Extended Abstract Track

sa-SVAE: a Shared and Aligned Structured Variational Autoencoder for Extracting Behaviorally Relevant and Preserved Neural Dynamics Across Animals

Editors: List of editors' names

Abstract

Understanding the preserved behaviorally-relevant neural dynamics across individuals when performing similar tasks presents a critical challenge. Current methods typically focus on analyzing subject-specific neural dynamics or employing post-training alignment to adapt latent dynamics across sessions and individuals. Yet, establishing a shared latent space that effectively captures the continuous nature of behavioral data remains elusive. In this study, we introduce sa-SVAE, a Shared and Aligned Structural Variational AutoEncoder that integrates neural recordings from multiple subjects and uncovers the shared, behaviorally-relevant latent dynamics, facilitating the prediction of corresponding behaviors through a universal decoder. Utilizing a Structured Variational AutoEncoder (SVAE), our approach infers nonlinear latent factors and learns tractable dynamics driven by behavior on a circuit-level manifold. We employ contrastive learning to align low-dimensional, behaviorally-relevant geometries across subjects, thereby preserving the integrity of neural representations linked to specific behaviors across different sessions and subjects. This alignment enables the development of a unified behavior decoder that outperforms previous methods. Our model demonstrates robust decoding of task-relevant behaviors by capturing these preserved latent dynamics, underscoring the factors essential for cross-subject generalization. This study highlights the potential for building a universal behavior decoder and provides neuroscience insights into preserved and behaviorally constrained neural representations.

Keywords: Structured variational autoencoder, contrastive learning, latent dynamics.

1. Introduction

Brain-computer Interfaces (BCIs) have gained significant attention due to their potential to enable direct communication between the brain and machines (Hochberg et al., 2006; Chaudhary et al., 2016; Maiseli et al., 2023). One of the major challenges is decoding neural activity in a way that generalizes across multiple subjects when behaving under a similar context. Recent research suggests that despite individual differences in neural circuitry, preserved latent dynamics—common patterns of neural activity shared across individuals—are responsible for producing similar behavioral outputs (Safaie et al., 2023; Degenhart et al., 2020; Saha et al., 2017; Ray et al., 2015). This indicates that BCIs could be designed to decode the behavior of these conserved neural trajectories, enabling generalization between subjects under similar behavioral tasks. A universal behavioral decoder that can integrate and align neural data from different subjects with high robustness and performance is highly desirable. This necessitates the extraction of the preserved behaviorally-relevant latent dynamics governing specific behaviors across individuals. This capability is crucial not only for practical applications like neuroprosthetics but also for uncovering fundamental principles of brain function (Koralek et al., 2012; Portes et al., 2022).

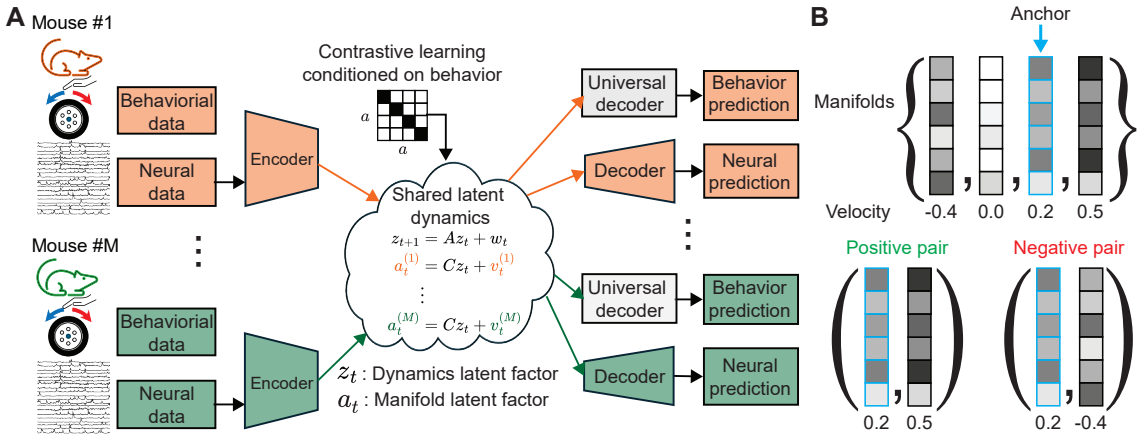

Figure 1: **A schematic overview of sa-SVAE architecture.** **A** The sa-SVAE model extracts low-dimensional latent factors from neural recordings using session-specific encoders, models their dynamics with a shared linear system, and aligns them to behavioral data via contrastive learning. **B** Contrastive learning for regression tasks by ranking the manifold latent factors based on their distances.

Aligning behaviorally-relevant latent dynamics across subjects remains a significant challenge. Traditionally, methods like canonical correlation analysis (CCA) have been employed to align the latent spaces of neural activities, facilitating the development of decoders that generalize across different subjects (Safaie et al., 2023). More recently, transformer-based neural foundation models have been utilized to harness large-scale neural data from various subjects, aiming to establish a universal neural representation (Azabou et al., 2024; Ye et al., 2024; Antoniades et al., 2023). However, these methods often do not provide a mathematical description of the temporal dynamics of neural activity, which is crucial for understanding continuous behavioral processes. Additionally, some approaches have attempted to project neural activities from different subjects into a common space using subject-specific projectors, such as multi-layer perceptrons (MLP), and then trained a shared dynamical system to capture these dynamics (Schneider et al., 2023; Pandarinath et al., 2018). Despite these efforts, the resulting latent dynamics still vary significantly between subjects, complicating the interpretation of shared behaviorally-relevant components.

In this study, we introduce the Shared and Aligned Structured Variational AutoEncoder (sa-SVAE), a novel model designed to address the integration of neural recordings from multiple subjects and the extraction of behaviorally-relevant latent dynamics during tasks. Our architecture incorporates the Structured Variational AutoEncoder (SVAE) framework, which supports tractable linear dynamics in the latent space, allowing us to capture the temporal structure of shared manifold factors. The process begins by projecting the neural activity from different subjects onto a common manifold using a subject-specific encoder implemented with MLPs (Figure 1A). We then model the latent factors with a linear dynamical system that operates consistently across subjects and sessions (Figure 1A). Finally,

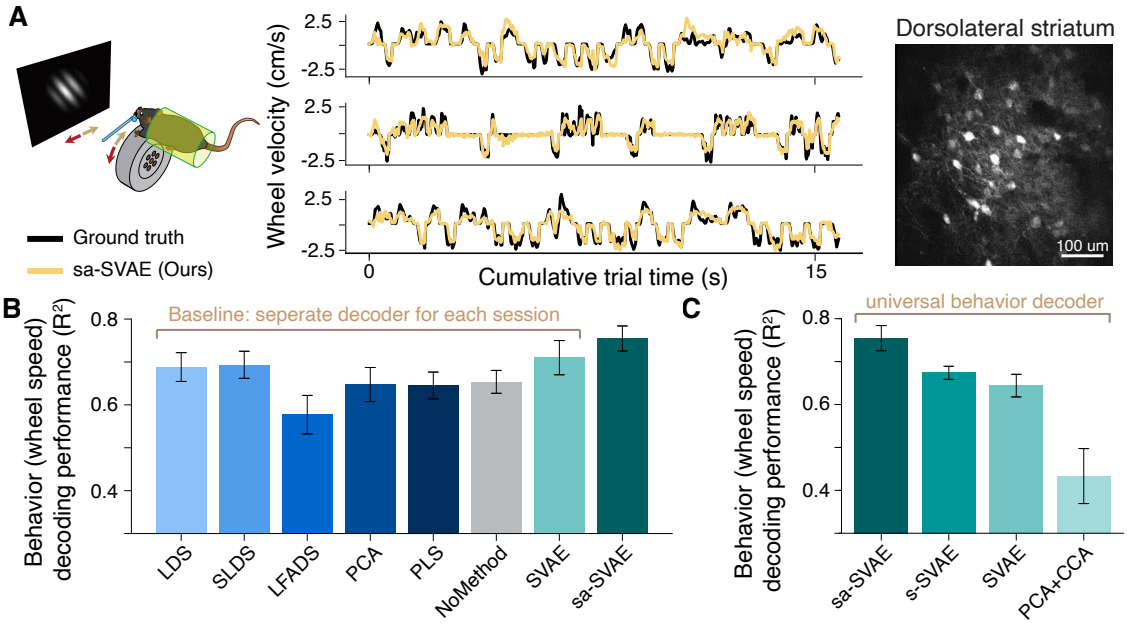

Figure 2: **Behavioral decoding performance from a universal linear decoder. A** Left: the steering wheel turning task. Middle: the ground truth and the predicted wheel velocities. Right: 2P calcium imaging in the dorsolateral striatum. **B** Ablation analysis. Averaged decoding R^2 with sa-SVAE, s-SVAE without contrastive learning, SVAE trained on each individual session, and PCA with CCA alignment. **C** Averaged decoding R^2 with LDS, SLDS, PCA, PLS, and decoding from raw neural activity. (Error bar: s.e.m across three sessions.)

we employ contrastive learning, guided by continuous behavioral data, to align these latent factors across various subjects and sessions (Figure 1B).

Our key contributions include: 1) A novel architecture that deciphers the shared and behaviorally-relevant neural dynamics to uncover underlying preserved patterns across subjects. 2) An innovative application of contrastive learning for regressions to generate latent factors conditioned on behavior, aligning the latent space across sessions and subjects for improved behavior decoding. 3) Demonstration of competitive performance and interpretability on real neural data. By addressing these challenges, our model provides an accurate and identifiable framework for universal BCIs, advancing our understanding of the preserved dynamics of the brain between subjects that drive the downstream behavior and paving the way for generalizable BCIs.

2. Experiments and Results

Datasets. All the methods are trained on two-photon calcium imaging data from the dorsolateral striatum when a mouse is conducting a steering wheel turning task, where the mouse needs to turn the wheel to move the visual cue from either left or right to the center

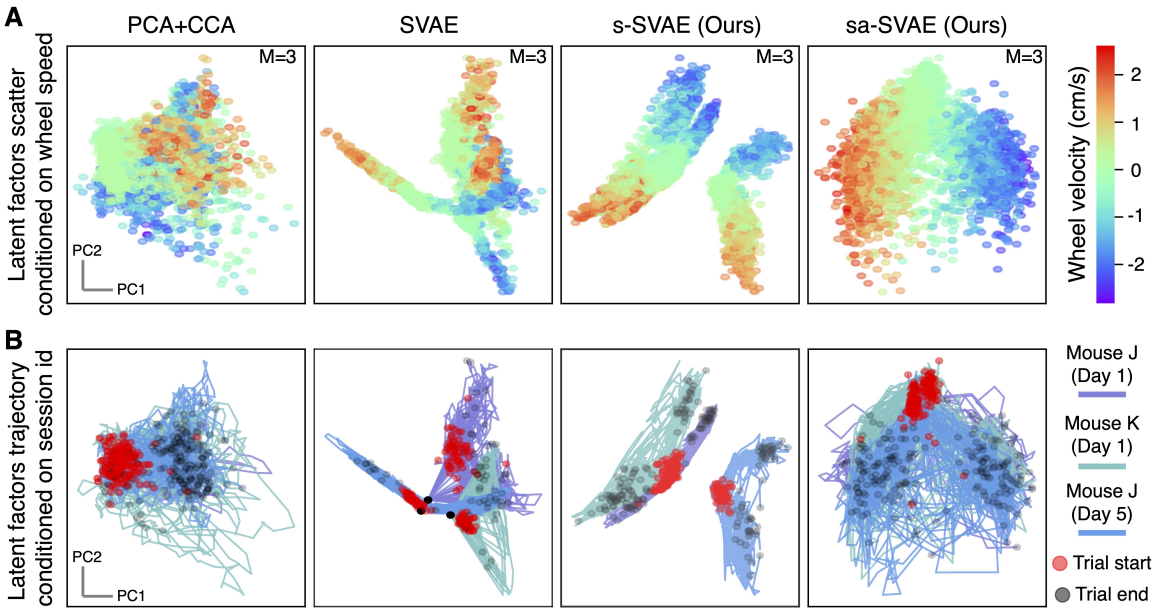

Figure 3: **Visualization of the latent factors.** The first two PC dimensions of the latent factors color-coded by (**A**) the wheel velocity and (**B**) session ID.

of the screen (Figure 2A). We used two mice and three sessions in total, with one session from mouse K and two sessions from mouse J. Other details regarding data collection and preprocessing details are detailed in Appendix C.

Behavior decoding performance comparison We first compared the performance of behavior decoding from latent dynamics between our sa-SVAE, linear dynamical systems (LDS), switching linear dynamical systems (SLDS), latent factor analysis via dynamical system (LFADS), principle component analysis (PCA) and partial least square regression (PLS) and compared with decoding directly from the raw neural activity (NoMethod). We trained a Lasso linear regression on the latent/original dynamics as the behavior decoder and showed that the sa-SVAE outperformed all other methods in this dataset Figure 2B.

Ablation analysis We analyzed the contribution of the two key components of the sa-SVAE framework through ablation analysis: 1) the shared dynamics across sessions and subjects; and 2) the contrastive learning between latent dynamics and behavior. The highest decoding performance can only be achieved with both components (Figure 2C).

Neural manifold analysis We visualized these manifolds to assess the sa-SVAE model's ability to extract the preserved latent dynamics across sessions and achieve effective alignment in the latent space. In Figure 3A, we plot the first two PC dimensions of the manifold latent factors a of four representative methods, color-coded by the behavior and the session. Notably, sa-SAVE demonstrated a better alignment of neural latent dynamics to behavior (Figure 3A) which is shared across all the sessions (Figure 3B) than other methods.

References

Hamidreza Abbaspourazad, Eray Erturk, Bijan Pesaran, and Maryam M. Shanechi. Dynamical flexible inference of nonlinear latent factors and structures in neural population activity. *Nature Biomedical Engineering*, 8(1):85–108, 2024. ISSN 2157-846X. doi: 10.1038/s41551-023-01106-1. URL https://doi.org/10.1038/s41551-023-01106-1.

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data. *arXiv preprint arXiv:2311.00136*, 2023.

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, 2024.

Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9): 513–525, September 2016. ISSN 1759-4766. doi: 10.1038/nrneurol.2016.113. URL https://doi.org/10.1038/nrneurol.2016.113.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

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. ISSN 2157-846X. doi: 10.1038/s41551-020-0542-9. URL https://doi.org/10.1038/s41551-020-0542-9.

Joshua Glaser, Matthew Whiteway, John P Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. *Advances in neural information processing systems*, 33:14867–14878, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 2016. doi: 10.1098/rsta.2015.0202. URL https://doi.org/10.1098/rsta.2015.0202.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Aaron C. Koralek, Xin Jin, John D. Long II, Rui M. Costa, and Jose M. Carmena. Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills. *Nature*, 483(7389):331–335, 2012. ISSN 1476-4687. doi: 10.1038/nature10845. URL https://doi.org/10.1038/nature10845.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. doi: 10.1109/ACCESS.2020.3031549.

Chengrui Li, Soon Ho Kim, Chris Rodgers, Hannah Choi, and Anqi Wu. One-hot generalized linear model for switching brain state discovery. *arXiv preprint arXiv:2310.15263*, 2023.

Baraka Maiseli, Abdi T. Abdalla, Libe V. Massawe, Mercy Mbise, Khadija Mkocha, Nassor Ally Nassor, Moses Ismail, James Michael, and Samwel Kimambo. Brain–computer interface: trend, challenges, and threats. *Brain Informatics*, 10(1):20, 2023. ISSN 2198-4026. doi: 10.1186/s40708-023-00199-3. URL https://doi.org/10.1186/s40708-023-00199-3.

Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.

Jacob Portes, Christian Schmid, and James M Murray. Distinguishing learning rules with brain machine interfaces. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25937–25950. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a6d94c38506f16fb50894a5b555f2c9a-Paper-Conference.pdf.

Andreas M. Ray, Ranganatha Sitaram, Mohit Rana, Emanuele Pasqualotto, Korhan Buyukturkoglu, Cuntai Guan, Kai-Keng Ang, Cristián Tejos, Francisco Zamorano, Francisco Aboitiz, Niels Birbaumer, and Sergio Ruiz. A subject-independent pattern-based brain-computer interface. *Frontiers in Behavioral Neuroscience*, 9, 2015. ISSN 1662-5153. doi: 10.3389/fnbeh.2015.00269. URL https://www.frontiersin.org/journals/behavioral-neuroscience/articles/10.3389/fnbeh.2015.00269.

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. ISSN 1476-4687. doi: 10.1038/s41586-023-06714-0. URL https://doi.org/10.1038/s41586-023-06714-0.

Simanto Saha, Khawza I. Ahmed, Raqibul Mostafa, Ahsan H. Khandoker, and Leontios Hadjileontiadis. Enhanced inter-subject brain computer interface with associative sensorimotor oscillations. *Healthcare Technology Letters*, 4(1):39–43, 2017. doi: https://doi.org/10.1049/htl.2016.0073. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/htl.2016.0073.

Omid G Sani, Hamidreza Abbaspourazad, Yan T Wong, Bijan Pesaran, and Maryam M Shanechi. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*, 24(1):140–149, 2021.

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

David Sussillo, Rafal Jozefowicz, L. F. Abbott, and Chethan Pandarinath. Lfads - latent factor analysis via dynamical systems, 2016. URL https://arxiv.org/abs/1608.06315.

Parsa Vahidi, Omid G Sani, and Maryam M Shanechi. Modeling and dissociation of intrinsic and input-driven neural population dynamics underlying behavior. *Proceedings of the National Academy of Sciences*, 121(7):e2212887121, 2024.

Yule Wang, Zijing Wu, Chengrui Li, and Anqi Wu. Extraction and recovery of spatio-temporal structure in latent dynamics alignment with diffusion model. *Advances in Neural Information Processing Systems*, 36, 2024.

Tengfei Xue, Fan Zhang, Leo R Zekelman, Chaoyi Zhang, Yuqian Chen, Suheyla Cetin-Karayumak, Steve Pieper, William M Wells, Yogesh Rathi, Nikos Makris, et al. Tractoscr: a novel supervised contrastive regression framework for prediction of neurocognitive measures using multi-site harmonized diffusion mri tractography. *Frontiers in Neuroscience*, 18:1411797, 2024.

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, 2024.

Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009. doi: 10.1152/jn.90941.2008. URL https://doi.org/10.1152/jn.90941.2008. PMID: 19357332.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36, 2024.

Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. *arXiv preprint arXiv:2301.08915*, 2023.

Appendix A. Methodology

In this section, we introduce our model architectures, which include the shared SVAE part and the contrastive learning section. Our model leverages the shared linear dynamical system in the SVAE structure to emphasize the preserved latent dynamics and we use a contrastive learning objective to align between two modalities: latent factors and behaviors.

A.1. Shared structured variational autoencoder

We denote the population neural activities for a single session $s \in \{1, \ldots, M\}$ as $\mathbf{X}^{(s)} = \left\{X_1^{(s)}, \ldots, X_{L^{(s)}}^{(s)}\right\}$, where $L^{(s)}$ is the number of trials in session s and each trial data is denoted as $X_i^{(s)} = \left[x_1^{(s,i)}, \ldots, x_{T_i^{(s)}}^{(s,i)}\right]^\top \in \mathbb{R}^{T^{(s)} \times N^{(s)}}$, where $N^{(s)}$ is the number of observed neurons in session s and $T_i^{(s)}$ is the trial length (*i.e.* number of time bins) for the i^{th} trial in session s.

We denote the associated dynamics latent factors for session s as $\mathbf{Z}^{(s)} = \left\{Z_1^{(s)}, \ldots, Z_{L^{(s)}}^{(s)}\right\}$ and the dynamics latent factors for each trial i of session s are $Z_i^{(s)} = \left[z_1^{(s,i)}, \ldots, z_{T_i^{(s)}}^{(s,i)}\right]^\top \in \mathbb{R}^{T_i^{(s)} \times n_z}$. We denote the manifold latent factors as $\mathbf{A}^{(s)} = \left\{A_1^{(s)}, \ldots, A_{L^{(s)}}^{(s)}\right\}$ and $A_i^{(s)} = \left[a_1^{(s,i)}, \ldots, a_{T_i^{(s)}}^{(s,i)}\right]^\top \in \mathbb{R}^{T_i^{(s)} \times n_a}$. Note that n_z and n_a are the factors dimensionality to be picked, and we chose $n_z = n_a = 8$ in this study. We further define $\mathcal{X} = \left\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}\right\}$ as the set of neural activities from all sessions, and $\mathcal{Z} = \left\{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}\right\}$ and $\mathcal{A} = \left\{\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(M)}\right\}$ as the corresponding set of dynamics latent factors and manifold factors.

In this work, we consider an SVAE with a linear Gaussian dynamical system prior for sequential data. This generative model defines a joint distribution over the latent factors and the observed neural activity as

$$p_{\theta,\phi}(\mathcal{X}, \mathcal{Z}) = \prod_{\mathbf{X},\mathbf{Z} \in \mathcal{X},\mathcal{Z}} \prod_{X,Z \in \mathbf{X},\mathbf{Z}} p_\theta(z_1) \prod_{t=2}^{T} p_\theta(z_t \mid z_{t-1}) \prod_{t=2}^{T} p_\phi(x_t \mid z_t) \tag{1}$$

where θ denotes the prior parameters and ϕ denotes the parameters of the decoder. Note that the outer summation is over all M sessions and we remove the superscript (s) in the above equation for simplicity. The preserved linear dynamical system (LDS) satisfies

$$\begin{aligned} p_\theta\left(z_1^{(s)}\right) &= \mathcal{N}\left(z_1^{(s)}; \mu_1, Q_1\right) \\ p_\theta\left(z_t^{(s)} \mid z_{t-1}^{(s)}\right) &= \mathcal{N}\left(z_t^{(s)}; A z_{t-1}^{(s)}, Q\right) \\ p_\theta\left(a_t^{(s)} \mid z_t^{(s)}\right) &= \mathcal{N}\left(a_t^{(s)}; C z_t^{(s)}, R\right) \\ p_\phi\left(x_t^{(s)} \mid a_t^{(s)}\right) &= \mathcal{N}\left(x_t; f_\phi(a_t^{(s)}), V\right) \end{aligned} \tag{2}$$

where $f_\phi(\cdot)$ represents the decoder in the autoencoder structure, which can be parameterized with a deep neural network with weights ϕ, and θ represents the LDS parameters $\{A, C, \mu_1, Q_1, Q, R, V\}$.

For efficiency, simplicity, and the capability to include behavior supervision learning, we leveraged k-step-ahead prediction error (Abbaspourazad et al., 2024) together with a penalty on behavior prediction using a linear decoder that maps from the manifold latent factors to the behavior. We set $k = 4$ in this study. As a result, the overall loss function is written as

$$\mathcal{L}_{\text{model}} = \frac{1}{M}\sum_{s=1}^{M}\frac{1}{L^{(s)}}\sum_{\kappa=1}^{K}\sum_{i=1}^{T_i^{(s)}}\text{MSE}(x_{i+k|i}, x_{i+k}) + \lambda_{\theta,\phi}L_2(\theta,\phi) \tag{3}$$

$$\mathcal{L}_{\text{behavior}} = \frac{1}{M}\sum_{s=1}^{M}\frac{1}{L^{(s)}}\text{MSE}\left(g_\gamma\left(a_i^{(s)}\right), y_i^{(s)}\right) + \lambda_\gamma L_2(\gamma) \tag{4}$$

where $g_\gamma(\cdot)$ is a universal behavior decoder shared across subjects. In this work, we use a Lasso linear regression as the behavior decoder for all methods.

A.2. Contrastive learning for regressions tasks

To align the distances in the neural manifold space ordered by the distances in the corresponding behavioral data, we leveraged one of the contrastive learning techniques for regression tasks - the Rank-N-Contrast loss ($\mathcal{L}_{RNC}$), which ranks the latent manifold factors according to their target distances, and then contrasts them against each other based on their relative rankings (Zha et al., 2024). These techniques ensure that closer labels in the target space are also closer in the representation space.

For a given anchor $a_h^{(s,i)}$, *i.e.*, a manifold factor, from session s with associated $y_h^{(s,i)}$ as the behavior data, we define a set of samples that have higher *ranks* than the anchor $a_h^{(s,i)}$ as $\mathcal{S}_{h,j} := \left\{a_k \mid k \neq h, d\left(y_h^{(s,i)}, y_k^{(s,i)}\right) \geq d\left(y_h^{(s,i)}, y_j^{(s,i)}\right)\right\}$, where $d(\cdot,\cdot)$ is the distance measure between two behavioral labels (e.g., L_1 distance). The normalized likelihood of $a_j^{(s,i)}$ given the anchor and the positive set is

$$\mathbb{P}\left(a_j^{(s,i)} \mid a_h^{(s,i)}, \mathcal{S}_{h,j}\right) = \frac{\exp\left(\text{sim}\left(a_h^{(s,i)}, a_j^{(s,i)}\right)/\tau\right)}{\sum_{a_k^{(s,i)}\in\mathcal{S}_{\tau,j}}\exp\left(\text{sim}\left(a_h^{(s,i)}, a_k^{(s,i)}\right)/\tau\right)} \tag{5}$$

where $\text{sim}(\cdot,\cdot)$ is the similarity measure between two manifold factors (e.g., negative L_2 norm) and τ as the temperature parameter.

As a result, we define a contrastive loss based on the ranking of the continuous behavior samples (we obviate the upper script (s,i) for simplicity):

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{N}\sum_{h-1}^{N}\frac{1}{N-1}\sum_{j=1,j\neq h}^{N} -\log\frac{\exp(\text{sim}(a_h,a_j)/\tau)}{\sum_{a_k\in\mathcal{S}_{h,j}}\exp(\text{sim}(a_h,a_k)/\tau)} \tag{6}$$

A.3. Extraction of shared and behaviorally aligned latent factors

Combining the shared SVAE together with behavior supervision and contrastive learning, we propose an architecture that aims to identify the preserved and behaviorally-relevant neural manifold under a specific behavioral task:

$$\mathcal{L}_{\text{sa-svae}} = \mathcal{L}_{\text{model}} + \lambda_{\text{behavior}}\mathcal{L}_{\text{behavior}} + \lambda_{\text{contrastive}}\mathcal{L}_{\text{contrastive}} \tag{7}$$

where λ_{behv} is the scaling factor the behavior decoding loss and λ_{con} is the scaling factor for the contrastive learning loss.

Appendix B. Model architecture and training details

The latent dimension for all methods is set to 8.

B.1. SVAE family models

For all the SVAE structures, we used a two-layer MLP, where the number of nodes is 64 and 16 respectively. The number of steps for calculating the reconstruction loss function is 4. The scale for regularization $\lambda_{\theta,\phi} = 1 \times 10^{-5}$. The temperature for the RNC-loss is 0.2 and contrastive learning loss scaler λ_{con} is 1.0. The batch size for training is 32 and the maximum number of time points for contrastive learning is 2048. We used Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 1×10^{-3}. The learning rate decay is 0.9 every 15 steps.

We trained all the models for 500 epochs and chose the model with the best validation loss to perform all the analysis in this work.

B.2. LFADS

We individually trained an LFADS model and the downstream behavior decoder for each session. We used a two-layer MLP for the encoder and decoder in LFADS, where the hidden layer size is 100 and 50 respectively. The latent dimension of the distribution is equal to 32. The dropout is set to 0.1 and the training batch size is set to 20. We used Adam optimizer with an initial learning rate set to 1×10^{-3}. We trained the model for 500 epochs and chose the model with the best validation loss to perform the analysis. Note that in future work, we will perform a more thorough sweep of the hyperparameters to prevent the overfitting issue as depicted in Fig. 2 B. Moreover, we will try the stitch-LFADS (Pandarinath et al., 2018) to leverage all sessions for training and compare the performance with the sa-SVAE in the future.

B.3. LDS and SLDS

We set the dynamics matrix to be Gaussian and the emission matrix to be orthogonal Gaussian. The number of states in the SLDS model is selected between 1, 2, and 3 based on the behavior decoding performance in the validation set.

Appendix C. Data collection and processing

To study the relationship between neural activity and motor behavior, we head-fixed the mice and had them manipulate a steering wheel to control a visual cursor on a screen during neural recordings. The initial position of the cursor on the iPad screen was randomly selected. Concurrently, neural activities were collected using a two-photon microscope, providing a 400-µm wide field of view at a sampling rate of 30 Hz.

Table 1: Neural data summary

Session	# of neurons	# of trials	# of frames
Mouse J - Day 1	239	99	2,119
Mouse K - Day 1	173	89	2,434
Mouse J - Day 5	259	118	3,822
Total	671	306	8,375

The neural recordings of each session captured ~ 220 neurons resulting in a total number of 671 individual neurons across the three sessions. Each session included approximately 100 trials and the trial length varies (usually between 0.5 seconds to 4 seconds, *i.e.*, 15 - 120 time bins) as the reward was initiated when the cursor is successfully turned to the center of the screen. We also binned the velocity of the mouse's behavior to account for 30 Hz sampling rate, resulting in time-aligned neural and behavior data.

Appendix D. Related work

Populational neural activities and neural manifold The rich spatiotemporal correlation in population activity demonstrates the potential to model the high-dimensional nature of neural activity in terms of low-dimensional latent factors constrained within a neural manifold. The temporal structure of the latent factors, *i.e.,* latent dynamics, can be modeled by simple linear dimensionality reduction methods (Jolliffe and Cadima, 2016; Yu et al., 2009; Safaie et al., 2023), linear dynamical systems (Sani et al., 2021; Vahidi et al., 2024), switching linear dynamical systems (Glaser et al., 2020; Li et al., 2023), variational autoencoders (Sussillo et al., 2016; Abbaspourazad et al., 2024), transformers (Azabou et al., 2024; Ye et al., 2024; Antoniades et al., 2023), and diffusion models (Wang et al., 2024). These models help uncover the underlying structure of population activity, allow for effective behavior decoding, and gain insights into brain functions.

Constrastive learning in data fusion Contrastive learning has emerged as a popular technique for self-supervised representation learning (Le-Khac et al., 2020). Previously, it gained more attention in the context of discrete classification or segmentation tasks (Chen et al., 2020; He et al., 2020; Khosla et al., 2020). Recently, machine learning scientists have developed regression-aware representation learning techniques by contrasting samples against each other based on their ranking in the target space (Zha et al., 2024; Xue et al., 2024; Zhang et al., 2023). For the field of computational neuroscience, the contrastive learning for discrete cases has been applied in multimodal models to align neural activity with external sensory inputs, *e.g.*, behavioral and/or visual stimuli (Antoniades et al.,

2023; Schneider et al., 2023) and demonstrated promising outcomes in finding behaviorally-relevant latent dynamics, resulting in high behavioral decoding performance.

Appendix E. Discussion

In this work, we presented a novel Shared and Aligned Structured Variational AutoEncoder (sa-SVAE) model designed to capture behaviorally-relevant latent dynamics across multiple subjects performing similar tasks. The core strength of our model lies in its ability to integrate continuous behavioral data and effectively decode behavior from neural activity across different subjects. Our method outperformed traditional approaches in decoding performance including a diverse range of dynamical system methods and alignment methods. Furthermore, by employing contrastive learning, we improved the alignment of neural manifolds across subjects, demonstrating a significant advantage over traditional PCA+CCA approaches and recent shared dynamical system approaches. Therefore, the sa-SVAE framework can potentially provide a more interpretable and identifiable solution, particularly through the neural manifold alignment to reduce noise from individual variability, offering insights into the neural mechanisms underlying behavior.

Our current study focused on calcium imaging data from the striatum of mice. Future work will include: 1) thoroughly investigating the effect of the contrastive loss and the behavior decoding loss on the model training using synthetic datasets, 2) studying multiple brain regions, including motor cortex and cerebellum to capture more comprehensive neural dynamics, 3) expanding the model to other species, including primates or humans, for broader applicability. Furthermore, building on our linear behavior decoder, we plan to explore nonlinear decoding methods to gain deeper insights into how complex behaviors are encoded. By addressing these areas, we seek to further enhance both the accuracy and interpretability of neural-behavioral models and advance our understanding of the neural basis of behavior across species and brain regions.