Model Improvements from Reverse-Engineering Natural Intelligence

Aran Nayebi

Google Foundation Models Workshop

Mountain View, CA

2024.11.15

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Outline

- Role of Recurrent Processing During Object Recognition
- Visually-Grounded Mental Simulation

Outline

- Role of Recurrent Processing During Object Recognition
- Visually-Grounded Mental Simulation

Role of Recurrent Processing During Object Recognition

A. Nayebi*, D. Bear*, J. Kubilius*, *et al.*
Task-Driven Convolutional Recurrent Models of the Visual System. *NeurIPS 2018*

A. Nayebi, *et al.*
Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition.
Neural Computation 2022

Daniel Yamins

Daniel Bear

Jonas Kubilius

Kohitij Kar

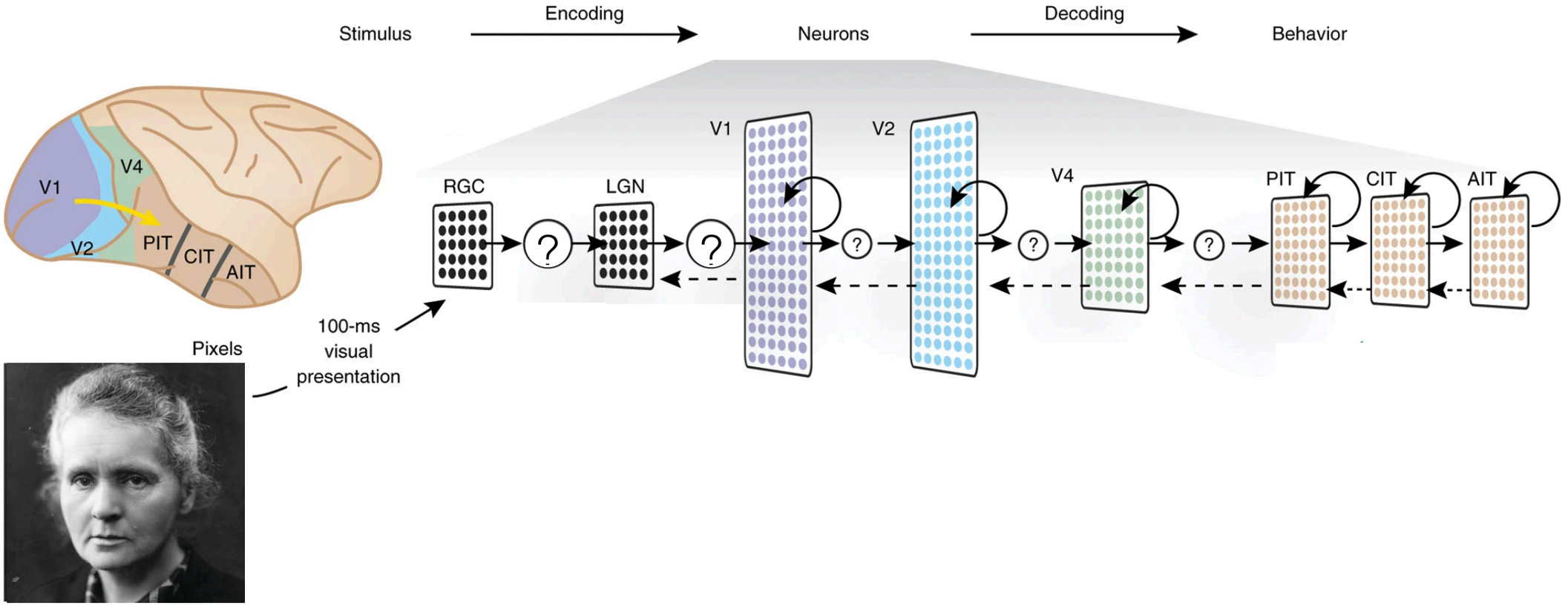
Surya Ganguli

Javier Sagastuy

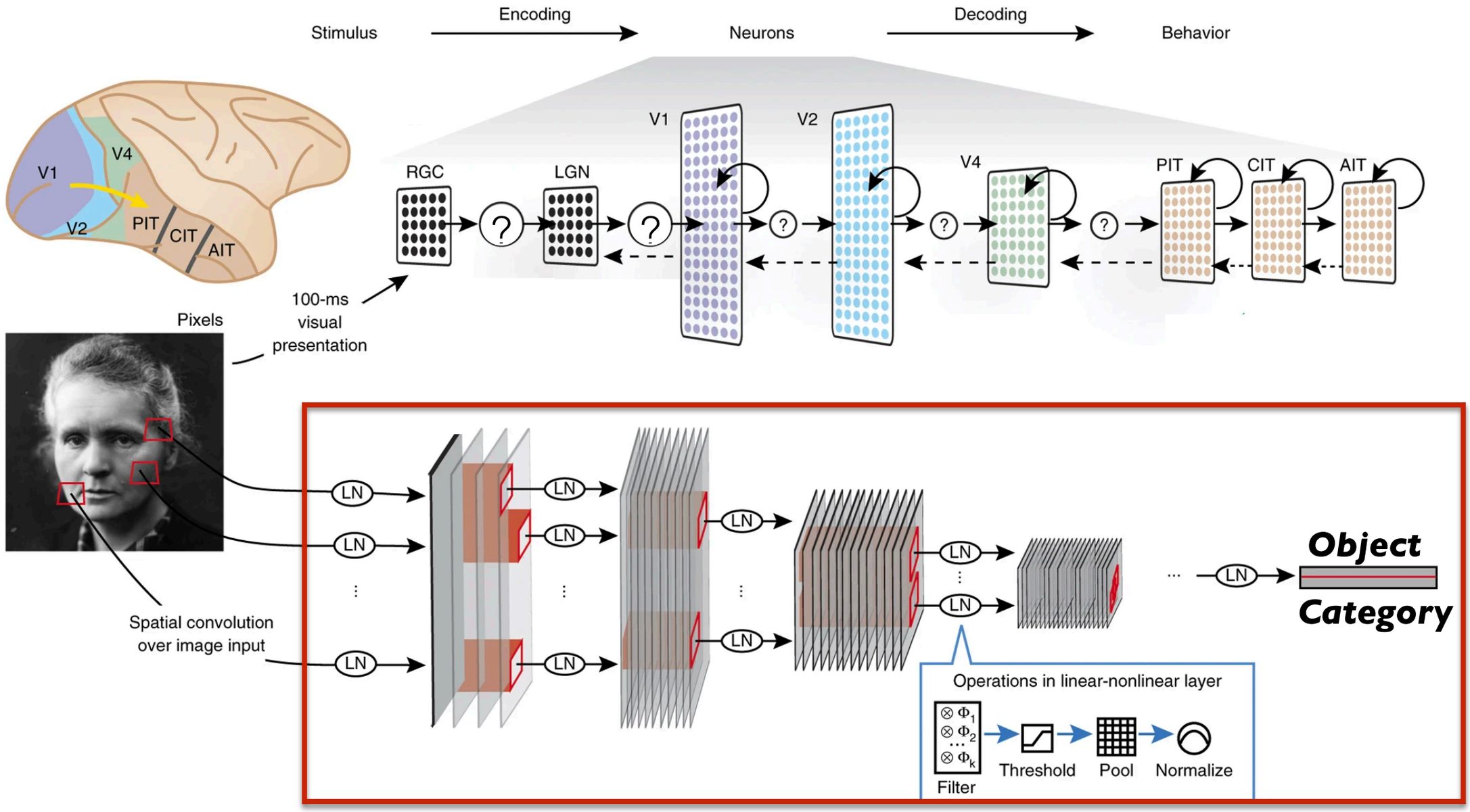
David Sussillo

Jim DiCarlo

Primate Ventral Stream Implements Object Recognition



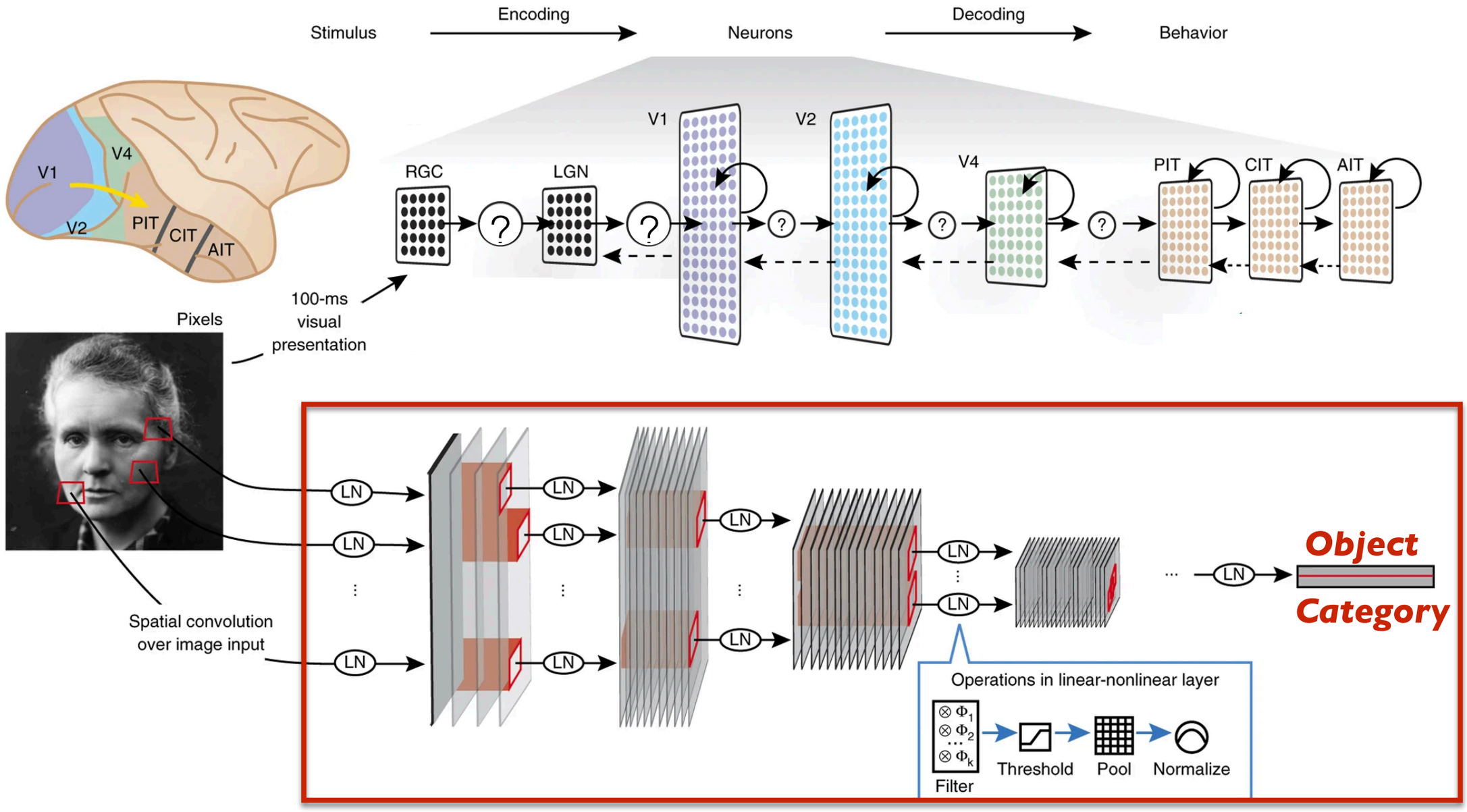
CNNs as Models of Primate Object Recognition



CNNs are inspired by visual neuroscience:

1) **hierarchy**
2) **retinotopy** (spatially tiled)

CNNs as Models of Primate Object Recognition

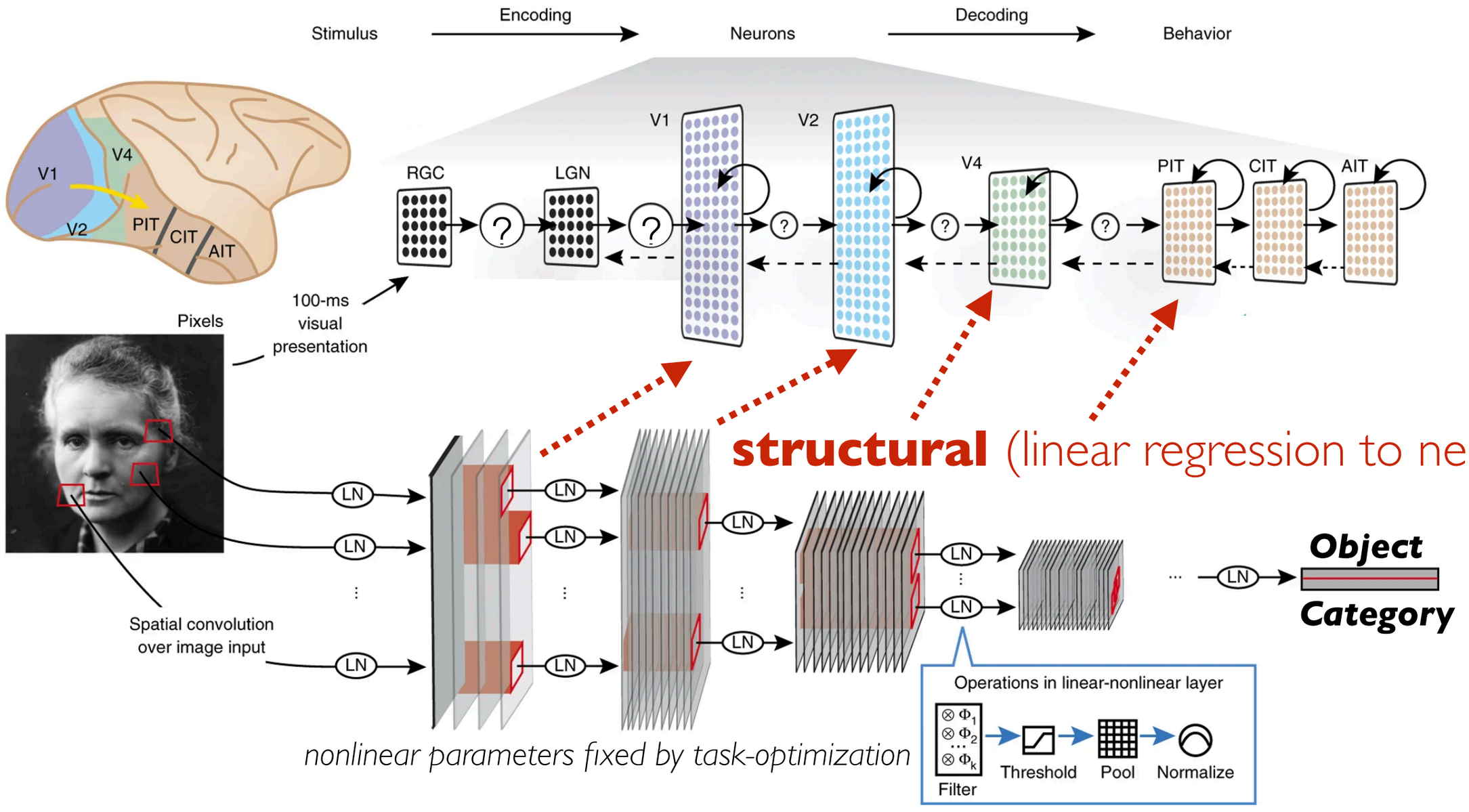


CNNs are inspired by visual neuroscience:

1) **hierarchy**
2) **retinotopy** (spatially tiled)

functional (performs behavior)

CNNs as Models of Primate Object Recognition



structural (linear regression to neurons)

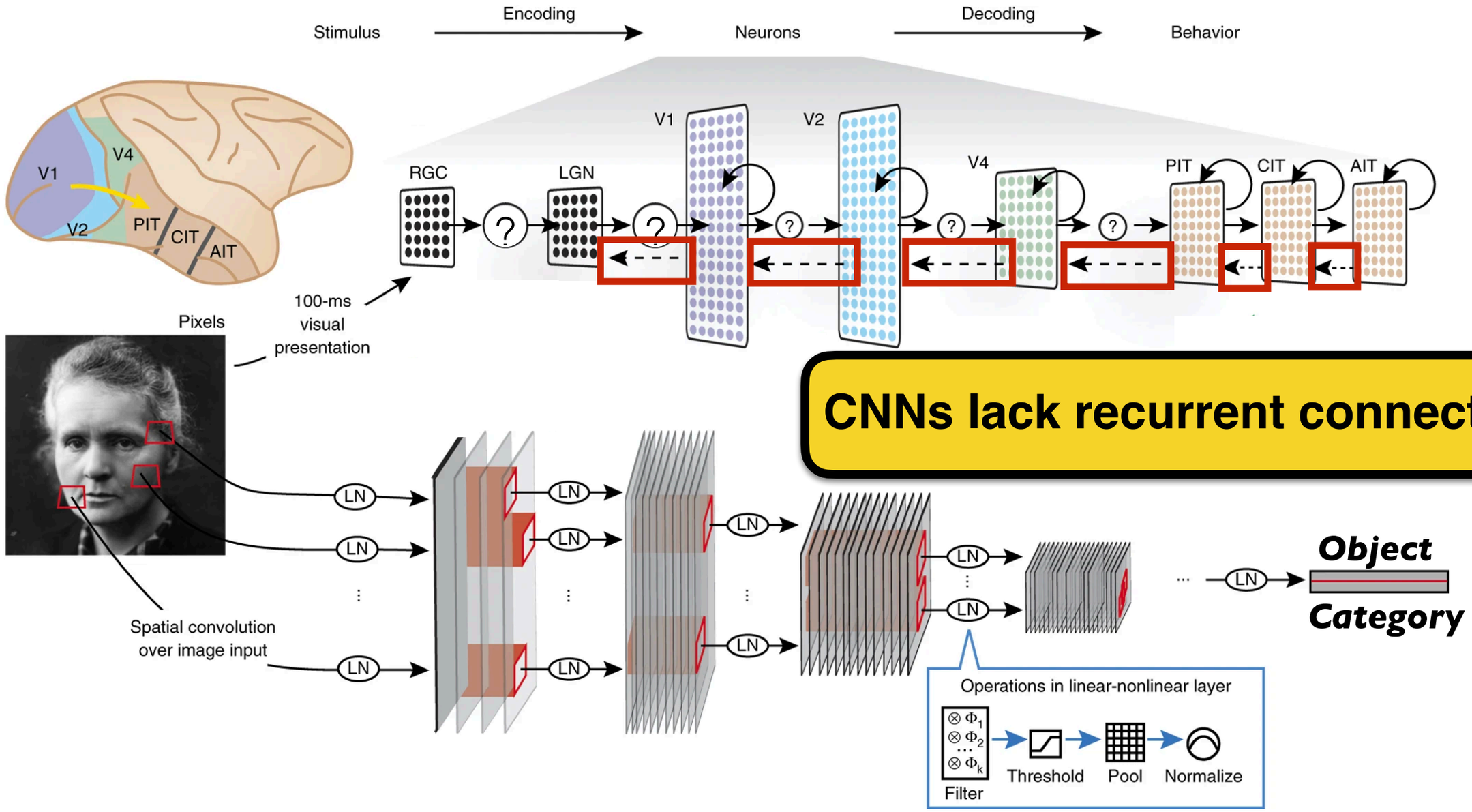
nonlinear parameters fixed by task-optimization

functional (performs behavior)

CNNs are inspired by visual neuroscience:

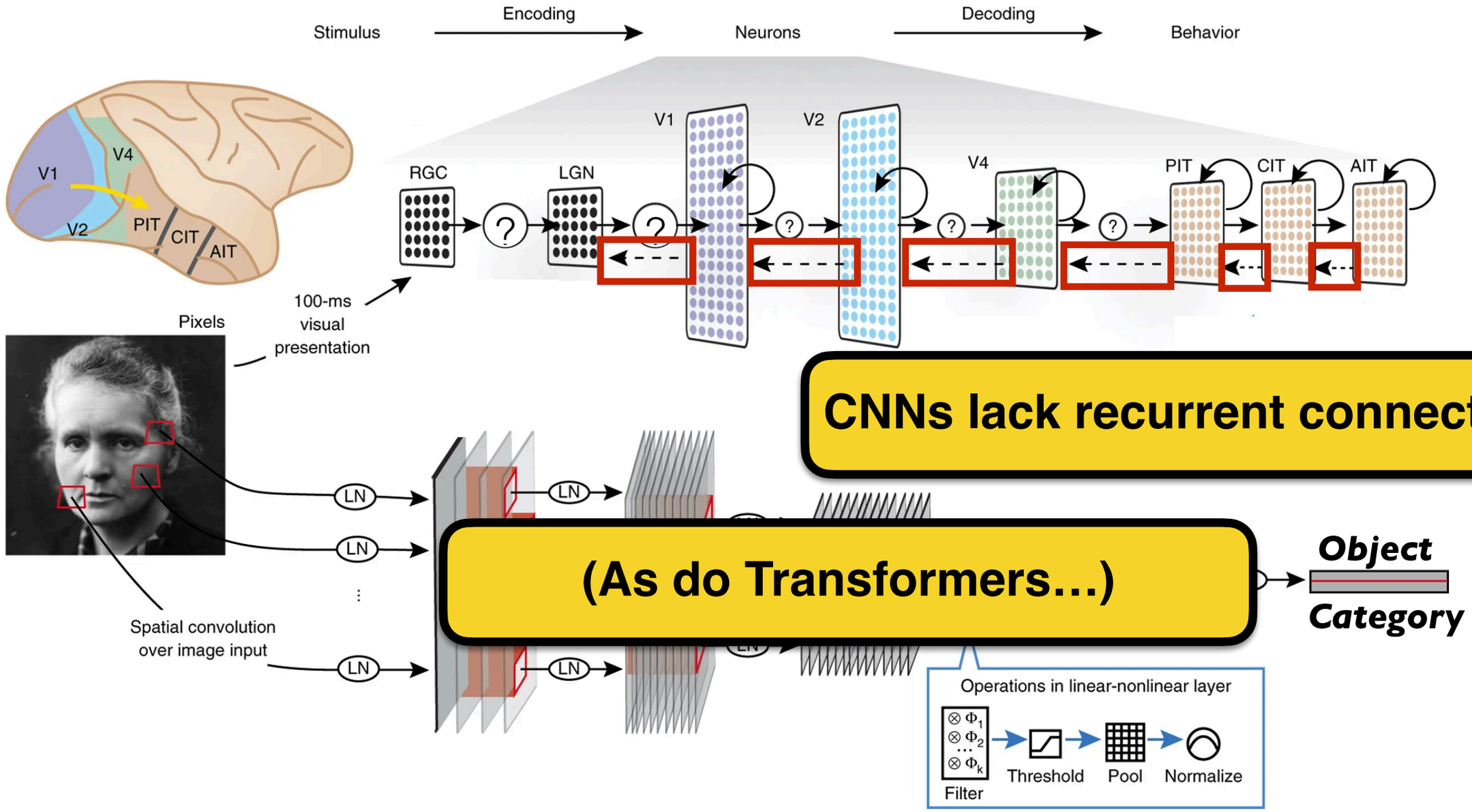
1) **hierarchy**
2) **retinotopy** (spatially tiled)

CNNs as Models of Primate Object Recognition

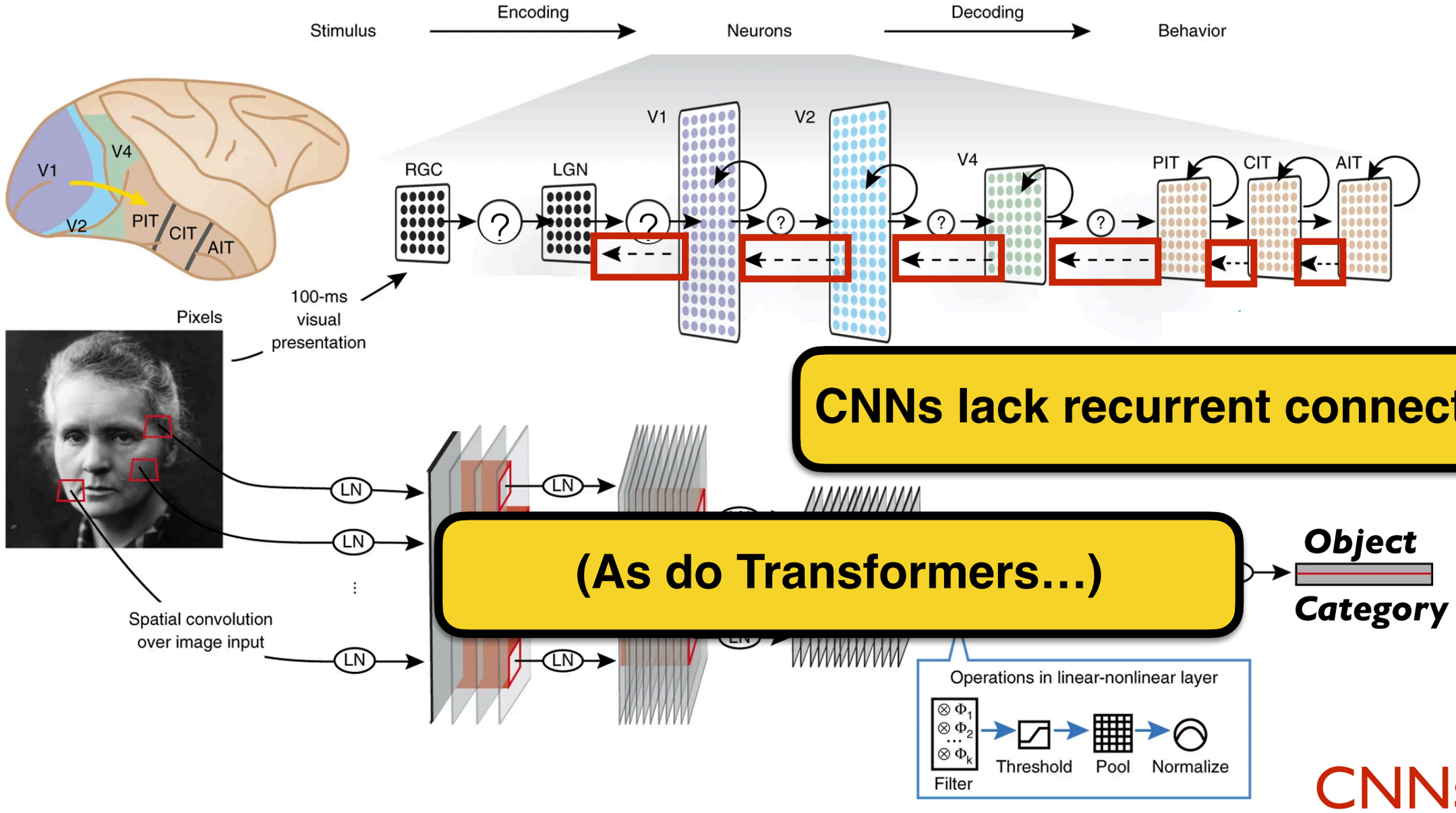


CNNs lack recurrent connections!

CNNs as Models of Primate Object Recognition



CNNs as Models of Primate Object Recognition



CNNs lack recurrent connections!

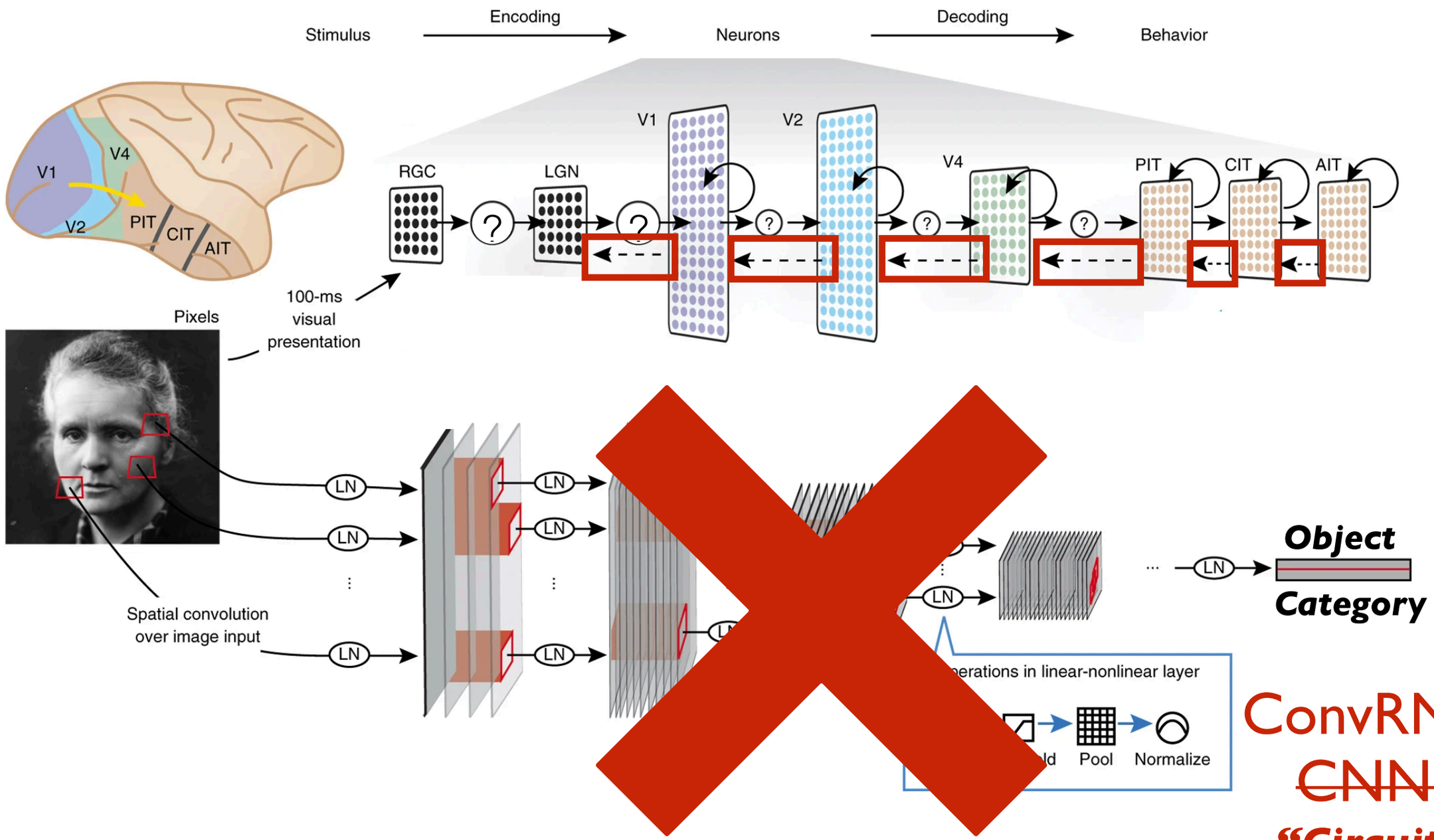
(As do Transformers…)

CNNs

"Circuit"

A = *architecture class*

Convolutional Recurrent Networks (ConvRNNs)



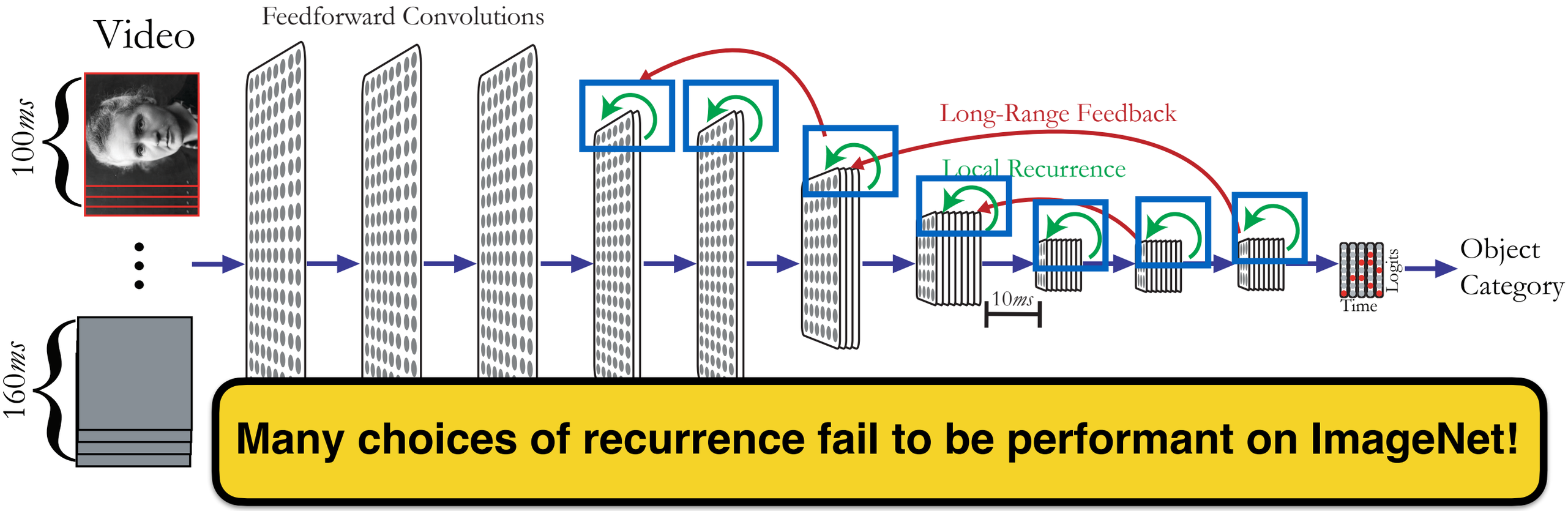
ConvRNNs

~~CNNs~~

"Circuit"

A = *architecture class*

Convolutional Recurrent Networks (ConvRNNs)



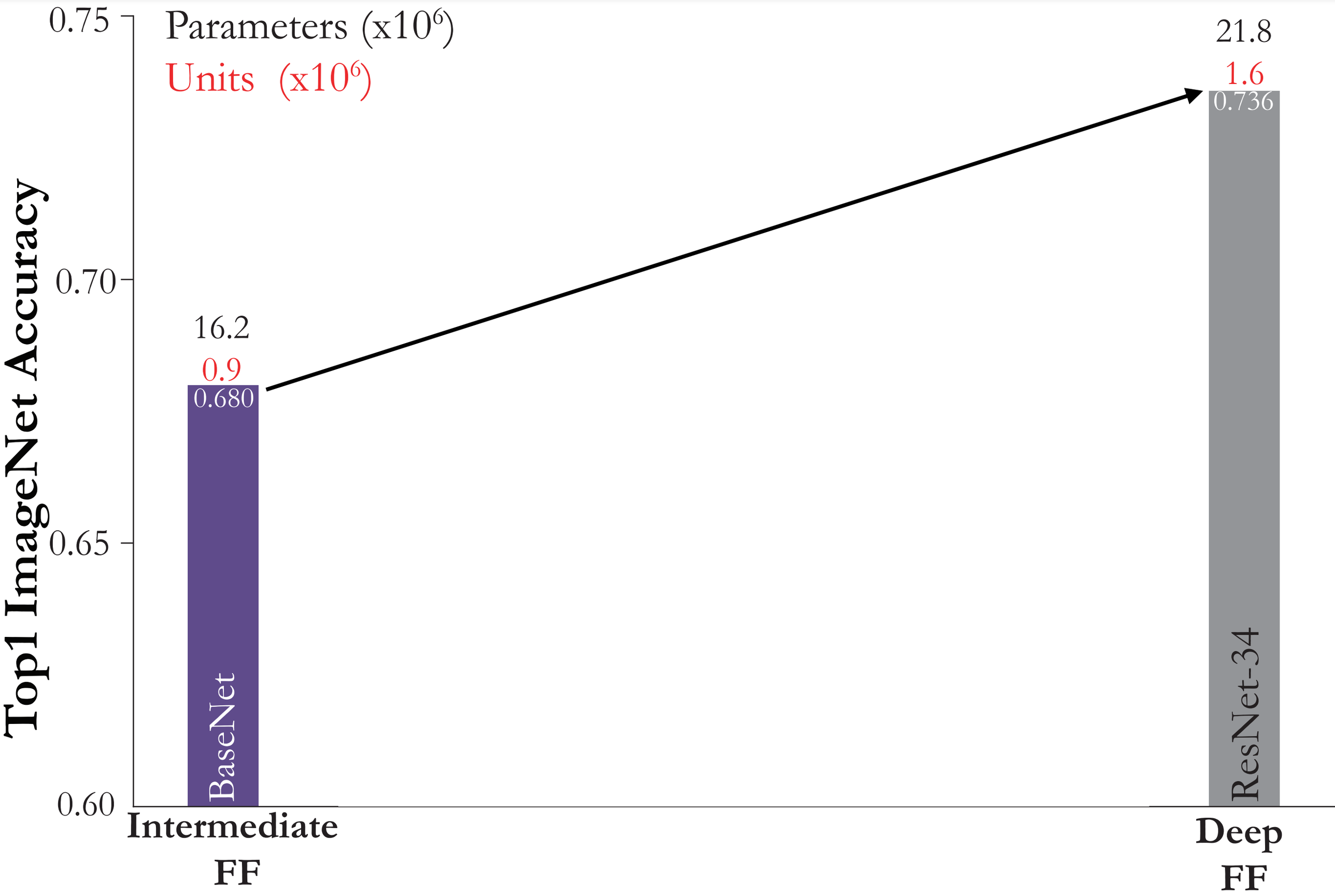
Many choices of recurrence fail to be performant on ImageNet!

Each time-step (10 *ms*) is treated equally
— including feedforward steps

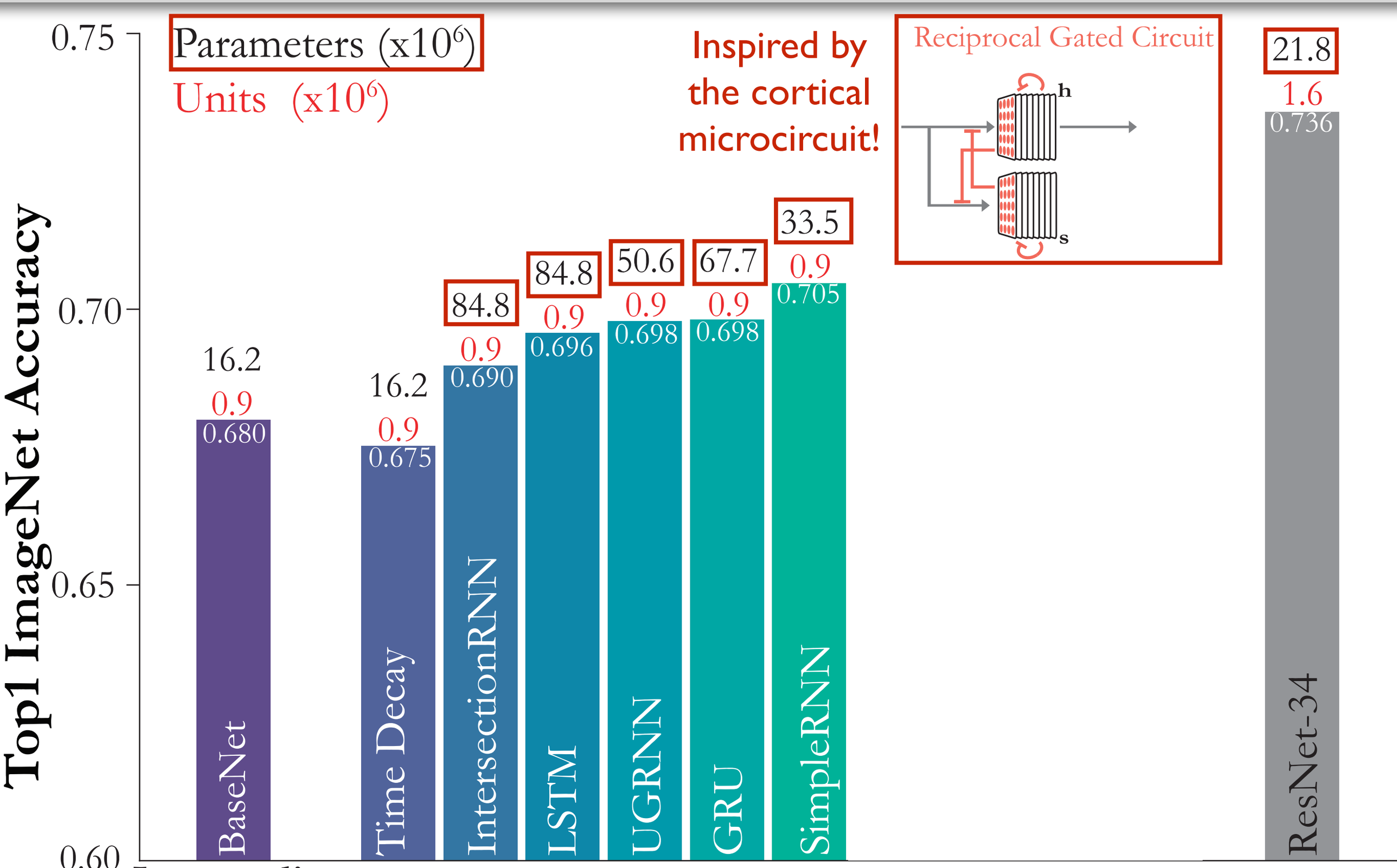
ConvRNNs
~~CNNs~~
"Circuit"

A = *architecture class*

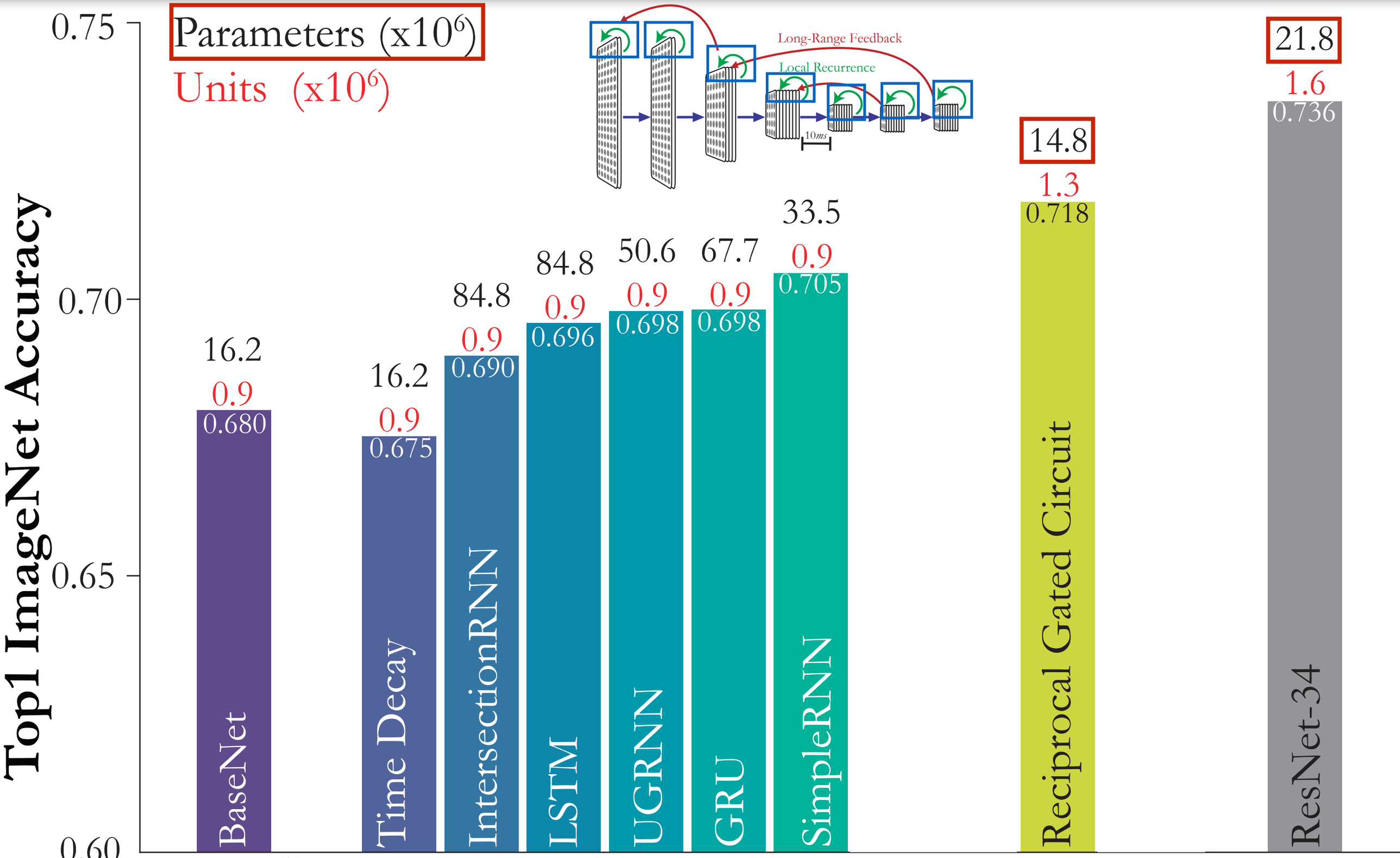
Implanting Local Recurrence into Feedforward CNNs



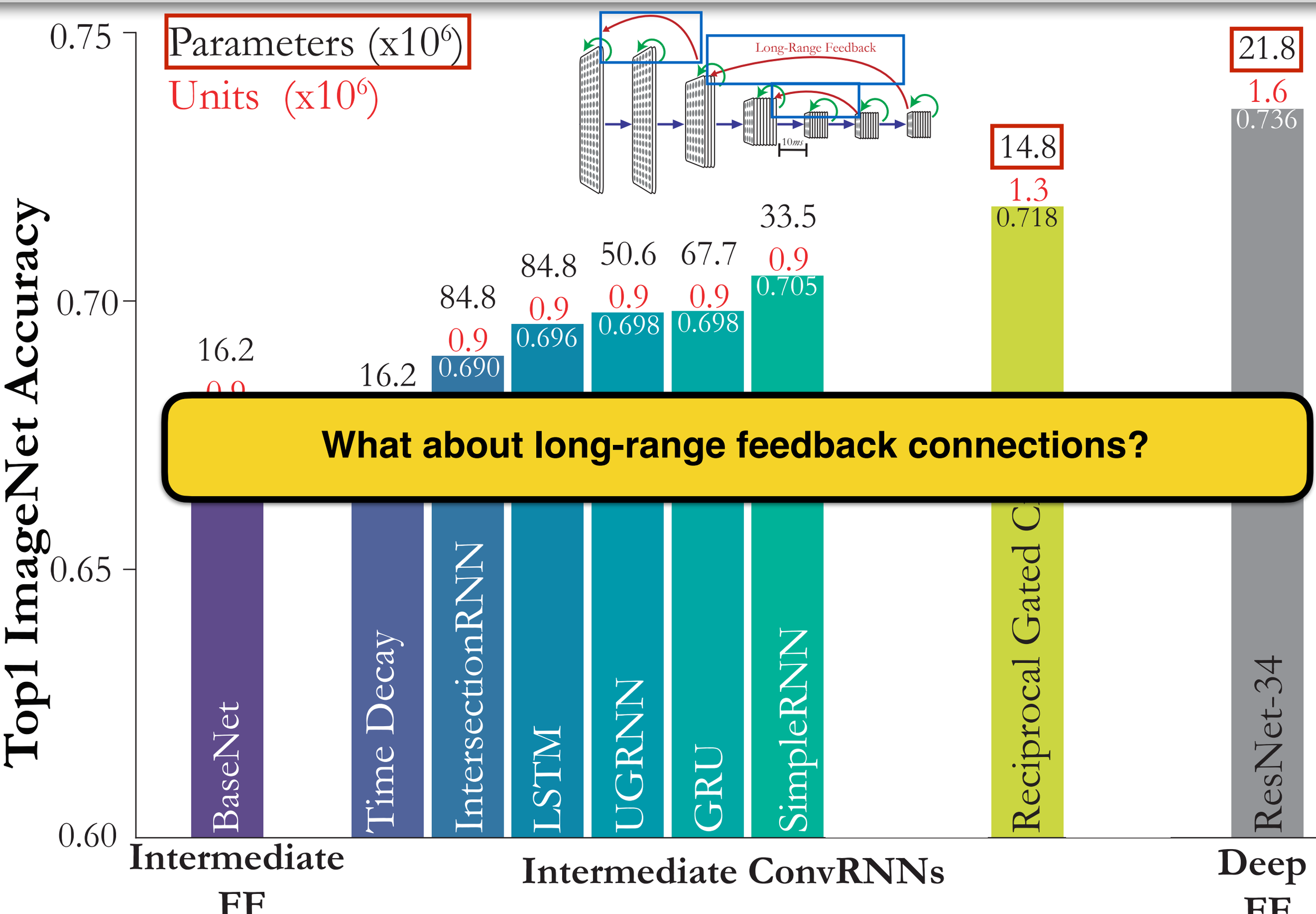
Adding Standard RNNs Helps Incrementally, but Add **Lots** of Parameters!



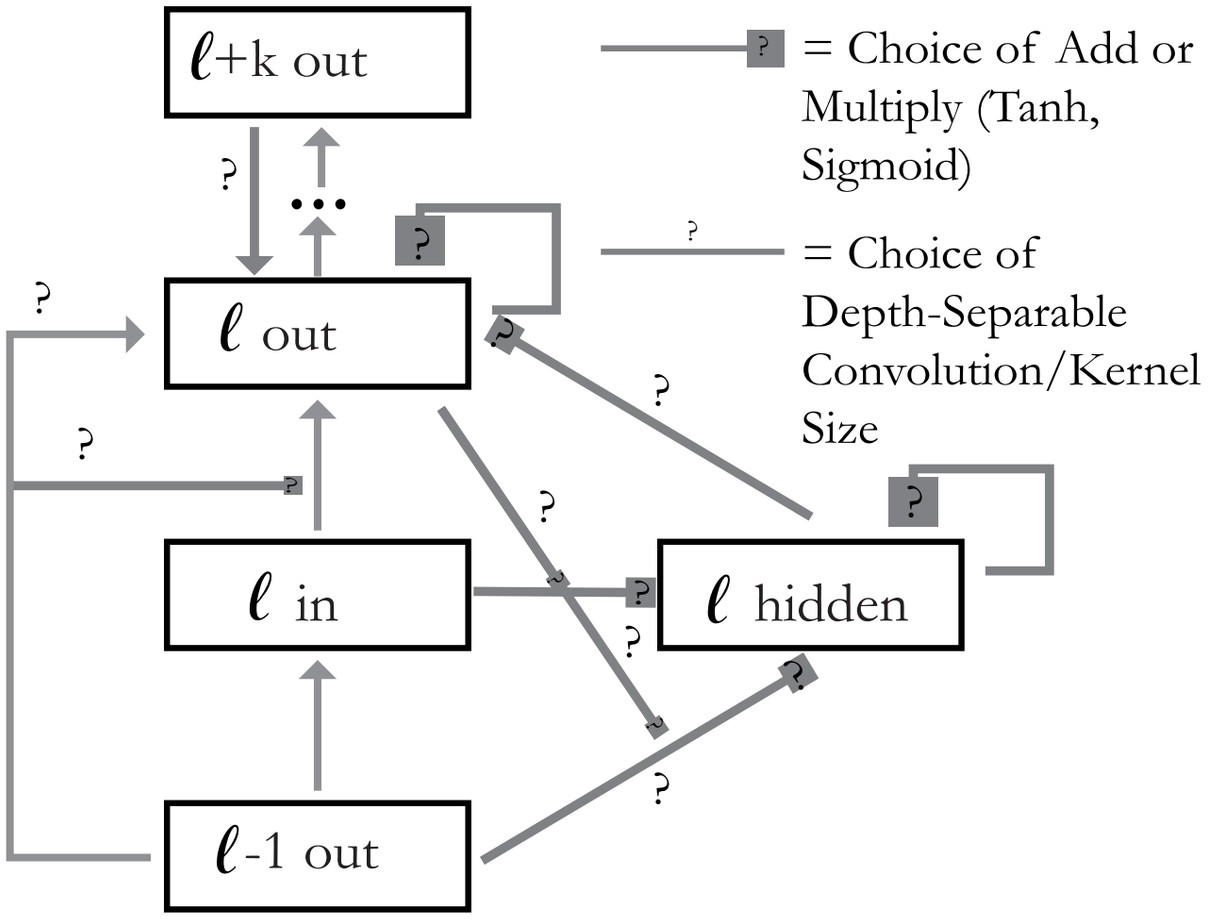
Novel Recurrent Cells Yield Improved ImageNet Performance

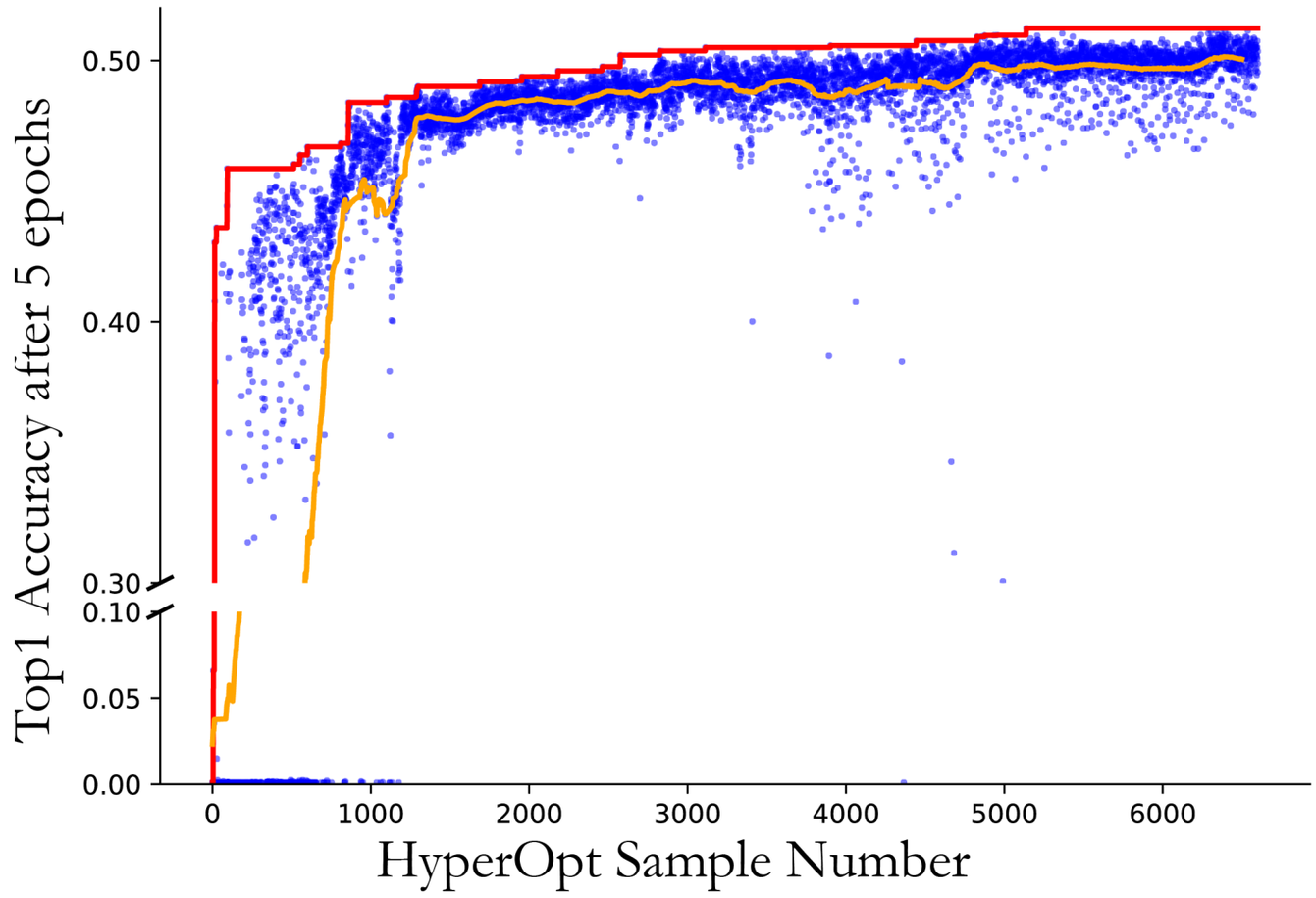


Novel Recurrent Cells Yield Improved ImageNet Performance

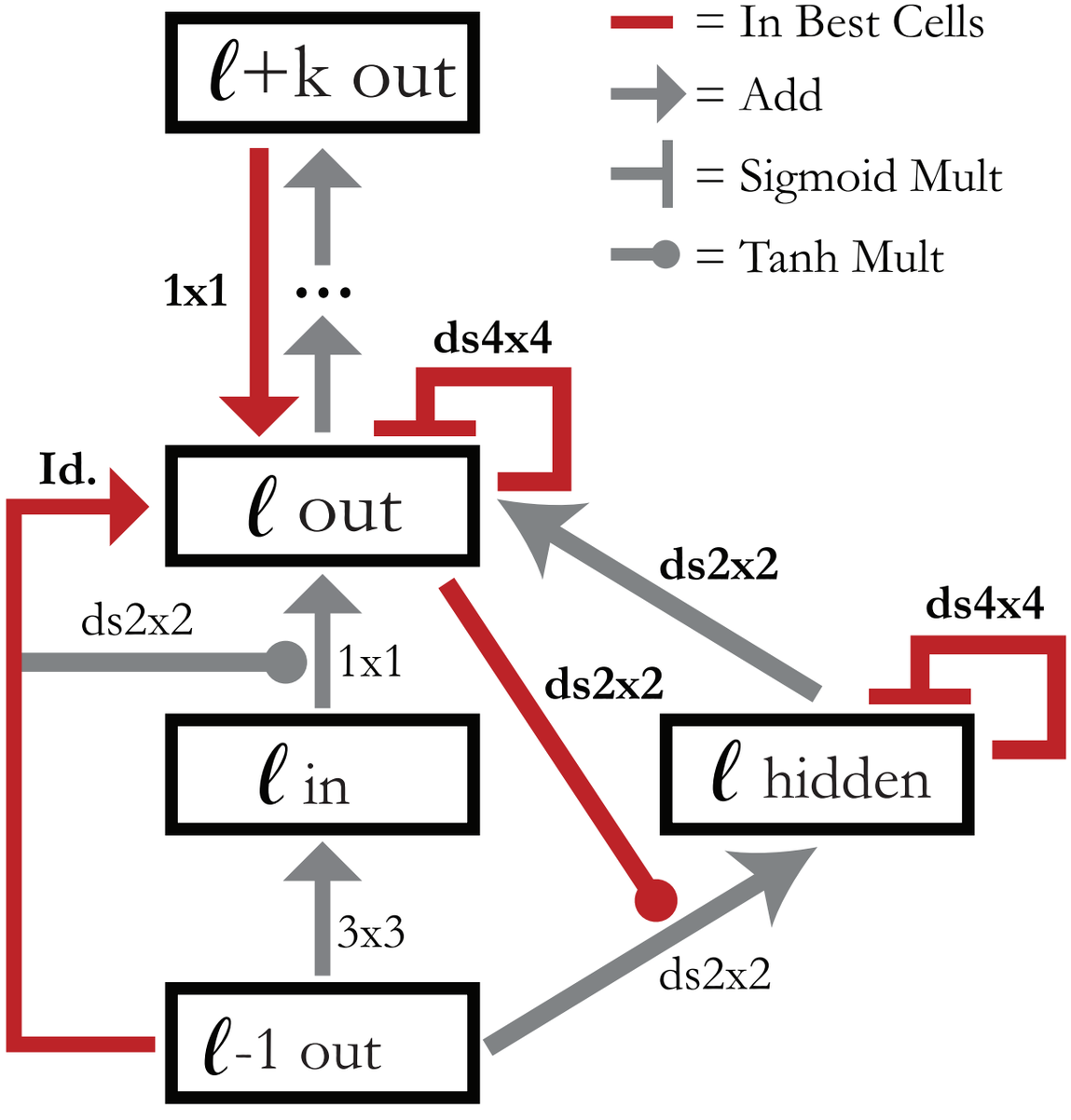


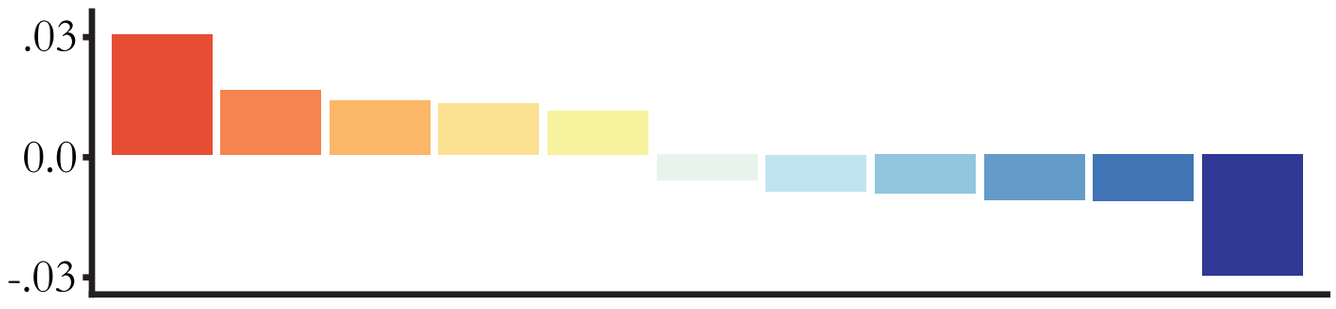
Large-Scale Search Over Long-Range Feedback Connections





Emergent Global Connectivity Patterns

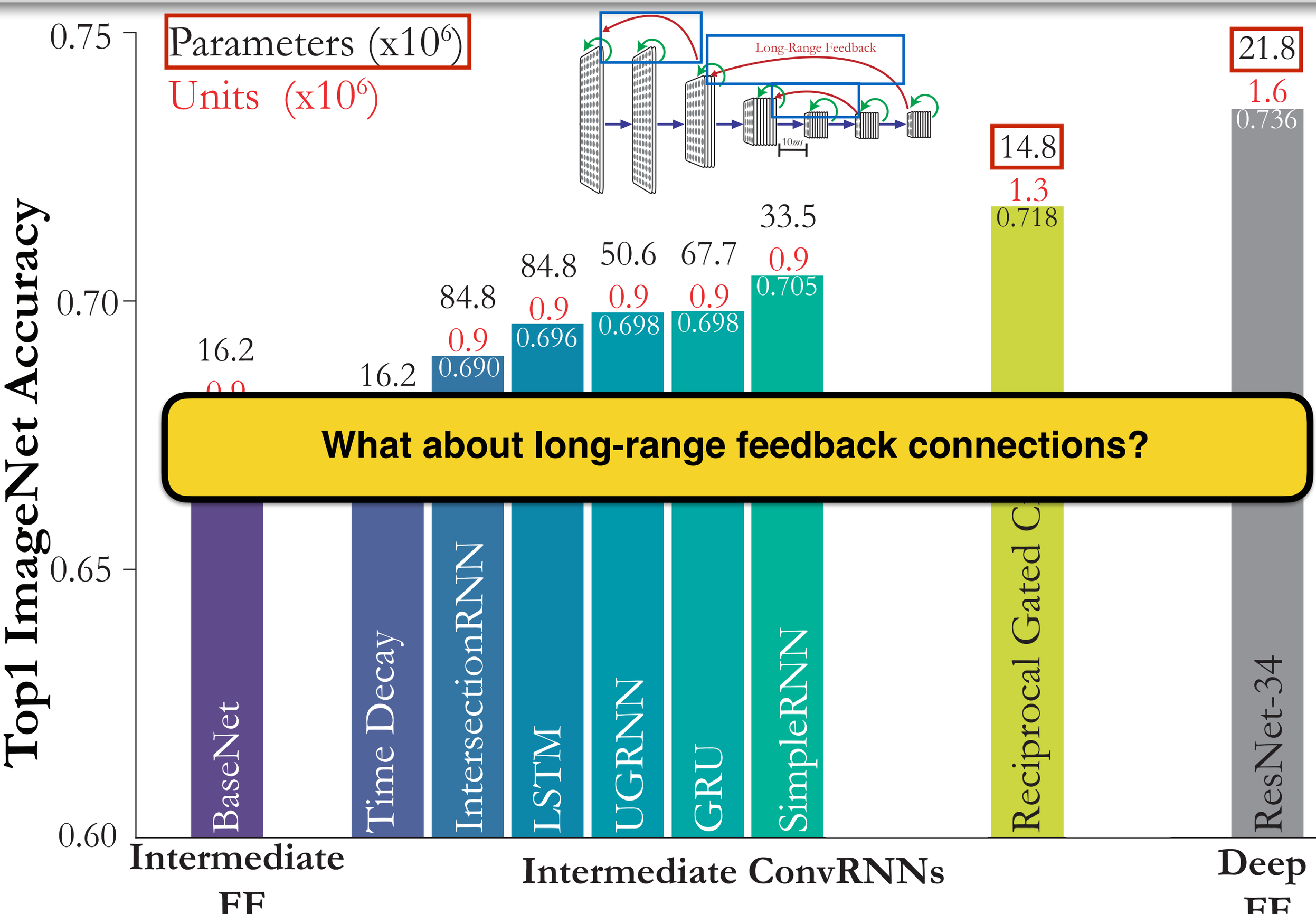




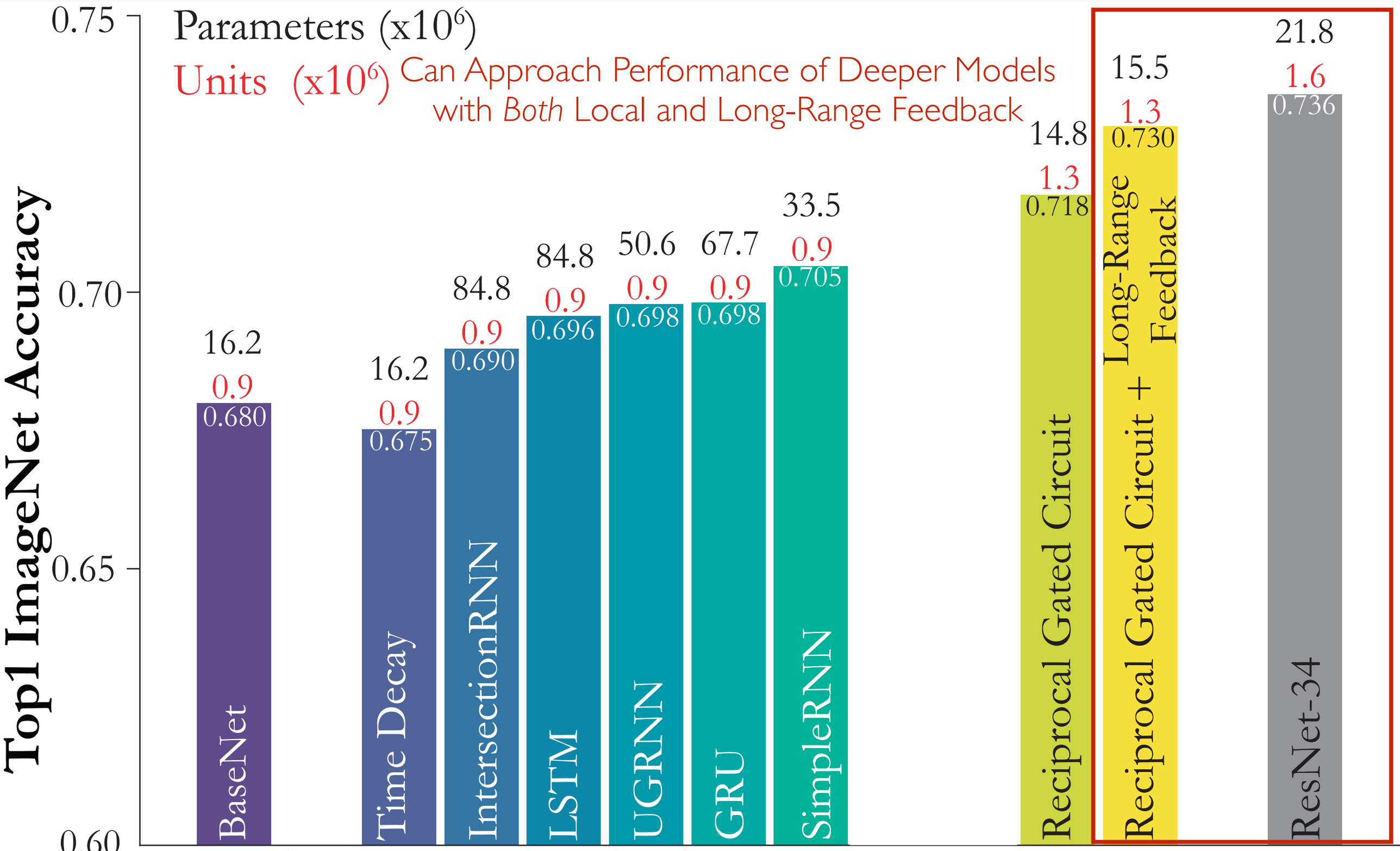
Top1 Gain from Feedback (Best to Worst)

Conservation on parameter count as a byproduct of evolutionary optimization

Novel Recurrent Cells Yield Improved ImageNet Performance



Long-Range Feedback Connections Matter



Outline

- Role of Recurrent Processing During Object Recognition

Enables more parameter/unit efficient models that gain object recognition performance by unrolling "deeper" in time, rather than adding more layers.

Moreso than simply "convolutionizing" standard LSTMs/GRUs.

- Visually-Grounded Mental Simulation

Outline

- Role of Recurrent Processing During Object Recognition

Enables more parameter/unit efficient models that gain object recognition performance by unrolling "deeper" in time, rather than adding more layers.

Moreso than simply "convolutionizing" standard LSTMs/GRUs.

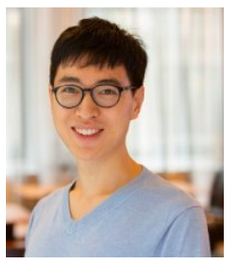
- Visually-Grounded Mental Simulation

Visually-Grounded Mental Simulation

A. Nayebi, R. Rajalingham, M. Jazayeri, G.R. Yang
Neural foundations of mental simulation: future prediction of latent representations on dynamic scenes.
NeurIPS 2023 (spotlight)


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

Visually-Grounded Mental Simulation

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

Plan:
How would I take these hats off the rack?

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

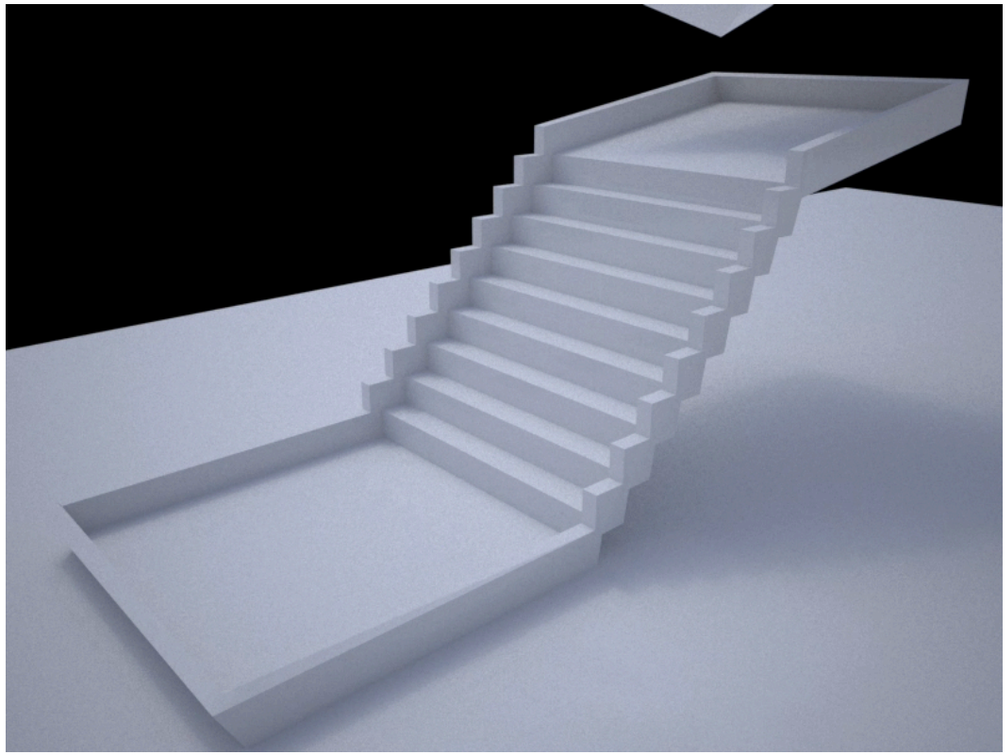
Infer:
Has this ice block been out longer?

Plan:
How would I take these hats off the rack?

Predict:
Will this box support me?

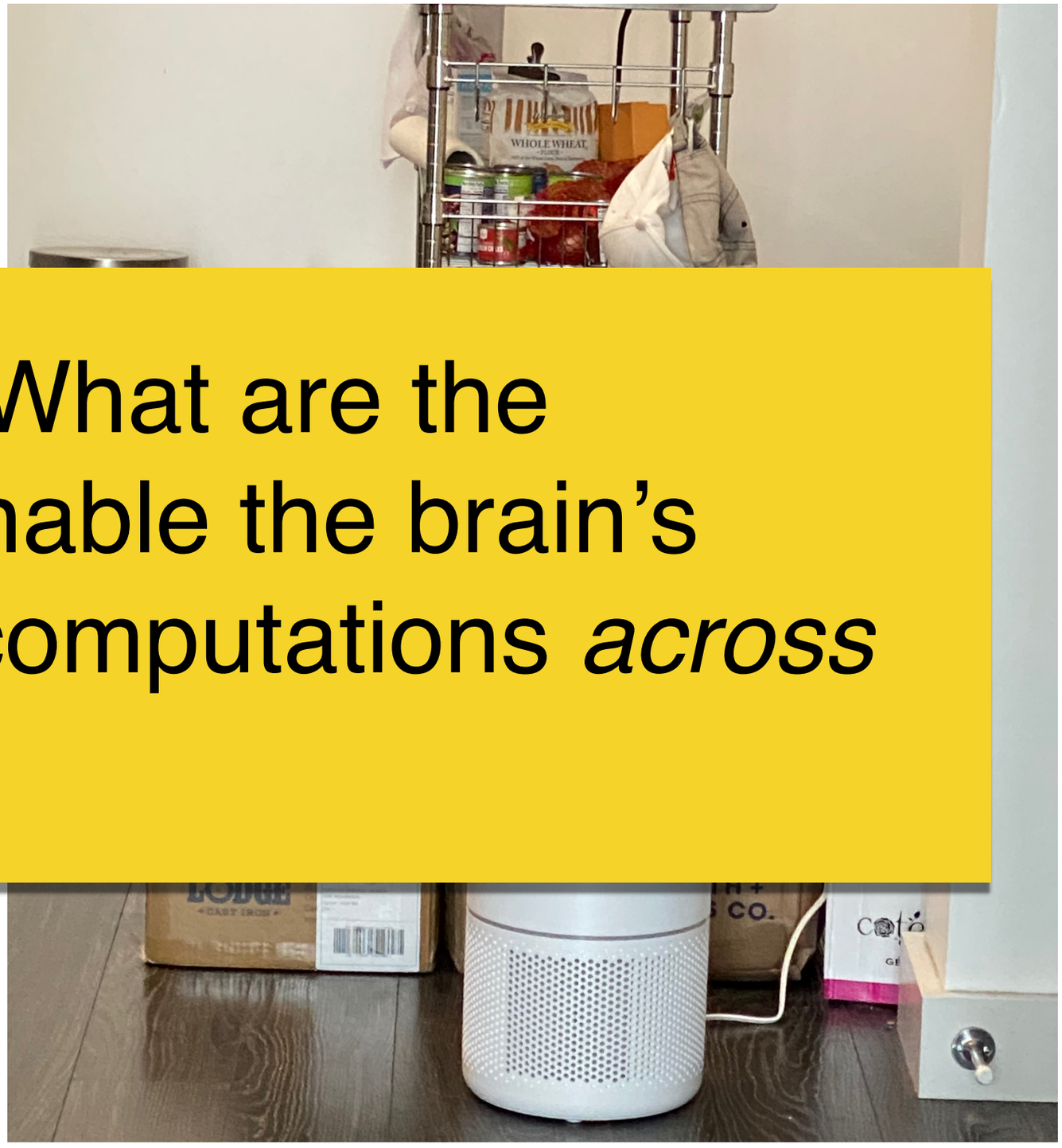
Visually-Grounded Mental Simulation

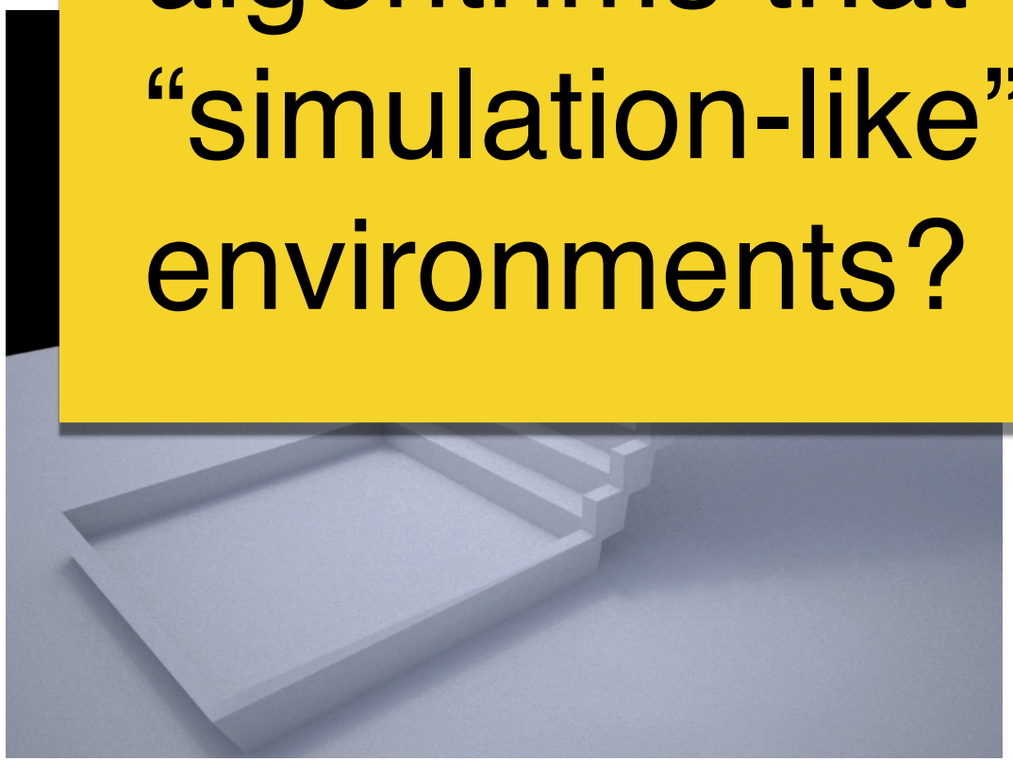
Infer:
Has this ice block been out longer?



Plan:
How would I take these hats off the rack?

Predict:
Will this box support me?





Crux Question: What are the algorithms that enable the brain's "simulation-like" computations *across* environments?

Defining Hypotheses

Defining Hypotheses

"Sensory-Cognitive Networks"

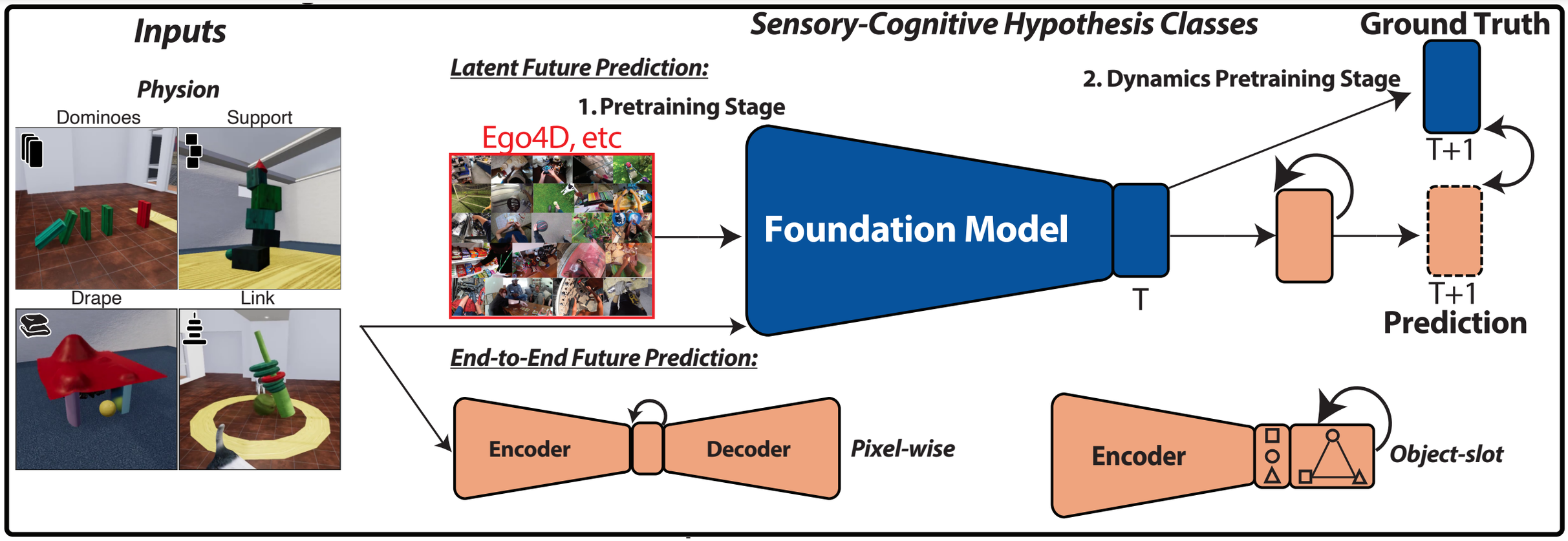
R1 (Input-Driven): Take in unstructured visual inputs across a range of physical phenomena.

R2 (Behavioral Outputs): Generate physical predictions for each scenario ("behavior").

R3 (Neural Representations): Consist of internal units that can be compared to biological units (e.g. containing "artificial neurons").

Overall Approach: Sensory-Cognitive Hypotheses

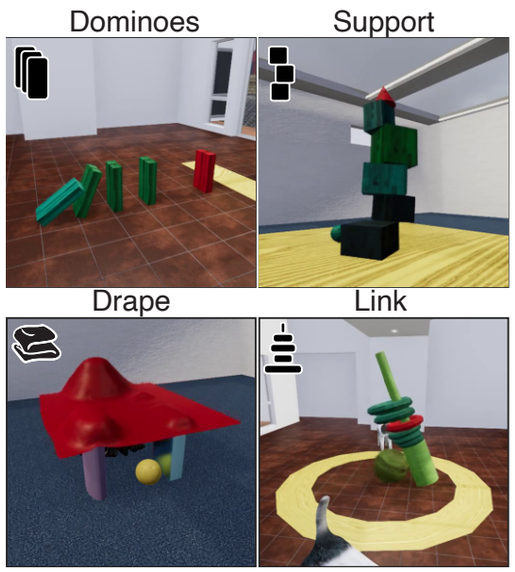
Overall Approach: Sensory-Cognitive Hypotheses



Overall Approach: Sensory-Cognitive Hypotheses

Inputs

Physion



Sensory-Cognitive Hypothesis Classes

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

T

2. Dynamics Pretraining Stage

T+1

T+1

Prediction

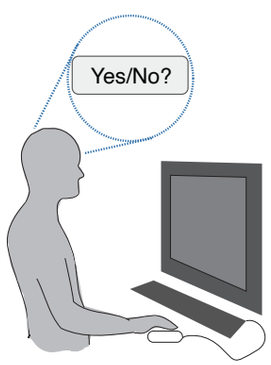
Ground Truth

End-to-End Future Prediction:

Encoder **Decoder** ***Pixel-wise***

Encoder ***Object-slot***

Human Behavior: Physion Object Contact Prediction (OCP)



Observed Stimuli

Unobserved Outcome

Time

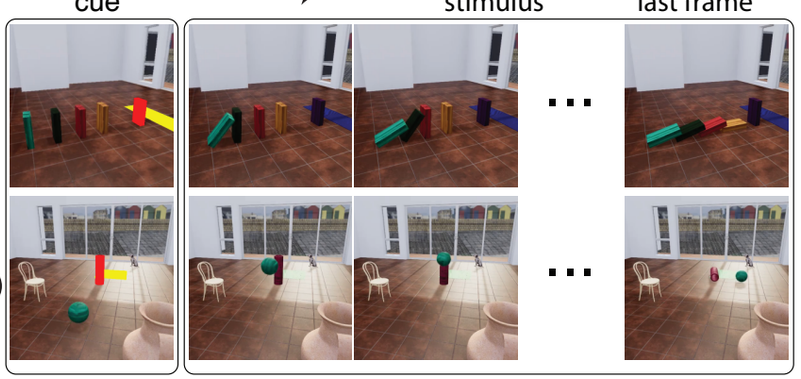
cue

stimulus

last frame

true label

Example Scenarios



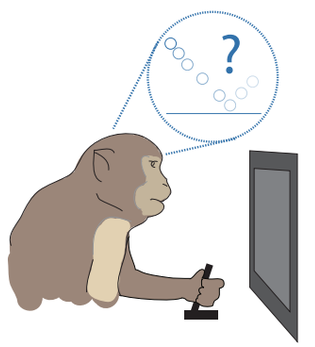
NO

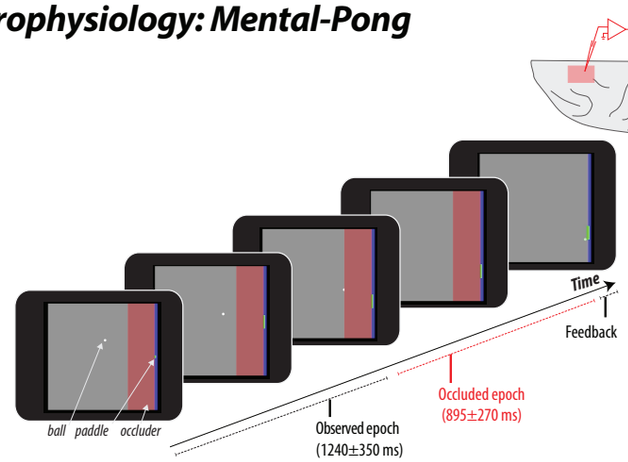
acc. = 0.89

YES

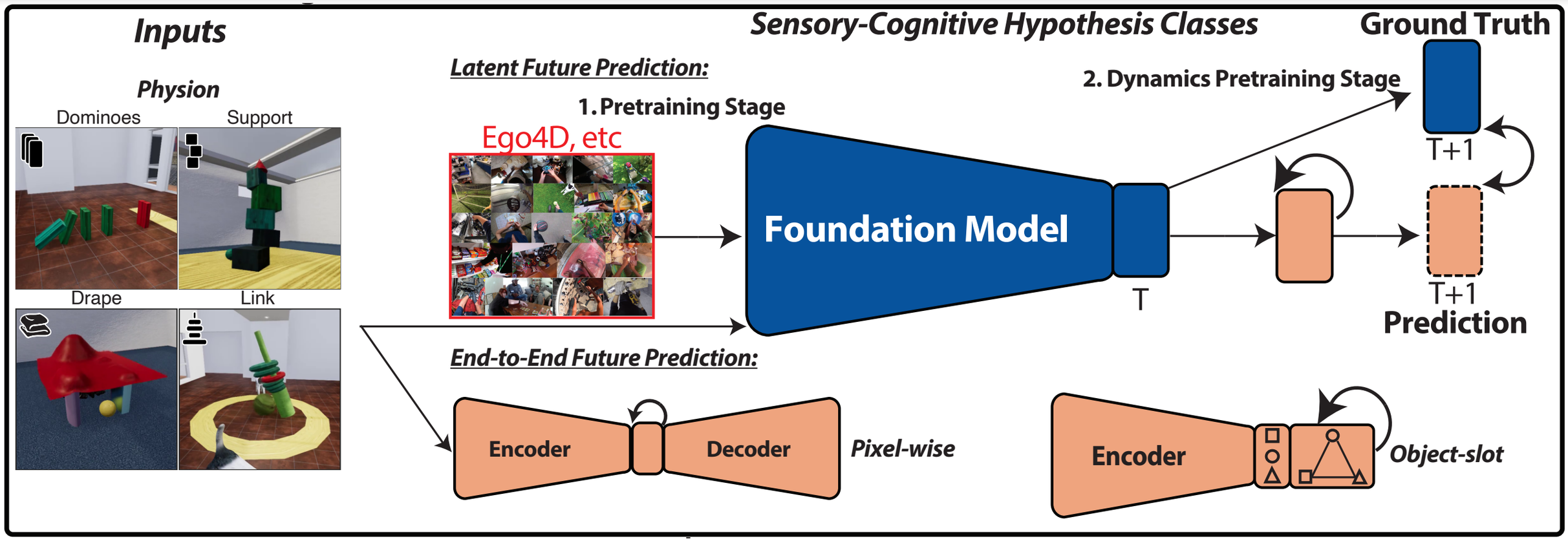
acc. = 0.96

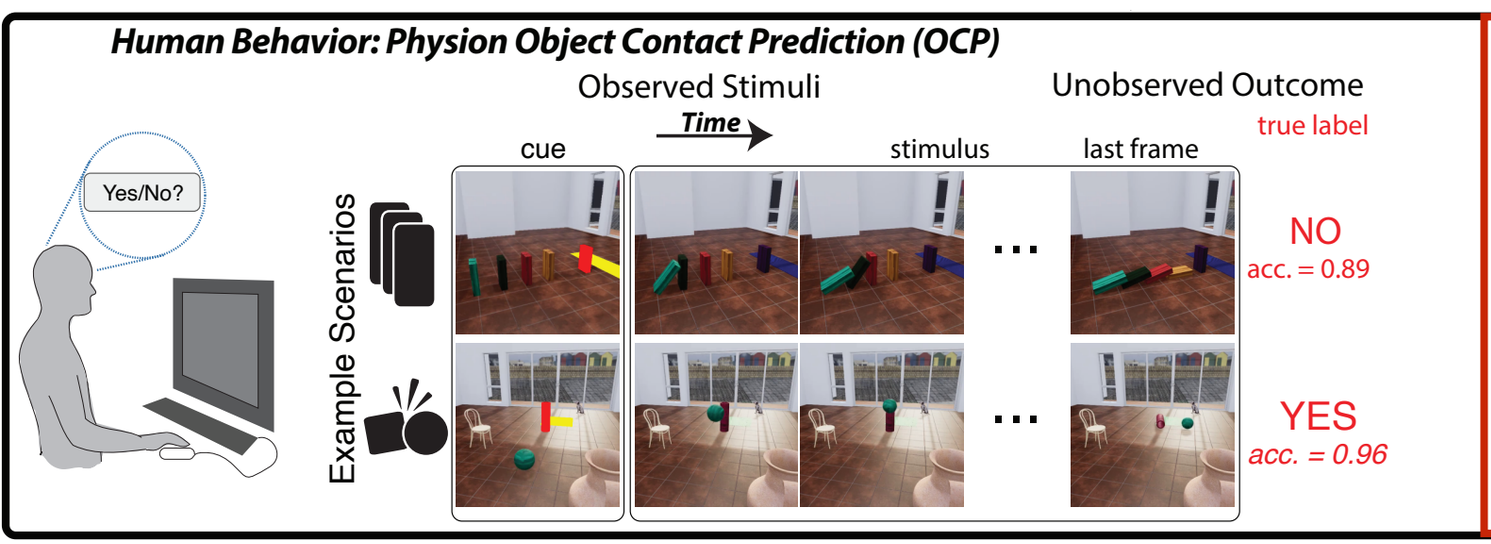
Macaque Neurophysiology: Mental-Pong

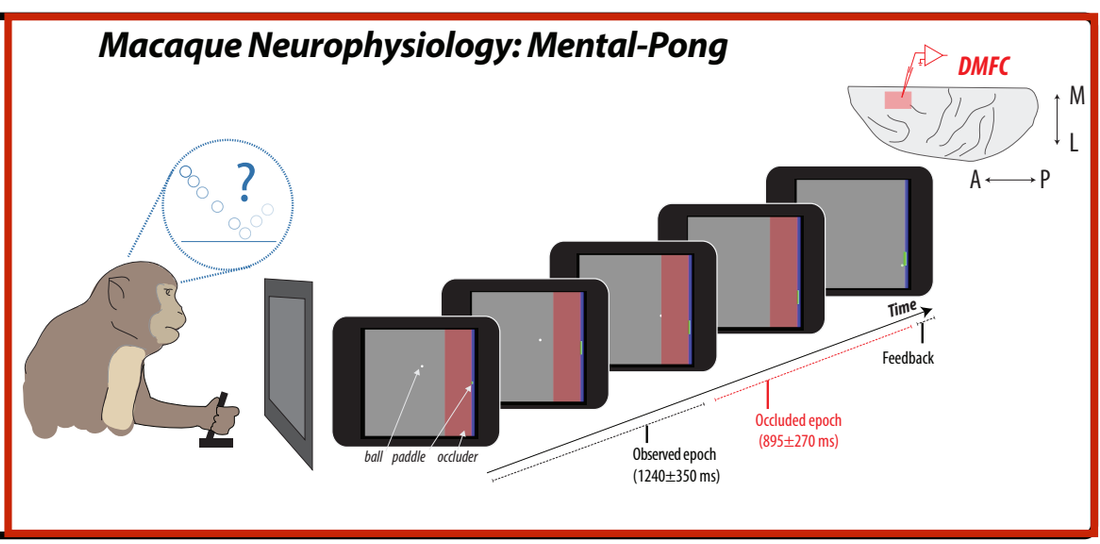




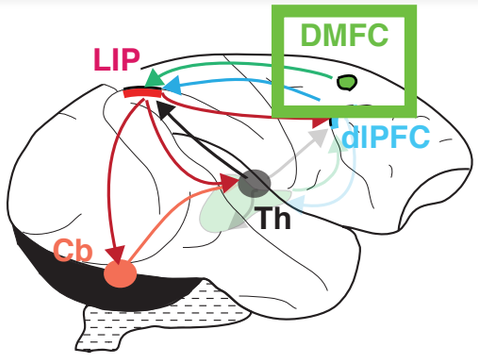
Macaque Neurophysiology: Mental Pong







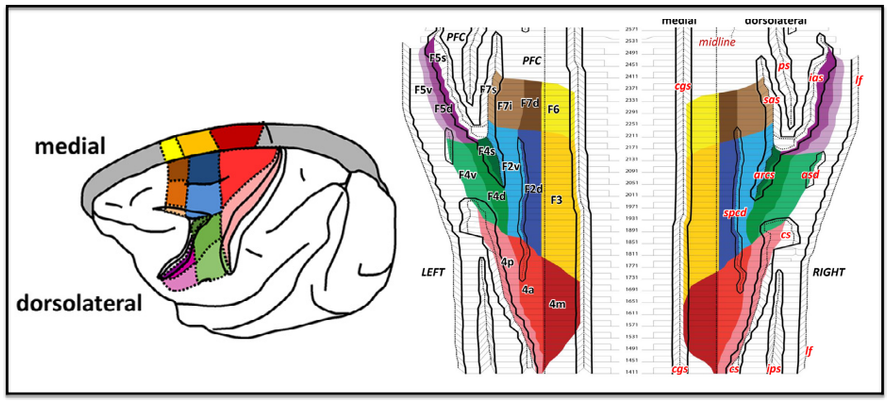
Macaque Neurophysiology: Mental Pong



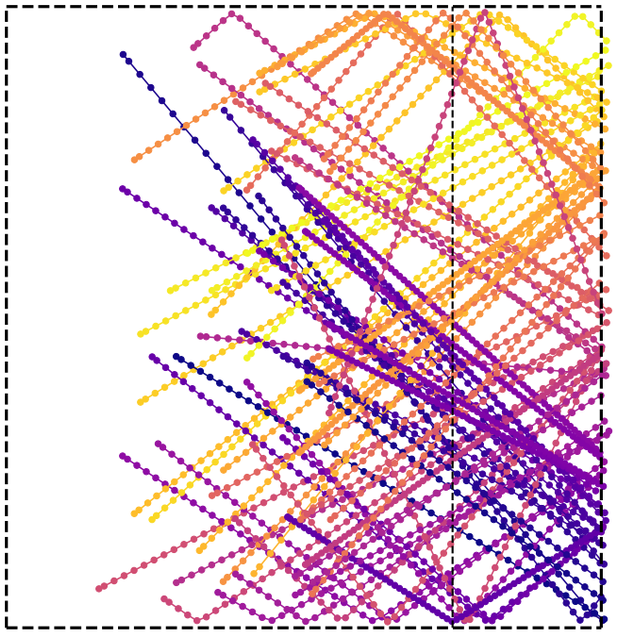
Fronto-Parietal Network

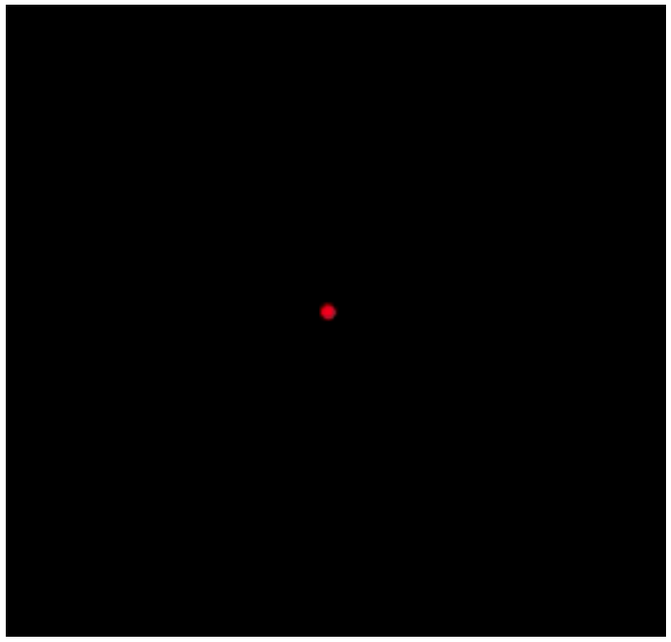
Dorsomedial frontal cortex (DMFC)

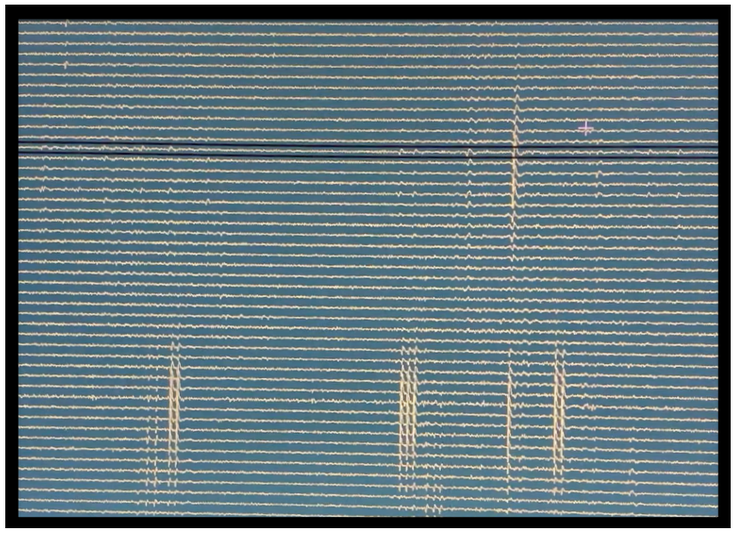
Monkey P *Monkey M*



79 conditions



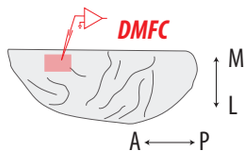


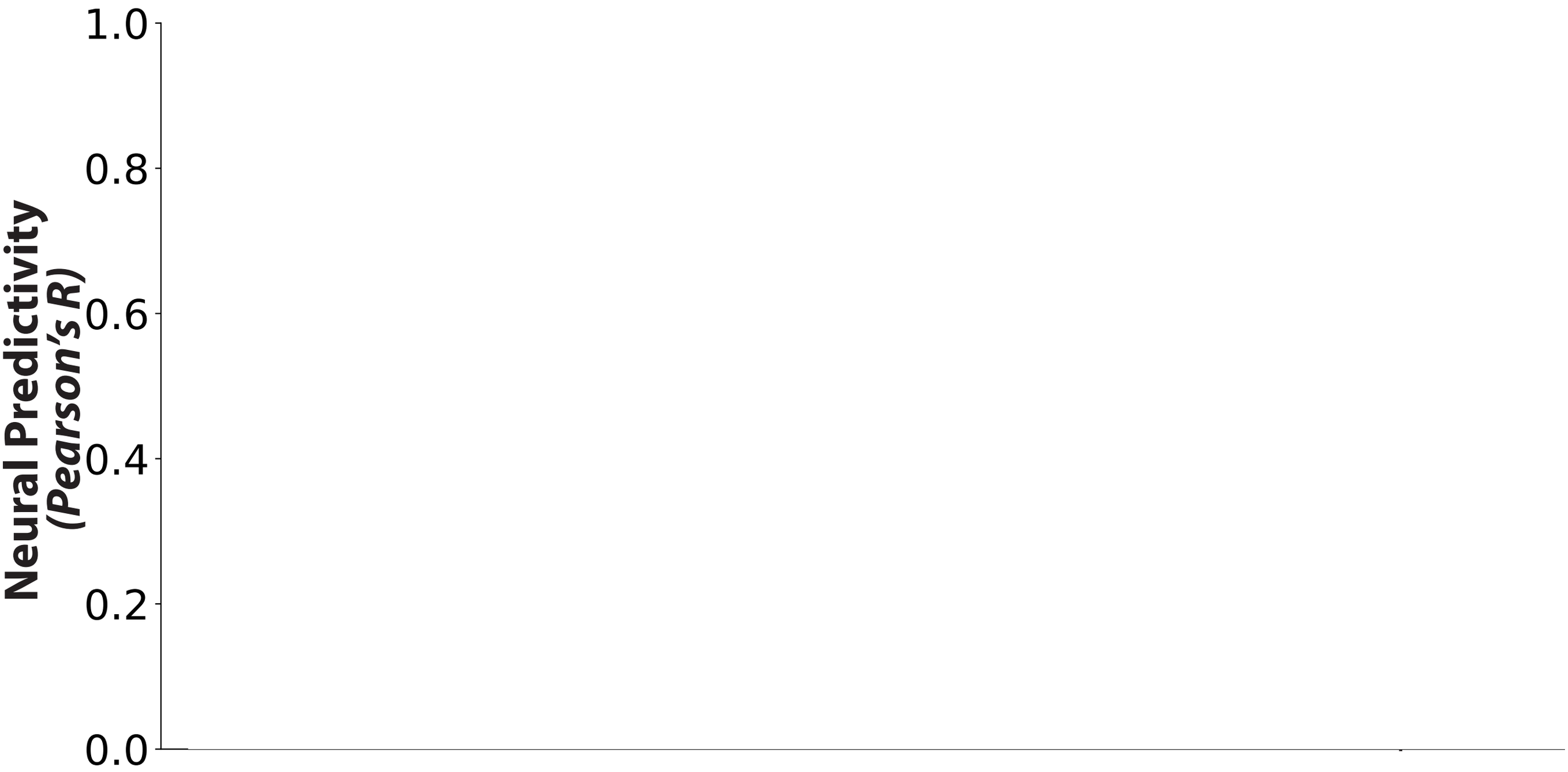


- *Data from two male adult monkeys*
- *79 subsampled M-Pong conditions*
- *64 channel v-probe (monkey P) and 384-channel Neuropixel probe (monkey M)*
- *Total of 1889 stable & reliable neurons recorded from DMFC*

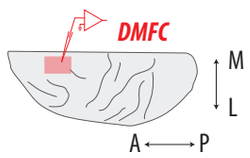
Rishi Rajalingham

Macaque Neurophysiology: Mental Pong





Macaque Neurophysiology: Mental Pong

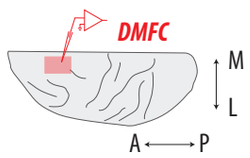


Inter-animal Consistency

Neural Predictivity (Pearson's R)

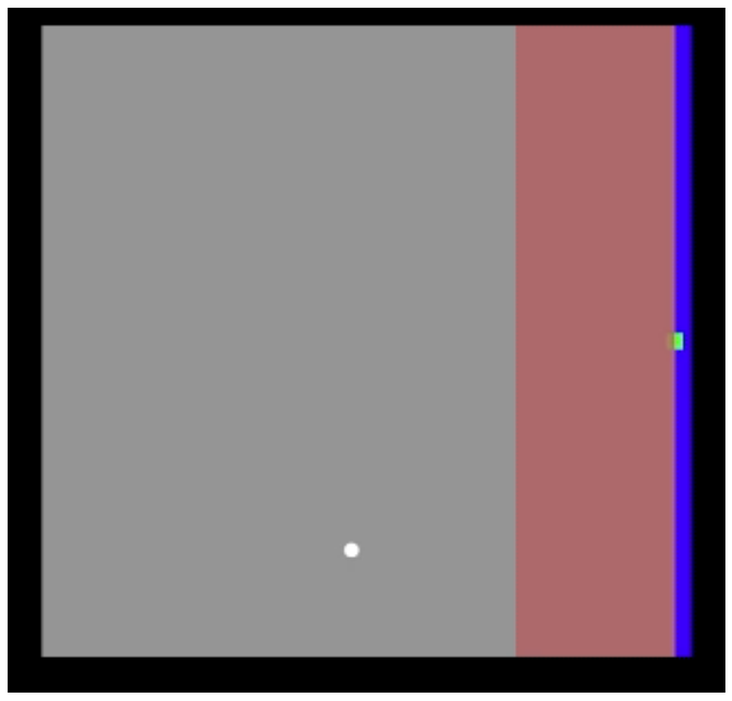
1.0
0.8
0.6
0.4
0.2
0.0

Perfect Simulation Oracle Predicts Neural Data Well



Inter-animal Consistency

Perfect simulation oracle



Ball Position + Velocity

Neural Predictivity (*Pearson's R*)

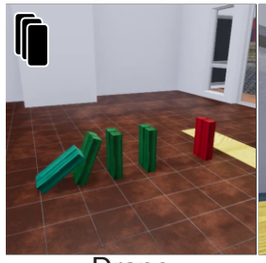
1.0
0.8
0.6
0.4
0.2
0.0

Functional Constraint Hypotheses

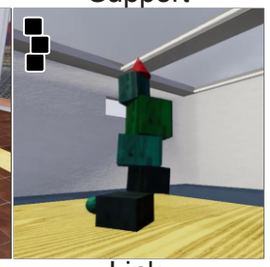
Inputs

Physion

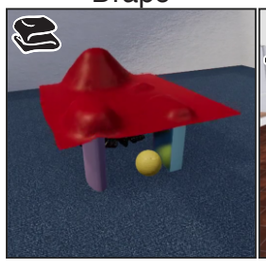
Dominoes



Support



Drape



Link

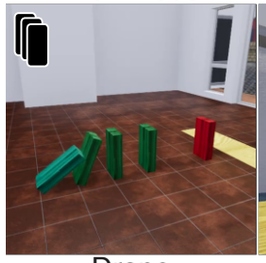
Sensory-Cognitive Hypothesis Classes

Hypothesis Class 1: Pixel-wise Future Prediction

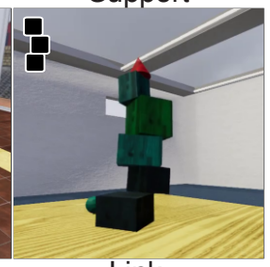
Inputs

Physion

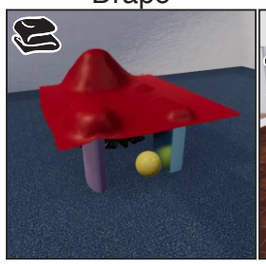
Dominoes



Support



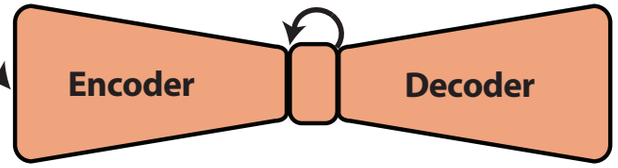
Drape



Link

Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



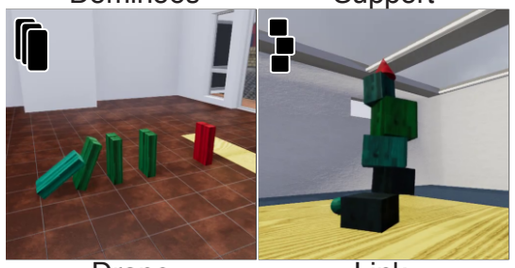
Pixel-wise

Hypothesis Class 1: Pixel-wise Future Prediction

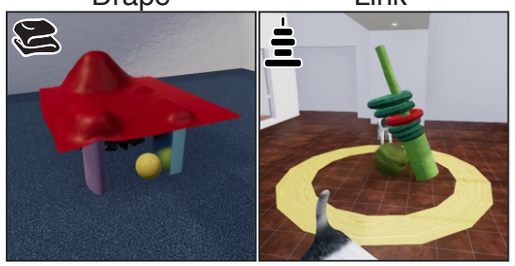
Inputs

Physion

Dominoes | Support

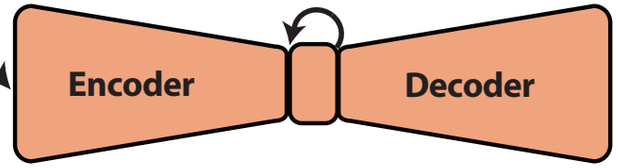


Drape | Link



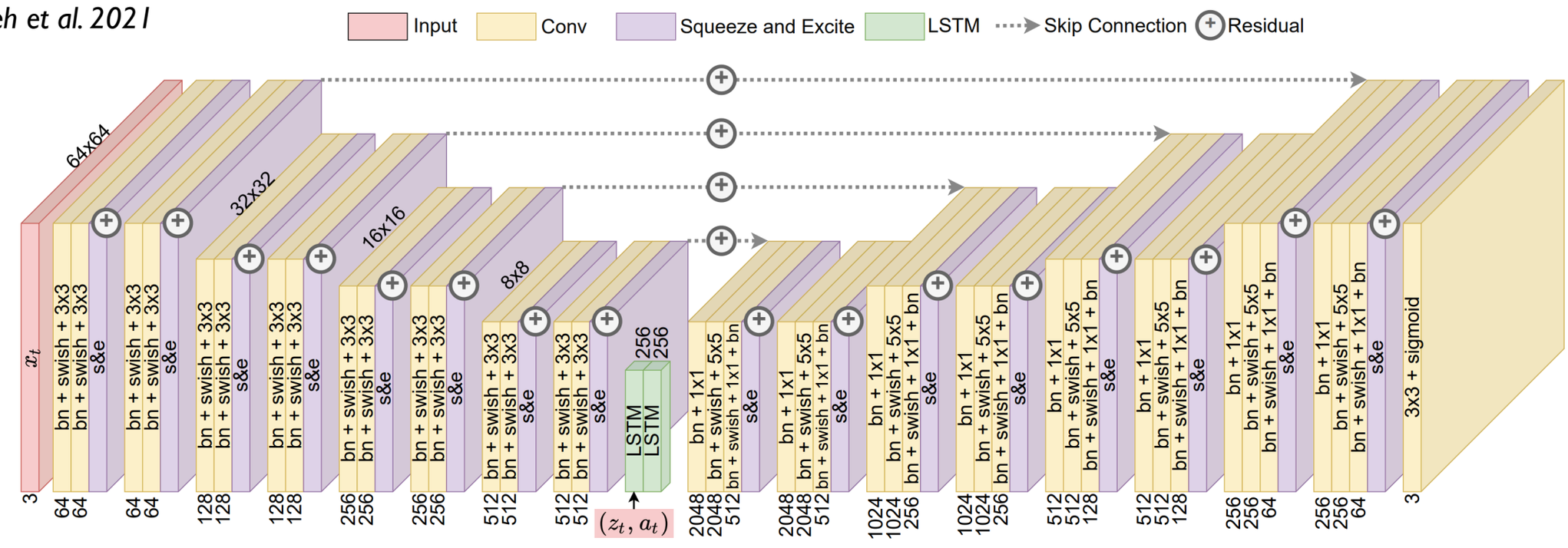
Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



Pixel-wise

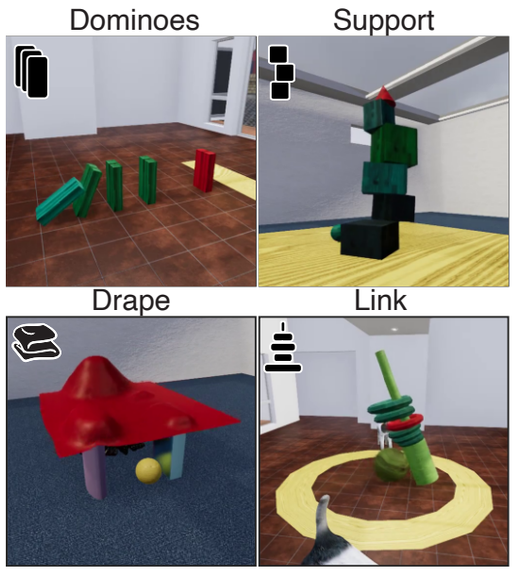
Babaeizadeh et al. 2021



Hypothesis Class 1: Pixel-wise Future Prediction

Inputs

Physion

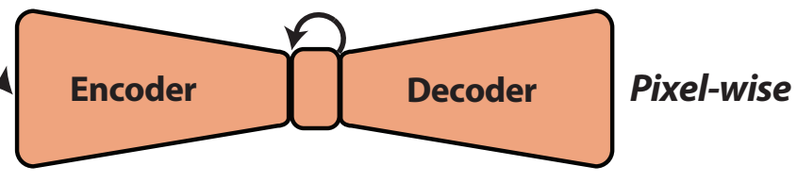


Sensory-Cognitive Hypothesis Classes

Predicts the future at the *pixel* resolution of the sensory input
(very detailed)

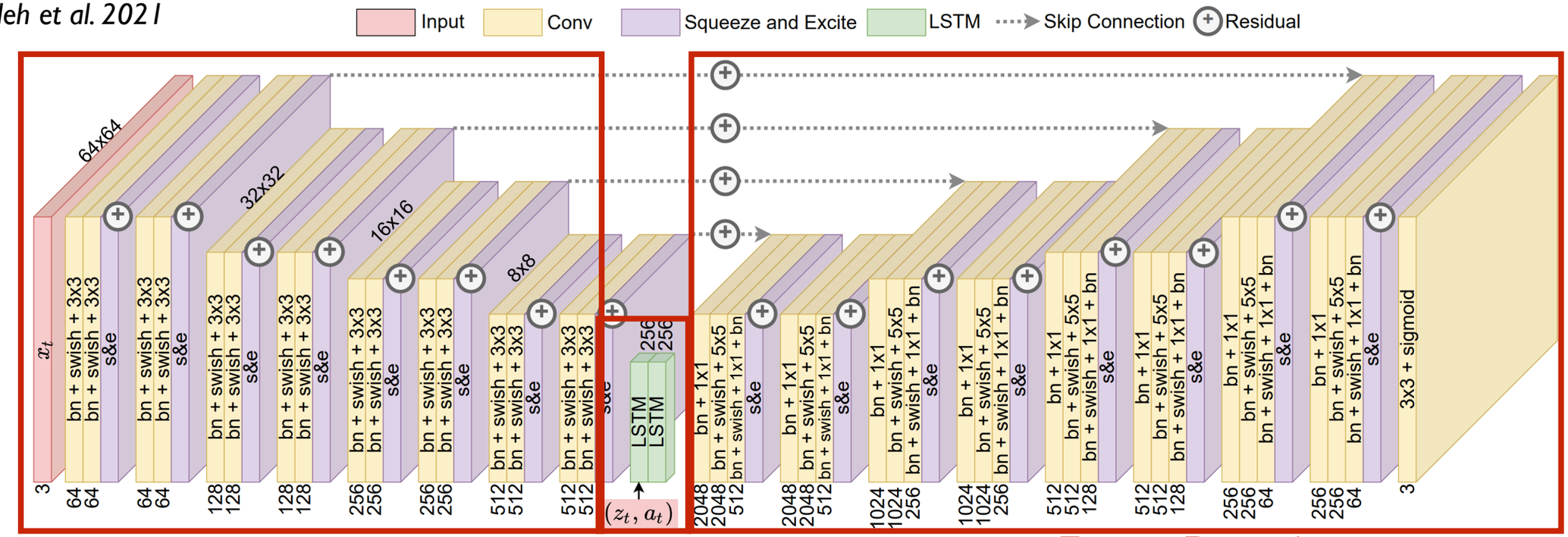
$$\mathcal{L} = \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\|_2^2 + \beta D_{KL}(\mathcal{N}(\mu, \sigma)\|\mathcal{N}(0, \mathbf{I}))$$

End-to-End Future Prediction:

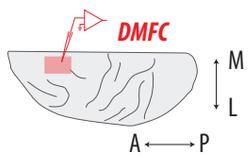


Pixel-wise

Babaeizadeh et al. 2021



Physical Simulation Oracles Predict Neural Data Well



Inter-animal Consistency

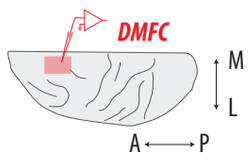
Perfect simulation oracle

Neural Predictivity (*Pearson's R*)

1.0
0.8
0.6
0.4
0.2
0.0

Ball Position + Velocity

Pixel-wise Future Prediction Poorly Predicts Neurons



Inter-animal Consistency

Perfect simulation oracle

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise

64x64
128x128
SVG

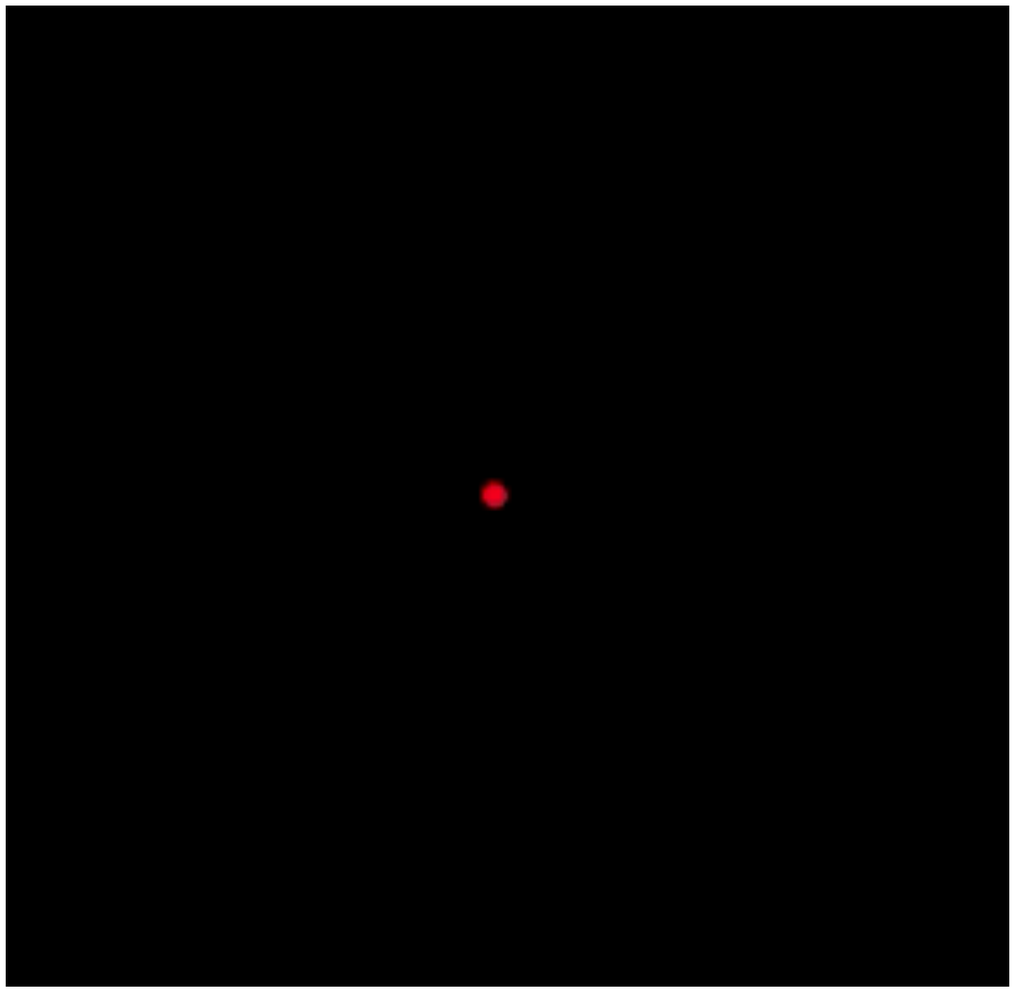
64x64
64x64+ RandAugment
FitVid

End-to-End

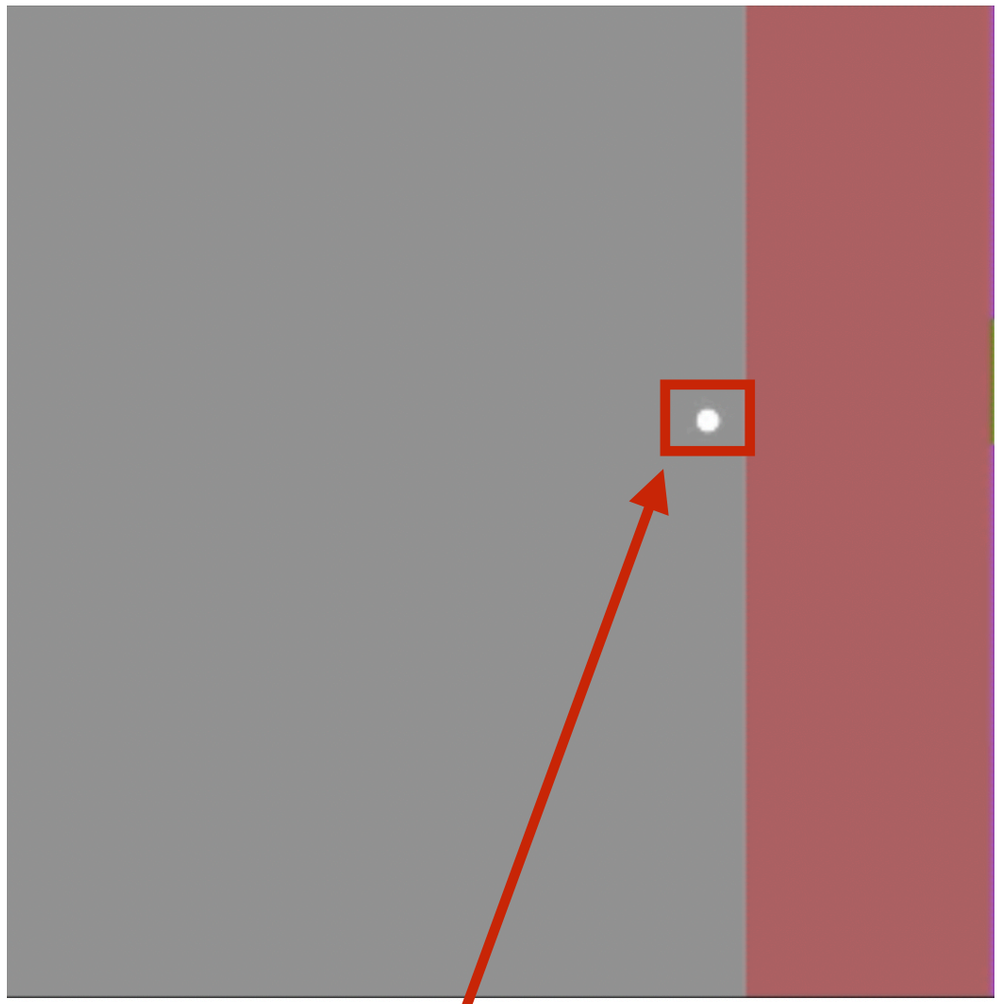
Ball Position + Velocity

…and they struggle to generalize to Pong

Input Frames



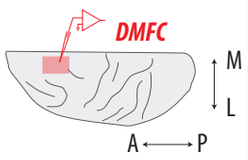
Predicted Frames



Ball stops at final input frame, in the model's "imagination"

Pixel-wise Future Prediction Poorly Predicts Neurons

Perhaps DMFC predicts a "factorized" version of the scene? How?



Inter-animal Consistency

Perfect simulation oracle

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise

64x64
128x128

64x64
64x64+ RandAugment

SVG FitVid

End-to-End

Ball Position + Velocity

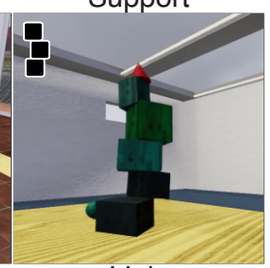
Hypothesis Class 2: Object Slots

Inputs

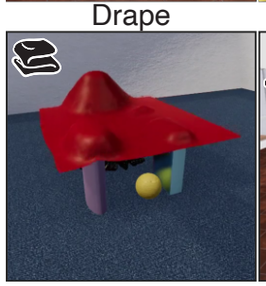
Physion

Dominoes

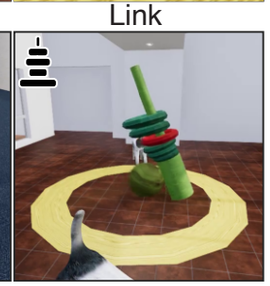
Support



Drape

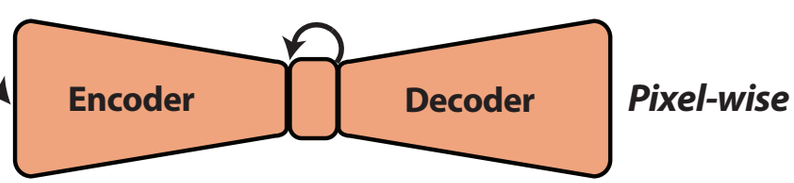


Link

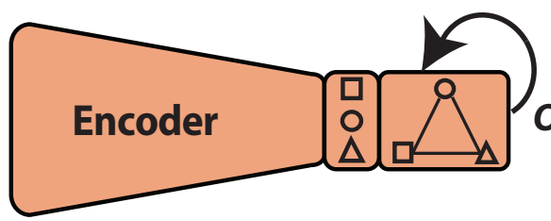


Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



Pixel-wise

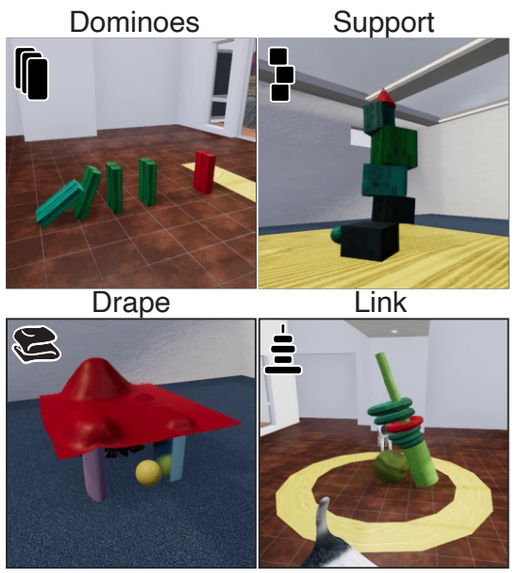


Object-slot

Hypothesis Class 2: Object Slots

Inputs

Physion

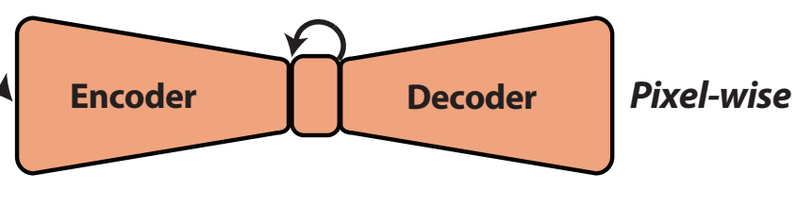


Sensory-Cognitive Hypothesis Classes

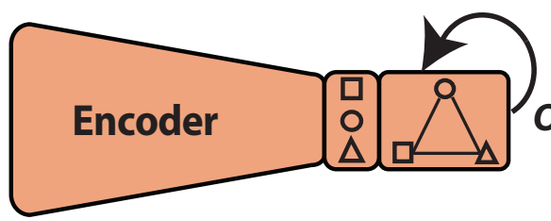
$$\mathcal{L} = \overbrace{\mathcal{H}}^{\text{attract}} + \max(0, 1 - \overbrace{\widetilde{\mathcal{H}}}^{\text{repel}})$$

$$\mathcal{H} = \frac{1}{K} \sum_{k=1}^{K} \left\| \left(\mathbf{z}_t^k + \Delta \mathbf{z}_t^k \right) - \mathbf{z}_{t+1}^k \right\|_2^2$$

End-to-End Future Prediction:



Pixel-wise

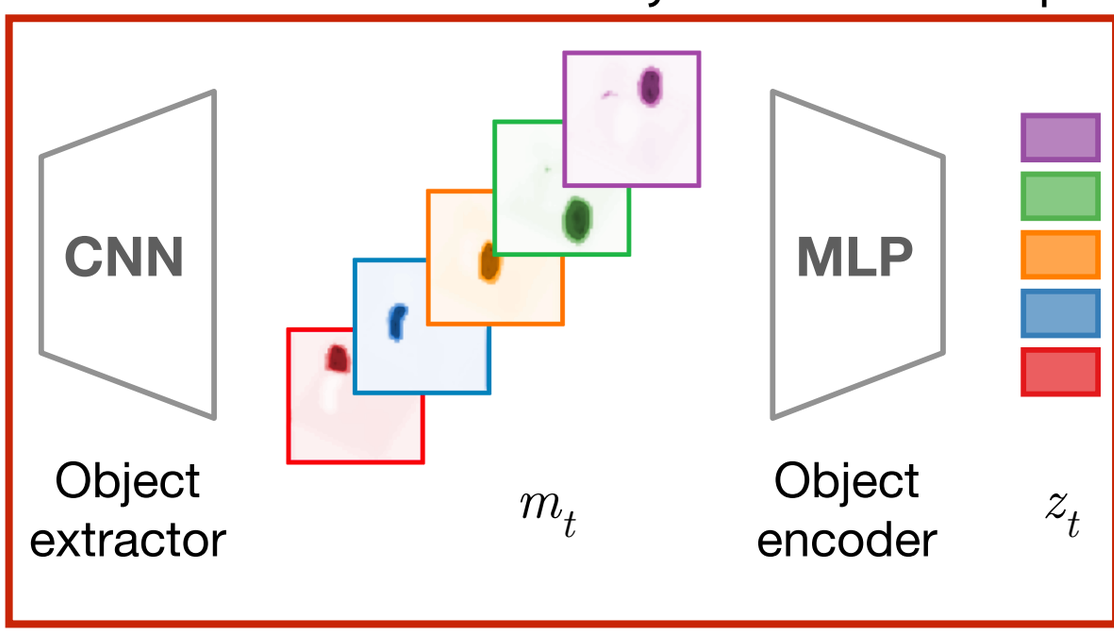


Object-slot

Predicts at the level of object slot representations and their relations

Kipf et al. 2020

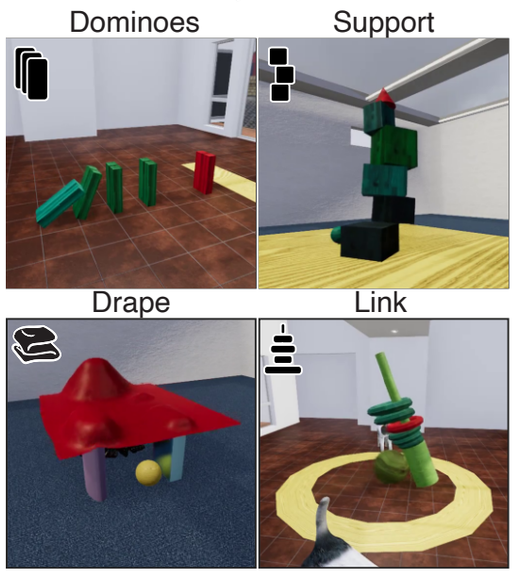
s_t



Hypothesis Class 2: Object Slots

Inputs

Physion

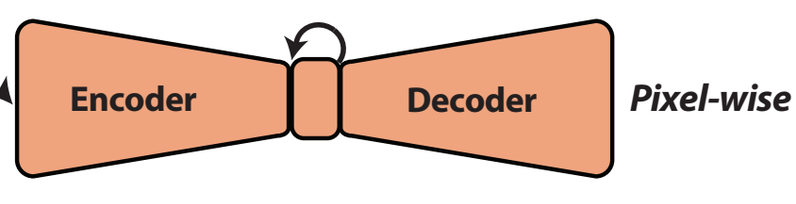


Sensory-Cognitive Hypothesis Classes

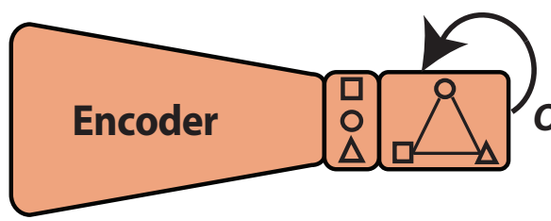
$$\mathcal{L} = \overbrace{\mathcal{H}}^{\text{attract}} + \max(0, 1 - \overbrace{\widetilde{\mathcal{H}}}^{\text{repel}})$$

$$\mathcal{H} = \frac{1}{K}\sum_{k=1}^{K} \left\| \left(\mathbf{z}_t^k + \Delta\mathbf{z}_t^k\right) - \mathbf{z}_{t+1}^k \right\|_2^2, \quad \widetilde{\mathcal{H}} = \frac{1}{K}\sum_{k=1}^{K} \left\| \widetilde{\mathbf{z}}_t^k - \mathbf{z}_{t+1}^k \right\|_2^2$$

End-to-End Future Prediction:



Pixel-wise

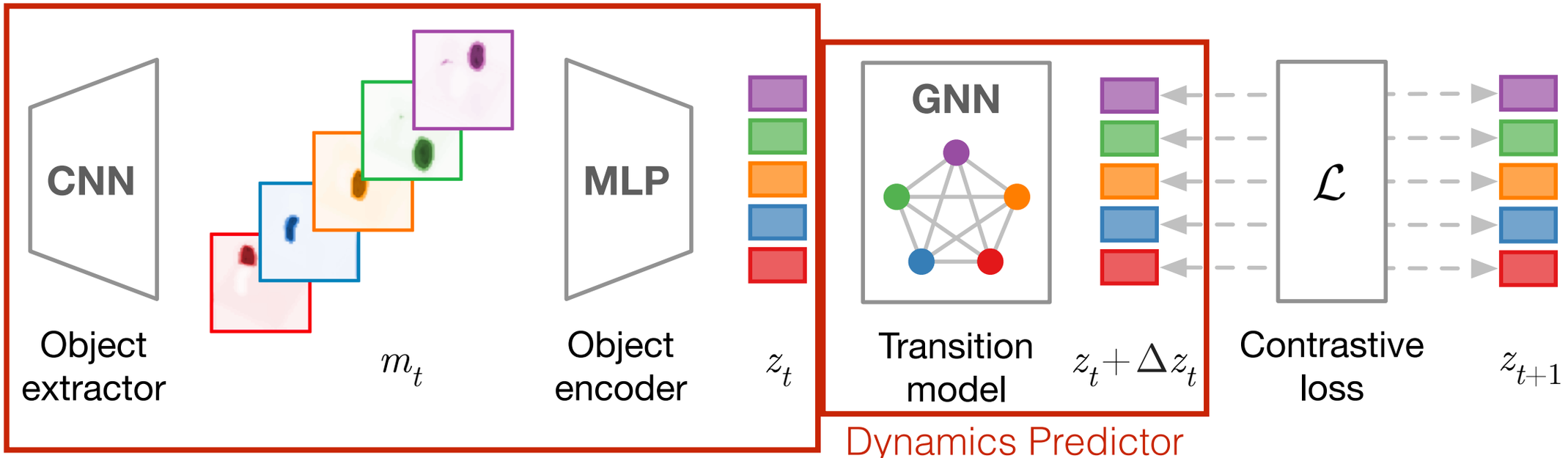


Object-slot

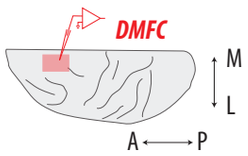
Predicts at the level of object slot representations and their relations

Kipf et al. 2020

s_t



Pixel-wise Future Prediction Poorly Predicts Neurons



Inter-animal Consistency

Perfect simulation oracle

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise

64x64
128x128

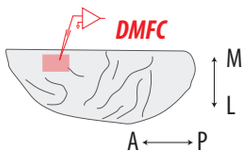
64x64
64x64+ RandAugment

SVG
FitVid

End-to-End

Ball Position + Velocity

Object Slot Future Prediction Poorly Predicts Neurons



Inter-animal Consistency

Perfect simulation oracle

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise
Object -slot

64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large

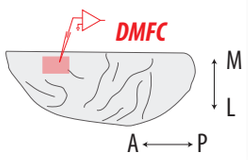
SVG
FitVid
C-SWM

End-to-End

Ball Position + Velocity

Object Slot Future Prediction Poorly Predicts Neurons

Perhaps DMFC predicts a "factorized" version of the scene? How?



Inter-animal Consistency

Perfect simulation oracle

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise
Object -slot

64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large

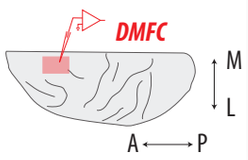
SVG
FitVid
C-SWM

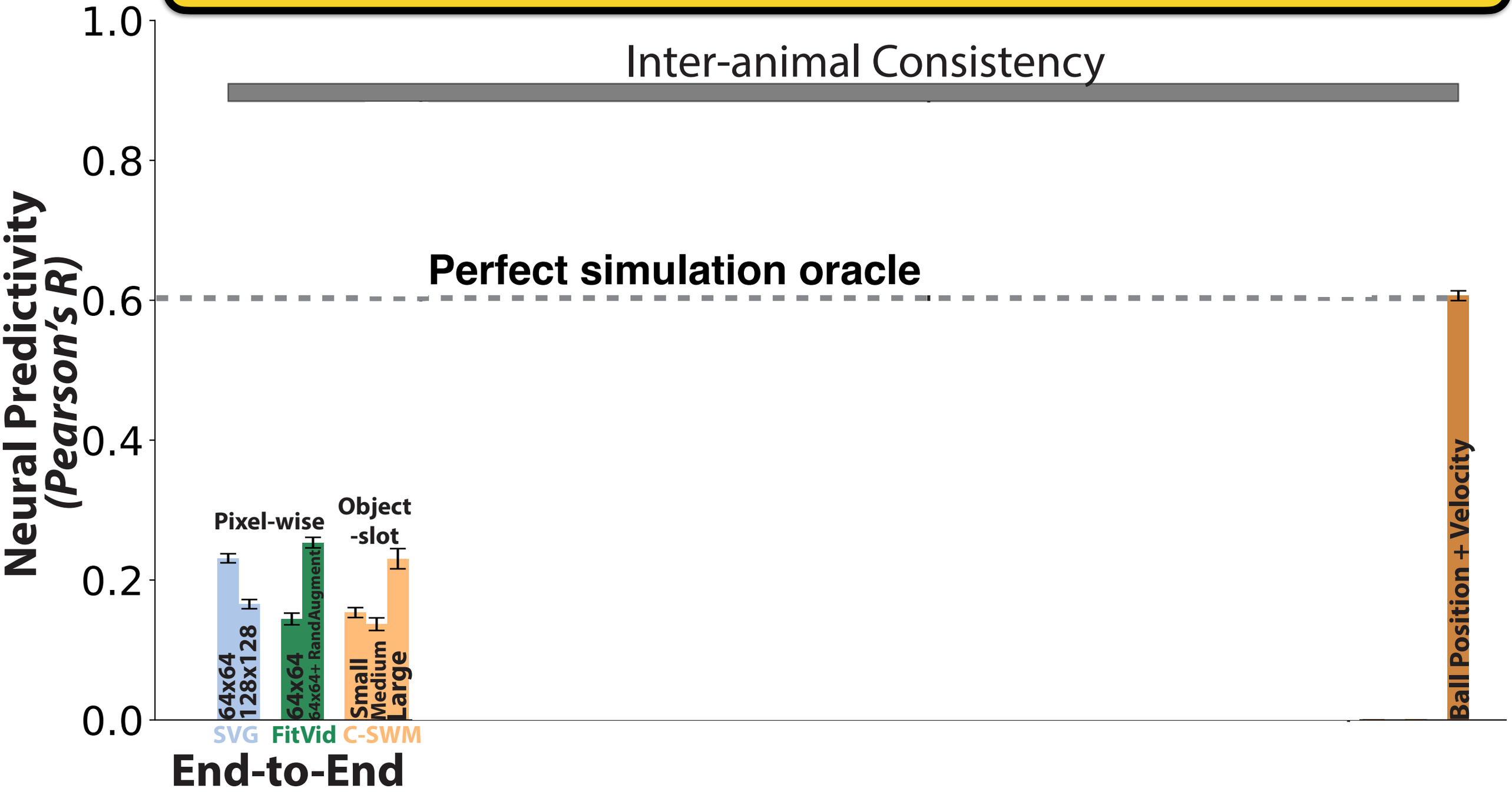
End-to-End

Ball Position + Velocity

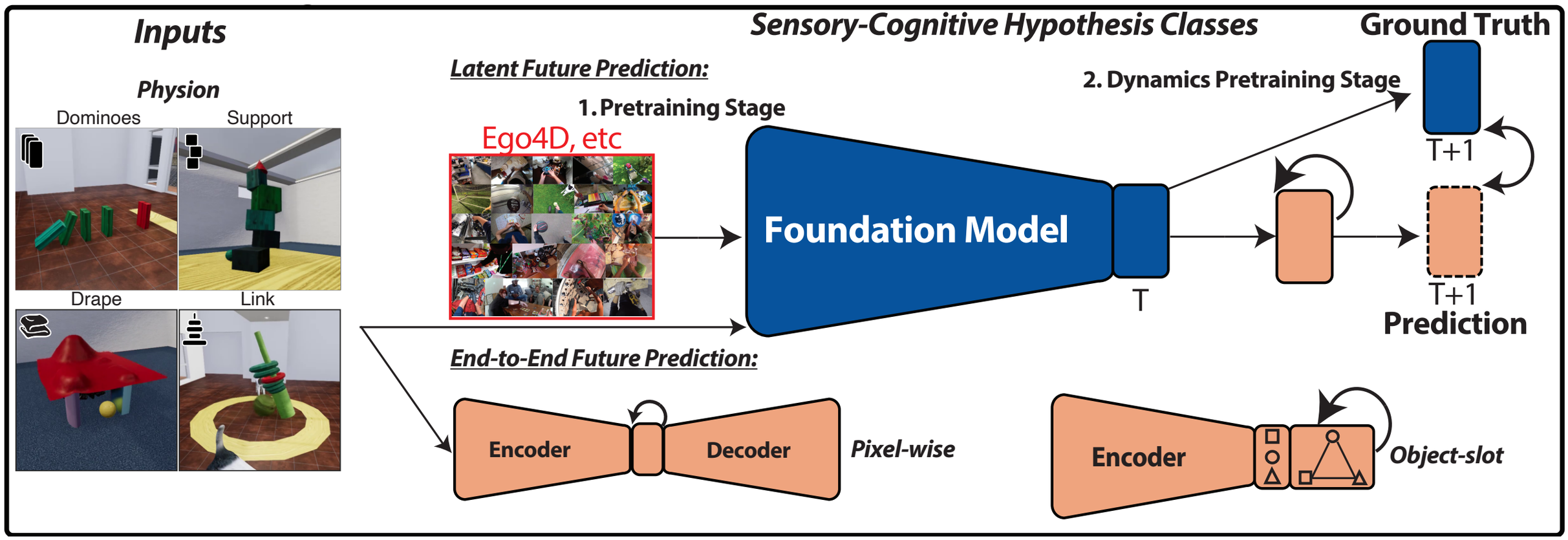
Object Slot Future Prediction Poorly Predicts Neurons

Perhaps DMFC predicts a "factorized" version of the scene?
How? Not by allocating fixed object slots!

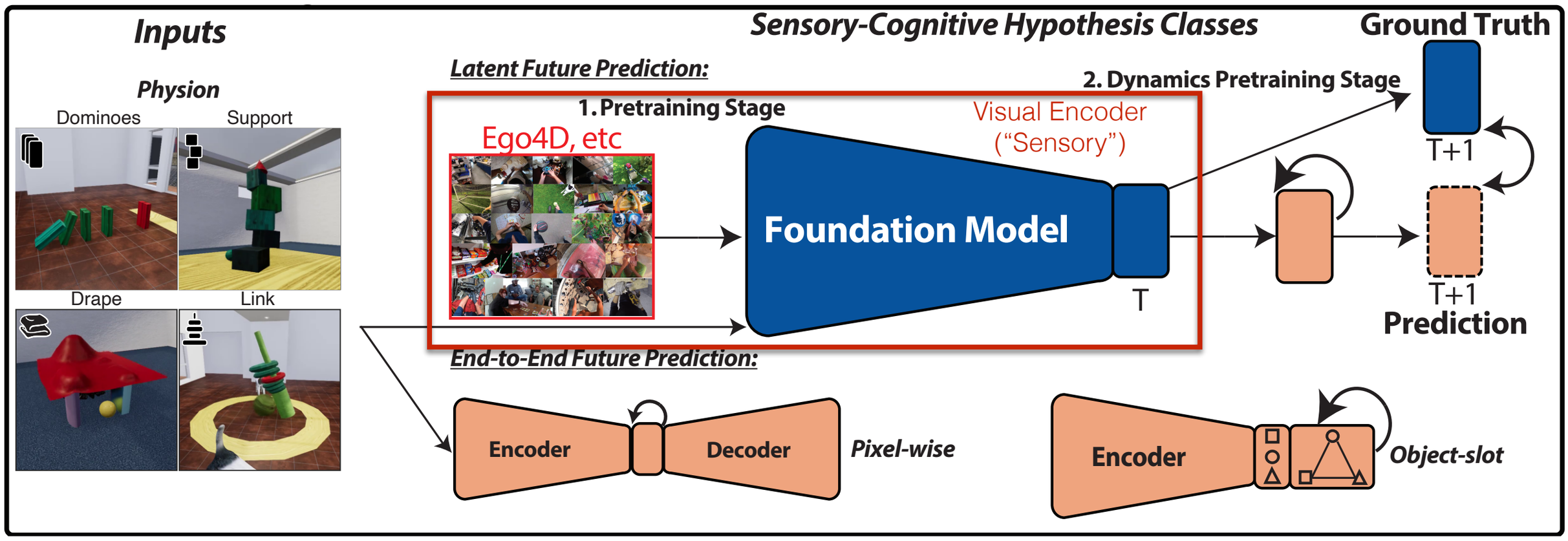




Hypothesis Class 3: Latent Future Prediction

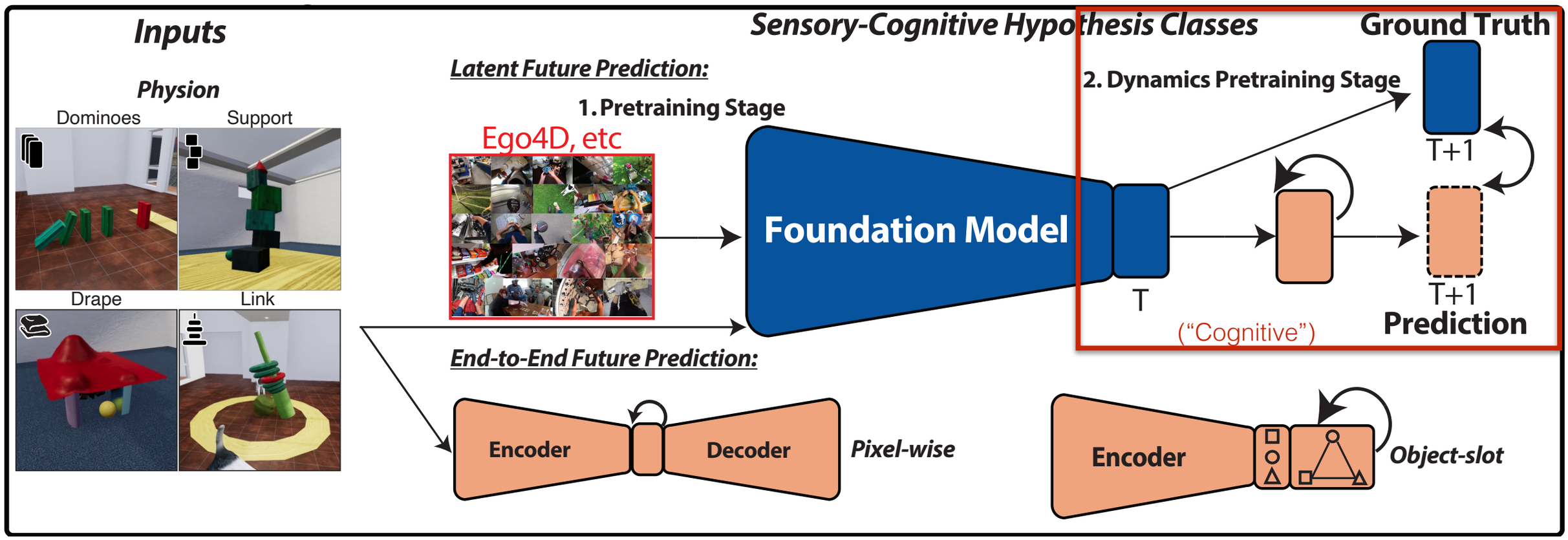


Hypothesis Class 3: Latent Future Prediction



Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

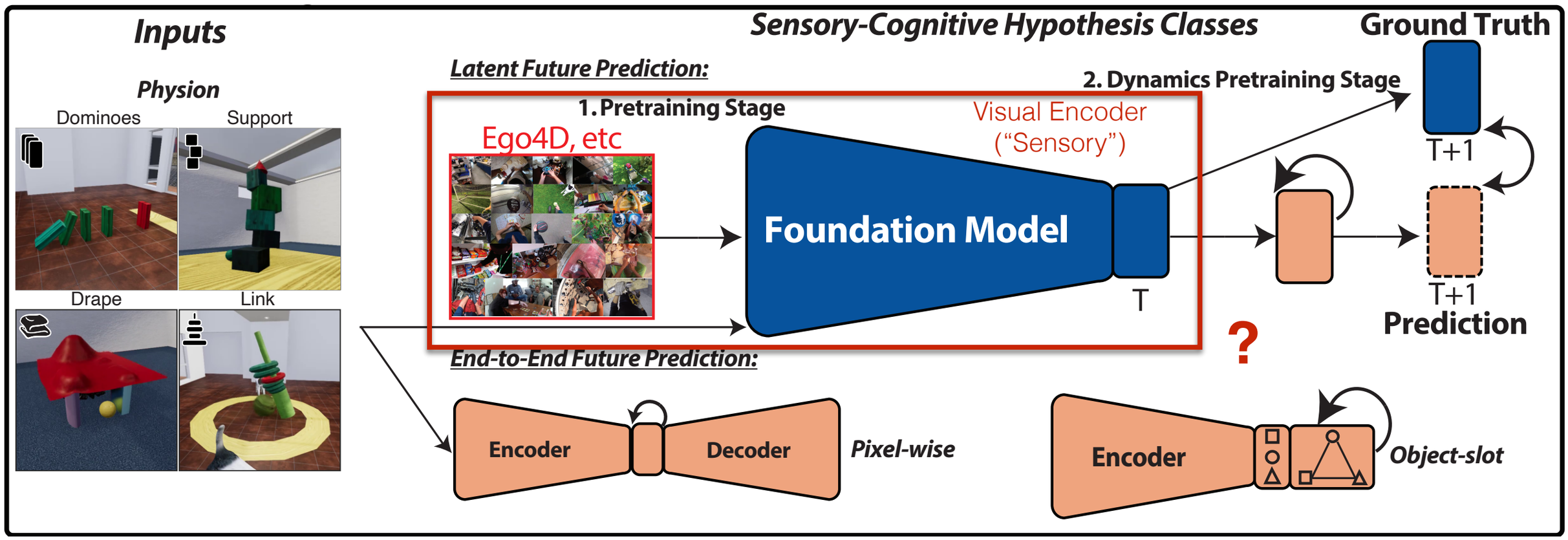
Hypothesis Class 3: Latent Future Prediction



Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

Leverage these dynamics to do explicit future prediction

Hypothesis Class 3: Foundation Models

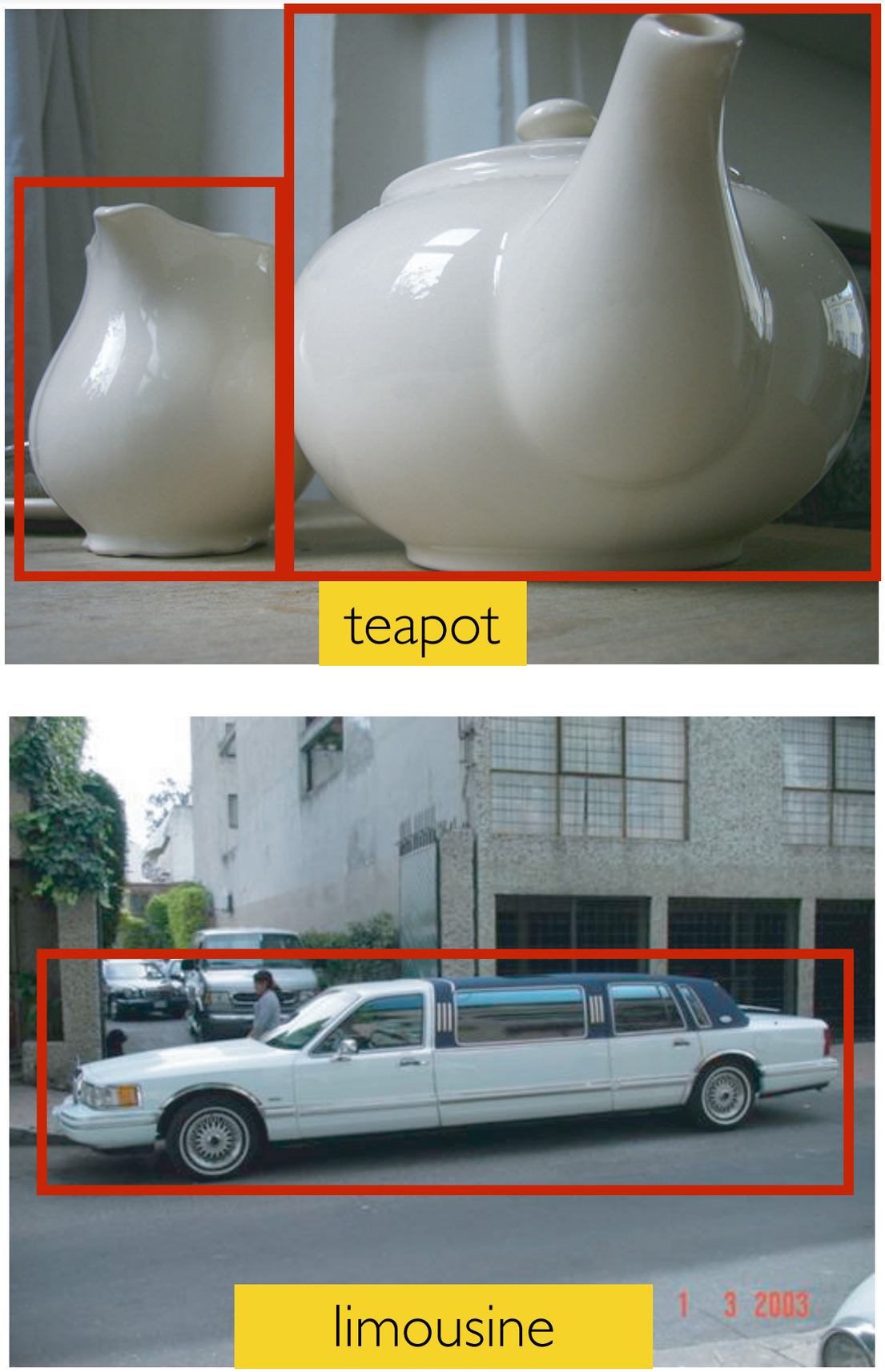


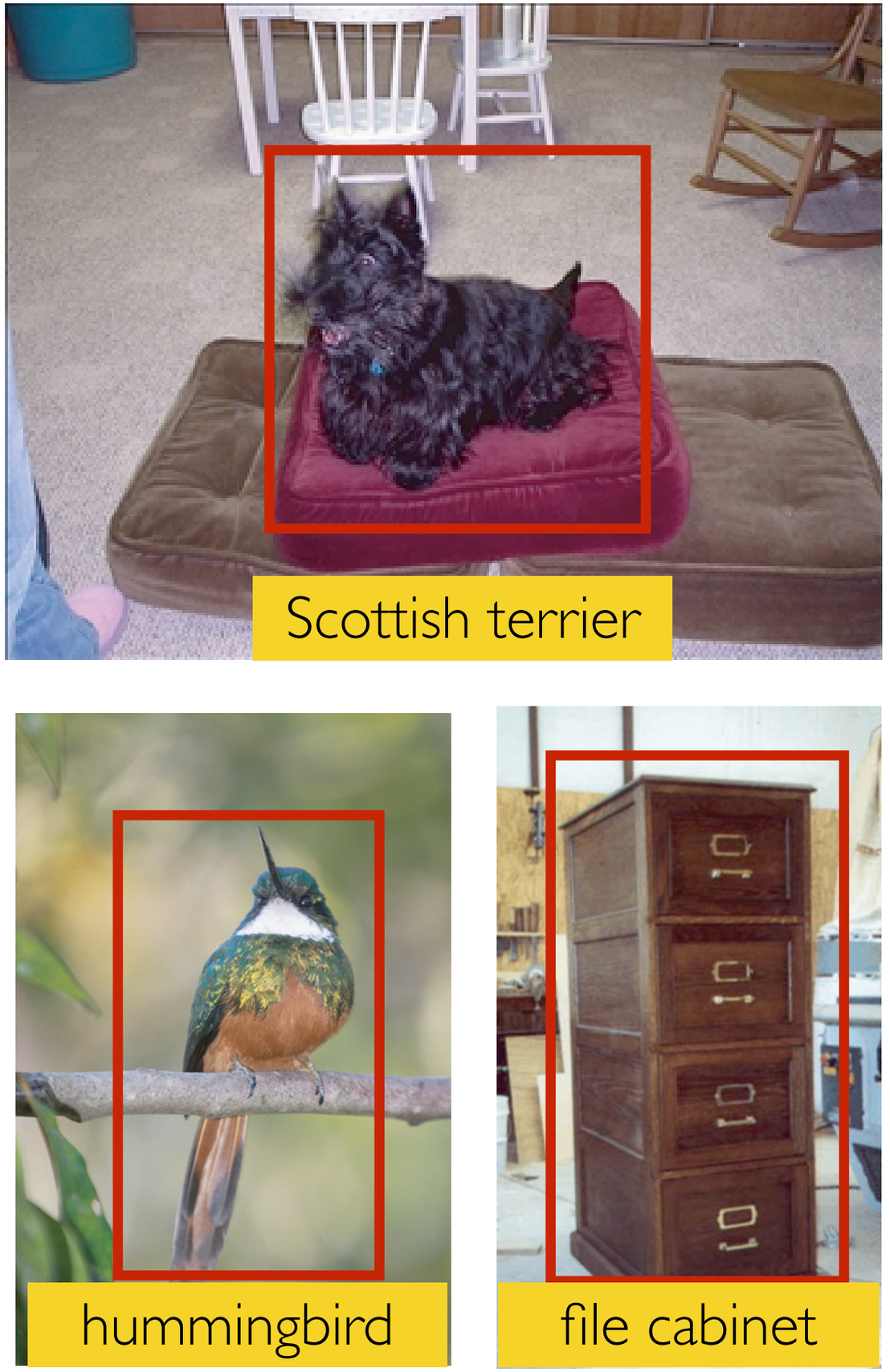
Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

What vision task?

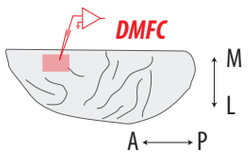
Leverage these dynamics to do explicit future prediction

Hypothesis Class 3: Static Image Foundation Models





Object Slot Future Prediction Poorly Predicts Neurons



Inter-animal Consistency

Perfect simulation oracle

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise
Object -slot

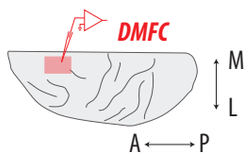
64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large

SVG
FitVid
C-SWM

End-to-End

Ball Position + Velocity

Static Image Foundation Future Prediction Poorly Predicts Neurons



Inter-animal Consistency

Perfect simulation oracle

Useful for image segmentation, classification, etc tasks

Image Foundation Models

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

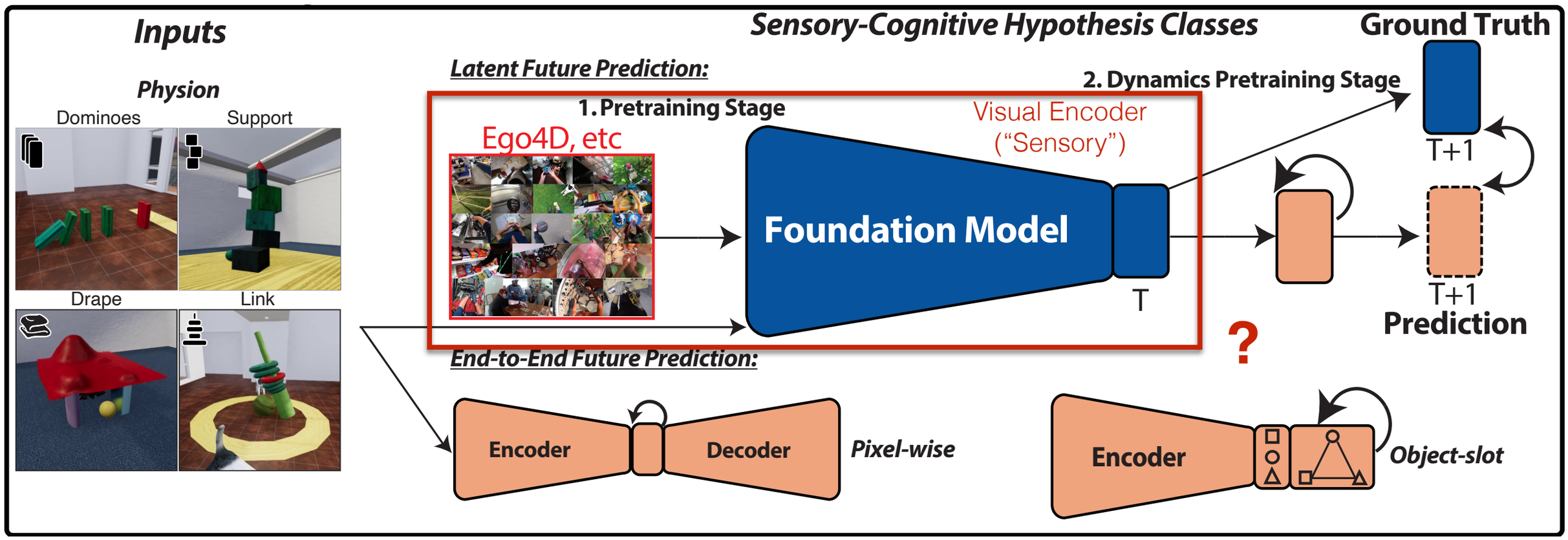
Pixel-wise
Object -slot
64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large
No Dynamics
LSTM Dynamics
CTRNN Dynamics
Ball Position + Velocity

SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP

End-to-End

Latent Future Prediction

Hypothesis Class 3: Foundation Models



Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

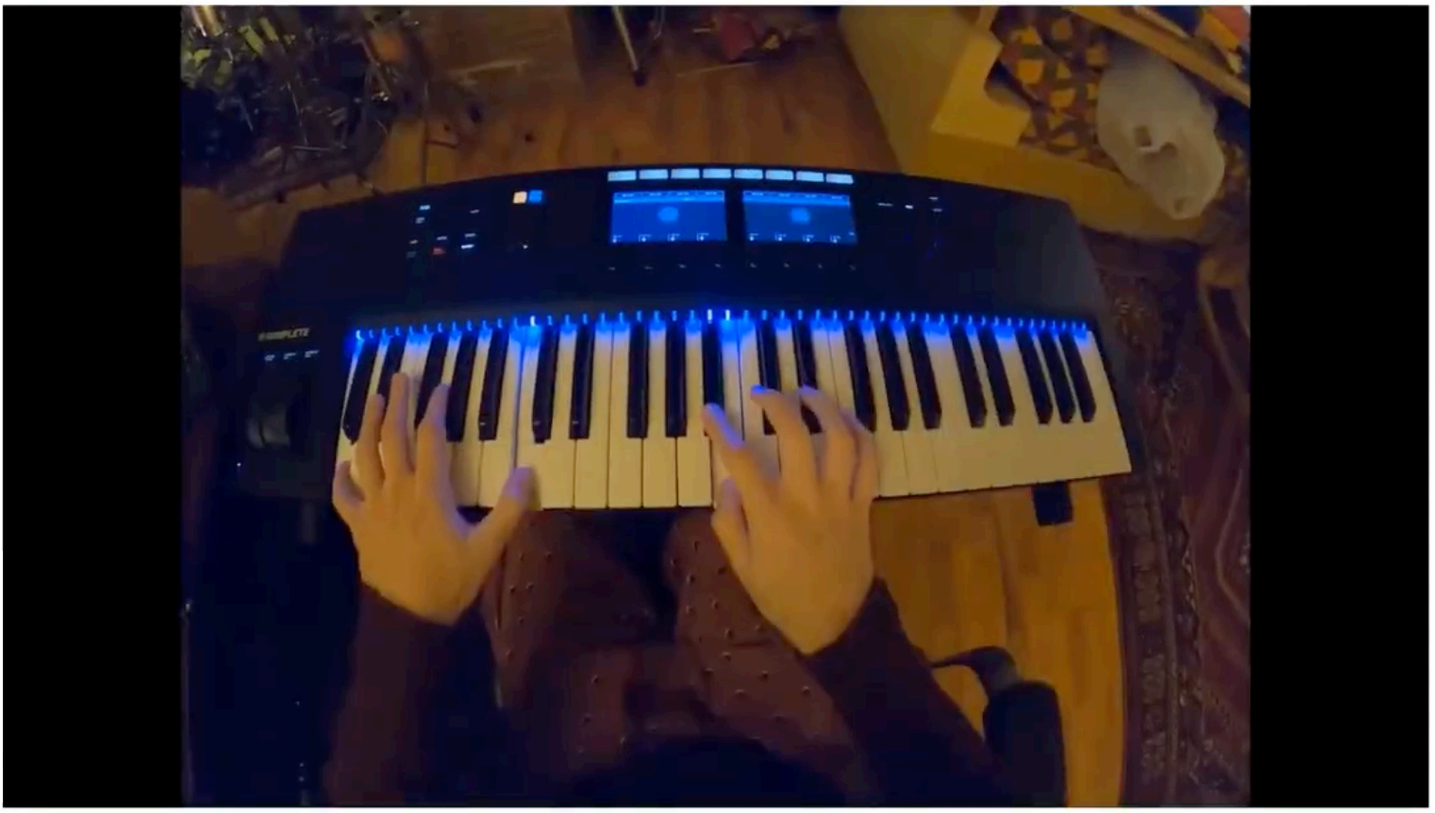
What vision task?

We do far more than engage with static images!

Leverage these dynamics to do explicit future prediction

Hypothesis Class 3:Video Foundation Models

Ego4D: everyday activity around the world



$$\mathcal{L}_{contrastive} = -\sum_{b \in B} \log \frac{e^{\overbrace{\mathcal{S}(\mathbf{z}_i^b, \mathbf{z}_j^b)}^{\text{attract}}}}{e^{\mathcal{S}(\mathbf{z}_i^b, \mathbf{z}_j^b)} + e^{\overbrace{\mathcal{S}(\mathbf{z}_i^b, \mathbf{z}_k^b)}^{\text{repel}}} + e^{\overbrace{\mathcal{S}(\mathbf{z}_i^b, \widetilde{\mathbf{z}}_i^b)}^{\text{repel}}}}$$

$$[I_i, I_{j>i}, I_{k>j}]^{1:B}$$

Ego4D: A massive-scale egocentric dataset

3,670 hours of in-the-wild daily life activity

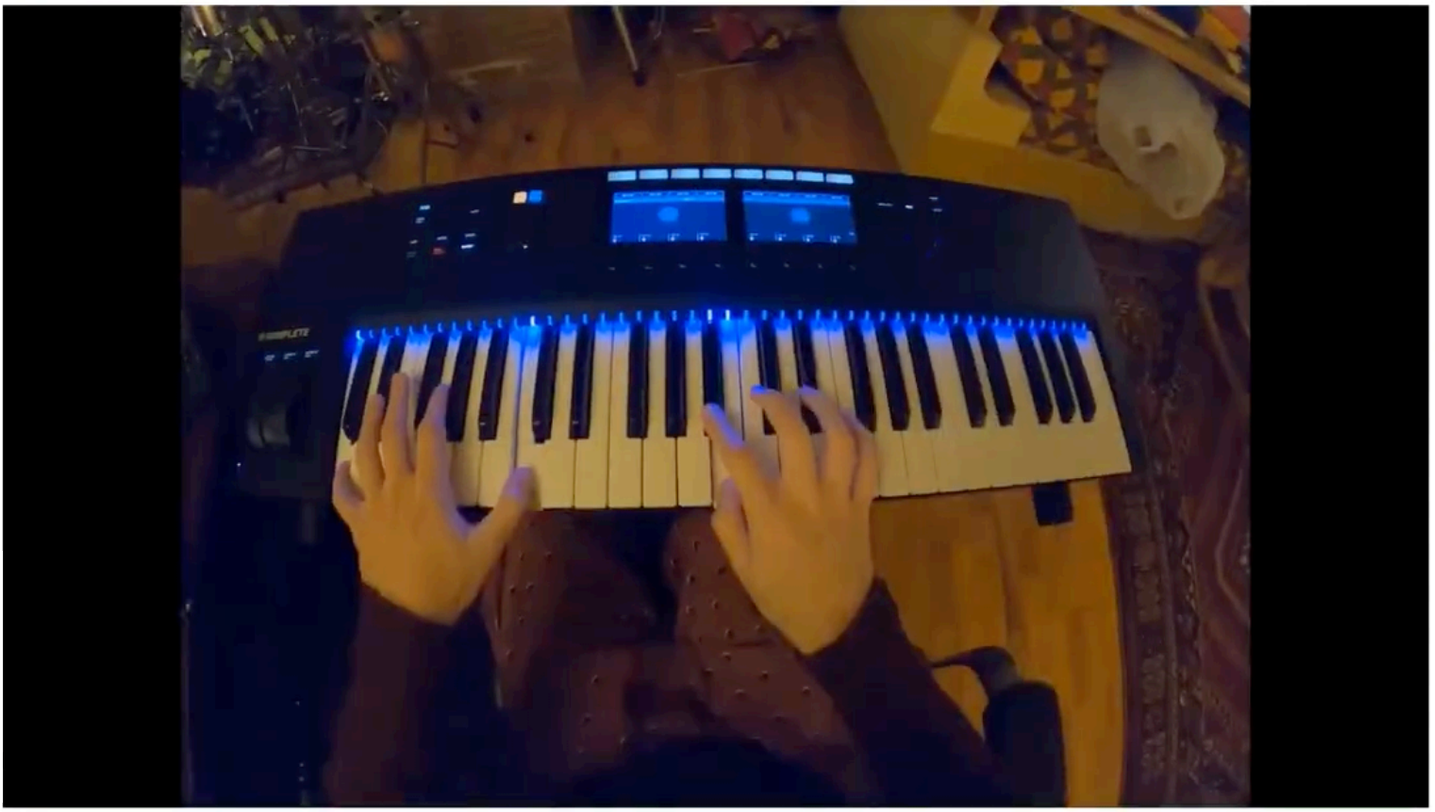
931 participants from 74 worldwide locations

Multimodal: audio, 3D scans, IMU, stereo, multi-camera

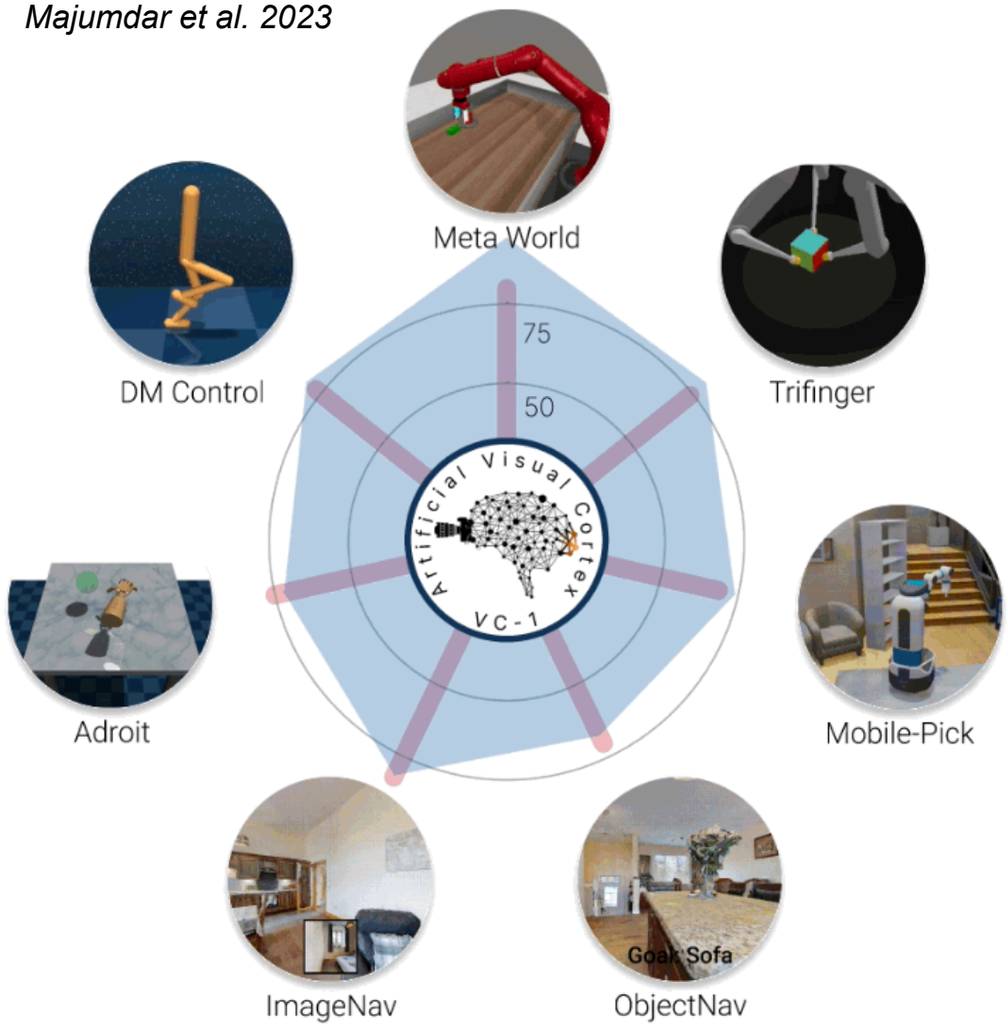
Grauman et al. 2022

Hypothesis Class 3: Video Foundation Models

Ego4D: everyday activity around the world



Majumdar et al. 2023



Ego4D: A massive-scale egocentric dataset

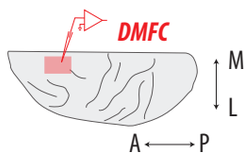
3,670 hours of in-the-wild daily life activity

931 participants from 74 worldwide locations

Multimodal: audio, 3D scans, IMU, stereo, multi-camera

Grauman et al. 2022

Static Image Foundation Future Prediction Poorly Predicts Neurons



Inter-animal Consistency

Perfect simulation oracle

Useful for image segmentation, classification, etc tasks

Image Foundation Models

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

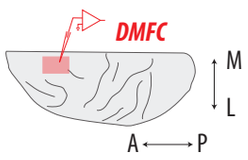
Pixel-wise
Object -slot
64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large
No Dynamics
LSTM Dynamics
CTRNN Dynamics
Ball Position + Velocity

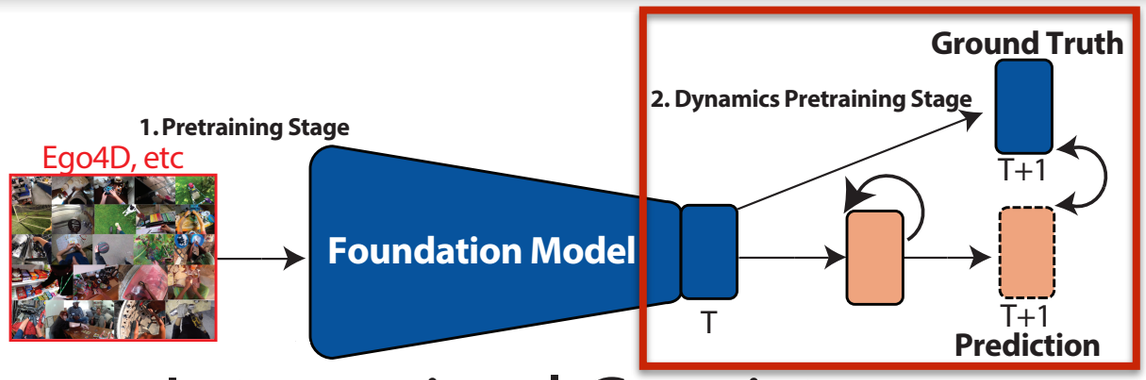
SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP

End-to-End

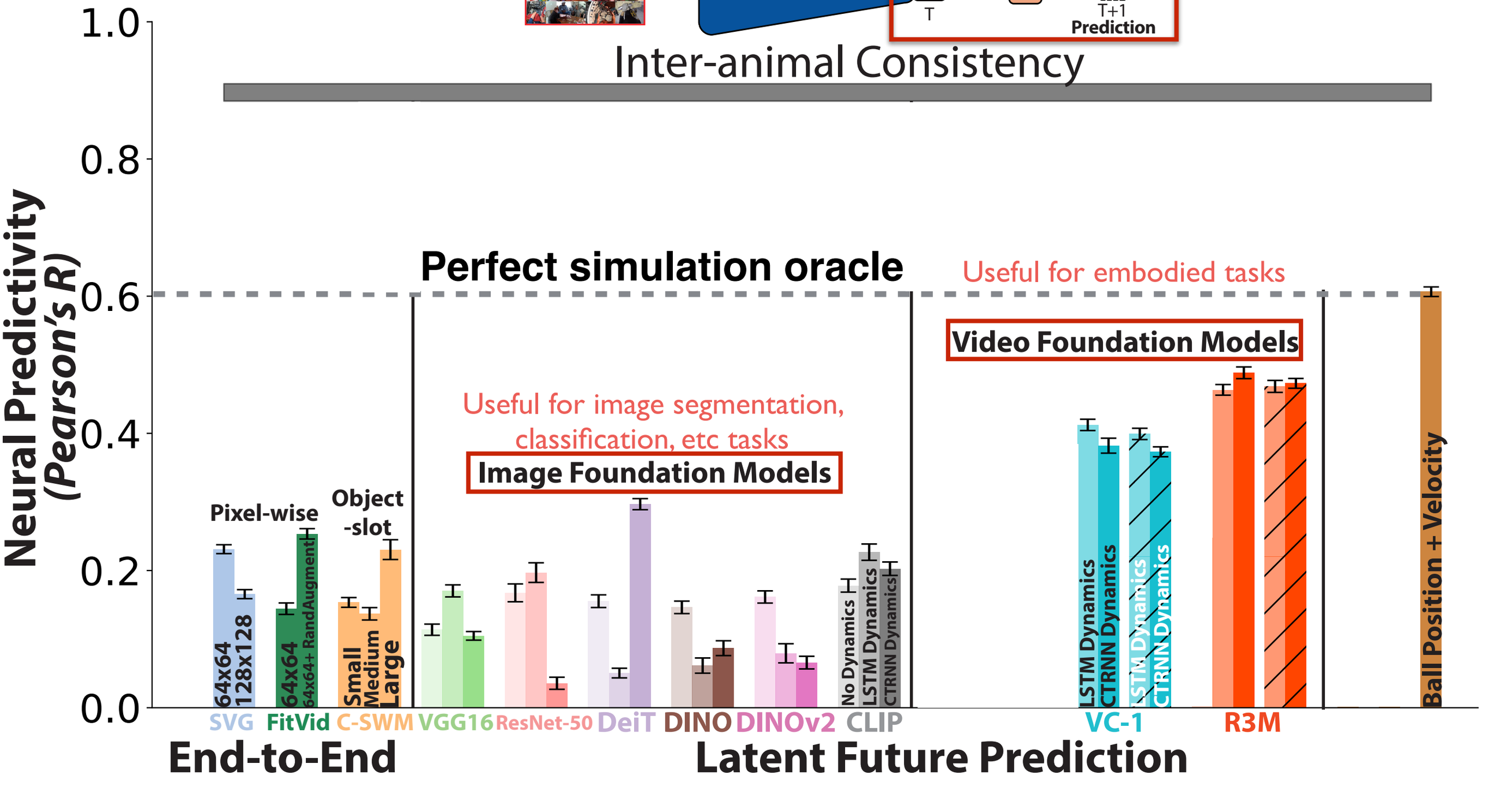
Latent Future Prediction

Video Foundation Future Prediction Best Predict Neurons

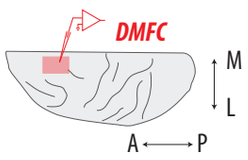


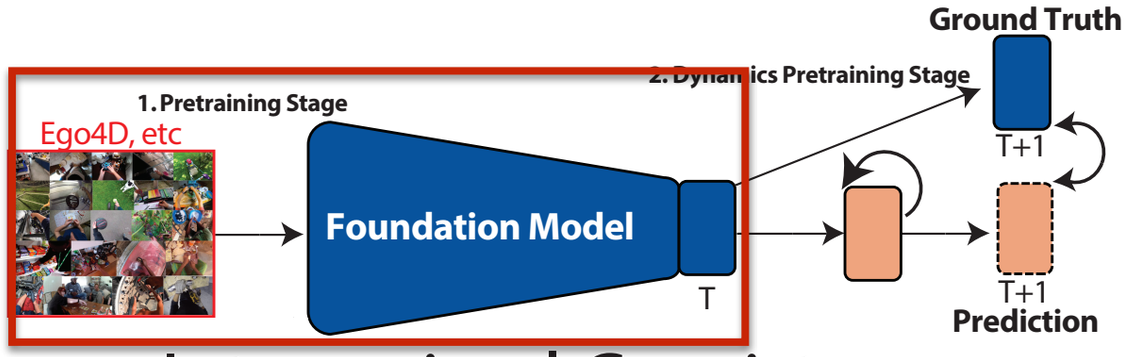


Inter-animal Consistency

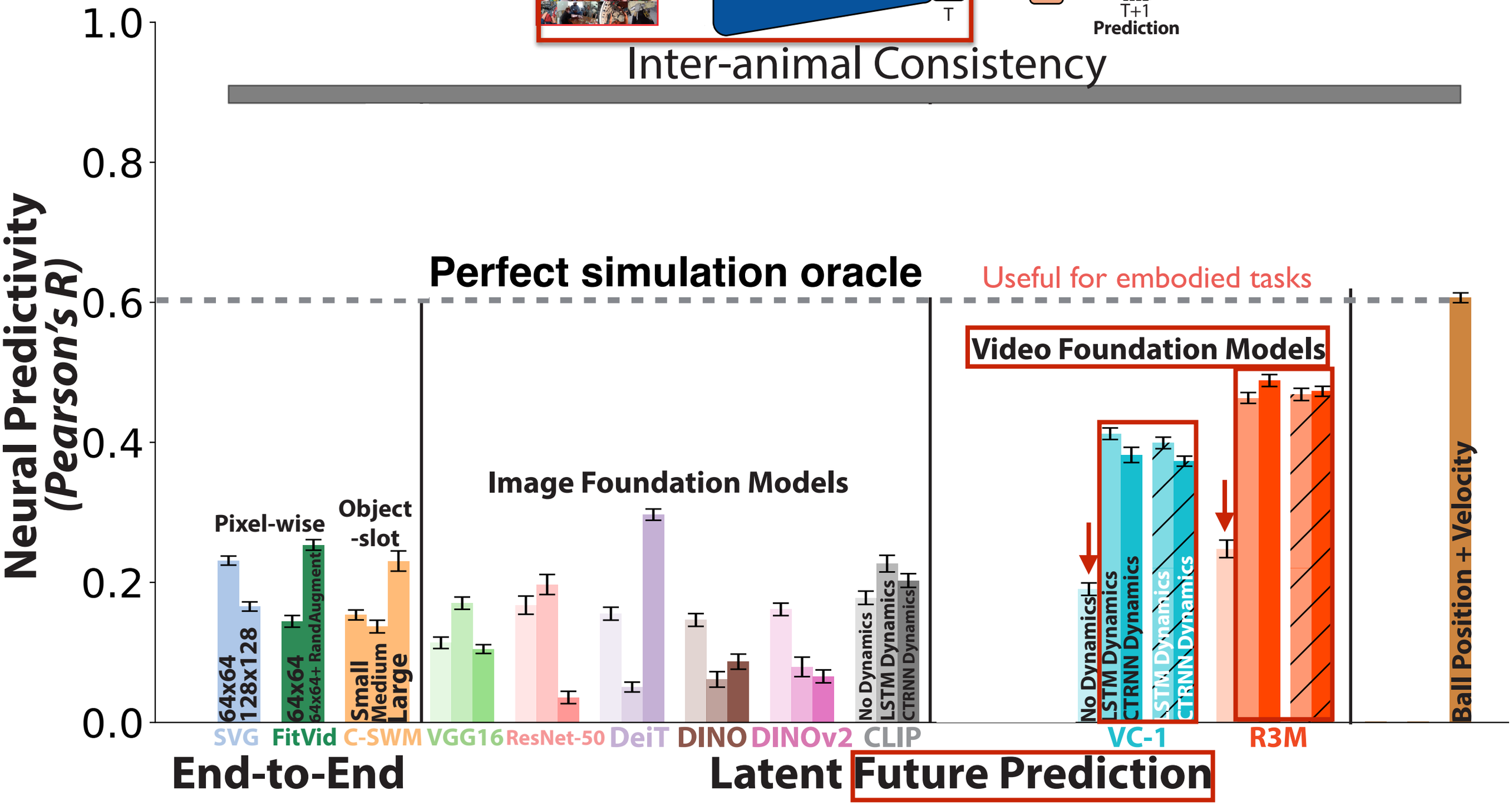


Video Foundation Future Prediction Best Predict Neurons

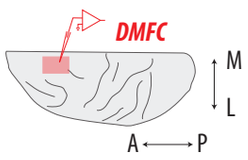


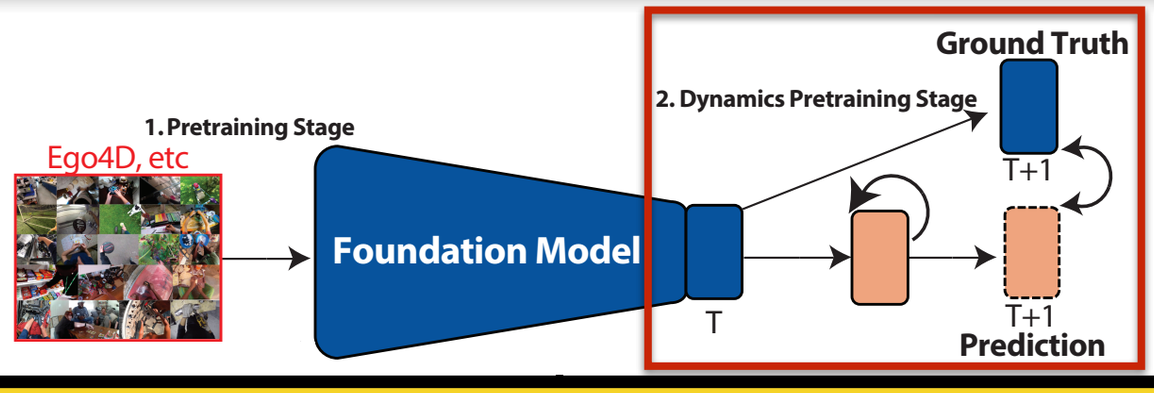


Inter-animal Consistency

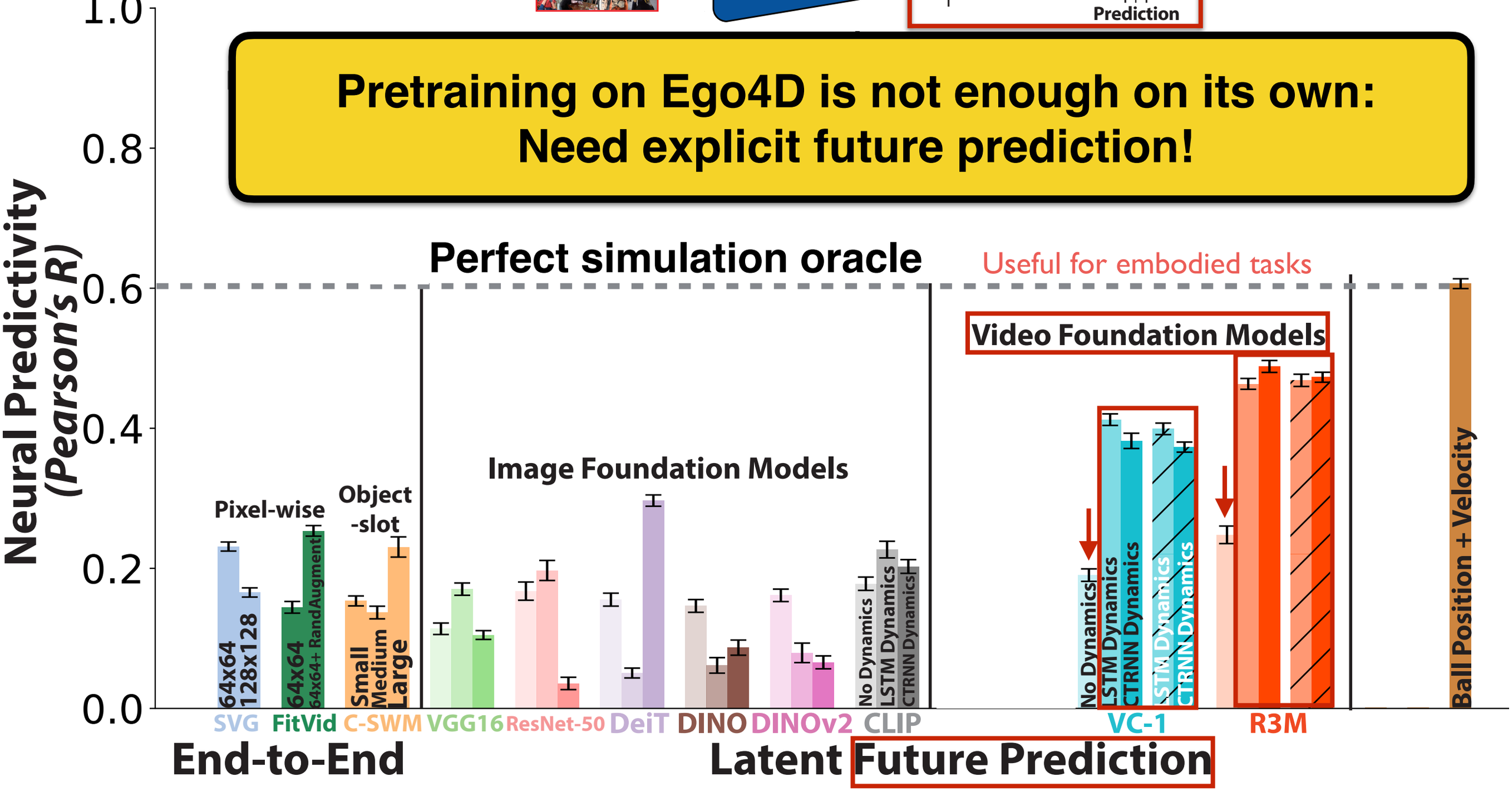


Video Foundation Future Prediction Best Predict Neurons





Pretraining on Ego4D is not enough on its own:
Need explicit future prediction!



Outline

- Role of Recurrent Processing During Object Recognition

Enables more parameter/unit efficient models that gain object recognition performance by unrolling "deeper" in time, rather than adding more layers.

Moreso than simply "convolutionizing" standard LSTMs/GRUs.

- Visually-Grounded Mental Simulation

The brain's mental simulations crucially involve explicit future prediction of a *factorized* visual scene description (*not* pixel-level!).

This factorization is strongly constrained. It does *not* appear to represent fixed object slots, but rather a critical component is for it to enable a wide range of OOD embodied abilities.

Acknowledgements

Daniel Yamins

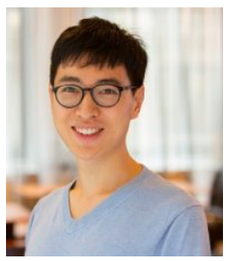
Daniel Bear

Rishi Rajalingham

Mehrdad Jazayeri

Surya Ganguli

Javier Sagastuy



Guangyu Robert Yang

Jonas Kubilius

Kohitij Kar

David Sussillo

Jim DiCarlo

Contact:

anayebi@cs.cmu.edu

@aran_nayebi

https://cs.cmu.edu/~anayebi

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Funding:

Burroughs Wellcome Fund CASI Award

K. Lisa Yang ICoN Postdoctoral Fellowship,
McGovern Institute, MIT

Stanford Neurosciences PhD Program

Stanford Mind, Brain, Computation and
Technology Training Program,
Wu Tsai Neurosciences Institute