interaction with a *within-session* ablation on 226 IBL sessions (an earlier version of our session list; the analysis takes only the learned gate offsets of the Table 1 model and keeps trials whose first movement falls inside the window) ($N \geq 20$ neurons, ≥ 100 valid trials each). For each session, we compute state means $\bar{\boldsymbol{x}}_k$ under the four conditions of Table 15 and train a 5-fold cross-validated Ridge regressor (trial split within session) to predict b_{move}. The evaluation is within-session because the state means $\bar{\boldsymbol{x}}_k \in \mathbb{R}^N$ are neuron-indexed and two sessions with different neuron sets cannot be compared in that space. The cross-session values of Table 1 need the full TWS pipeline, cross-attention into $\mathbb{R}^d$, to make the representations comparable. The four conditions are:

1. **Real** (b_{move}-aligned gate, matched neural data): standard TWS forward pass.
2. **Gate-only** (trial-shuffled $\boldsymbol{x}$, real gate): preserves the gate window shape but breaks the correspondence between neural content and b_{move}. Any remaining R^2 reflects temporal fingerprinting: generic temporal modulation (e.g., post-stimulus rate increase common to all trials) interacting with gate position.

3. **Neural-only** (shuffled b_{move}, real $\boldsymbol{x}$): applies a mismatched gate to each trial. Any remaining R^2 reflects neural temporal structure that encodes movement timing independently of the gate.
4. **No-gate baseline**: time-averaged firing rate $\bar{\boldsymbol{x}} = T^{-1} \sum_t \boldsymbol{x}(t)$ (no windowing).

Table 15: Regime boundary timing decomposition (226 IBL sessions, within-session 5-fold CV Ridge on raw state means $\bar{\boldsymbol{x}}_k \in \mathbb{R}^{3N}$, mean $\pm$ std across sessions). Not directly comparable to cross-session R^2 in Table 1.

Condition	R^2 (predicting b_{move})	Interpretation
Real (aligned gate + neural)	0.281 ± 0.192	Full signal
Gate-only (shuffled $\boldsymbol{x}$)	0.149 ± 0.105	Temporal fingerprint
Neural-only (shuffled gate)	0.274 ± 0.160	Neural temporal content
No-gate (time average)	0.104 ± 0.116	Firing rate baseline

Results (Table 15). The gate-only signal (R^2=0.149) shows that temporal fingerprinting exists. Temporal modulation common to all trials produces state means that correlate with b_{move} once the window moves with it. The neural-only signal (R^2=0.274) is nearly twice as large and is the larger of the two in 195 of 226 sessions (86%). Trial-specific neural dynamics, not the geometry of the gate, carry most of the movement timing.

Event-time information. The decomposition above is within-session and restricted to trials with a first movement. To measure how much of each label the event time holds in the cross-session setting of Table 1, we fit the same leave-one-session-out Ridge probe on a one-hot encoding of the first-movement bin alone, with no neural data. It scores movement MCC 0.908 and RT R^2=0.443, against 0.565 and 0.402 for TWS trained from scratch (Table 1). Movement is defined by that time: a first movement 0–2 s after stimulus onset (Table 10). RT runs from stimulus onset to the IBL response time, which follows the first movement by a median of 0.19 s. The boundary, in contrast, is the first-movement time wherever it falls. In 11% of the held-out trials the mouse moves within 0.5 s before the stimulus (median 64 ms): the label is then 0 and the boundary falls before the stimulus. The rule "a boundary after the stimulus means a movement" alone scores MCC 0.90, and "any boundary means a movement" scores 0.56. We therefore score the saved embeddings of Table 1 again without these trials (Table 16). TWS does not read the label from them: without them it rises from 0.565 to 0.723, close to the 0.776 that the event time then holds, whereas PerceiverIO changes little. The boundary-placement ablation adds that TWS keeps about a third of its movement decoding when the boundary holds no label information (Table 2b).

Table 16: Cross-session decoding on the 53 held-out IBL sessions with and without the 11% of trials whose first movement precedes the stimulus (probe of Table 1; mean $\pm$ SD over 5 model seeds).

	All trials		Without pre-stimulus movements	
	Movement MCC	RT R^2	Movement MCC	RT R^2
TWS from scratch	0.565 ± 0.006	0.402 ± 0.003	0.723 ± 0.004	0.418 ± 0.003
PerceiverIO	0.303 ± 0.009	0.169 ± 0.003	0.324 ± 0.018	0.181 ± 0.004
Event-time information	0.908	0.443	0.776	0.462

Every representation given the boundary. TWS receives the first-movement time as input, and POYO, CEBRA and PerceiverIO do not. We append the one-hot boundary to every representation and refit the probe of Table 1, with the features z-scored over trials and no labels involved. With the boundary, every representation decodes movement at the level of the boundary alone, and every representation but POYO does so for RT (Table 17). Over the 53 held-out sessions, TWS from scratch plus the boundary differs from the boundary alone by +0.001 in movement MCC (p=0.23, paired Wilcoxon test over sessions) and +0.006 in RT R^2 (p=0.56), and PerceiverIO with the same backbone plus the boundary differs from the boundary alone by +0.001 (p=0.57) and +0.007 (p=0.54). TWS pretrained + finetuned adds 0.002 in movement MCC (p=8×10^{-4}) and nothing in RT. Block is the only target where TWS plus the boundary exceeds the boundary alone by more than 0.01 (+0.014 and +0.018), and neither difference is significant (p=0.41 and 0.21).

Interpretation. Segmenting at the first-movement boundary gives TWS tokens a timing prior that any windowed-averaging scheme at that boundary would produce, and within sessions trial-specific neural content adds to it (neural-only 0.274 against gate-only 0.149). Methods without a gate keep

Table 17: Cross-session decoding on the 53 held-out IBL sessions when each representation is given the one-hot first-movement bin (Table 1 probe with z-scored features; mean ± SD over 5 model seeds; CEBRA is a single run). PerceiverIO uses the same CNN backbone as TWS.

Representation + boundary	Movement MCC	RT R^2	Choice MCC	Block MCC
Event-time information	0.908	0.445	0.032	0.019
TWS from scratch	0.910 ± 0.000	0.450 ± 0.002	0.029 ± 0.004	0.033 ± 0.005
TWS pretrained + finetuned	0.911 ± 0.000	0.450 ± 0.001	0.031 ± 0.006	0.037 ± 0.007
PerceiverIO	0.909 ± 0.001	0.452 ± 0.002	0.021 ± 0.006	0.025 ± 0.005
POYO	0.900 ± 0.008	0.408 ± 0.015	0.021 ± 0.007	0.022 ± 0.005
CEBRA	0.909	0.444	0.027	0.015

no timing structure across sessions: POYO and CEBRA both decode reaction time at R^2=0.00 (Table 1). Movement and RT share their defining event with the boundary. Neither target therefore measures neural decoding beyond the boundary. They measure how much regime-aligned timing structure survives the TWS compression.

S Backbone Sweep

We replace the backbone while keeping TWS, the pretraining and finetuning recipe, and the evaluation fixed: mean pooling followed by layer normalization (no token mixing), an MLP on the pooled tokens, a CNN (a kernel-3 convolution over the six tokens in each of four blocks), a Transformer, Mamba-1 and Mamba-2. Table 3 gives cross-session decoding on IBL for each of them. Decoding on IBL is the same within 0.03 MCC for every backbone that mixes the tokens. Transfer to the other datasets differs between backbones. On Steinmetz and NLB we compare backbones trained on all 243 IBL sessions. From the same TWS tokens, the convolutional backbone achieves Steinmetz choice 0.155 ± 0.006 and NLB movement 0.378 ± 0.008, and the Mamba backbones land within 0.03 of it (Mamba-1 over 4 seeds). The Transformer falls behind on both (0.122 ± 0.054 and 0.299 ± 0.060), with seed-to-seed spreads an order of magnitude wider. On reach direction, scored with the held-out models, the Transformer (0.201 ± 0.006) stays close to the convolutional backbone (0.232 ± 0.010). Figures 5, 6 and 7 show the representation of every tokenizer, backbone and training combination: its session means, the composition of its variance, and how similar the combinations are to one another. Every trained TWS combination separates the two movement states more than every trained PerceiverIO one (1.03–1.88 against 0.48–0.72), and the tokenizer also decides how similar two embeddings are. Among the four token-mixing backbones trained from scratch, the TWS embeddings agree with one another at a linear CKA of 0.87–0.92 and the PerceiverIO embeddings at 0.44–0.64, while a TWS embedding and a PerceiverIO embedding agree at 0.20–0.32. Backbone and tokenizer trade off in the variance as well: with TWS tokens the convolutional backbone keeps 23% of the variance between sessions and the Transformer 18%, against 31% and 28% for the same two backbones on PerceiverIO tokens.

T Coding Rate Reduction of the Representations

Low-dimensional structure is the property TWS is built to extract, so we also measure it directly with the coding rate of Yu et al. (2020). For n trial embeddings normalized to unit length and stacked in $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$,

$$R(\boldsymbol{Z}) = \tfrac{1}{2} \log\det\!\left(\boldsymbol{I} + \tfrac{d}{n\epsilon^2} \boldsymbol{Z}^\top \boldsymbol{Z}\right), \qquad \Delta R(\boldsymbol{Z}, \Pi) = R(\boldsymbol{Z}) - \sum_j \tfrac{n_j}{n} R(\boldsymbol{Z}_j), \tag{3}$$

where Π partitions the trials into groups $\boldsymbol{Z}_j$ of size n_j and ϵ^2=0.5. R is the number of nats needed to encode the embeddings up to precision ϵ, a log-volume of the embedding cloud. ΔR is how much of that volume the grouping explains. It is large when each group is compact and the groups occupy different directions. It is near zero when every group is spread like the whole set. Unlike the variance shares of Figure 2B, which compare group means, ΔR also counts differences in the directions a group occupies. We group trials by movement (behavioral structure) and by session (session identity), using the same trials for both.

Table 18 shows that learning puts behavior into the structure of the representation, and that the tokenizer decides how much. At random initialization neither tokenizer organizes its volume by

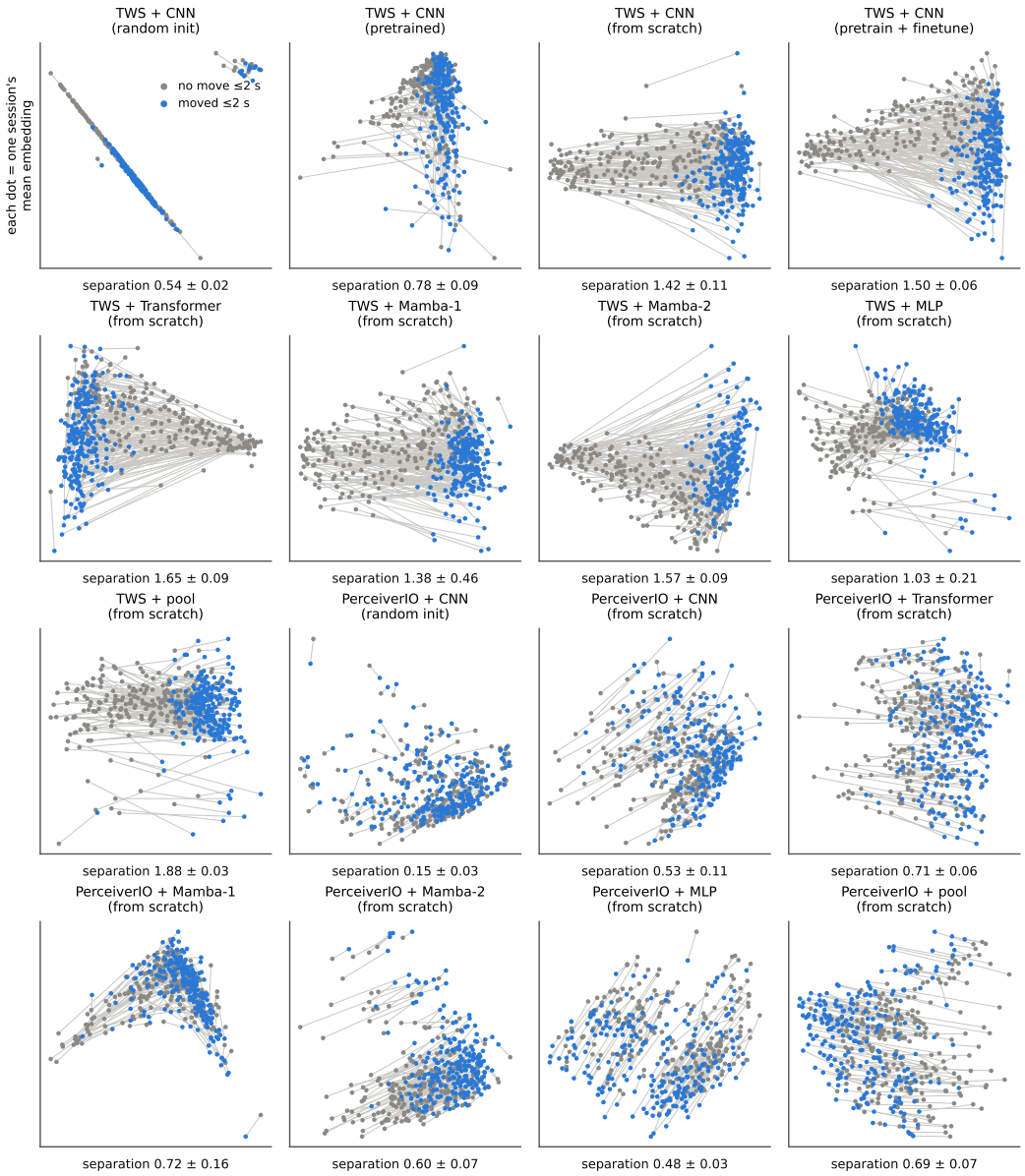

Figure 5: Session means of every tokenizer × backbone × training combination, drawn as in Figure 2A.

behavior: the movement ΔR is 0.05 for TWS and 0.03 for PerceiverIO. Training from scratch raises it to 0.84 for TWS and to 0.38 for PerceiverIO. The PerceiverIO cloud is also the larger one (R=56 against 40), so behavior explains a smaller share of its volume. Training adds session structure as well: the session ΔR rises from 3.0 to 12.7 for TWS and from 3.0 to 14.3 for PerceiverIO. Pretraining moves TWS toward invariance. After pretraining and finetuning the session term falls to 7.0 and most of the movement term remains (0.61). Absolute values depend on ϵ and on the overall volume R, which differs between models, so we compare only large differences. Yu et al. (2023) derive self-attention as a step that compresses tokens against a set of subspaces, which matches TWS tokens being subspace bases. We did not test that link.

U Cross-Dataset Transfer Controls

Level and pattern of each regime. TWS summarizes the population of every regime without neuron identity. Two features of the same three regimes separate the parts of that summary.

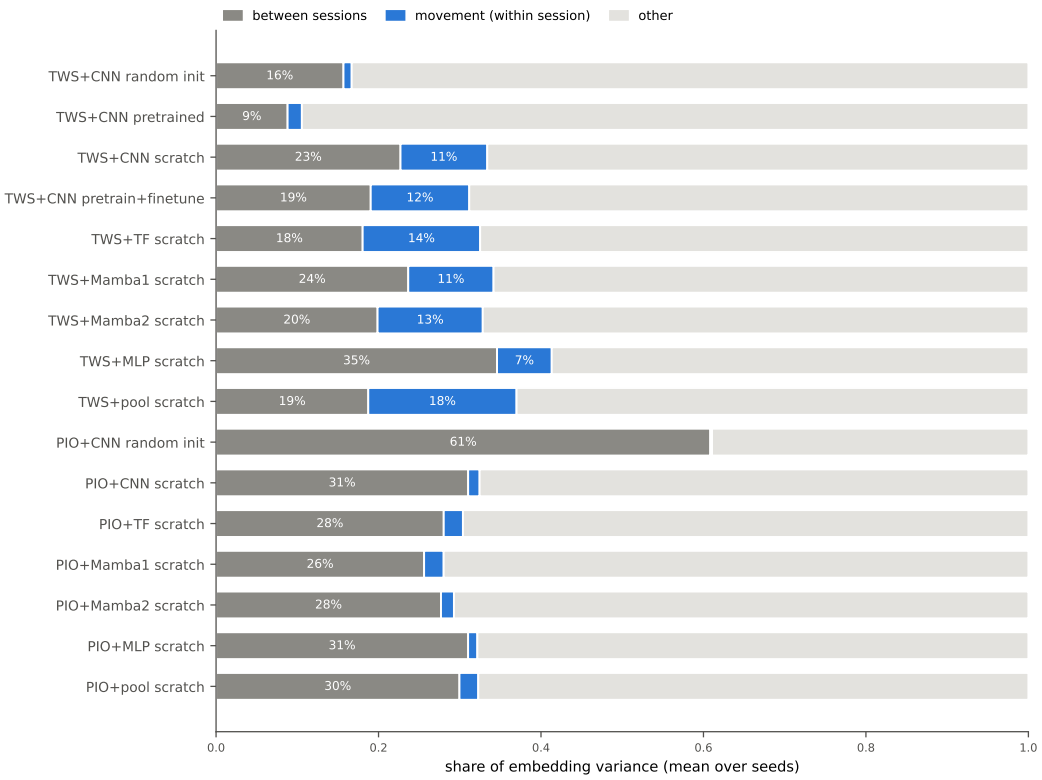

Figure 6: What the trial-embedding variance of each combination is made of, drawn as in Figure 2B.

Table 18: Coding rate R and its reduction ΔR (nats) when IBL trial embeddings are grouped by movement or by session (243 sessions, all trials with a movement label; ϵ^2=0.5). Models trained without the 53 held-out sessions and with the clipped boundary. Mean $\pm$ SD over 5 model seeds. Other backbones follow with Table 3.

Tokenizer (backbone)	Training	R	ΔR (movement)	ΔR (session)
TWS (CNN)	random init	11.65 ± 1.27	0.049 ± 0.004	3.00 ± 0.32
	from scratch	39.71 ± 6.27	0.836 ± 0.108	12.67 ± 2.29
	pretrained + finetuned	28.27 ± 4.16	0.607 ± 0.079	6.99 ± 1.51
PerceiverIO (Transformer)	random init	10.13 ± 0.59	0.030 ± 0.005	3.03 ± 0.17
PerceiverIO (Transformer)	from scratch	56.07 ± 4.97	0.382 ± 0.029	14.26 ± 1.06

Level is the mean activity over all recorded neurons. *Pattern* adds the standard deviation and the 10/25/50/75/90% quantiles over neurons. Both are invariant to the order of the neurons, and both use the boundaries the model receives. The probe is the one of Table 4; Table 19 gives the result. *Level over time* keeps the population mean in every time bin instead of every regime, so it adds the time course inside each regime that TWS averages away. Over Level, the time course raises NLB movement by a third and doubles reach direction. It raises choice tenfold and lowers feedback. PerceiverIO keeps this time course in its patches and leads TWS on NLB movement (Table 20); on reach direction TWS leads despite averaging it away (`scripts/what_transfers_time.py`).

Table 19: Cross-dataset MCC of model-free population summaries, against TWS (Table 4).

Dataset	Target	Level	Pattern	Level over time	TWS
Steinmetz	movement	0.419	0.441	0.475	**0.581**
	feedback	0.160	0.174	0.030	**0.188**
	choice	0.012	0.027	0.126	**0.151**
NLB Gallego et al.	movement	0.354	0.421	**0.476**	0.392
	reach direction	0.057	0.083	0.116	**0.232**

A second identity-free tokenizer and a random-bias reference. Table 20 adds two references to Table 4, under the same probe. PerceiverIO with the CNN backbone, trained from scratch on IBL

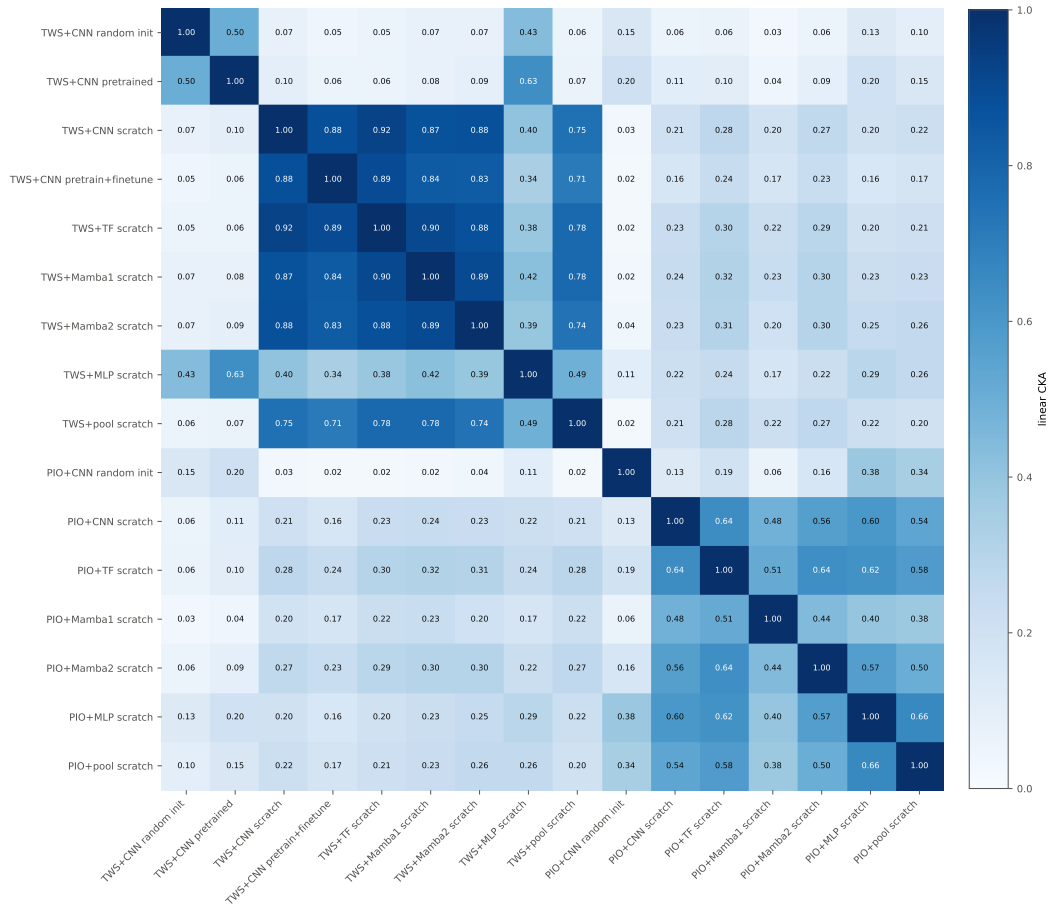

Figure 7: Linear CKA between the trial embeddings of every combination (first seed, 20,000 shared trials).

like TWS and then frozen, has no neuron or session embedding either. It transfers as well as TWS: it leads on Steinmetz choice (p=0.018, paired Wilcoxon test over sessions) and on NLB movement (p=0.006), TWS leads on reach direction (p=2×10^{-6}), and the two do not differ on Steinmetz feedback and movement. Untrained, PerceiverIO already decodes choice at 0.169. TWS (random init) has zero encoder biases. With random biases of the norm of the trained model's biases it reaches three quarters of the trained TWS on choice and on reach direction, and training still adds to choice (p=0.018), reach direction ($p<10^{-10}$) and NLB movement (p=0.012).

Table 20: Cross-dataset MCC of two frozen tokenizers without neuron or session embeddings, both with the CNN backbone (protocol of Table 4; mean ± SD over 5 seeds). The two TWS rows are the runs of Table 4.

	Steinmetz 2019			Gallego et al.	NLB MC_Maze
Method	Choice	Feedback	Movement	Reach direction	Movement
TWS from scratch	0.151 ± 0.015	$\mathbf{0.188 \pm 0.008}$	0.581 ± 0.003	$\mathbf{0.232 \pm 0.010}$	0.392 ± 0.010
TWS (random init)	0.036 ± 0.016	0.077 ± 0.017	0.550 ± 0.007	0.128 ± 0.002	0.247 ± 0.010
TWS (random init, random bias)	0.114 ± 0.009	0.178 ± 0.005	$\mathbf{0.592 \pm 0.002}$	0.173 ± 0.004	0.361 ± 0.005
PerceiverIO, trained on IBL	$\mathbf{0.190 \pm 0.016}$	0.176 ± 0.004	0.572 ± 0.004	0.197 ± 0.013	$\mathbf{0.443 \pm 0.020}$
PerceiverIO (random init)	0.169 ± 0.024	0.158 ± 0.014	0.570 ± 0.009	0.160 ± 0.009	0.419 ± 0.017

Time-shuffled pretraining. We permute the 100 time bins of the pretraining data independently for every neuron and trial. The permutation keeps each trial's spike counts and removes temporal order and within-trial co-firing. The pretraining recipe, the backbone and the 53 excluded sessions are those of TWS pretrained, and each model is scored with the probe of Table 4. Time-shuffled pretraining transfers at least as well as intact pretraining (Table 21): it is slightly higher on Steinmetz choice and feedback and on reach direction ($p\leq0.03$, paired Wilcoxon test over sessions) and equal on both movement targets.

Table 21: Pretrained-only TWS (no behavioral finetuning), cross-dataset MCC, mean ± SD over 5 seeds.

Task	Intact pretraining	Time-shuffled pretraining	TWS (random init)
NLB movement	0.385 ± 0.018	0.385 ± 0.007	0.254 ± 0.004
Gallego et al. reach direction	0.215 ± 0.014	0.232 ± 0.009	0.126 ± 0.004
Steinmetz choice	0.109 ± 0.016	0.125 ± 0.015	0.031 ± 0.014
Steinmetz feedback	0.142 ± 0.011	0.163 ± 0.007	0.085 ± 0.027
Steinmetz movement	0.563 ± 0.003	0.568 ± 0.005	0.552 ± 0.004

V Event-Time Information on Cross-Dataset Tasks

Why the boundary decodes IBL but not the other datasets. The boundary is the first-movement time of the trial. How much it decodes follows from how each target is defined (Table 22). On IBL movement is a function of that time and RT follows it. On the other datasets movement follows the speed of the wheel or the cursor over the whole trial, which the onset time only constrains, and choice and reach direction do not involve it at all. Each reference uses the probe of its own table: the IBL row the leave-one-session-out Ridge probe of Table 1, the cross-dataset row the per-session leave-one-out RidgeClassifierCV of Table 4.

Table 22: How each target is defined and what the boundary position alone decodes (MCC; RT: R^2).

Dataset	Target	Definition	Event-time information
IBL	movement	a first movement occurs within 2 s of stimulus onset	0.908
	RT	stimulus onset to the IBL response time	0.443
Steinmetz	movement	mean wheel speed of the trial above the session median	0.319
	feedback	reward or no reward	0.149
	choice	wheel turned left, against right or no-go	−0.009
NLB	movement	mean cursor speed of the trial above the median	0.260
Gallego et al.	reach direction	which of the reach targets the trial used	0.010

On IBL the event that places the first-movement boundary defines movement and RT, so the boundary alone decodes them better than any model (Table 2). The other datasets have targets that the boundary does not define, and we ask there whether TWS adds anything beyond it. We fit the cross-dataset probe (per-session leave-one-out, standardized features, RidgeClassifierCV, MCC) on a one-hot encoding of each trial's boundary bin, the boundary TWS itself receives. Steinmetz movement follows wheel speed rather than the boundary event, and Steinmetz choice and Gallego et al. reach direction are unrelated to boundary timing by design (Table 23).

Table 23: Cross-dataset MCC of the boundary position alone versus neural representations (TWS: the model of Table 4, trained without the 53 held-out IBL sessions; mean ± SD over 5 seeds; PCA and TWS (random init): mean over 5 seeds).

Dataset	Target	Event-time information	PCA(50)	TWS (random init)	TWS from scratch
Steinmetz	choice	−0.009	0.011	0.036	$\mathbf{0.151 \pm 0.015}$
	feedback	0.149	0.012	0.077	$\mathbf{0.188 \pm 0.008}$
	movement	0.319	0.390	0.550	$\mathbf{0.581 \pm 0.003}$
Gallego et al.	reach direction	0.010	0.090	0.128	$\mathbf{0.232 \pm 0.010}$
NLB	movement	0.260	0.144	0.247	$\mathbf{0.392 \pm 0.010}$

Choice within movement groups. Steinmetz choice is correlated with movement, because the fraction of trials with a movement differs between the two classes. A representation that decodes movement could therefore decode choice through it. We decode choice separately within the trials with a movement and within those without, with the probe of Table 4 (Table 24). TWS decodes choice in both groups, while TWS (random init) and PCA stay near chance among the trials without a movement. The boundary position alone decodes choice in neither group. Steinmetz choice is a left turn against a right turn or no turn, so the absence of a movement boundary on no-go trials could also carry it. On the go trials alone, left against right, TWS still leads TWS (random init), PCA and the boundary position.

Table 24: Steinmetz choice MCC within movement groups and on go trials, left against right (cross-session linear probe as in Table 4; mean ± SD over 5 seeds; Event-time information as in Table 23).

Method	All trials	Moved	Did not move	Go trials
TWS from scratch	$\mathbf{0.151 \pm 0.015}$	$\mathbf{0.122 \pm 0.010}$	$\mathbf{0.091 \pm 0.011}$	$\mathbf{0.105 \pm 0.007}$
TWS (random init)	0.032 ± 0.012	0.086 ± 0.015	0.004 ± 0.007	0.076 ± 0.005
PCA (per session)	0.011 ± 0.010	0.054 ± 0.015	-0.001 ± 0.003	0.043 ± 0.009
Event-time information	-0.009	0.005	-0.005	0.031

W Pretraining Loss Ablation

Backbone note. The analyses in this section were run with an earlier backbone on the same tokenizer (a multi-scale state-space model) and report accuracy where marked; all rows of the section share that backbone.

Table 25: Pretraining loss ablation (IBL, 243 sessions, finetuned). Movement and feedback accuracy.

Configuration	Movement	Feedback
Full (MSE + diversity + geodesic + state masking)	0.873	0.817
MSE + diversity (no geodesic)	0.875	0.822
MSE + geodesic (no diversity)	0.876	0.821
MSE only (no geometric losses)	0.883	0.823
No pretraining (from scratch)	0.889	0.824
+ Behavioral supervision during pretraining	0.887	0.825

With 243 sessions every pretraining configuration converges to the same place after finetuning (Table 25). Movement accuracy spans 0.016 (0.873–0.889) and feedback 0.008 (0.817–0.825). Training from scratch, with no pretraining at all, achieves 0.889 movement against 0.873 for the full pipeline. The amount of data decides this convergence. On one of the smaller datasets the IBL checkpoint still adds something: it improves on training from scratch on NLB, and not on Steinmetz or Gallego et al. (Appendix N). With 243 sessions, finetuning finds an effective representation from any initialization, which leaves the geometric prior nothing to add.

X Residual Decoding: Orthogonality of Task and Session Information

We ask whether movement and session identity occupy orthogonal subspaces. We remove the linear session component from the trial embeddings by subtracting each session's mean embedding, computed from that session's own unlabeled trials. We then decode movement from the residuals with the leave-one-session-out probe of Table 1 (Table 26; first model seed).

Table 26: Cross-session movement MCC before (Original) and after (Residual) removing each session's mean embedding. CNN backbone for TWS and PerceiverIO; models trained without the 53 held-out sessions, scored on them (first model seed).

Model	Original ($\mathbf{h}$)	Residual ($\mathbf{h} - \bar{\mathbf{h}}_s$)
TWS pretrained + finetuned	0.555	0.561
TWS from scratch	0.564	0.584
TWS (random init)	0.301	0.294
PerceiverIO	0.310	0.324
POYO	0.000	0.001
CEBRA	0.000	0.000

Y Session-Level Statistics

The ± in the main tables is the spread over 5 model seeds with the 53 held-out sessions fixed. It says how reproducible a number is, not how consistent it is across sessions. Table 27 tests the main comparisons over sessions instead. Each held-out session contributes the cross-session score of each representation, averaged over model seeds. We then report the mean paired difference with a 95%

bootstrap interval over sessions, a Wilcoxon signed-rank test, and the fraction of sessions where the first representation scores higher. TWS beats PerceiverIO in most sessions. Neither TWS nor PerceiverIO adds to the boundary alone on movement or on reaction time.

Seed spread in Figure 3A. Over 5 model seeds TWS + CNN from scratch keeps session identifiability 0.489 ± 0.017 and decodes cross-session movement at MCC 0.565 ± 0.006 on the held-out sessions. At random init it keeps 0.150 ± 0.004 and decodes 0.274 ± 0.018.

Table 27: Paired comparisons over the 53 held-out IBL sessions (cross-session protocol of Table 1; TWS is the from-scratch model with the convolutional backbone unless named). Difference: first minus second, mean over sessions with a bootstrap 95% interval; p: Wilcoxon signed-rank; Sessions: fraction of sessions where the first is higher. Each session's score is averaged over 5 model seeds; PerceiverIO uses the same CNN backbone.

	Movement MCC			RT R^2		
Comparison	Difference [95% CI]	p	Sessions	Difference [95% CI]	p	Sessions
TWS vs. PerceiverIO	$+0.261$ [$+0.210$, $+0.313$]	3.3×10^{-9}	91%	$+0.233$ [$+0.170$, $+0.299$]	2.4×10^{-8}	91%
TWS vs. aligned PCA	$+0.216$ [$+0.159$, $+0.273$]	1.2×10^{-7}	83%	$+0.176$ [$+0.103$, $+0.254$]	4.0×10^{-7}	85%
TWS + boundary vs. event-time information	$+0.001$ [-0.000, $+0.003$]	0.23	51%	$+0.006$ [-0.004, $+0.016$]	0.56	49%
TWS + boundary vs. PerceiverIO + boundary	$+0.001$ [-0.001, $+0.002$]	0.35	57%	-0.001 [-0.015, $+0.014$]	0.72	51%
PerceiverIO + boundary vs. event-time information	$+0.001$ [-0.001, $+0.003$]	0.57	49%	$+0.007$ [-0.010, $+0.027$]	0.54	57%
TWS vs. TWS (random init)	$+0.290$ [$+0.250$, $+0.330$]	4.7×10^{-10}	98%	$+0.319$ [$+0.245$, $+0.402$]	6.9×10^{-9}	92%
From scratch vs. pretrained + finetuned	$+0.012$ [-0.005, $+0.029$]	0.18	57%	$+0.004$ [-0.003, $+0.011$]	0.02	70%

Z Session Directions and Behavior Decoding Directions

We ask whether the directions that separate sessions overlap the directions a linear probe uses to decode behavior. We work on the z-scored trial embeddings of each model, first model seed. One pooled linear regression first removes the part of every embedding that movement and reaction time explain, so that sessions differing only in behavior, a slower mouse for instance, do not define the session directions. The session directions are the top principal directions of the 243 session means of this residual, as many as explain half of their variance (k). The behavior decoding directions are the weights of Ridge probes for movement and for reaction time, fit on the 190 training sessions.

Table 28 reports two tests on the 53 held-out sessions. In the first, projecting out the k session directions leaves movement decoding almost unchanged for every model, -3% for TWS and for PerceiverIO, and projecting out k random directions changes nothing. Reaction time overlaps more with the session directions: TWS loses 37% and PerceiverIO 17%. What TWS keeps still exceeds everything PerceiverIO decodes. In the second test, the two behavior decoding directions identify the session of a trial no better than two random directions. Session information therefore stays out of the behavior decoding directions, and it is not fully separable from reaction time. The regression step matters for this conclusion. Without it the session directions also carry the behavior differences between sessions, and projecting them out costs TWS 18% of movement and 59% of reaction-time decoding.

Table 28: Session directions (after removing what behavior explains) and behavior decoding directions on the 53 held-out IBL sessions, first model seed. Arrows: cross-session decoding before and after projecting out the k session directions; projecting out k random directions leaves it unchanged. Session identification: nearest-centroid accuracy over the 53 sessions (chance 0.019).

		Remove session directions		Session identification from	
Model	k	Movement MCC	RT R^2	behavior 2D	random 2D
TWS from scratch	3	$0.564 \to 0.546$	$0.401 \to 0.252$	0.050	0.061
TWS pretrained + finetuned	3	$0.550 \to 0.526$	$0.398 \to 0.251$	0.054	0.072
TWS (random init)	1	$0.308 \to 0.302$	$0.071 \to 0.106$	0.080	0.087
PerceiverIO	2	$0.312 \to 0.303$	$0.170 \to 0.141$	0.043	0.065

Where session information sits. We fit the session directions on the 190 training sessions alone, after the same regression step, and identify the session of each trial of the 53 held-out sessions from different subsets of dimensions (Table 29). In the trained models the top session directions carry more session information than as many random directions, and removing them does not lower session identification. Any two random directions identify the session above chance. Session information is therefore spread over the whole embedding rather than held in a few directions. This holds

for TWS and for PerceiverIO, which differ in how much of it they keep. The two behavior decoding directions carry no more of it than two random directions.

Table 29: Nearest-centroid session identification on the 53 held-out IBL sessions from subsets of embedding dimensions (chance 0.019; two random directions give 0.05–0.10). First model seed.

		Session directions (90% of the variance of session means)				
Model	All	k	Only these	All but these	k random	Behavior directions (2D)
TWS from scratch	0.232	8	0.187	0.248	0.145	0.050
TWS pretrained + finetuned	0.212	7	0.185	0.200	0.132	0.054
TWS (random init)	0.126	3	0.102	0.098	0.107	0.080
PerceiverIO	0.417	12	0.332	0.425	0.220	0.043

AA Pretraining Information Decomposition

We measure how much nuisance information each embedding keeps at three levels: lab, mouse and session (Table 30). For each level we train a 5-fold cross-validated logistic regression probe (ℓ_1 penalty, saga solver) on 50K subsampled trial embeddings. MI is $H(Y) - \mathrm{CE}_{\mathrm{probe}}$, with $H(Y)$ the entropy of the uniform label distribution at that level: 12 labs, 109 mice, 243 sessions.

Table 30: Nuisance information retention: $\mathrm{MI}(\mathbf{h}; Y)/H(Y)$ at each hierarchical level. Lower values indicate greater invariance to the corresponding factor. TWS and PerceiverIO are trained without the 53 held-out sessions, POYO pretrained on all sessions and CEBRA fits one model per session; PerceiverIO uses the same CNN backbone as TWS.

Factor	TWS pretrained + finetuned	TWS from scratch	TWS (random init)	PerceiverIO	POYO	CEBRA
Lab	0.174	0.209	0.078	0.281	0.426	0.024
Mouse	0.334	0.361	0.111	0.570	0.809	0.113
Session	0.420	0.440	0.137	0.733	0.959	0.220

AB Site Invariance Across Recording Labs

The IBL Brain Wide Map spans 12 laboratories in 4 countries, each with its own Neuropixels rigs. Recording site is a common source of batch effects in multi-site studies. We test whether TWS state tokens cluster by recording lab or by computational state (Figure 8). In the t-SNE map the tokens group by state (silhouette 0.53) and not by lab (−0.08), subject (−0.49) or session (−0.62). The session information that a linear probe still finds sits in directions that behavior decoding does not use (Section 4.7).



Figure 8: State tokens of TWS pretrained (the 53 held-out sessions excluded from training) for all 243 IBL sessions. **(A)** t-SNE of 2,400 state tokens (800 per state) colored by state. **(B)** The same coordinates colored by recording lab. **(C)** Silhouette scores of the t-SNE coordinates for state, lab, subject and session.

AC Trial Embeddings of Held-Out Sessions

Figure 9 maps the trial embeddings of 10 held-out sessions in two dimensions, for every method of Table 1 that produces trial embeddings.

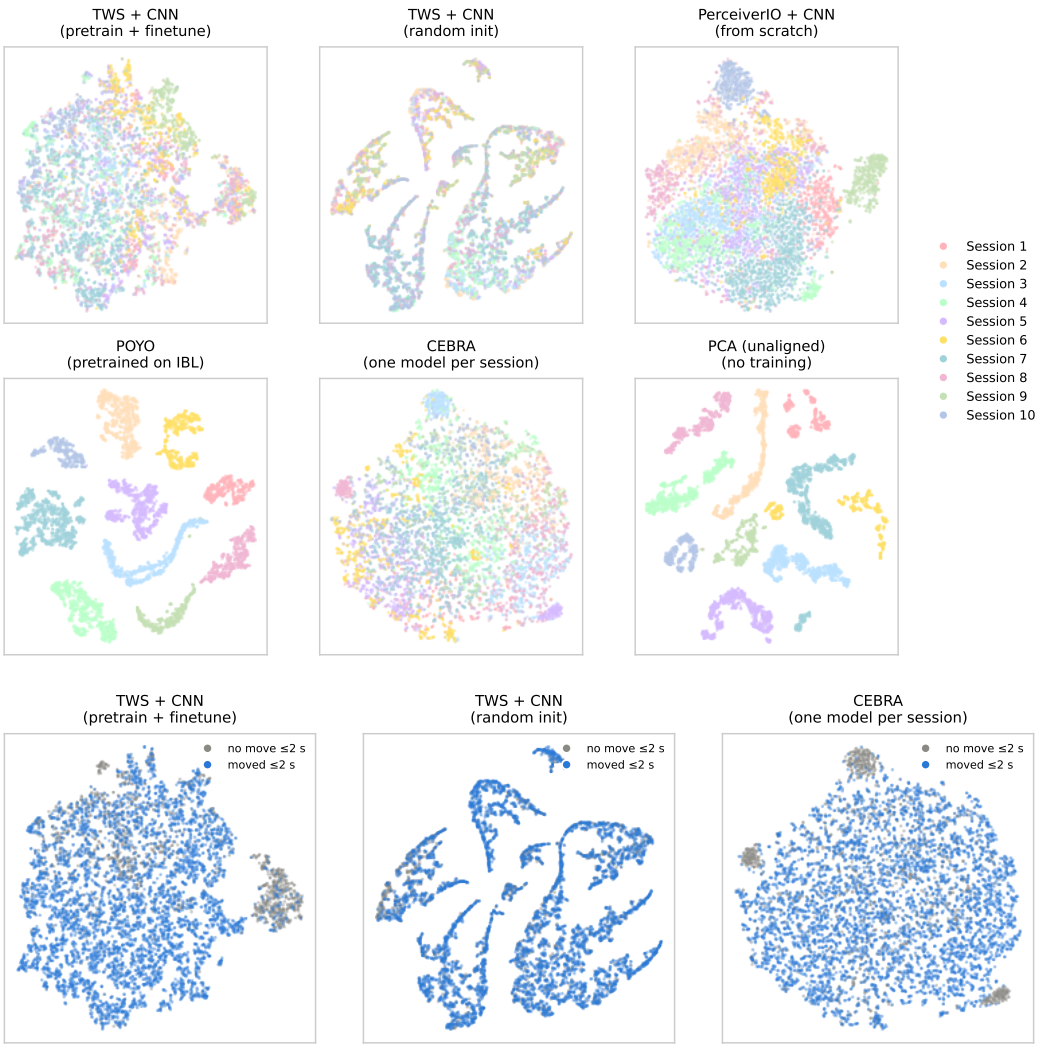

Figure 9: 2D t-SNE of the trial embeddings of 10 held-out IBL sessions. Top: one color per session. Bottom: the same maps of TWS pretrained + finetuned, TWS (random init) and CEBRA, colored by whether the mouse moved within 2 s. POYO and PCA separate the sessions, while TWS mixes them and still separates the two behaviors. Nearby trials from other sessions share the movement label above chance for TWS pretrained + finetuned ($+0.13$), but hardly for TWS (random init) ($+0.03$) or CEBRA ($+0.02$), whose separate per-session models overlap only by chance (15 nearest neighbors from other sessions).

AD Cross-Dataset Replication: Steinmetz 2019

We repeat the session identifiability analysis of Figure 3A on Steinmetz 2019 with the frozen TWS of Table 4, which never saw this dataset (Figure 10). Per-neuron features identify every session and decode movement close to chance (50%, a median split of wheel speed). TWS trained on IBL keeps a third of that session information and decodes movement best; TWS (random init) retains almost none and decodes movement nearly as well, as in Table 4.

AE A Shared Decoding Direction Explains Data Efficiency

For every model and each of the 53 held-out sessions we z-score the trial embeddings and take one decoding direction of the session. For movement it is the mean embedding of the trials with a first movement within 2 s minus the mean of the trials without. For reaction time it is the covariance of each embedding dimension with reaction time, z-scored within the session. *Direction consistency* is

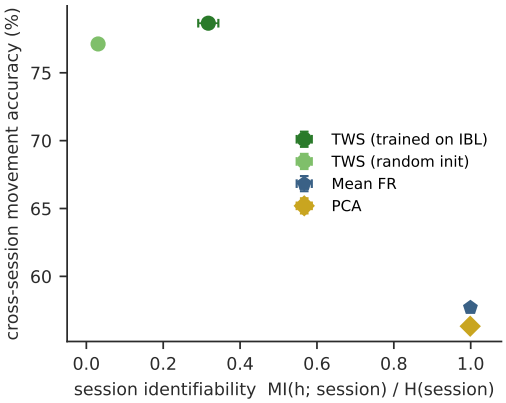

Figure 10: Session identifiability (normalized MI of a 39-class probe) against cross-session movement accuracy on the 39 Steinmetz sessions. Mean ± SD over 5 model seeds.

the cosine between the unit decoding direction of a session and the mean direction of the other held-out sessions. The *five-session score* is the movement MCC or the reaction-time R^2 of a Ridge probe fit on 5 sessions drawn from the 190 training sessions, with the regularization chosen on a sixth session and the score averaged over 20 draws; R^2 is clipped to $[-1, 1]$ per session, as in Table 1. *Session identification* is the accuracy of a nearest-centroid classifier over the 53 held-out sessions, 5-fold within each session. Table 31 gives all three quantities.

Across the six models the five-session score follows direction consistency for both targets (Spearman ρ=0.94 for each). It does not follow session identification (ρ=−0.09 for movement, 0.03 for reaction time). The same relation holds inside every model: sessions whose decoding direction agrees with the others also score higher on five sessions (ρ from 0.58 to 0.80, all $p<10^{-5}$). Part of that within-model relation can come from sessions with a stronger behavior signal, which have both a steadier decoding direction and better decoding, so we rest the claim on the comparison across models.

Table 31: Direction consistency, five-session decoding and session identification on the 53 held-out IBL sessions (mean over 5 model seeds). PerceiverIO uses the same CNN backbone as TWS.

	Movement		Reaction time		
Model	Consistency	Five-session MCC	Consistency	Five-session R^2	Session identification
TWS from scratch	0.82	0.513	0.93	0.313	0.272
TWS pretrained + finetuned	0.82	0.507	0.91	0.330	0.228
TWS pretrained	0.55	0.383	0.56	0.009	0.132
TWS (random init)	0.50	0.182	0.62	0.021	0.123
PerceiverIO	0.64	0.254	0.73	0.073	0.406
PerceiverIO (random init)	0.40	0.138	0.40	−0.027	0.347