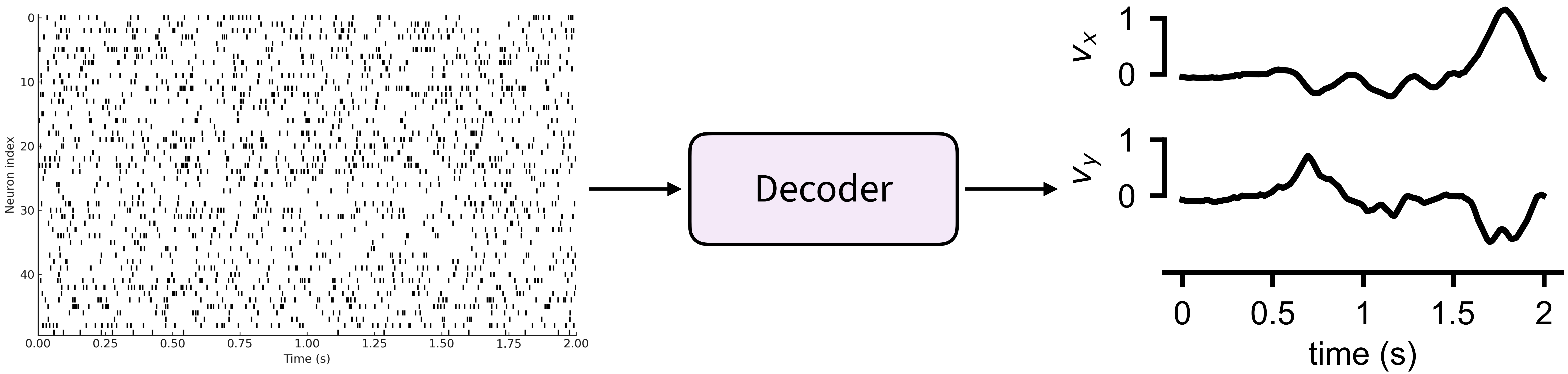
Generalizable, real-time neural decoding with hybrid state-space models

Avery Hee-Woon Ryoo*, Nanda H Krishna*, Ximeng Mao*, Mehdi Azabou,
Eva L Dyer, Matthew G Perich†, Guillaume Lajoie†

NeurIPS 2025

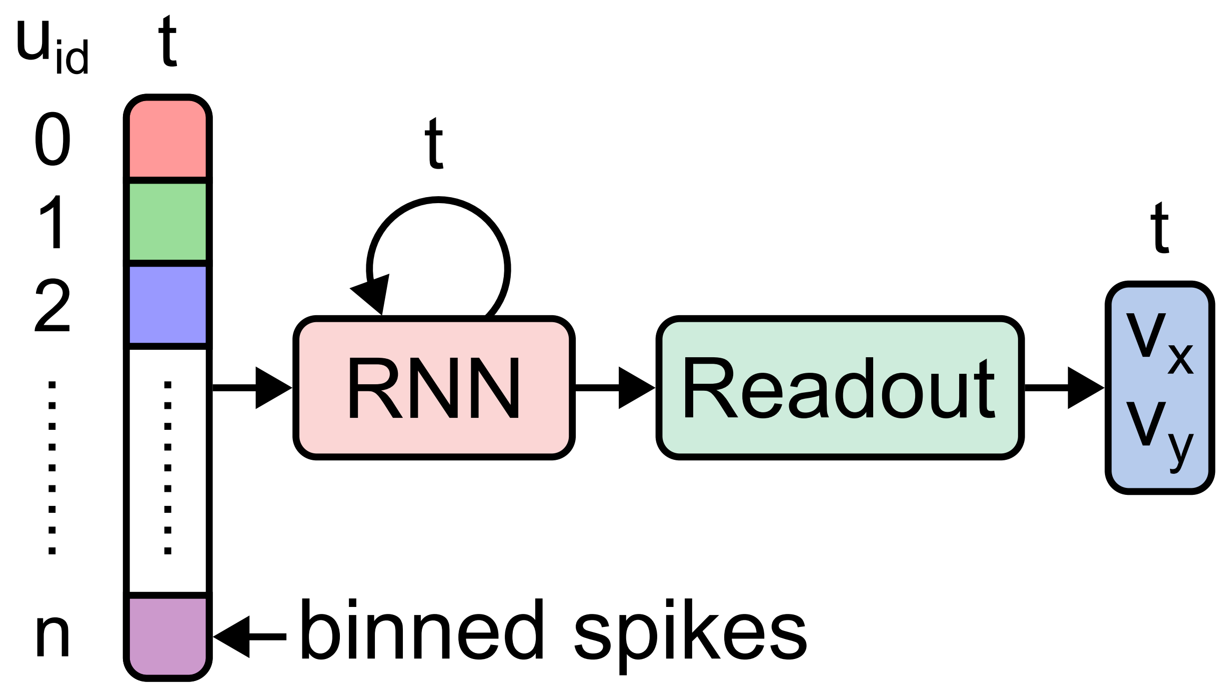
Introduction

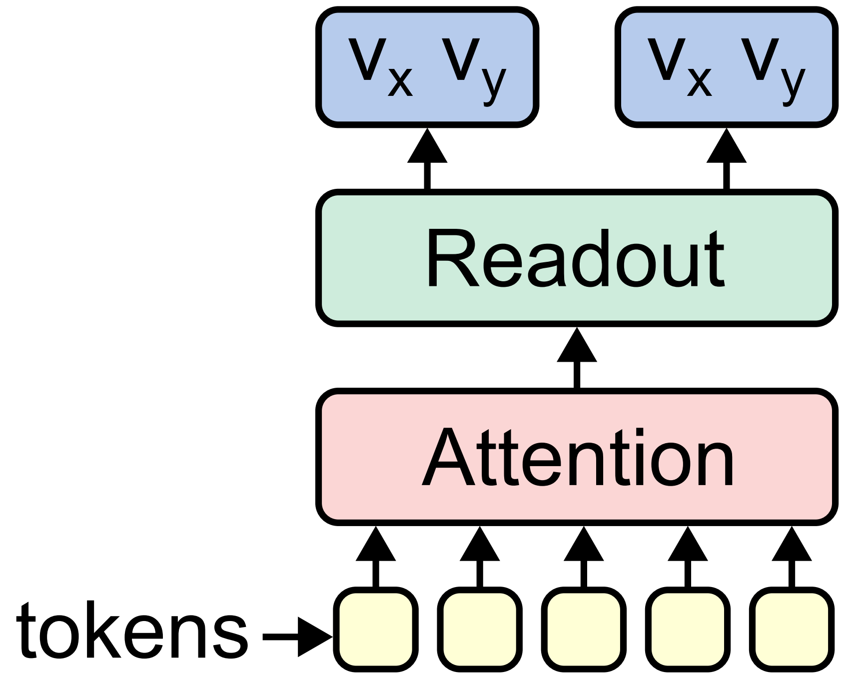
- Neural decoders translate neural activity into behavioural action or control signals



Introduction

- Typical deep learning-based approaches to neural decoding:
  - **Recurrent neural networks**
    - Efficient at inference
    - Rigid input specification
    - Limited generalizability
  - **Transformer-based approaches**
    - Slower at inference
    - Flexible input specification
    - Large-scale pre-training → strong generalization

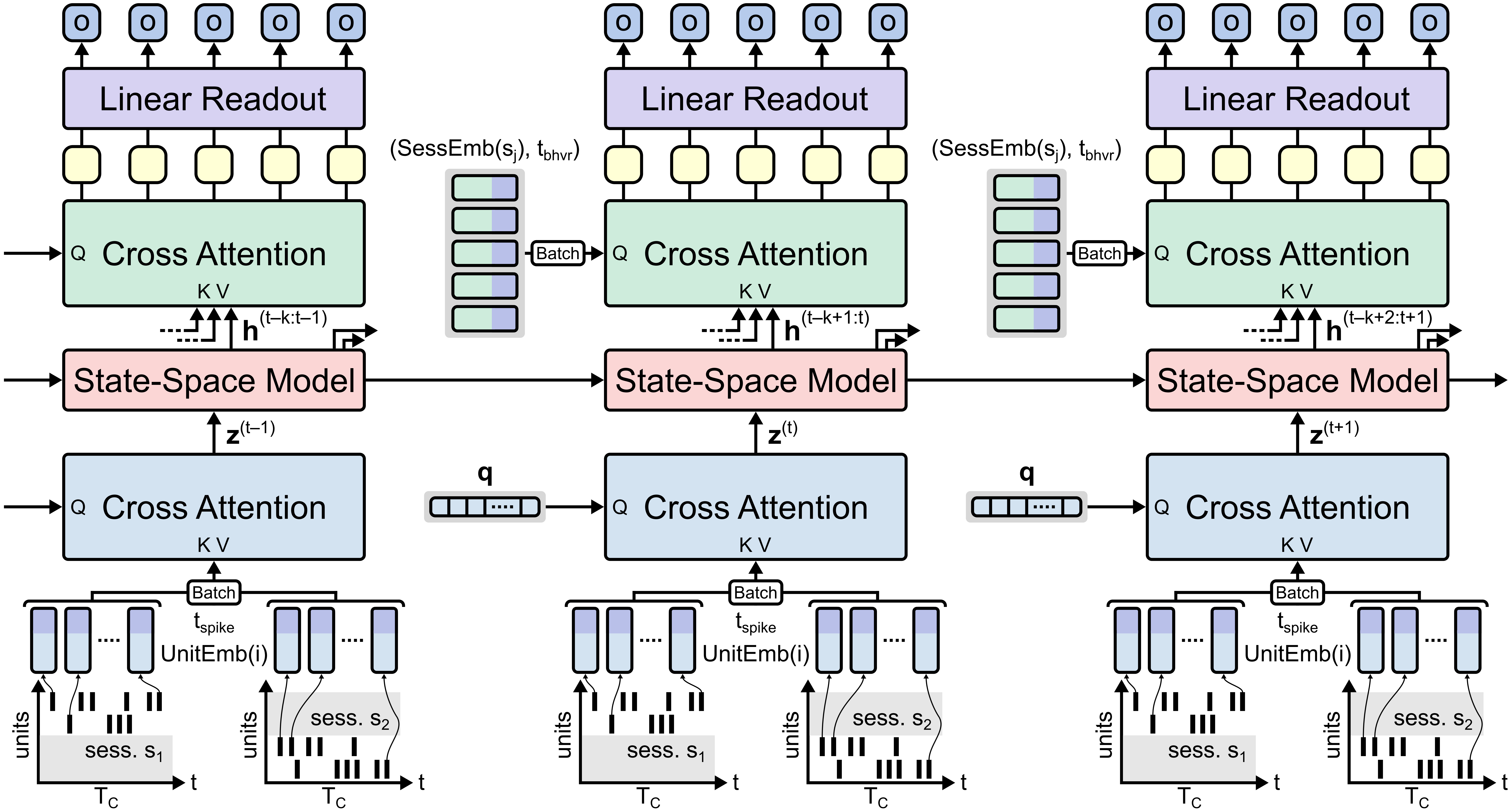




Introduction

- Online decoding + interfacing are important for therapeutic neurotechnologies
- We want to build neural decoders that are:
  - **Highly performant** in terms of decoding accuracy
  - **Efficient**, enabling real-time inference and control
  - **Generalizable** to new days and individuals with minimal calibration

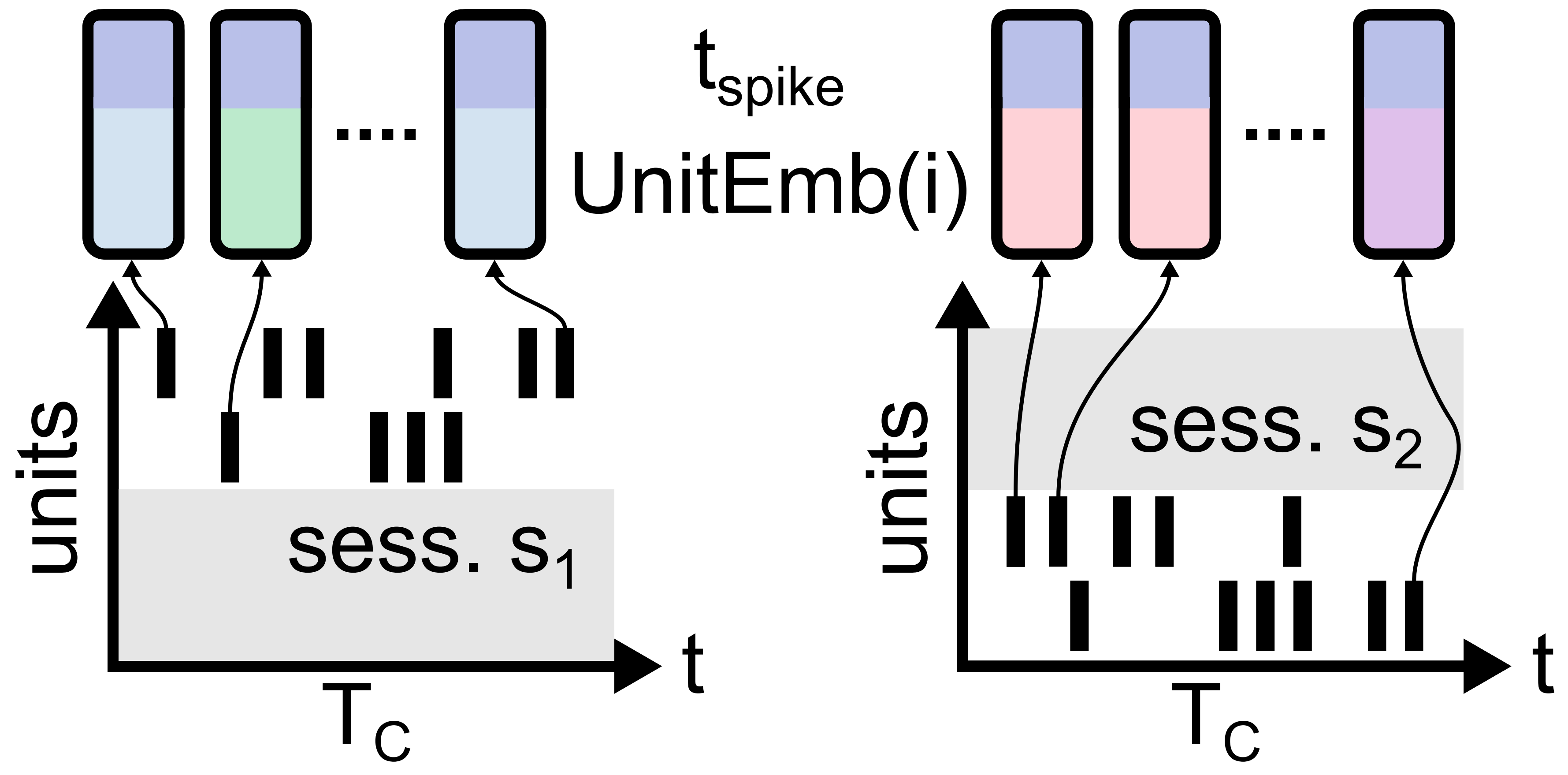
POSSM



o
Linear Readout
(SessEmb(s_j), t_bhvr)
Batch
Q
Cross Attention
K V
h^(t–k:t–1)
h^(t–k+1:t)
h^(t–k+2:t+1)
State-Space Model
z^(t–1)
z^(t)
z^(t+1)
q
Batch
t_spike
UnitEmb(i)
units
sess. s_1
sess. s_2
T_C
t

Spike tokenization

- Based on POYO (Azabou et al., 2023):
  - Embed each *unit* uniquely using learnable *unit embeddings*
  - Provide spike timing information through rotary positional embeddings (RoPE)

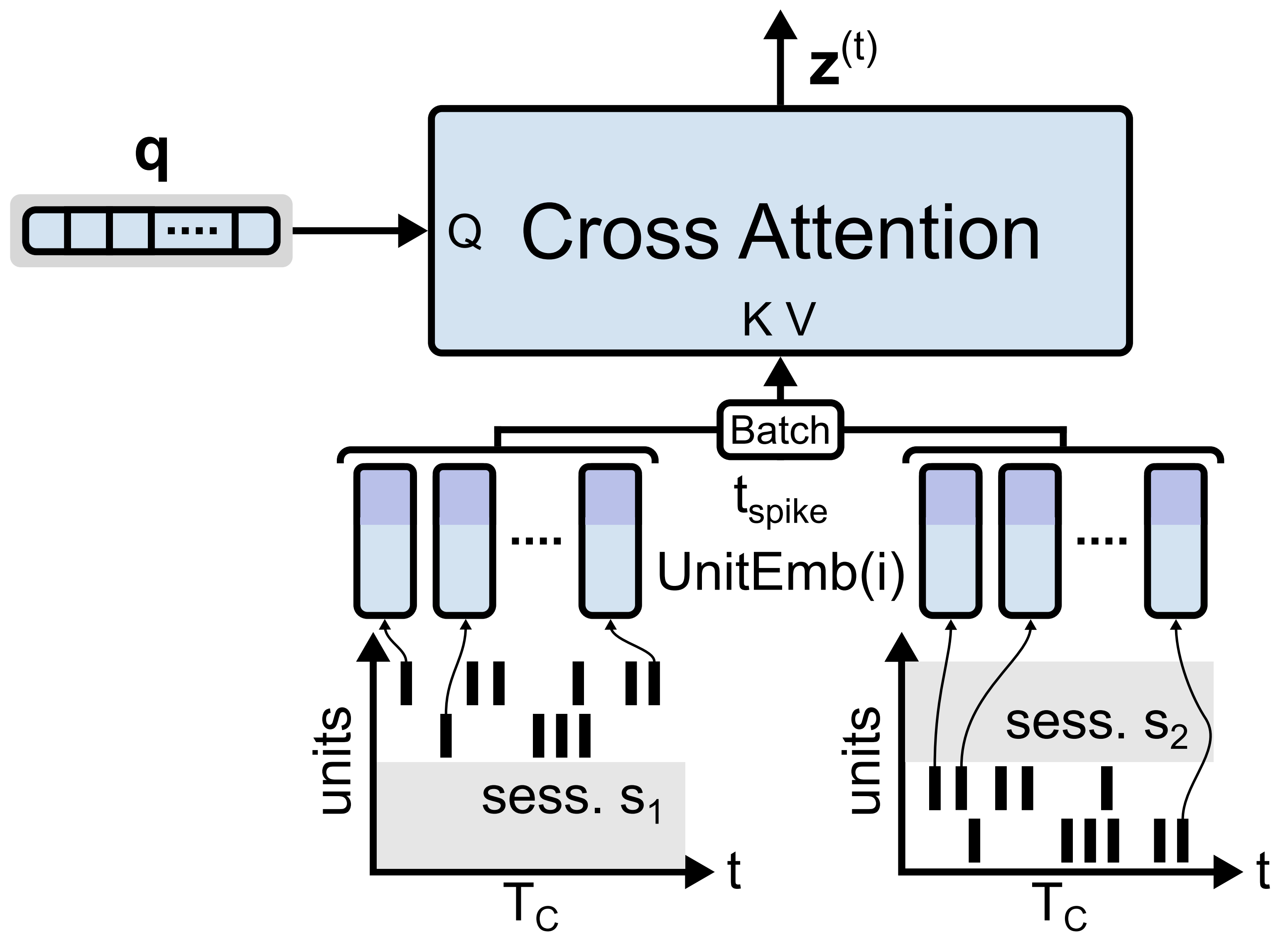


Input cross-attention

- Converts variable-length sequences of spike tokens into a fixed-size latent sequence:

$$\mathbf{z}^{(t)} = \mathrm{softmax}\left(\frac{\mathbf{q}\mathbf{K}_t^\top}{\sqrt{d_k}}\right)\mathbf{V}_t$$

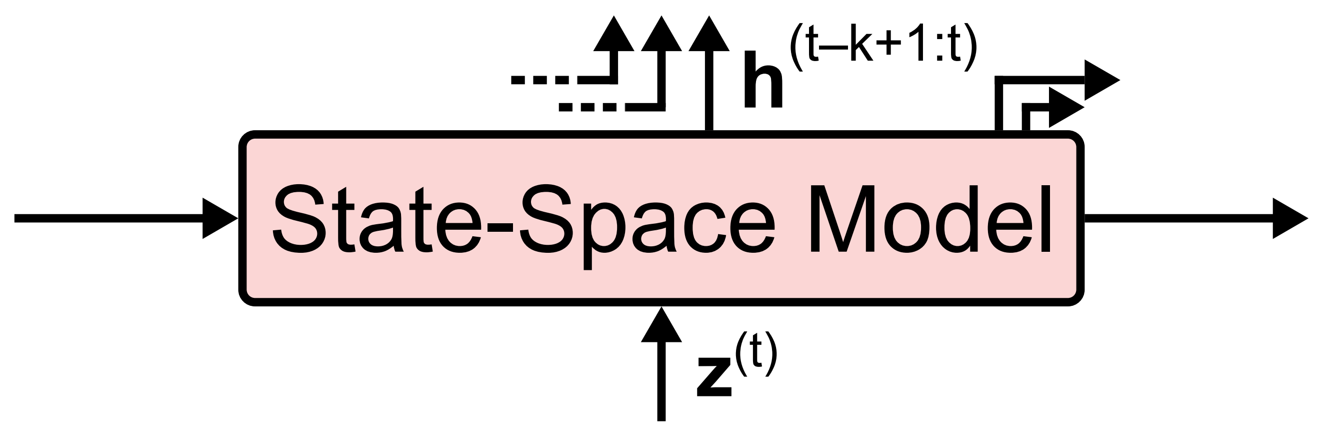
- We use one learnable query vector **q** in most of our experiments, so $\mathbf{z}^{(t)}$ is a single vector
- Compression which can represent more information than naïve time-binning



Recurrent backbone

- Maintains a hidden state $\mathbf{h}^{(t)}$ across time chunks
- This is updated based on the output of the cross-attention, $\mathbf{z}^{(t)}$, at each time chunk:

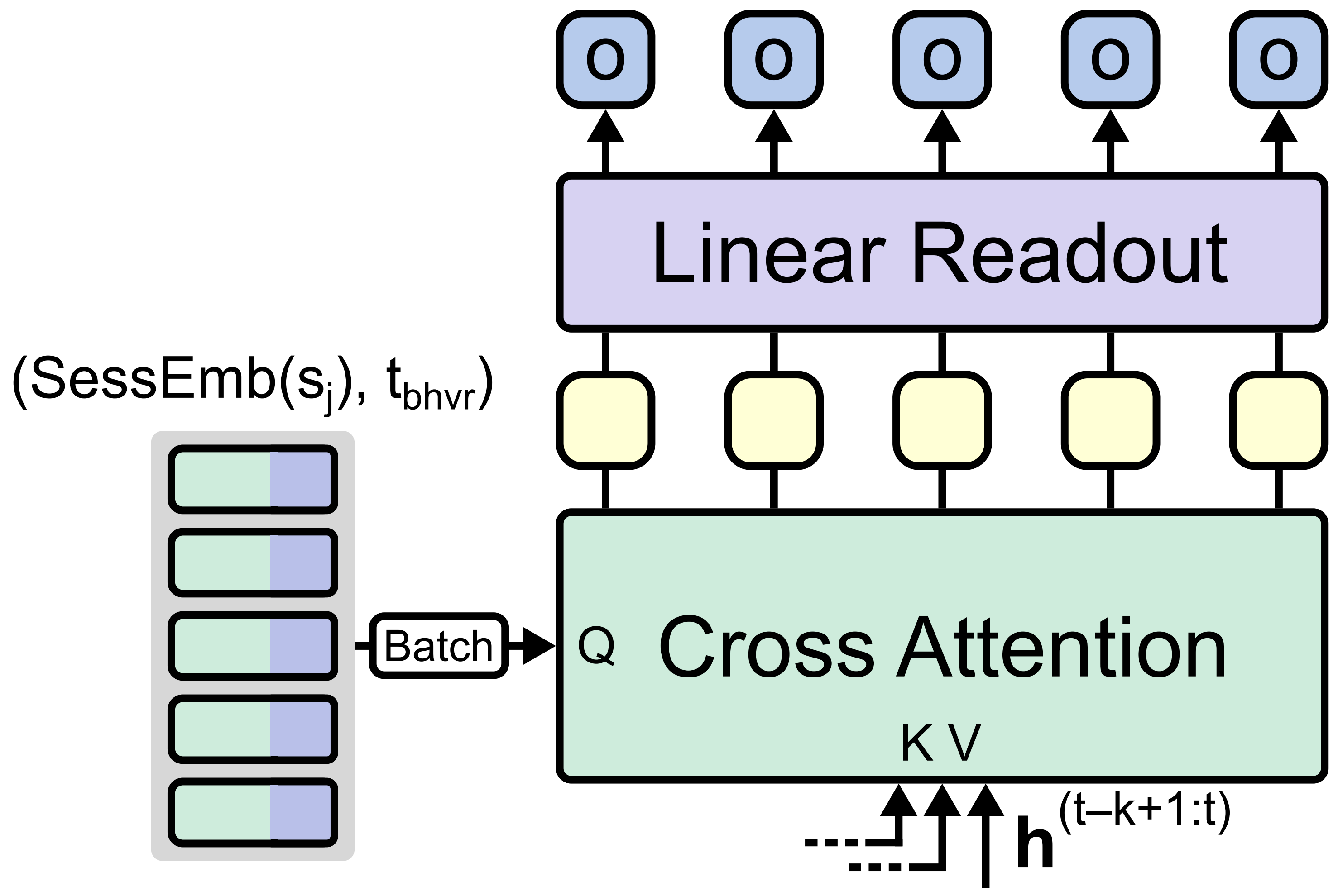
$$\mathbf{h}^{(t)} = f_{\text{SSM}}\left(\mathbf{z}^{(t)}, \mathbf{h}^{(t-1)}\right)$$



- In our experiments, we use:
  - GRU
  - S4D
  - Mamba

Output cross-attention & readout

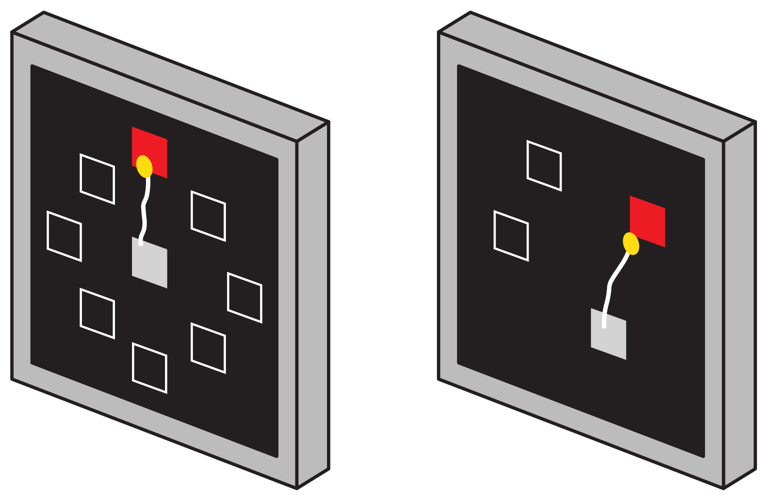
- We decode behaviour using $\{\mathbf{h}^{(t-k+1):(t)}\}$, i.e., the k most recent hidden states as keys and values
- Queries consist of learnable session embeddings (captures latent session-specific factors) and timestamp to predict behaviour at (via RoPE)
- Advantages:
  - Can predict multiple behaviours per chunk
  - No trial-alignment
  - Can predict beyond the current chunk



Multi-session pretraining & finetuning

- We pretrain **o-POSSM** on several datasets from different labs
- Two finetuning schemes:
  - **Unit identification** (UI)
    - Freeze all parameters, relearn unit + session embeddings
    - Parameter- and compute-efficient, matches single-session performance
  - **Full finetuning** (FT)
    - UI for first *k* epochs, unfreeze, finetune all parameters
    - Maximizes performance on target sessions

NHP reaching tasks



- POSSM is either competitive with or outperforms all baselines
- Pretrained o-POSSM outperforms single-session POSSM

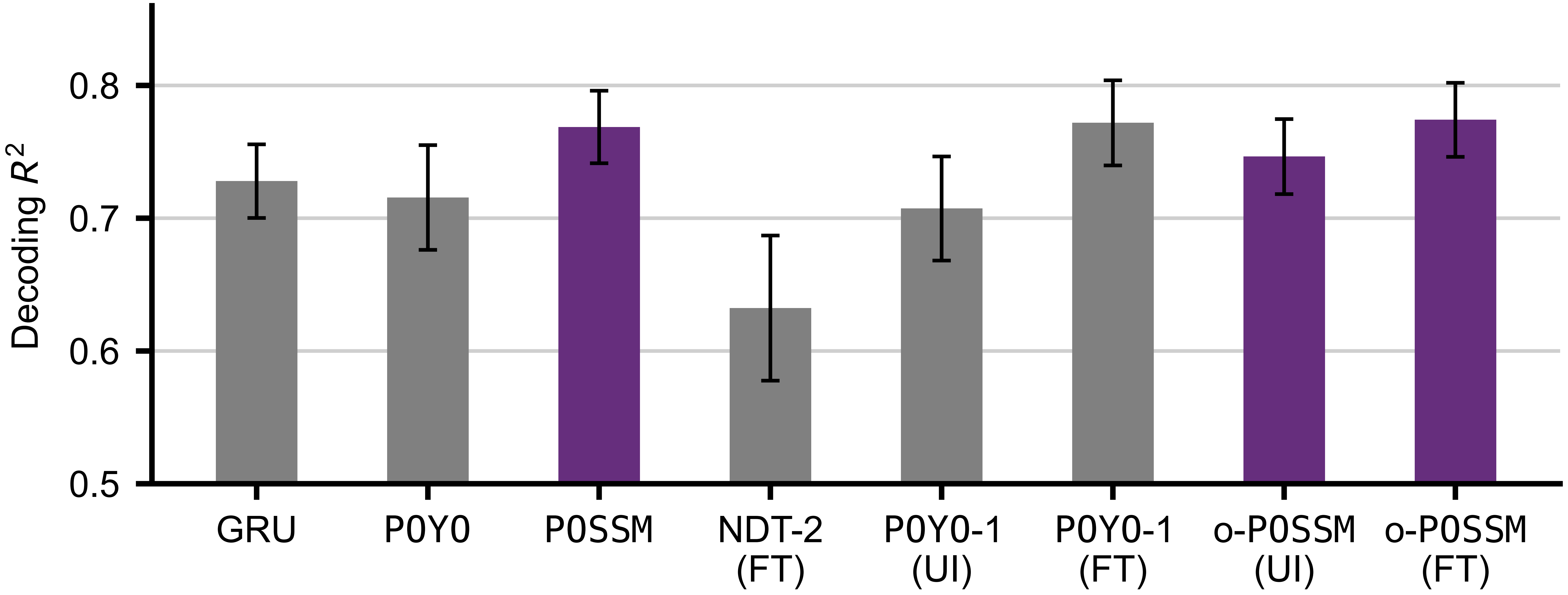
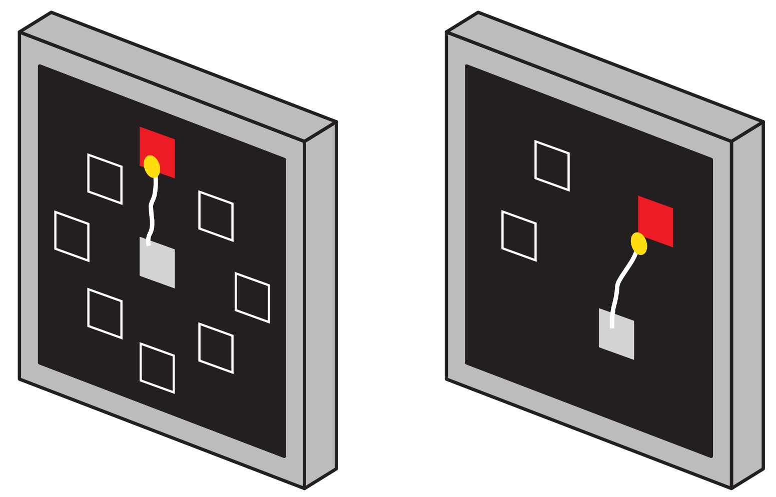


Figure: Decoding performance on held-out subject, RT task. We report the mean R^2 ± SD over 6 sessions.

Computational efficiency



- Training data + compute efficiency

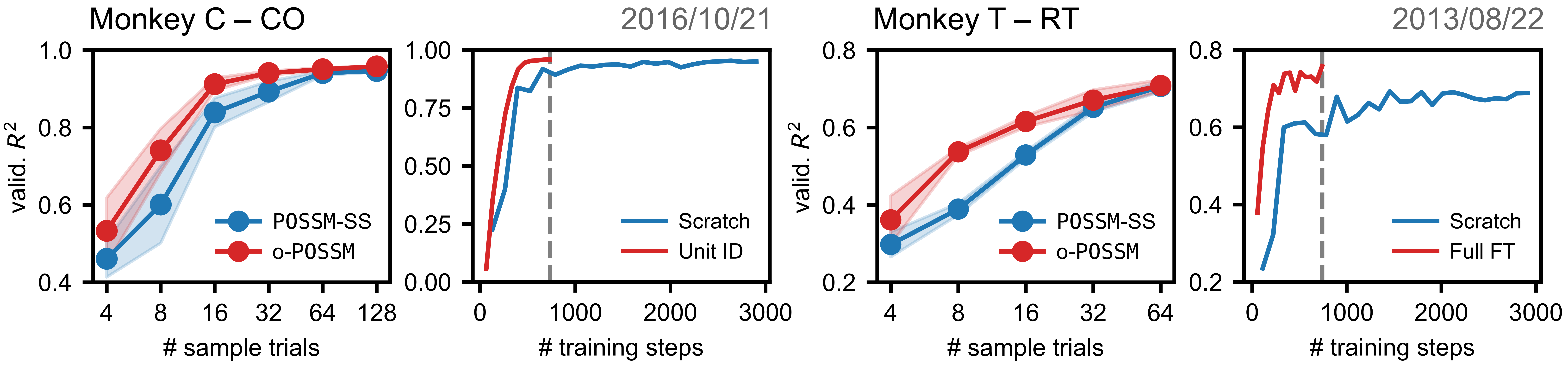
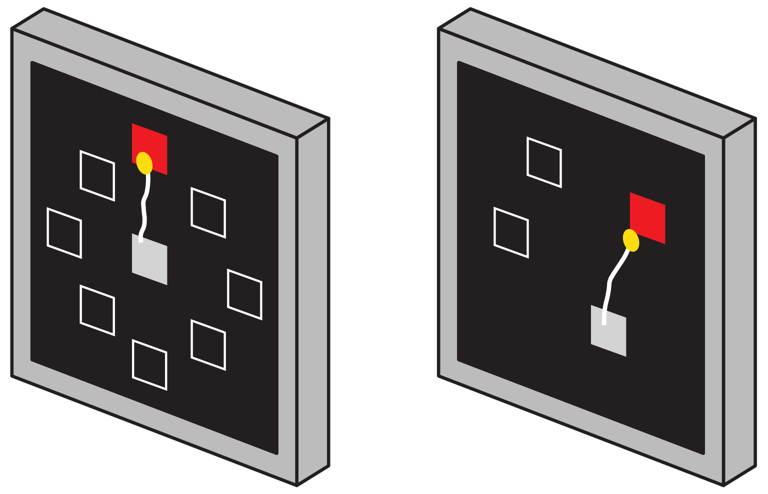


Figure: Demonstrating POSSM's data- and compute-efficiency during finetuning.

Computational efficiency



- Low latency at inference, comparable to GRU & MLP

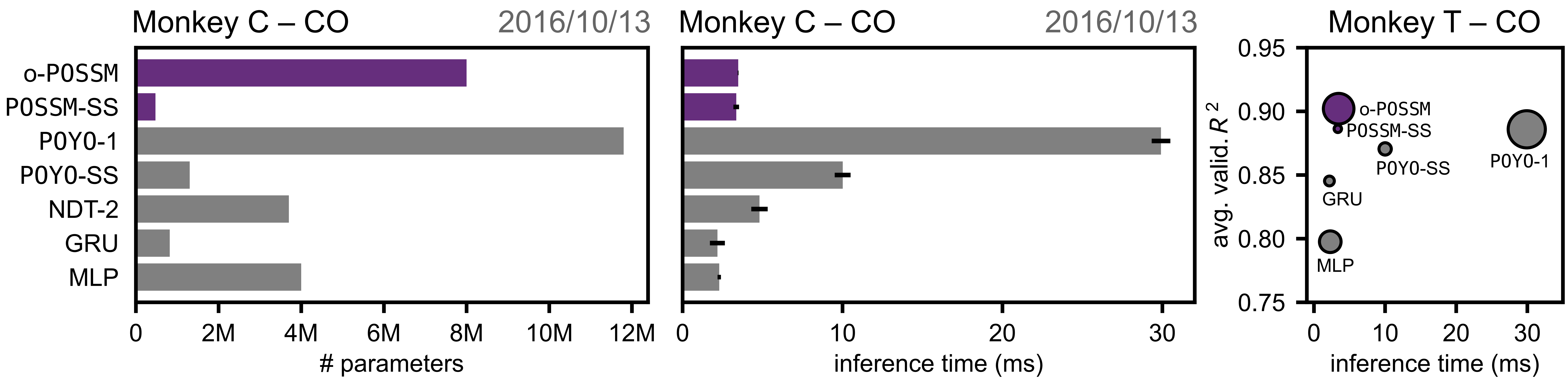
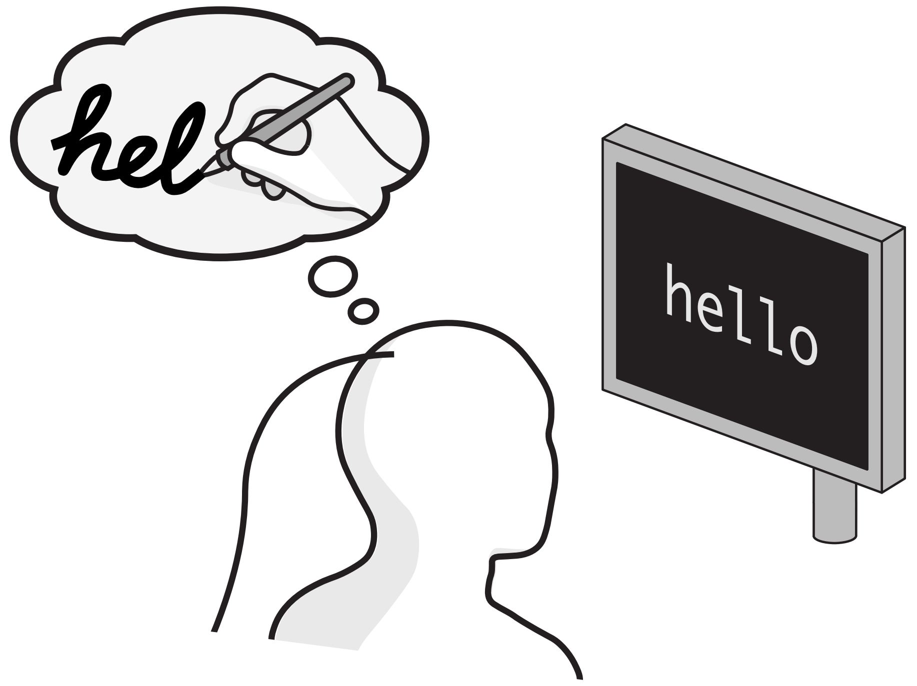


Figure: Demonstrating POSSM's efficiency at inference.

Human handwriting task



- Imagined handwriting character classification task
- Finetuning NHP-pretrained o-POSSM improves overall performance

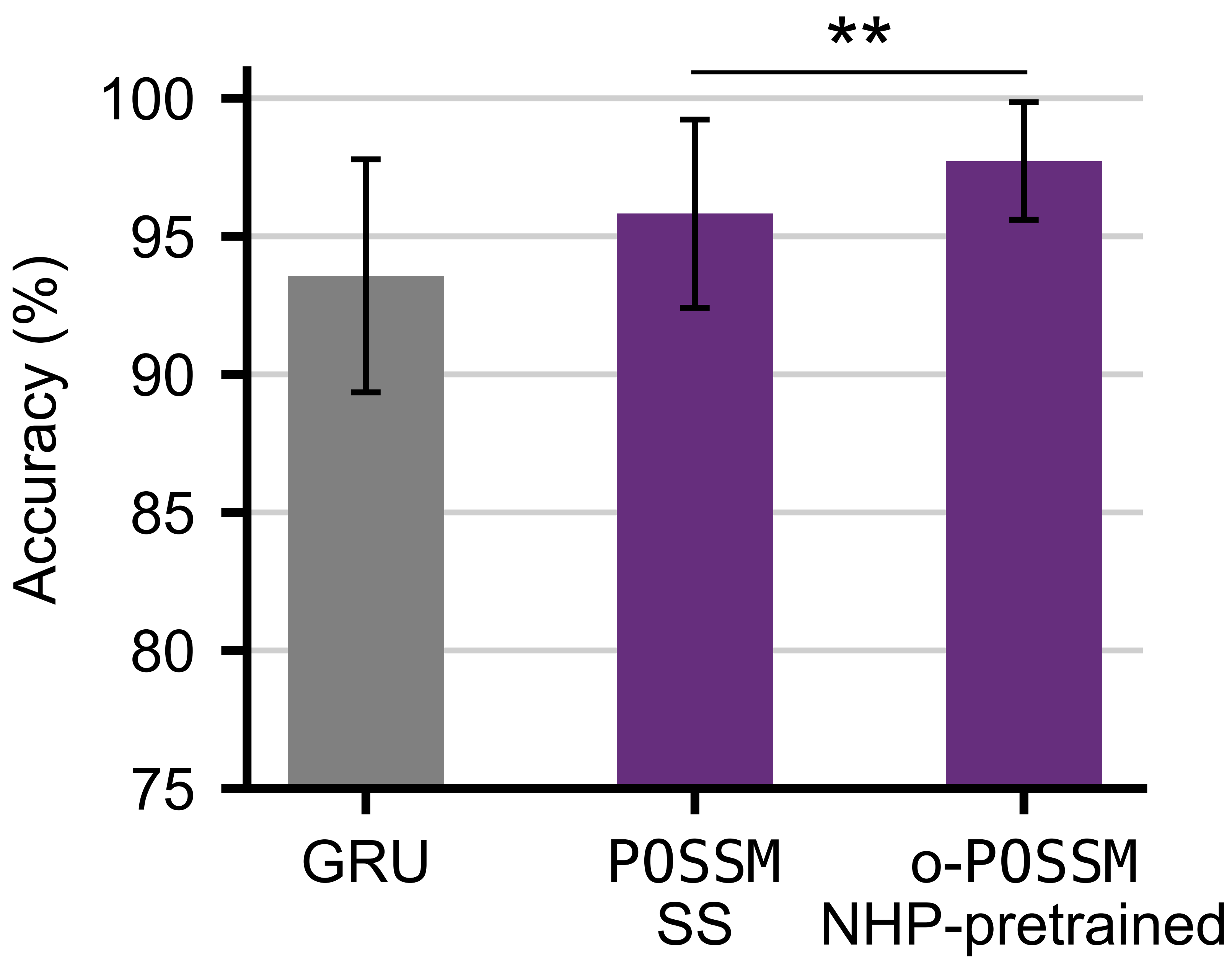
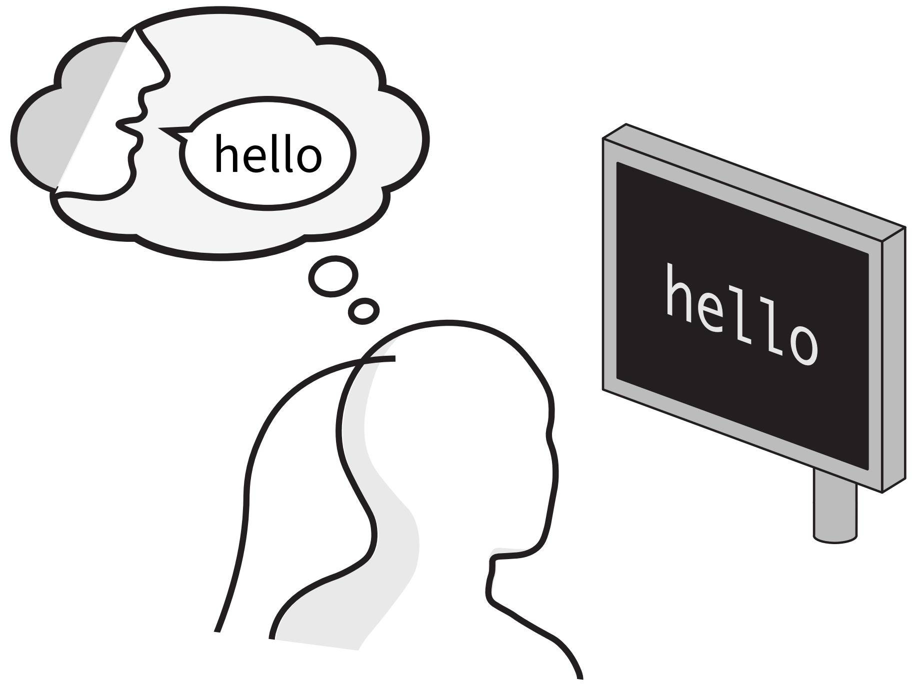


Figure: Decoding accuracy on human handwriting task. We report the mean $R^2 \pm$ SD over sessions.

Human speech task



- Long trials (2–16 seconds) with only sentence-level labels
- POSSM outperforms a strong GRU baseline
- Performance improves when using spiking-band powers

Model	Validation PER (%)
GRU (spikes only)	30.06
POSSM (spikes only)	**27.32**
GRU (spikes + powers)	21.74
POSSM (spikes + powers)	**19.80**

Conclusion

- POSSM is highly performant on several neural decoding tasks
- We show the benefits of pretraining in enabling efficient generalization
- Even with increased model sizes, inference times are low
- Results indicate positive cross-species transfer (NHPs to humans)
- POSSM also excels at tasks involving long trials, such as speech decoding
- Future directions:
  - Incorporating self-supervised learning objectives
  - Processing multiple neural recording modalities