A Methods Details

A.1 Training Details and Hyperparameters

For each session of data, we split the population activity into three continuous partitions: approximately 90% for training, 10% for validation, and 10% for testing. The validation set was used to select the epoch with the best performance and to tune hyperparameters, while the test set was reserved for the final evaluation of the models and reporting performance metrics.

For SynapsNet, we used context window size $W = 5$, neuron embedding size of 32, and a single-layer GRU with hidden layer size of 40 as the dynamical model. During training, the training set is divided into identical (but overlapping) partitions of size $W + 1$, where the first W time points are the input and the last time point is the target. For batch sampling, we first select a random session from which all batch samples are chosen. This approach ensures that all samples in a batch have identical dimensionality, simplifying implementation, as different sessions contain different numbers of neurons.

For all datasets, we use a batch size of 32 and an initial learning rate of 10^{-3}. The models are trained for 100 epochs, with the learning rate halved every 10 epochs. Additionally, dropout layers with a rate of 0.1 are employed in all neural networks. SYNAPSNET is implemented using `PyTorch` in `Python`, and evaluations are run on a Linux machine with a GPU and 16GB of RAM. Training SYNAPSNET on each session of data (single-session training) takes approximately half an hour, depending on the number of recorded neurons and the session length.

For the sake of fair comparison, we use the same training, validation, and testing data for all benchmark models and ablated versions of SYNAPSNET.

A.2 Models Used for Comparison

The RNN, GRU, and LSTM models each consist of 2 layers with a hidden layer size of 100. The LFADS model we used has the dimensionality of the generator, encoder, and controller all set to 256, and the dimensionality of the factor and inferred inputs to the generator set to 128. The GWNet model is defined with 32 residual channels, 32 dilation channels, 128 skip channels, and 256 end channels, with a single layer, 4 blocks, and a kernel size of 2.

B Additional Results

B.1 Sensitivity Analysis

Effect of Context Window Length on Forecasting Performance. We evaluated the effect of different context window lengths (W) on the forecasting performance of both SYNAPSNET and NeuPRINT (the second-best-performing model). The results for Ca imaging and Neuropixels datasets are presented separately in Figure 6. The context window length determines how far back in the past the model can access neural activity and behavioral data to predict the population's activity at the next time point. The results indicate that at least two past time steps are required to accurately forecast neural activity, as evidenced by the substantial performance difference between $W = 1$ and $W = 2$. This aligns with the observations made by the authors of NeuPRINT, who identified $W = 2$ as the optimal choice (Mi et al., 2024). Performance plateaus after $W = 5$, with the negligible difference observed among context window sizes of 5, 10, and 20. Therefore, all results reported in the main paper are based on $W = 5$ to maintain model simplicity without significantly sacrificing performance.

Effect of Neuronal Embedding Dimensionality on Forecasting Performance. We investigated the impact of varying neuronal embedding sizes (D) on the forecasting performance of both SYNAPSNET and NeuPRINT. Figure 7 displays the results for the Ca imaging and Neuropixels datasets separately. Our analysis reveals that neuronal embedding dimensionality does not significantly affect neural activity forecasting accuracy. This aligns with the findings from the ablation study (Table 2), which indicated that neuronal embeddings contribute minimally to the performance of SYNAPSNET. Consequently, we opted for $D = 32$ to remain consistent with the choice made by Mi et al. (2024).

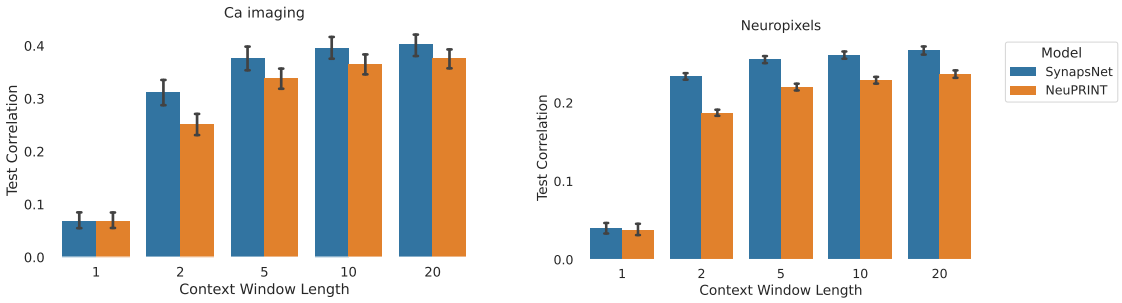

Figure 6: Sensitivity Analysis of Context Window Length (W). Test correlation scores for different context window lengths are plotted separately for the Ca imaging dataset (a) and the Neuropixels dataset (b).

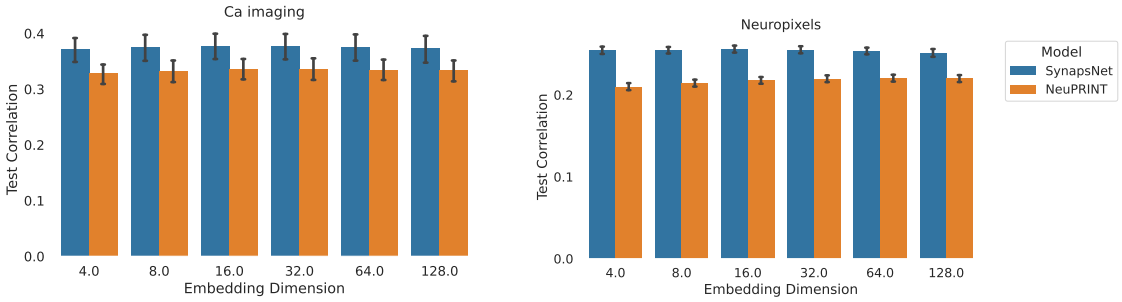

Figure 7: Sensitivity Analysis of Neuronal Embedding Dimensionality (D). Test correlation scores for different neuronal embedding sizes are plotted separately for the Ca imaging dataset (a) and the Neuropixels dataset (b).