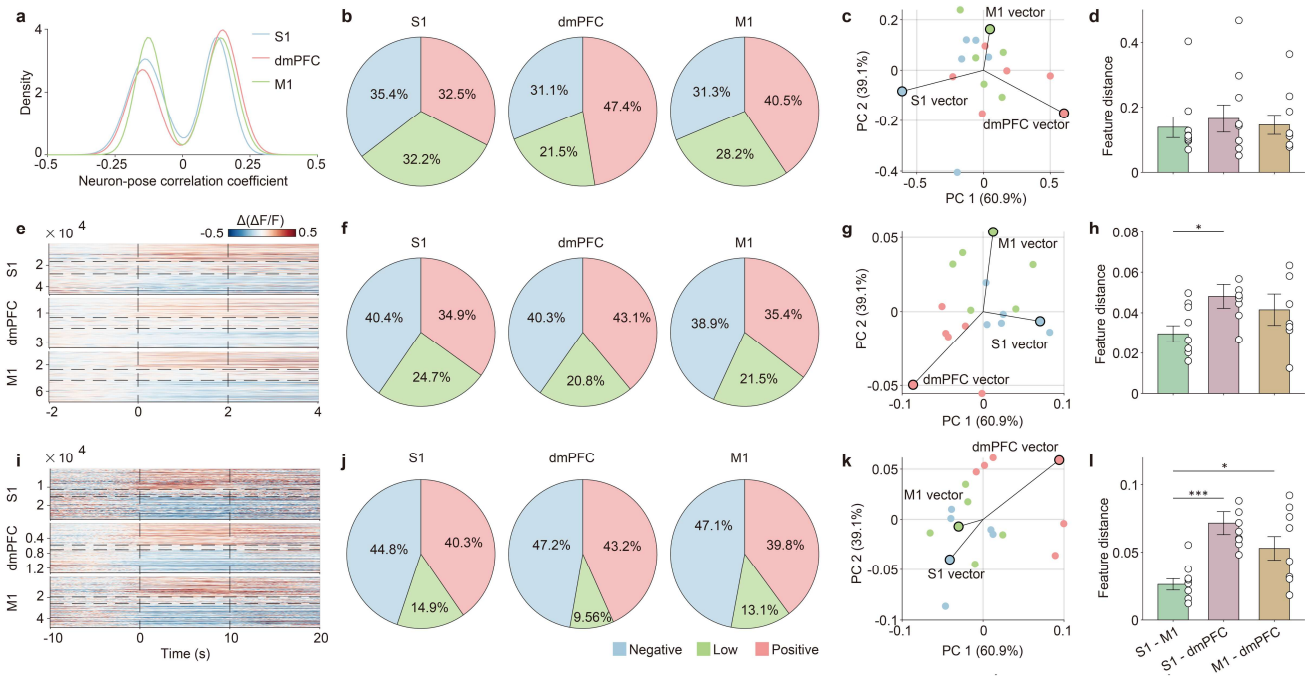
Table-of-contents

Supplementary Figures 2
Supplementary Notes 27
Single-neuron representations diversify across behavioral timescales 27
Neural population activity captures behavioral representations across multiple timescales 28
Global continuity and local conservation of cortical dynamics 30
Behavioral episodes constrain cortical activity into reproducible latent dynamical regimes 31
A conserved neural dimension captures recurrent dynamics during social interaction 33
A brain-inspired recurrent model recapitulates neural dynamics and their perturbation-dependent behavioral effects 34
Supplementary Methods 37
The correlation between poses and neural activities 37
The alignment of neural activities of movements and sequences 37
Neural activity consistency analysis 37
Neural embedding using CEBRA 38
The parameter settings and validation of NPC-EDREN 38
Optogenetic perturbation in NPC-EDREN 40
Reduced-rank regression 42
Dynamic quantification using Lyapunov exponents 43
References 45

Supplementary Figures



Supplementary Fig. 1| Single-neuron encoding to poses, movements, and sequences

a, Distribution of neuron-pose correlation coefficients in S1, dmPFC, and M1.

b, Fractions of neurons with negative, low, or positive pose-related correlations in each cortical area. The threshold of correlation coefficients is ±0.1.

c, Principal-component projection of cortical feature vectors for pose encoding. Dots represent pose-related features, and arrows indicate the feature-vector directions associated with S1, dmPFC, and M1 in the shared feature space.

d, Pairwise feature distances between cortical areas for pose encoding (one-way ANOVA, mean±SEM, N=10, P value between S1-M1 and S1-dmPFC is 0.5858, between S1-M1 and M1-dmPFC is 0.8806, and between S1-dmPFC and M1-dmPFC is 0.6924).

e, Neural activity changes associated with mid-hierarchy movement motifs. Heatmaps show event-aligned calcium activity changes across neurons in S1, dmPFC, and M1.

f, Fractions of neurons with negative, low, or positive movement-related responses in each cortical area.

g, Principal-component projection of cortical feature vectors for movement-motif encoding.

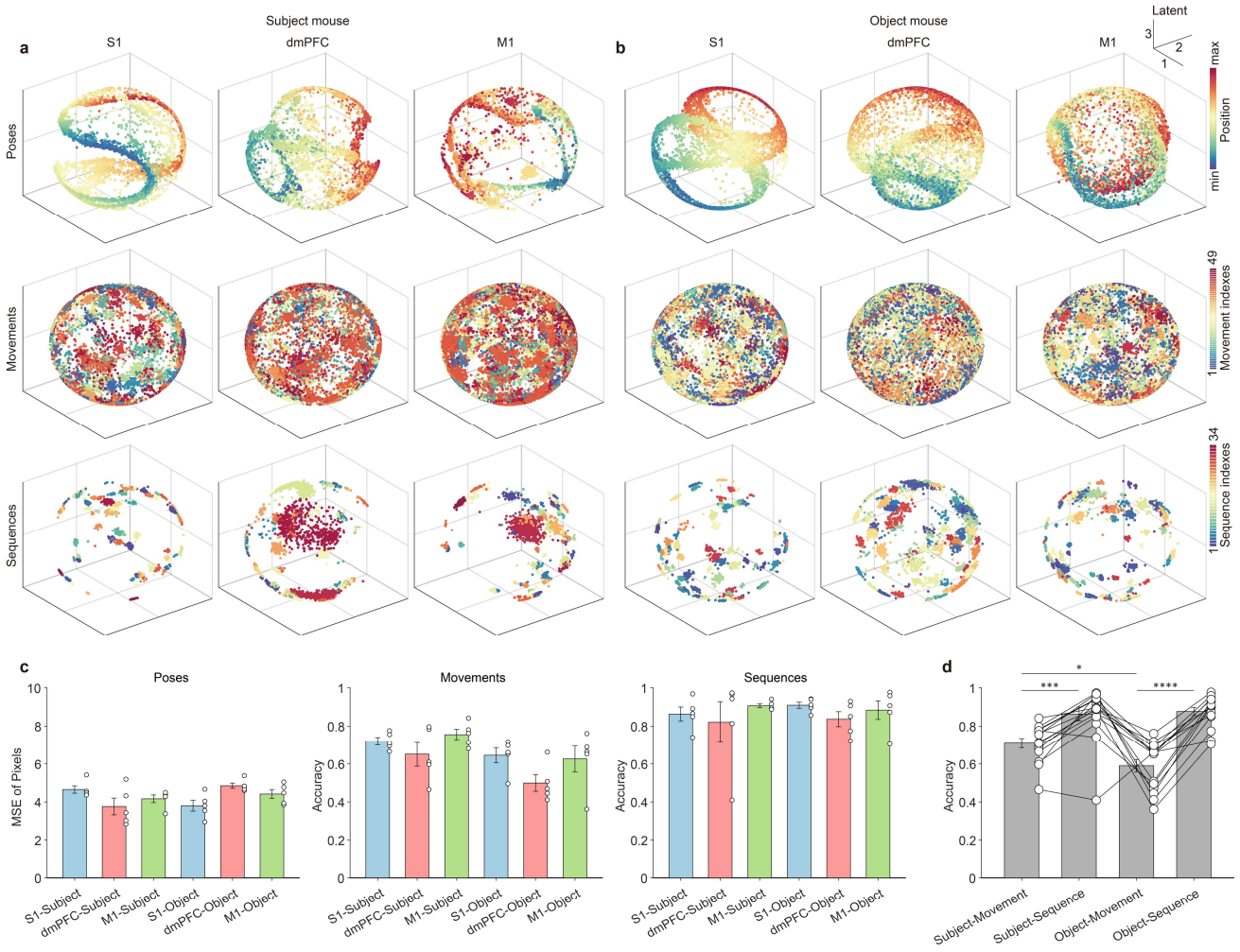
h, Pairwise feature distances between cortical areas for movement-motif encoding (one-way ANOVA, mean±SEM, N=10, P value between S1-M1 and S1-dmPFC is 0.0403, between S1-M1 and M1-dmPFC is 0.1775, and between S1-dmPFC and M1-dmPFC is 0.4485).

i, Neural activity changes associated with high-hierarchy behavioral sequence composites. Heatmaps show event-aligned calcium activity changes across neurons in S1, dmPFC, and M1 over a longer temporal window.

j, Fractions of neurons with negative, low, or positive sequence-related responses in each cortical area.

k, Principal-component projection of cortical feature vectors for sequence-composite encoding (one-way ANOVA, mean±SEM, N=10, P value between S1-M1 and S1-dmPFC is 0.0002, between S1-M1 and M1-dmPFC is 0.0177, and between S1-dmPFC and M1-dmPFC is 0.0866).

l, Pairwise feature distances between cortical areas for sequence-composite encoding. *$P < 0.05$, ***$P < 0.001$.



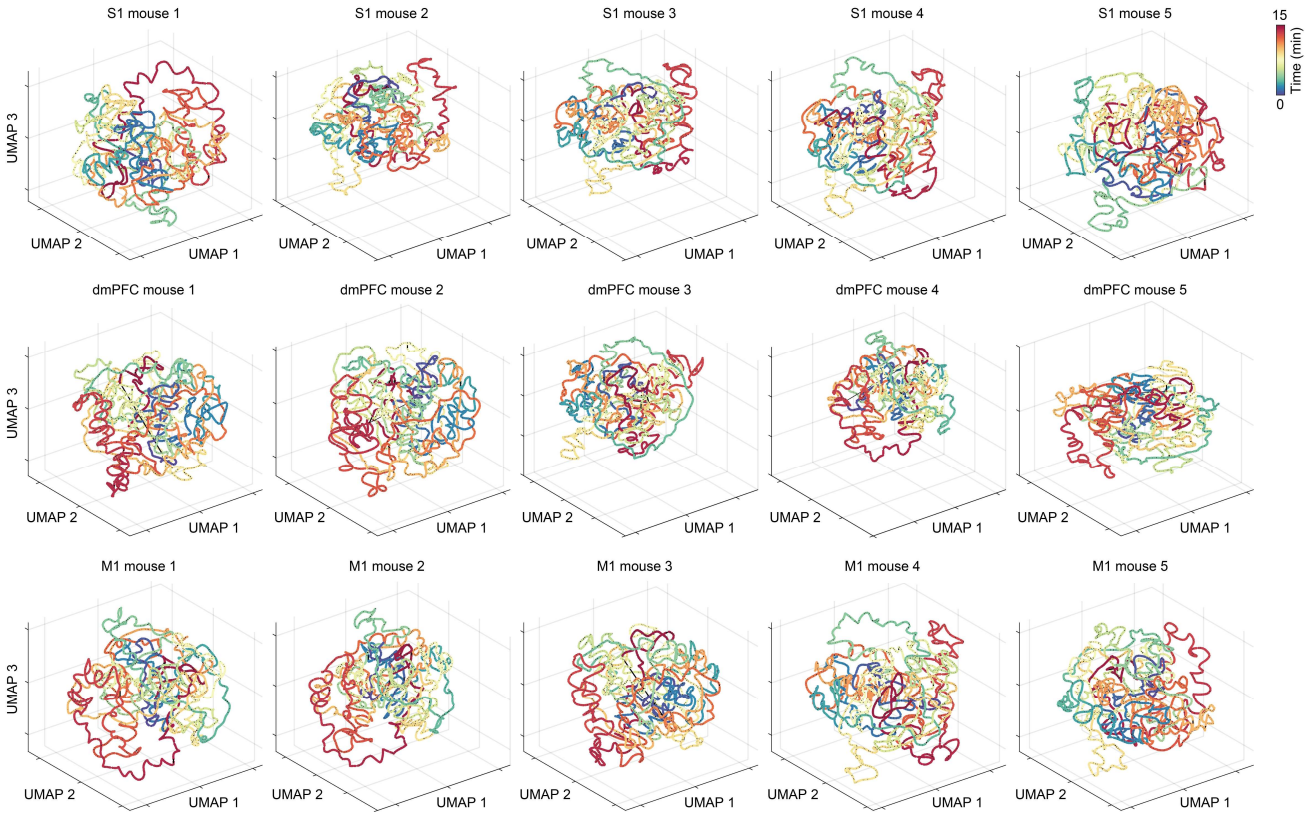
Supplementary Fig. 2| Neural-behavioral joint embeddings and decoding performance across behavioral timescales, cortical regions, and individuals.

a, CEBRA embeddings of the subject mouse's behaviors and the subject mouse's neural activity across three behavioral timescales, including pose, movement, and sequence.

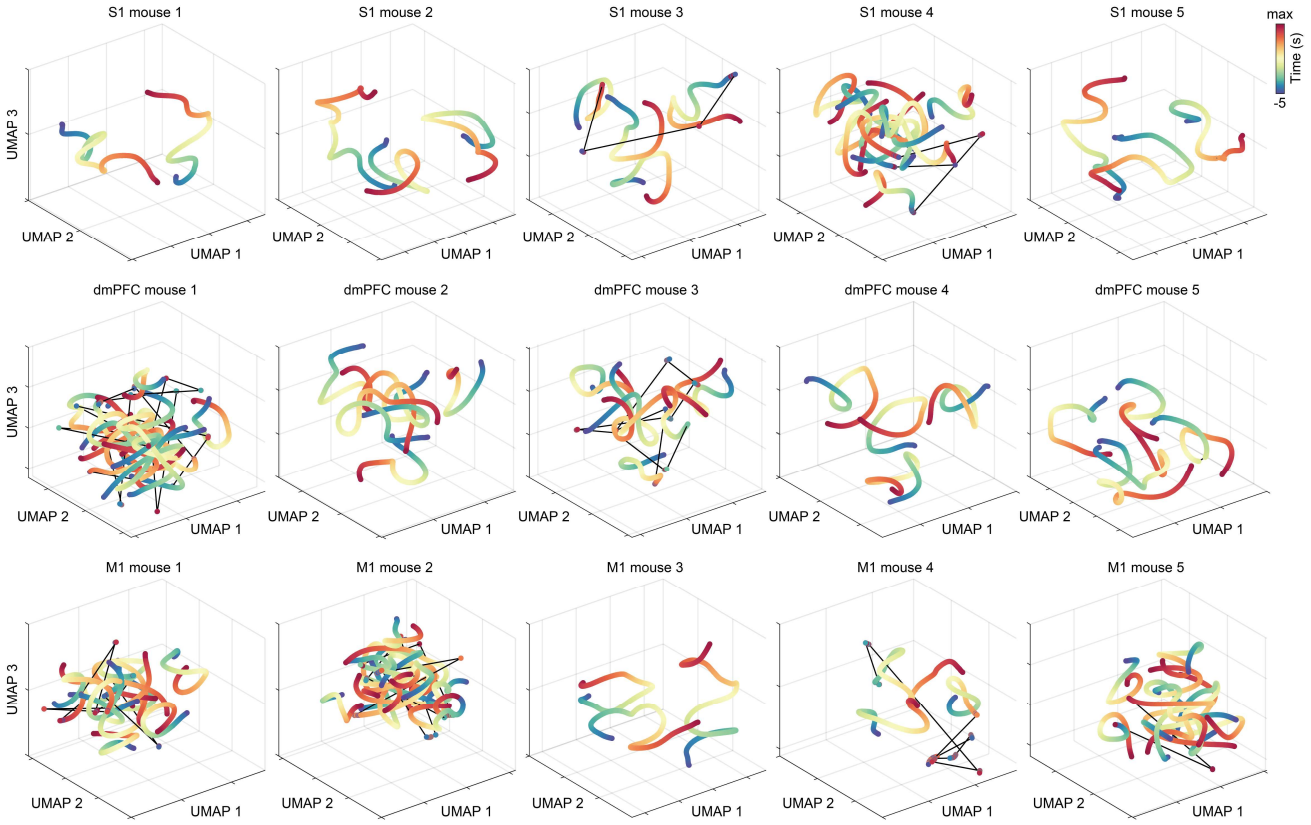
b, CEBRA embeddings of the object mouse's behaviors and the subject mouse's neural activity, using the same behavioral timescales and cortical inputs as in panel **a**.

c, Reconstruction performance of CEBRA embeddings across behavioral timescales. From left to right are the reconstruction error of poses, decoding accuracy for movement, and decoding accuracy for sequence (mean±SEM, N=5, the values are 4.27±0.12, 0.65±0.02, and 0.87±0.02).

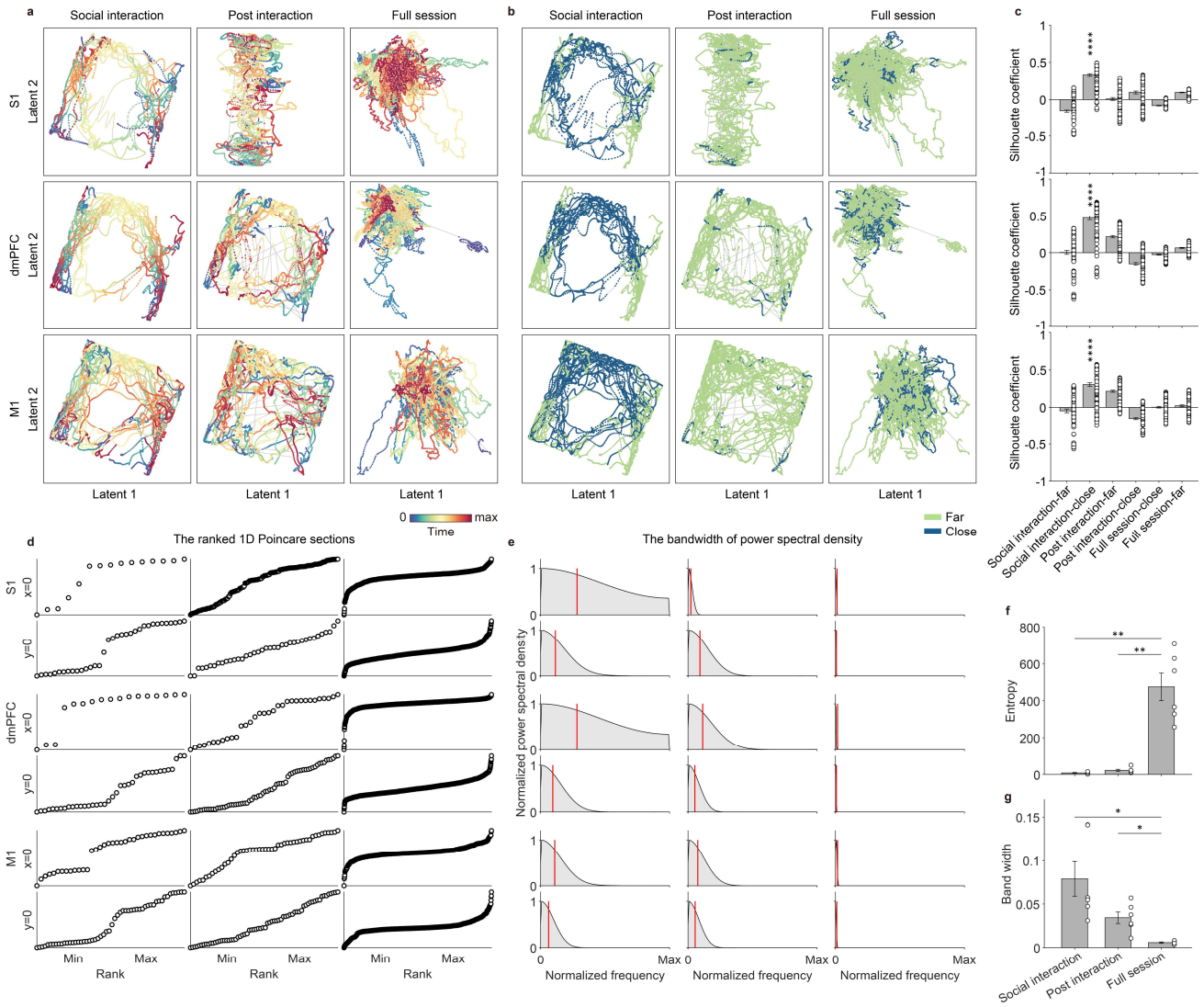
d, Comparison of decoding accuracy across subject/object and movement/sequence embeddings (One-way ANOVA followed by Tukey multiple comparison test, mean±SEM, N=15, P values are 0.0026, 0.0291, 0.0010, <0.0001, 0.9891, and <0.0001). *P < 0.05, ***P < 0.001, ****P < 0.0001.



Supplementary Fig. 3| UMAP embeddings of full behavior session neural dynamics. Three-dimensional UMAP projections of full-length neural activity from individual trials. Color gradients indicate time progression within each trial, from blue (start) to red (end).



Supplementary Fig. 4| UMAP embeddings of neural dynamics during the social interaction episode. Three-dimensional UMAP projections of neural activity restricted to the social interaction episode. Time progression within each trial is encoded by a color gradient, from blue (start) to red (end).



Supplementary Fig. 5| Latent dynamics of rSLDS models across behavioral phases.

a, Latent embeddings colored by trial time (blue to red) for social interaction episodes, post-interaction behavior, and the full behavioral session.

b, Latent embeddings colored by spatial proximity category. Blue represents close-distance segments, and green represents far-distance segments.

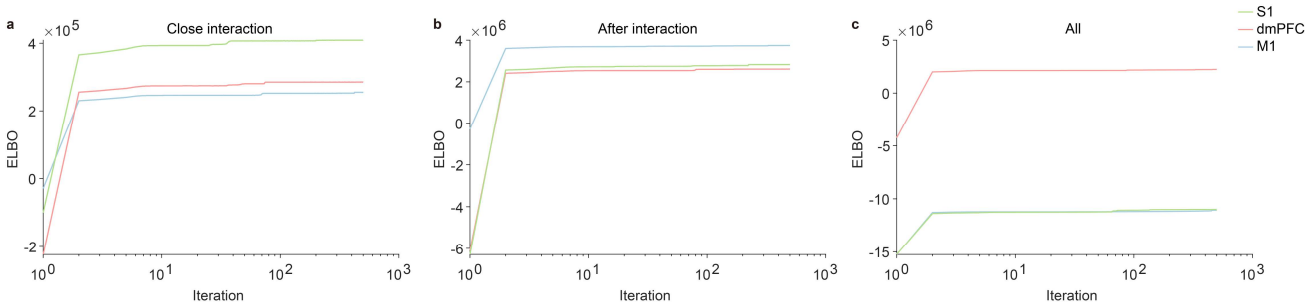
c, Silhouette coefficients quantifying cluster separability based on distance labels in **b**. (One-way ANOVA followed by Tukey multiple comparison test, mean±SEM, N=100, P values are <0.0001,

<0.0001, <0.0001, <0.0001, <0.0001, <0.0001, <0.0001, <0.0001, <0.0001, <0.0001, <0.0001, 0.0059, <0.0001, <0.0001, and <0.0001).

d, Ranked one-dimensional Poincaré sections (at x = 0 and y = 0) of latent trajectories from **a**, highlighting structural regularity.

e, Power spectral density (PSD) of the Poincaré maps in **d**, with bandwidths marked in red.

f, Entropy of the ranked 1D Poincaré sections across behavioral phases. (One-way ANOVA followed by Dunnett's T3 multiple comparisons test, mean±SEM, N=6, P values are 0.1798, 0.0041, and 0.0047). (G) Bandwidth of the PSD curves from (E) across behavioral phases. (One-way ANOVA followed by Dunnett's T3 multiple comparisons test, mean±SEM, N=6, P values are 0.1997, 0.0393, and 0.0212). *P < 0.05, **P < 0.001, ****P < 0.0001.

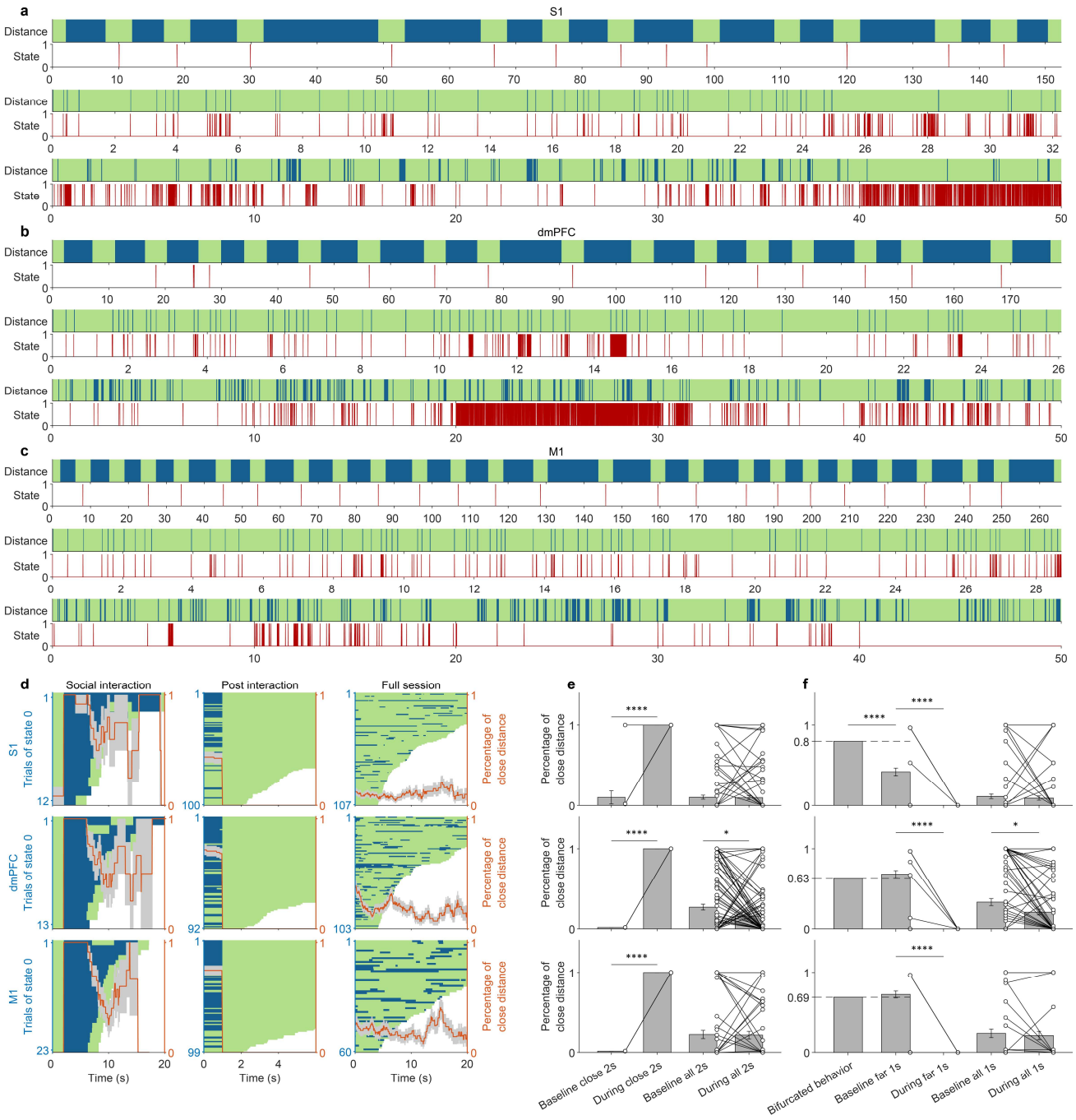


Supplementary Fig. 6| Evidence Lower Bound (ELBO) of rSLDS models fitted to different behavioral phases.

a, ELBO values of rSLDS models trained on neural activity from social interaction episodes in S1, dmPFC, and M1.

b, ELBO values of rSLDS models trained on neural activity during post-interaction behavior.

c, ELBO values of rSLDS models trained on neural activity from the full behavioral session.



Supplementary Fig. 7| Predicted states of rSLDS models fitted to different behavioral phases.

a, State predictions of the rSLDS model trained on S1 data. From top to bottom: results from models trained on social interaction episodes, post-interaction behavior, and the full behavioral session.

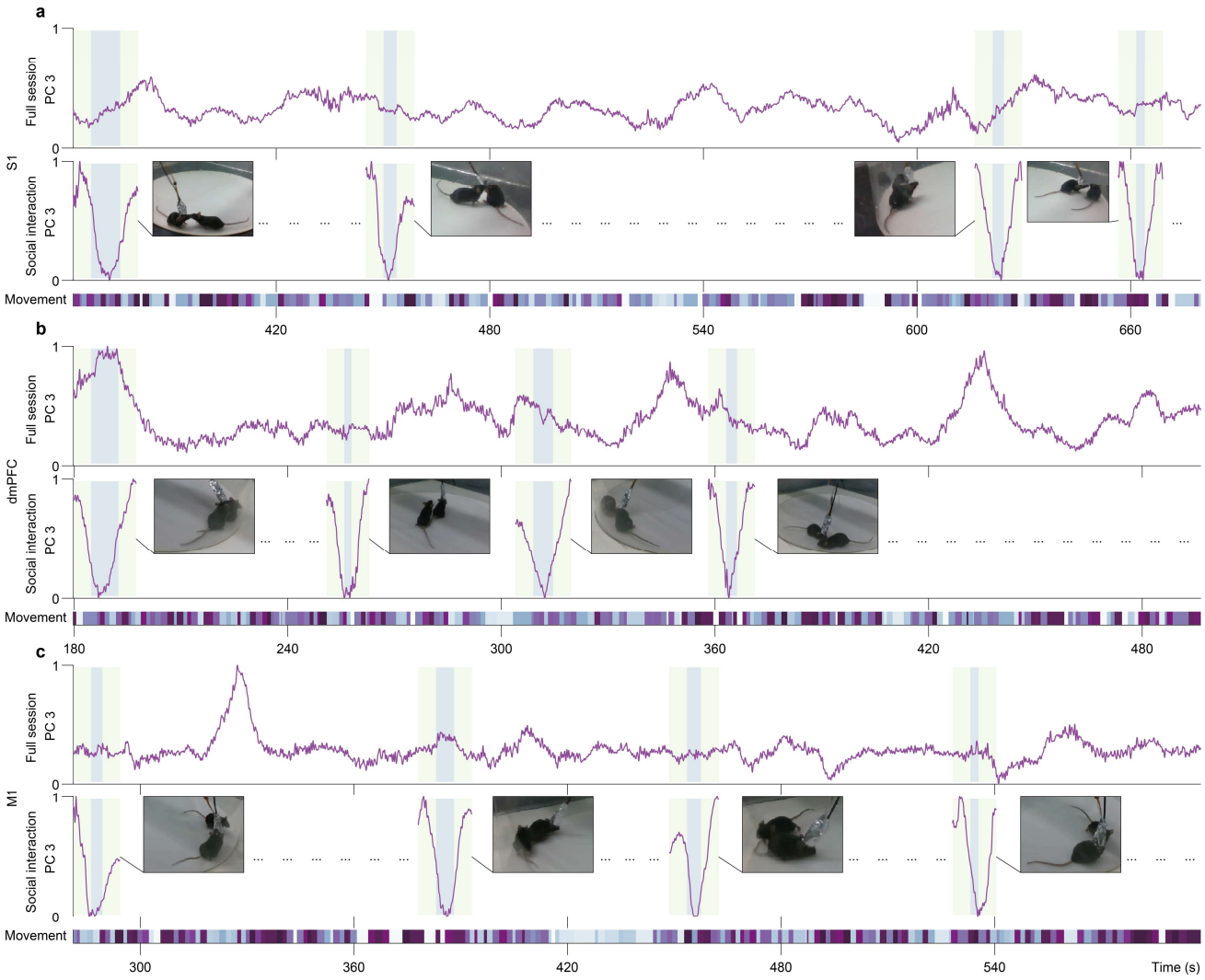
b, State predictions from rSLDS trained on dmPFC data.

c, State predictions from rSLDS trained on M1 data.

d, Temporal alignment of state 0 predictions across brain regions. The orange line represents the percentage of state 0 corresponding to close distance, as plotted on the right y-axis.

e, Consistency comparison of social interaction episodes using close distance percentages. From top to bottom are S1 (Rank sum test, mean±SEM, N=12 and 107, P values are <0.0001 and 0.7286), dmPFC (Rank sum test, mean±SEM, N=13 and 103, P values are <0.0001 and 0.01579), and M1 (Rank sum test, mean±SEM, N=23 and 60, P values are <0.0001 and 0.9010).

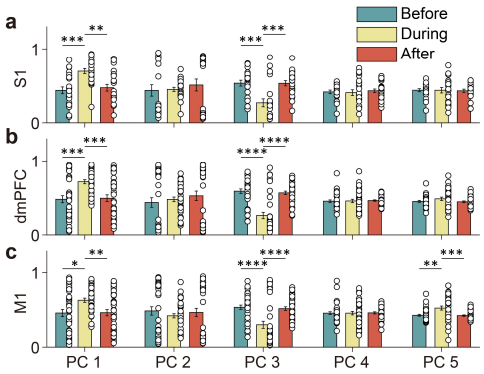
f, Consistency comparison of post-interaction behavior using close distance percentages. From top to bottom are S1 (Student t-test, rank sum test, and rank sum test, mean±SEM, N=100 and 107, P values are <0.0001, <0.0001 and 0.4451), dmPFC (Student t-test, rank sum test, and rank sum test, mean±SEM, N=92 and 103, P values are 0.3090, <0.0001 and 0.0307), and M1 (Student t-test, rank sum test, rank sum test, mean±SEM, N=99 and 60, P values are 0.4106, <0.0001 and 0.5393). The percentages of bifurcated behavior of individual behavior are from Fig. 3J. *P < 0.05, ****P < 0.0001.



Supplementary Fig. 8| Reproducible PC3 dynamics across repeated social interaction episodes in S1, dmPFC and M1.

a-c, Representative examples from S1 (**a**), dmPFC (**b**) and M1 (**c**), showing the temporal dynamics of the third principal component (PC3) across the full behavioral session (top) and during social interaction episodes (middle). Shaded regions in the full-session traces indicate the occurrence of social interaction episodes, with the corresponding PC3 trajectories expanded below. Representative video frames illustrate the behavioral context of selected episodes; ellipses indicate intervening episodes or time intervals omitted for visualization. Bottom, movement identities throughout the session, color-coded according to the behavioral classification. Despite substantial variation in neural

dynamics across the full session, repeated social interaction episodes were associated with stereotyped PC3 trajectories within each cortical region.

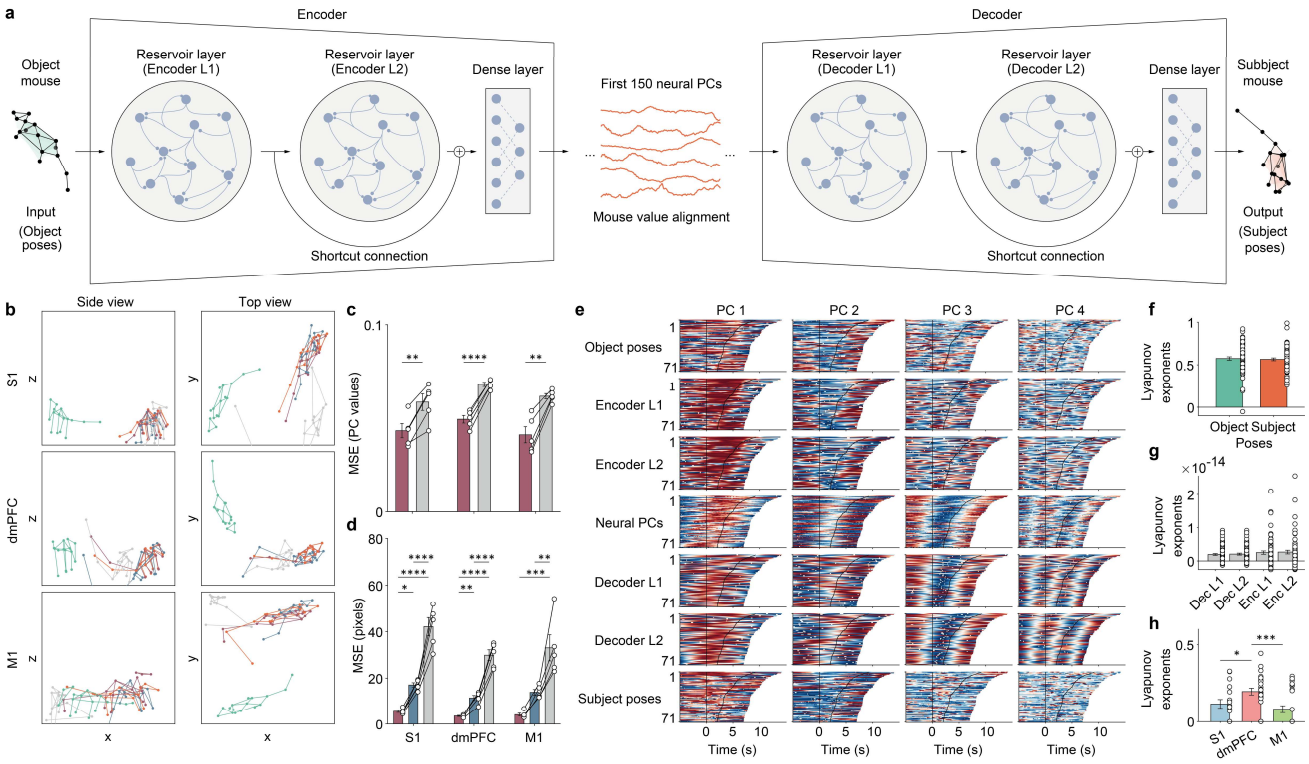


Supplementary Fig. 9| Comparison of neural principal components (PCs) during social interaction episodes across brain regions.

a, Distribution of neural PC trajectories from S1. (Wilcoxon rank sum test, mean±SEM, N=98, P values are 0.0003, 0.0016, 0.4047, 0.5652, 0.0009, 0.0004, 0.6812, 0.4740, 0.9532, and 0.8053).

b, Distribution of neural PC trajectories from dmPFC. (Wilcoxon rank sum test, mean±SEM, N=98, P values are 0.0006, 0.0005, 0.4313, 0.4593, <0.0001, <0.0001, 0.9438, 0.5183, 0.2747, and 0.2402).

c, Distribution of neural PC trajectories from M1. (Wilcoxon rank sum test, mean±SEM, N=98, P values are 0.0117, 0.0064, 0.6764, 0.9926, <0.0001, <0.0001, 0.9556, 0.8095, 0.0022, and 0.0009). *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.



Supplementary Fig. 10| A brain-inspired AI model recapitulates neural-behavioral alignment.

a, Architecture of the Neural PC-aligned Encoder-Decoder Reservoir Network (NPC-EDREN), which includes an encoder predicting neural PCs from object poses, and a decoder reconstructing subject poses from the neural PCs. The encoder and decoder each contain two-layer reservoirs with residual connections.

b, Representative 3D skeletons showing: the actual pose of the object mouse (green), the actual pose of the subject mouse (orange), the predicted subject pose decoded from recorded neural PCs (red), the predicted subject pose decoded from neural PCs predicted by the encoder (blue), and the predicted subject pose decoded from shuffled object pose inputs (gray).

c, Mean squared errors (MSEs) of neural PCs predicted by the encoder from object poses versus shuffled object poses (t-test, mean±SEM, N=5, P values are 0.0042, <0.0001, and 0.0025).

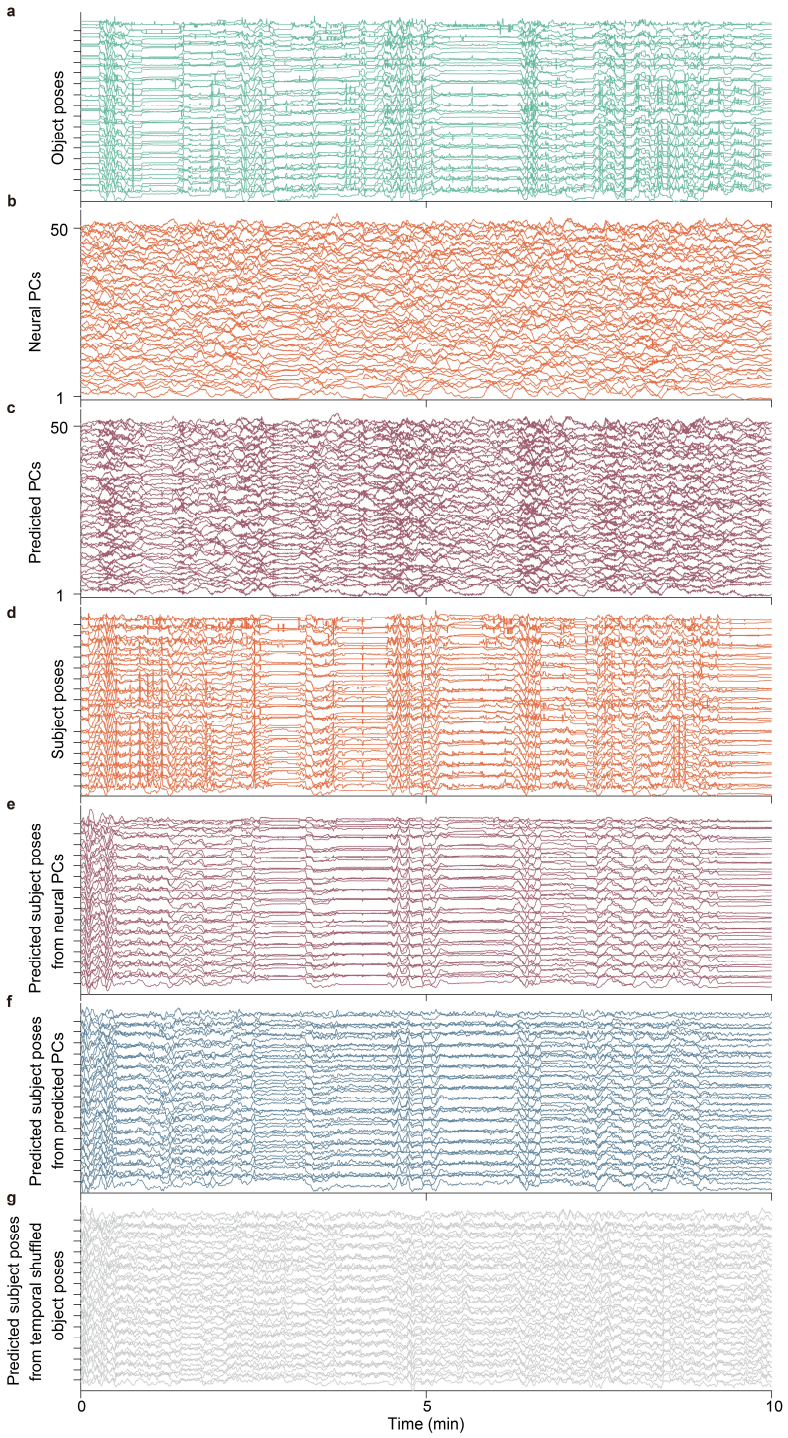
d, MSEs of subject poses predicted by the decoder using different inputs: recorded neural PCs (red), predicted PCs from object poses (blue), and shuffled object poses (gray) (one-way ANOVA, mean±SEM, N=5, P values are 0.0136, <0.0001, <0.0001, 0.0079, <0.0001, <0.0001, 0.1425, 0.00015, and 0.0043).

e, The first four PCs of poses, reservoir neural activity, and brain-derived neural PCs of social interaction episodes.

f, Lyapunov exponents during social interaction episodes of the comparison between object and subject poses (rank sum test, mean±SEM, N=71, P value is 0.5192).

g, Lyapunov exponents comparison across reservoir layers (Dec L1, Decoder L1; Dec L2, Decoder L2; Enc L1, Encoder L1; Enc L2, Encoder L2; one-way ANOVA followed by Dunn's multiple comparisons test, mean±SEM, N=71, adjusted P values are 0.7049, 0.5085, 0.3648, >0.9999, >0.9999, and >0.9999).

h, Lyapunov exponents from the neural PCs of S1, dmPFC, and M1 (One-way ANOVA followed by a two-stage linear step-up procedure of Benjamini, Krieger, and Yekutieli-corrected multiple comparisons test, mean±SEM, N=17, 26, and 28, the individual P values are 0.0445, 0.3907, and 0.0011).



Supplementary Fig. 11| Input, prediction, and output examples of the NPC-EDREN model.

a, Input object poses.

b, Input neural principal components (PCs).

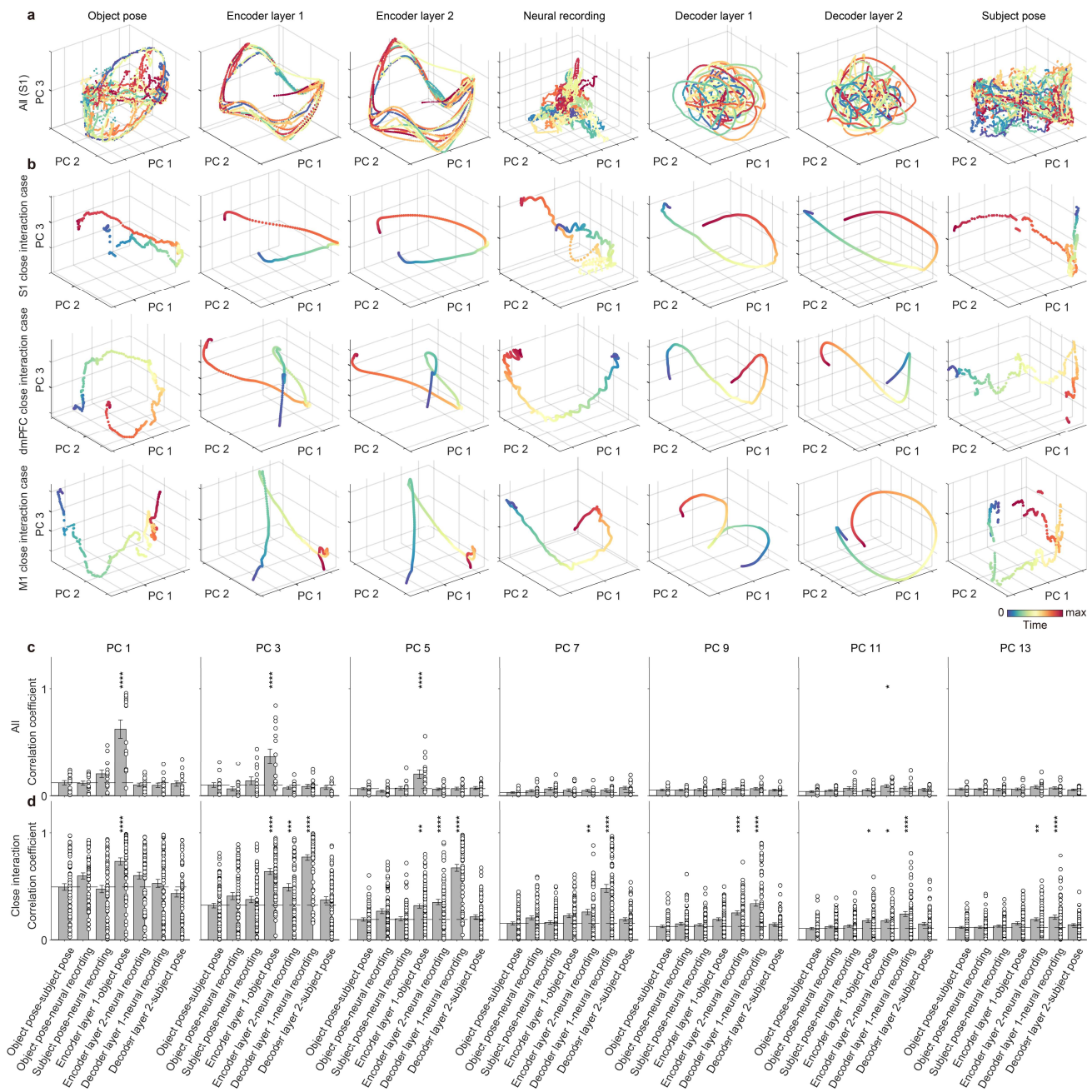
c, Predicted neural PCs generated by the encoder.

d, Ground-truth subject poses.

e, Predicted subject poses decoded from real neural PCs.

f, Predicted subject poses decoded from the encoder-predicted PCs.

g, Predicted subject poses decoded from temporally shuffled object poses (control condition).



Supplementary Fig. 12| Dynamics and cross-region correlations in the NPC-EDREN model.

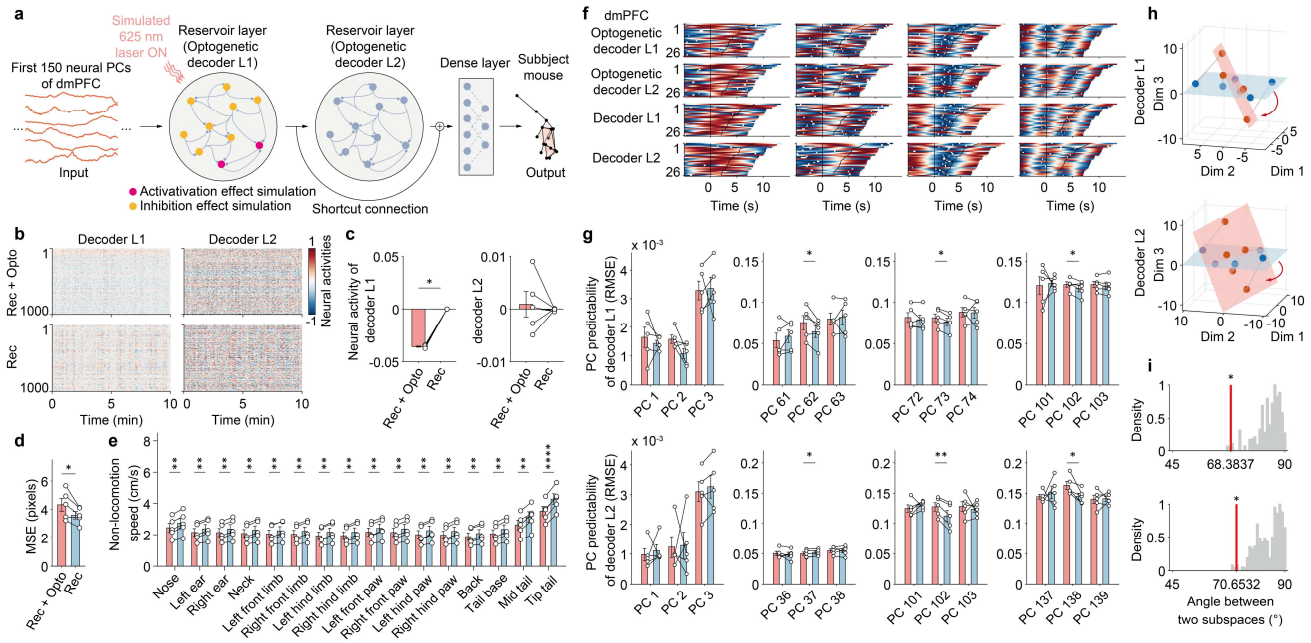
a, PC dynamics across the full behavioral session from one S1 mouse recording.

b, PC dynamics during social interaction episodes across S1, dmPFC, and M1.

c, Correlation coefficients between latent elements of the NPC-EDREN model computed over the full behavioral session. (One-way ANOVA followed by Tukey multiple comparison test, mean±SEM, N=15, P values of PC1 are >0.9999, 0.7295, <0.0001, 0.9998, 0.9995, and >0.9999, P values of PC3

are 0.9893, 0.9823, <0.0001, 0.9979, >0.9999, and 0.9980, P values of PC5 are 0.9816, >0.9999, <0.0001, >0.9999, >0.9999, and >0.9999, P values of PC7 are 0.9670, 0.4443, 0.8913, 0.9262, 0.8491, and 0.1434, P values of PC9 are >0.9999, 0.9999, 0.9972, 0.9831, 0.9642, and >0.9999, P values of PC11 are 0.9968, 0.5380, 0.9640, 0.0430, 0.4752, and 0.9103, and P values of PC13 are >0.9999, 0.9999, >0.9999, 0.8614, 0.9945, and 0.9991.).

d, Correlation coefficients computed during social interaction episodes. (One-way ANOVA followed by Tukey multiple comparison test, mean±SEM, N=71, P values of PC1 are 0.2705, 0.9996, <0.0001, 0.2335, 0.9901, and 0.8433, P values of PC3 are 0.2946, 0.7802, <0.0001, 0.0007, <0.0001, and 0.8218, P values of PC5 are 0.2333, >0.9999, 0.0020, <0.0001, <0.0001, and 0.9894, P values of PC7 are 0.6135, 0.9999, 0.2108, 0.0078, <0.0001, and 0.8861, P values of PC9 are 0.9620, 0.9975, 0.1185, <0.0001, <0.0001, and 0.9846, P values of PC11 are 0.9865, 0.9273, 0.0103, 0.0116, <0.0001, and 0.4163, P values of PC13 are >0.9999, 0.9966, 0.4128, 0.0021, 0.0000, 0.8780.). *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.



Supplementary Fig. 13| NPC-EDREN reproduces dimension-specific effects of cortical perturbation on behavior.

a, Schematic of neural perturbation simulation in NPC-EDREN. Simulated optogenetic perturbation was introduced into the first decoder reservoir layer (Decoder L1). The input and output to the model consist of the recorded neural PCs of dmPFC and the subject mouse poses from Fig. 2 (without perturbation).

b, Neural dynamics of Decoder L1 and Decoder L2 under simulated optogenetic perturbation (Rec + Opto) and control (Rec) conditions.

c, Neural activity of decoder L1 and L2 for the optogenetic activation (Rec + Opto) and the control (Rec) conditions (Sign-rank test, mean±SEM, N=5, P values are 0.0312 and >0.9999).

d, MSEs of the subject-pose decoding under simulated optogenetic perturbation versus no-perturbation control (paired t-test, mean±SEM, N=5, P value is 0.0379).

e, Non-locomotion speed across body parts during the extracted social interaction episodes for the simulated optogenetic activation and control conditions (Paired t-test, mean±SEM, N=5, P values from

left to right are 0.0030, 0.0032, 0.0049, 0.0046, 0.0042, 0.0068, 0.0048, 0.0053, 0.0032, 0.0046, 0.0020, 0.0083, 0.0065, 0.0029, 0.0014, and <0.0001).

f, The first four PCs of the optogenetic decoder L1, optogenetic decoder L2, decoder L1, and decoder L2, trained by the dmPFC neural PCs during social interaction episodes.

g, The RMSE of PC predictability of decoder L1 (Top) and L2 (Bottom) under simulated optogenetics. Only PC dimensions exhibiting significant differences between the optogenetic activation (Rec + Opto) and control (Rec) conditions are displayed (Paired t-test, mean±SEM, N=5, P values of Top from left to right are 0.4375, 0.0870, 0.8645, 0.4008, 0.0453, 0.7988, 0.4064, 0.0364, 0.9008, 0.8285, 0.0457, and 0.2868, and P values of Bottom from left to right are 0.6250, >0.9999, 0.4063, 0.5808, 0.0339, 0.9893, 0.3901, 0.0032, 0.6484, 0.6225, 0.0322, and 0.8650).

h, Reduced-rank regression for modeling the coupling between PC predictability and non-locomotion speed of decoder L1 (Top) and L2 (Bottom). The rank was set to 2. The blue plane represents the coupling subspace without optogenetic stimulation, and the red plane represents the coupling subspace with optogenetic stimulation. Optogenetic activation changes the coupling directions between PC predictability and non-locomotion speed.

i, Quantification of the angle between the low-rank coupling subspaces under the no-optogenetic and optogenetic conditions for decoder L1 (Top) and L2 (Bottom). Gray histograms show the null distributions of subspace angles obtained by permutation (permutation test, N=1000, P values are 0.0120 and 0.0220). *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

Behavior	**Definition**
Creep	Slow forward movement with low body posture and minimal limb extension.
Creep_sniff	Creep while continuously sniffing the ground, the air, or the partner.
Creep_turn	Creep combined with a gradual body axis rotation or orientation change.
Dash	Rapid, short-distance locomotion with burst-like acceleration.
Dive	Sudden downward motion of the head or forelimbs.
Groom	Repetitive self-directed cleaning behavior using forepaws or mouth.
Hunch	A crouched posture with a curved back and compressed limbs.
Immobility	Absence of detectable movement except for respiration.
Micromovement	Small, localized motions without displacement.
Nod	Brief vertical head movement.
Poke	Forward thrusting of the snout or head toward any target or direction.
Poke_retreat	Poke behavior immediately followed by backward movement or withdrawal.
Rear	Vertical elevation of forelimbs off the ground while maintaining balance.
Rear_dive	Rear followed by a sudden dive or head-down plunge.
Rear_sniff	Rearing posture combined with sniffing behavior in elevated position.
Rear_walk	Forward walking behavior accompanied by a progressive elevation of the forelimbs.
Right_turn	Body axis rotation to the right.
Rise	Upward elevation of the head and/or upper body without fully transitioning into a rearing posture.

Rise_sniff	Rise motion accompanied by exploratory sniffing.
Run	High-speed locomotion.
Sniff	Rhythmic nasal movements.
Sniff_rise	Sniffing behavior combined with partial vertical postural shift.
Sniff_turn	Head or body turning while maintaining sniffing.
Sniff_walk	Low-speed forward locomotion then sniffing.
Step	A forward stride involving one or a few limb movements, without sustained locomotion.
Step_rear	Stepping motion leading into a rear posture.
Step_rise	Stepping motion initiating a rise in body elevation.
Step_sniff	Small step taken while engaging in sniffing.
Step_turn	Localized stepping movement accompanied by body rotation.
Stretch_sniff	Forward elongation of the body with extended neck during sniffing.
Trot	Rhythmic, medium-speed forward locomotion with diagonal limb movements resembling quadrupedal trotting in horses.
Turn	Body axis rotation without translational displacement.
Turn_creep	Creep behavior accompanied by slow angular body turning.
Turn_rise	Body rotation occurring during a rise motion.
Turn_sniff	Turning while maintaining sniffing.
Turn_step	Stepwise rotation involving pivoting motions.
Turn_walk	Turning motion occurring during walking.
Walk	Forward locomotion at low to medium speed with regular alternating limb movements.
Walk_rise	Walking motion with body elevation.

Walk_sniff	Walking while continuously engaging in sniffing.
Walk_turn	Forward locomotion combined with gradual turning.
Walk_turn_rise	Walk and turn followed by body elevation.
Walk_turn_sniff	Walk and turn accompanied by sniffing behavior.

Supplementary Tab. 1| Movement definitions for manually labelled behavior

Supplementary Notes

Single-neuron representations diversify across behavioral timescales

To examine how cortical representations vary across different behavioral timescales, we quantified single-neuron encoding of behavioral events spanning short-, intermediate-, and long-timescale structures, including instantaneous poses, movement motifs, and behavioral sequences. We compared encoding properties across S1, dmPFC, and M1 to determine whether cortical representations become progressively specialized as behavioral timescales expand.

At short timescales, pose-related neural representations were broadly shared across cortical regions. Neuron-wise pose correlations showed similar distributions across S1, dmPFC, and M1 (Fig. S1a), with comparable fractions of positively, negatively, and weakly modulated neurons (Fig. S1b). Pose-related feature vectors from the three cortical areas occupied overlapping regions in a shared principal component space (Fig. S1c), and pairwise distances between cortical representations were not significantly different (Fig. S1d). These results indicate that instantaneous body configurations are represented through a distributed cortical code with limited regional differentiation.

At intermediate timescales, cortical representations began to diverge. Movement motif-related activity revealed structured population responses aligned to behavioral events across all three regions (Fig. S1e). Although movement-related responses were observed throughout the cortex, the distribution of responsive neurons differed between areas (Fig. S1f). Feature representations associated with movement motifs showed increased separation in principal component space (Fig. S1g), with greater divergence between cortical areas compared with pose encoding, including a significant difference between S1 and dmPFC representations (Fig. S1h). These results suggest that cortical specialization emerges as neural activity integrates information over longer behavioral intervals.

At long timescales, sequence-related representations exhibited the strongest regional differentiation. Extended event-aligned neural activity revealed distinct temporal response patterns associated with behavioral sequences (Fig. S1i), accompanied by more pronounced differences in the proportions of sequence-responsive neurons across cortical areas (Fig. S1j). Sequence-related feature vectors showed clear separation among S1, dmPFC, and M1 in the shared feature space (Fig. S1k), with significantly increased inter-regional distances compared with shorter-timescale representations (Fig. S1l).

Together, these analyses demonstrate a timescale-dependent organization of cortical behavioral representations. Short-timescale behavioral variables are encoded by broadly shared distributed cortical activity, whereas longer-timescale behavioral structures recruit increasingly differentiated representations across cortical regions. This progressive divergence suggests that cortical specialization emerges from the temporal scale over which neural populations organize behavior.

Neural population activity captures behavioral representations across multiple timescales

To examine whether cortical population activity contains information about behavioral structures across different temporal scales, we constructed neural-behavioral joint embeddings using CEBRA and evaluated the correspondence between neural dynamics and behavioral variables at short, intermediate, and long timescales[1]. We analyzed embeddings derived from the subject mouse's neural activity and evaluated their alignment with both the subject mouse's own behavior and the behavior of the interacting object mouse.

At short timescales, CEBRA embeddings based on neural activity captured continuous pose dynamics. Pose-related embeddings exhibited smooth low-dimensional trajectories, with similar

structures observed across S1, dmPFC, and M1 (Fig. S2a). When neural activity was aligned with the object mouse's poses rather than the recorded subject mouse's own poses, the resulting embeddings retained structured organization, although with reduced correspondence (Fig. S2b). Quantitative reconstruction analysis demonstrated reliable recovery of pose information from cortical activity across regions, with comparable reconstruction errors between subject and object conditions (Fig. S2c).

At intermediate timescales, neural embeddings captured discrete movement motifs. Compared with pose representations, movement embeddings showed more clustered organization, reflecting the segmentation of continuous behavior into recurring movement patterns (Fig. S2a,b). Decoding accuracy for movement motifs was consistent across cortical regions, indicating that population activity contains robust information about intermediate behavioral structures (Fig. S2c).

At longer timescales, neural activity provided increasingly strong representations of behavioral sequences. Sequence-related embeddings exhibited more distinct clustering patterns compared with pose and movement embeddings, suggesting that cortical dynamics capture extended behavioral organization over longer temporal windows (Fig. S2a,b). Sequence decoding achieved high accuracy across cortical regions, exceeding movement decoding performance (Fig. S2c,d). This enhancement was observed for both subject and object mouse behaviors, indicating that cortical population dynamics encode not only self-generated actions but also behaviorally relevant information from interacting conspecifics.

Together, these analyses demonstrate that cortical population activity contains behaviorally meaningful information across multiple temporal scales. While short-timescale variables are represented through continuous neural-behavioral trajectories, longer-timescale behavioral structures emerge as more separable neural representations. These results support that cortical population

dynamics provide a continuous substrate from which behavioral structures at different temporal scales can be recovered, linking transient sensorimotor states with extended behavioral sequences.

Global continuity and local conservation of cortical dynamics

To further examine how distributed cortical dynamics are organized across timescales, we visualized full-session and episode-restricted neural activity using three-dimensional UMAP embeddings[2]. This analysis revealed a clear distinction between global session dynamics and local episode dynamics.

Across full behavioral sessions, neural trajectories in S1, dmPFC, and M1 evolved continuously over time, forming extended paths in latent space without obvious re-entry into highly conserved states (Fig. S3). Although natural behavior contains many repeated actions and motifs, the corresponding full-session neural embeddings did not collapse into stable, repeatedly occupied global states. Instead, the trajectories progressed smoothly over time, indicating that at the whole-session level, cortical population activity is dominated by temporal continuity rather than global state conservation. In other words, neural dynamics remain continuous across the session, but their precise embedding geometry is not strongly preserved across repeated behavioral occurrences.

In contrast, when the analysis was restricted to the social interaction episode, neural dynamics exhibited a markedly different structure (Fig. S4). Within this specific social behavior context, trajectories from different trials and animals showed similar geometric forms, despite being located at different positions in latent space. Thus, the absolute position of the trajectory was variable, but its overall shape was conserved. This indicates that distributed cortical dynamics are reproducible in an episode-specific manner, revealing conserved trajectory motifs that are not apparent at the level of the

entire session.

Together, these results suggest that distributed cortical dynamics exhibit two complementary organizational principles across timescales. At the global scale, neural activity forms a temporally continuous flow through latent space, with continuity exceeding conservation. At the local scale, behaviorally specific episodes recruit conserved dynamical trajectories, whose geometry is preserved even when their global embedding positions vary. These findings support the idea that reproducible cortical dynamics emerge not as fixed global states across an entire natural session, but as locally conserved trajectory structures tied to specific behavioral episodes.

Behavioral episodes constrain cortical activity into reproducible latent dynamical regimes

To further characterize the dynamical organization underlying episode-specific neural trajectories, we fitted recurrent switching linear dynamical system (rSLDS) models[3-6] separately to neural PCs during social interaction, post-interaction behavior, and the full behavioral session. Across S1, dmPFC, and M1, the inferred latent dynamics differed between behavioral phases (Fig. S5a). During social interaction, trajectories occupied structured regions of latent space and evolved along reproducible paths, whereas post-interaction and full-session dynamics were progressively less constrained. Labeling the same trajectories according to spatial proximity revealed that close- and far-distance periods were strongly segregated during social interaction but became more intermixed outside this episode (Fig. S5b). Consistently, silhouette analysis showed stronger separation of proximity-defined states during social interaction than during post-interaction behavior or the full session across all three cortical regions (Fig. S5c).

We next quantified the internal organization of these latent trajectories using Poincaré sections.

Ranked one-dimensional sections revealed strongly phase-dependent structures. Social-interaction trajectories exhibited relatively ordered distributions, post-interaction dynamics showed intermediate organization, whereas full-session trajectories produced broader and less constrained distributions (Fig. S5d). This difference was also evident in their spectral properties (Fig. S5e). Full-session dynamics showed increased entropy of the ranked Poincaré sections and reduced spectral bandwidth relative to social-interaction dynamics, with post-interaction activity again occupying an intermediate regime (Fig. S5f, g). Thus, the transition from a specific behavioral episode to unconstrained natural behavior was accompanied by a systematic reorganization of the underlying latent dynamics rather than simply a change in their position in state space. ELBO analysis across behavioral phases and cortical areas further provided a model-fit control for these rSLDS representations (Fig. S6a-c).

The discrete states inferred by the rSLDS models further linked these dynamical regimes to ongoing behavior (Fig. S7a-c). During social interaction, one recurrent latent state was consistently associated with close-distance periods across S1, dmPFC, and M1, and its temporal occurrence aligned across cortical regions (Fig. S7d). This closely associated state was highly reproducible across social-interaction episodes (Fig. S7e), indicating that the conserved trajectory structure identified at the episode level was accompanied by a reproducible sequence of latent dynamical states. After interaction, however, this correspondence was reorganized and became more region-dependent, consistent with the heterogeneous behavioral trajectories observed during the post-interaction period (Fig. S7f).

Together, these analyses show that cortical population activity is not governed by a single stationary dynamical landscape throughout natural behavior. Instead, specific behavioral episodes transiently constrain distributed cortical activity into reproducible latent regimes in which neural trajectories, dynamical states, and behavior become tightly organized. Outside these episodes, cortical

activity explores a substantially larger and less conserved portion of state space. These results provide a dynamical account of the local reproducibility observed in the preceding analyses, suggesting that conserved cortical organization emerges locally within behaviorally defined episodes rather than globally across the entire session.

A conserved neural dimension captures recurrent dynamics during social interaction

To determine whether the reproducible dynamics associated with social interaction were concentrated within specific dimensions of population activity, we examined principal-component trajectories across repeated interaction episodes. Despite substantial fluctuations in neural activity over the full behavioral session, the third principal component (PC3), for example, exhibited a strikingly reproducible temporal pattern whenever social interaction occurred. In S1, dmPFC, and M1, PC3 consistently decreased as animals entered the interaction episode and recovered as the episode ended, producing stereotyped trajectories across repeated occurrences within the same session (Fig. S8a-c). Importantly, similar movement identities also occurred outside these episodes without producing the same PC3 trajectory, indicating that this low-dimensional pattern reflected the broader interaction episode rather than the recurrence of individual movements alone.

We next compared the first five principal components before, during, and after social interaction. PC3 showed the most consistent episode-dependent modulation across all three cortical regions, with activity significantly reduced during interaction relative to both flanking periods in S1, dmPFC, and M1 (Fig. S9a-c). PC1 was also consistently modulated across regions, whereas changes in the remaining components were weaker or region-dependent; for example, PC5 showed significant modulation in M1 but not in S1 or dmPFC. Thus, although social interaction recruits distributed

population activity, its reproducible temporal structure is concentrated within a restricted subset of neural dimensions rather than being expressed uniformly across the population space.

Together, these results show that repeated social interaction episodes repeatedly engage conserved low-dimensional cortical dynamics across S1, dmPFC, and M1. The recurrence of a stereotyped PC3 trajectory despite substantial variability in full-session activity further supports the idea that specific behavioral episodes locally constrain ongoing cortical dynamics into reproducible trajectories. Such conservation at the level of individual neural dimensions provides a low-dimensional substrate through which distributed cortical populations can maintain episode-specific temporal organization amid continuously varying natural behavior.

A brain-inspired recurrent model recapitulates neural dynamics and their perturbation-dependent behavioral effects

To test whether the observed neural-behavioral relationships could be reproduced by a recurrent dynamical architecture, we developed a Neural PC-aligned Encoder–Decoder Reservoir Network (NPC-EDREN). The model consisted of an encoder that transformed object-mouse poses into neural principal components, followed by a decoder that reconstructed subject-mouse poses from these neural representations (Fig. S10a). Both encoder and decoder contained two recurrent reservoir layers[7] with shortcut connections, thereby introducing an explicit neural latent representation between social input and behavioral output.

The encoder reliably predicted neural PCs from object-mouse poses, producing significantly lower prediction errors than temporally shuffled controls across S1, dmPFC, and M1 (Fig. S10b, c). These predicted neural PCs could subsequently be decoded into subject-mouse poses, with reconstruction

errors significantly lower than those obtained from shuffled object-pose inputs and approaching those obtained when recorded neural PCs were directly supplied to the decoder (Fig. S10d). Examination of the internal representations further showed structured temporal patterns across object poses, encoder layers, recorded neural PCs, decoder layers, and subject poses during social interaction episodes (Fig. S10e). Lyapunov analysis revealed comparable dynamical stability between object and subject behavioral representations and across intermediate reservoir layers, while brain-derived neural PCs showed region-dependent dynamical properties across S1, dmPFC, and M1 (Fig. S10f-h). These results indicate that the model can transform social behavioral input into subject behavior through a structured neural-like dynamical representation.

We next examined the temporal fidelity of this transformation across extended behavioral periods. Representative examples showed that object poses, recorded neural PCs, encoder-predicted neural PCs, and reconstructed subject poses preserved coherent temporal structure over the full session (Fig. S11a-f). In contrast, temporally shuffled object-pose inputs failed to reproduce the corresponding subject-pose dynamics (Fig. S11g). Thus, the model captured not only static correspondences between behavioral variables, but also the temporal organization linking social input, neural population activity, and behavioral output.

We then asked whether the model reproduced the episode-specific neural organization observed in the cortical recordings. Across the full behavioral session, neural PCs exhibited substantial temporal variability (Fig. S12a). When activity was restricted to social interaction episodes, both recorded neural PCs and intermediate model representations displayed reproducible temporal trajectories across S1, dmPFC, and M1 (Fig. S12b). Correlation analysis showed that relationships among object poses, recorded neural PCs, encoder states, decoder states, and subject poses were relatively limited and

dimension-specific over the full session (Fig. S12c), but became stronger and more structured during social interaction episodes (Fig. S12d). These results indicate that neural-behavioral coupling is dynamically reorganized during specific behavioral episodes rather than remaining fixed across the entire session.

Finally, we tested whether perturbing the model's internal dynamics could reproduce the behavioral consequences of cortical perturbation. Simulated optogenetic activation was introduced into the first decoder reservoir layer while recorded dmPFC neural PCs were used as model input (Fig. S13a). The perturbation altered neural activity in Decoder L1, whereas activity in Decoder L2 was comparatively preserved (Fig. S13b, c). Simulated activation significantly increased subject-pose reconstruction error and altered non-locomotor movement speeds across multiple body parts during social interaction (Fig. S13d, e). At the latent level, perturbation effects were dimension-selective rather than uniform across the decoder state space (Fig. S13f, g). Reduced-rank regression[8] further showed that simulated activation changed the low-dimensional coupling between PC predictability and behavioral output (Fig. S13h), with the resulting rotations between control and perturbed coupling subspaces exceeding those expected under permutation (Fig. S13i).

Together, these results show that NPC-EDREN reproduces several key properties of the experimentally observed neural-behavioral system: transformation of social input through intermediate neural dynamics, preservation of temporally structured behavioral relationships, stronger neural-behavioral coupling during specific social episodes, and dimension-selective behavioral consequences following perturbation. The model therefore provides a computational framework linking recurrent neural dynamics to the organization and causal modulation of natural behavior timescales.

Supplementary Methods

The correlation between poses and neural activities

The Pearson correlation coefficient is used to quantify the correlation between the poses of each coordinate and the neural activities of each neuron. The coefficients with $P<0.05$ are selected to calculate the distribution. The Gaussian Mixture Model (GMM)[9] is used to estimate peak centers of the distribution. The histogram shows two peaks, so the number of components of the GMM is set to 2.

The alignment of neural activities of movements and sequences

Each trial of movements and sequences has 3 parts, including baseline, behaving, and after behaving. Considering that the maximum time lengths of movements and sequences are 2 s and 10 s, their baseline, behaving, and after-behavior are set to 2 s and 10 s separately. Because behavior is continuous, the baseline of neural activities is not silence. Hence, each trial of the neural activities matching with behavior are transformed to $\Delta F / F(t) = \frac{F(t) - F_0(t)}{F_0(t)}$ format, where $F_0(t)$ is the mean value of the baseline. After the transformation, the differences of neural activity values between baseline and behaving are compared through a t-test. The trials are sorted according to their P values.

Neural activity consistency analysis

The consistencies of each neuron are quantified by their activity patterns. The neural activities of each neuron can be classified into positive response, low response, and negative response according to their behavioral aligned activities. In movements and sequences, the thresholds across positive, low,

and negative responses are determined by the P values and the differences of mean values. Every neuron has a percentage vector of three responses like $\left[a_p, a_l, a_n\right]$, where a_p is the percentage of positive response, a_l is the percentage of low response, and a_n is the percentage of negative response. Because the summation of $\left[a_p, a_l, a_n\right]$ equals to 1, one of the percentages of reponse is information redundancy. They can be represented into a 2-dimensional space. Hence, principal component analysis (PCA) can be used to reduce their redundancies. The percentages of each neuron are stacked to $[a_{p1}...a_{pm}, a_{l1}...a_{lm}, a_{n1}...a_{nm}]^T$, where m is the number of neurons. The dimensional reduction was done across the percentage dimension. The feature vectors represent the coordinates of the raw positive, low, and negative responses. The response feature distances were calculated by the Euclidean distance between PCA scores and feature vectors.

Neural embedding using CEBRA

CEBRA models were trained using a multi-session mode[1]. The combinations of three factors of behaviors from 3 hierarchies, neural activities from 3 brain areas, and the social pairs of 2 mice are trained using separate CEBRA models. The model architecture is 'offset10-model'. The batch size is 1024. The learning rate is 0.0003. The temperature is 1. The output dimension is 3. The maximum iteration is 15,000. The distance is 'cosine'. The conditional distribution is 'time_delta'. The time offset is 10. The random seed is 1234.

The parameter settings and validation of NPC-EDREN

The main body of NPC-EDREN is the encoder-decoder structure, but they are separately trained. NPC-EDREN is built using BrainPy[10]. The encoder has 3 layers. The first 2 layers are the reservoir

layer, and the output layer is the dense layer. The input dimension of the first reservoir layer is 48. The number of its neurons is 1000. Its leaky rate is 0.1. The connectivity of its input neurons is 0.1. The connectivity of its recurrent weights matrix is 0.1. Its output dimension is 1000. Other parameters of the first reservoir layer are set to default. The first layer accepts the input of the pose trajectory of the object mouse and outputs to the second layer and the dense layer. The input dimension of the second layer is 1000. The number of its neurons is 1000. Its leaky rate is 0.1. The connectivity of its input neurons is 0.1. The connectivity of its recurrent weights matrix is 0.1. Its output dimension is 1000. Other parameters of the second reservoir layer are set to default. The second layer accepts the output of the first layer and outputs to the dense layer. The dense layer is a linear transformation $y = wx + b$, where x is the input, w is the trainable weight matrix, b is the bias, and y is the output. Its input dimension is 1000, and its output dimension is 150. The dense layer accepts the input from the first and the second reservoir layers. The input weights of these two reservoir layers to the dense layer are 0.5. The encoder is trained to fit the first 150 neural PCs of the subject mouse from the input of the poses of the object mouse. The training algorithm is ridge regression, and its regularization coefficient is 10^{-7}.

The decoder has 3 layers. The first 2 layers are the reservoir layer, and the output layer is the dense layer. The input dimension of the first reservoir layer is 150. The number of its neurons is 1000. Its leaky rate is 0.1. The connectivity of its input neurons is 0.1. The connectivity of its recurrent weights matrix is 0.1. Its output dimension is 1000. Other parameters of the first reservoir layer are set to default. The first layer accepts the input of the 150 neural PCs of the subject mouse and outputs to the second layer and the dense layer. The input dimension of the second layer is 1000. The number of its neurons is 1000. Its leaky rate is 0.1. The connectivity of its input neurons is 0.1. The connectivity of

its recurrent weights matrix is 0.1. Its output dimension is 1000. Other parameters of the second reservoir layer are set to default. The second layer accepts the output of the first layer and outputs to the dense layer. The dense layer is a linear transformation $y = wx + b$, where x is the input, w is the trainable weight matrix, b is the bias, and y is the output. Its input dimension is 1000, and its output dimension is 48. The dense layer accepts the input from the first and the second reservoir layers. The input weights of these two reservoir layers to the dense layer are 0.5. The encoder is trained to fit the subject pose trajectories from the input of the 150 neural PCs of the subject mouse. The training algorithm is ridge regression, and its regularization coefficient is 10^{-7}.

The training data are organized according to each experimental trial. Each trial has three parts, including the object pose trajectories as input, the neural PCs as the alignment, and the subject pose trajectories as the output. The neural PCs are normalized using the min-max method through the PC dimensions to balance the range of error in training to get better optimization. Each trial of data is used to train an NPC-EDREN. There are 15 well-trained NPC-EDRENs in total. Each of the brain areas has 5 NPC-EDREN models as a group for further analysis.

Optogenetic perturbation in NPC-EDREN

To simulate optogenetic perturbation within the NPC-EDREN framework, we extended the above design with an explicit optogenetic bias pattern and a slow activity-dependent adaptation current. The reservoir consisted of N recurrent units with sparse feedforward and recurrent connectivity. Optogenetic perturbation was implemented by assigning a fixed perturbation pattern $p \in \mathbb{R}^N$ to the reservoir units, with the first $N_{ex} = 100$ units assigned +1 (photo-activation), the next $N_{ih} = 900$ units

assigned -1 (photo-inhibition), and the remaining units assigned 0. The optogenetic bias vector was fixed throughout the simulation and scaled by a global light strength parameter $\alpha = 0.07$:

$$u_{opto} = \alpha p .$$

At each time step, the reservoir state was updated according to:

$$h_t = \tanh(x_t W_{in} + s_{t-1} W_{rec} + \alpha p - \beta a_{t-1}) ,$$

$$s_t = (1-\lambda) s_{t-1} + \lambda h_t ,$$

where s_t denotes the reservoir state, λ is the leaky integration rate, and $\beta = 0.5$ controls the strength of an adaptation current. To model slow activity-dependent adaptation, we introduced a per-neuron adaptation variable a_t updated as:

$$a_t = (1-\eta) a_{t-1} + \eta s_{t-1} ,$$

where $\eta = 0.001$ sets the adaptation time scale. This adaptation term provides a slow negative feedback that accumulates recent activity and counteracts sustained excitation.

The optogenetic perturbation was applied inside the neuronal activation function, effectively modulating the membrane potential before the nonlinearity, rather than directly altering the output. This design mimics the biophysical effect of optogenetic stimulation on neuronal excitability. We also tested alternative implementations, including applying the perturbation after the nonlinearity or omitting the adaptation current, but none of these variants reproduced the selective dynamical effects observed in biological experiments. Only the combination of pre-activation perturbation and slow negative feedback recapitulated the optogenetic-induced destabilization of high-dimensional dynamics. The light strength parameter α was systematically scanned from 0.05 to 0.30 in steps of 0.01. We found that $\alpha = 0.07$ most closely matched the magnitude of dynamical perturbation observed under the experimental optogenetic stimulation, and this value was used in all subsequent simulations.

Reduced-rank regression

To quantify the low-dimensional linear mapping between neural representations and behavioral variables, we applied reduced-rank regression to relate neural predictors $X \in \mathbb{R}^{N \times p}$ to behavioral targets $Y \in \mathbb{R}^{N \times q}$. Both X and Y were first z-scored across samples. Let

$$C_{xx} = \frac{1}{N} X^T X ,$$

$$C_{xy} = \frac{1}{N} X^T Y ,$$

denote the covariance and cross-covariance matrices.

To ensure numerical stability in the presence of rank deficiency, we performed an eigen-decomposition of C_{xx} and retained only eigen-directions with eigenvalues exceeding a tolerance threshold. The retained subspace was whitened using:

$$W_x = U_x D_x^{-\frac{1}{2}},$$

where U_x and D_x denote the eigenvectors and eigenvalues of C_{xx} in the non-degenerate subspace. We then formed the whitened cross-covariance matrix:

$$M = W_x^T C_{xy},$$

and computed its singular value decomposition:

$$M = \hat{U} S V^T .$$

The top r singular vectors define the reduced-rank subspaces:

$$U_r = W_x \hat{U}_1^r ,$$

$$V_r = V_1^r .$$

The final low-rank regression mapping was constructed as:

$$W_{RRR} = U_r U_r^T (X \setminus Y),$$

where $X \setminus Y$ denotes the ordinary least-squares solution. Predicted behavior was obtained as:

$$\hat{Y} = XW_{RRR}.$$

This procedure extracts a rank r ($r = 2$) linear mapping that captures the dominant low-dimensional coupling between neural and behavioral variables.

Dynamic quantification using Lyapunov exponents

To quantify the dynamical regimes of both biological and model-derived time series, we computed the maximum Lyapunov exponent as a measure of trajectory divergence in state space. A positive maximum Lyapunov exponent indicates exponential sensitivity to initial conditions, values between 0 and 1 indicate mesostasis, and values near zero indicate marginal or fixed-point stability. For each trial, we extracted the original behavioral poses, neural population activity and reservoir state vector over time as a multivariate time series to calculate their maximum Lyapunov exponents using MATLAB's lyapunovExponent function. The sampling frequency is 30 Hz, with all other default parameters like an embedding dimension of 2, a time delay of 1, and an expansion range of [1, 5] steps. For each embedded point, the algorithm identifies its nearest neighbor in reconstructed phase space and evaluates the mean logarithmic divergence over the expansion range; the slope of this divergence curve, divided by the sampling interval.

To assess differences in dynamical regimes between cortical regions and model components, we performed pairwise comparisons of the maximum Lyapunov exponents across conditions. Considering the large number of comparisons and the focus on internal structural similarity rather than absolute statistical thresholds, we prioritized computational efficiency and interpretability. Therefore, we

directly applied paired two-sided Student's t-tests and Wilcoxon rank-sum tests to each pair of conditions. The sequence of comparisons followed the ordering of mean Lyapunov exponent values across conditions.

References

1 Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023). https://doi.org/10.1038/s41586-023-06031-6

2 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

3 Linderman, S. *et al.* in *Artificial intelligence and statistics.* 914–922 (PMLR).

4 Liu, M., Nair, A., Coria, N., Linderman, S. W. & Anderson, D. J. Encoding of female mating dynamics by a hypothalamic line attractor. *Nature* **634**, 901–909 (2024). https://doi.org/10.1038/s41586-024-07916-w

5 Vinograd, A., Nair, A., Kim, J. H., Linderman, S. W. & Anderson, D. J. Causal evidence of a line attractor encoding an affective state. *Nature* **634**, 910–918 (2024). https://doi.org/10.1038/s41586-024-07915-x

6 Mountoufaris, G. *et al.* A line attractor encoding a persistent internal state requires neuropeptide signaling. *Cell* **187**, 5998–6015 e5918 (2024). https://doi.org/10.1016/j.cell.2024.08.015

7 Lin, X. *et al.* A brain-inspired computational model for spatio-temporal information processing. *Neural Netw* **143**, 74–87 (2021). https://doi.org/10.1016/j.neunet.2021.05.015

8 Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, eaav7893 (2019).

9 McLachlan, G. J., Lee, S. X. & Rathnayake, S. I. Finite mixture models. *Annual review of statistics and its application* **6**, 355–378 (2019).

10 Wang, C. *et al.* BrainPy, a flexible, integrative, efficient, and extensible framework for general-purpose brain dynamics programming. *Elife* **12** (2023). https://doi.org/10.7554/eLife.86365