SPHERICAL RECOGNITION-PARAMETERIZED STATE SPACE MODELS FOR LEARNING SHARED NEURAL AND BEHAVIORAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Simultaneous recordings of neural activity and behavior offer a window into how brain dynamics relate to an animal's actions. However, recovering shared dynamical structure from multimodal timeseries is challenging when each modality contains substantial private variation. Generative approaches must reconstruct this variation, while standard contrastive methods learn shared representations without explicitly modeling their evolution. We introduce sRP-SSM, a spherical recognition-parameterized state space model that jointly learns cross-modal representations and stochastic latent dynamics without reconstructing observations. The spherical geometry allows recognition factors to remain log-linear in the latent state even with observation-dependent confidence, so evidence from each modality and from past timesteps combines by simple addition. We derive a scalable contrastive training objective that approximates the model's evidence lower bound, connecting contrastive representation learning with probabilistic modeling of shared dynamics. In synthetic experiments, sRP-SSM combines strong shared-state recovery and cross-modal retrieval with more accurate dynamics recovery than competing generative and contrastive methods. In simultaneous neural and 3D pose recordings from a freely moving mouse, sRP-SSM improves cross-modal retrieval, future prediction, and keypoint decoding. During rearing, it detects behavior from neural activity and recovers bout-averaged neural spectral patterns from pose. These results show how using shared dynamics to guide representation learning can improve the recovery of neural–behavioral relationships, establishing a foundation for dynamics-informed learning across complementary modalities.

1 INTRODUCTION

Many scientific systems are observed through different synchronized modalities. For instance, in neuroscience, behavior may be measured through video-derived pose trajectories while neural activity is measured through electrophysiology or calcium imaging. These measurements reflect a shared underlying process, but each modality also contains substantial structured private variation, such as pose estimation noise or neural activity unrelated to movement. A central modeling problem is to recover the *shared dynamical structure* from observations so that we can infer how the common state evolves and what it encodes from both modalities.

A common approach is to fit a generative latent-variable model in which a shared latent state gives rise to all observed modalities (Wu & Goodman, 2018). Multimodal variational autoencoders and generative state space models can incorporate uncertainty, missing observations, and temporal dynamics. Their likelihood objectives, however, require them to account for variation specific to each modality. When that private variation is high-dimensional or structured, modeling it can degrade recovery of the shared latent. Introducing shared and private latent variables can help, but requires assumptions about how variation factorizes across modalities, ad-hoc regularization terms, and sufficiently expressive generative models for each.

Contrastive methods offer an appealing alternative to reconstructing observations: they learn representations to distinguish matching samples from mismatched ones (van den Oord et al., 2018; Radford et al., 2021; Schneider et al., 2023). In cross-modal matching, features shared across modalities

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.

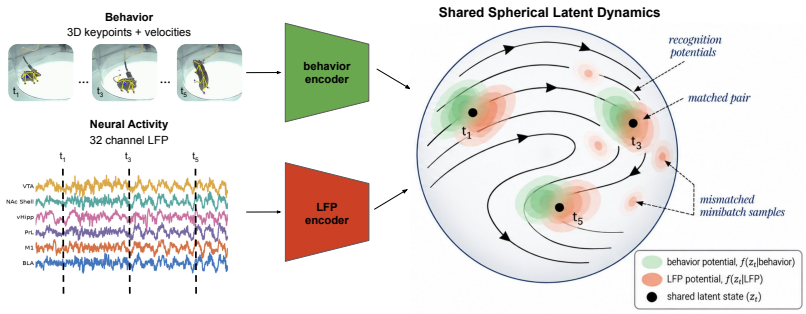

Figure 1: sRP-SSM encodes two timeseries modalities using modality-specific encoders (**left**) to a shared dynamical latent space (**right**), explicitly modeling their shared dynamic variability without reconstruction. Each encoder maps observations to modality- and timepoint-specific recognition potentials, which are combined across modalities with a product of experts (PoE) construction. The approximate ELBO objective uses mismatched within-batch recognition potentials as contrasts.

help identify a match, while modality-specific features offer less help. Applied to multimodal timeseries, FOCAL learns shared and private features and encourages temporal locality in embeddings without an explicit model of dynamics (Liu et al., 2023). CANDY uses behavioral similarity to align neural embeddings across sessions while jointly fitting a linear dynamical system (Jiang et al., 2025). We seek a model in which both modalities provide evidence about the same evolving latent state, with explicit stochastic dynamics and the ability to infer the state from either modality or both together.

Recognition-parameterized models (RPMs) connect contrastive learning to probabilistic latent variable models (Walker et al., 2023; Aitchison & Ganev, 2024). Instead of learning a decoder that reconstructs an observation from a latent state, an RPM uses an encoder to score plausible latent states for an observation. This permits inference about latent variables without having to model irrelevant variability in the observations. Hromadka et al. (2026) extend this approach to timeseries with a Gaussian latent dynamical model, but study a single observation modality and use an objective that does not readily decompose over minibatches (Hromadka et al., 2026). We build on this foundation to combine evidence from multiple modalities about a shared evolving state and train with a minibatch objective.

We introduce sRP-SSM (spherical recognition-parameterized state space model) to recover shared dynamical structure from multimodal timeseries. Modality-specific encoders map observations or temporal windows to evidence about a shared latent state, which evolves according to stochastic dynamics on a hypersphere (Figure 1). On the sphere, recognition factors can express observation-dependent confidence while retaining a simple form for combining evidence across modalities and time. Comparing matched observations with mismatched observations in a minibatch then yields a CLIP-like training objective approximating the model's evidence lower bound.

Our primary contributions are:

1. **Multimodal latent dynamics without reconstruction:** We extend RPMs to fuse evidence from either or both modalities into a shared evolving latent state, with spherical recognition factors enabling simple, confidence-weighted fusion.
2. **Minibatch training with a probabilistic objective:** We derive a CLIP-like minibatch approximation to the sRP-SSM evidence lower bound and verify its agreement with a reference ELBO estimate in simulation.
3. **Recovery of shared dynamics:** On synthetic data with known shared and private variation, sRP-SSM combines strong latent recovery and cross-modal retrieval with more accurate estimates of the underlying dynamics than the tested baselines.
4. **Neural-behavioral application:** In synchronized mouse LFP and pose recordings, sRP-SSM improves cross-modal retrieval and neural-to-behavior decoding. We use its learned state to examine shared neural-behavioral structure during rearing.

2 BACKGROUND

Setup and Notation: We consider two synchronized modalities that have both shared and modality-specific variation. Our goal is to learn a shared dynamic representation that can be inferred from either modality or both, without reconstructing the observations. Let $x_t^{(a)} \in \mathbb{R}^{D_a}$ and $x_t^{(b)} \in \mathbb{R}^{D_b}$ for $t = 1, \ldots, T$ be the two synchronized modalities. We assume a shared latent variable z_t on the unit sphere $\mathbb{S}^{d-1}$ that explains the dependence between the observations across modalities at the same time t, and across time via a dynamical prior on latent trajectories. Conditional on the latent trajectory, observations are independent across modalities and time, with each $x_t^{(m)}$ depending only on z_t.

Recognition-parameterized models: A recognition-parameterized model (Walker et al., 2023) replaces generative observation terms $p_\phi(x^{(m)}|z)$ for modality $m \in \{a, b\}$ in a typical latent variable model by a learned *recognition factor* $f_\phi^{(m)}(z \mid x^{(m)})$, a normalized density over the latent. Let $p_0^{(m)}$ denote the empirical reference marginal of modality m, i.e. $p_0^{(m)}(x) = {}^1\!/\!_T \sum_t \delta(x - x_t^{(m)})$. Using these pieces, the RPM observation likelihood is given by

$$\tilde{p}_\phi^{(m)}(x \mid z) = p_0^{(m)}(x) \frac{f_\phi^{(m)}(z|x)}{F_\phi^{(m)}(z)} , \qquad F_\phi^{(m)}(z) = \mathbb{E}_{x \sim p_0^{(m)}}[f_\phi^{(m)}(z \,|\, x)] , \tag{1}$$

where the *aggregate recognition density* $F_\phi^{(m)}$ normalizes the recognition factors over observations. The RPM likelihood then specifies which observations are compatible with a latent state relative to the empirical reference distribution.

InfoNCE and mixture estimation. For a fixed latent target z and J candidate observations, let $s_j(z)$ denote the compatibility score between z and candidate x_j, with x_+ the positive observation. The one-way InfoNCE loss is $\mathcal{L}_{\text{InfoNCE}} = -s_+(z) + \log \sum_{j=1}^{J} \exp s_j(z)$ (van den Oord et al., 2018; Poole et al., 2019). Methods like CLIP (Radford et al., 2021) and CEBRA (Schneider et al., 2023) use a two-way symmetrized version of InfoNCE by applying the one-way classification loss in both modality directions.

von Mises–Fisher distributions. Let ν denote the uniform distribution on $\mathbb{S}^{d-1}$. Let $z \in \mathbb{S}^{d-1}$ and $e \in \mathbb{R}^d$. The von Mises–Fisher density with respect to ν is given by

$$\text{vMF}(z; e) = C_d(\|e\|) \exp(e^\top z) , \qquad C_d(\|e\|) = \left[\mathbb{E}_{z \sim \nu} \exp(e^\top z)\right]^{-1} , \tag{2}$$

where $e/\|e\|$ gives the direction and $\|e\|$ gives the concentration of the distribution, and $e = 0$ gives the uniform distribution. Products of vMF potentials remain proportional to a vMF density, with natural parameters that add: $\prod_i \text{vMF}(z; e_i) \propto \text{vMF}(z; \sum_i e_i)$. Additional vMF identities used in Section 3 are collected in Appendix B.

3 SPHERICAL RECOGNITION-PARAMETERIZED STATE SPACE MODELS

Here we introduce sRP-SSM, which uses encoders to map observations to evidence about a shared spherical latent state. A causal recurrence combines available evidence with transported past evidence. We train representations and dynamics jointly using an approximate RPM evidence lower bound (ELBO), while masking enables inference from either modality.

3.1 MODEL

Each modality emits a recognition potential via a neural encoder with parameters ϕ,

$$f_\phi^{(m)}(z_t \,|\, x_t^{(m)}) = \text{vMF}\Big(z_t; e_m(x_t^{(m)})\Big) , \quad m \in \{a, b\}, \quad t \in \{1, \ldots, T\} . \tag{3}$$

We choose a spherical rather than a Euclidean latent geometry for our model. The spherical vMF construction supports additive evidence fusion, a uniform stationary prior, and an analytic KL to the

dynamical prior under our variational family. A further advantage in contrastive settings is discussed in Section 4. The shared latent on $\mathbb{S}^{d-1}$ rotates under a learned orthogonal transform R, keeping the latent on the sphere, then diffuses with vMF concentration κ_w:

$$p_\psi(z_t \mid z_{t-1}) = \mathrm{vMF}\big(z_t; \kappa_w R z_{t-1}\big), \qquad R = \exp(\Omega), \qquad \Omega = -\Omega^\top, \tag{4}$$

where $\psi = (\Omega, \kappa_w)$. Large κ_w produces highly predictable transitions, while $\kappa_w = 0$ makes consecutive states independent. Rotation plus isotropic noise leaves the uniform distribution on the sphere invariant, so we take $p_\psi(z_1) = \nu$.

Writing $X = \{x_t^{(a)}, x_t^{(b)}\}_{t=1}^T$, the multimodal RPM joint takes the form

$$\tilde{p}_\theta(X, z_{1:T}) = p_\psi(z_{1:T}) \prod_{t=1}^T \prod_{m\in\{a,b\}} \frac{f_\phi^{(m)}(z_t \mid x_t^{(m)})\, p_0^{(m)}(x_t^{(m)})}{F_\phi^{(m)}(z_t)}, \tag{5}$$

where $\theta = (\phi, \psi)$ and $p_\psi(z_{1:T}) = p_\psi(z_1) \prod_{t>1} p_\psi(z_t \mid z_{t-1})$. Integrating out the observations recovers the dynamical prior $p_\psi(z_{1:T})$, while coupling both modalities to the same latent states induces their statistical dependence.

3.2 Variational Inference

We introduce a tractable variational approximation to the latent posterior, enabling optimization of an evidence lower bound (ELBO) on the observation log likelihood (Blei et al., 2017). To infer the shared state from either or both modalities, we construct a variational posterior using causal vMF beliefs $q_t(z_t) = \mathrm{vMF}(z_t; \lambda_t)$, whose parameters depend only on the observations up to time t. We choose the factorized form $q(z_{1:T} \mid X, \delta) = \prod_t q(z_t | X_{1:t}, \delta_{1:t})$, where δ specifies which modalities enter inference. Although q is factorized over time, its parameters are coupled by a recurrence that combines current encoder evidence with rotated and discounted previous beliefs:

$$\lambda_t = \underbrace{\delta_t^{(a)} e_a(x_t^{(a)})}_{\text{modality-}a\text{ evidence}} + \underbrace{\delta_t^{(b)} e_b(x_t^{(b)})}_{\text{modality-}b\text{ evidence}} + \underbrace{cR\lambda_{t-1}}_{\text{transported past}}, \qquad \lambda_0 = 0\,, \tag{6}$$

where $\delta_t^{(a)}, \delta_t^{(b)} \in \{0, 1\}$ are masks indicating which modalities are available to the filter. The fixed gain $c \in [0, 1)$ controls discounting of past evidence: $c = 0$ uses only current evidence while larger values integrate evidence over longer horizons. The addition of natural parameters corresponds to multiplication of vMF factors, so this update preserves a normalized vMF belief. When both modalities enter inference, the first two terms implement a product-of-experts (PoE) construction that combines evidence of the shared latent variable. When only one is present, the same recursion gives the single-modality beliefs required for cross-modal prediction.

The true Bayesian filtering update is given by

$$p(z_t|X_{1:t}) \propto \int p(z_{t-1}|X_{1:t-1}) p(z_t|z_{t-1}) \mathrm{d}z_{t-1} \prod_m \frac{f_\phi^{(m)}(z_t|x_t^{(m)})}{F_\phi^{(m)}(z_t)}\,. \tag{7}$$

Neither the integral, which describes the latent dynamics, nor the product, which describes the observation–latent compatibility, is generally a vMF factor. Therefore, the recurrence that defines our variational family is not an exact Bayesian update, but a simple and practical choice. Causal conditioning supports online inference, while temporal factorization makes the objective below tractable. Alternative filtering updates are discussed in Appendix C.

3.3 Objective

We jointly learn the encoders and dynamics by maximizing a tractable approximation to the RPM ELBO. Omitting the parameter-independent term $p_0(X) = \prod_{t,m} p_0(x_t^{(m)})$, the ELBO consists of a recognition reward and a KL penalty to the dynamical prior:

$$\mathcal{L}_{\mathrm{ELBO}} = \underbrace{\sum_{t,m} \mathbb{E}_{q_t}\left[\log \frac{f_\phi^{(m)}(z_t \mid x_t^{(m)})}{F_\phi^{(m)}(z_t)}\right]}_{\text{recognition}} - \underbrace{\mathrm{KL}(q_\theta \,\|\, p_\psi)}_{\text{dynamics}}\,, \tag{8}$$

where $\log \tilde{p}_\theta(X) \geq \log p_0(X) + \mathcal{L}_{\mathrm{ELBO}}$. We discuss the two terms in turn: the recognition term, whose intractable normalizer we estimate with contrastive negatives, and the dynamics term, which is available in closed form.

Recognition terms as contrastive classification. The recognition term contains the aggregate recognition density $F_\phi^{(m)}$, a mixture over the entire dataset. We replace it by a minibatch estimate over J candidates, including the paired observation $x_t^{(m)}$: $\widehat{F}_\phi^{(m)}(z) = \frac{1}{J}\sum_j f_\phi^{(m)}(z \mid x_j^{(m)})$. Writing $e_j^{(m)}$ for the natural parameter emitted by candidate j, its compatibility score with a latent target z is

$$s_j^{(m)}(z) = e_j^{(m)\top} z + \log C_d(\|e_j^{(m)}\|) \,, \tag{9}$$

where the normalizer remains in the score because encoder concentrations vary across candidates. The recognition reward for target t is then

$$\ell_t^{(m)} = \mathbb{E}_{q_t}\big[s_t^{(m)}(z_t) - \log \tfrac{1}{J}\textstyle\sum_{j=1}^{J} \exp s_j^{(m)}(z_t)\big] \,, \tag{10}$$

where $s_t^{(m)}$ denotes the matching candidate compatibility score. For each latent sample, the reward is $\log J$ minus the one-way InfoNCE loss that identifies the matching observation among same-modality candidates (Aitchison & Ganev, 2024). Crucially, both modalities contribute to the recognition objective, regardless of which modalities enter the filter. When only modality a is available, the resulting belief must identify the paired modality b observation and vice versa. Evidence masking therefore turns the recognition objective into cross-modal prediction without a separate alignment loss.

Expectation over the belief. The positive score is affine in z_t, so its expectation is exactly $s_t^{(m)}(\bar{z}_t)$, writing $\bar{z}_t = \mathbb{E}_{q_t}[z_t]$ for the belief mean, available analytically for vMF distributions. Jensen's inequality bounds the remaining log mixture, and each overlap in the resulting expression is analytic (Appendix B):

$$\hat{\ell}_t^{(m)} = s_t^{(m)}(\bar{z}_t) - \log\big[\tfrac{1}{J}\textstyle\sum_j \mathbb{E}_{q_t} f_\phi^{(m)}(z_t|x_j^{(m)})\big] \le \ell_t^{(m)} \,. \tag{11}$$

This is a tractable lower bound on the recognition contribution for a fixed candidate set, though it does not remove the finite-candidate approximation to the full aggregate recognition density. Alternative expectation estimators are compared in Appendix E.

Closed-form dynamics contribution. Temporal factorization of q and the bilinear transition log density give an analytic KL contribution:

$$-\mathrm{KL}(q_\theta \| p_\psi) = -\textstyle\sum_{t=1}^{T} \mathrm{KL}(q_t \| \nu) + \sum_{t=2}^{T} \left[\kappa_w \left\langle R\bar{z}_{t-1}, \bar{z}_t \right\rangle + \log C_d(\kappa_w)\right] , \tag{12}$$

with the KL to uniform available in closed form (Appendix B). The KL to uniform penalizes concentrated beliefs, while the transition term encourages predictable consecutive beliefs. Unlike the recognition normalizer, the transition normalizer is state-independent, so this term requires no contrastive approximation. Since R enters both here and in the filter (Eq. 6), the transition model is shaped both by how well it predicts and by how useful it is for accumulating evidence. Our training objective is $\mathcal{J} = \sum_{t,m} \hat{\ell}_t^{(m)} - \mathrm{KL}(q_\theta \,\|\, p_\psi)$, combining the candidate-based recognition bound with the analytic KL for the chosen variational family.

3.4 Training

Modality subsampling. During training we stochastically mask the evidence entering the filter, sampling only modality a with probability 0.4, only b with probability 0.4, and both with probability 0.2. The mask is held fixed across each window and affects only the filter; both modalities always remain in the recognition objective. Each mask defines a valid variational distribution. Then, for the exact RPM objective, averaging over masks averages valid variational bounds.

Windows and candidates. The objective is a sum of local terms, but each depends on a filter state that incorporates the entire past and on an aggregate recognition density over the entire dataset. We therefore estimate it from windows. A minibatch consists of B windows of W contiguous timesteps with start indices drawn uniformly. Within each window the filter is run from the uniform prior, $\lambda = 0$, and the first t_{burn} timesteps warm the filter and are excluded from the scored objective. Because R is orthogonal and $c < 1$, the influence of the unknown pre-window state decays exponentially with burn-in length. In practice we ramp a weight β_{dyn} on the KL penalty from 0 to 1 early in training. $\beta_{\mathrm{dyn}} = 1$ is the setting under which the objective approximates the ELBO.

4 Discussion of Model Design Choices

We next discuss two design choices underlying sRP-SSM: why spherical latent geometry is particularly well-suited to first-order recognition potentials, and why RPMs avoid the reconstruction-driven pressure to encode modality-specific variation that has been identified in the multimodal VAE literature.

Latent geometry and first-order recognition. For first-order recognition potentials, in which the conditioning variable interacts with the target only through a linear statistic, latent geometry determines which density ratios can be represented exactly. In Euclidean Gaussian models, curvature in the target arises in two relevant settings. When candidate observations are contrasted against a fixed latent state, as in the RPM recognition logit (Eq. 9), observation-dependent precision introduces a quadratic term in z. When candidate targets are contrasted, as in transition-density ratios or CPC-style critics, nonzero Gaussian dependence also introduces quadratic target curvature. On a fixed-radius sphere, isotropic quadratic curvature is constant, so vMF factors can represent both per-observation confidence and nonzero dependence while remaining first-order. This is particularly useful in sRP-SSM: vMF natural parameters add under product-of-experts fusion, and the vMF transition yields a stationary process whose transition density ratio is affine in the next state (Eq. 24). Appendix D.1 gives the corresponding Gaussian calculations and characterizes the anisotropic curvature that fixed radius does not remove. Euclidean RPMs remain possible, but exact Gaussian formulations require critics with at least second-order dependence on the target.

RPMs avoid reconstruction-driven private-variation bias. Reconstruction-based multimodal models can place pressure on a nominally shared latent variable to encode modality-specific variation, particularly when private variation is large relative to shared signal (Daunhawer et al., 2022; Märtens & Yau, 2024). This is a concern for multimodal VAE models structurally similar to sRP-SSM. The MVAE (Wu & Goodman, 2018), for instance, also combines modality-specific evidence via a product-of-experts construction, trains on modality subsets so encoders remain usable when modalities are missing, and has been applied to joint neural-behavioral representation (Singh Alvarado et al., 2021), yet has been shown particularly susceptible to this bias (Märtens & Yau, 2024). RPMs such as sRP-SSM avoid this issue because the shared latent is not required to reconstruct either modality. Modality-specific variability can remain in the empirical marginal distributions, while recognition potentials only need to represent how each observation changes beliefs about the shared state. Thus, sRP-SSM obtains product-of-experts evidence fusion without requiring explicit shared/private latent decomposition or selective gradient blocking, as in MMVAE++ (Märtens & Yau, 2024). Appendix D.2 discusses the connection to MVAE, MMVAE++, and related reconstruction-based objectives in more detail.

5 Synthetic Experiments

We first ask whether our minibatch objective tracks the evidence lower bound (ELBO) it is designed to approximate. Using a synthetic timeseries model with both shared and modality-specific dynamical latents (Figure 2A; full data and evaluation details in Appendix E), we tracked the minibatch training objective alongside a high accuracy reference estimate of the RPM ELBO. Their trajectories were nearly identical across training (Pearson $r = 0.9985$). At the end of training, the values differed by 0.047 nats per timestep compared with an ELBO improvement of 6.54-nat-per-timestep during training (Figure 2B).

We next evaluate sRP-SSM's recovery of the ground truth dynamics (Figure 2C). In a high-SNR condition with low-amplitude private latents, sRP-SSM recovers dynamics with an order of magnitude lower error than CLIP-posthoc and also outperforms the two VAE models. Under low-SNR conditions, sRP-SSM maintains its dynamics recovery performance, CLIP-posthoc deteriorates slightly, and the VAEs deteriorate to chance performance. We next evaluate latent recovery and cross-modal retrieval together (Figure 2D-E). Varying the filter gain produces a trade-off between these two metrics. We use an oracle with exact observation evidence and true dynamics to trace the attainable tradeoffs between these two metrics within this filtering family, the upper right boundary of which is the Pareto frontier, where neither metric can improve without reducing the other. In the high-SNR condition we find that sRP-SSM, CLIP, and FOCAL variants sit near the Pareto frontier, while

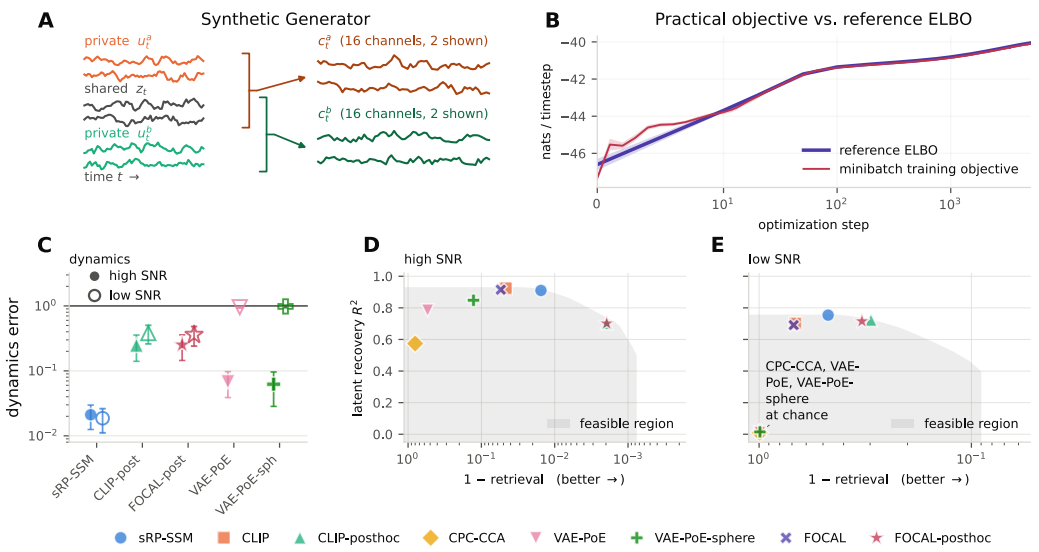

Figure 2: Synthetic experiments. **A:** The data generating process constructs two timeseries modalities by mixing shared and modality-specific latent processes. **B:** The minibatch-targeted objective closely tracks a high-quality reference ELBO estimate throughout optimization (Pearson $r = 0.9985$; the final discrepancy is 0.047 nats, compared with 6.54 nats of improvement during training). Curves show mean $\pm$ standard deviation for 5 seeds. **C:** sRP-SSM achieves the lowest dynamics recovery error among the compared methods under both high- and low-SNR conditions. **D-E**: Varying a scalar filter gain parameter with oracle observation evidence and dynamics traces a feasible region of latent recovery and cross-modal retrieval accuracy for both high- (**D**) and low-SNR (**E**) conditions. sRP-SSM, CLIP, and FOCAL variants lie near the frontier.

contrastive predictive coding with post-hoc alignment (CPC-CCA) and the two VAEs do not (Figure 2D). In a low-SNR condition, only sRP-SSM, CLIP-posthoc, and FOCAL-posthoc remain near the frontier. CLIP and FOCAL drop off the frontier, while the remaining baselines fall to chance levels on both metrics (Figure 2E). sRP-SSM ablation experiments are presented in Appendix E.7.

6 Mouse Neural-Behavioral Experiments

Having established sRP-SSM can accurately learn the shared dynamical features in simulation, we now test whether these gains persist in real neural-behavioral recordings. We use six 30-minute sessions from one mouse, with synchronized 32-channel LFP recordings across 13 regions and camera-derived 3D egocentric keypoints (Figure 3A). We compare sRP-SSM against CLIP and CLIP-posthoc, the strongest baselines from the synthetic experiments, using cross-modal retrieval, temporal retrieval, and bidirectional decoding. Full dataset and evaluation details are contained in Appendix F.

We find that sRP-SSM achieves better cross-modal retrieval performance than CLIP variants, with average LFP-to-keypoint and keypoint-to-LFP accuracies of 35% (chance: 0.4%), followed by CLIP-posthoc (33%), and then CLIP (14%; Figure 3B). As in the synthetic experiments, giving CLIP the ability to accumulate beliefs over time via an assumed density filter more than doubles its cross-modal retrieval accuracies. Even accounting for these differences in simultaneous retrieval by normalizing by chance and zero-horizon prediction, we find that sRP-SSM retains comparatively more of its cross-modal retrieval performance out to a horizon of approximately three seconds, at which point the methods converge (Figure 3C). This indicates that sRP-SSM representations retain cross-modal predictive information over longer timescales, beyond their slight advantage at simultaneous retrieval.

When decoding egocentric 3D keypoints from LFP-derived latents, we find that sRP-SSM outperforms the CLIP models for eight of nine individual keypoints (Figure 3D). Averaging over keypoints,

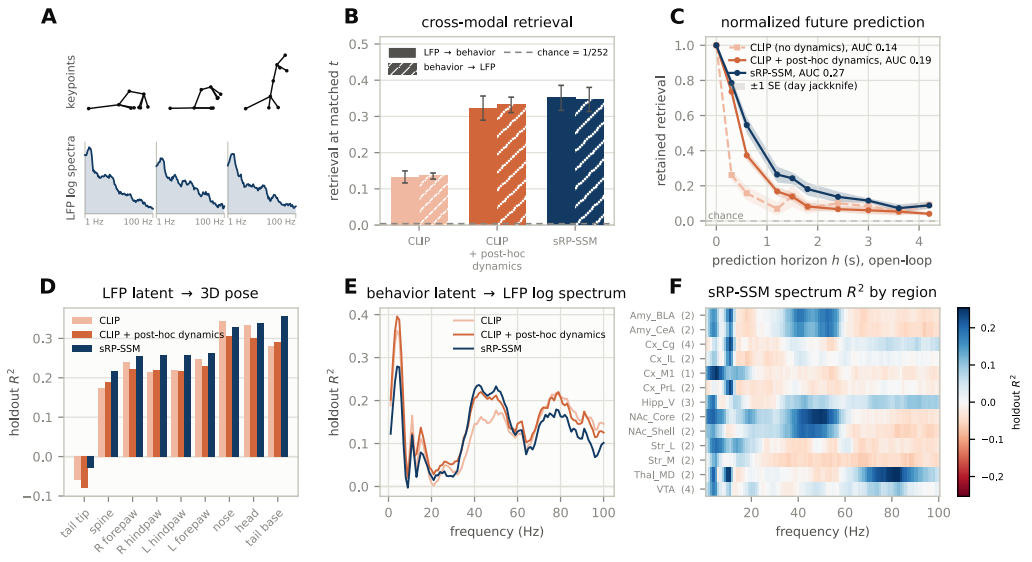

Figure 3: Quantitative neural-behavioral results. **A:** Example mouse poses and LFP log spectra. **B:** Cross-modal retrieval metrics for CLIP, CLIP with post-hoc dynamics, and sRP-SSM. Whiskers denote leave-one-session-out jackknife standard errors. **C:** Normalized and chance-corrected cross-modal retrieval by horizon. Shaded bands denote leave-one-session-out jackknife standard errors. **D:** Decoding egocentric keypoint positions from LFPs. **E:** Decoding LFP log spectrum from egocentric keypoints. **F:** sRP-SSM-decoded region-specific LFP log spectra from egocentric keypoints. Appendix F contains full region names.

this translates to an average R^2 of 0.249 for sRP-SSM, 0.222 for CLIP, and 0.211 for CLIP-posthoc. We then tested decoding in the other direction, from keypoints to LFPs, summarized as log spectra. In this direction, sRP-SSM falls behind CLIP at $R^2 = 0.124$ versus 0.143 while CLIP-posthoc achieves $R^2 = 0.166$ (Figure 3E). We speculate that CLIP preserves behavioral detail useful for reconstructing LFP spectra even though sRP-SSM retains more retrieval accuracy at longer horizons.

For decoding among all models (Figure 3E), we observe interesting spectral patterns, with delta (0-4 Hz) and gamma (30-100 Hz) bands showing the highest predictability. Region-specific sRP-SSM decoding reveals several notable patterns. Many regions show distinct predictability (R^2) peaks at delta (1-4 Hz) and theta (6-12 Hz) frequencies, nucleus accumbens and the amygdala show low gamma (40-60 Hz) peaks, and the medial dorsal thalamus shows a large peak at high gamma (70-100 Hz; Figure 3F).

To illustrate what these decoding numbers mean for a discrete, interpretable behavior, we examine rearing, in which a mouse lifts its forelimbs off the ground and raises its head. An example rearing bout along with sRP-SSM-decoded poses from the LFP is shown in Figure 4A. The height of the animal is underestimated in the decoding, but shows the same forelimb lifting over much of the bout. Collecting 105 rearing bouts from the validation set and sorting by duration, we see that this underestimation of body height is a general phenomenon, despite decoded body height having a substantial correlation with true height (Pearson $r = 0.41$) and reliably distinguishing rearing from non-rearing timepoints (AUC $= 0.84$). Lastly, we look at the average LFP spectral signatures of rearing, aligned to both the onset and offset of the bout, finding distinct spectrotemporal signatures around both onset and offset (Figure 4C). Using sRP-SSM to decode keypoints to LFP spectra, we find very similar signatures for both onsets (Pearson $r = 0.81$ for -1.5 to 1.5 s, 1-100 Hz) and offsets (Pearson $r = 0.83$). Together, these results show sRP-SSM predicts rearing bouts from LFPs and a spectral neural signature of rearing from keypoints, concrete instances of the bidirectional decoding structure quantified in Figure 3.

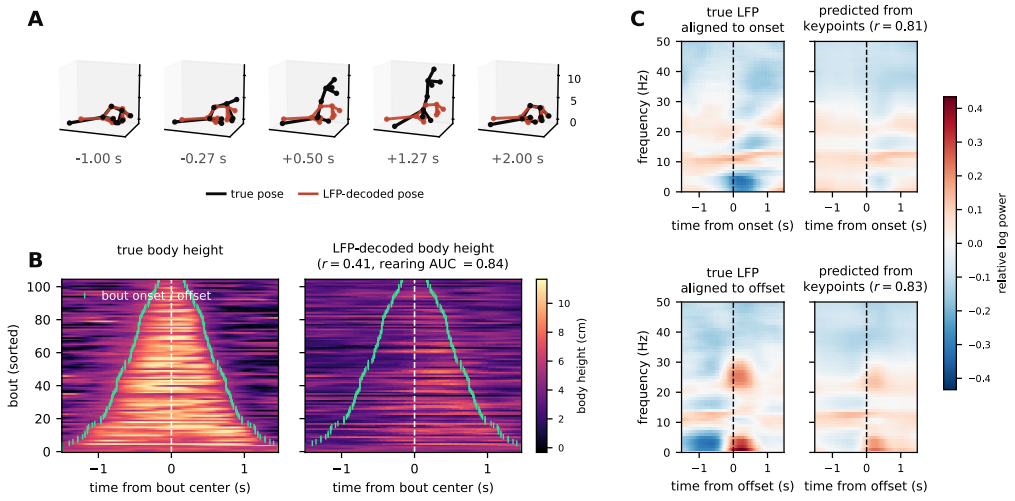

Figure 4: Mouse rearing behavior. **A:** Stills from an example rearing bout (black) and LFP-decoded keypoints from the same timepoints (red). **B:** At left, a collection of 105 validation set rearing bouts, sorted from shortest to longest, with color indicating body height. At right, the same bouts colored by LFP-decoded body height displays a similar pattern (Pearson $r = 0.41$), but attenuated compared to the true body heights. **C:** LFP log average spectrum aligned to rearing bout onset (first column) and offset (third column), showing distinct spectrotemporal patterns. The keypoint-decoded LFP spectra (second and fourth columns) largely reconstruct the patterns (onset-aligned Pearson $r = 0.81$, 1-100 Hz; offset-aligned Pearson $r = 0.83$).

7 Discussion and Limitations

We propose sRP-SSM, a recognition-parameterized model that combines contrastive-style representation learning with a probabilistically grounded dynamical objective to learn the shared latent dynamical structure among multimodal timeseries. Comparing sRP-SSM to the most competitive baseline model, CLIP with post-hoc fitted dynamics, helps us understand the benefits of jointly learning dynamics and representations. We find that static contrastive alignment can already recover much of the shared instantaneous structure and post-hoc dynamics can exploit temporal smoothness and effectively accumulate evidence over time, further improving cross-modal alignment. The advantage of jointly learning dynamics with representations is its ability to preserve signal that is weakly informative for cross-modal matching, but consistently predictive over time. Conversely, it can ignore nuisance variation that is strongly informative for cross-modal matching, but uninformative about dynamics.

sRP-SSM is similar to contrastive models in that it avoids specifying an observation model, while it is similar to probabilistic generative models in that it explicitly represents uncertainty and optimizes an approximation to the ELBO. This choice is especially useful in a neuroscience context, where the shared information between neural and behavioral data is the object of study, while generative models spend capacity on reconstructing high-dimensional signals.

Several limitations remain in the proposed method. First, our neural-behavioral data results are from a single animal and so cannot establish cross-animal generalization. Additionally, we find that CLIP-posthoc is competitive in some regimes, so we cannot uniformly recommend joint learning of dynamics for all downstream uses. The dynamics and filtering in this work are fairly simple and may not capture highly nonlinear or switching behavior well. Lastly, the learned associations between modalities reflect simple association and predictability, not necessarily causal coupling.

AI USE STATEMENT

We used generative AI tools to discuss aspects of the model and its mathematical claims, discuss, suggest, and evaluate experimental designs, and assist with code. We also used them to help identify relevant literature, find typos, edit the manuscript for clarity and concision, and write an initial draft of the introduction section from a provided outline. The authors reviewed the AI-assisted work, checked mathematical and empirical claims against their derivations and results, and take responsibility for the final paper.

ETHICS STATEMENT

All animal care and experimental procedures were conducted according to IACUC-approved protocols at Anonymous University.

REPRODUCIBILITY STATEMENT

Appendix D contains the full arguments summarized in Section 4. Appendix E contains details of the synthetic data experiments. Appendix F describes mouse data preprocessing. The mouse keypoint and LFP data from Section 6 will be uploaded to a public data depository upon publication. Full code to generate synthetic data, preprocess real data, fit models, and produce figures will be released upon publication.

REFERENCES

Laurence Aitchison and Stoil Ganev. InfoNCE is variational inference in a recognition parameterised model. *Transactions on Machine Learning Research*, 2024.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Imant Daunhawer, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E. Vogt. On the limitations of multimodal vaes. In *International Conference on Learning Representations (ICLR)*, 2022.

Timothy W. Dunn, Jesse D. Marshall, Kyle S. Severson, Diego E. Aldarondo, David G. C. Hildebrand, Selmaan N. Chettih, William L. Wang, Amanda J. Gellis, David E. Carlson, Dmitriy Aronov, Winrich A. Freiwald, Fan Wang, and Bence P. Ölveczky. Geometric deep learning enables 3d kinematic profiling across species and environments. *Nature Methods*, 18:564–573, 2021. doi: 10.1038/s41592-021-01106-6.

Samo Hromadka, Kai Biegun, Lior Fox, James Heald, and Maneesh Sahani. Maximum-likelihood learning of latent dynamics without reconstruction. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Cole Hurwitz, Akash Srivastava, Kai Xu, Justin Jude, Matthew Perich, Lee Miller, and Matthias H. Hennig. Targeted neural dynamical modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yiqi Jiang, Kaiwen Sheng, Yujia Gao, Estefany Kelly Buchanan, Yu Shikano, Seung Je Woo, Yixiu Zhao, Tony Hyun Kim, Fatih Dinc, Scott Linderman, and Mark J. Schnitzer. Extracting task-relevant preserved dynamics from contrastive aligned neural recordings. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5376.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. FOCAL: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kaspar Märtens and Christopher Yau. Disentangling shared and private latent factors in multimodal variational autoencoders. In *Proceedings of the 18th Machine Learning in Computational Biology Meeting*, volume 240 of *Proceedings of Machine Learning Research*, pp. 60–75. PMLR, 2024.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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*, 24:140–149, 2021. doi: 10.1038/s41593-020-00733-0.

Omid G. Sani, Bijan Pesaran, and Maryam M. Shanechi. Dissociative and prioritized modeling of behaviorally relevant neural dynamics using recurrent neural networks. *Nature Neuroscience*, 27: 2033–2045, 2024. doi: 10.1038/s41593-024-01731-2.

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

Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jonnathan Singh Alvarado, Jack Goffinet, Valerie Michael, William A. Liberti, Jordan Hatfield, Timothy Gardner, John Pearson, and Richard Mooney. Neural dynamics underlying birdsong practice and performance. *Nature*, 599(7886):635–639, 2021. doi: 10.1038/s41586-021-04004-1. URL `https://doi.org/10.1038/s41586-021-04004-1`.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal ELBO. In *International Conference on Learning Representations*, 2021.

David J Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9): 1055–1096, 1982.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

William I. Walker, Hugo Soulat, Changmin Yu, and Maneesh Sahani. Unsupervised representation learning with recognition-parametrised probabilistic models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4209–4230, 2023.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

A Related Work

Neural-behavioral modeling It is well known that neural dynamics contain both components related to behavior and others that are not clearly related. Several methods have been proposed to isolate the shared dynamics. PSID uses linear-Gaussian dynamical systems and preferentially identifies behavior-relevant subspaces (Sani et al., 2021). TNDM uses shared and private latents in a generative timeseries model (Hurwitz et al., 2021). DPAD prioritizes behaviorally relevant nonlinear dynamics, and has been applied to both spikes and LFPs (Sani et al., 2024). These methods generally separate shared and private dynamics by representing both. In this work, we make the shared neural-behavioral process itself the object of study and do not explicitly model modality-private variation.

Multimodal latent variable models There is a substantial literature on multimodal VAEs for jointly learning and performing inference in multimodal settings (Shi et al., 2019; Sutter et al., 2021). For example, Wu & Goodman (2018) introduce modality-specific encoders and combine evidence across modalities using a product of experts (PoE) construction. However, because a generative model must account for the observed data, substantial private variation must also be represented. As shown by Märtens & Yau (2024), multimodal VAE performance degrades under large private variability regimes.

Contrastive multiview learning Contrastive learning methods target predicting matching data against mismatched negatives, a learning process which targets shared information (van den Oord et al., 2018; Radford et al., 2021; Schneider et al., 2023). FOCAL combines contrastive learning with multimodal timeseries and shared and private latents (Liu et al., 2023). Liang et al. (2023) argue that shared information is not necessarily the right target for downstream tasks, proposing FactorCL to segregate the embedding space. CEBRA and CANDY are particularly relevant to a neuroscience audience (Schneider et al., 2023; Jiang et al., 2025). Temporal structure in these methods is primarily learned through contrastive matching, with the exception of CANDY, which jointly learns a linear dynamical system in the embedding space. Our model learns a latent dynamical model with an explicit transition model and filtering distribution.

Recognition-parameterized models RPMs replace the difficult generative likelihood with recognition potentials together with the empirical marginal data distribution, offering principled probabilistic inference without generation (Walker et al., 2023). Aitchison & Ganev (2024) show that InfoNCE admits an RPM variational inference interpretation, a technique we use in deriving the sRP-SSM objective. The recent work of Hromadka et al. (2026) uses this connection to build a latent dynamical model without data reconstruction. We extend this line of work to multimodal time series by letting modality-specific recognition factors provide evidence about the same evolving latent state, using PoE constructions to obtain scalable cross-modal contrastive terms, and retaining explicit spherical dynamics and filtering.

B von Mises–Fisher identities

Let ν denote the uniform probability measure on $\mathbb{S}^{d-1}$. We define the von Mises–Fisher (vMF) normalizer $C_d(\kappa)$ relative to ν, so that

$$\frac{d\,\mathrm{vMF}(\mu,\kappa)}{d\nu}(z) = C_d(\kappa)\exp(\kappa\mu^\top z), \qquad C_d(\kappa) = \left[\mathbb{E}_{z\sim\nu}\exp(\kappa\mu^\top z)\right]^{-1}. \tag{13}$$

Here $\mu \in \mathbb{S}^{d-1}$ and $\kappa \geq 0$. We also write $\mathrm{vMF}(z; e)$ for the density with natural parameter $e = \kappa\mu$. Products of vMF potentials remain proportional to a vMF density, with natural parameters that add:

$$\prod_i \mathrm{vMF}(z; e_i) \propto \mathrm{vMF}\left(z; \sum_i e_i\right). \tag{14}$$

The posterior moment and KL divergence to uniform are

$$\mathbb{E}_{\mathrm{vMF}(e)}[z] = \mathrm{A}_d(\|e\|)\frac{e}{\|e\|}, \qquad \mathrm{A}_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}, \tag{15}$$

$$\mathrm{KL}(\mathrm{vMF}(\mu,\kappa)\|\nu) = \kappa\ \mathrm{A}_d(\kappa) + \log C_d(\kappa). \tag{16}$$

The moment is zero when $e = 0$. Lastly, for a vMF belief with natural parameter λ and a recognition factor with natural parameter e,

$$\mathbb{E}_{z\sim\mathrm{vMF}(\lambda)}[\mathrm{vMF}(z;e)] = \frac{C_d(\|\lambda\|)\, C_d(\|e\|)}{C_d(\|\lambda + e\|)}. \tag{17}$$

C Alternative filtering gains

By default, we set the propagation gain c in equation 6 to a fixed constant during training. We also experimented with tying it directly to the learned transition model. If z_{t-1} has mean $\bar{z}_{t-1}$, the vMF transition in equation 4 implies the exact first-moment propagation

$$\mathbb{E}_{q_{t-1}}[\mathbb{E}[z_t \mid z_{t-1}]] = \mathrm{A}_d(\kappa_w)\, R\, \bar{z}_{t-1}. \tag{18}$$

This motivates the tied gain

$$c = \mathrm{A}_d(\kappa_w), \tag{19}$$

which connects the filter's temporal persistence directly to the learned transition concentration. Because equation 6 propagates natural parameters rather than moments, however, this tied gain does not in general reproduce equation 18 exactly.

A moment-matched assumed-density prediction instead propagates the previous vMF belief through the transition and matches the resulting first moment with a vMF distribution. If $\alpha_{t-1} = \|\lambda_{t-1}\|$, the predicted concentration κ_t^- is defined implicitly by

$$\mathrm{A}_d(\kappa_t^-) = \mathrm{A}_d(\kappa_w)\, \mathrm{A}_d(\alpha_{t-1}), \tag{20}$$

giving the predicted natural parameter and state-dependent gain

$$\lambda_t^- = c_t R \lambda_{t-1}, \qquad c_t = \frac{\kappa_t^-}{\alpha_{t-1}}. \tag{21}$$

When $\alpha_{t-1} = 0$, the predicted natural parameter is zero. The limit of c_t as $\alpha_t \to 0$ is $\mathrm{A}_d(\kappa_w)$, showing the tied gain is the low-confidence limit of the moment-matched gain. The observation evidence is then added to λ_t^- as in equation 6. In ablation experiments presented in Appendix E, the fixed gain performed similarly to the tied gain and this state-dependent gain, so we opted for the simple fixed-gain setting.

D Model Design Appendix

D.1 Latent geometry and first-order recognition

The choice of latent geometry decides whether a first-order recognition score as in Eq. 9 is well specified. Throughout we consider first-order critics. We call a log potential first-order in its target when the conditioning variable and target interact only through a linear statistic of the target. For example, the sRP-SSM critic is $\log f(z|x) = s_j(z) = e_j^\top z + b_j$ after removing z-only terms shared across candidates (Eq. 9), while the form used by CPC, $\log p(v \mid u) - \log p(v)$ must be affine in the candidate target v up to u-only terms. A first-order critic is exact if and only if the relevant log density is affine in the target once such shared terms are removed.

Proposition 1 (Target curvature and its removal by fixed radius)**.** (i) Candidate observations, as in equation 10. *Under the sRP-SSM joint (Eq. 5),* $\log p(x \mid z) - \log p_0(x) = \log f(z \mid x) - \log F(z)$, *and the z-only term* $\log F(z)$ *cancels across candidates. For an isotropic Gaussian recognition factor* $f(z \mid x) = \mathcal{N}(z; \mu(x), \tau(x)^{-1} I)$,

$$\log f(z \mid x) = \tau(x)\mu(x)^\top z - \tfrac{1}{2}\tau(x)\|z\|^2 + b(x), \tag{22}$$

so a first-order critic is exact on a Euclidean latent space if and only if $\tau(x)$ is constant across observations. On $\mathbb{S}^{d-1}$ the radial term is candidate-only and is absorbed into the normalizer. The resulting family is vMF, $\log f(z \mid x) = e(x)^\top z + \log C_d(\|e(x)\|)$, *which is first-order for arbitrary per-observation concentration* $\|e(x)\|$.

(ii) Candidate targets, as in CPC/CLIP. *Let (U, V) be jointly Gaussian with positive-definite joint covariance. After whitening and rotating into canonical-correlation coordinates with correlations ρ_k,*

$$\log \frac{p(v \mid u)}{p(v)} = \sum_k \frac{\rho_k}{1-\rho_k^2}\, u_k v_k - \frac{1}{2} \sum_k \frac{\rho_k^2}{1-\rho_k^2}\, v_k^2 + a(u), \tag{23}$$

and $I(U;V) = -\frac{1}{2}\sum_k \log(1-\rho_k^2) = \frac{1}{2}\log\det(I+M)$ with $M = \operatorname{diag}\left(\rho_k^2/(1-\rho_k^2)\right)$. Hence a first-order critic is exact if and only if they have zero mutual information, $I(U;V) = 0$. If instead $V \mid U \sim \mathrm{vMF}(\mu(u), \kappa(u))$ on $\mathbb{S}^{d-1}$ with uniform or vMF reference, the ratio is affine in v for every $\kappa(u)$.

Both cases expose the same underlying issue. Gaussian dependence introduces quadratic curvature when candidate targets are classified, while observation-dependent Gaussian precision introduces quadratic curvature in the fixed latent target when candidate observations are classified, as in the sRP-SSM. For isotropic factors this curvature is radial, and fixing the target norm makes it constant. Consequently, vMF factors can represent both nonzero dependence and per-observation confidence while remaining first-order in the latent state. Fixing the radius does not remove anisotropic curvature, however: writing $M = \alpha I + A$, with $\operatorname{tr} A = 0$, leaves the second-order Bingham term $z^\top A z$ on the sphere.

The transition prior provides a second instance of case (ii). Under equation 4, orthogonal R, and the uniform stationary law,

$$\log \frac{p(z_t \mid z_{t-1})}{\nu(z_t)} = \kappa_w \langle R z_{t-1}, z_t \rangle + \log C_d(\kappa_w) \tag{24}$$

is affine in z_t for every κ_w, while every $\kappa_w > 0$ defines a dependent stationary process. By contrast, a dependent stationary linear-Gaussian chain may be stable and contractive, but its transition density ratio necessarily contains a quadratic term in z_t. The distinction is therefore not stationarity or dynamical stability, but whether nonzero temporal dependence is compatible with a first-order density ratio.

Per-observation confidence is central to our model: $\kappa(x)$ controls each recognition factor's contribution to product-of-experts fusion and temporal filtering. An isotropic Gaussian recognition model with observation-dependent precision would require a quadratic latent statistic, whereas the vMF family permits variable confidence, additive natural-parameter fusion, and exactly normalized first-order recognition logits simultaneously. We therefore use spherical vMF factors as the default. This is not a claim that Euclidean RPMs are intrinsically misspecified. Exact Gaussian formulations retain the quadratic statistic $zz^\top$, but constitute a more complex critic and inference family.

D.2 RPMs avoid reconstruction-driven private-variation bias

Like sRP-SSM, multimodal variational autoencoders (VAEs) are probabilistic latent-variable models trained by maximizing an ELBO. Unlike sRP-SSM, however, they specify observation likelihoods and learn latent representations through reconstruction. Among these models, the MVAE of Wu & Goodman (2018) is particularly similar to sRP-SSM in that it combines modality-specific evidence about a shared latent variable using a product-of-experts (PoE) construction, optimizes modality-subset objectives so that individual encoders remain usable when modalities are missing, and has been applied to joint neural-behavioral representation learning (Singh Alvarado et al., 2021). However, reconstruction-based multimodal objectives exhibit important limitations. Daunhawer et al. (2022) show that reconstructing all modalities from an observed subset introduces an irreducible gap governed by the information in the omitted modalities that cannot be predicted from the observed ones. This gap therefore grows with modality-specific variation.

Märtens & Yau (2024) examine this gap's consequences for recovering shared latent structure when modality-specific variation is much stronger than shared variation. In this setting, reconstructing many private features can improve the ELBO more than recovering a weaker shared signal, causing nominally shared latent variables to absorb modality-specific structure. They find MVAE particularly susceptible and propose MMVAE++, which explicitly separates shared and private latent variables and prevents same-modality reconstruction gradients from updating the shared encoder.

RPMs in general, and sRP-SSM in particular, do not require the shared latent variable to reconstruct either modality. Private variability can instead remain in the empirical marginal distribution of each

observation modality, eliminating the reconstruction-driven incentive to encode it in the shared latent state. Recognition potentials are rewarded for representing how observations change beliefs about that shared state, rather than for preserving information needed to reconstruct the observations. Thus, sRP-SSM avoids the specific private-variation bias identified above without requiring gradients from within-modality reconstruction to be selectively blocked.

E Synthetic Experiments

E.1 Synthetic data-generating process

Here we describe the synthetic data used in Figure 2. We first construct three independent latent processes, one shared (z_t) and two private (p_t^m for modalities $m = 0, 1$), all defined on an 8-dimensional spherical latent space. These are stationary vMF chains on the sphere with $\kappa_w = 12$ and independently drawn rotational dynamics. For each modality, we first construct a noisy 16-dimensional core observation

$$c_t^m = \lambda M_m[z_t;\ a p_t^m] + \sigma \epsilon_t \ , \tag{25}$$

where M_m is a modality-specific orthogonal mixing matrix and $\sigma = 0.3$. The core is then deterministically padded to 64 dimensions by a fixed random linear projection of c_t^m. We consider two nuisance conditions, with nuisance amplitude $a = 1$ in the high-SNR condition and $a = 3$ in the low-SNR condition, with the shared scale λ adjusted so that the total observation variance remains comparable across conditions.

Training samples are generated online, so every minibatch is newly simulated. Each random seed redraws dynamics, modality-specific mixing and padding matrices, and model initialization. Methods evaluated under the same seed use matched generator instances and data.

E.2 Compared methods

sRP-SSM: sRP-SSM uses a single shared spherical vMF latent chain. Modality-specific encoders contribute recognition potentials that are combined via a product-of-experts (PoE) construction. We use fixed gain dynamical filtering with $c = 0.3$. The dynamics are trained jointly with the encoders. We use the analytic contrastive denominator and modality masking during training.

CLIP and CLIP-posthoc: CLIP uses the standard symmetric InfoNCE objective to align contemporaneous observations Radford et al. (2021). There is no dynamical model during CLIP training. However, after training, CLIP-posthoc uses the same frozen CLIP model and estimates the orthogonal dynamics matrix R and the filter gain from the fixed embeddings.

CPC-CCA: We use independent per-modality CPC (van den Oord et al., 2018) objectives with temporal prediction horizons of 1, 2, and 3 timesteps. After training, we use canonical correlation analysis (CCA) to align the two modalities.

VAE-PoE and VAE-PoE-Sphere: We use product-of-experts VAEs in the style of MVAE (Wu & Goodman, 2018), but with dynamical priors. VAE-PoE uses a standard linear-Gaussian dynamical prior, while VAE-PoE-Sphere uses the same spherical vMF chain prior used by sRP-SSM. Therefore VAE-PoE-Sphere provides a closely matched generative counterpart to sRP-SSM.

FOCAL and FOCAL-posthoc FOCAL (Liu et al., 2023) learns separate shared and modality-specific embeddings for timeseries using contrastive objectives. We adapt FOCAL to our synthetic observations using simple label-preserving augmentations (multiplicative scaling, Gaussian jitter, and random feature masking) and evaluate only its shared representation. FOCAL contains no explicit dynamical model during training and FOCAL-posthoc fits the same orthogonal dynamics and fixed-gain filter used for CLIP-posthoc to fixed embeddings.

E.3 Matched Architecture and Training Protocol

All methods use the same encoder architecture, a 3-layer MLP with width 256 and latent dimension 8. A window length of $W = 16$ is used with a burn-in of 4 samples to initialize the filters. Each

batch consists of 256 windows. All methods are optimized for 5000 iterations using Adam with a fixed learning rate of 10^{-3}. Three seeds determining generated data and model initialization are run for each method.

Note that the encoder architectures are matched, but not the total parameter budget. The VAEs require decoders, so they use 600k parameters compared to sRP-SSM's 300k parameters. Also note that sRP-SSM uses a fixed filter gain of $c = 0.3$, the VAEs learn their gains, CLIP has no gain, and CLIP-posthoc estimates one after training.

E.4 Evaluation Metrics

Cross-modal retrieval: We use an item-stratified cross-modal retrieval metric. For each scored within-window timepoint, an embedding from one modality is used to retrieve the corresponding observation from the other modality among all items in the batch at that same timepoint. Retrieval is evaluated symmetrically in both modality directions and averaged. With a batch of n evaluation windows, chance performance is $1/n$.

Latent recovery R^2: We use the following latent recovery metric:

$$R^2 = 1 - \frac{\min_{s,O} \|s\hat{Z}O - Z\|^2}{\|Z\|^2} , \tag{26}$$

with orthogonal O and scalar s accounting for the spherical models' rotational indeterminacy and the Euclidean models' scale indeterminacy, respectively. Here, Z is the ground truth latents and $\hat{Z}$ is the trained model's posterior mean latent representations.

Dynamics recovery: To quantify recovery of the latent dynamics, we compare the invariant two-dimensional rotation subspaces of the learned transition matrix with those of the ground-truth rotation. Learned and true rotation planes are matched by the assignment that maximizes subspace alignment, with nearly degenerate true rotation planes grouped and compared jointly. We report the mean principal-angle error after accounting for the model's global orthogonal latent alignment, and normalize this error so that 0 corresponds to the oracle error achievable at the model's filtering gain and 1 corresponds to the expected error from an unrelated dynamical frame. Lower values therefore indicate more accurate recovery of the ground-truth rotational dynamics.

E.5 Oracle performance and feasible region

For the data generating process in these synthetic experiments, the observation mixing is known so we can calculate the instantaneous evidence for the shared latent analytically. The oracle therefore uses exact evidence, the true latent dynamics, and the same fixed-gain filtering family used for evaluation. Because the gain that optimizes retrieval does not optimize latent recovery, the oracle defines a trade-off frontier rather than a single optimal operating point. We evaluate the oracle on a fine grid of gain values, producing a family of attainable retrieval/recovery tradeoffs. The shaded regions in Figure 2D-E are the downward-closed feasible regions.

E.6 Evaluation filter gain moves the sRP-SSM operating point

Figure 5 shows the effect of the filter evaluation gain c (Eq. 6) on latent recovery, retrieval, and dynamics error. Dynamics error and retrieval are generally improved with higher evaluation gains whereas latent recovery deteriorates with sufficiently high gains. Notably, these different operating points are achieved after model training, allowing the same model to closely track portions of the latent recovery/retrieval Pareto frontier.

E.7 Ablations

We present several model ablations and compare their performance to the reference sRP-SSM with fixed filter gain $c = 0.3$. We evaluate five changes to the reference model:

- `no_transition`: removing the transition term from the objective KL decomposition (second sum on RHS of Eq. 12)

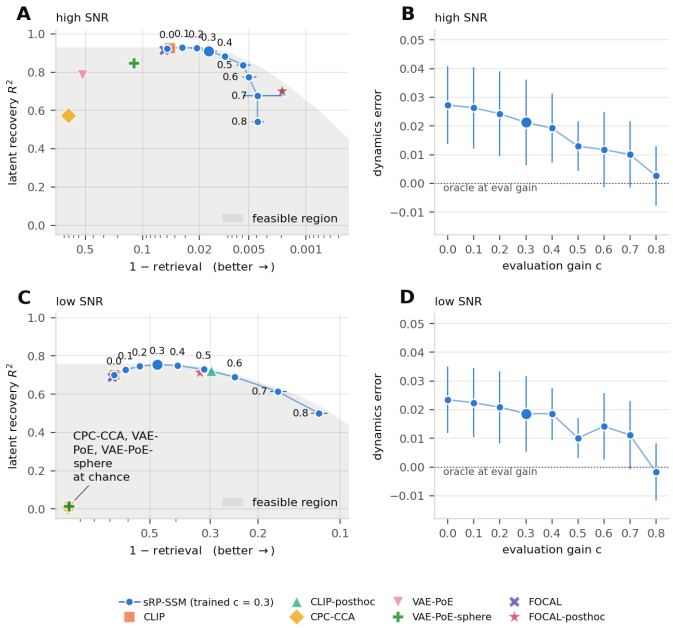

Figure 5: Evaluation filter gain moves the sRP-SSM operating point, trading off latent recovery and cross-modal retrieval. In both high- (**A**) and low-SNR (**C**) conditions evaluation gain changes sRP-SSM latent recovery–retrieval tradeoff: higher gains generally produce better retrieval at the expense of latent recovery. Dynamics error generally decreases as evaluation gain increases (**B,D**). All error bars denote sample standard deviation across 3 seeds.

- `plug-in_denom`: replace the log term in Eq. 10 with a plug-in approximation instead of the Jensen bound in Eq. 11
- `mc_denom`: replace the log term in Eq. 10 with a Monte-Carlo estimate instead of the Jensen bound in Eq. 11
- `tied_gain`: replace the fixed filter gain with a gain that is tied to the learned transition concentration, κ_w. See Appendix C for details.
- `iterated_adf`: replace the fixed gain filter with a 3-iteration assumed density filter. See Appendix C for details.

Each variant was trained for three seeds in both SNR conditions, following the protocol described above, and at two filter gain settings: the model's native gain and a fixed $c = 0.3$ setting. The second gain setting allows us to isolate the effect of the training filter on the learned representations.

The ablation results are shown in Table 1. All five variants produce nearly identical retrieval and latent recovery metrics in the fixed gain setting. The tied gain and iterated ADF variants move to different operating points along the Pareto frontier: tied gain produces better retrieval and worse latent recovery and iterated ADF results in the opposite trend. The dynamics error is also largely similar, with the tied gain filter resulting in slightly larger errors and the iterated ADF filter resulting in slightly smaller errors compared to the reference and other ablation settings.

Given the small difference between the reference model with the Jensen denominator (Eq. 11), `plug-in_denom`, and `mc_denom`, we believe the reference model provides a reasonable choice. It does not resort to sampling, unlike `mc_denom`, and maintains the lower bound property on the fixed candidate set, unlike `plug-in_denom`.

To understand why the `no_transition` ablation did not disrupt sRP-SSM's performance, we introduced an additional 8-dimensional latent component, matching the dimensionality of the existing latent, that is shared across modalities but non-dynamical in that it is drawn i.i.d. from the uniform distribution each timestep. This makes cross-modal agreement alone insufficient to recover the

Model and readout	Gain	Retrieval	Latent R^2	Dynamics error
High SNR				
sRP-SSM (reference)	0.30	.985 (.003)	.910 (.001)	.021 (.015)
`no_transition`	0.30	.984 (.003)	.910 (.001)	.020 (.015)
`plug-in_denom`	0.30	.984 (.004)	.910 (.001)	.022 (.017)
`mc_denom`	0.30	.984 (.003)	.910 (.001)	.021 (.014)
`tied_gain`, fixed readout	0.30	.984 (.003)	.910 (.001)	.025 (.018)
`iterated_adf`, fixed readout	0.30	.984 (.004)	.910 (.001)	.018 (.012)
`tied_gain`, native readout	0.42	.990 (.003)	.874 (.001)	.021 (.018)
`iterated_adf`, native readout	–	.972 (.001)	.929 (.000)	.019 (.014)
Low SNR				
sRP-SSM (reference)	0.30	.529 (.009)	.754 (.002)	.019 (.013)
`no_transition`	0.30	.529 (.004)	.754 (.002)	.016 (.010)
`plug-in_denom`	0.30	.527 (.007)	.754 (.002)	.018 (.013)
`mc_denom`	0.30	.526 (.011)	.754 (.002)	.018 (.013)
`tied_gain`, fixed readout	0.30	.528 (.010)	.754 (.002)	.023 (.015)
`iterated_adf`, fixed readout	0.30	.528 (.010)	.754 (.002)	.016 (.012)
`tied_gain`, native readout	0.42	.613 (.005)	.747 (.002)	.020 (.016)
`iterated_adf`, native readout	–	.415 (.011)	.737 (.002)	.018 (.012)

Table 1: Synthetic ablations. Entries are mean (standard deviation) over three seeds. Fixed readouts evaluate the trained models at $c = 0.3$ without retraining; native readouts use each model's own filtering rule. Models whose native gain is 0.3 are shown once. Lower dynamics error indicates better recovery.

	transition term	no transition term	CLIP
$c = 0.3$	$.911 \pm .001$	$.911 \pm .001$	$.305 \pm .468$
$c = 0$	$.924 \pm .001$	$.462 \pm .456$	$.311 \pm .475$

Table 2: Latent R^2 in the shared i.i.d. experiment (high SNR, mean $\pm$ std over 6 seeds). CLIP at $c = 0$ is plain CLIP; at $c = 0.3$ it is CLIP-posthoc with R fitted post hoc and the gain fixed at 0.3. sRP-SSM is scored at its training gain ($c = 0.3$ or $c = 0.0$).

shared dynamical state. We crossed the transition term (present or absent) with the training filter gain ($c = 0.3$ or $c = 0.0$, disabling the filter; Table 2). Models retaining either the filter or the transition term recovered the shared dynamical latents (latent $R^2 > 0.9$), while the model with both removed did not ($R^2 = 0.462$, $R^2 > 0.5$ for 3 of 6 seeds). CLIP and CLIP-posthoc, which both lack dynamical cues during representation learning, likewise failed ($R^2 \approx 0.31$, $R^2 > 0.5$ for 2 of 6 seeds each). These results suggest that the transition objective term and backpropagation through the recurrent filter play largely redundant roles in the sRP-SSM; they both produce an inductive bias in representation learning towards temporal structure.

E.8 ELBO ACCOUNTING DIAGNOSTIC

This diagnostic, presented in Figure 2B, is designed to evaluate whether the minibatch objective optimized by sRP-SSM tracks the ELBO it is designed to approximate.

Densities are evaluated on the noisy 16-dimensional core rather than the full 64-dimensional observations. The additional 48 observation dimensions are deterministic linear functions of the noisy core, so the full 64-dimensional observation distribution is singular with respect to Lebesgue measure and does not admit an ordinary density. We therefore express the objective as densities over the same noisy core variable and with respect to the same base measure.

We report $\mathcal{J}$, the minibatch objective optimized during training, alongside a reference ELBO estimate computed using a fixed bank of 10^5 stationary observations to estimate the aggregate recognition density. Changing the reference bank size from 10^4 to 10^6 changes the final estimate by only 0.005 nats.

Region	Abbreviation	Channels
Basolateral amygdala	Amy_BLA	2
Central amygdala	Amy_CeA	2
Cingulate cortex	Cx_Cg	4
Infralimbic cortex	Cx_IL	2
Primary motor cortex	Cx_M1	2
Prelimbic cortex	Cx_PrL	2
Ventral hippocampus	Hipp_V	3
Nucleus accumbens core	NAc_Core	2
Nucleus accumbens shell	NAc_Shell	2
Lateral striatum	Str_L	2
Medial striatum	Str_M	2
Medial dorsal thalamus	Thal_MD	3
Ventral tegmental area	VTA	4

Table 3: Recorded LFP regions

F Neural-Behavioral Experiment Details

F.1 Recording setup

We use a dataset of one mouse recorded across 6 30-minute sessions, totaling about 3 hours. LFPs and 6-view videos are captured simultaneously with a sync trace recorded alongside the LFPs to synchronize the two modalities. LFPs consist of 32 channels across 13 brain structures, listed in Table 3 captured at 1 kHz. Videos are captured at 30 Hz.

F.2 3D pose estimation

We use DANNCE, a markerless 3D pose-estimation framework that combines calibrated multiview images within a volumetric convolutional neural network, to estimate 3D keypoints for each video frame (Dunn et al., 2021). We estimated nine anatomical landmarks: the nose, head, spine, tail base, tail tip, left and right forepaws, and left and right hindpaws.

We manually annotated these 9 keypoints for 1,485 unique timepoints across seven video sets, including the six-session video set and six other video sets from mice with and without LFP headcaps. For each video set, 30 annotated timepoints are held out for validation. A center-of-mass (COM) network is first trained to localize the animal and center the 3D prediction volume. The DANNCE pose network is then trained on the remaining manually annotated landmarks with a batch size of 2 and a learning rate of 10^{-4}. We use the `compressed_dannce` architecture, a channel-compressed version of the original DANNCE encoder–decoder, and train the model for 950 epochs. DANNCE predictions are generated for each behavioral recording at 30 frames per second, producing a nine-keypoint 3D pose trajectory for every video frame.

F.3 Pose-estimation validation

We evaluate DANNCE pose-estimation accuracy on manually annotated held-out validation frames. For each anatomical landmark, predicted 3D coordinates are compared with the corresponding hand-labeled coordinates.

For keypoint k at validation frame t, let $\mathbf{p}_{t,k} \in \mathbb{R}^3$ denote the manually annotated 3D position and $\hat{\mathbf{p}}_{t,k}$ the corresponding DANNCE prediction. We first compute the 3D Euclidean prediction error as

$$e_{t,k} = \left\| \hat{\mathbf{p}}_{t,k} - \mathbf{p}_{t,k} \right\|_2 \ . \tag{27}$$

We report the mean $\pm$ standard deviation and the median 3D prediction error for each anatomical landmark.

Keypoint	Mean ± SD 3D Error (mm)	Median 3D Error (mm)
Nose	1.64 ± 0.83	1.53
Head	2.17 ± 1.17	1.96
Spine	4.61 ± 3.04	3.66
Tail Base	1.52 ± 0.77	1.40
Tail Tip	3.98 ± 4.53	2.22
Left Forepaw	2.00 ± 2.33	1.33
Right Forepaw	1.57 ± 0.99	1.31
Left Hindpaw	1.62 ± 0.86	1.41
Right Hindpaw	1.59 ± 1.04	1.34

Table 4: DANNCE 3D pose-estimation errors.

DANNCE predictions closely match the manually annotated poses across all nine landmarks. Mean 3D prediction error is approximately 1.6–2.1 mm for the nose, head, tail base, and paws. The largest errors occur for the spine and tail tip, with mean errors of 4.61 mm and 3.98 mm, respectively. For reference, the average annotated nose–head distance is 25.1 mm, while the average head–tail base distance is 58.3 mm.

F.4 LFP PREPROCESSING

We first detect contaminated channels by computing the mean positive excess (in decibels) of its Welch power spectrum over a ± 5 Hz running median of itself within 1–250 Hz. Channels whose score lies more than 5 robust (median/MAD) standard deviations above the array median are considered contaminated. We perform common median referencing using only the uncontaminated channels, but all channels are retained in the data. Next, we remove electrical line noise at 60 Hz and its harmonics using sliding-window multitaper harmonic regression (Thomson, 1982). The frequency of each line is estimated in 1 s windows with a 0.25 s step and 5 DPSS tapers ($NW = 3$). Then the line's sinusoid is subtracted only where Thomson's F-test is significant ($p < 0.05$, corrected for the frequency search), and the fits are combined across windows via Hann windows. Lastly, we bandpass filter each channel from 0.5 to 250 Hz with a fourth-order zero-phase Butterworth filter and perform robust median/MAD normalization for each channel independently.

F.5 KEYPOINT PREPROCESSING

For each session, we first perform floor alignment so that the vertical has a physical interpretation. To do this, we first fit the arena floor from paw locations and rigidly rotate so that the floor coincides with the $z = 0$ plane. We then center each pose on the torso (center of mass of tail base, spine, and head keypoints) and rotate about the vertical axis so that the tail base to head defines the forward direction. Importantly, the vertical direction is not altered by this egocentric transformation. The 3D keypoints and velocities for each of the nine keypoints are retained, giving 54 behavioral dimensions. The 54 dimensions are standardized before being passed to the encoder.

F.6 WINDOWING AND TRAIN/VALIDATION SPLITS

Each model timestep corresponds to 300 ms comprising 300 LFP samples and 9 behavioral frames. Consecutive timesteps form training chains. The middle 20% of timesteps (6 minutes) from every session form the validation split while the first and last 40% of timesteps (the remaining 24 minutes) form the training split. Items are constrained to not cross split boundaries. Training uses ± 10 ms LFP jitter. Validation does not.

F.7 MODELS AND TRAINING CONFIGURATION

We compare three models: CLIP, CLIP + post-hoc dynamics, and sRP-SSM. All models use the matched temporal convolutional network encoders and a 16-dimensional spherical latent space. The LFP encoder operates on 300-ms (300 sample) multichannel windows, while the behavior encoder operates over 9-frame position and velocity windows. Both use temporal convolutional networks followed by temporal pooling and projection into the latent space.

For sRP-SSM, we use a window size of $W = 10$ consecutive 300-ms timepoints to form a batch item. Burn-in is fixed at 4 timepoints to warm up the filter, which has a fixed training gain of 0.3, corresponding to the default value in the synthetic experiments. Batch size was tuned per-method. sRP-SSM uses a batch size of 1024 while CLIP uses a batch size of 2048. The prior term was linearly ramped from 0 to 1 over the first 8,000 training iterations for sRP-SSM. Both models were optimized with Adam using a fixed learning rate of 10^{-3} for 24k iterations.

CLIP uses symmetric InfoNCE and no temporal dynamics during training. Note that a CLIP batch of size B contains B 300-ms timepoints, whereas an sRP-SSM batch of size B contains BW timepoints. CLIP-posthoc is the same trained CLIP model, but with dynamics fit after the encoder on the training data.

F.8 Continuous held-out evaluation protocol

Our evaluations use a single continuous chain from each 6-minute validation portion per session instead of repeatedly restarting the filter. A common burn-in is removed from every model so that filtering history cannot change which observations are scored. Therefore all methods are evaluated on identical timestamps. CLIP-posthoc and sRP-SSM are evaluated at a common filter gain fixed at $c = 0.7$ so that differences between them are directly comparable. The value was chosen in advance to sit in the range that training-split post-hoc dynamics fits land (about 0.64–0.76). For this evaluation gain, we use a burn-in of 16 timesteps (4.8 s) to warm up the filters.

F.9 Cross-modal retrieval and future prediction

For the instantaneous retrieval evaluation presented in Figure 3B, candidate sets contain 252 held-out timestamps, balanced equally across the six recording sessions. Candidates within a recording are temporally separated by at least 6 s. Chance performance is therefore $1/252 \approx 0.004$. Results are presented for both decoding directions.

For the future prediction presented in Figure 3C, we first infer a latent at time t, propagate the learned dynamics to $t + h$, and retrieve the corresponding future representation. CLIP is assumed to have static latents, since it is not a dynamical model. We evaluate horizons of 0, 0.3, 0.6, 1.5, 3, and 6 s. The plotted accuracy is corrected for both chance and instantaneous retrieval accuracy:

$$\rho_h = \frac{A_h - A_{\text{chance}}}{A_0 - A_{\text{chance}}}\,, \tag{28}$$

where A_h is the accuracy at horizon h (averaged over LFP $\to$ behavior and behavior $\to$ LFP directions) and A_{chance} is chance performance.

F.10 Cross-modal decoding

For each decoding direction, we fit distance-weighted k-NN decoders with $k = 25$ on the training data and evaluate on the held-out continuous evaluation protocol described in Section F.8. We report hold-out R^2 metrics, evaluated relative to the training-set target mean.

For LFP $\to$ pose decoding presented in Figure 3D, we decode the full 300-ms, 9-frame egocentric keypoint 3D pose window from the LFP latent. Only keypoint locations are predicted, not keypoint velocities. For the behavior $\to$ LFP spectrum in Figure 3E, we decode the LFP 1–100 Hz log spectrum from the behavior latent. The spectral estimator for this prediction uses 900 ms context centered on the current 300-ms step. We estimate power spectral density for each channel at 1-Hz resolution using a multitaper estimator with $NW = 3$ and five tapers. The average is taken across channels before being log-transformed. The same procedure, but for region-specific spectra, is presented in Figure 3F. Note that region-wise R^2 values should not be compared numerically with the whole-brain R^2 because each region-specific spectrum averages over fewer channels and therefore contains more channel-specific noise.

F.11 Rearing analysis

Rearing definition: We define rearing bouts using 3D keypoints in the world-frame after floor alignment. A frame is considered rearing when both forepaws are > 2 cm above the floor and the

nose is > 5 cm above the floor, after smoothing with a 5-frame (167 ms) box filter. We close gaps of five frames or fewer and discard bouts shorter than 10 frames (333 ms).

LFP-derived pose and rearing readout: In Figure 4A-B, we show decoded body height (hind-paw to nose Δz) from LFPs. This uses the k-NN decoder described in Section F.10, not a separately trained LFP to height predictor, which could be better optimized for this task. Separately, we fit a two-feature logistic regression model to detect rearing behavior from pose, specifically using nose and forepaw height. This model was trained on true poses from the training split and applied on decoded poses from the validation split and evaluated using AUC on validation data.

Event-aligned analyses To decode quantities at finer resolutions than our model timestep (300 ms), we decode quantities using overlapping decoding grids. For the analyses in Figure 4, we decode using three overlapping grids with a stride of 100 ms. Predictions are then averaged among all three grid tilings. For the spectra in Figure 4C, we aligned true and behavior-predicted LFP spectra to rearing onsets and offsets, averaged across all detected rearing bouts, and subtracted the held-out mean so that the maps show event-locked spectral deviations.