Appendix

Comparing the MMGPVAE to competing approaches

We evaluated our MMGPVAE model alongside both Targeted Neural Dynamical Modeling (TNDM) (Hurwitz et al., 2021) and Preferential Subspace Identification (PSID) (Sani et al., 2021) using our synthetic data of a rotating and scaling MNIST digit '3' alongside Poisson neural rates. In Figure 1, we show the reconstruction of both the behavioral modality, the MNIST digit '3', and Poisson neural rates for each model. In addition to these, we also provide the error on both data modalities on held-out trials. We find that, in this setting, the MMGPVAE is able to better reconstruct data from both modalities.

To compare these models in a real-world setting, we implement PSID, TNDM, and MMGPVAE in a simple neuroscience example where all models are capable of recovering behavioral trajectories as well as neural rates. Since these competing approaches are not well-suited to complex behavioral modalities, we focus on a simpler behavioral task here. We use the primate reaching data used in the TNDM paper (Hurwitz et al., 2021). Here, the 2D reaching task is simple, and so linear dynamics as well as our deep neural network can reconstruct both modalities well. The results from this can be seen in Figure 2. While both TNDM and PSID can provide only two subspaces in which only one can separate the 8 reaching positions (relevant subspace), MM-GPVAE can provide visualization for 3 subspaces, 2 modality specific, and 1 shared, and we can see a clear separation with both the behavioral-only subspace as well as shared subspace.

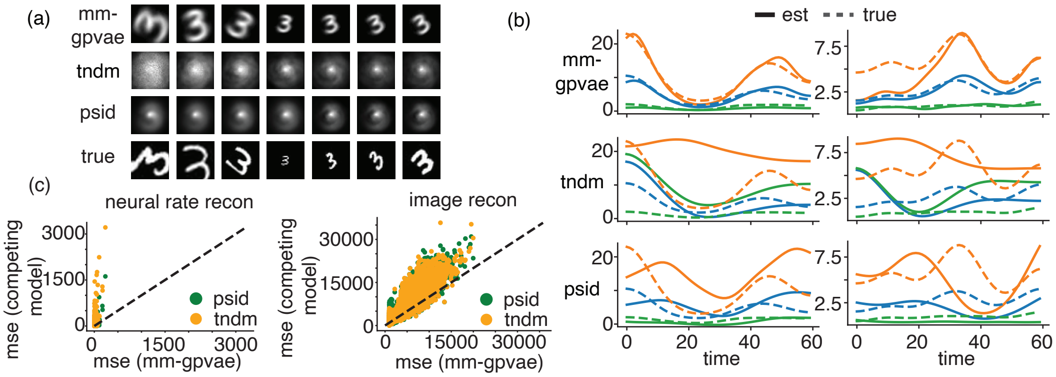

Figure 1: (a) Reconstruction of the scaling/rotating MNIST digit '3' with MM-GPVAE, TNDM and PSID. (b) Reconstruction of neural rates with MM-GPVAE, TNDM, and PSID. (c) MSE for neural rate reconstruction (left), and MSE for image reconstruction (right). Here, each dot indicates one trial mse from MMGPVAE vs a competing model. The majority of the trials errors fall above the unity line for both models, indicating overall better reconstruction with MMGPVAE.

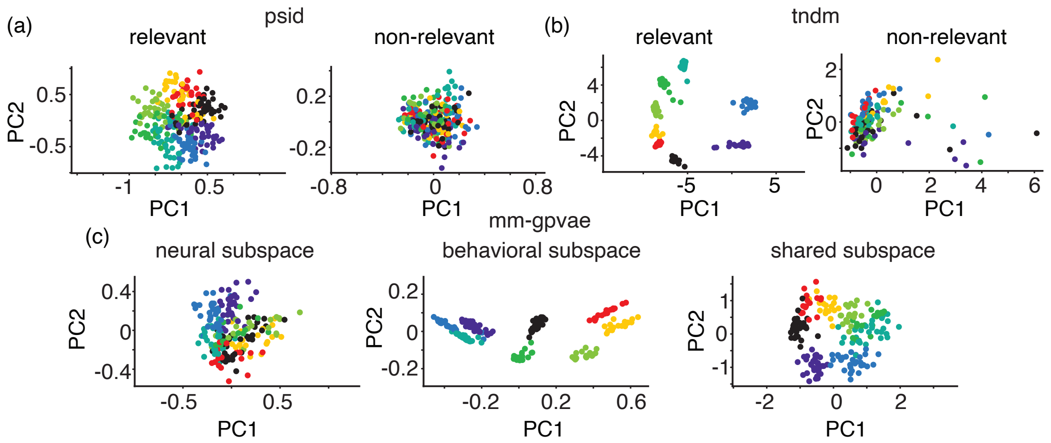

Figure 2: (a) Separation of all 8 reaching directions in the relevant subspace using PSID. (b) Separation of all 8 reaching directions in the relevant subspace using TNDM. (c) Separation of all 8 reaching directions in behavior-only and shared subspaces. Dots here indicate the mean latent value across the entire trial. The neural subspace shows no behavioral separation in the latent-space whereas the behavioral and shared subspaces show strong behavioral separation. This result closely parallels with (Hurwitz et al., 2021; Sani et al., 2021), which isolates behaviorally relevant and irrelevant neural subspaces. However, in contrast to (Hurwitz et al., 2021; Sani et al., 2021) the MM-GPVAE isolates a distinct shared subspace as well as both neural and behavioral independent subspaces from a raw unsupervised partitioning of both the behavioral and neural data.

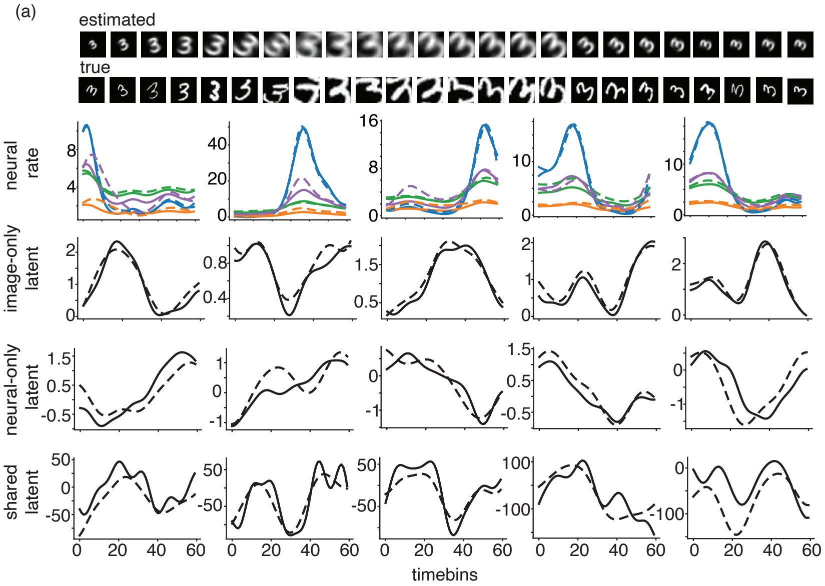

Figure 3: (a) More examples of estimated images and neural rates of the MNIST digit ’3’ as well as independent and shared latent estimation.

Additional evaluations of the MM-GPVAE on simulated data

To provide a more complete picture of the ability of the MM-GPVAE to both reconstruct simulated ’behavioral’ and neural data, and to accurately recover the true underlying latent variables, we show additional performance evaluations here. Figure 3 shows an example of 24 reconstructed 3s from the evaluation shown in Figure 3 of the manuscript. Again, here we reconstruct a scaling and rotating MNIST digit ’3’ alongside 100 neural spike trains, where one latent dimension is shared across modalities. Figure 3 also shows latent reconstructions and 5 example neural rates in 5 held out trials. Using the same set-up, we additionally run the MM-GPVAE with MNIST digit ’2’, and show an example of 24 reconstructed images as well as 3 example neural rates (of 100) on 5 held-out trials in Figure 4.

In addition to these, we also extend figure 3(a) from our manuscript with all the latent trajectories for our simulated data comparisons, including estimates from GPFA and GPVAE. Because in GPFA and GPVAE there is ambiguity as to which latent is “shared” (as these are unimodal models), attached Figure 5 panels (b) and (c) show the mean squared error for both of the possible latents for GPVAE and GPFA. Here, we note that the MMGPVAE is able to ‘disentangle’ the shared representation across the modalities, something that each unimodal model is unable to do. We visualize both the average mean-squared error (c) on held-out data as well as a per-trial mean-squared error scatter-plot (b) to provide a better sense of model performance.

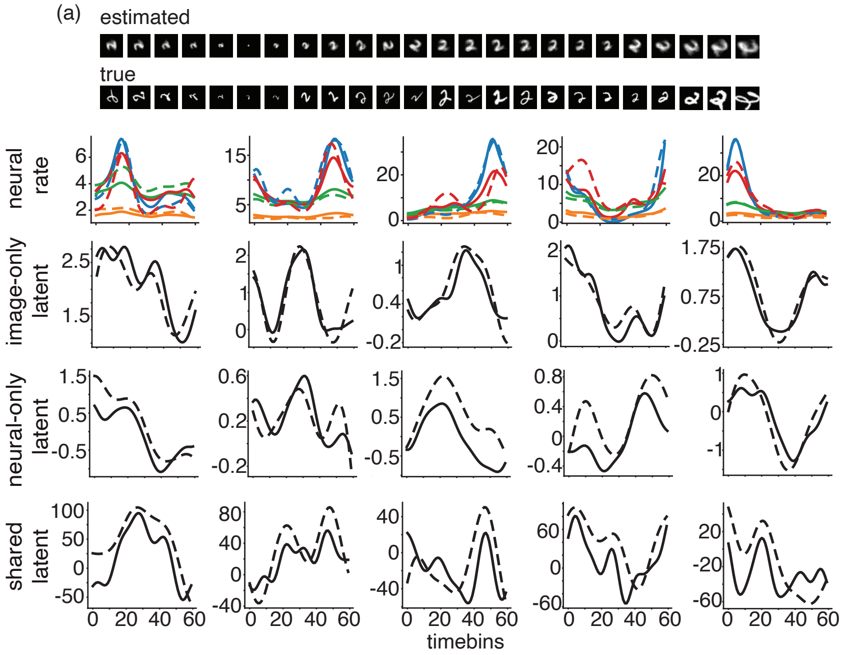

Figure 4: (a) An example of using the MNIST digit '2' in MM-GPVAE with reconstructed digits and neural rates as well as independent and shared latent estimation.

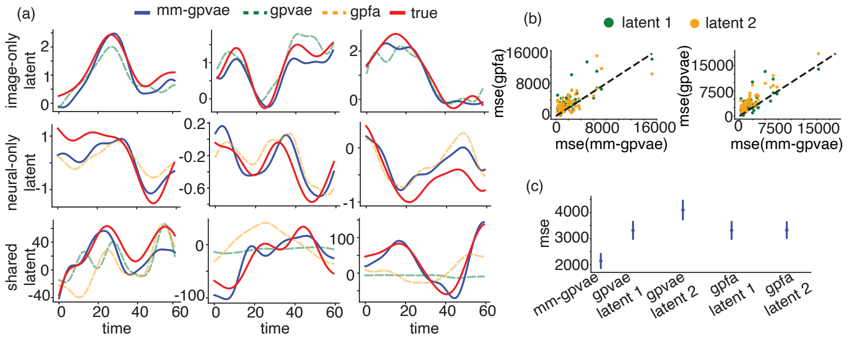

Figure 5: (a) Extension of Fig 3.(a) from the manuscript with all latent trajectories, showing the best matched latent from GPFA and GP-VAE to the true shared and independent latents. (b) MSE for both possible latents estimating the shared latent for GPFA (left) and GP-VAE (right). Each dot indicates a trial. (c) Average MSE for recovery of shared latent for the possible latents for all three models.

Assessing the importance of pruning the Fourier frequencies

Figure 6 demonstrates the necessity of pruning the Fourier frequencies when evaluating the MM-GVPAE on our simulated data. Here, we present pruned and not-pruned MM-GPVAE model fits (Figure 6a) and show their ability in reconstructing the true generative smooth latent across each latent subspace (Figure 6b). Without constraining the number of Fourier features for these data, the latent identification completely fails, akin to what is seen when we ignore the Fourier representation entirely. We note here that this feature is a function of the length of the trial, and the failure is not as stark for shorter-length trials. For these 40 time-point trials, our pruned condition the length scale of the GP to a value of 10 (i.e. the hyperparameter ℓ can go no smaller than 10).

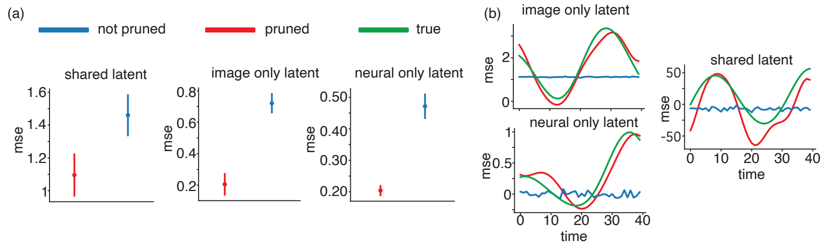

Figure 6: (a) MSE of all three latents for pruned and not pruned version of the Fourier MM-GPVAE. (b) Example latent trajectories for each subspace. Not pruned latents fail to represent the true latents accurately.

Varying the number of neurons

To demonstrate how the recovered latent of the shared and independent subspaces might trade-off with the amount of data a practitioner may have for a given modality, we consider fits of our MMGVAE in our simulated setting with a varying number of neurons. As expected, increasing the number of neurons increases the ability of the MMGPVAE to identify the neural latent, as well as its ability to accurately reconstruct neural data. Importantly, because structure is shared across modalities, increasing the number of neurons also increases the ability of the MMGPVAE to identify shared latent structure, as well as shows some improvement in the MMGPVAE to generate image data, as the model is able to extract cross-modality latent structure. Expectedly, increasing the number of neurons has no effect on recovering the image latent structure. (See figure 7)

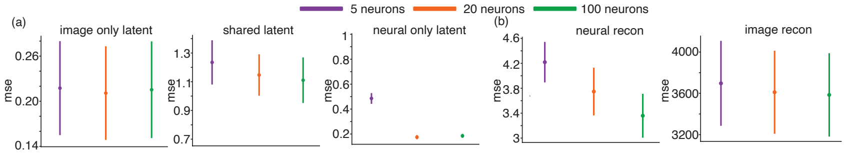

Figure 7: (a) Latent reconstructions for varying numbers of neurons. (b) Modality reconstructions for varying numbers of neurons.

Robustness to measurement error

To assess how the MMGPVAE might perform in the presence of noisy or imperfect behavioral measurements, we consider a simulated experiment where a randomly selected 20% of the images are greyed on either the top or bottom half. We fit the MMGPVAE jointly to the image data of this form as well as uncorrupted neural rates of the same form as in the main manuscript. We find that the MMGPVAE is able to exploit smoothness over time, rendering well-generated images even in during these occluded trials (Fig 8a). As expected, we find that a standard MM-VAE (without a GP prior over time) is not able to reconstruct the true image as well, but does a better job reconstructing the occluded image, as it does not exploit any smoothness over time and treats every frame independently (Fig 8b). We also show that this occlusion procedure has no bearing on the ability of the MMGPVAE or MMVAE to reconstruct the neural rates (Fig 8c)

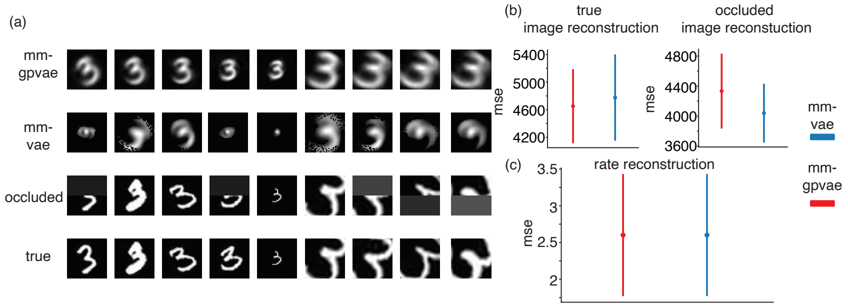

Figure 8: (a) Image reconstructions of MM-VAE and MM-GPVAE trained on occluded threes alongside the true and occluded threes. (b) (left) MSE of MM-GPVAE and MM-VAE for the true, not occluded data. (right) MSE of MM-GPVAE and MM-VAE for the occluded data. (c) MSE of rate reconstructions.

Neural Net Architectures

For the MM-GPVAE, both the simulated and real-world multi-modal evaluations used similar neural net architectures. However, there were some modifications of the nodes/layers that were unique to each evaluation. This was necessary as the latent dimensionality was different across our different evaluations, and certain behavioral reconstructions, especially those that were high-dimensional, required richer neural network parameterizations. The schematic for the MM-GPVAE neural network architecture in our simulated example can be found in Figure 9, and the schematics for the neural network architectures for the real-world multi-modal datasets can be found in Figure 10 (fly) and Figure 11 (hawkmoth). For our evaluation on the dataset used in (Hurwitz et al., 2021), whose results are above, the architecture can be found in Figure 12. Note that for all experiments, each modality is encoded to its own set of variational means and variances (transformed into the Fourier domain). The encoded means and variances representing the shared latents are then summed to give the encoded shared latents means and variances. Across all evaluations, we parameterized our latents in the Fourier domain and converted back to the time domain before decoding.

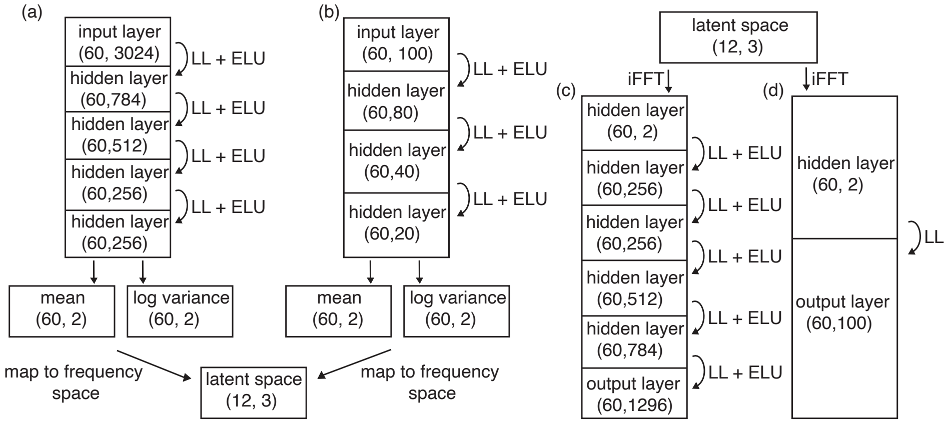

Figure 9: MM-GPVAE architecture for simulated data. (a) Encoder network of the MNIST digit. (b) Encoder network of the neural information. (c) Decoder network of the MNIST digit. (d) Decoder network of the neural information.

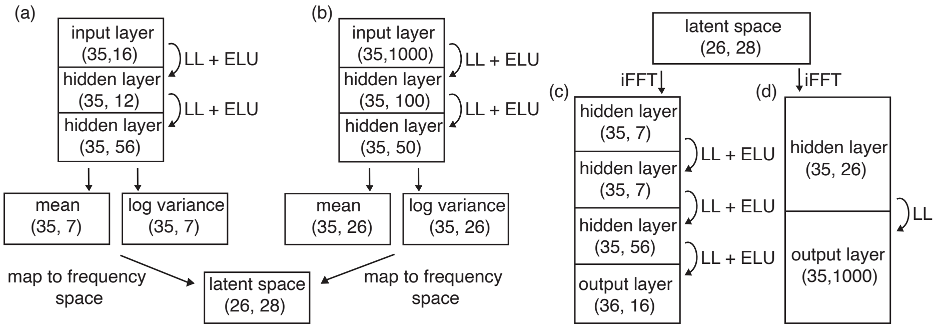

Figure 10: Neural network architecture for evaluation on fly dataset. (a) Encoder network for behavior. (b) Encoder network for neural data. (c) Decoder network for behavior. (d) Decoder network for neural data.

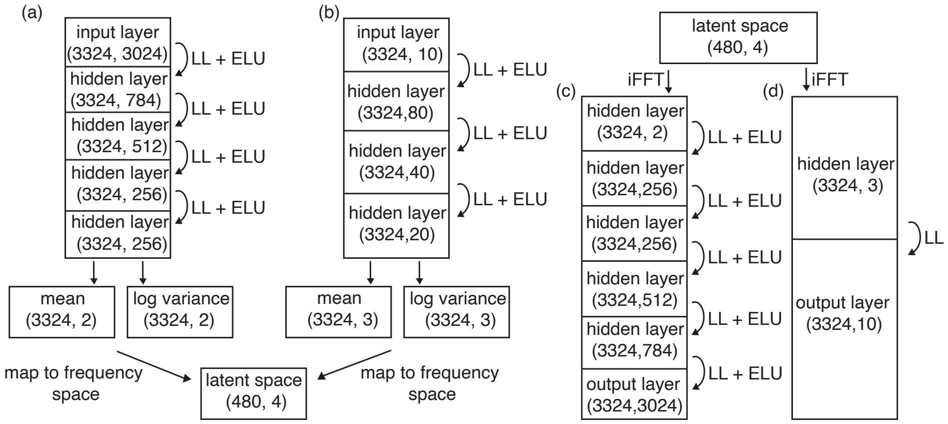

Figure 11: Neural network architecture for evaluation on hawkmoth dataset. (a) Encoder network for behavior. (b) Encoder network for neural data. (c) Decoder network for behavior. (d) Decoder network for neural data.

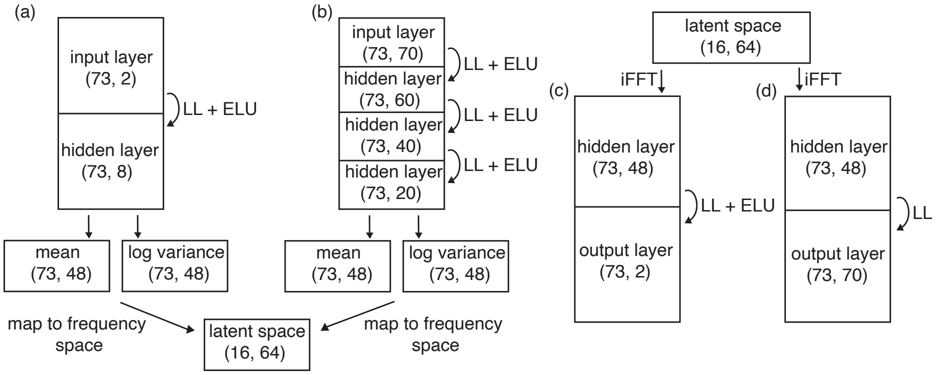

Figure 12: Neural network architecture for Monkey reaching data, evaluated above. (a) Encoder network for the behavior. (b) Encoder network for the neural activity. (c) Decoder network for the behavior. (d) Decoder network for the neural activity.

Additional information on fly experimental data

Pre-processing: We isolated 1000 raw calcium traces from (Schaffer et al., 2021) for use with our model for the evaluations in section 4.2 of the manuscript. For this set-up, we consider a variant of the MM-GPVAE where we have removed the non-linearity for the neural data modality, and instead consider Gaussian observations with an additional parameter controlling the observation variance for the fluorescence traces (akin to the original formulation of GPFA (Yu et al., 2009)). This dataset also contained x,y positions from 8 tracked limbs positions (Mathis et al., 2018). The data came as one continuous recording, and at every time point there was a behavioral categorization probability for one of 6 distinct behaviors (undefined, still, running, front grooming, back grooming, abdomen bending) determined via the algorithm outlined in (Whiteway et al., 2021). To split this continuous recording into trials, we found segments of the recording where, for 35 continuous samples, a single behavior was estimated at a $\geq \%60$ probability. This generated 318 total trials where each trial was of one of 5 possible behaviors. There was no section of the recording where there was an 'undefined' behavior for 35 continuous samples, so there are no undefined behavioral trials in our analysis.

Additional evaluation: Though we show the time-course of the behavioral reconstruction in the main manuscript, we show a 2-d reconstruction on 2 example trials in Figure 13. Here we plot true x,y positions for 2 trails for each of the 8 limb positions alongside our model's reconstruction. We can see here that our model also captures the spatial information in the behavioral data. We also show for comparison the 2-d depiction of the shared and independent latent representation of all trials in the dataset (Figure 14), with all five of the behaviors labelled. You can see here that the 'still' behavior separates well from all other behaviors in the shared and neural subspaces.

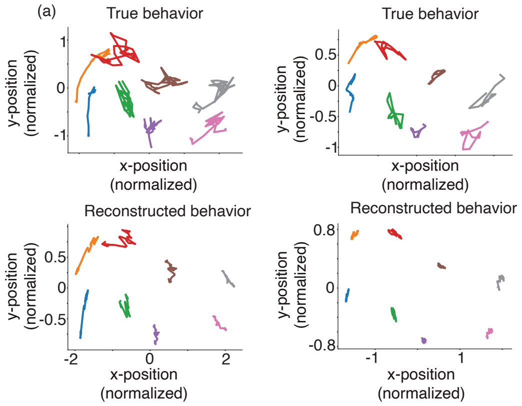

Figure 13: (a) Reconstruction of 8 fly limb positions in 2 held-out trials. Here, we see the MM-GPVAE is able to reconstruct the spatial information of the behavioral modality in the 8 tracked limb positions.

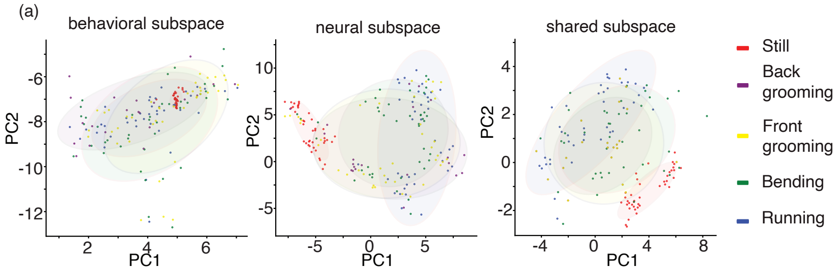

Figure 14: (a) Separation of all 5 fly behaviors in the behavior-only, neural-only, and shared subspaces. Similar to what is seen in the main manuscript, the 'still' behavior is well separated from trials with other behavioral labels in the neural and shared latent subspaces.

Additional information on hawkmoth experimental data

Pre-processing: For our hawkmoth data, the original synthetic visual stimuli were sampled at 125 Hz and the neural and torque recordings were sampled at 10K Hz (Sprayberry and Daniel, 2007; Sikandar et al., 2023; Putney et al., 2019). To prepare these data for evaluation with the MM-GPVAE, we first downsampled neural information and torque measurements to 1K Hz by binning the spike counts and averaging the torque measurements at this temporal resolution. To align these measurements with the visual stimuli, we upsampled the images 8-fold. The entire recording was 20 seconds long and was split into 6 evenly-divided trials. Here, it is worth noting that the trials are much longer ($\sim$ 3300 timebins) than in the other experiments in this work.

The visual stimuli contained 3024 pixels and there were ten total hawkmoth motor neurons. We set the neural-independent subspace to 2-dimensional, the images-independent subspace to 1-dimensional, and an additional 1 dimension for the shared subspace. To encourage slow-evolving smooth latents in the shared and image subspaces, and faster-evolving neural latents, we initialized the length scale parameters for each latent dimension to different values. The length scale was set to 10 for the neural latents, 150 for the shared latent, and 300 for the image latent. This biased the model fits to capture slower dynamics in the image subspace and the faster dynamics in the neural subspace.

Additional evaluation: The neural rates are well-captured by the MM-GPVAE for all 10 hawkmoth motor neurons, which each show strong periodicity. Figure 15 shows the strong fast-oscillation spike rates captured by the MM-GPVAE for all ten hawkmoth neurons alongside the recorded spikes for a 1.5 second period. Similar periodicity exists across the entire 20 second recording.

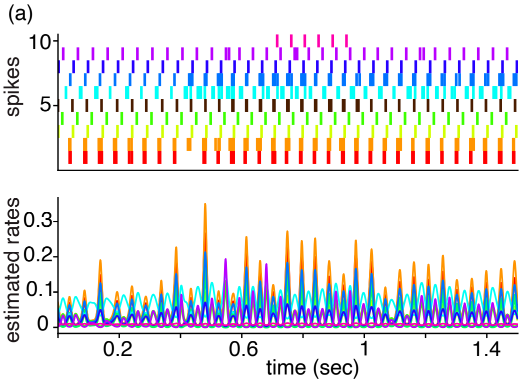

Figure 15: (a) True spikes and estimated rates of all 10 hawkmoth neurons over 1500 time points.

Additional model fitting information

All data had a 80-20 split for training and testing respectively. The Fourier frequency pruning was set to the minimum length scale of 10, 10, 3, and 16 for GP-VAE (simulated), MM-GPVAE (simulated), MM-GPVAE (fly), and MM-GPVAE (moth) respectively. GP length scales parameters were initialized to a value of 30 for all except for the hawkmoth evaluations (where initial values are indicated above), and jointly optimized with the ELBO. The covariance parameter α was set at a fixed value of 1e-2, 1e-2, 1e-3, and 1e-4 for GP-VAE (simulated), MM-GPVAE (simulated), MM-GPVAE (fly), and MM-GPVAE (moth) respectively. We additionally initialized the offsets $\mathbf{d}$ of the neural modality to the average log-rate of the neural data.

Latent dimensionality selection: In the case of the simulated examples in this work, the dimensionality was always set to the true generative dimensionality. For the real neural dataset examples in the main paper, the dimensionality was chosen using a different process for each modality. For the neural dataset, we use PCA and select the number of dimensions that capture greater than 97% of the variance. We choose PCA here because we use a linear decoder for this modality. For the other modality (limb position or visual image) the dimensionality was chosen by increasing the number of dimensions in a unimodal Fourier GPVAE until cross-validated performance did not show significant improvement. Then, we evaluate data from both modalities using the MMGPVAE and systematically increase the number of dimensions shared between datasets until predictive performance no longer improved. Across both datasets, we always saw that adding some shared dimensions increased overall predictive performance compared to none, suggesting shared structure across the data modalities.

We also note that the dimensionality selection procedure was used to approximate share and independent dimensionality as best as possible, but across CV folds the variance in cross validated log-likelihood was high. However, small changes to the number of shared and independent dimensions did not change latent representations or decoding performance much in our models.

For the evaluation where we compare MMGPVAE to TNDM and PSID in our attached figure, we choose approximately the same dimensionality as in the original TNDM paper so readers can easily compare fits across models.

Choice of Gaussian observation variance initialization, σ^2

We found that how we initially set the value of σ^2 could effect the performance of the MM-GPVAE, especially regarding the reconstruction of the behavioral data. Although we learned the σ^2 value jointly with the other parameters during optimization, the value of σ^2 tended to vary minimally from its initial value. Therefore, our setting of σ^2 tended to act as a reconstruction penalty, balancing the ability of the MM-GPVAE to prefer to reconstruct either the behavioral or neural data. Such a scaling term in the ELBO reconstruction has been seen in other models (Hurwitz et al., 2021), and in the original GP-VAE this parameter was chosen carefully through a specific cross-validation approach (Casale et al., 2018). Here, we simply set σ^2 to a value proportional to the dimensionality of the behavioral modality, which tended to balance the neural and behavioral terms in the ELBO, and generate good reconstructions for each modality. The inital values for σ^2 were 1000, 100, 1e-6, and 1 for GP-VAE (simulated), MM-GPVAE (simulated), MM-GPVAE (fly), and MM-GPVAE (moth) respectively.

Derivation of the evidence lower bound (ELBO)

Here we show the derivation of the evidence lower bound used for the MM-GPVAE. For clarity, we will start by deriving the ELBO for the MM-GPVAE just in the time domain. We set the GP prior parameters $\boldsymbol{\Theta} = \{\alpha, \theta\}$, $\boldsymbol{W}$ to be the loadings matrix and offsets from equation 4 in the manuscript, $\boldsymbol{T}$ to be the observed timepoints, and $\boldsymbol{Z}$ to be the collection of all the shared and independent latents $\boldsymbol{Z} = \{\mathbf{z_a}, \mathbf{z_s}, \mathbf{z_b}\}$.

$$
\begin{aligned}
&\log p\left(\boldsymbol{Y}^A, \boldsymbol{Y}^B \mid \boldsymbol{T}, \boldsymbol{\phi}, \boldsymbol{W}, \sigma^2, \boldsymbol{\Theta}\right) = \\
&\qquad \log \int \frac{p\left(\boldsymbol{Y}^A, \boldsymbol{Y}^B \mid \boldsymbol{Z}, \boldsymbol{\phi}, \boldsymbol{W}, \sigma^2\right) p(\boldsymbol{Z} \mid \boldsymbol{T}, \boldsymbol{\Theta})}{q_\psi(\boldsymbol{Z} \mid \boldsymbol{Y}^A, \boldsymbol{Y}^B)} q_{\boldsymbol{\psi}}(\boldsymbol{Z} \mid \boldsymbol{Y}^A, \boldsymbol{Y}^B) d\boldsymbol{Z} \\
&\quad \geq \int \log \left(\frac{p\left(\boldsymbol{Y}^A, \boldsymbol{Y}^B \mid \boldsymbol{Z}, \boldsymbol{\phi}, \boldsymbol{W}, \sigma^2\right) p(\boldsymbol{Z} \mid \boldsymbol{T}, \boldsymbol{\Theta})}{q_\psi(\boldsymbol{Z} \mid \boldsymbol{Y}^A, \boldsymbol{Y}^B)}\right) q_\psi(\boldsymbol{Z} \mid \boldsymbol{Y}^A, \boldsymbol{Y}^B) d\boldsymbol{Z} \\
&\quad = \mathbb{E}_{\boldsymbol{Z} \sim q_\psi}\left[\log p(\boldsymbol{Y}^A \mid \boldsymbol{\phi}, \sigma^2, \mathbf{z_s}, \mathbf{z_a}) + \log p(\boldsymbol{Y}^B \mid \boldsymbol{W}, \mathbf{z_s}, \mathbf{z_b}) + \log p(\boldsymbol{Z} \mid \boldsymbol{T}, \boldsymbol{\Theta})\right] \\
&\qquad\qquad - \int \log q_{\boldsymbol{\psi}}(\boldsymbol{Z} \mid \boldsymbol{Y}) q_\psi(\boldsymbol{Z} \mid \boldsymbol{Y}) d\boldsymbol{Z} \\
&\quad = \mathbb{E}_{\boldsymbol{Z} \sim q_\psi}\Bigg[\sum_t \log \mathcal{N}\left(\boldsymbol{y}_A \mid g_{\boldsymbol{\phi}}\left(\boldsymbol{x}_A\right), \sigma^2 \boldsymbol{I}_N\right) + \sum_t \log(\mathcal{P}(\boldsymbol{y}_B | f(\boldsymbol{x}_B)) \\
&\qquad\qquad + \log p(\boldsymbol{Z} \mid \boldsymbol{T}, \boldsymbol{\Theta})\Bigg] + H(q_\psi)
\end{aligned}
$$

The Fourier domain representation of the ELBO only requires sampling over variational parameters in the Fourier space, but only changes the expression of the GP prior term $p(\boldsymbol{Z})$.

GP prior: The expectation of the GP prior term can be expressed in the Fourier domain as:

$$\begin{aligned}
\mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}\left[p(\tilde{\boldsymbol{Z}} \mid \boldsymbol{\Theta}, \boldsymbol{\omega})\right] &= \sum_{p,\omega} \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}\left[\log \mathcal{N}\left(\tilde{z}_{p,\omega}|0, [\tilde{\boldsymbol{K}}_p]_{\omega,\omega}\right)\right] \\
&= \tfrac{1}{2}\sum_{p,\omega} \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}\left[\log([\tilde{\boldsymbol{K}}_p]_{\omega,\omega} + \alpha) + \frac{\tilde{z}^2_{p,\omega}}{([\tilde{\boldsymbol{K}}_p]_{\omega,\omega} + \alpha)}\right] \\
&= \tfrac{1}{2}\sum_{p,\omega} \log([\tilde{\boldsymbol{K}}_p]_{\omega,\omega} + \alpha) + \frac{\tilde{\sigma}^2_{p,\omega}(\boldsymbol{Y}) + \tilde{\mu}^2_{p,\omega}(\boldsymbol{Y})}{([\tilde{\boldsymbol{K}}_p]_{\omega,\omega} + \alpha)},
\end{aligned}$$

where the double sum is due to the variational distribution q being a mean field Gaussian, and p here indexes latents and ω indexes Fourier frequencies.

Neural likelihood: Since the estimated log-rates of the neural data is a linear transform of the shared and neural latent variables, we can also evaluate the expectation of the neural-modality likelihood term in closed form. Recall that

$$\boldsymbol{X} = \boldsymbol{Z} = \tilde{\boldsymbol{Z}}\boldsymbol{B}^\top,$$

where $\boldsymbol{X}$ is the matrix of embeddings, and $\tilde{\boldsymbol{Z}}$ is the $P \times \mathcal{F}$ matrix of Fourier-domain latent variables. We may therefore write the embedding for measurement i at time t as

$$x_{i,t} = \boldsymbol{w}_i^\top \tilde{\boldsymbol{Z}} \boldsymbol{b}_t,$$

where $\boldsymbol{w}_i$ is the i^{th} row of and $\boldsymbol{b}_t^\top$ is the t^{th} column of $\boldsymbol{B}^\top$. If we let $\tilde{\mu}_{p,\omega} = \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[\tilde{z}_{p,\omega}]$, and correspondingly let $\tilde{\boldsymbol{\mu}}$ be the $P \times \mathcal{F}$ matrix of $\tilde{\mu}_{p,\omega}$'s then $E_{\tilde{\boldsymbol{Z}}\sim q_\psi}[x_{i,t}] = \boldsymbol{w}_i^\top \tilde{\boldsymbol{\mu}} \boldsymbol{b}_t$. We may equivalently write $x_{i,t}$ as

$$\begin{aligned}
x_{i,t} &= \text{vec}(x_{i,t}) = \text{vec}(\boldsymbol{w}_i^\top \tilde{\boldsymbol{Z}} \boldsymbol{b}_t) \\
&= \text{vec}(\boldsymbol{b}_t^\top \tilde{\boldsymbol{Z}}^\top \boldsymbol{w}_i) \\
&= (\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)\tilde{\boldsymbol{z}},
\end{aligned}$$

where $\tilde{\boldsymbol{z}} = \text{vec}(\tilde{\boldsymbol{Z}}^\top)$. We may therefore we may derive the variance of $x_{i,t}$ as

$$\begin{aligned}
\text{Var}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[x_{i,t}] &= \text{Var}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[(\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)\tilde{\boldsymbol{z}}] \\
&= \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[((\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)\tilde{\boldsymbol{z}})^2] - \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[(\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)\tilde{\boldsymbol{z}}]^2 \\
&= \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[\text{Trace}[(\boldsymbol{w}_i \otimes \boldsymbol{b}_t)(\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)\tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}^\top]] - (\boldsymbol{w}_i^\top \tilde{\boldsymbol{\mu}} \boldsymbol{b}_t)^2 \\
&= \text{Trace}[(\boldsymbol{w}_i \otimes \boldsymbol{b}_t)(\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)(\boldsymbol{V} + \text{vec}(\tilde{\boldsymbol{\mu}})\text{vec}(\tilde{\boldsymbol{\mu}})^\top] - (\boldsymbol{w}_i^\top \tilde{\boldsymbol{\mu}} \boldsymbol{b}_t)^2 \\
&= (\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)\boldsymbol{V}(\boldsymbol{w}_i \otimes \boldsymbol{b}_t),
\end{aligned}$$

where $\boldsymbol{V}$ is the diagonal posterior covariance of $\tilde{\boldsymbol{z}}$ whose elements are the encoded Fourier variational variances, $\tilde{\sigma}^2_\omega(\boldsymbol{Y})$. Therefore, we observe that under the variational posterior $x_{i,t}|\boldsymbol{Y} \sim \mathcal{N}(m_{i,t}, v_{i,t})$, where $m_{i,t} \equiv \boldsymbol{w}_i^\top \tilde{\boldsymbol{\mu}} \boldsymbol{b}_t$ and $v_{i,t} \equiv (\boldsymbol{w}_i^\top \otimes \boldsymbol{b}_t^\top)\boldsymbol{V}(\boldsymbol{w}_i \otimes \boldsymbol{b}_t)$.

We note that for $\lambda_{i,t} = e^{x_{i,t}}$ follows a long-normal distribution, meaning that, for $x_{i,t} \sim \mathcal{N}(m_{i,t}, v_{i,t})$ then $\mathbb{E}[\lambda_{i,t}] = e^{m_{i,t} + \frac{1}{2}v_{i,t}}$. This allows us to specify the posterior expectation of the Poisson likelihood in closed form. Namely,

$$\begin{aligned}
\mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}\left[\log \mathcal{P}(y_{i,t}|f(x_{i,t}))\right] &= \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}\left[y_{i,t}\log\lambda_{i,t} + \lambda_{i,t}\right] - \log y_{i,t}! \\
&= y_{i,t}\mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[x_{i,t}] + \mathbb{E}_{\tilde{\boldsymbol{Z}}\sim q_\psi}[\lambda_{i,t}] + \text{const}_{\tilde{\boldsymbol{Z}}} \\
&= y_{i,t}m_{i,t} + e^{m_{i,t} + \frac{1}{2}v_{i,t}} + \text{const}_{\tilde{\boldsymbol{Z}}}
\end{aligned}$$

References

Casale, F. P., Dalca, A. V., Saglietti, L., Listgarten, J., and Fusi, N. (2018). Gaussian process prior variational autoencoders. *arXiv preprint arXiv:1810.11738*.

Hurwitz, C., Srivastava, A., Xu, K., Jude, J., Perich, M., Miller, L., and Hennig, M. (2021). Targeted neural dynamical modeling. *Advances in Neural Information Processing Systems*, 34:29379–29392.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289.

Putney, J., Conn, R., and Sponberg, S. (2019). Precise timing is ubiquitous, consistent, and coordinated across a comprehensive, spike-resolved flight motor program. *Proceedings of the National Academy of Sciences*, 116(52):26951–26960.

Sani, O. G., Abbaspourazad, H., Wong, Y. T., Pesaran, B., and Shanechi, M. M. (2021). Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*, 24(1):140–149.

Schaffer, E. S., Mishra, N., Whiteway, M. R., Li, W., Vancura, M. B., Freedman, J., Patel, K. B., Voleti, V., Paninski, L., Hillman, E. M., et al. (2021). Flygenvectors: the spatial and temporal structure of neural activity across the fly brain. *bioRxiv*, pages 2021–09.

Sikandar, U. B., Choi, H., Putney, J., Yang, H., Ferrari, S., and Sponberg, S. (2023). Predicting visually-modulated precisely-timed spikes across a coordinated and comprehensive motor program. In *2023 International joint conference on neural networks (IJCNN)*. IEEE.

Sprayberry, J. D. and Daniel, T. L. (2007). Flower tracking in hawkmoths: behavior and energetics. *Journal of Experimental Biology*, 210(1):37–45.

Whiteway, M. R., Schaffer, E. S., Wu, A., Buchanan, E. K., Onder, O. F., Mishra, N., and Paninski, L. (2021). Semi-supervised sequence modeling for improved behavioral segmentation. *bioRxiv*, pages 2021–06.

Yu, B., Cunningham, J., Santhanam, G., Ryu, S., Shenoy, K., and Sahani, M. (2009). Gaussian-process factor analysis for low-d single-trial analysis of neural population activity. In *Frontiers in Systems Neuroscience. Conference Abstract: Computational and systems neuroscience*.