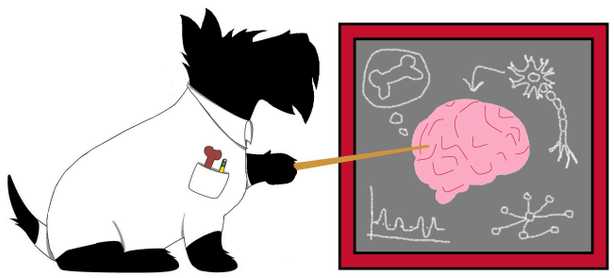
Using Embodied Agents for "Why" Questions in Systems Neuroscience

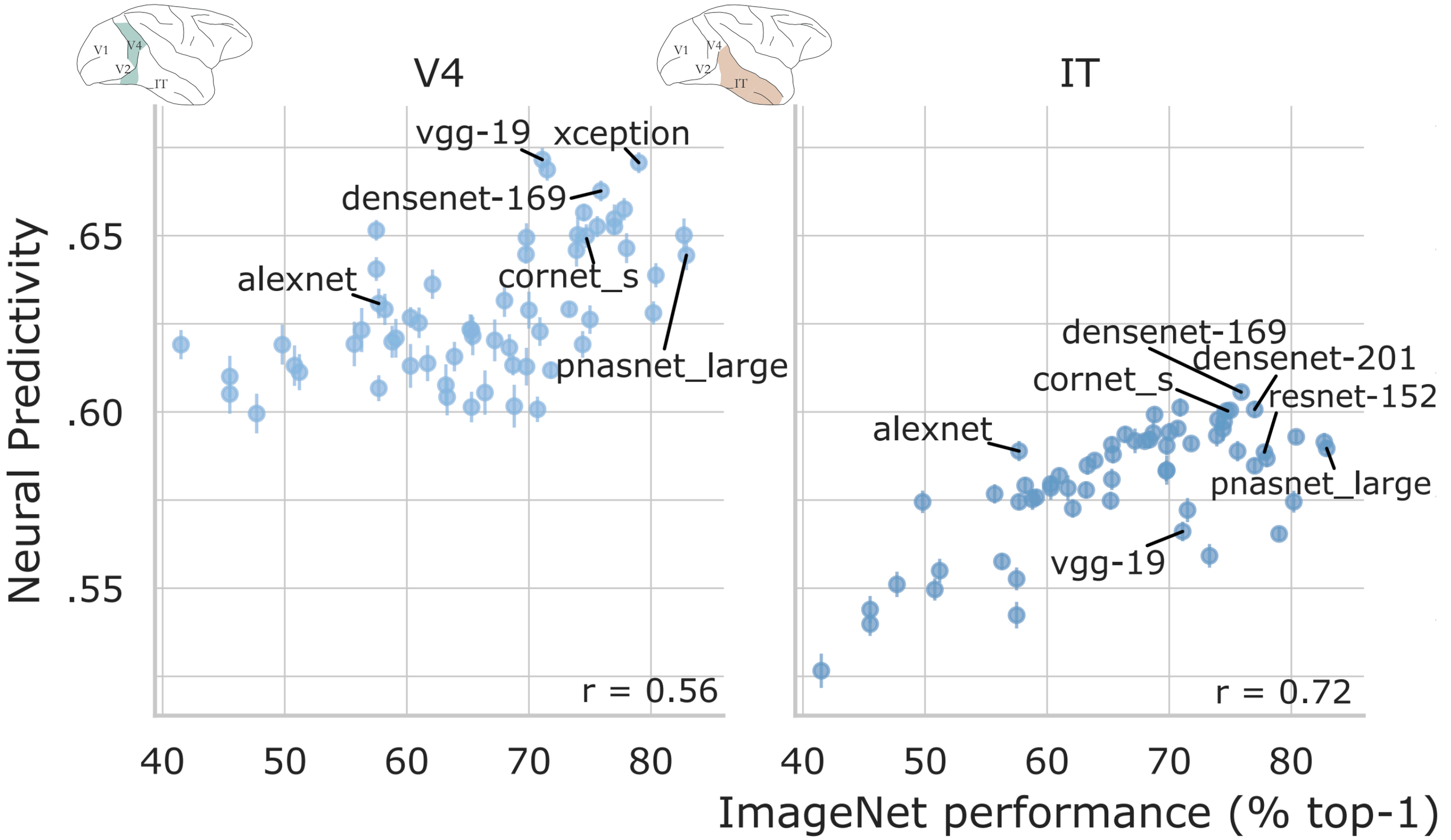
Aran Nayebi
Assistant Professor
Machine Learning Department
Neuroscience Institute (core faculty), Robotics Institute (courtesy)

EPFL BMI Seminar
2025.02.12



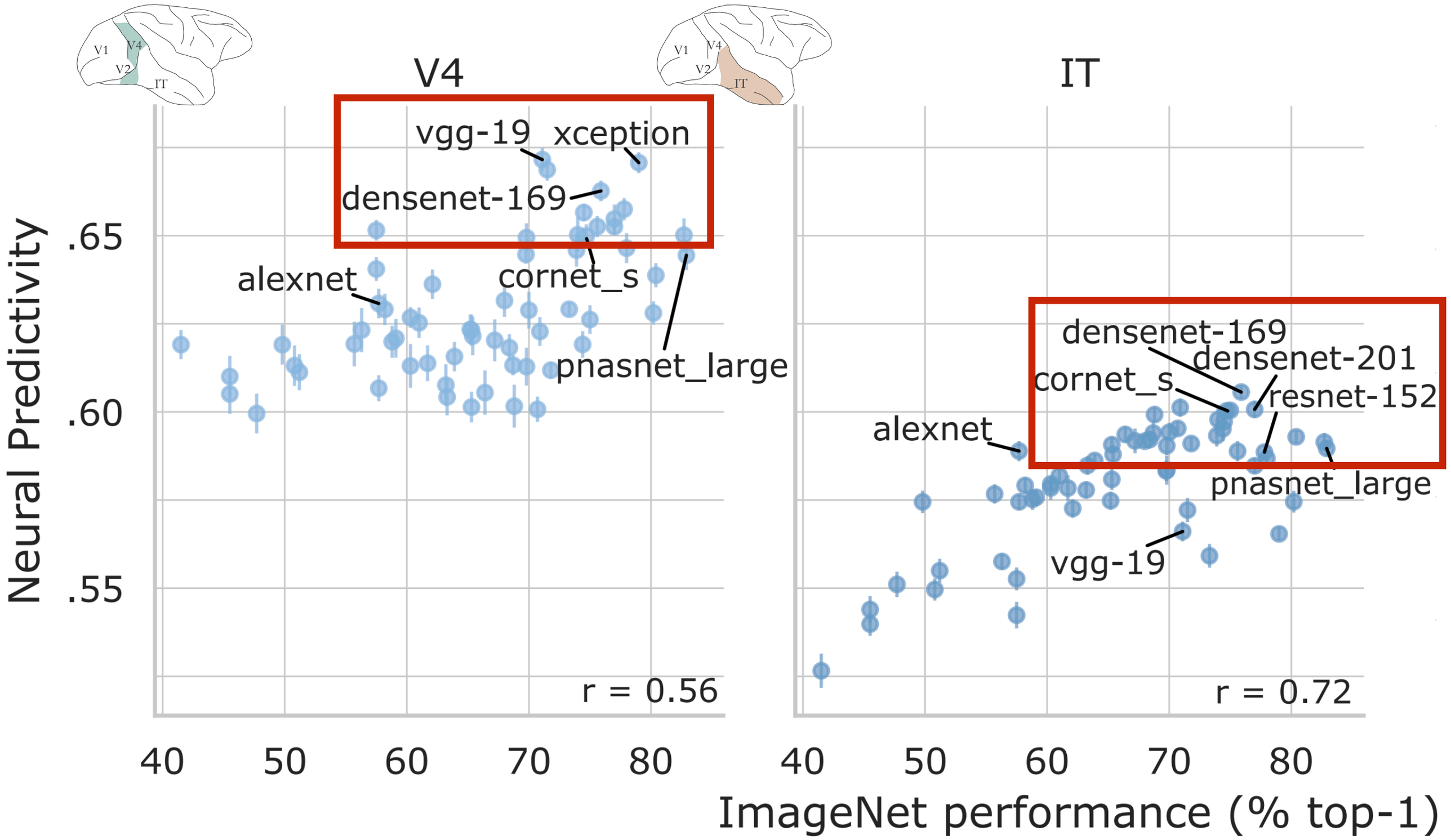
Carnegie Mellon University
Neuroscience Institute

Similar predictivities among very different CNN architectures



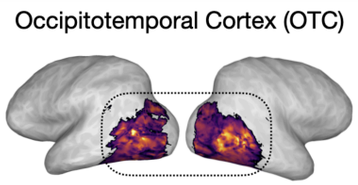
Schrimpf, Kubilius* et al. 2018*

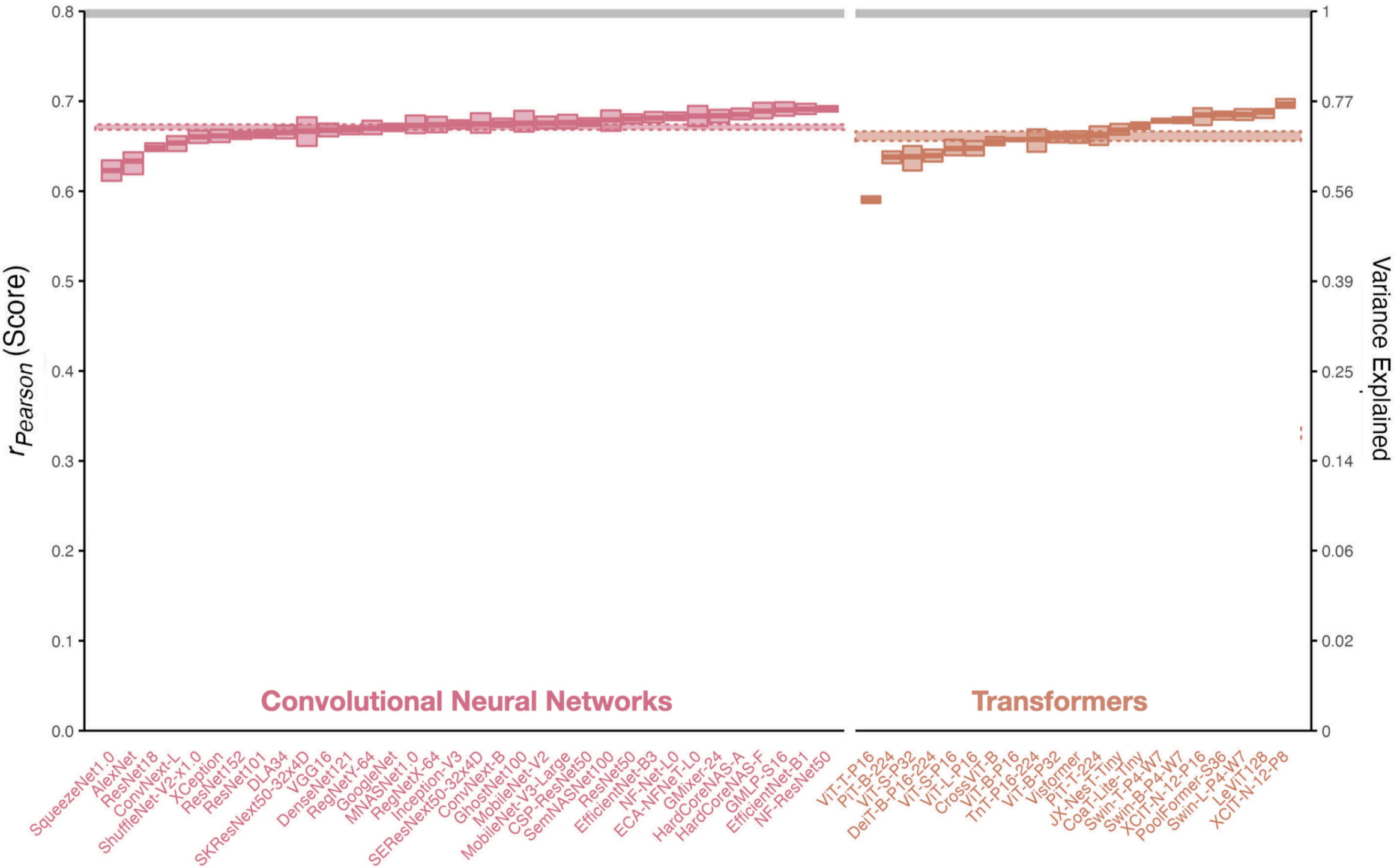
Similar predictivities among very different CNN architectures



Schrimpf, Kubilius* et al. 2018*

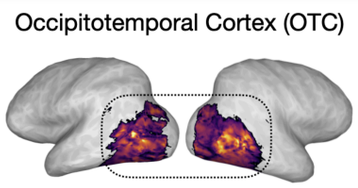
Similar predictivities between CNNs vs. Transformers

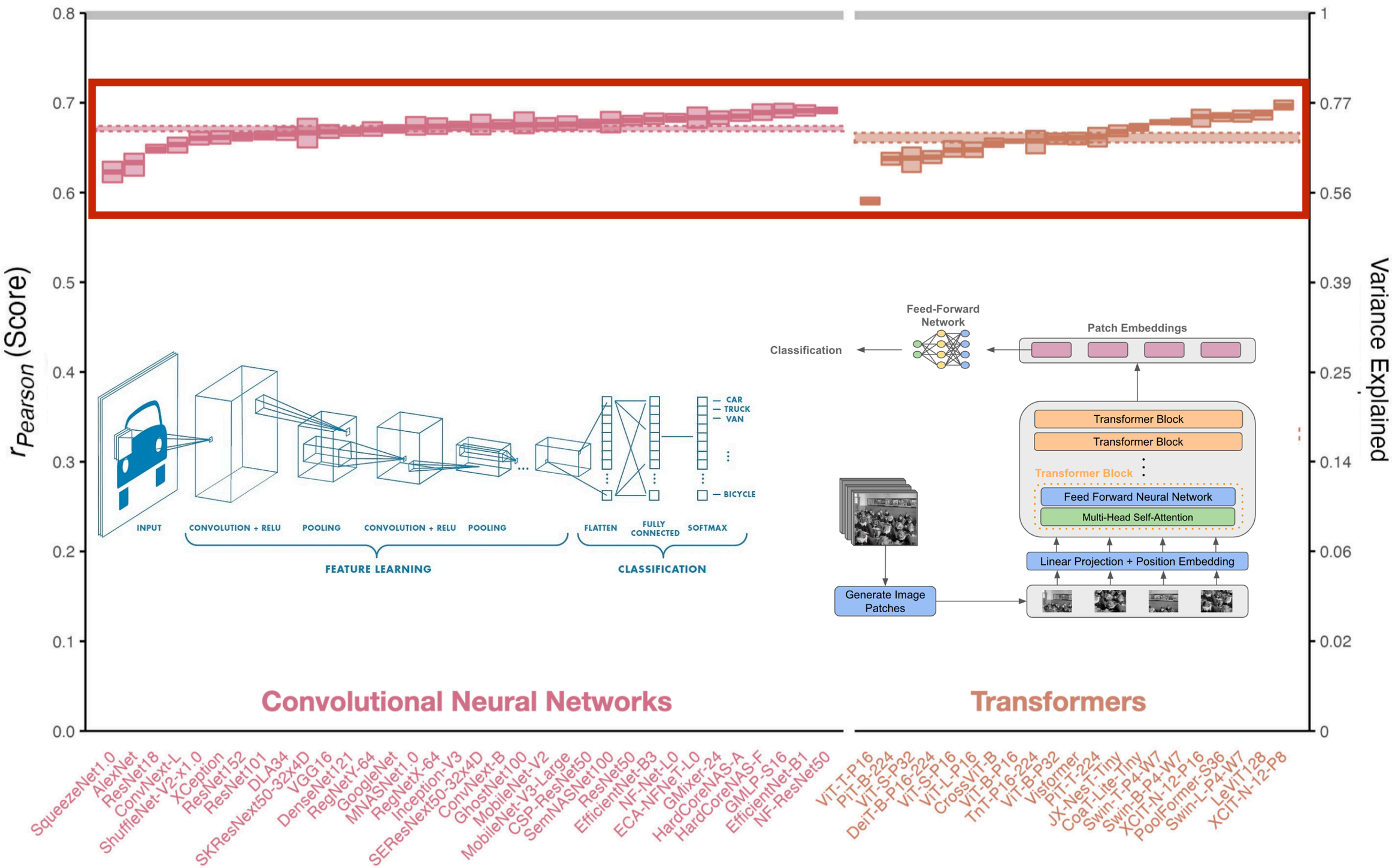




Conwell et al. 2023

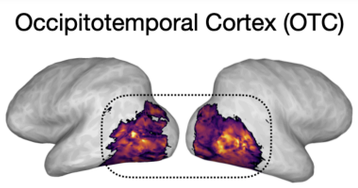
Similar predictivities between CNNs vs. Transformers

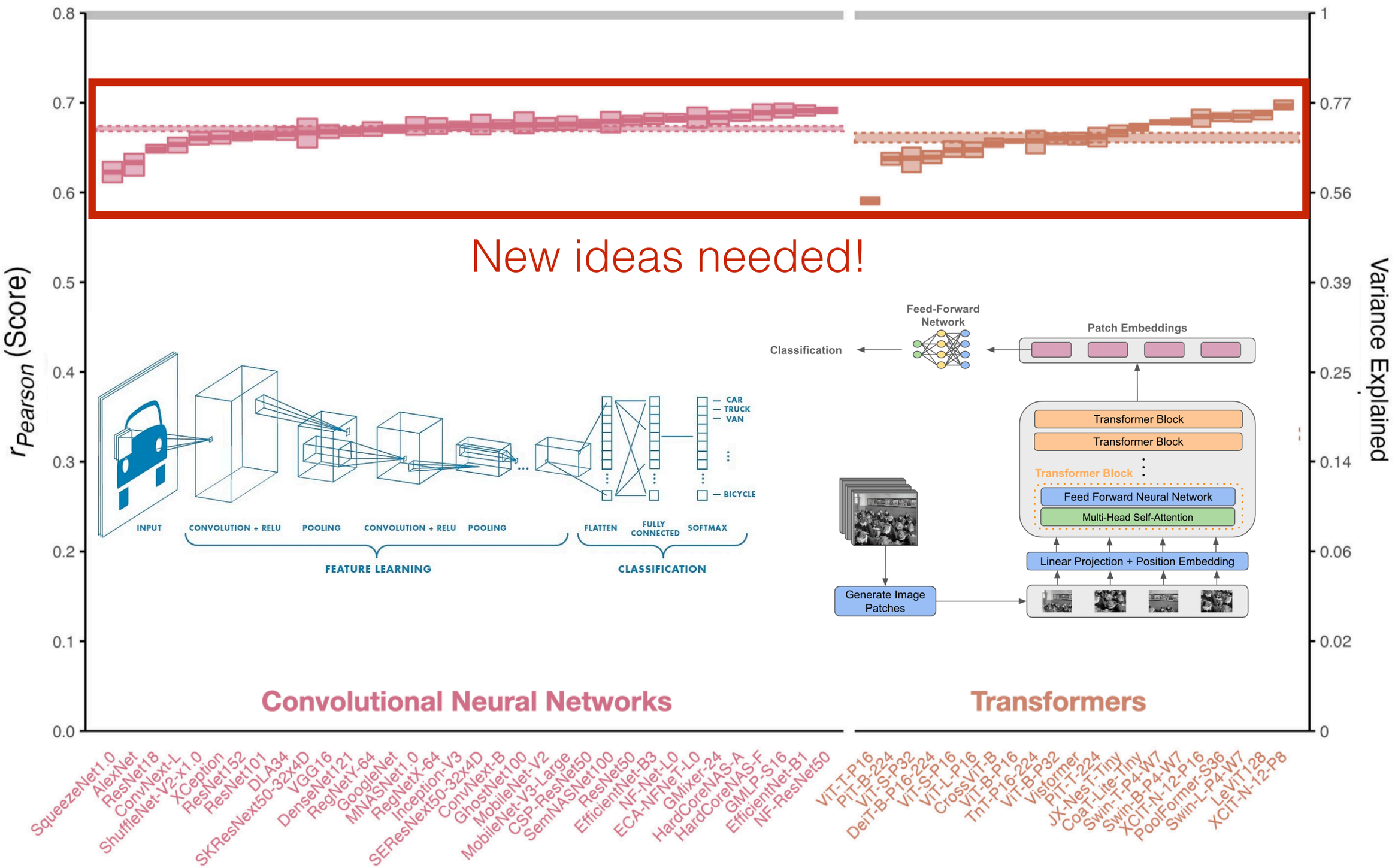




Conwell et al. 2023

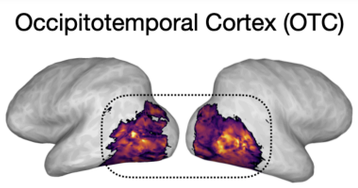
Similar predictivities between CNNs vs. Transformers





Conwell et al. 2023

Similar predictivities between CNNs vs. Transformers



Input (Diet) — ImageNet-1k

Architecture — CNNs versus Transformers

Task — 1K Classification

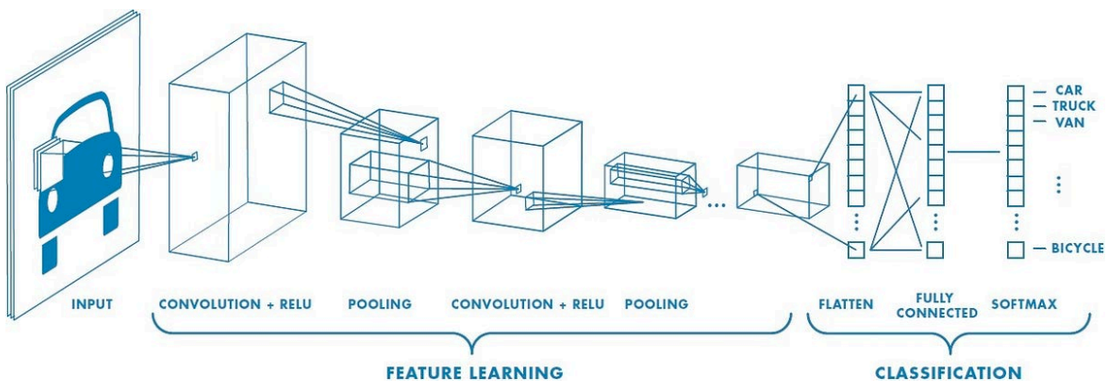
Models? New ideas needed!

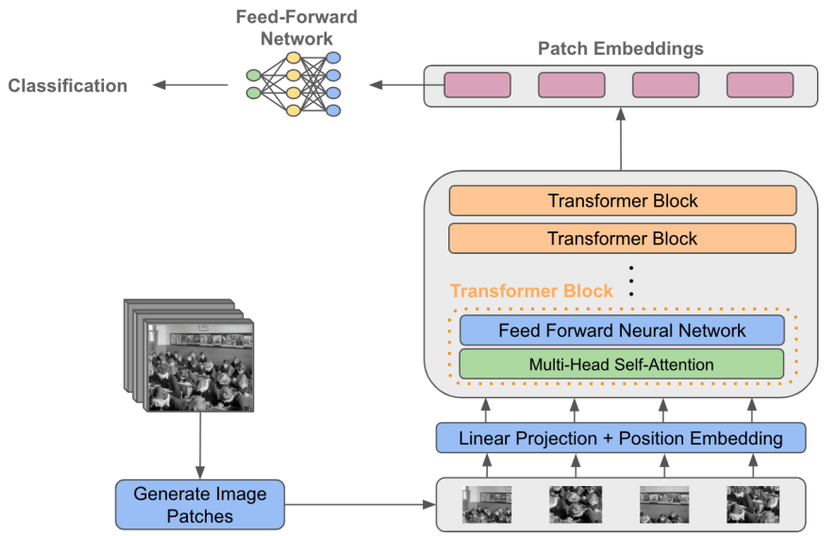
$r_{Pearson}$ (Score)

Variance Explained

0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

1 0.77 0.56 0.39 0.25 0.14 0.06 0.02 0





Convolutional Neural Networks

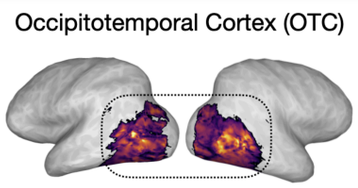
SqueezeNet1.0, AlexNet, ResNet18, ConvNext-L, ShuffleNet-V2-x1.0, XCeption, ResNet152, ResNet101, DLA34, SKResNext50-32x4D, VGG16, DenseNet121, RegNetY-64, GoogleNet, MNASNet1.0, RegNetX-64, Inception-V3, SEResNext50-32x4D, ConvNext-B, GhostNet100, MobileNet-V2, MobileNet-V3-Large, CSP-ResNet50, SemNASNet100, ResNet50, EfficientNet-B3, NF-Net-L0, ECA-NFNeT-L0, GMixer-24, HardCoreNAS-A, HardCoreNAS-F, GMLP-S16, EfficientNet-B1, NF-ResNet50

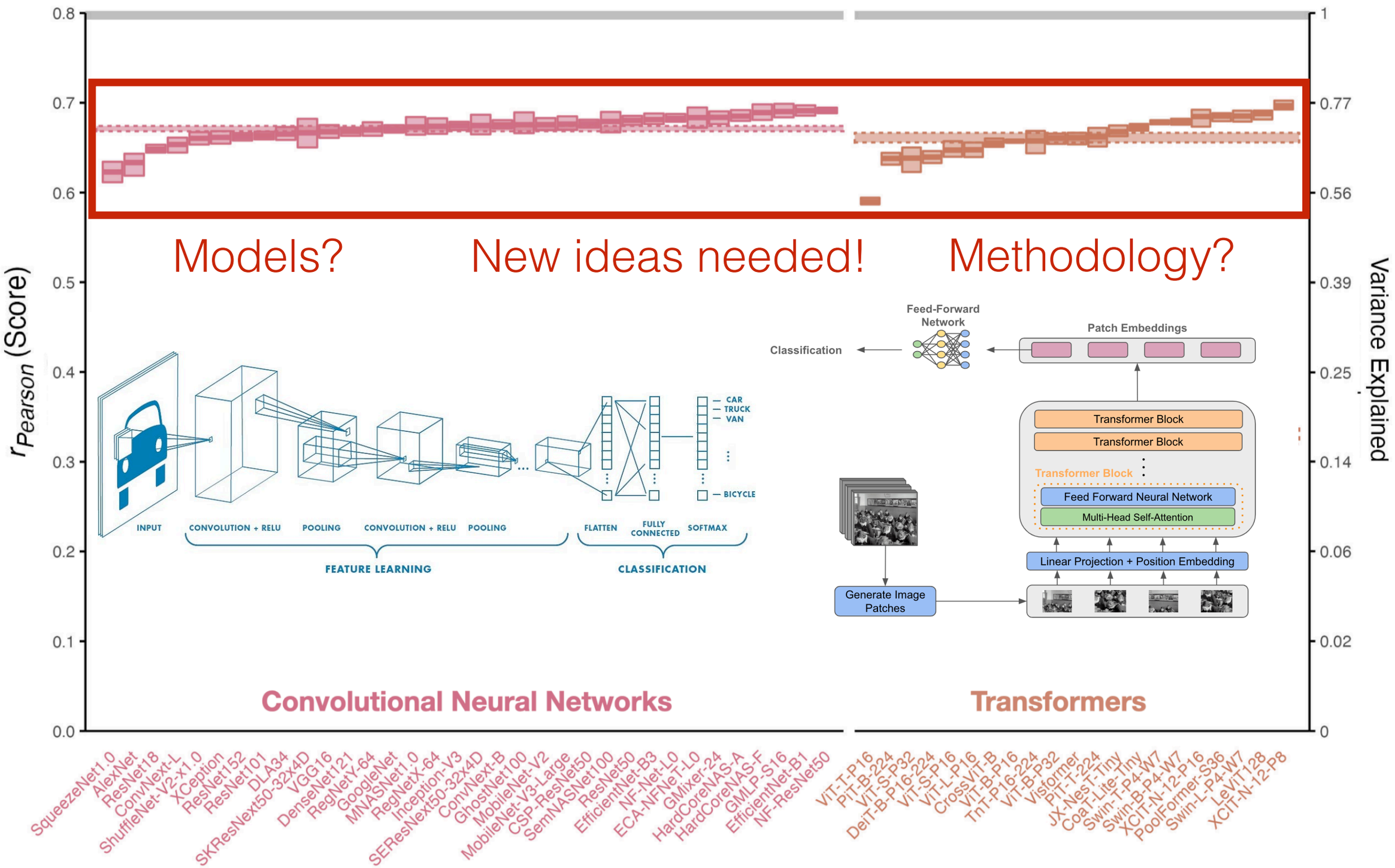
Transformers

ViT-T-P16, PiT-B-224, ViT-S-P32, DeiT-B-P16-224, ViT-S-P16, ViT-L-P16, CrossViT-B, ViT-B-P16, TnT-P16-224, ViT-B-P32, Visformer, PiT-T-224, JX-NesT-Tiny, CoaT-Lite-Tiny, Swin-T-P4-W7, Swin-B-P4-W7, XCiT-N-12-P16, PoolFormer-S36, Swin-L-P4-W7, LeViT128, XCiT-N-12-P8

Conwell et al. 2023

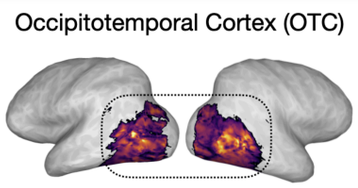
Similar predictivities between CNNs vs. Transformers





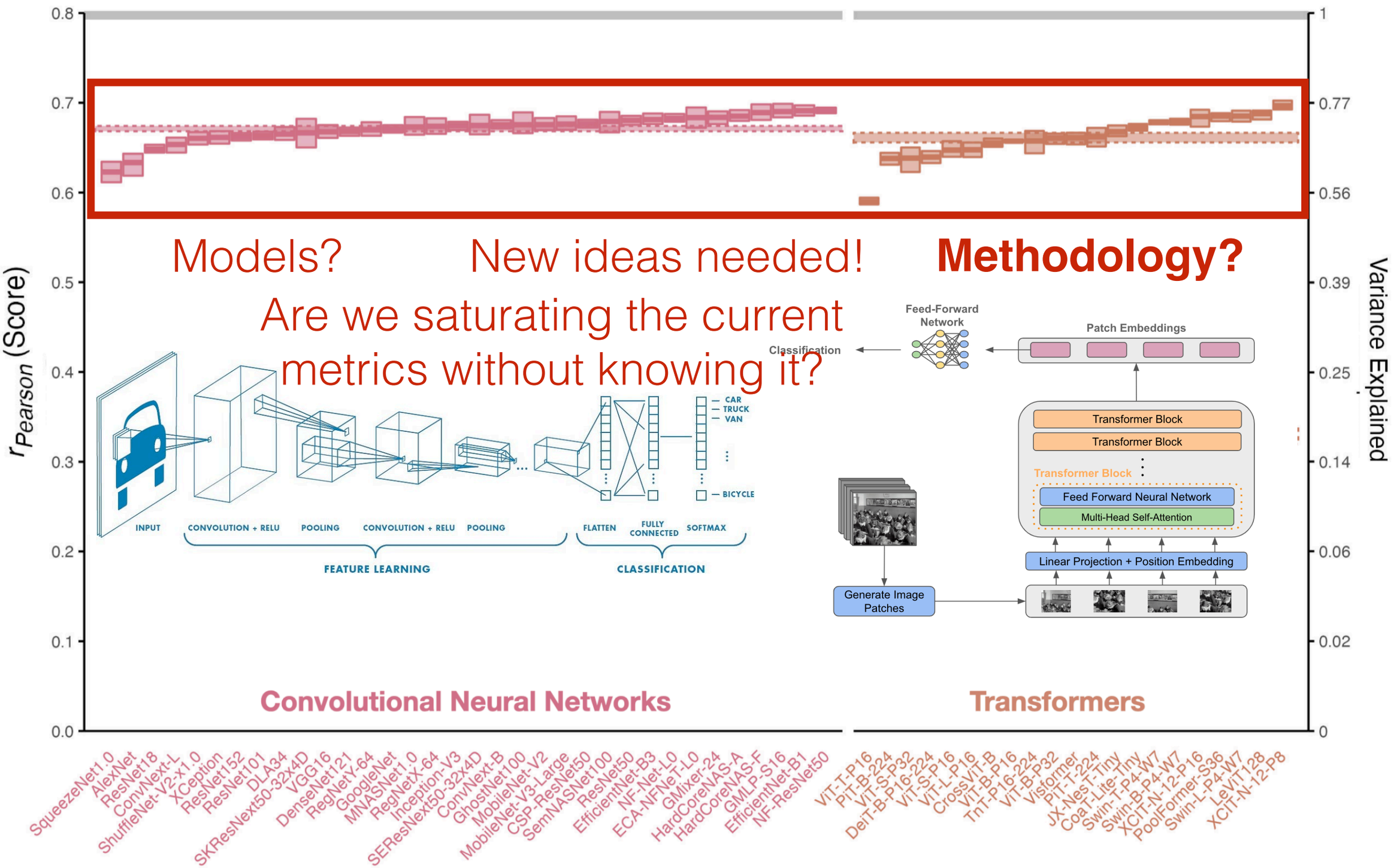
Conwell et al. 2023

Similar predictivities between CNNs vs. Transformers



Models? New ideas needed! **Methodology?**

Are we saturating the current metrics without knowing it?



Conwell et al. 2023

Determining the Methodology

Determining the Methodology

We don't have an *a priori* notion of what constitutes a "good" brain-model in an *absolute* sense

Determining the Methodology

We don't have an *a priori* notion of what constitutes a "good" brain-model in an *absolute* sense

…if we did, we would just put it in the model and be done already!

Determining the Methodology

Brain-Model Evaluations Need the NeuroAI Turing Test

Jenelle Feather [*1] **Meenakshi Khosla** [*2] **N. Apurva Ratan Murty** [*3] **Aran Nayebi** [*4]

We don't have an *a priori* notion of what constitutes a "good" brain-model in an *absolute* sense

…if we did, we would just put it in the model and be done already!

Determining the Methodology

Brain-Model Evaluations Need the NeuroAI Turing Test

Jenelle Feather [*1] **Meenakshi Khosla** [*2] **N. Apurva Ratan Murty** [*3] **Aran Nayebi** [*4]

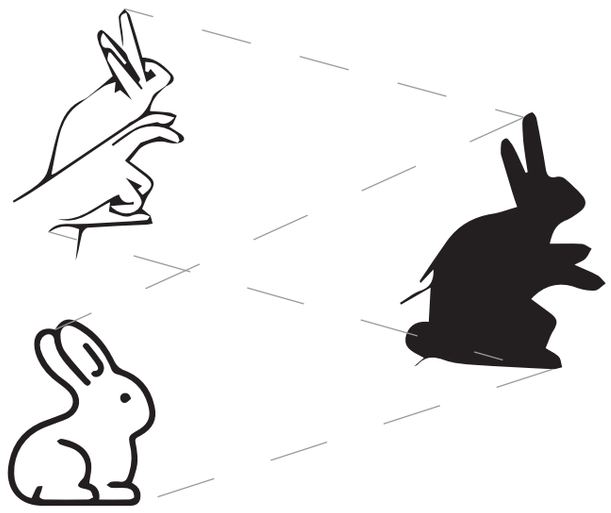
We don’t have an *a priori* notion of what constitutes a “good” brain-model in an *absolute* sense

…if we did, we would just put it in the model and be done already!

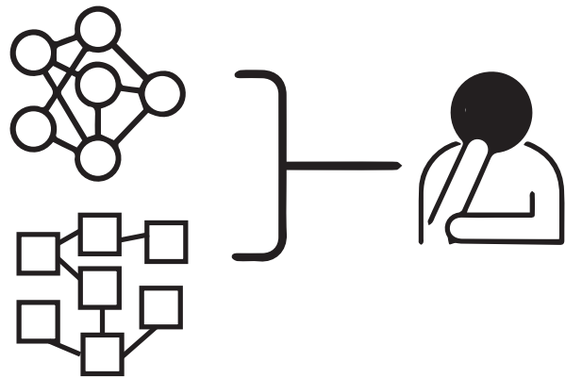
But, we do know that animals/subjects contain those desired properties (even if we can’t describe them), so we instead take a *relative* perspective

NeuroAI Turing Test

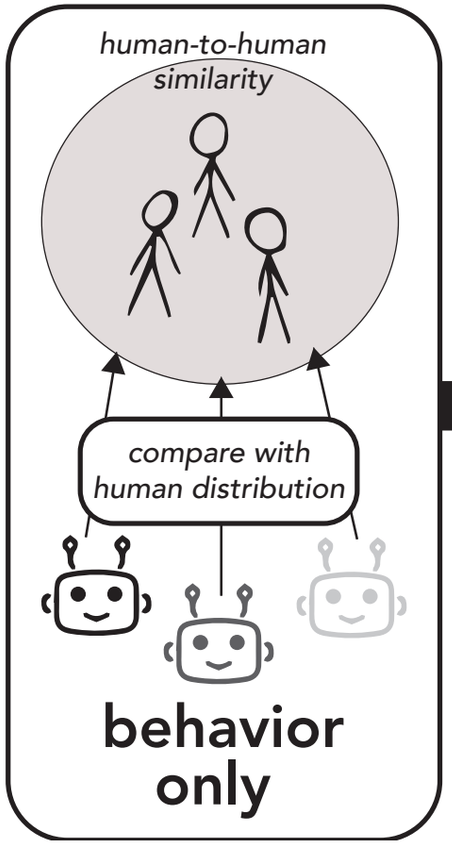
Just as distinct objects can cast the same shadow...



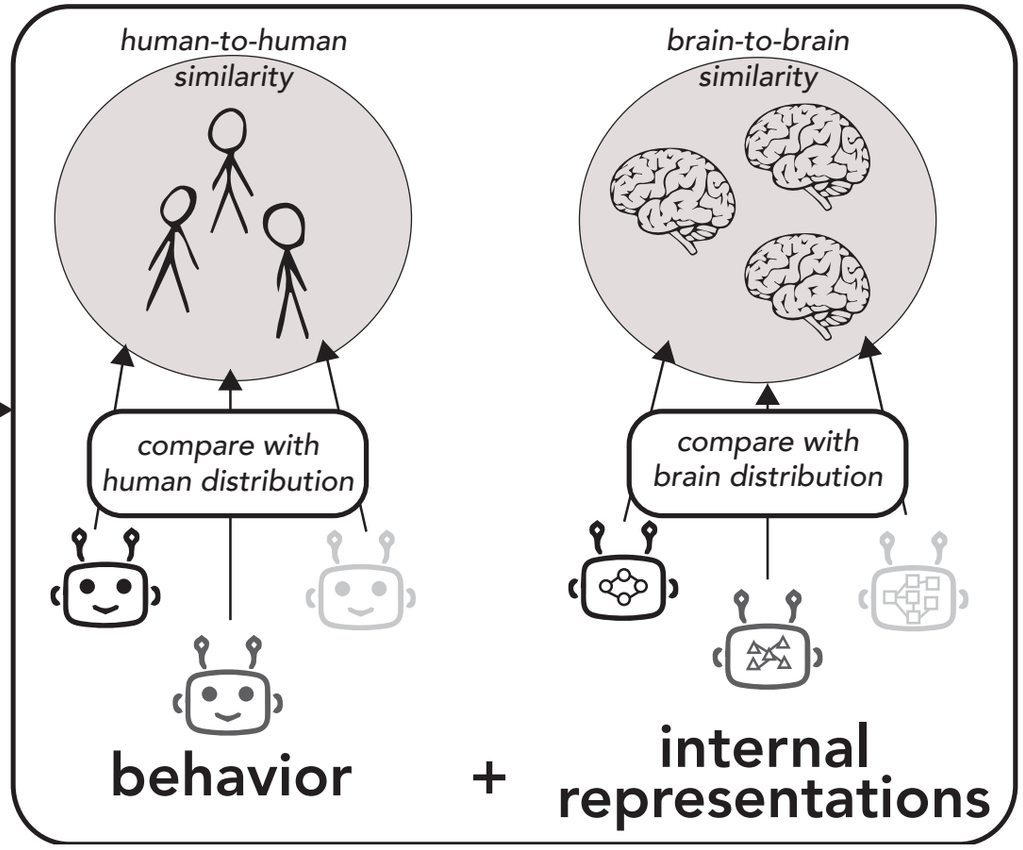
...distinct internal processes (representations) can produce similar outputs (behavior)



Turing Test

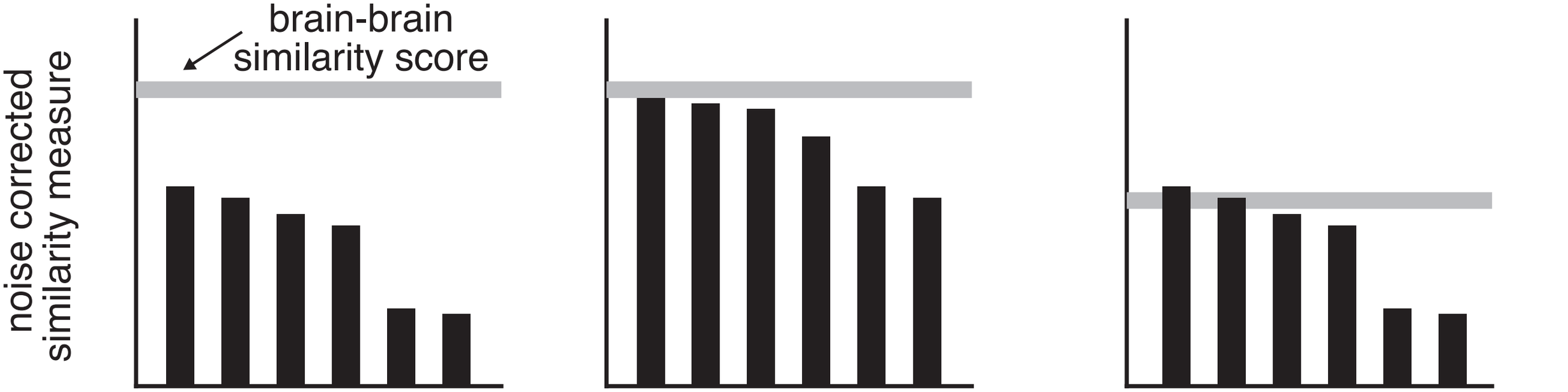


The NeuroAI Turing Test



NeuroAI Turing Test: Possible Scenarios

NeuroAI Turing Test: Possible Scenarios



No models reach brain-brain similarity

↓

Continue to refine models to improve on the benchmark

Models reach brain-brain similarity, brain-brain similarity close to noise ceilng

↓

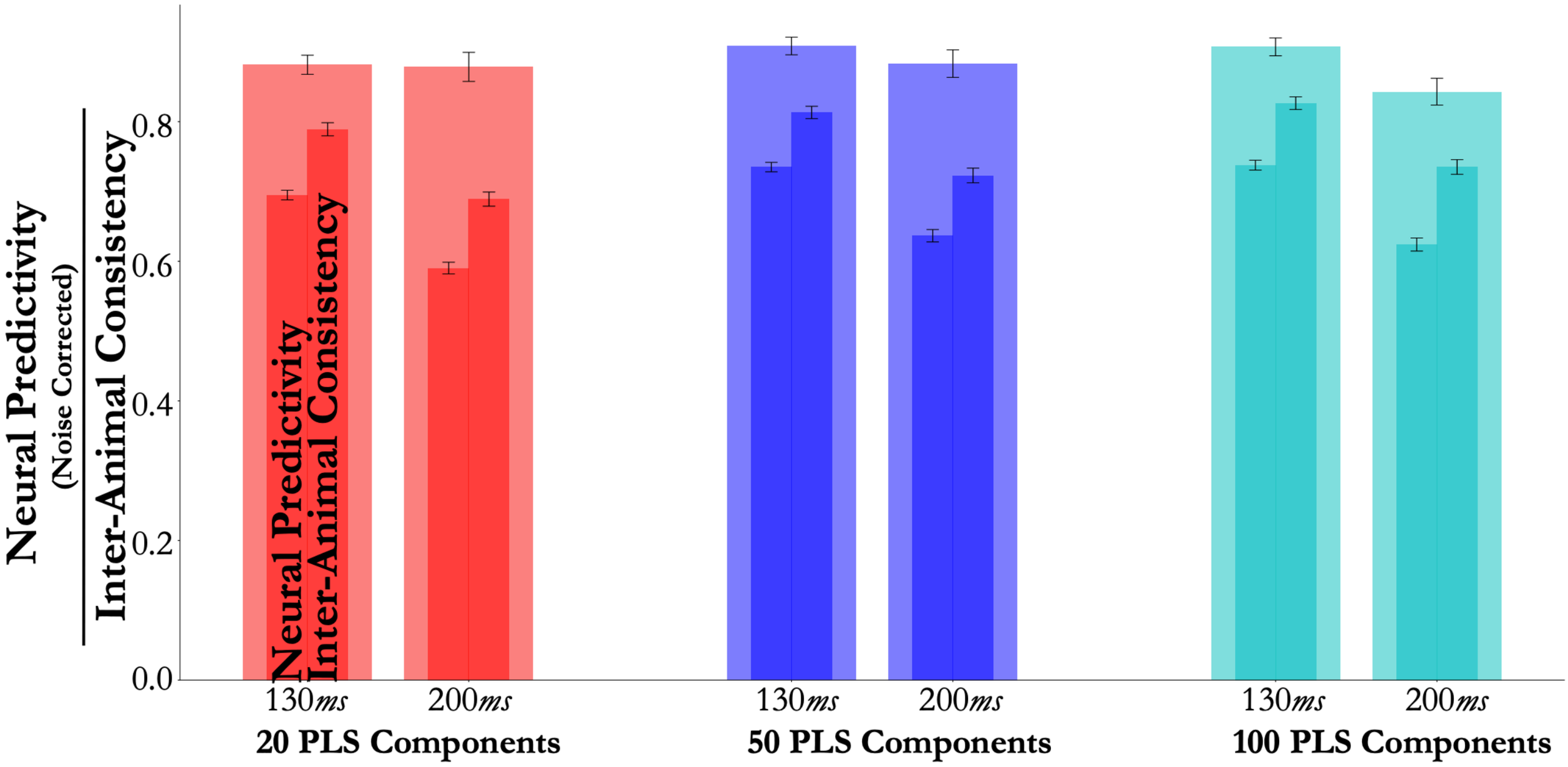
Benchmark is saturated, move on to other benchmarks.

Models reach or surpass brain-brain similarity, but brain-brain similarity is low

↓

If high data quality, this brain region or behavior has individual variability, and the model reflects the average.

NeuroAI Turing Test: An Example in 200 ms Vision

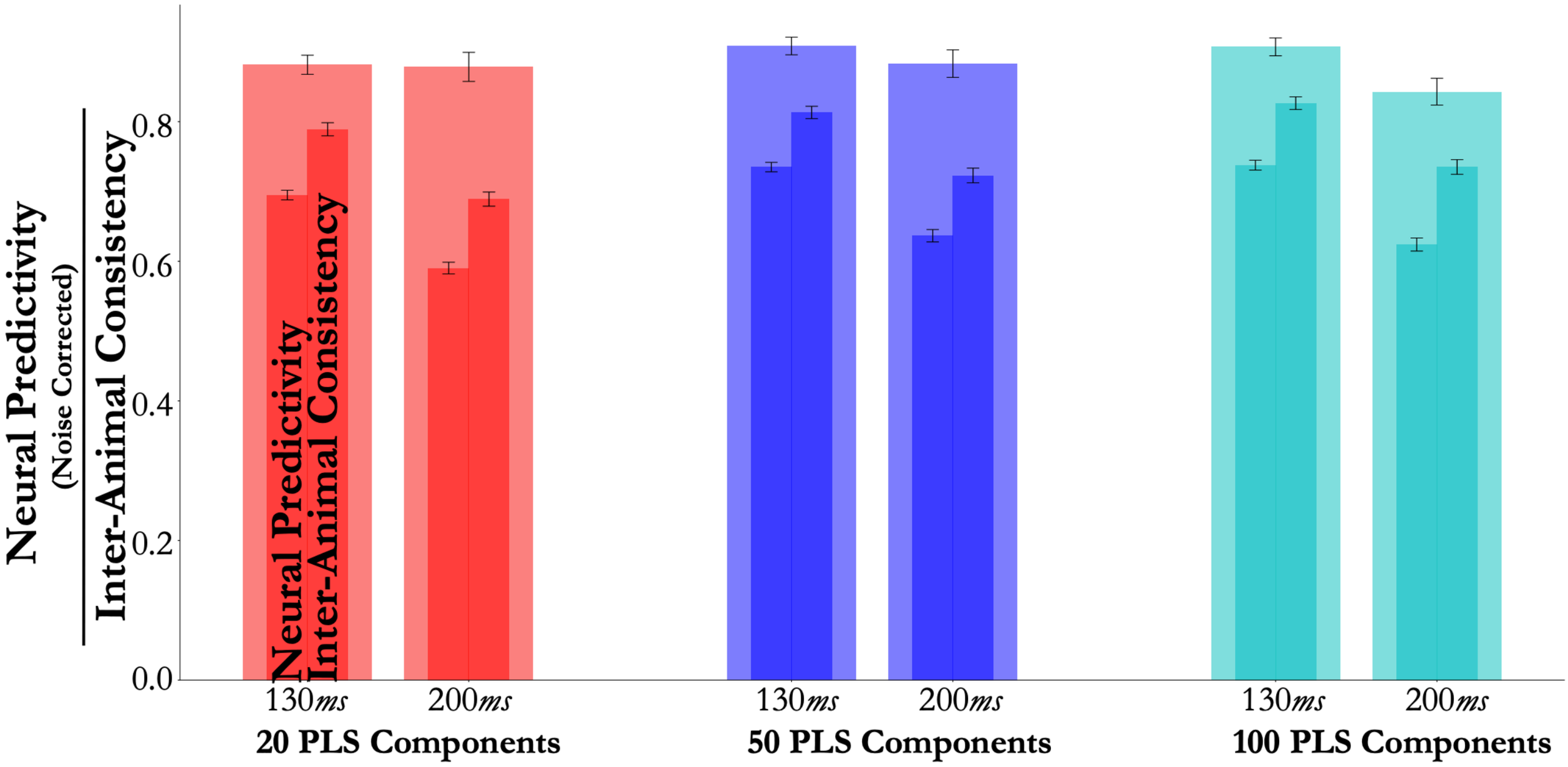


Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition

Aran Nayebi[1,*], Javier Sagastuy-Brena[1], Daniel M. Bear[1], Kohitij Kar[2], Jonas Kubilius[2,3], Surya Ganguli[1], David Sussillo[1], James J. DiCarlo[2], and Daniel L. K. Yamins[1]

NeuroAI Turing Test: An Example in 200 ms Vision

So we may have achieved it for the classic “HvM” dataset (it’s tapped out!)



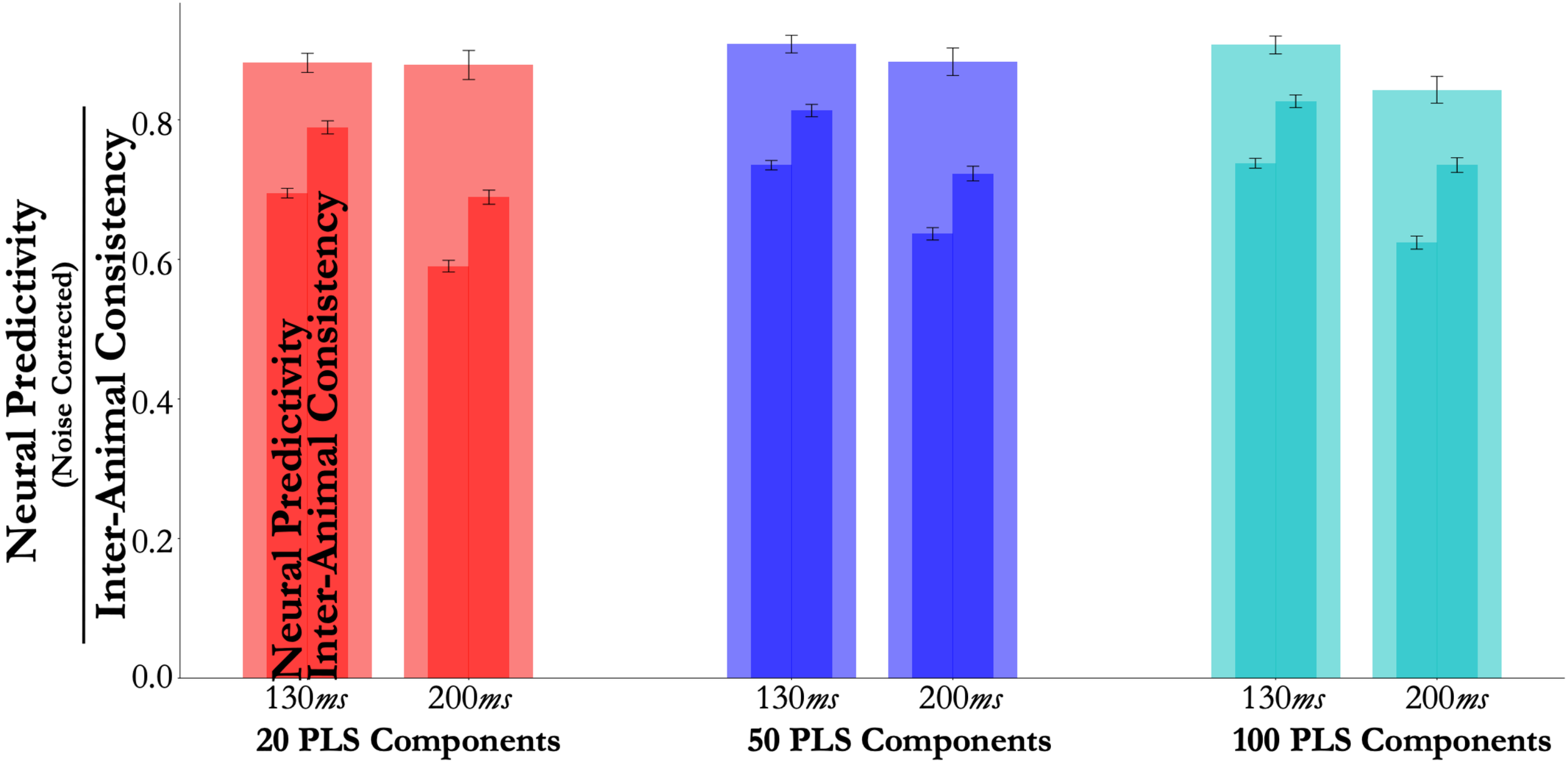
Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition

Aran Nayebi[1,*], Javier Sagastuy-Brena[1], Daniel M. Bear[1], Kohitij Kar[2], Jonas Kubilius[2,3], Surya Ganguli[1], David Sussillo[1], James J. DiCarlo[2], and Daniel L. K. Yamins[1]

NeuroAI Turing Test: An Example in 200 ms Vision

So we may have achieved it for the classic “HvM” dataset (it’s tapped out!)

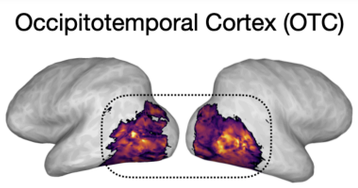
But I will show you that we haven’t yet in embodied intelligence…

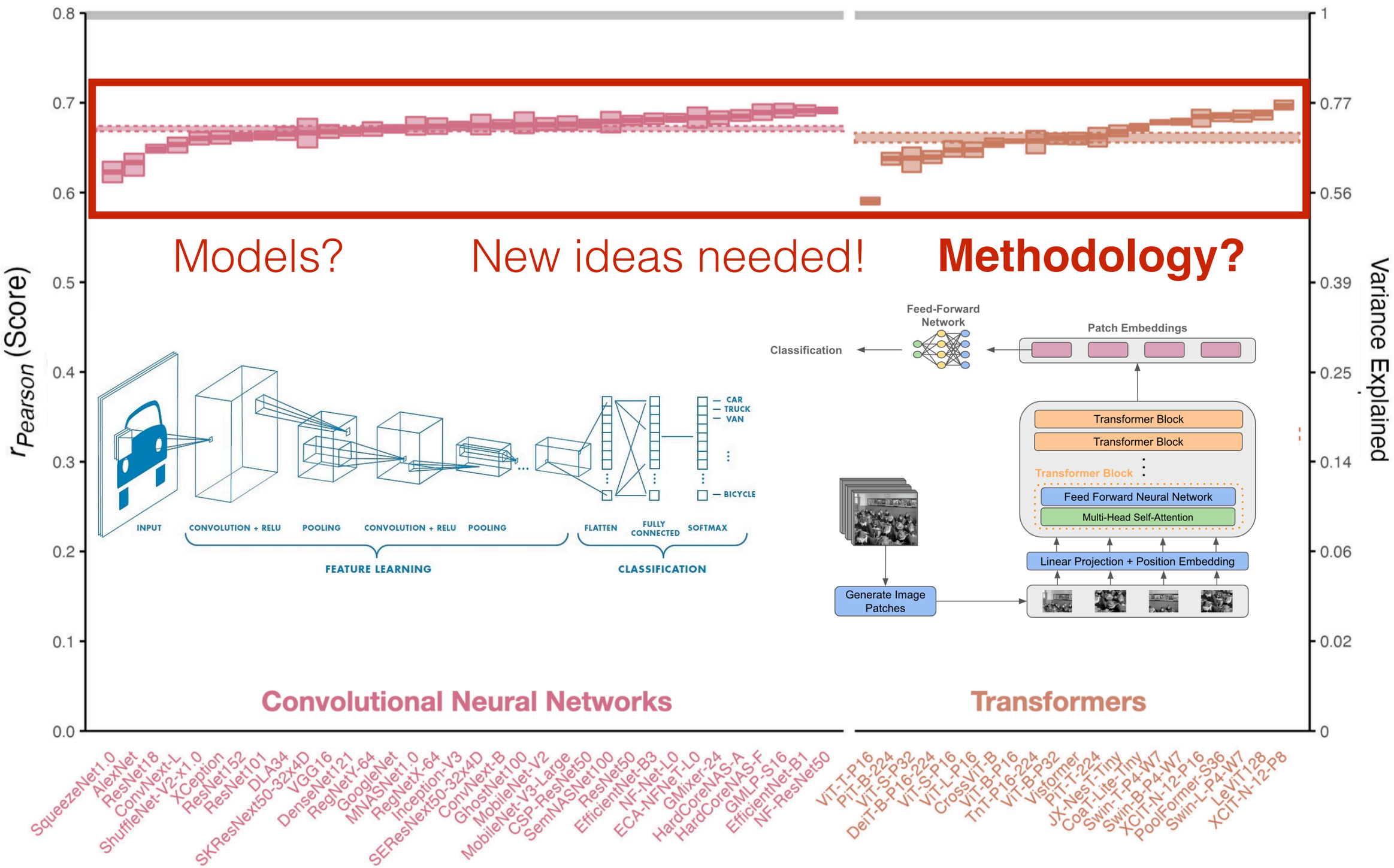


Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition

Aran Nayebi[1,*], Javier Sagastuy-Brena[1], Daniel M. Bear[1], Kohitij Kar[2], Jonas Kubilius[2,3], Surya Ganguli[1], David Sussillo[1], James J. DiCarlo[2], and Daniel L. K. Yamins[1]

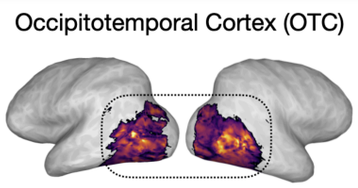
Similar predictivities between CNNs vs. Transformers

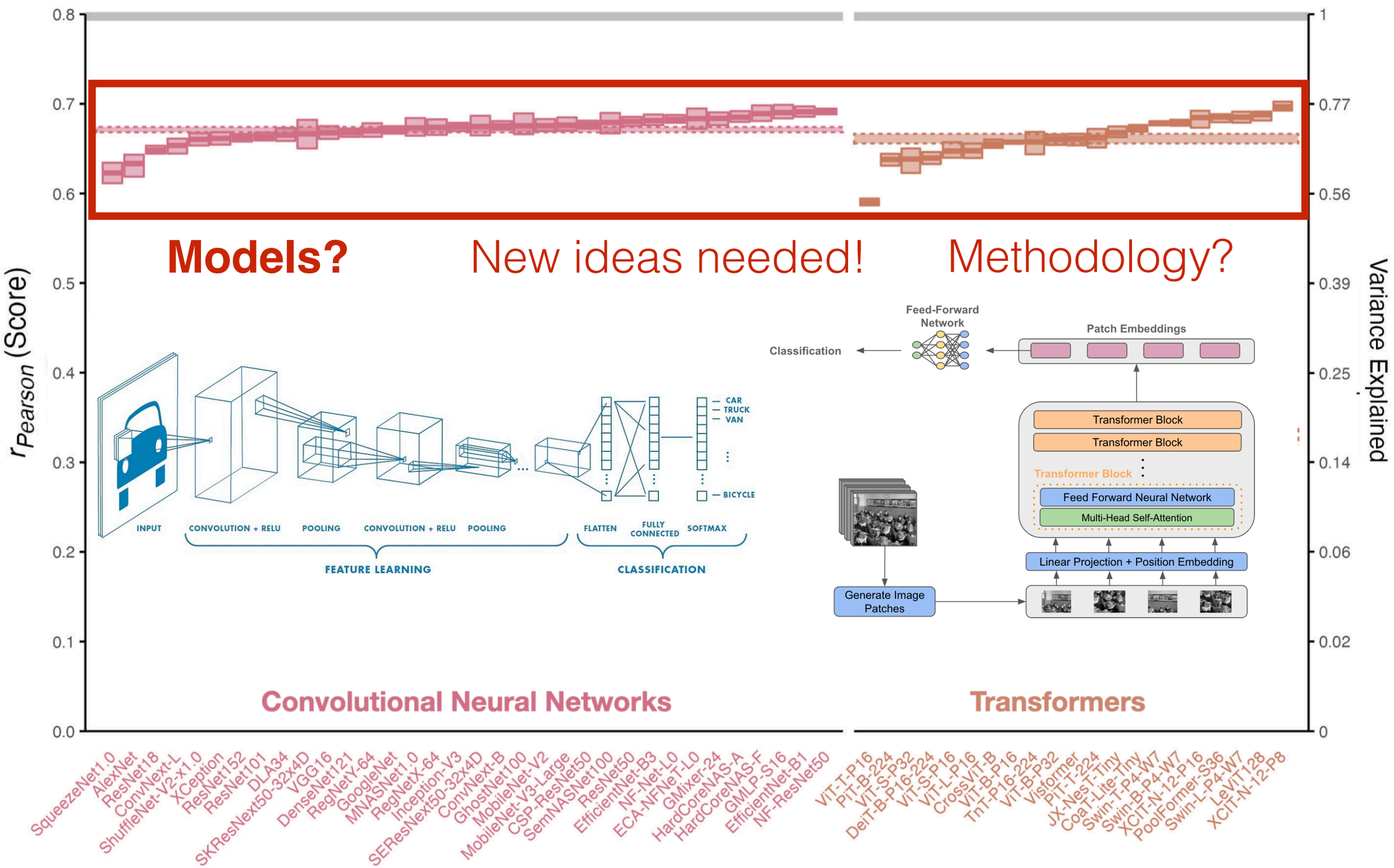




Conwell et al. 2023

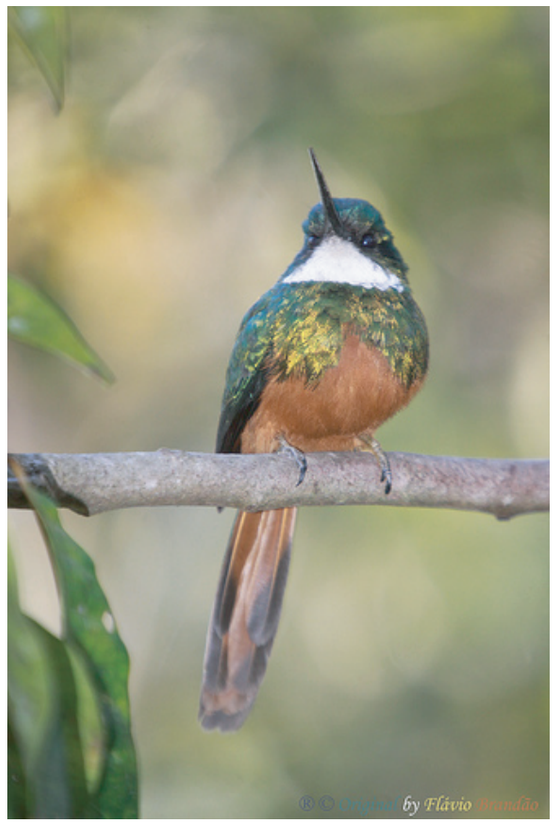
Similar predictivities between CNNs vs. Transformers



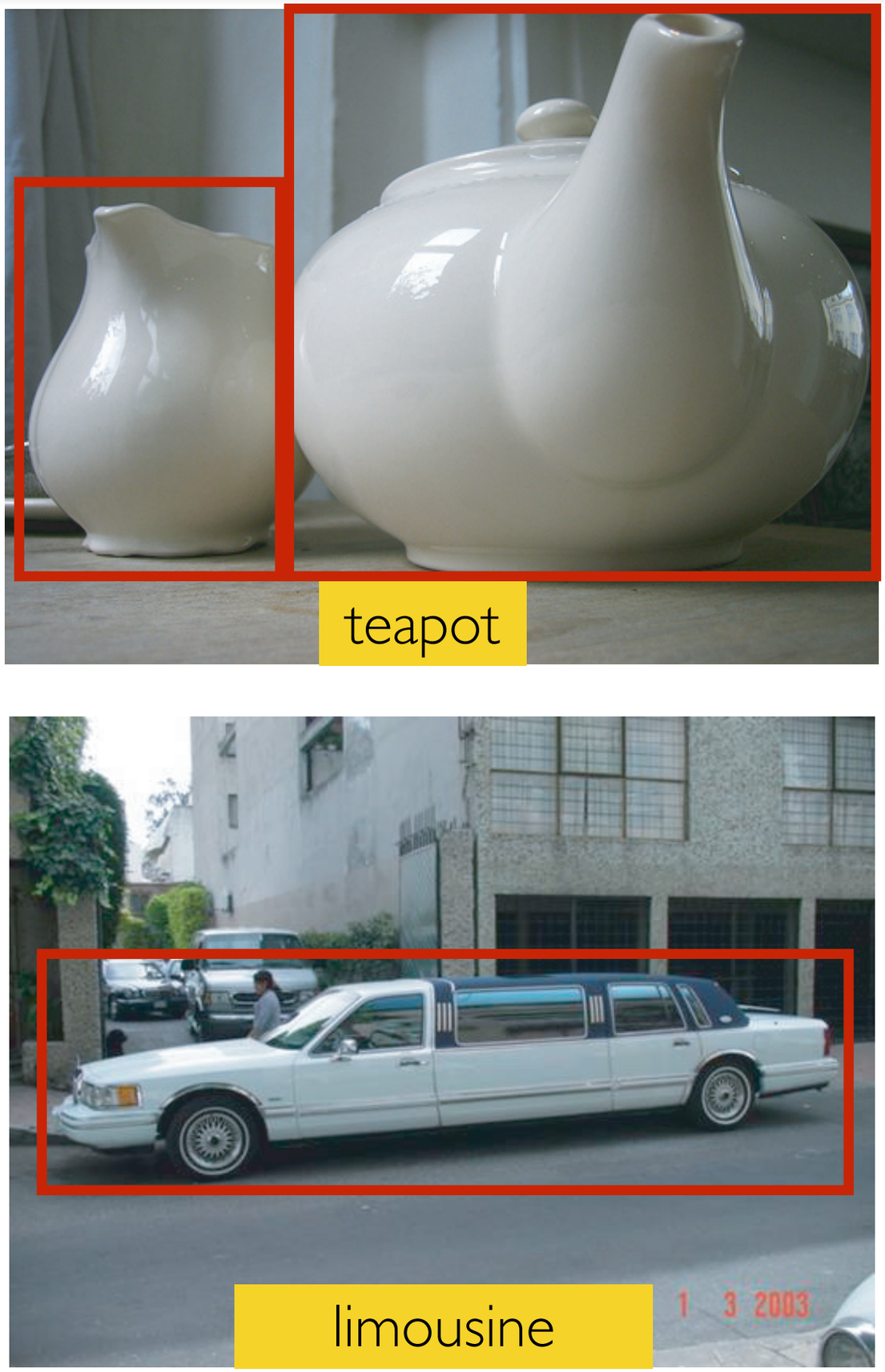


Conwell et al. 2023

We do a lot more than passive viewing…

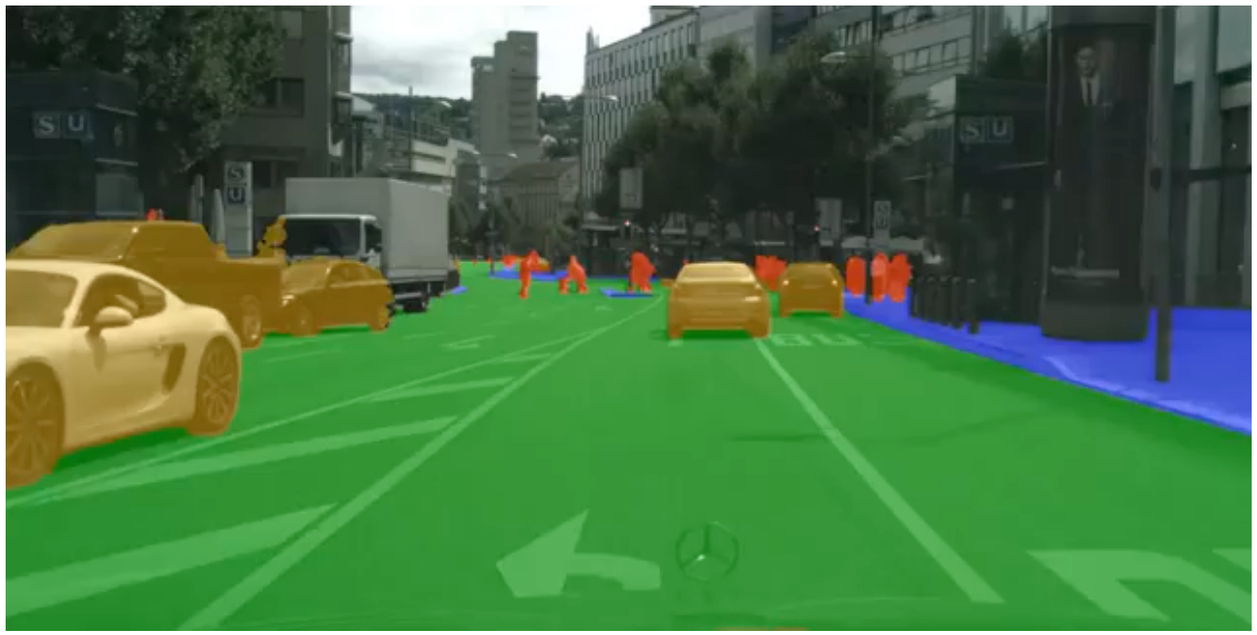


We do a lot more than passive viewing…



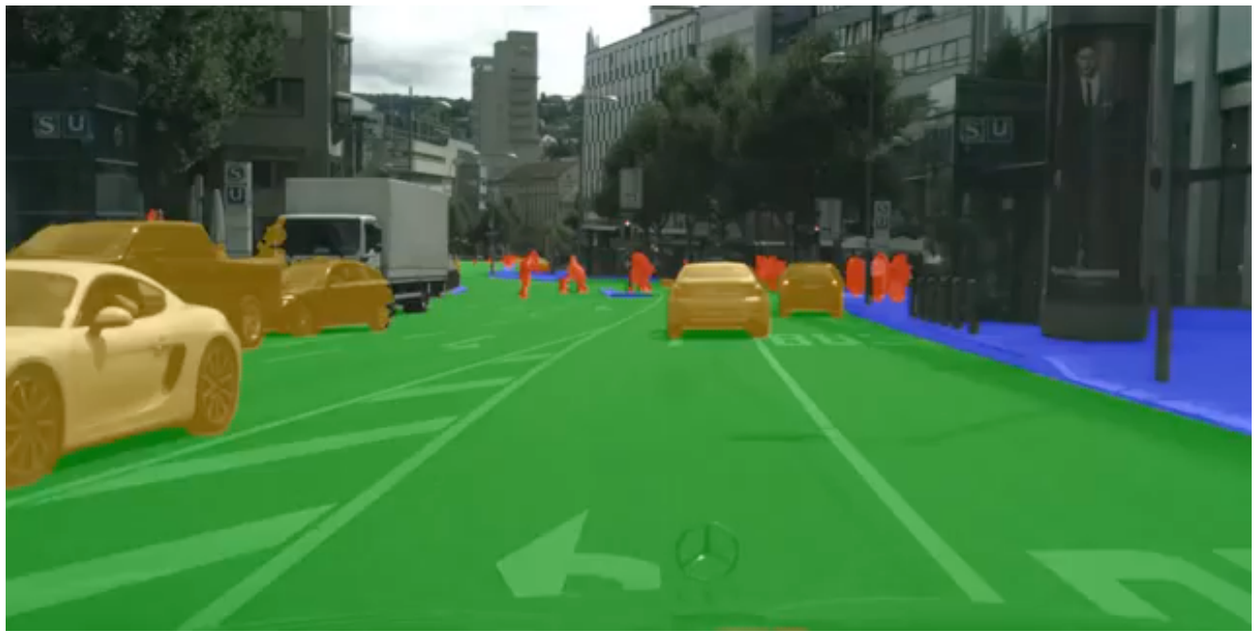
We do a lot more than passive viewing…

Scene Understanding



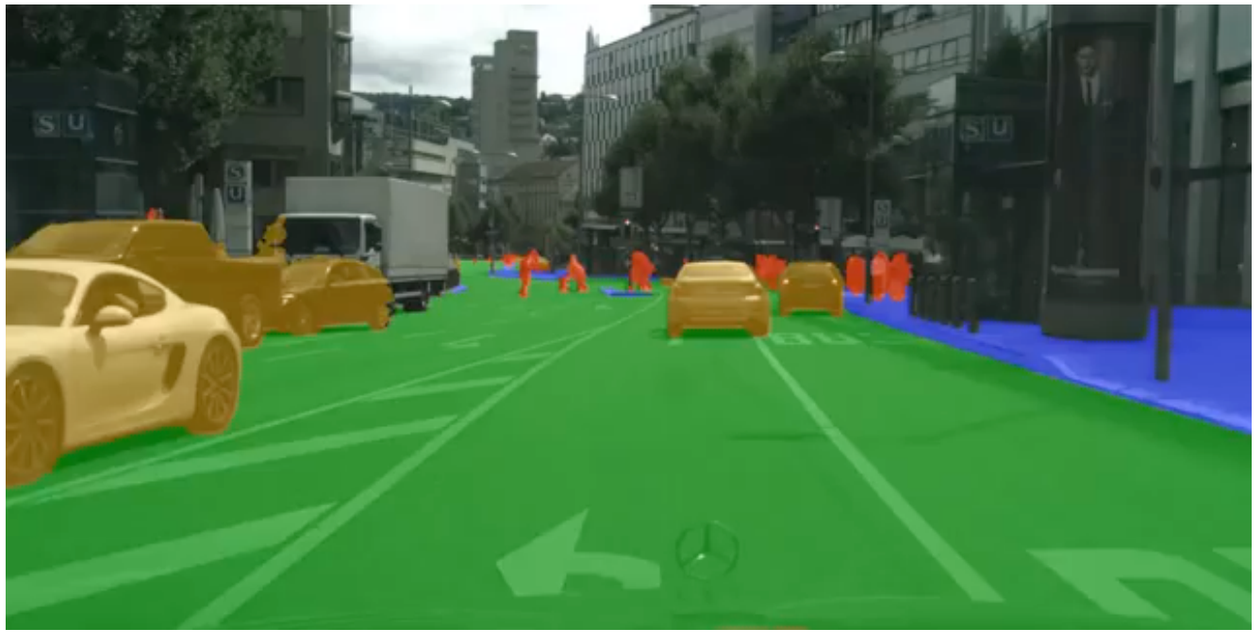
We do a lot more than passive viewing…

Scene Understanding



We do a lot more than passive viewing…

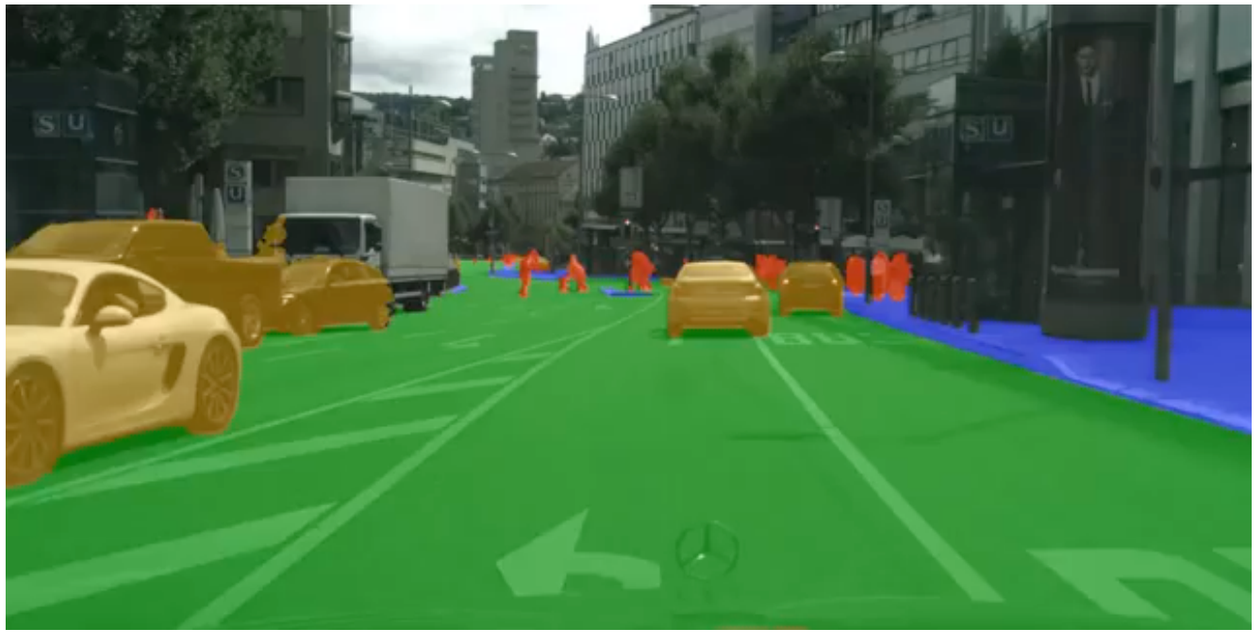
Scene Understanding



Multi-Step Planning

We do a lot more than passive viewing…

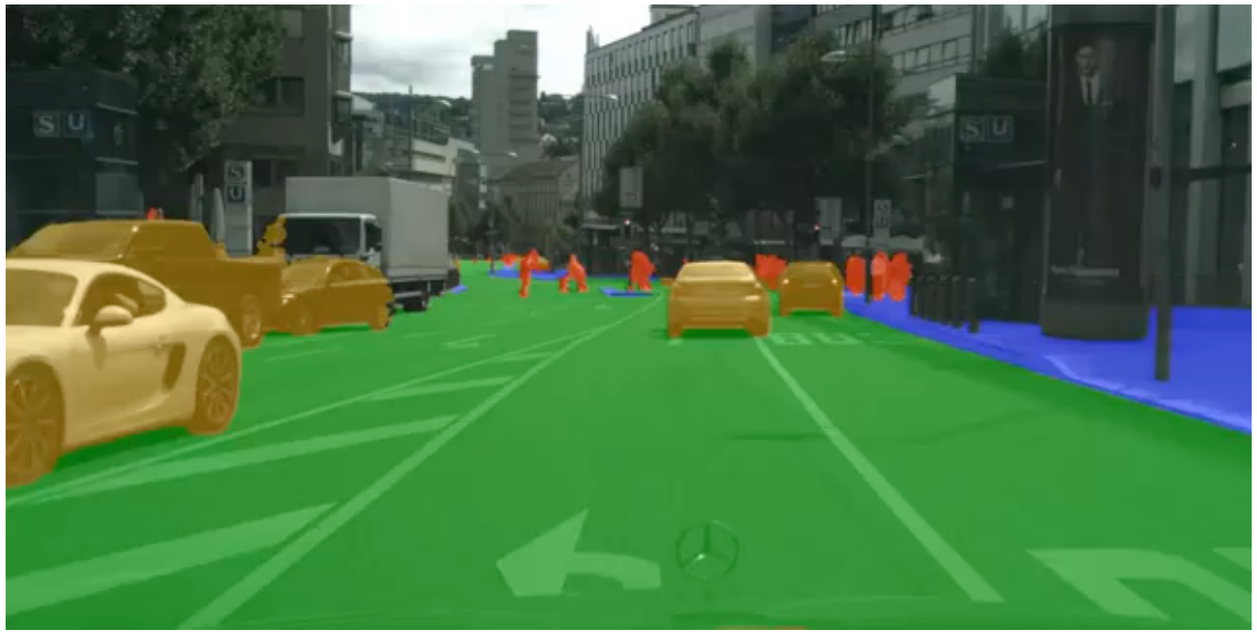
Scene Understanding



Multi-Step Planning

We do a lot more than passive viewing…

Scene Understanding

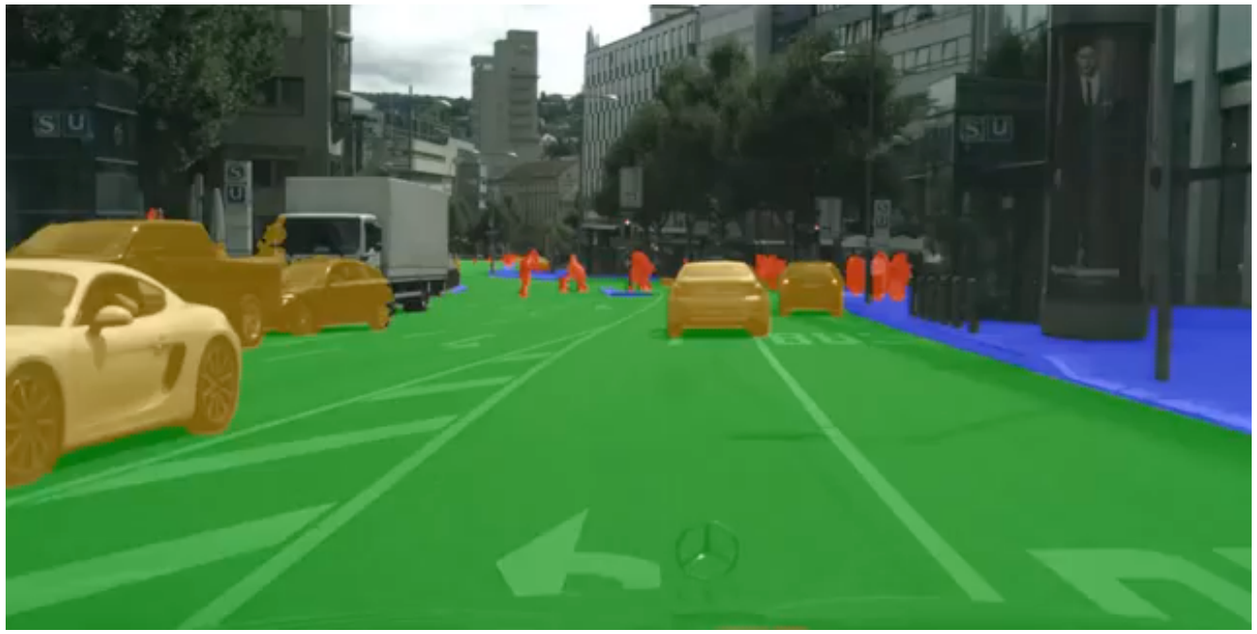


Multi-Step Planning

Navigation

We do a lot more than passive viewing…

Scene Understanding

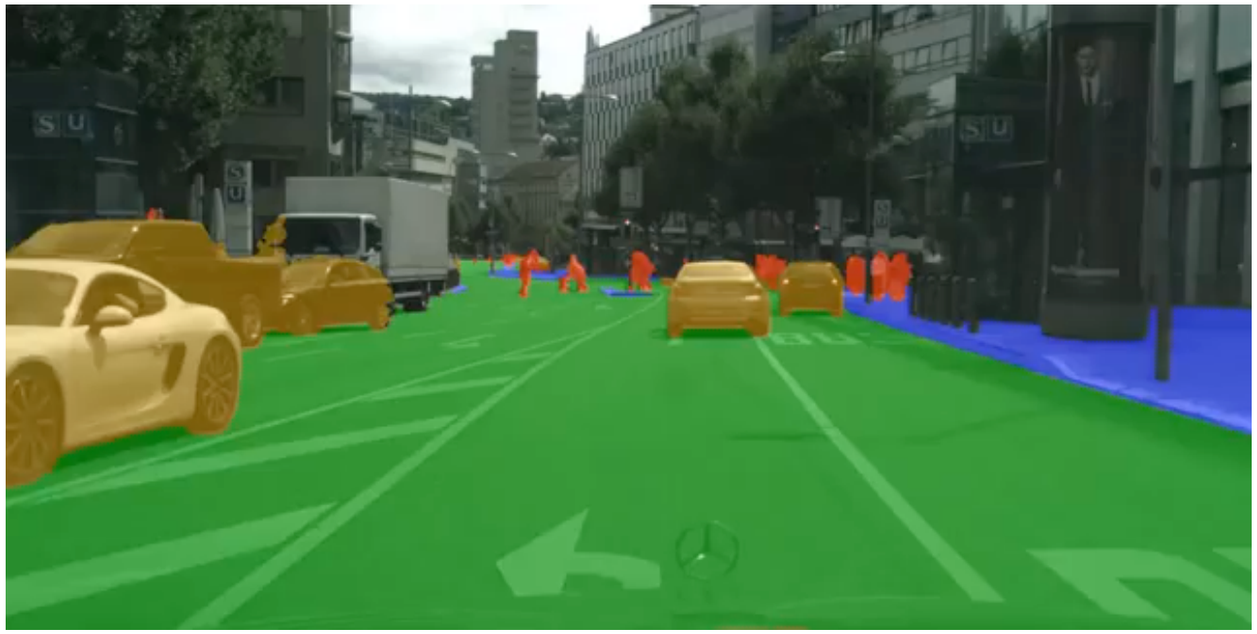


Multi-Step Planning

Navigation

We do a lot more than passive viewing…

Scene Understanding



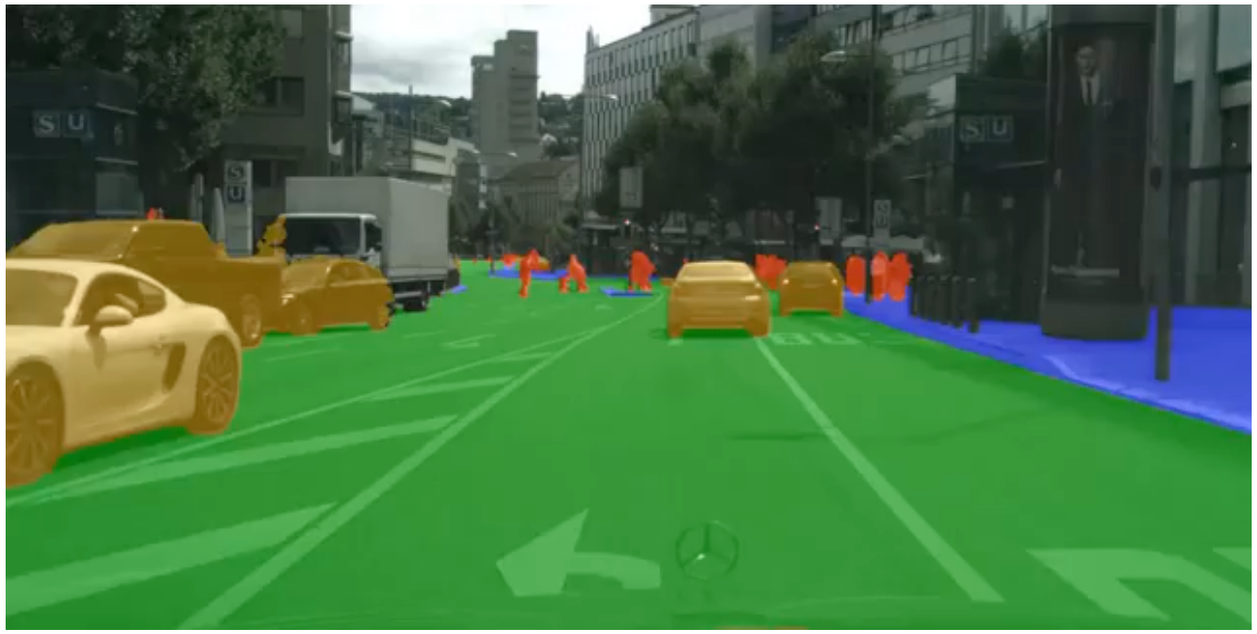
Multi-Step Planning

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding



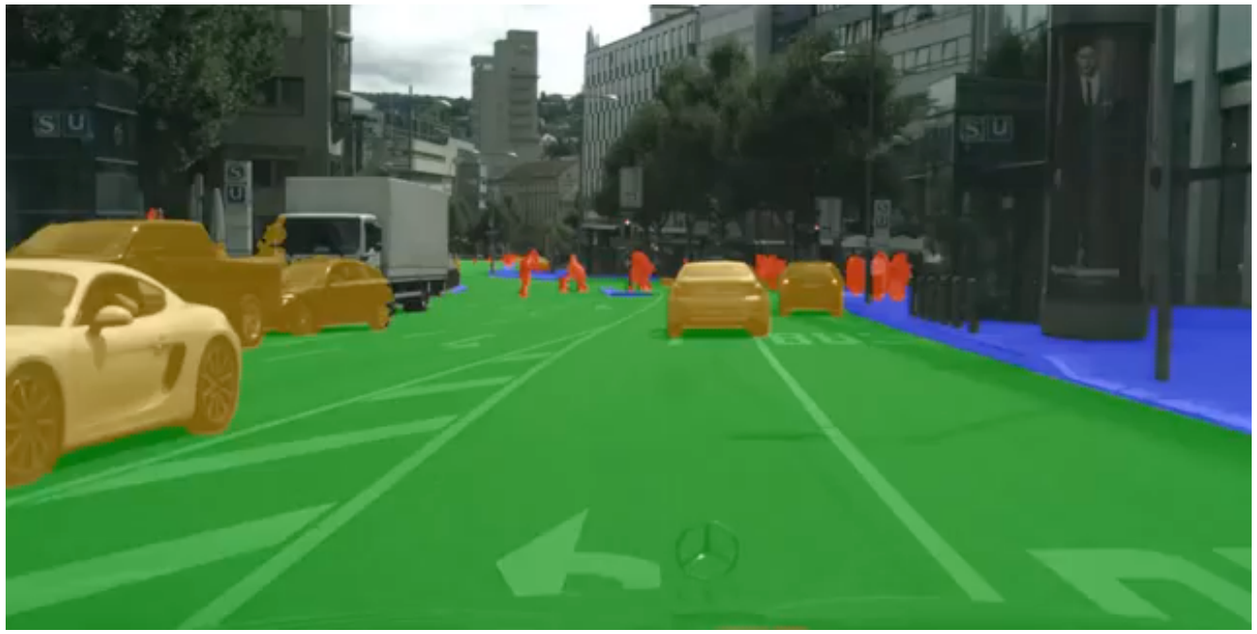
Multi-Step Planning

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding



Multi-Step Planning

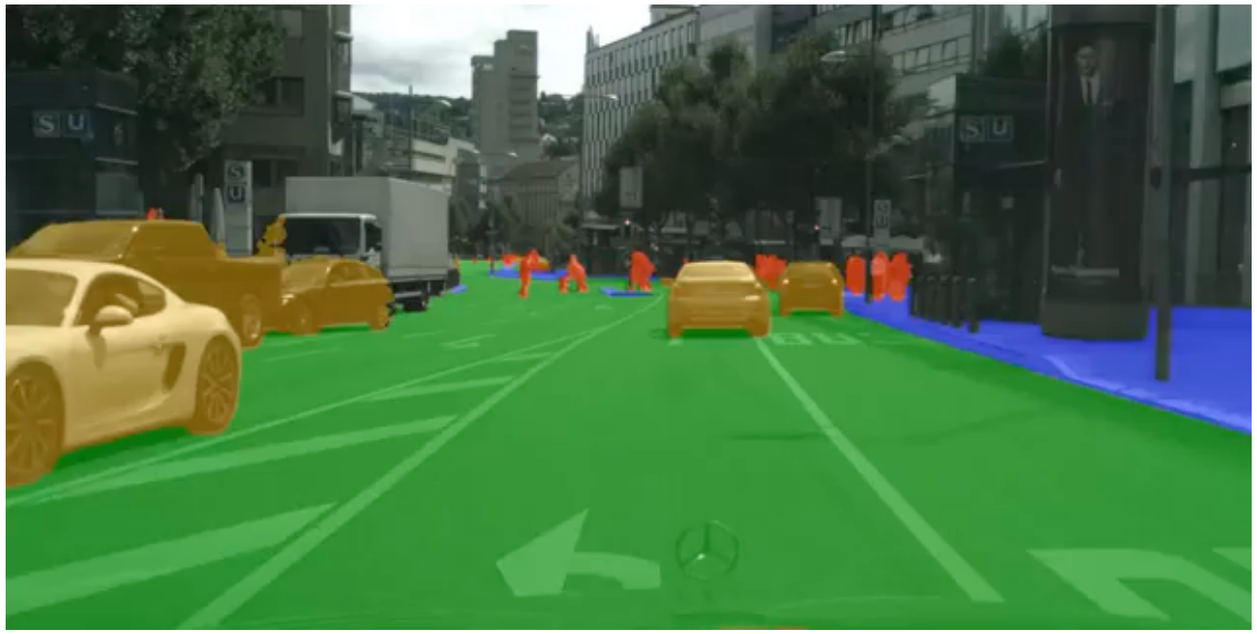
All of these behaviors are done in a body!

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding



Multi-Step Planning

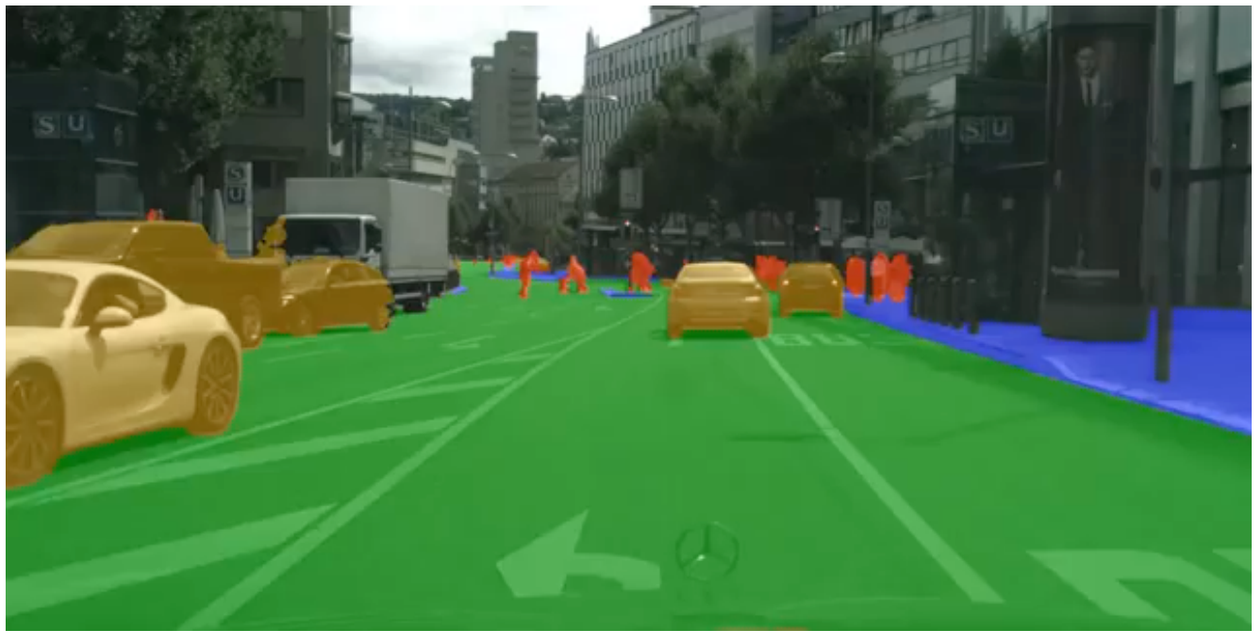
All of these behaviors are done in a body!

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding



Multi-Step Planning

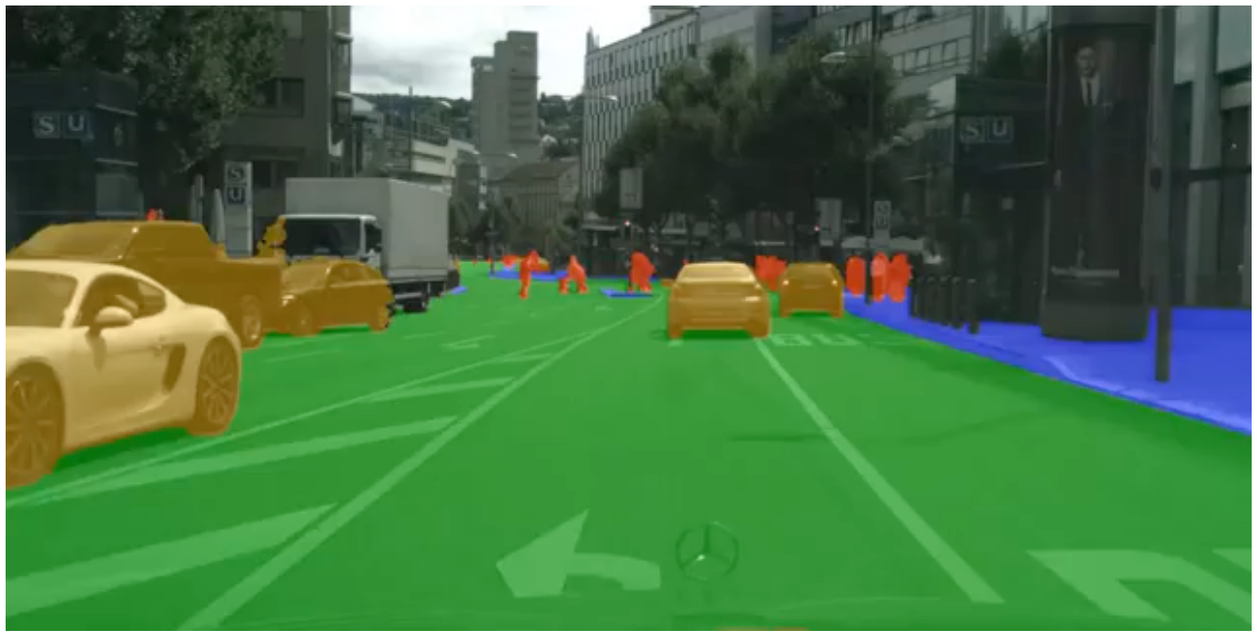
What are the core design principles that give rise to these abilities?

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding



Multi-Step Planning

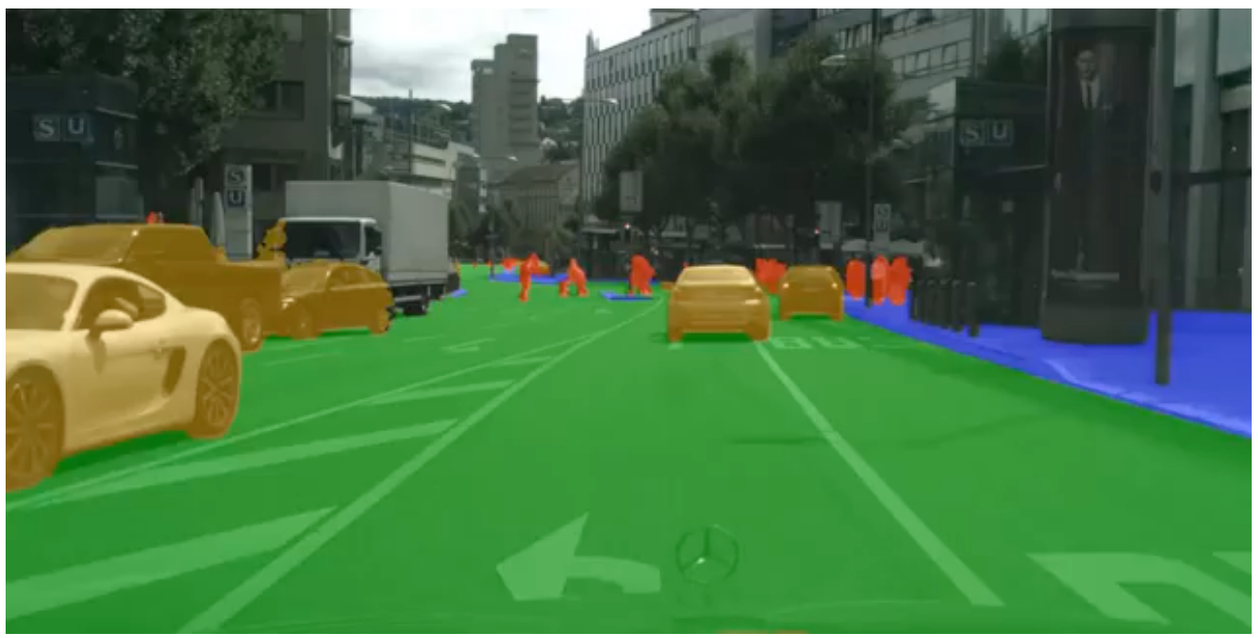
What are the core design principles that give rise to these abilities?

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding



Multi-Step Planning

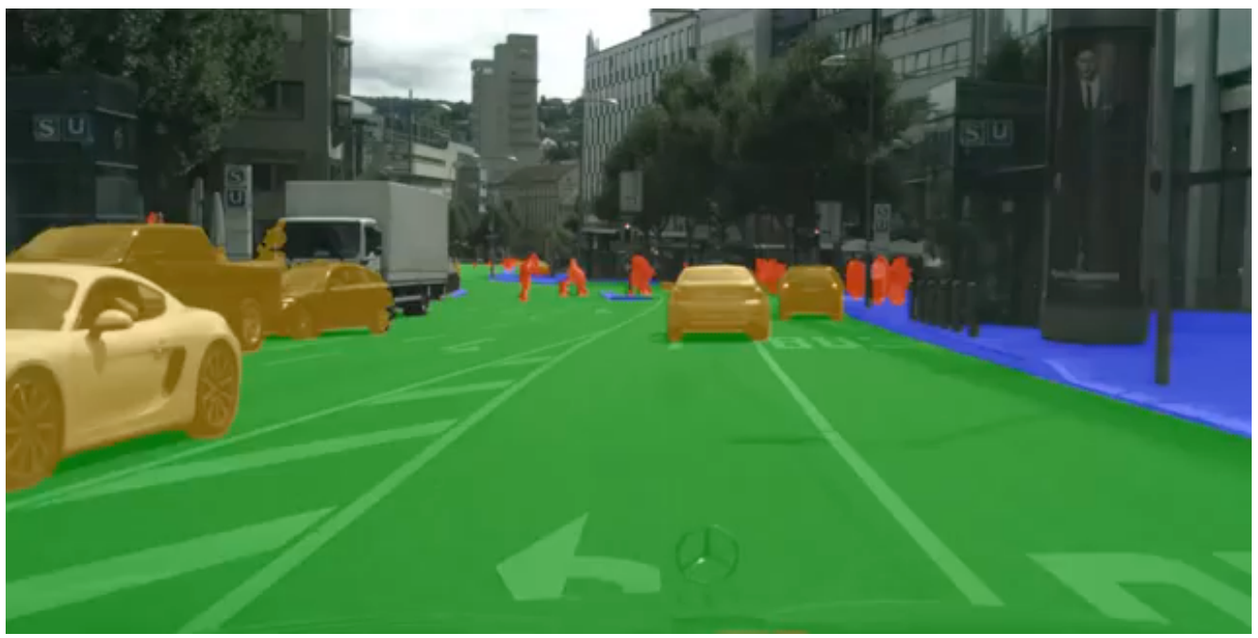
What are the core design principles that give rise to these abilities?

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding



Multi-Step Planning

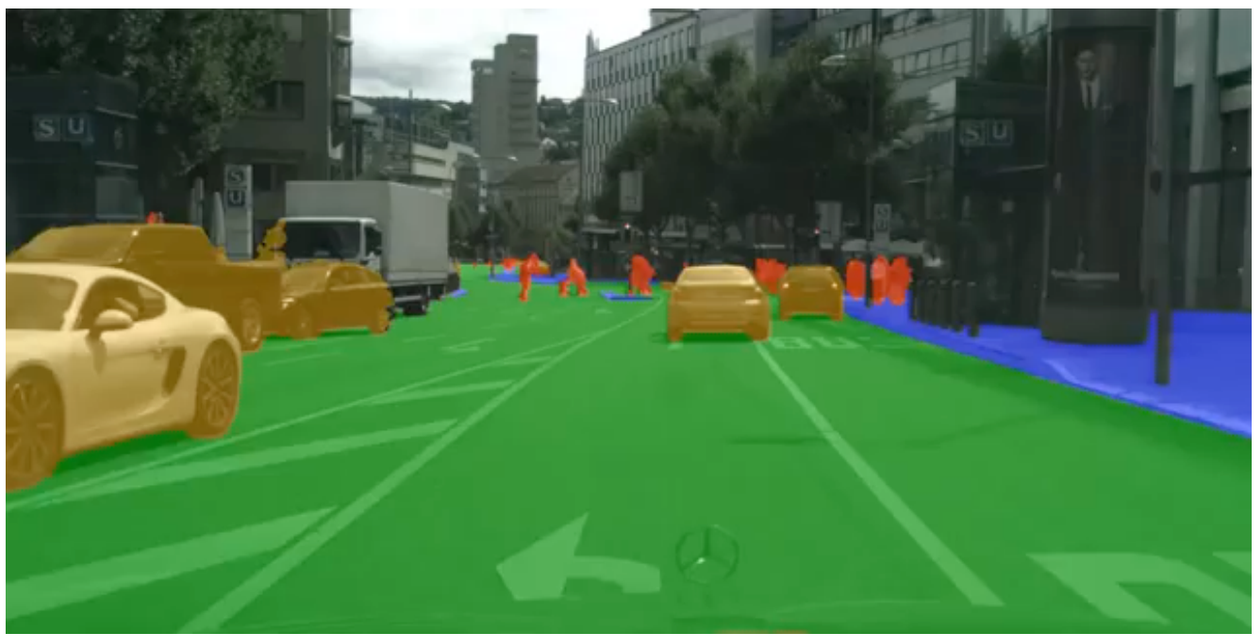
What are the core design principles that give rise to these abilities?

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding

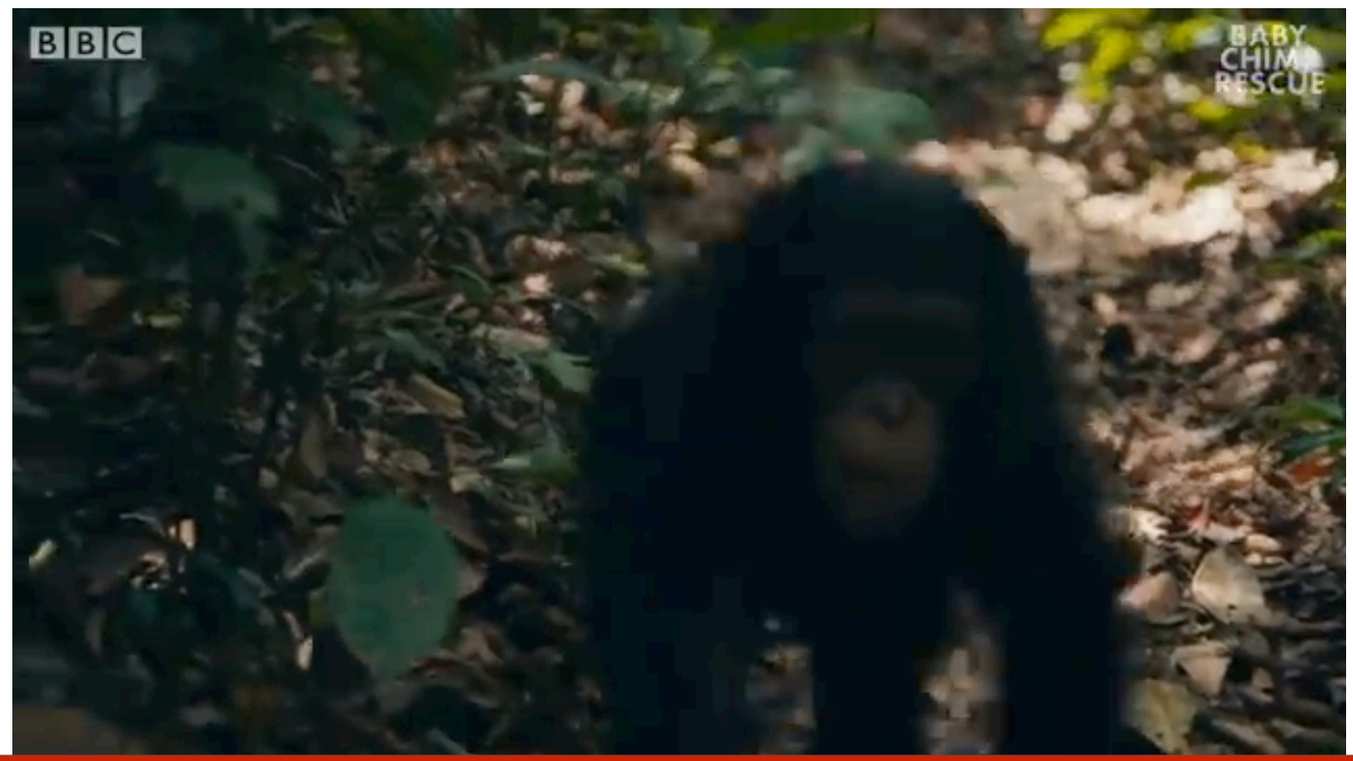


Multi-Step Planning

What are the core design principles that give rise to these abilities?

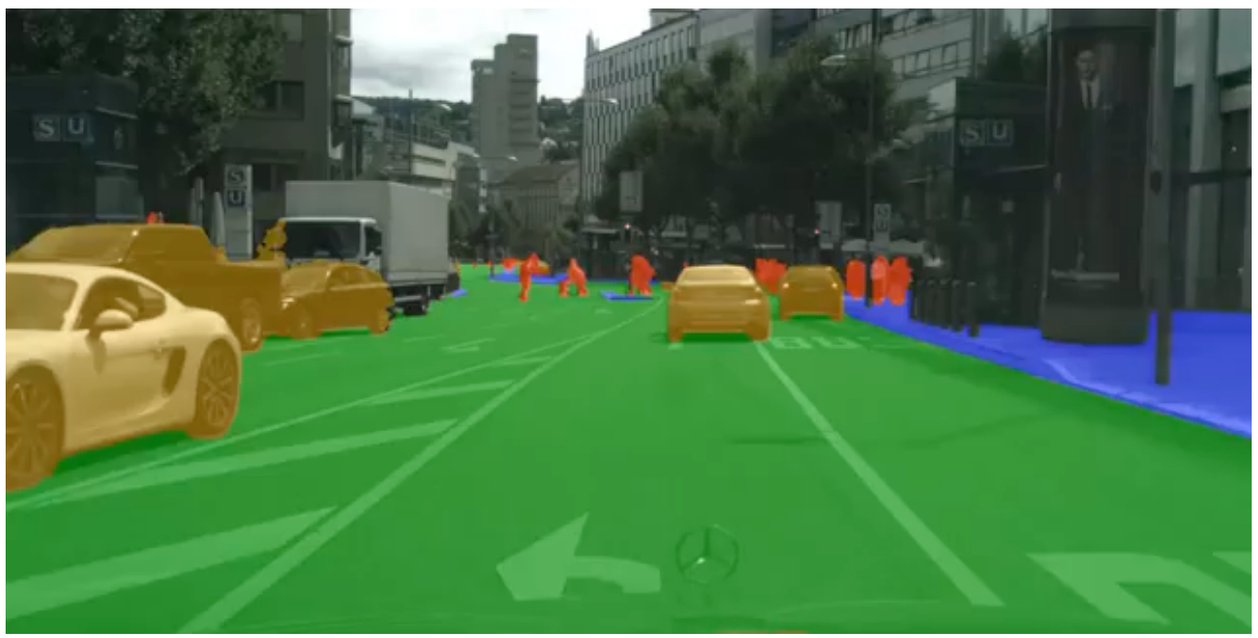
Navigation

Flexible Embodiment



We do a lot more than passive viewing…

Scene Understanding

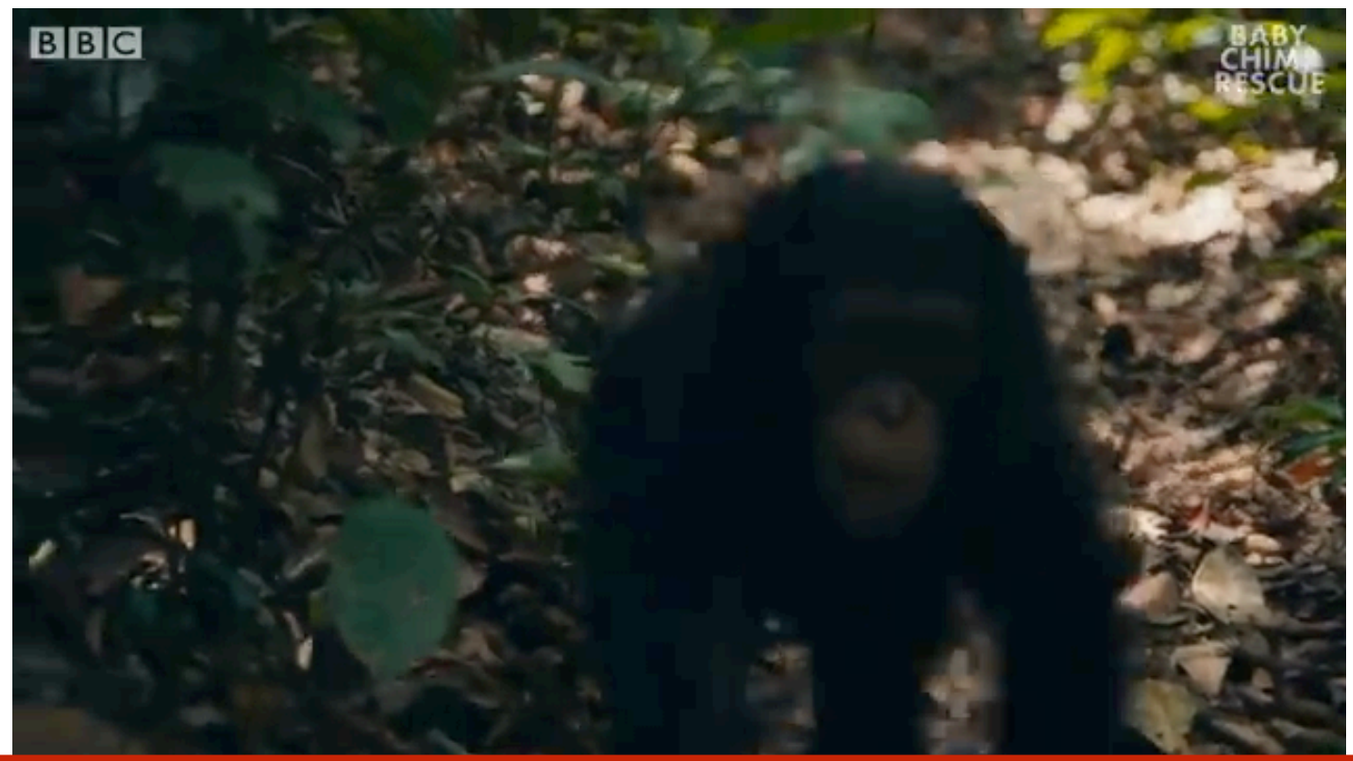


Multi-Step Planning

What are the core design principles that give rise to these abilities?

Navigation

Flexible Embodiment



Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System
- Reusable Latent Representations for Primate Mental Simulation
- Heuristics for Interrogating Natural Intelligence

Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System
- Reusable Latent Representations for Primate Mental Simulation
- Heuristics for Interrogating Natural Intelligence

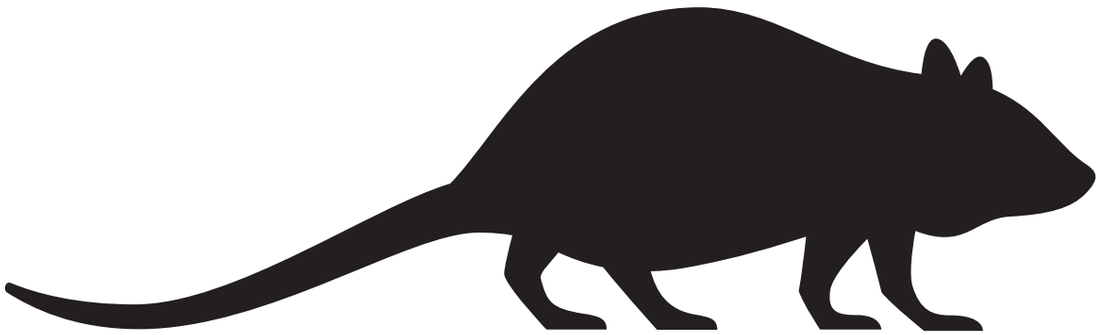
Mouse Visual Cortex as a Task-General, Limited Resource System

A. Nayebi*, N.C.L. Kong*, C. Zhuang, J.L. Gardner, A.M. Norcia, D.L.K. Yamins

Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation.

PLOS Computational Biology 2023

Nathan C.L. Kong*



Chengxu Zhuang

Justin L. Gardner

Anthony M. Norcia

Daniel Yamins

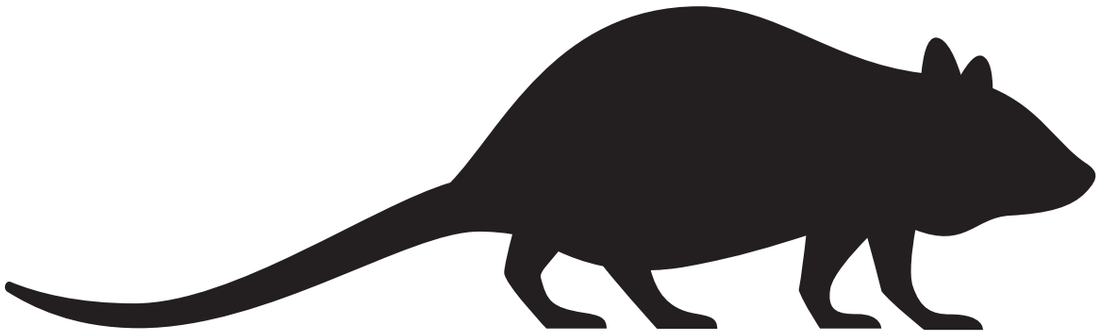
Mouse Visual Cortex as a Task-General, Limited Resource System

A. Nayebi*, N.C.L. Kong*, C. Zhuang, J.L. Gardner, A.M. Norcia, D.L.K. Yamins

Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation.

PLOS Computational Biology 2023

Nathan C.L. Kong*



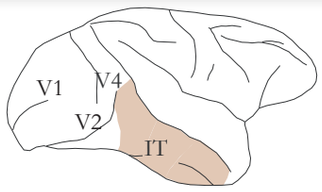
Chengxu Zhuang

Justin L. Gardner

Anthony M. Norcia

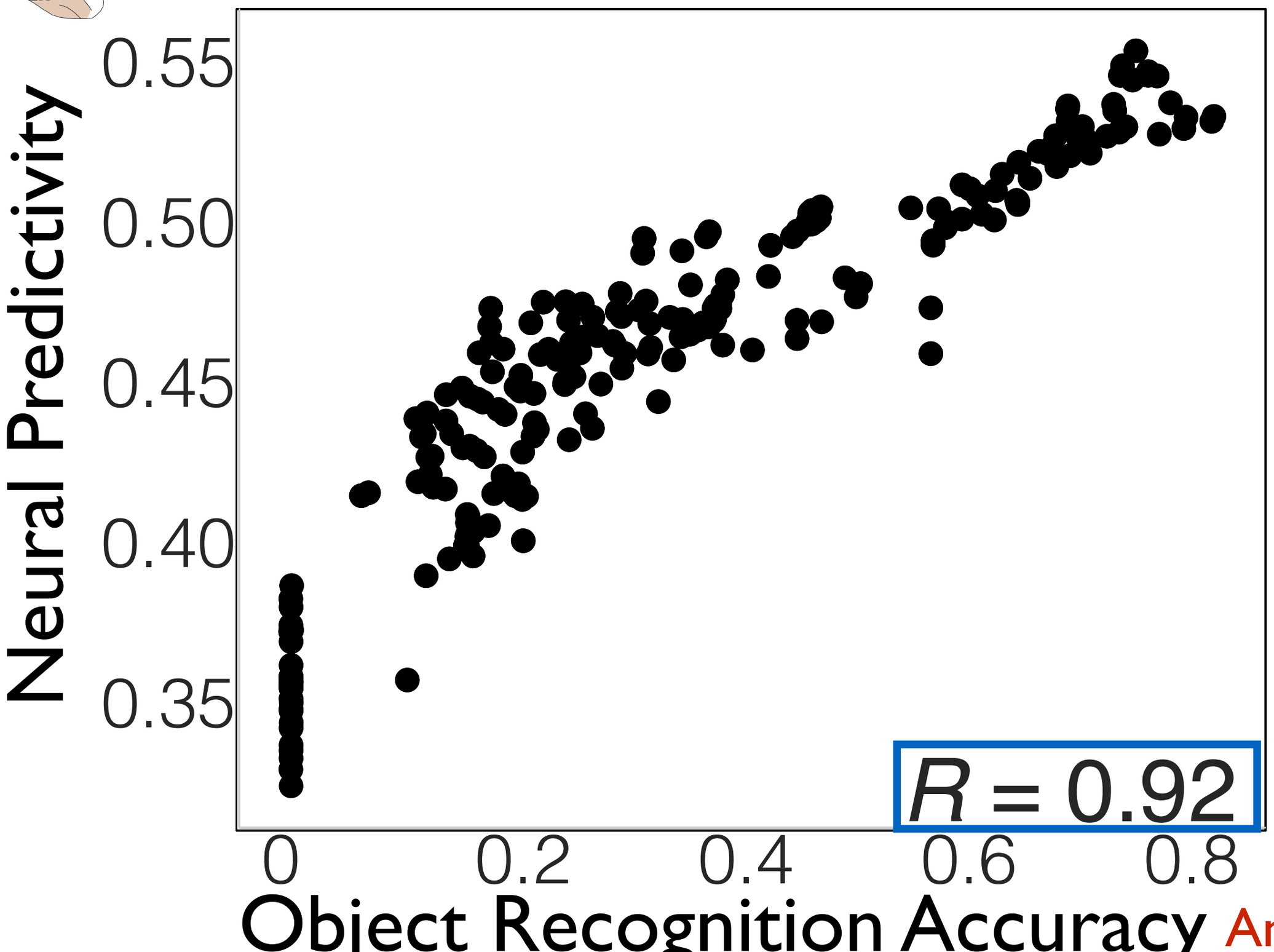
Daniel Yamins

Task Performance Correlated with Neural Predictivity



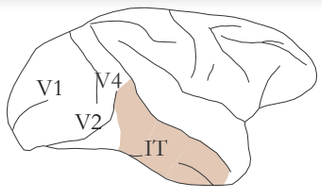
A Neuroscience Goal

Schrimpf, Kubilius* et al. 2018*



An AI Goal

Task Performance Correlated with Neural Predictivity



Schrimpf, Kubilius* et al. 2018*

Neural Predictivity

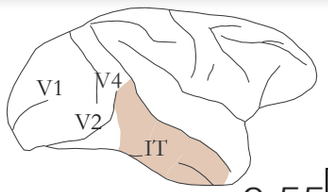
0.55
0.50
0.45
0.40
0.35

$R = 0.92$

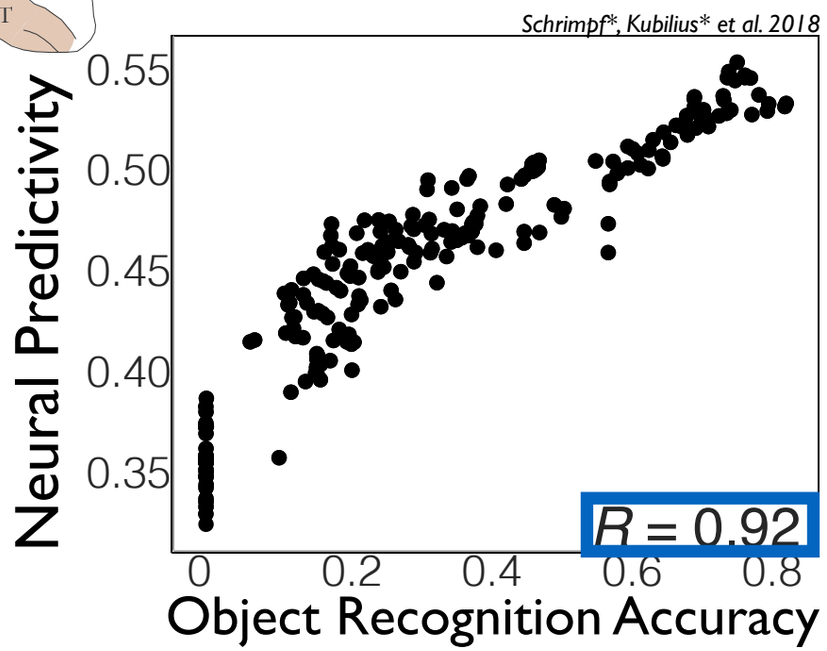
0
0.2
0.4
0.6
0.8

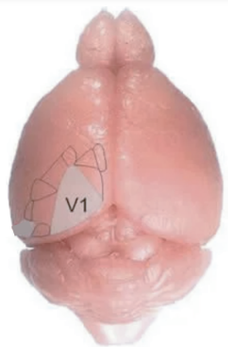
Object Recognition Accuracy

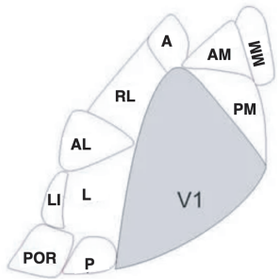
Task Performance Correlated with Neural Predictivity



Primates

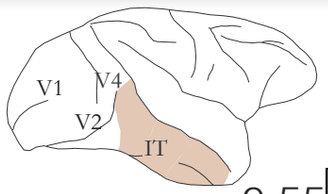






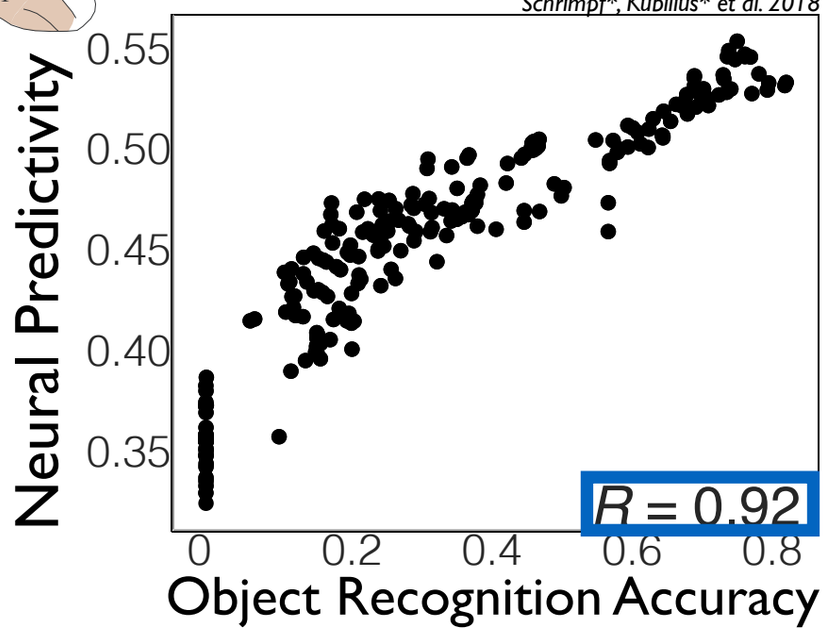
Mouse?

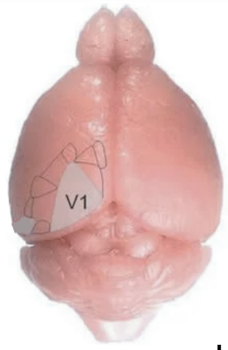
Object Categorization Ability **NOT** Correlated with Neural Predictivity

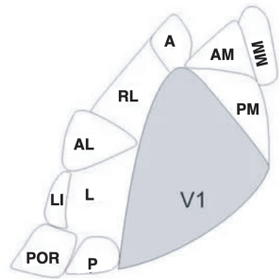


Primates

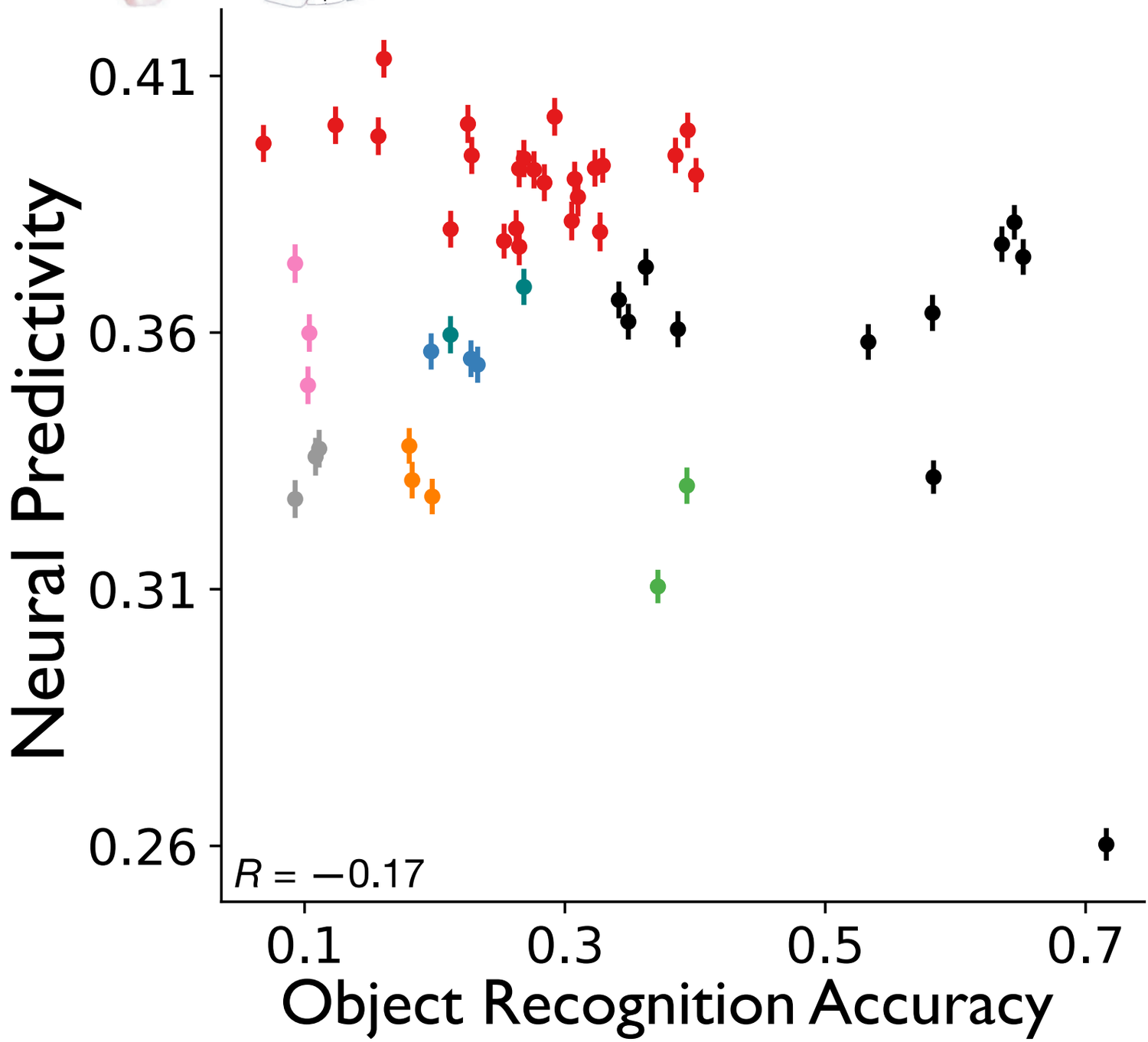
Schrimpf, Kubilius* et al. 2018*



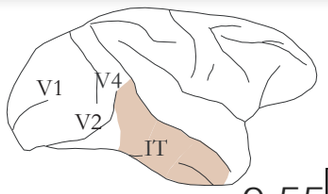




Mouse

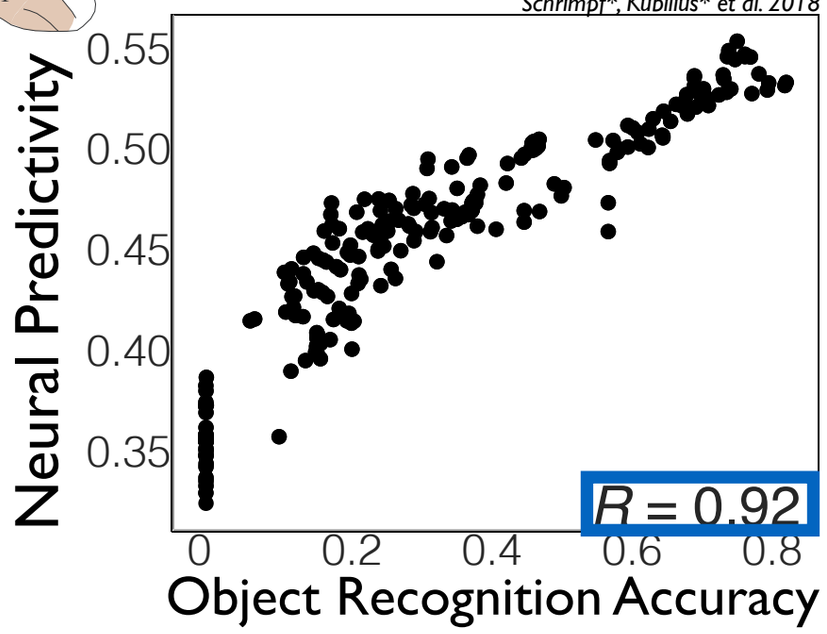


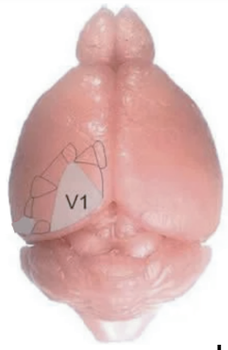
Object Categorization Ability **NOT** Correlated with Neural Predictivity

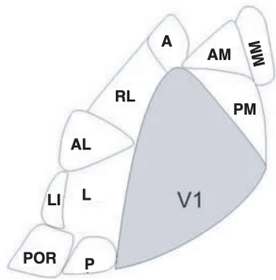


Primates

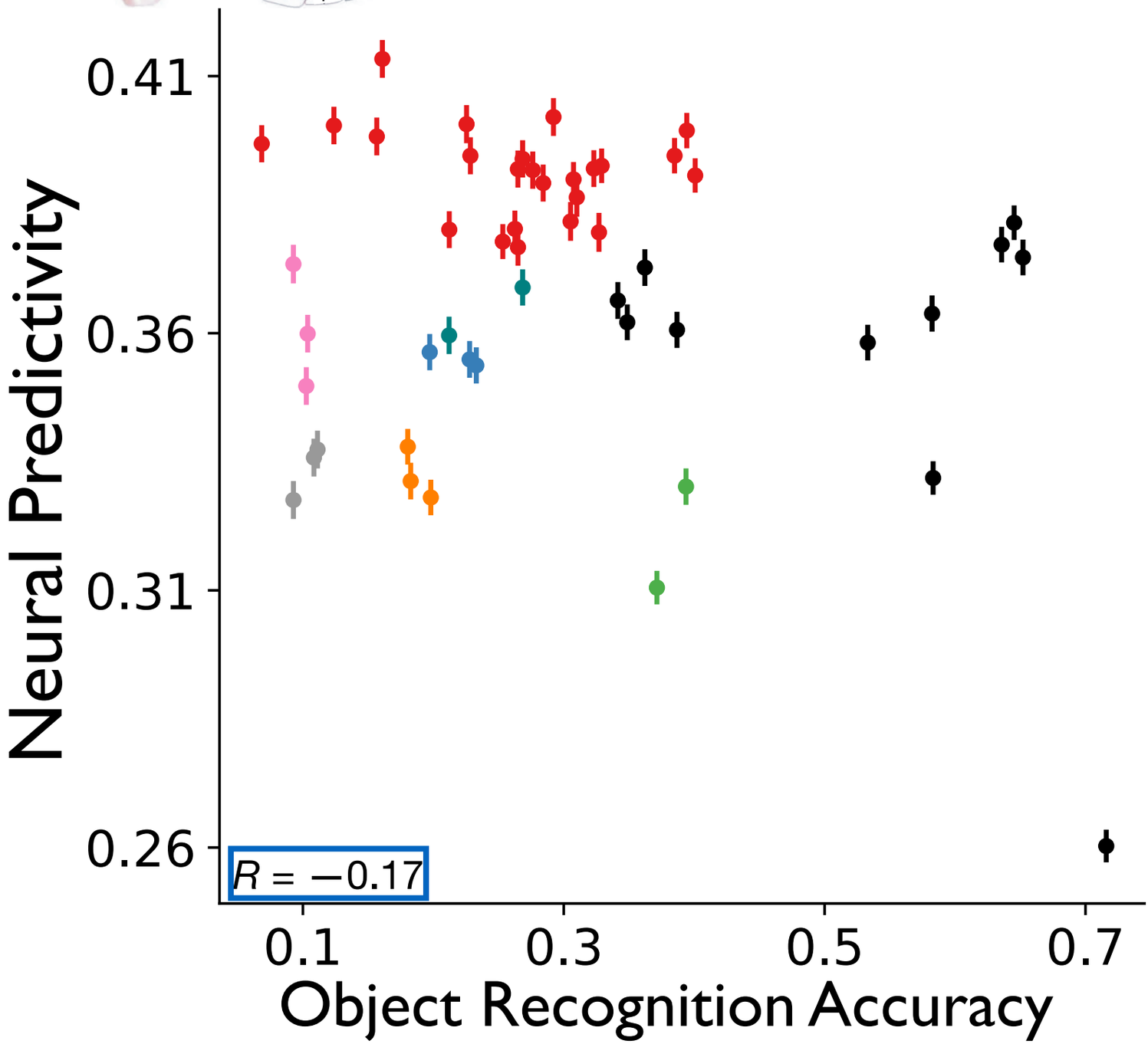
Schrimpf, Kubilius* et al. 2018*



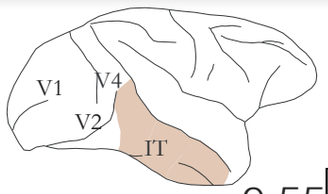




Mouse

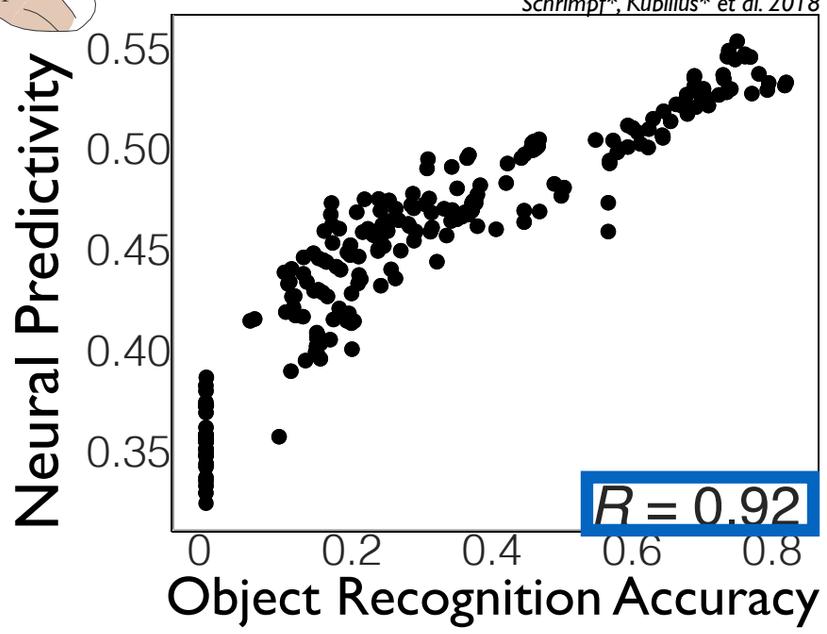


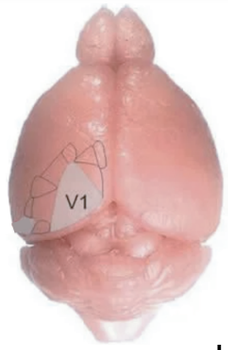
Object Categorization Ability **NOT** Correlated with Neural Predictivity

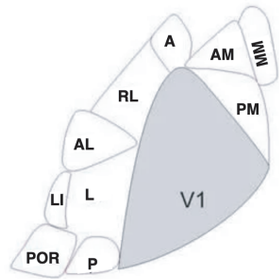


Primates

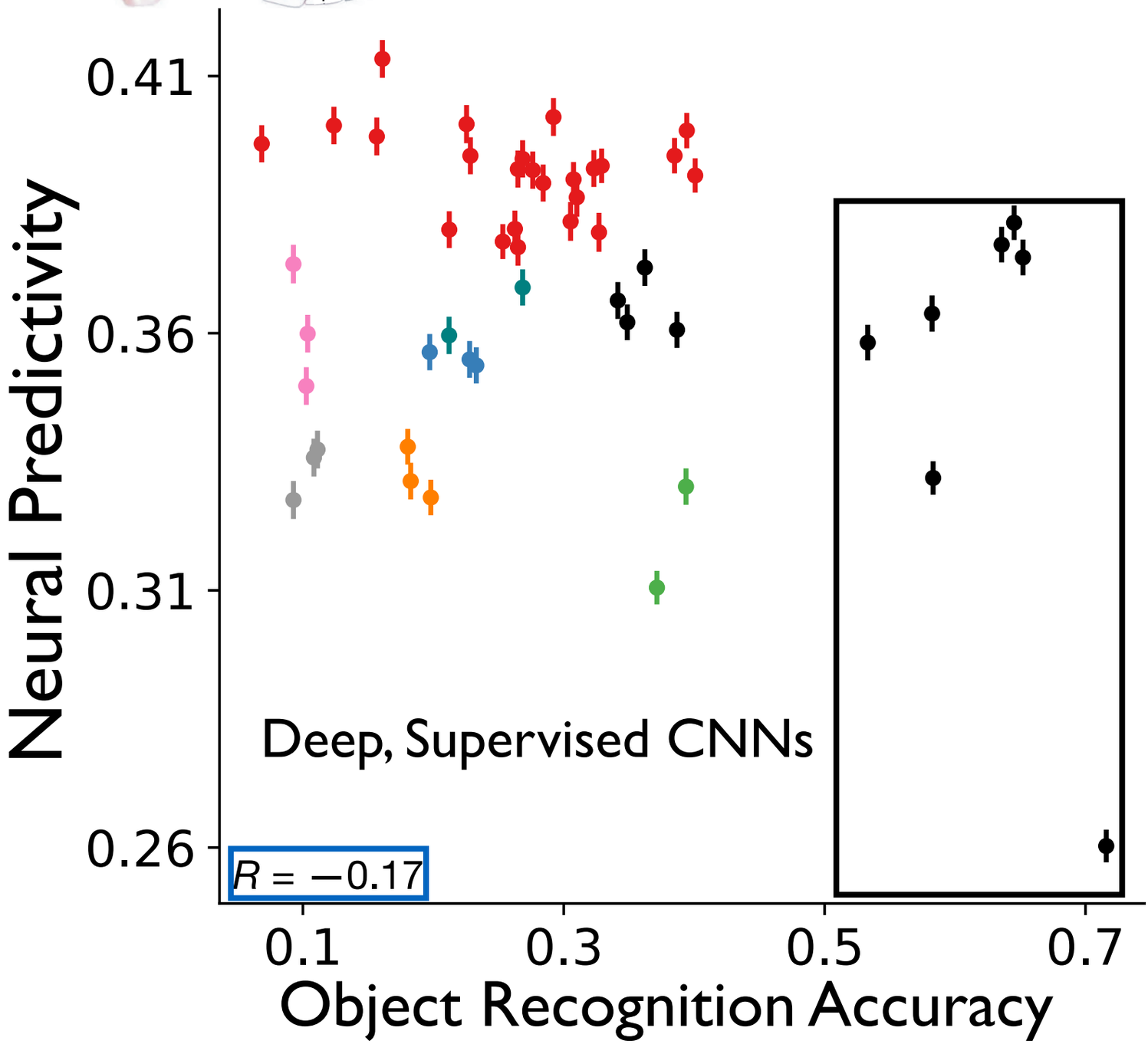
Schrimpf, Kubilius* et al. 2018*



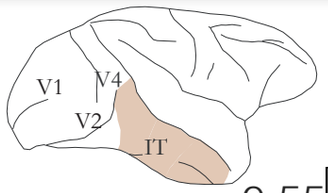




Mouse

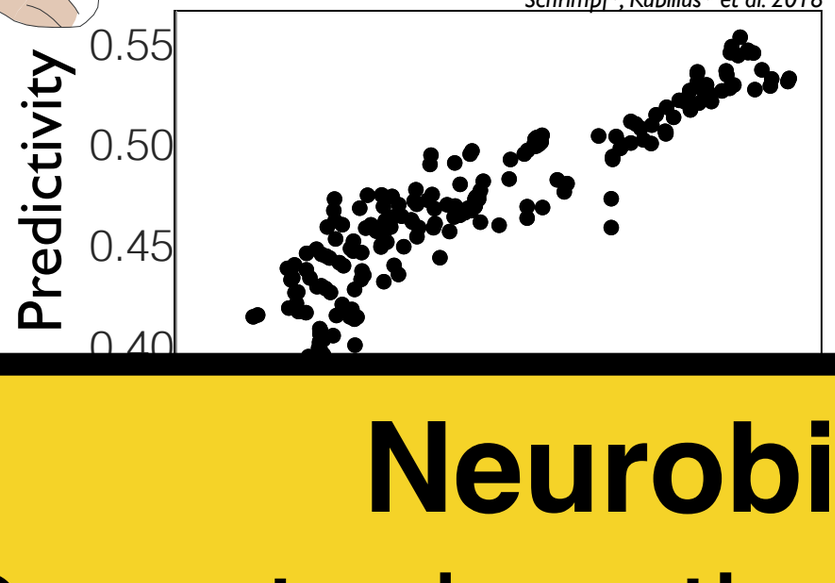


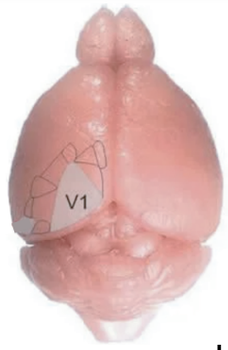
Object Categorization Ability **NOT** Correlated with Neural Predictivity

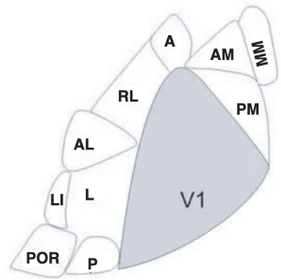


Primates

Schrimpf, Kubilius* et al. 2018*



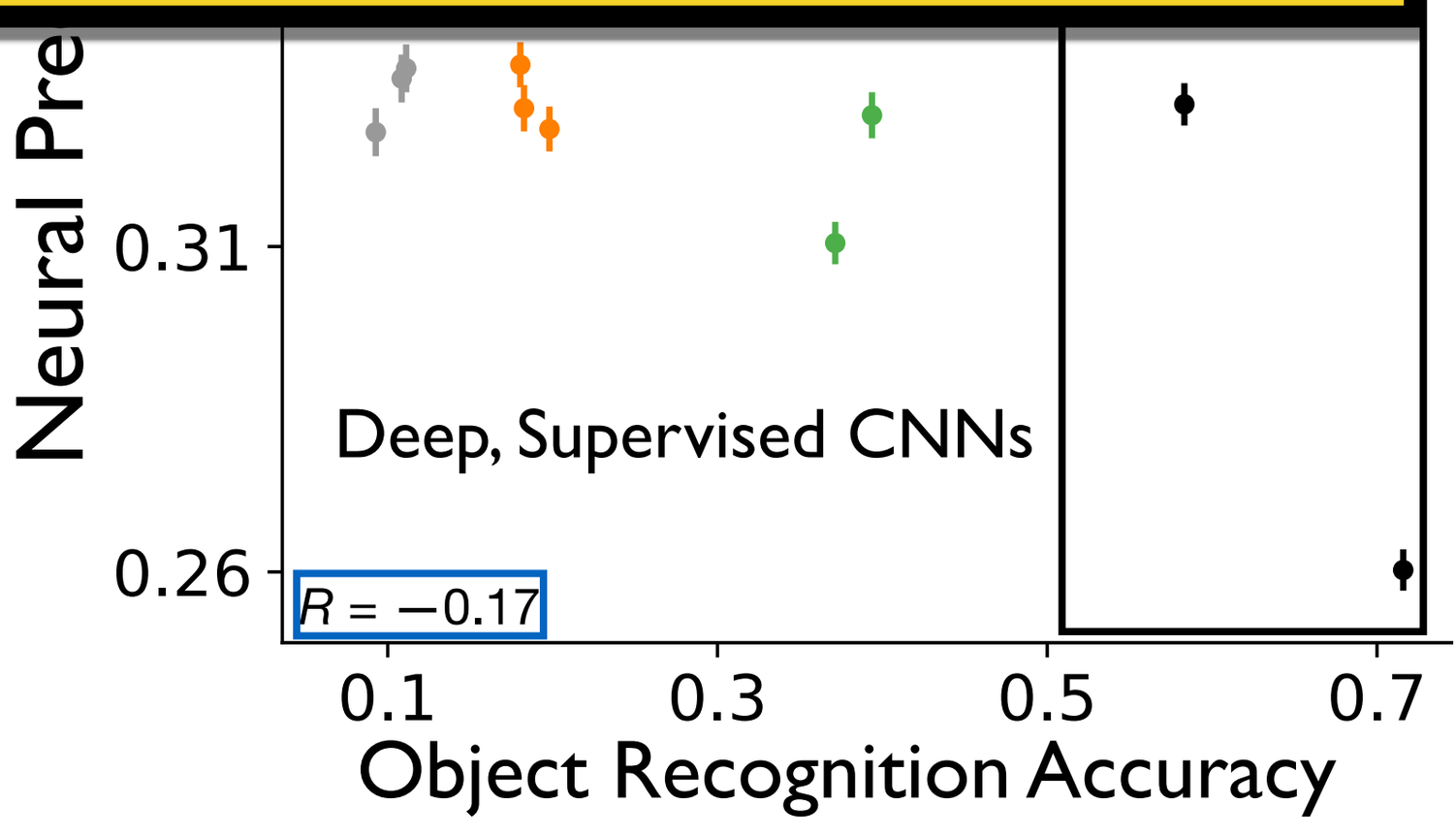




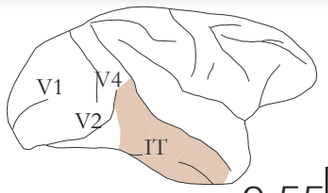
Mouse

Neurobiological Puzzle:

Does task-optimization apply to rodents?

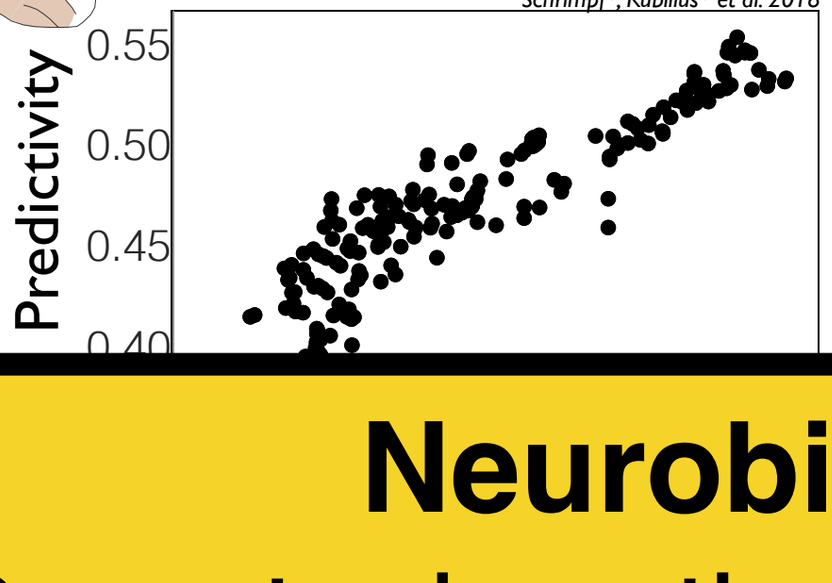


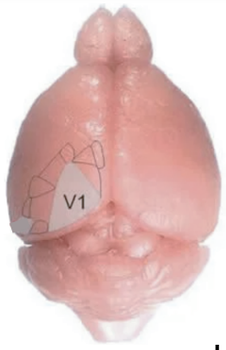
Object Categorization Ability **NOT** Correlated with Neural Predictivity

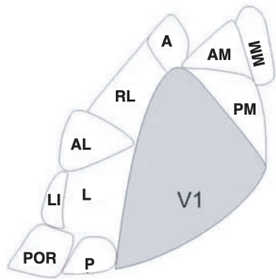


Primates

Schrimpf, Kubilius* et al. 2018*





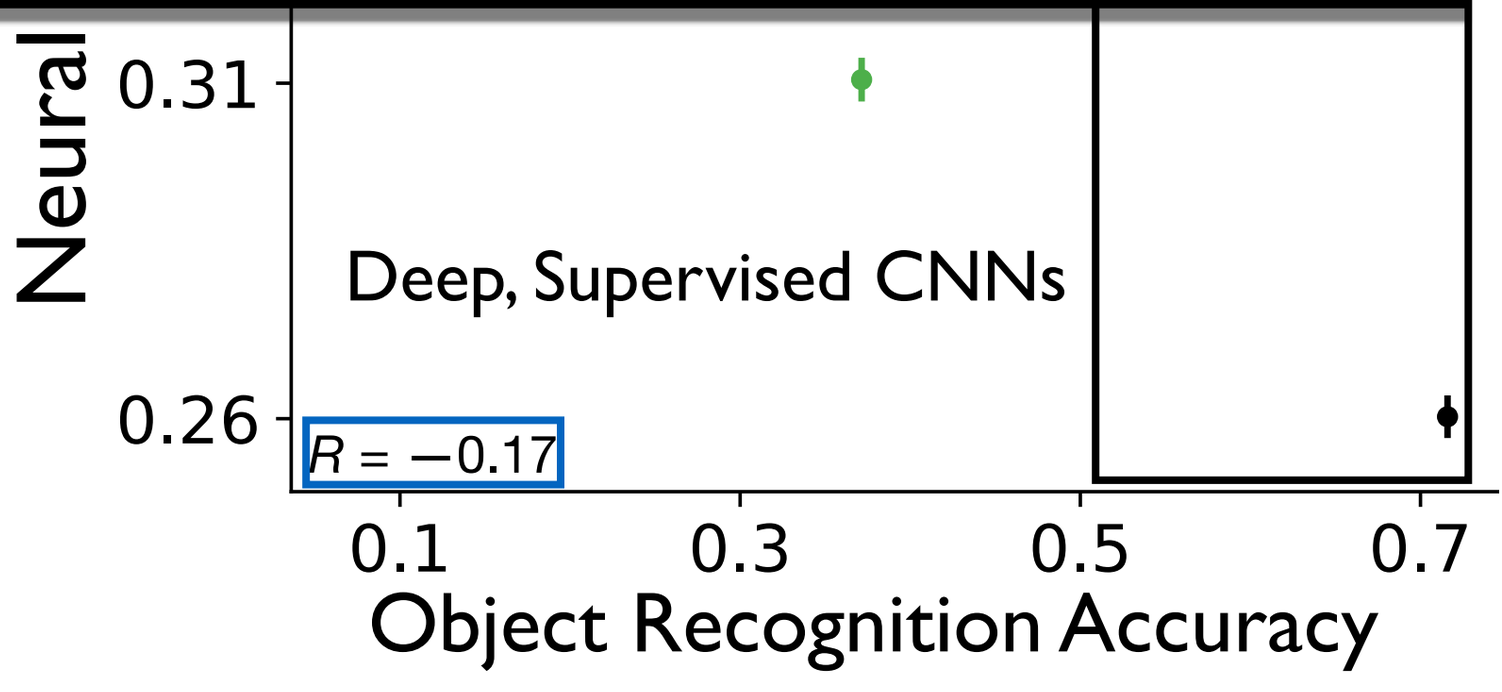


Mouse

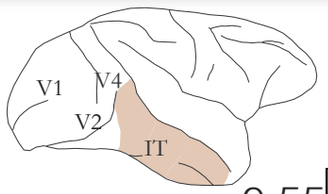
Neurobiological Puzzle:

Does task-optimization apply to rodents?

Yes!

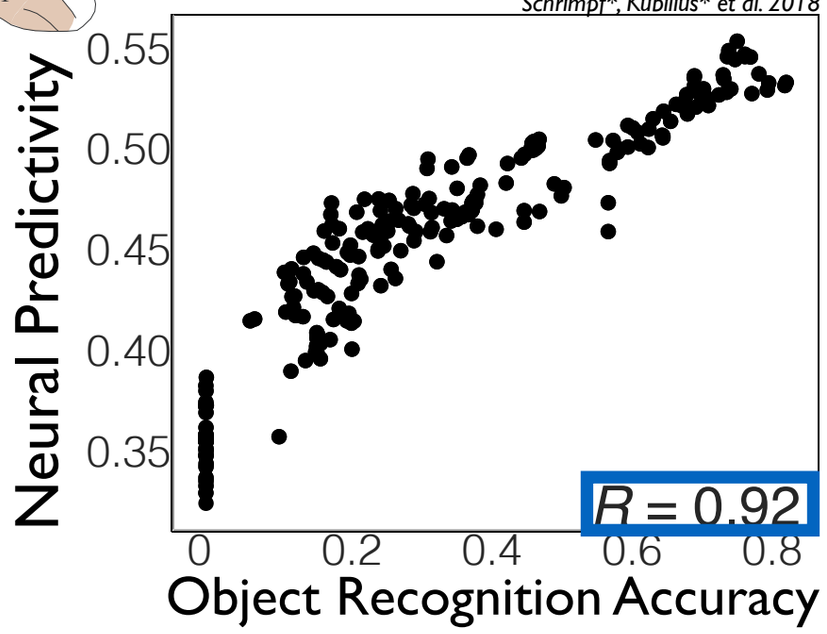


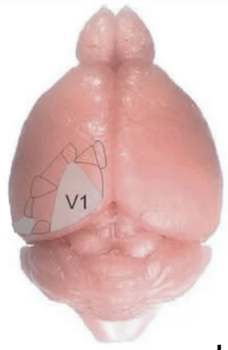
Object Categorization Ability **NOT** Correlated with Neural Predictivity

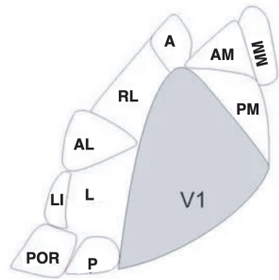


Primates

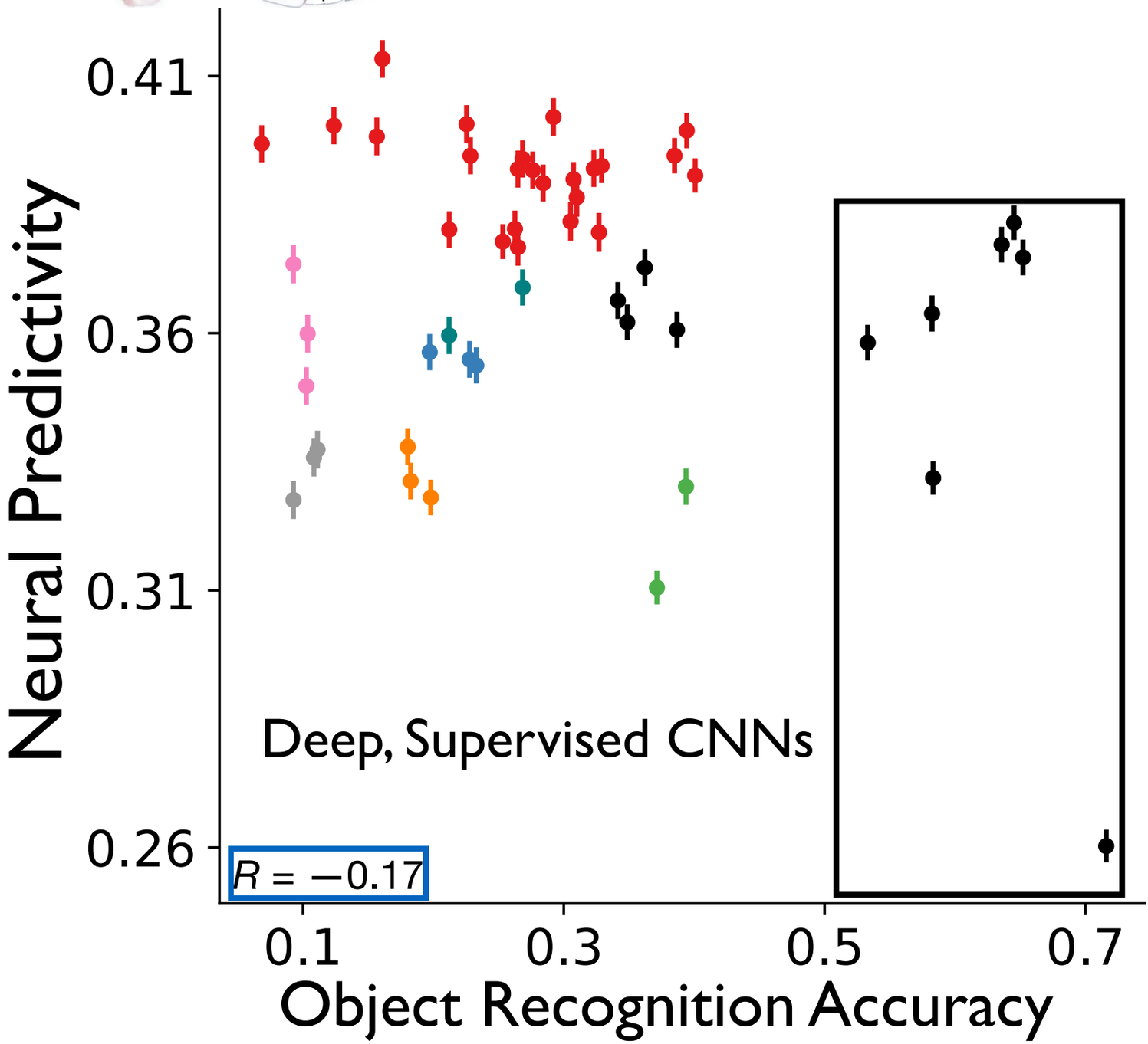
Schrimpf, Kubilius* et al. 2018*



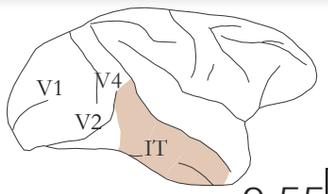




Mouse

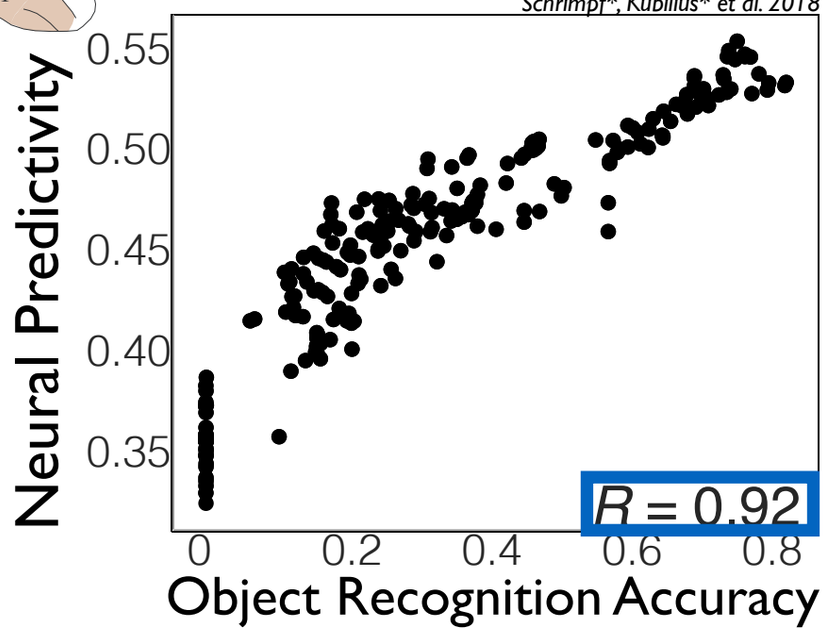


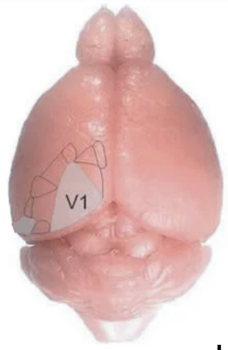
Object Categorization Ability **NOT** Correlated with Neural Predictivity

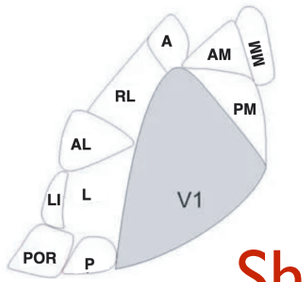


Primates

Schrimpf, Kubilius* et al. 2018*

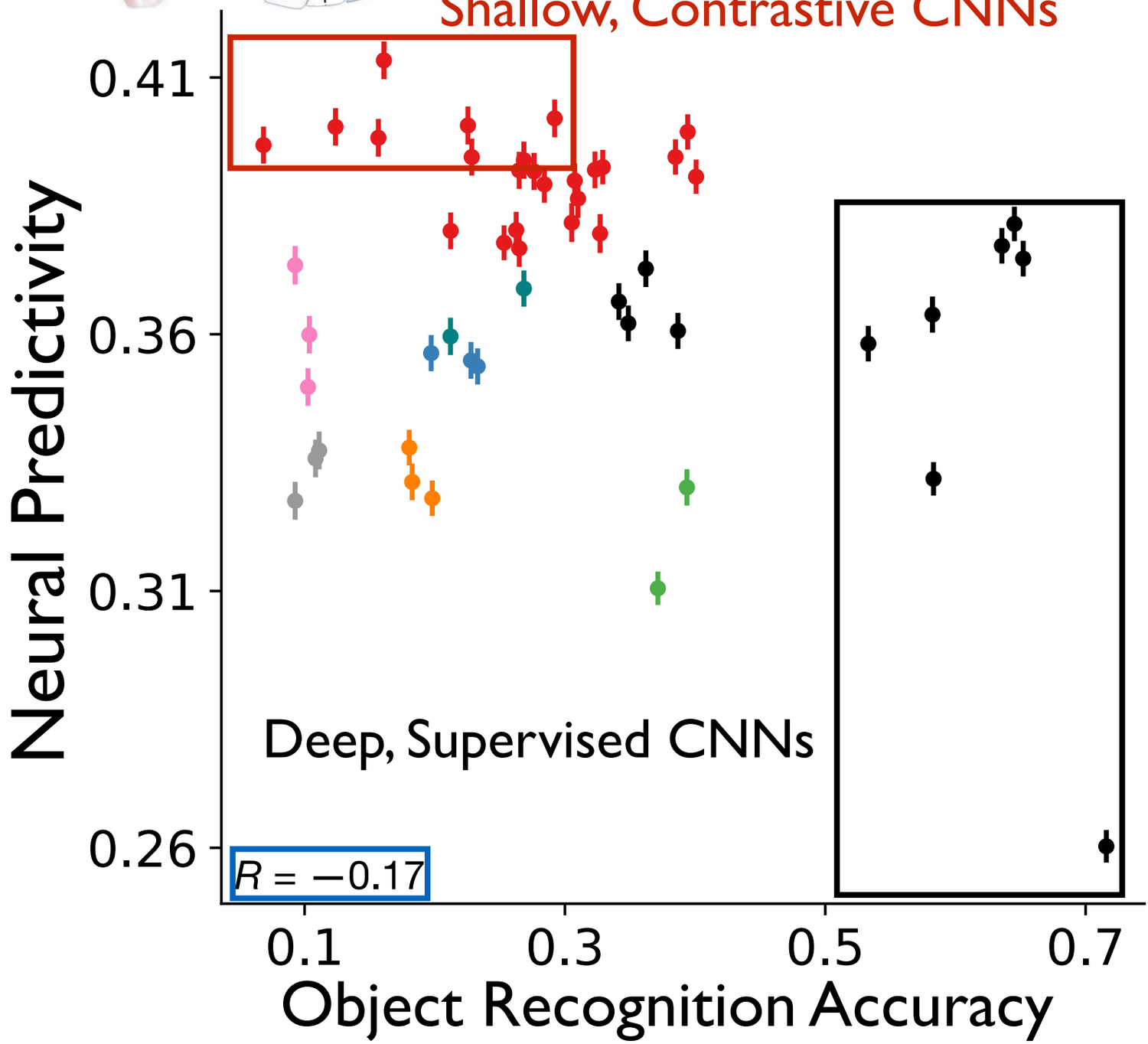




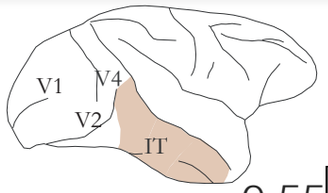


Mouse

Shallow, Contrastive CNNs

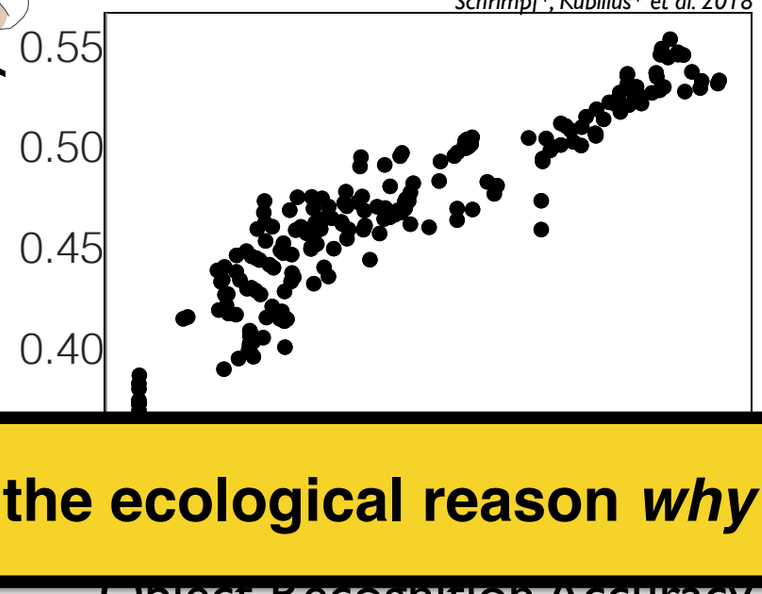


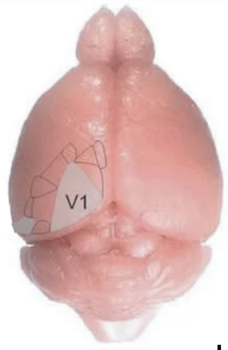
Object Categorization Ability **NOT** Correlated with Neural Predictivity

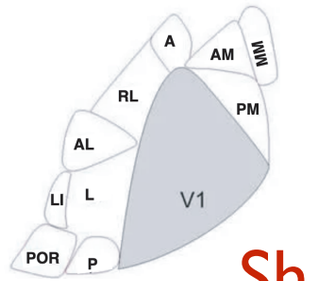


Primates

Schrimpf, Kubilius* et al. 2018*



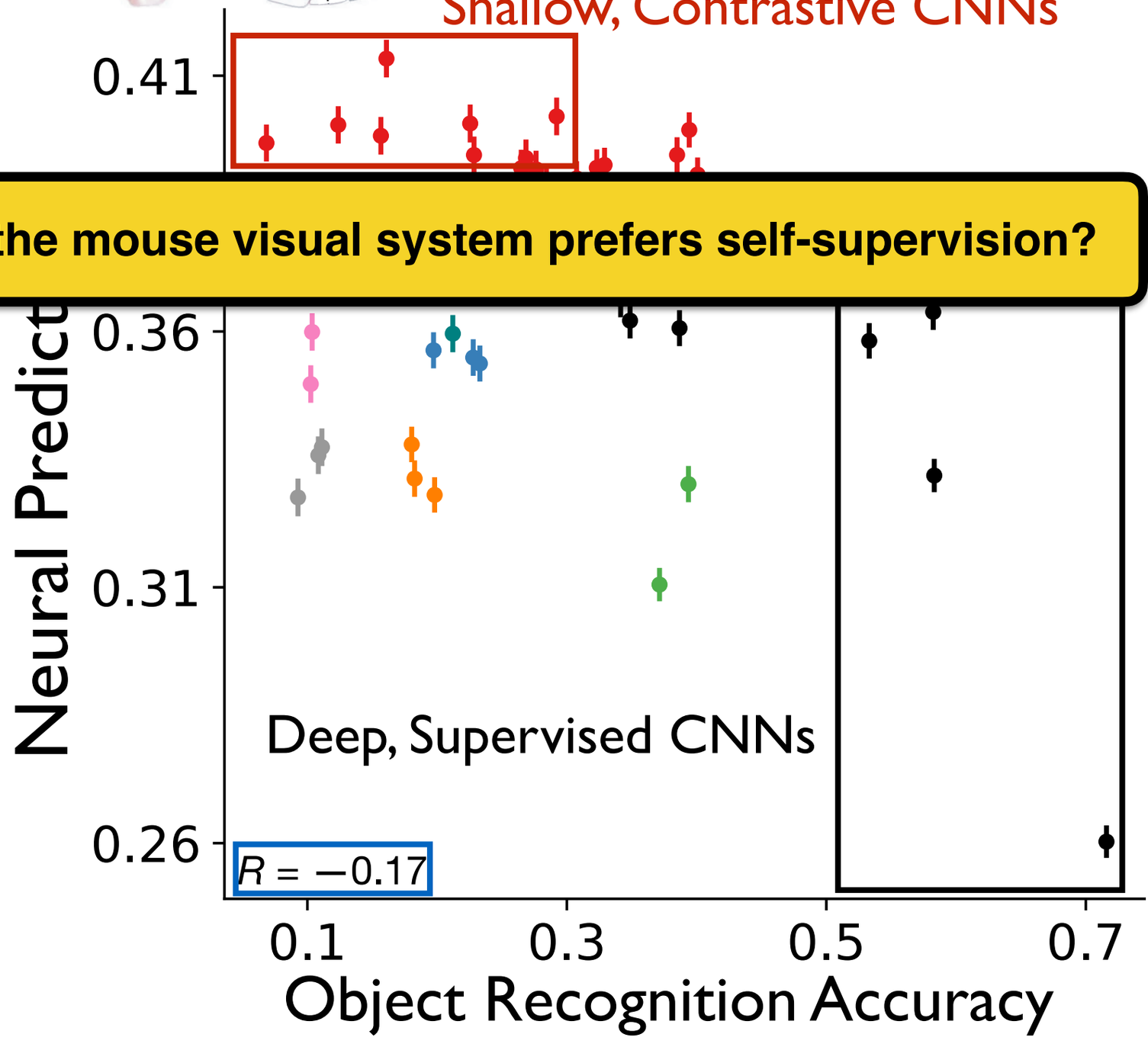




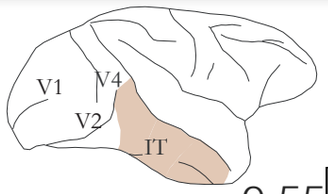
Mouse

Shallow, Contrastive CNNs

What is the ecological reason *why* the mouse visual system prefers self-supervision?

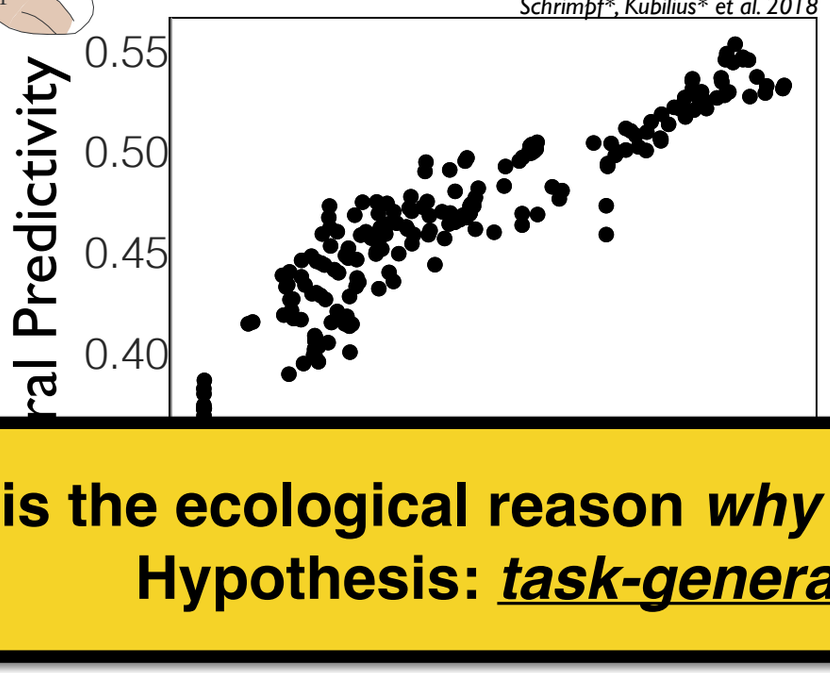


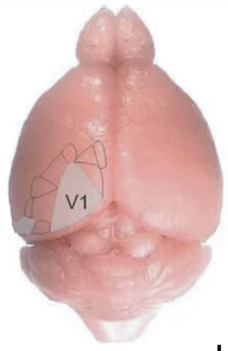
Object Categorization Ability **NOT** Correlated with Neural Predictivity

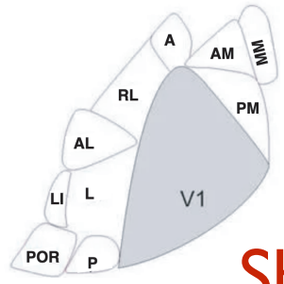


Primates

Schrimpf, Kubilius* et al. 2018*



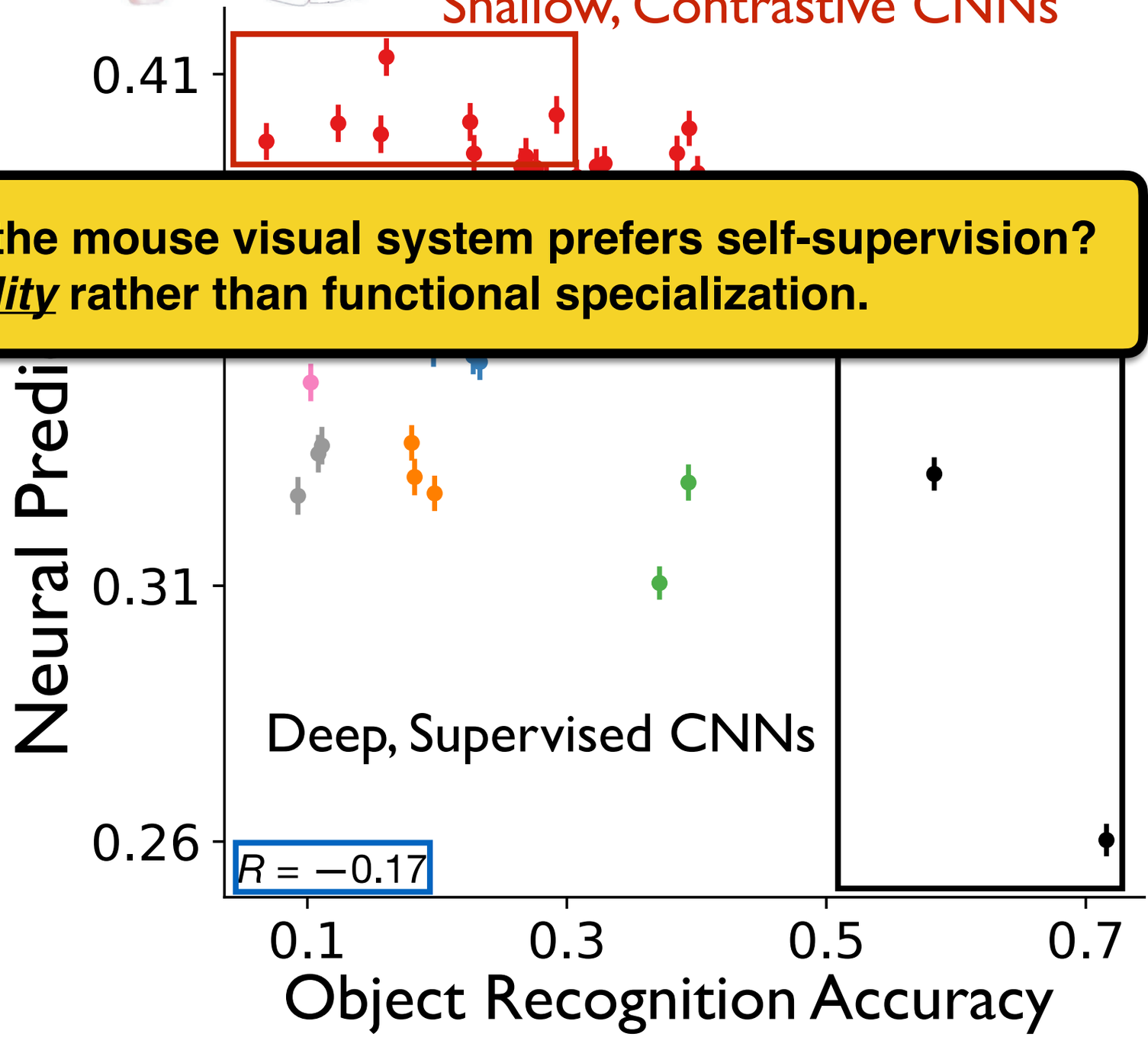




Mouse

Shallow, Contrastive CNNs

What is the ecological reason *why* the mouse visual system prefers self-supervision?
Hypothesis: *task-generality* rather than functional specialization.



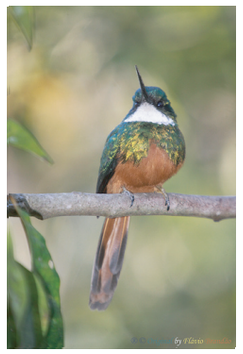
Assessing Task-Generality

Assessing Task-Generality

Train

ImageNet



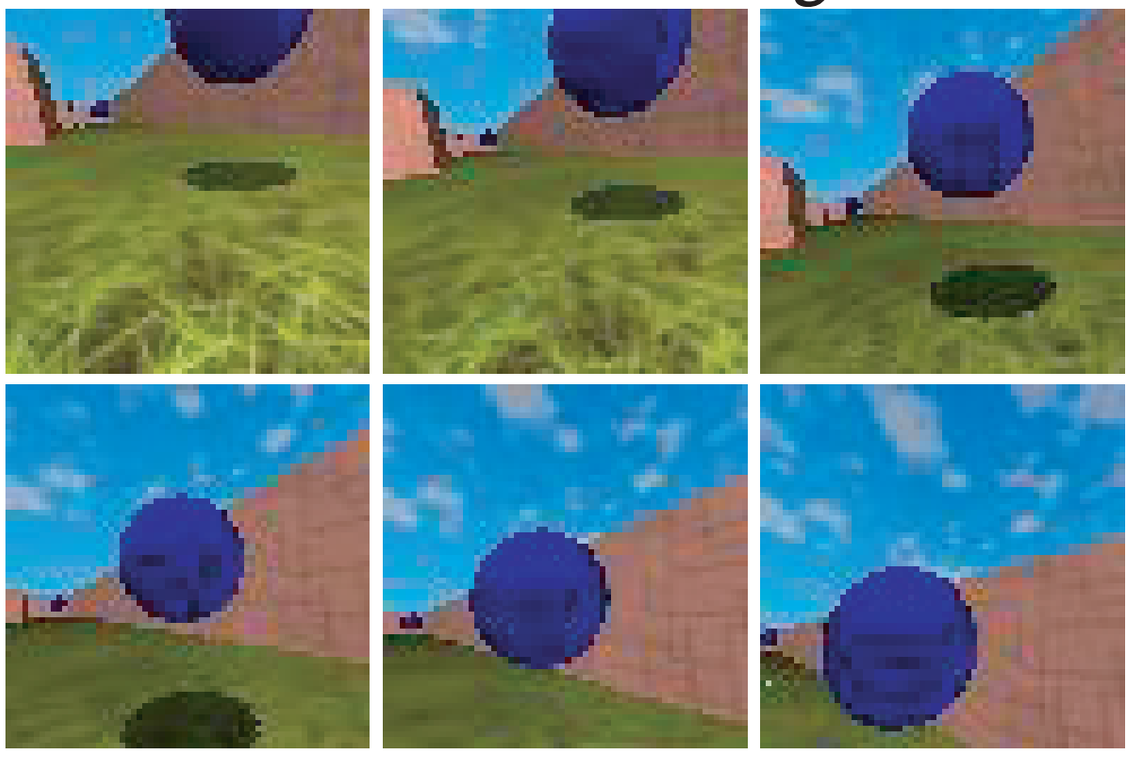

Assessing Task-Generality

Train

ImageNet

Evaluate

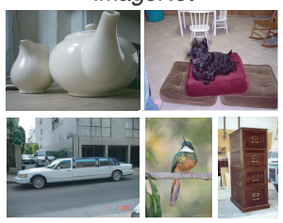
Reward-Based Navigation



Assessing Task-Generality

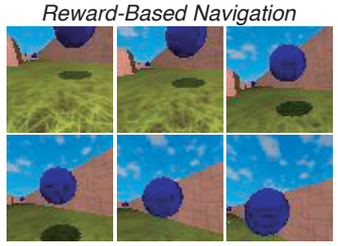
Train

ImageNet



Evaluate

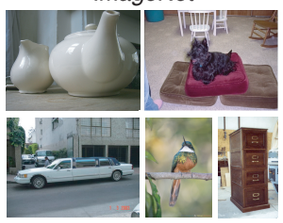
Reward-Based Navigation



Assessing Task-Generality

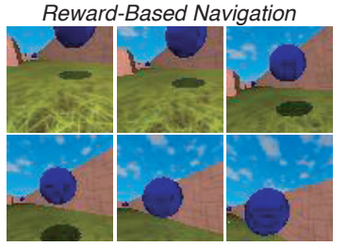
Train

ImageNet

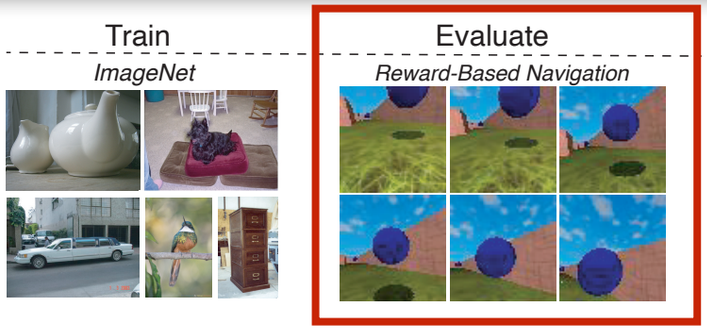


Evaluate

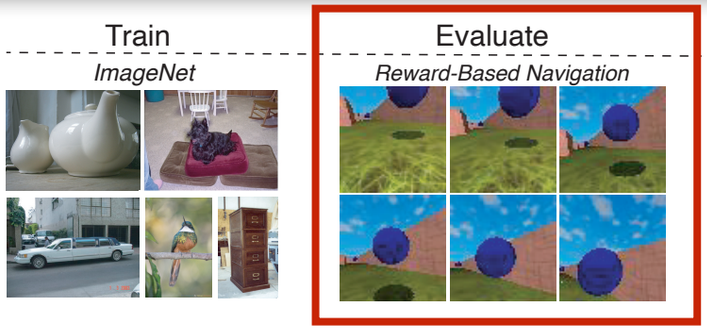
Reward-Based Navigation

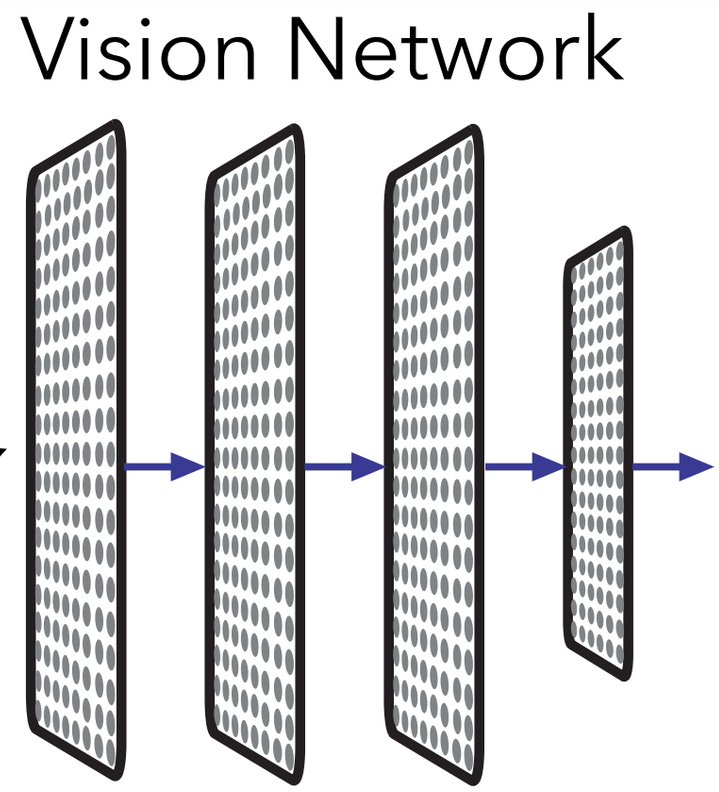


Embodied Virtual Rodent Navigation

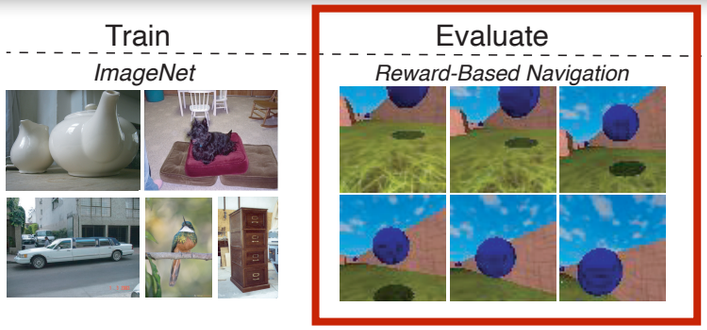


Embodied Virtual Rodent Navigation

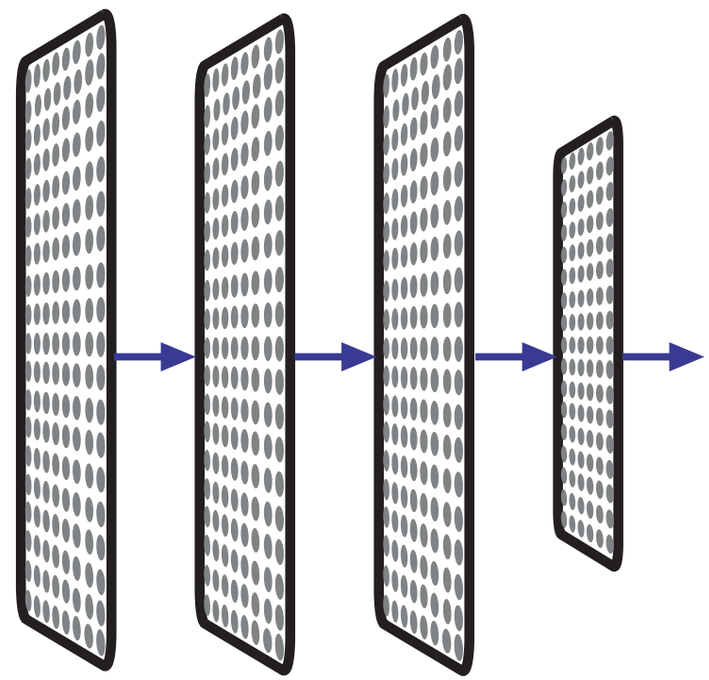


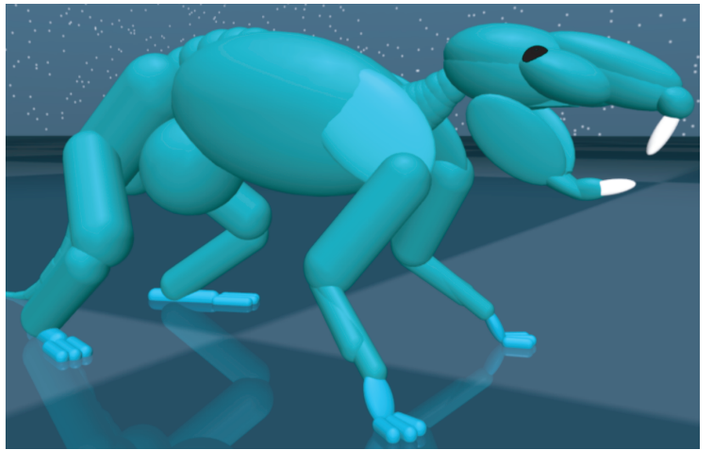


Embodied Virtual Rodent Navigation



Vision Network

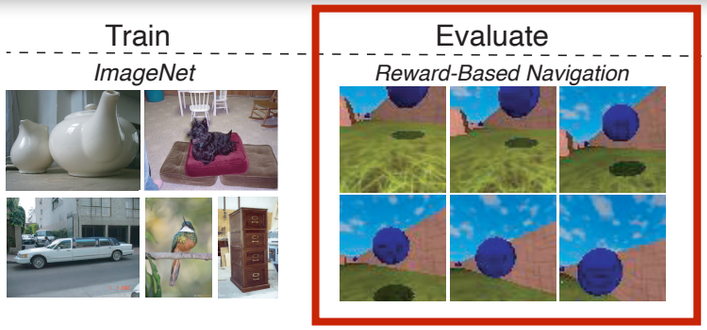




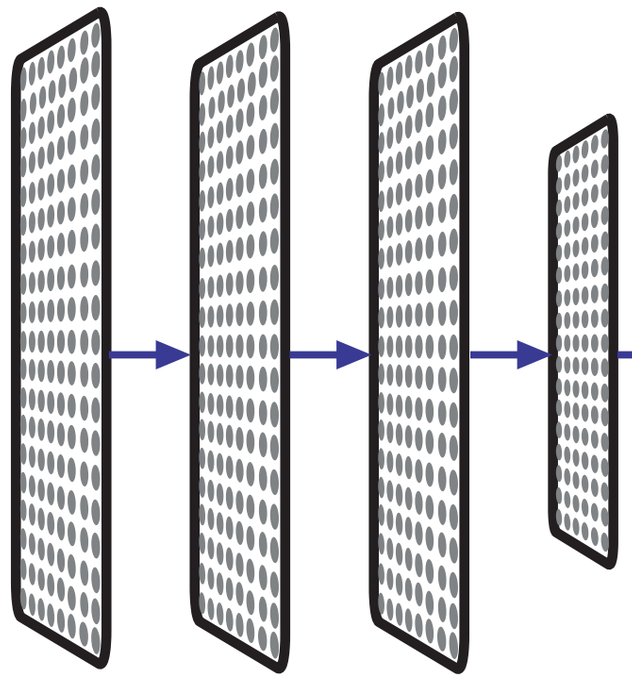
Biomechanical Model
(Joint angles, accelerometer, etc.)

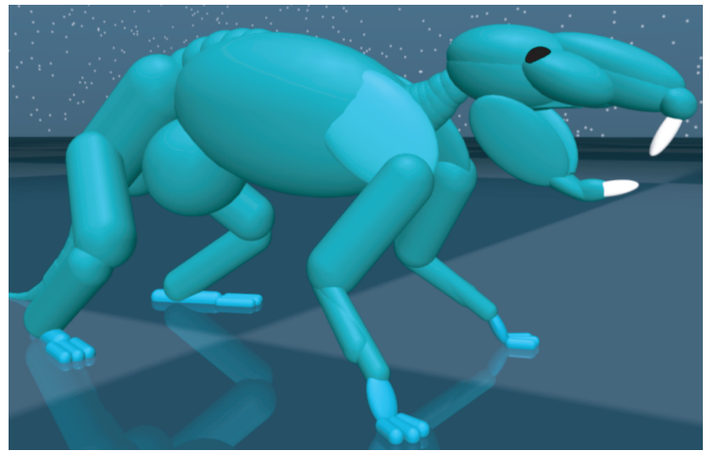
Bence Ölveczky

Embodied Virtual Rodent Navigation



Vision Network

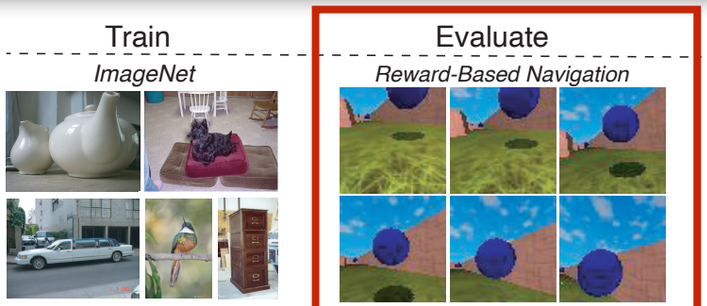




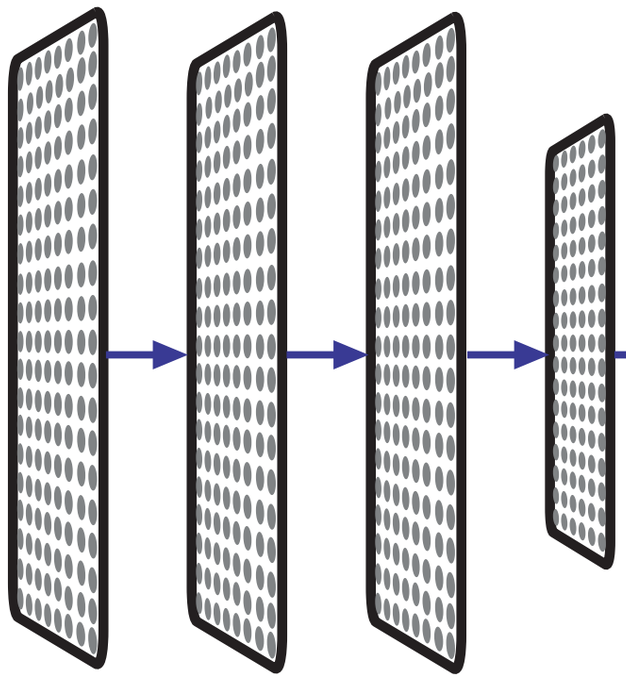
Biomechanical Model
(Joint angles, accelerometer, etc.)

Bence Ölveczky

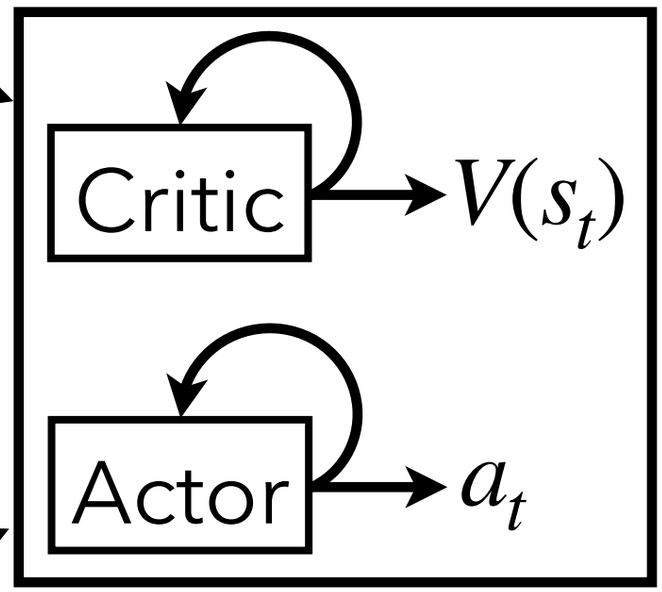
Embodied Virtual Rodent Navigation

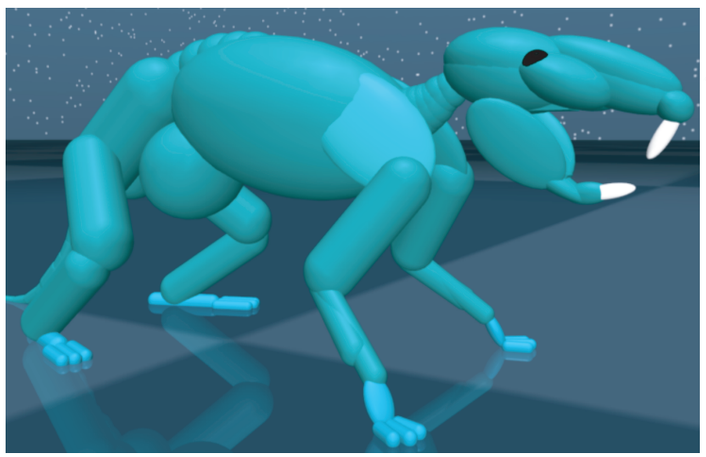


Vision Network



Decision Making

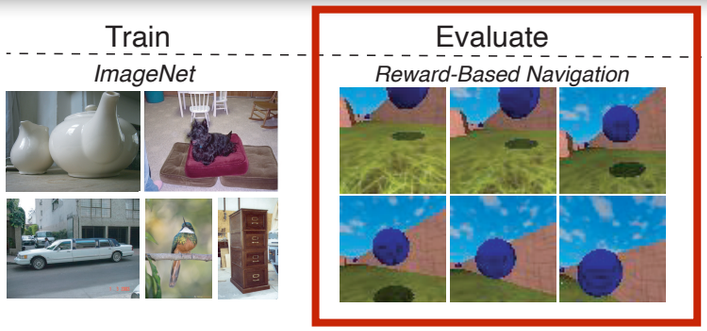




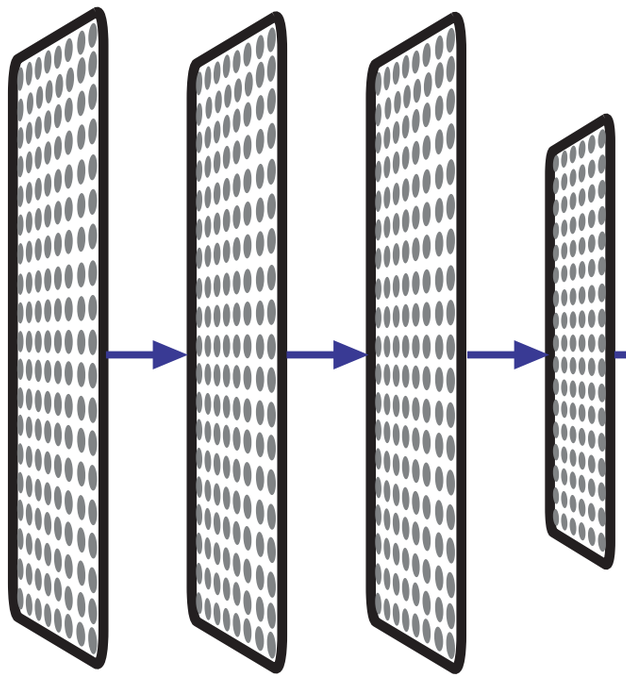
Biomechanical Model
(Joint angles, accelerometer, etc.)

Bence Ölveczky

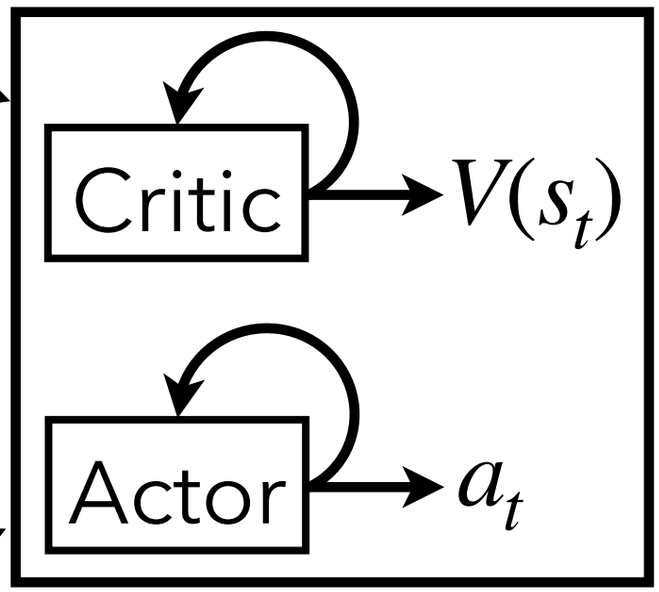
Embodied Virtual Rodent Navigation

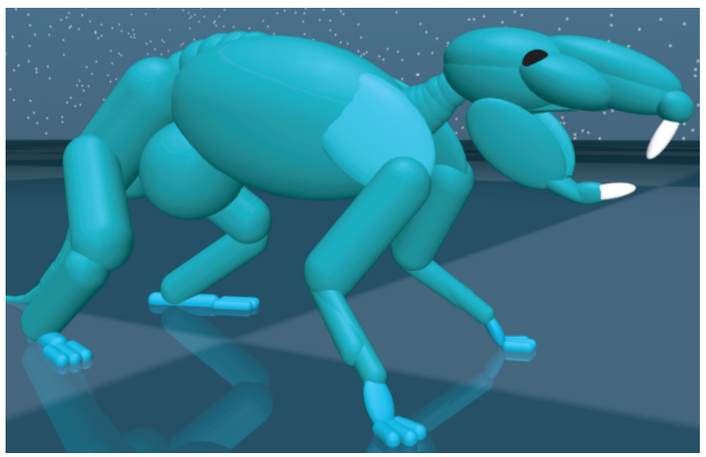


Vision Network



Decision Making

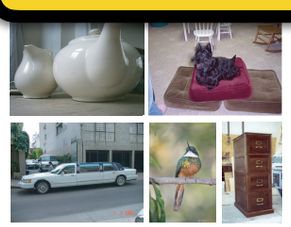


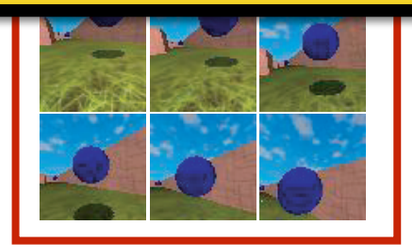


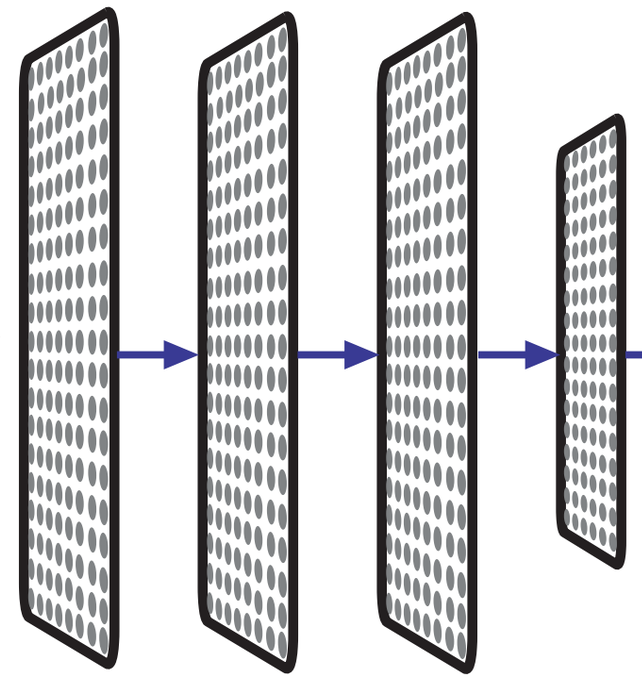
Biomechanical Model
(Joint angles, accelerometer, etc.)

Bence Ölveczky

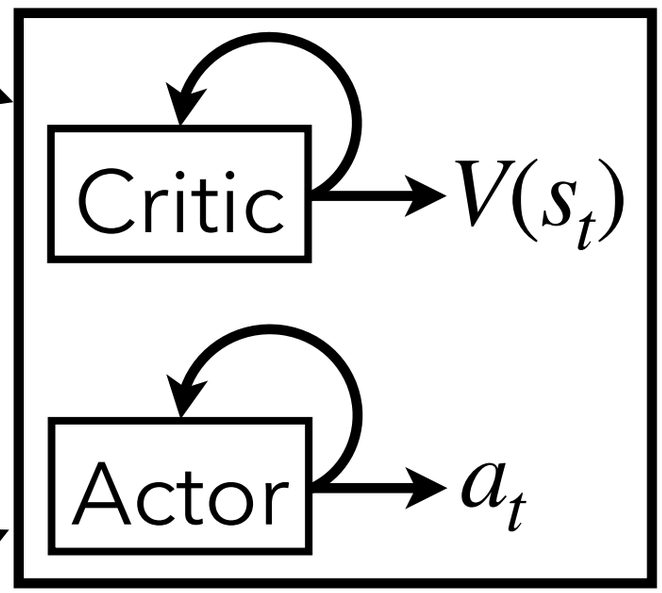
High degree-of-freedom body, keeping track of history over long timescales with high-dimensional, continuous inputs

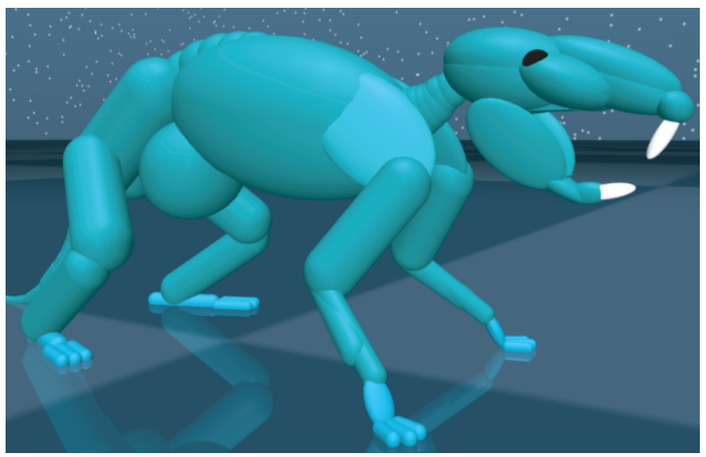






Decision Making



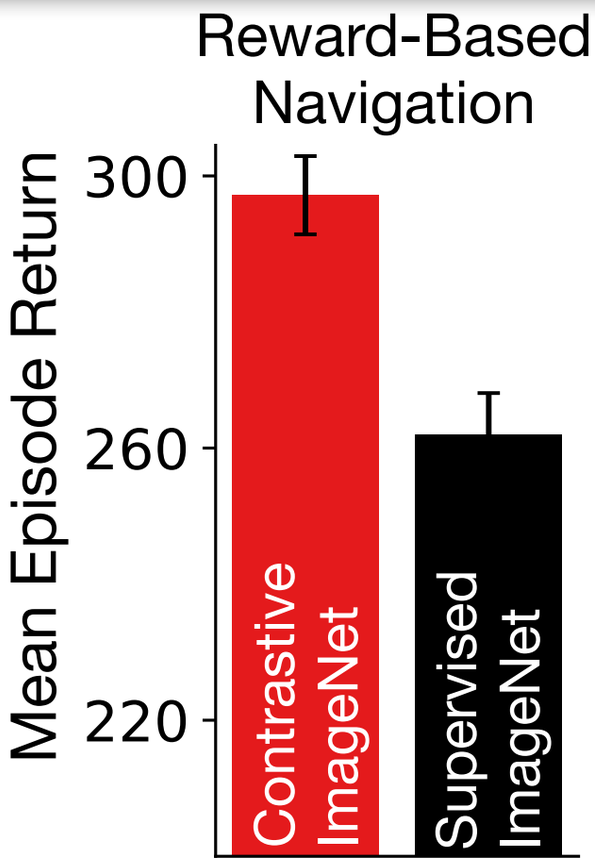


Biomechanical Model
(Joint angles, accelerometer, etc.)

Bence Ölveczky

Contrastive Models Yield Better Transfer Performance

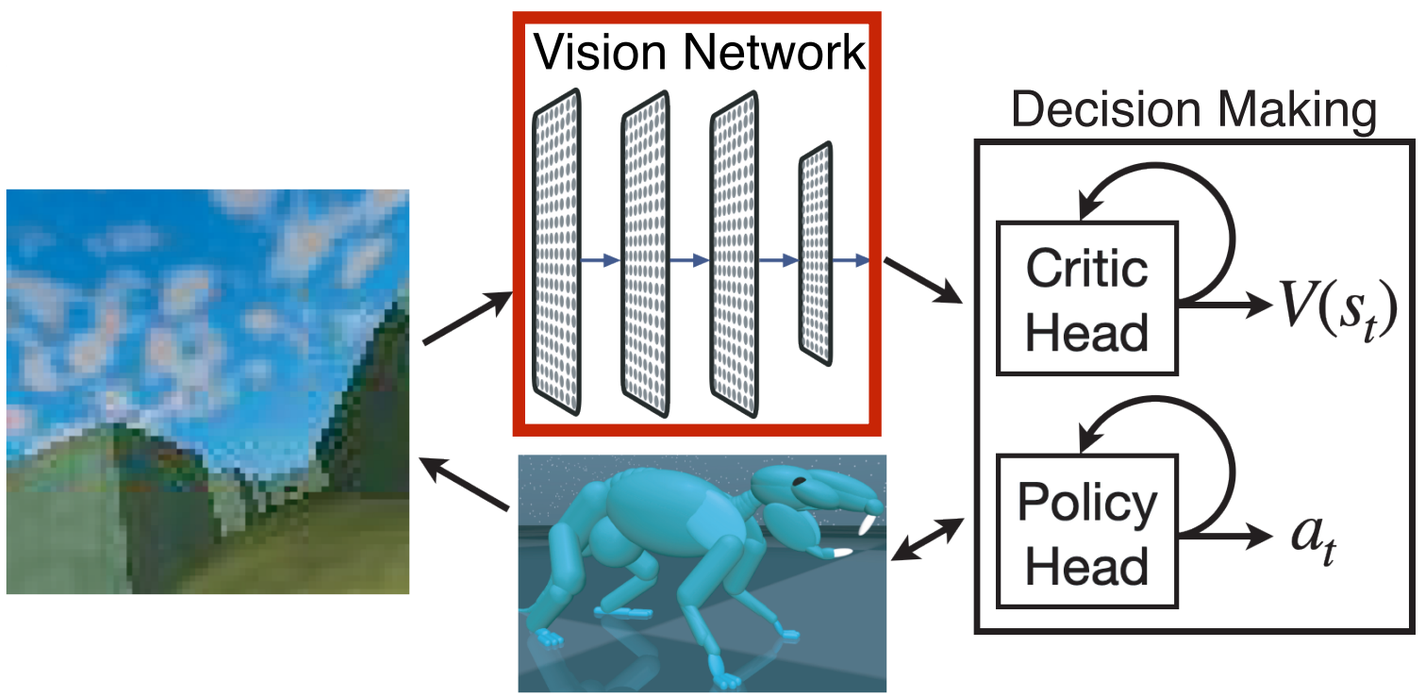
Contrastive Models Yield Better Transfer Performance



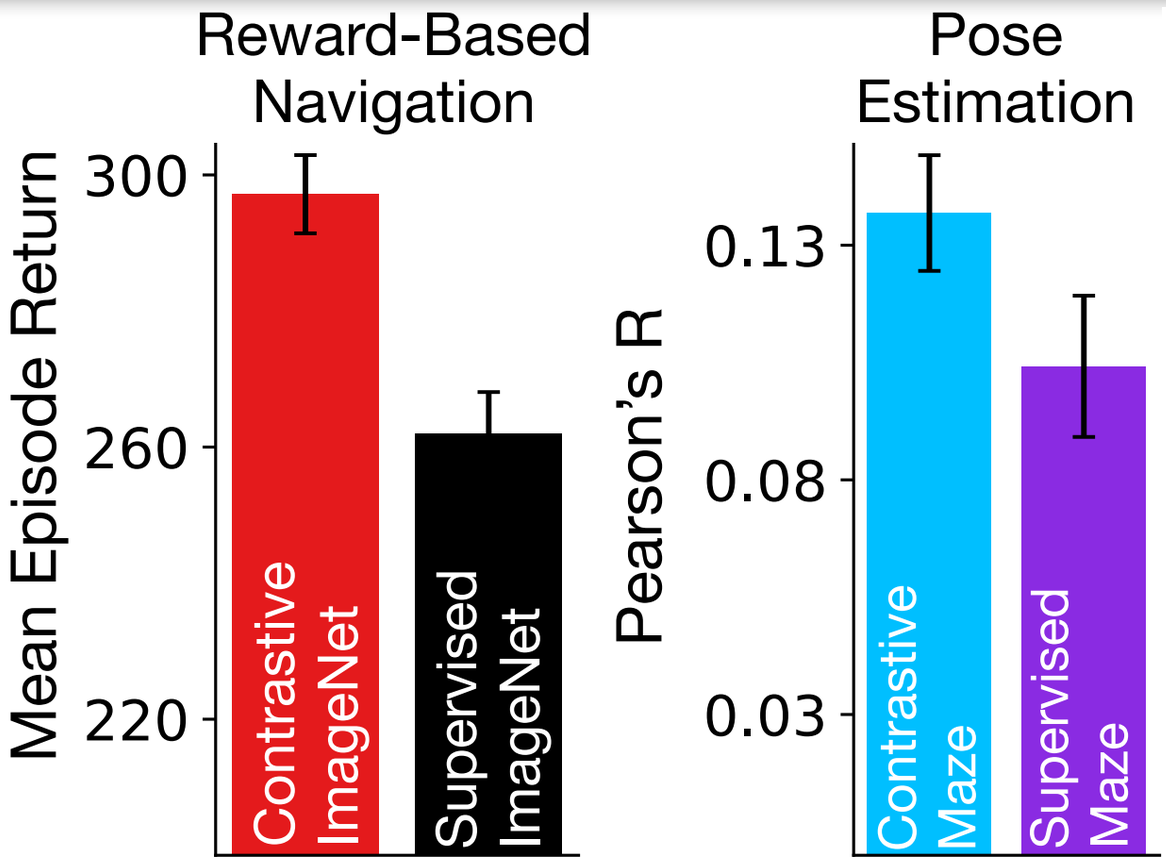
Train

ImageNet

Evaluate

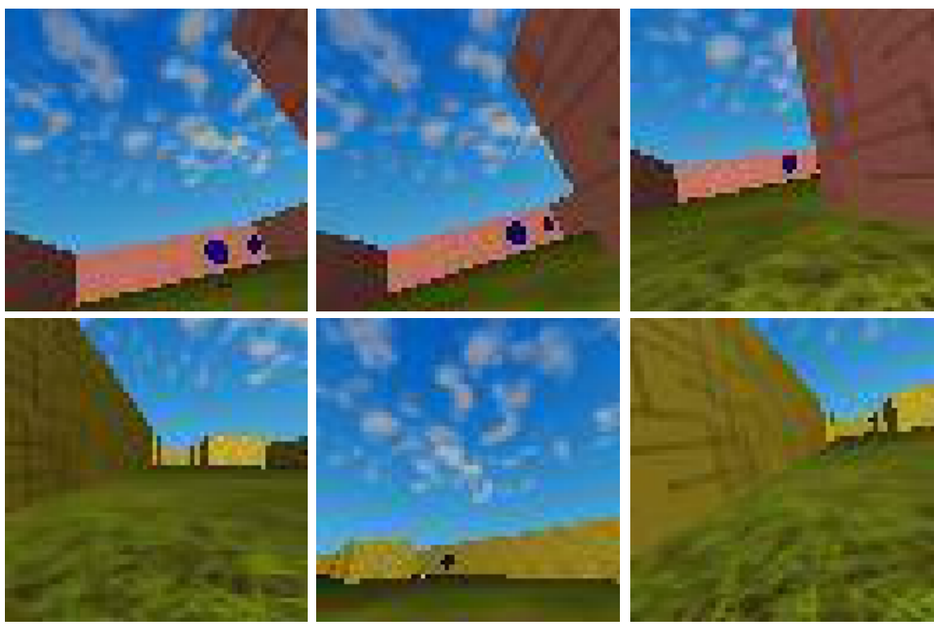


Contrastive Models Yield Better Transfer Performance



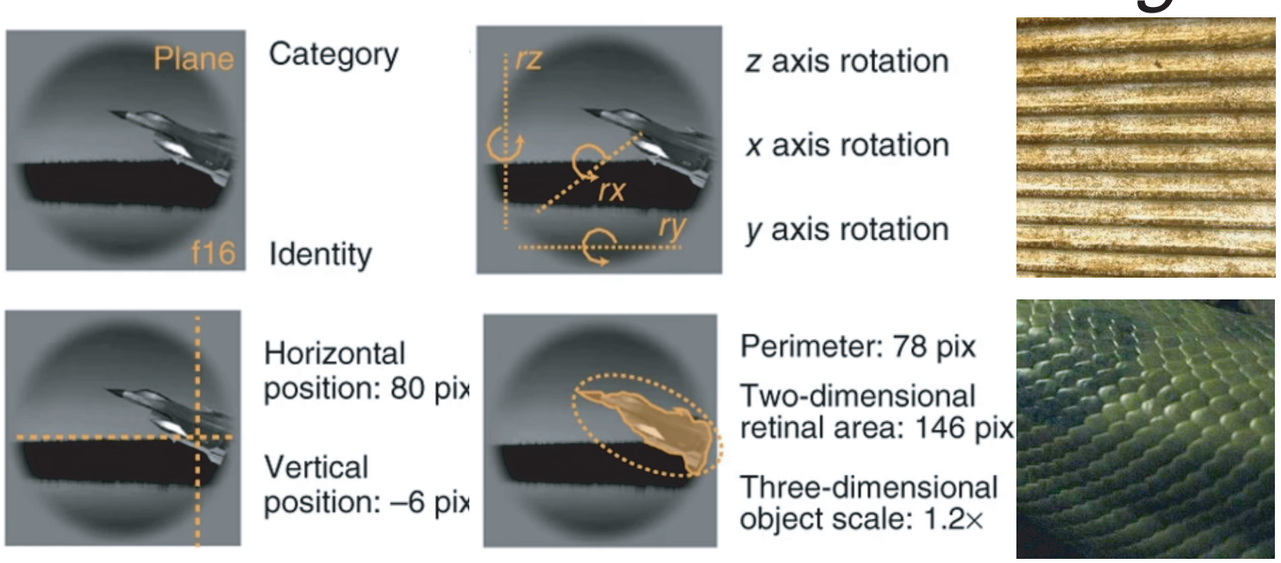
Train

Maze Environment



Evaluate

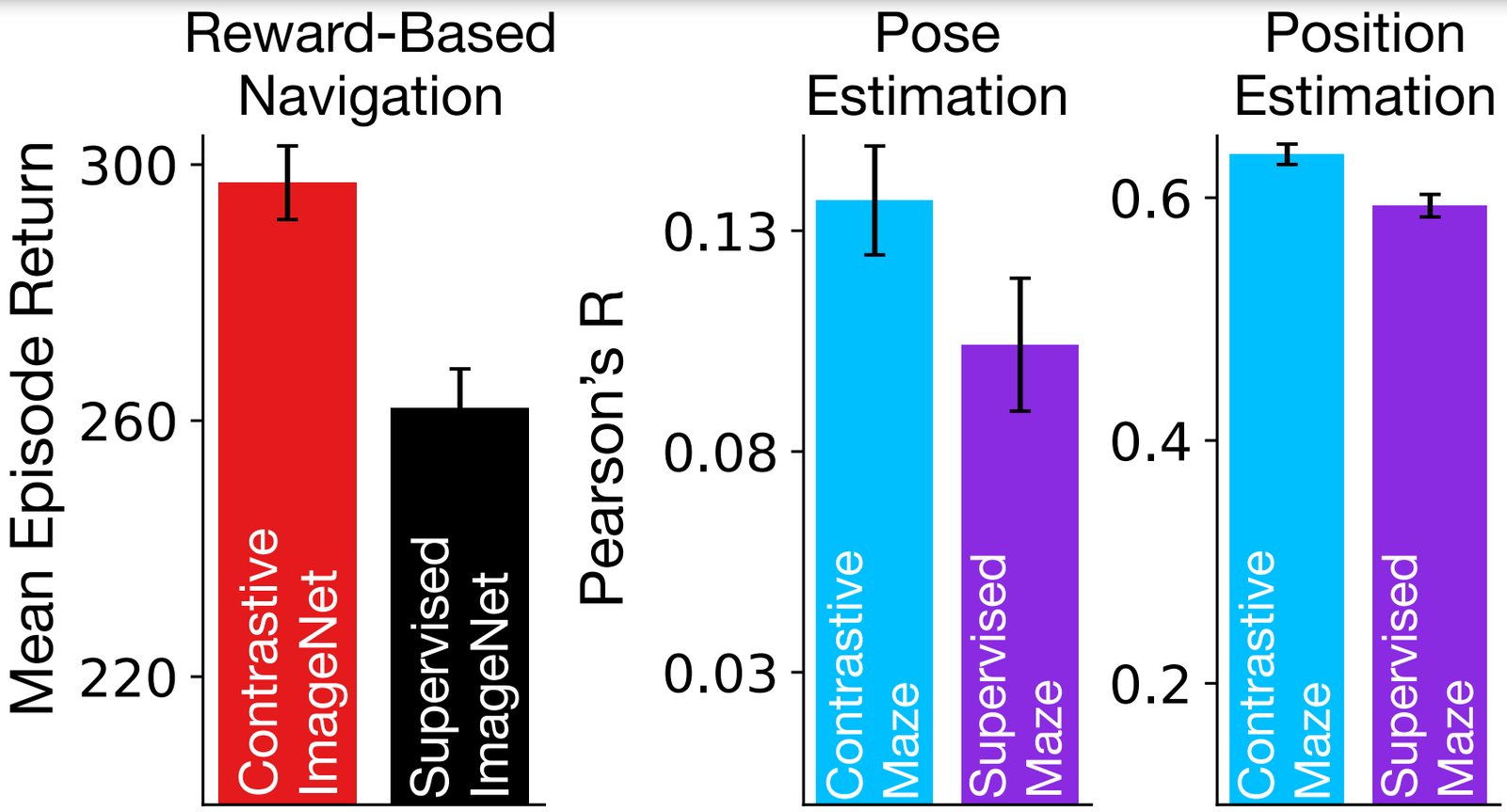
Visual Scene Understanding



Object properties

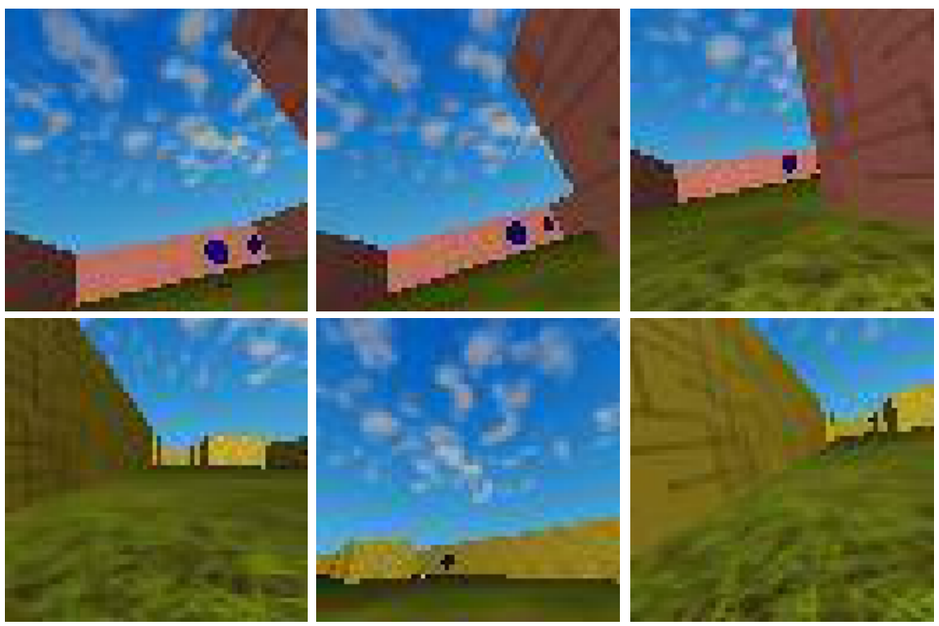
Texture

Contrastive Models Yield Better Transfer Performance



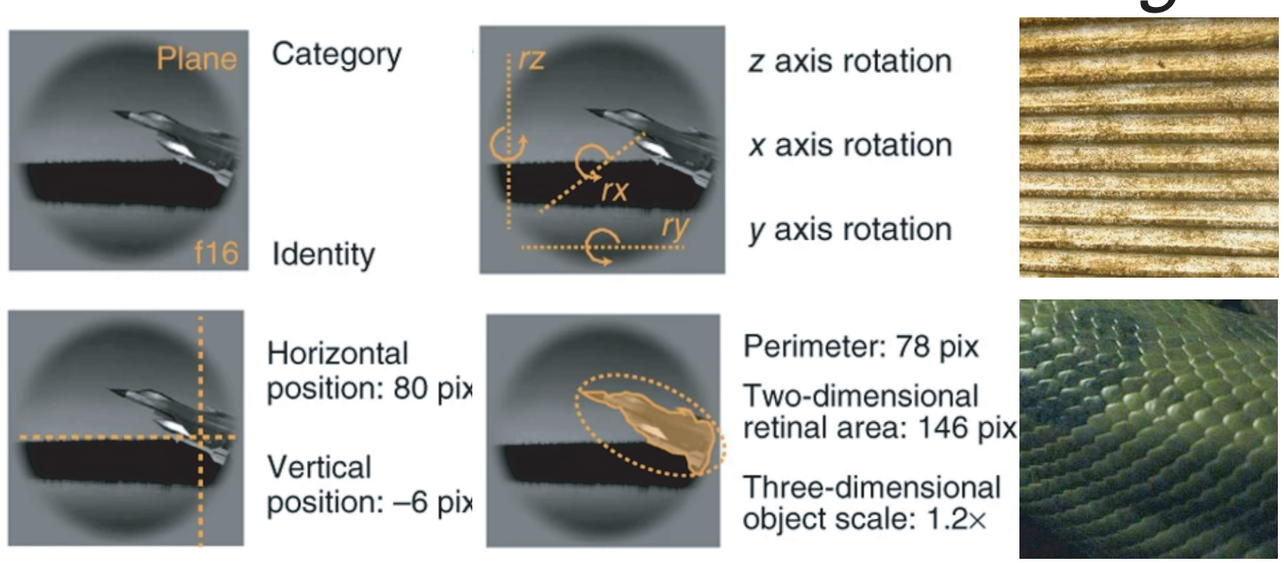
Train

Maze Environment



Evaluate

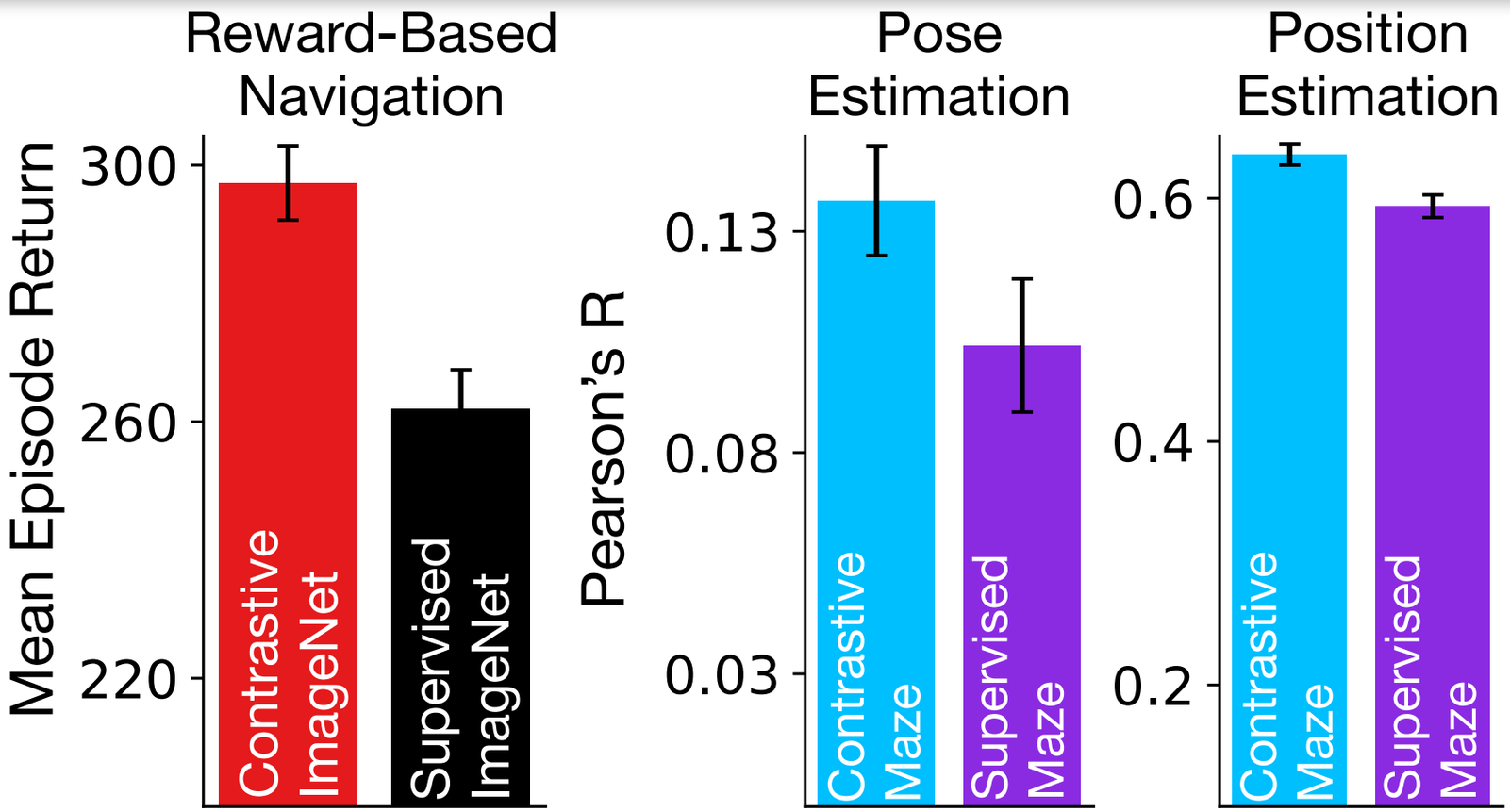
Visual Scene Understanding

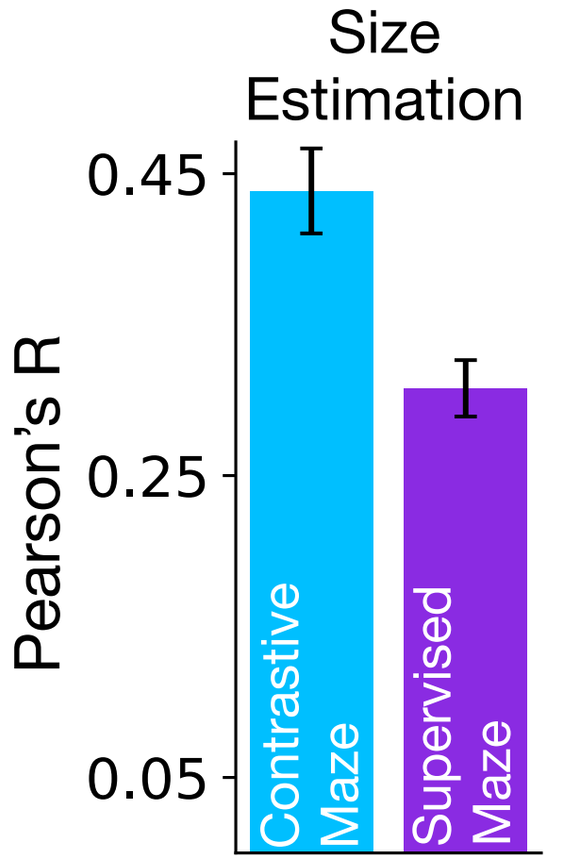


Object properties

Texture

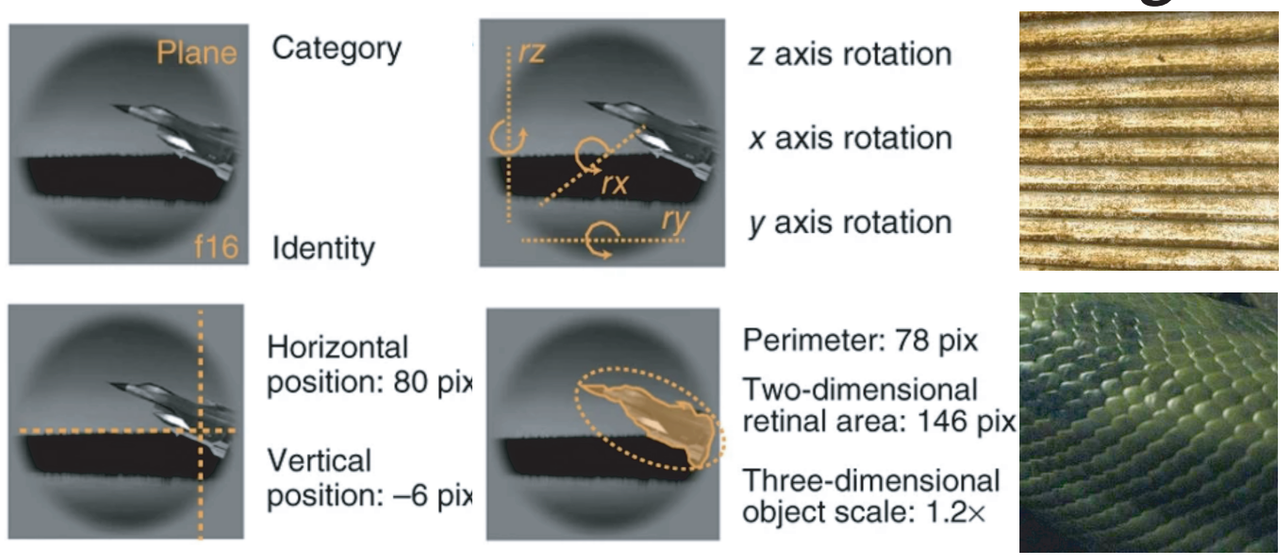
Contrastive Models Yield Better Transfer Performance





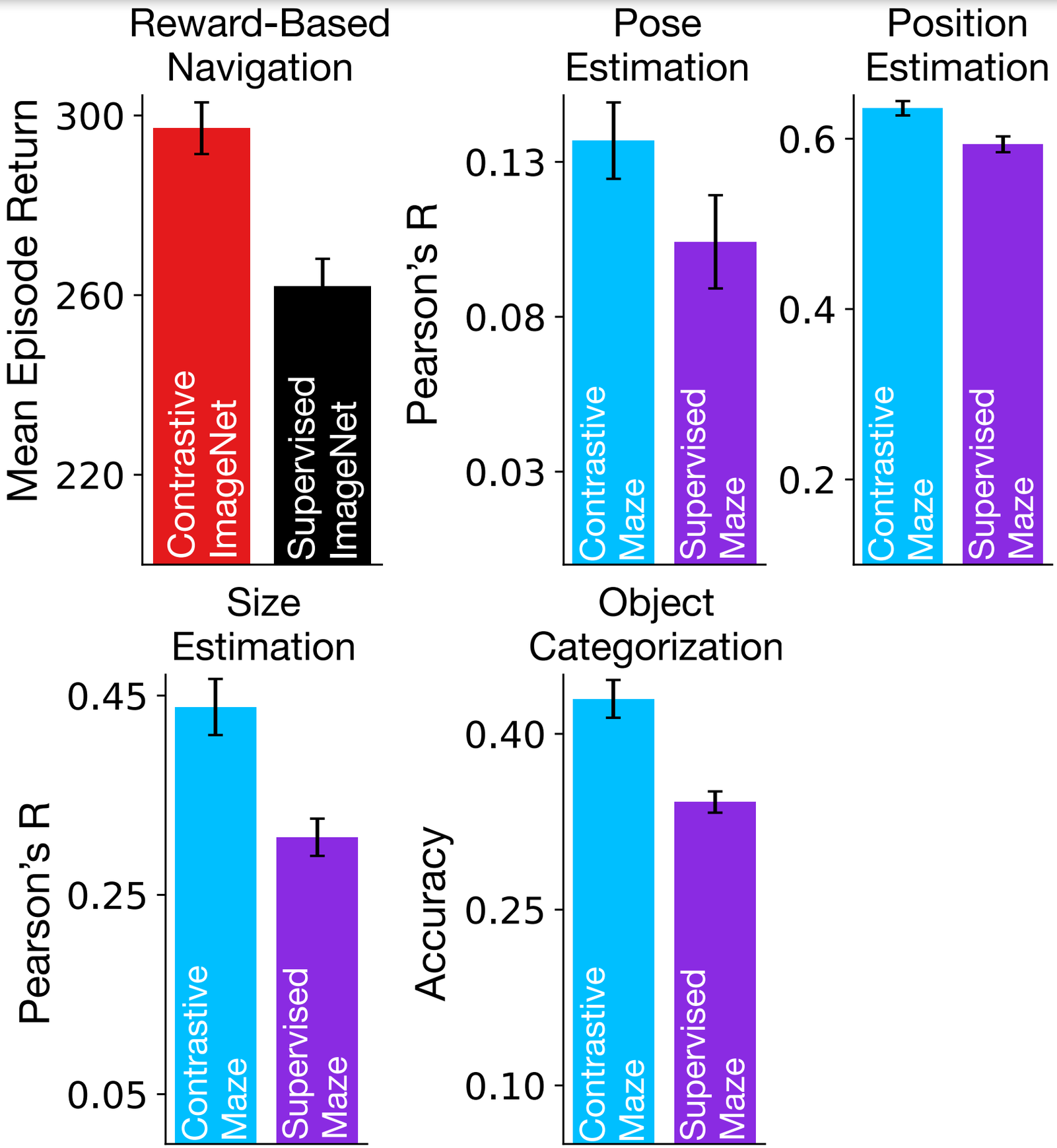
Evaluate

Visual Scene Understanding

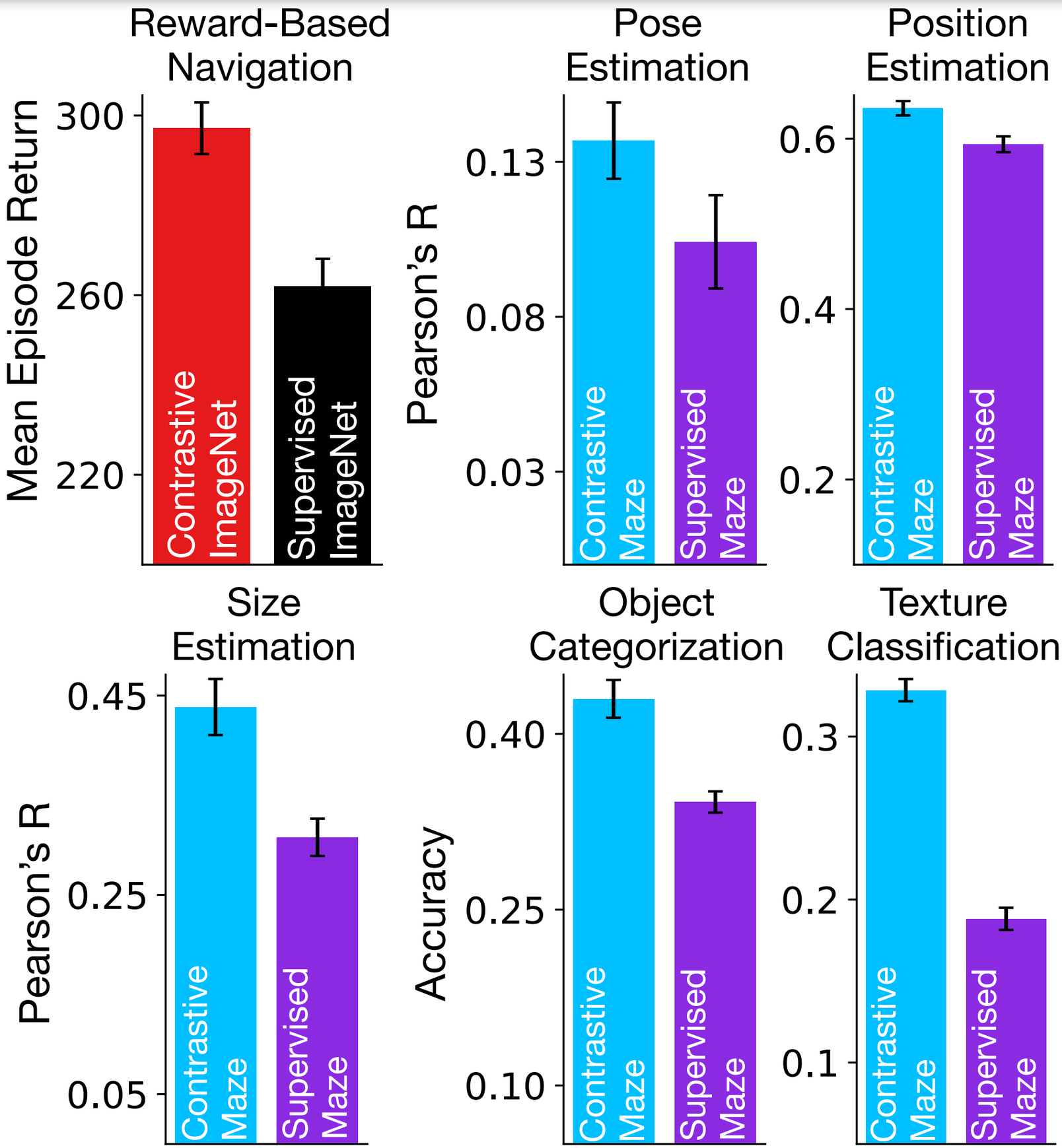


Object properties *Texture*

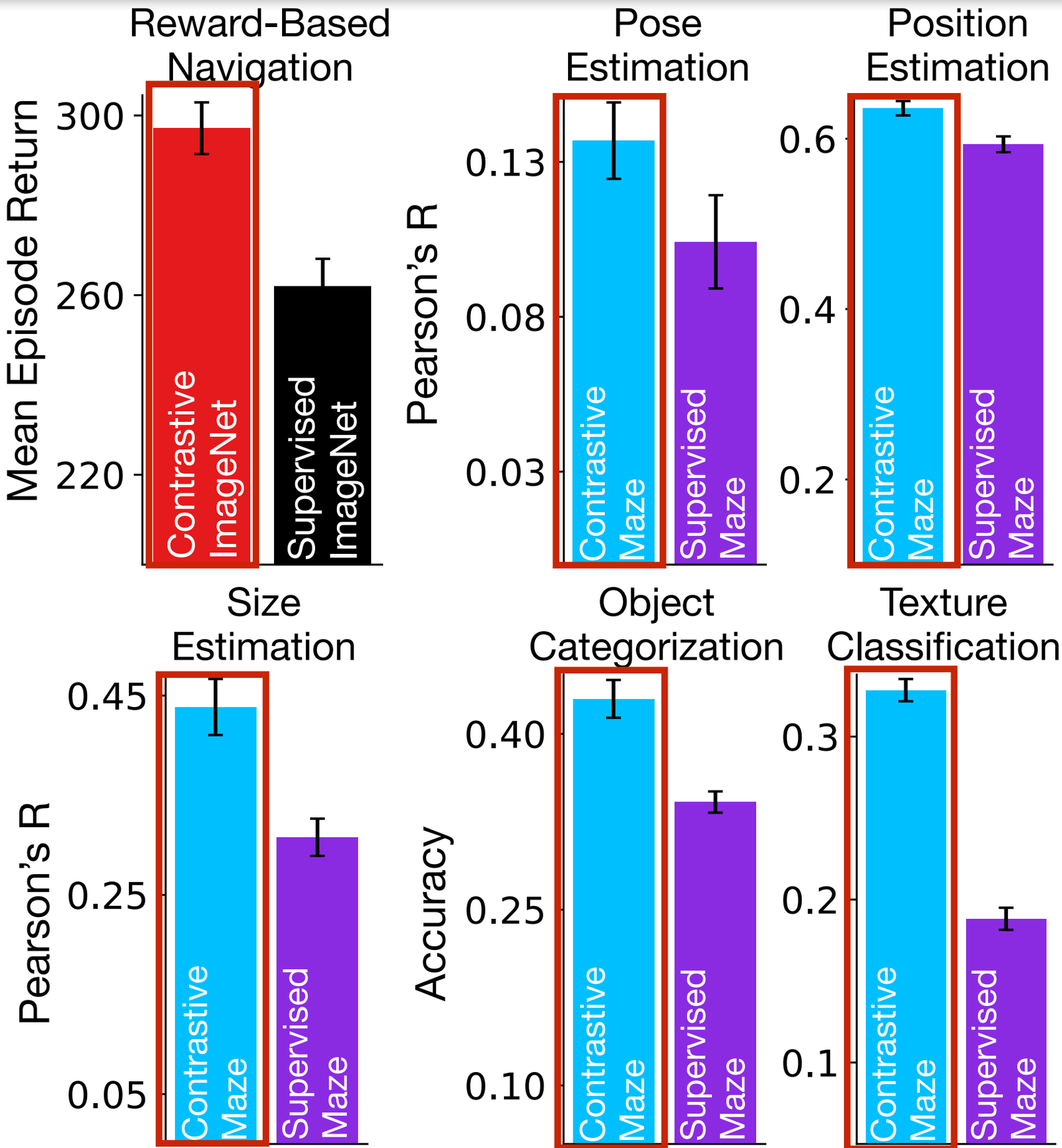
Contrastive Models Yield Better Transfer Performance



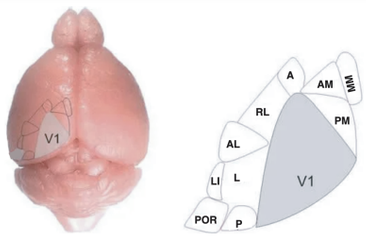
Contrastive Models Yield Better Transfer Performance



Contrastive Models Yield Better Transfer Performance



The best neural models have the best task transfer



Noise-Corrected Neural Predictivity (Pearson's R)

0.5

0.4

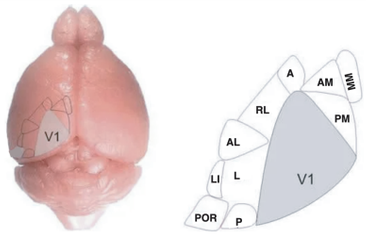
0.3

0.2

Inter-animal Consistency

Do the contrastive methods that task generalize best, also match the neurons better?

The best neural models have the best task transfer



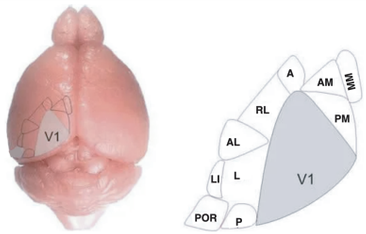
Noise-Corrected Neural Predictivity (Pearson's R)

0.5
0.4
0.3
0.2

Inter-animal Consistency

Contrastive ImageNet
Contrastive Maze
Supervised ImageNet
Supervised Maze

The best neural models have the best task transfer



Noise-Corrected Neural Predictivity (Pearson's R)

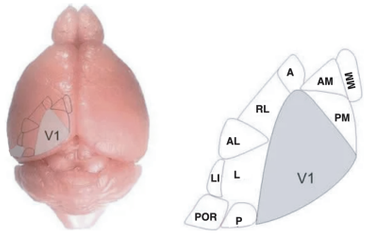
0.5
0.4
0.3
0.2

Inter-animal Consistency

Contrastive loss better than supervised or RL loss in *same* environment

Contrastive ImageNet
Contrastive Maze
Supervised ImageNet
Supervised Maze

The best neural models have the best task transfer



Noise-Corrected Neural Predictivity (Pearson's R)

0.5
0.4
0.3
0.2

Inter-animal Consistency

Contrastive loss better than supervised or RL loss in *same* environment

Contrastive ImageNet
Contrastive Maze
Supervised ImageNet
Supervised Maze

Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System
- Reusable Latent Representations for Primate Mental Simulation
- Heuristics for Interrogating Natural Intelligence

Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System

Mouse visual cortex (so far) is a low-acuity, shallow network that makes best use of the mouse's limited resources to create a general-purpose visual system, that can be deployed in novel environments and <u>embodied</u> contexts.

- Reusable Latent Representations for Primate Mental Simulation

- Heuristics for Interrogating Natural Intelligence

Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System

Mouse visual cortex (so far) is a low-acuity, shallow network that makes best use of the mouse's limited resources to create a general-purpose visual system, that can be deployed in novel environments and <u>embodied</u> contexts.

- Reusable Latent Representations for Primate Mental Simulation

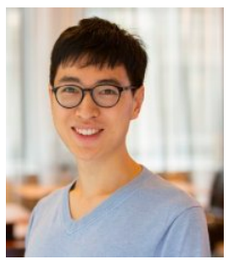
- Heuristics for Interrogating Natural Intelligence

Reusable Latent Representations for Primate 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

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

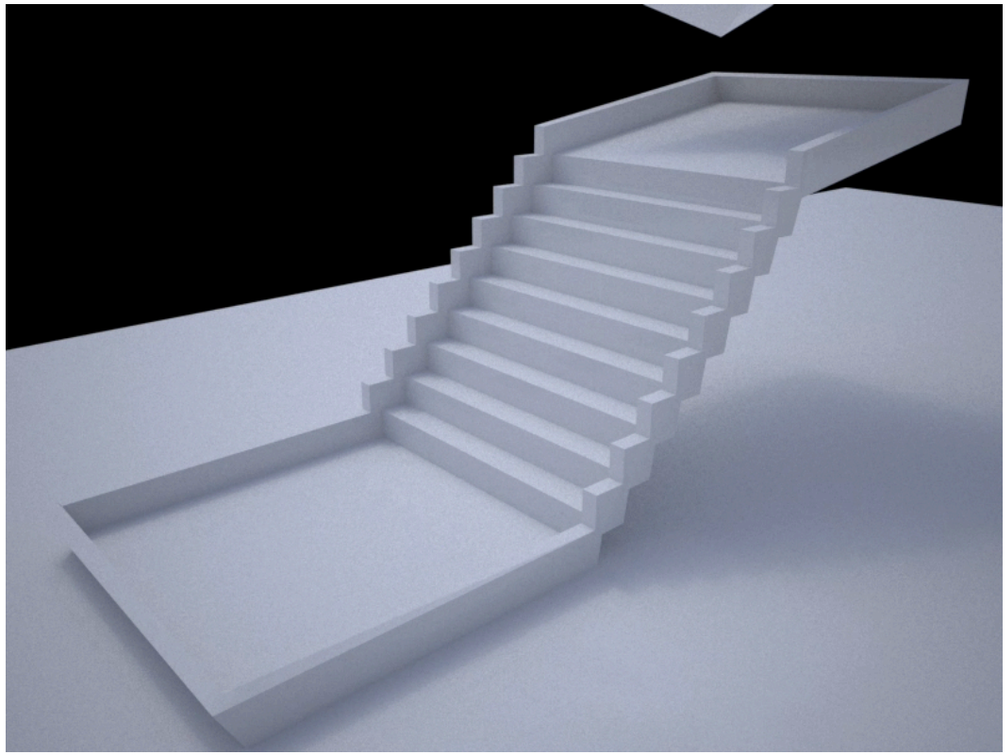
Infer:
Has this ice block been out longer?

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

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

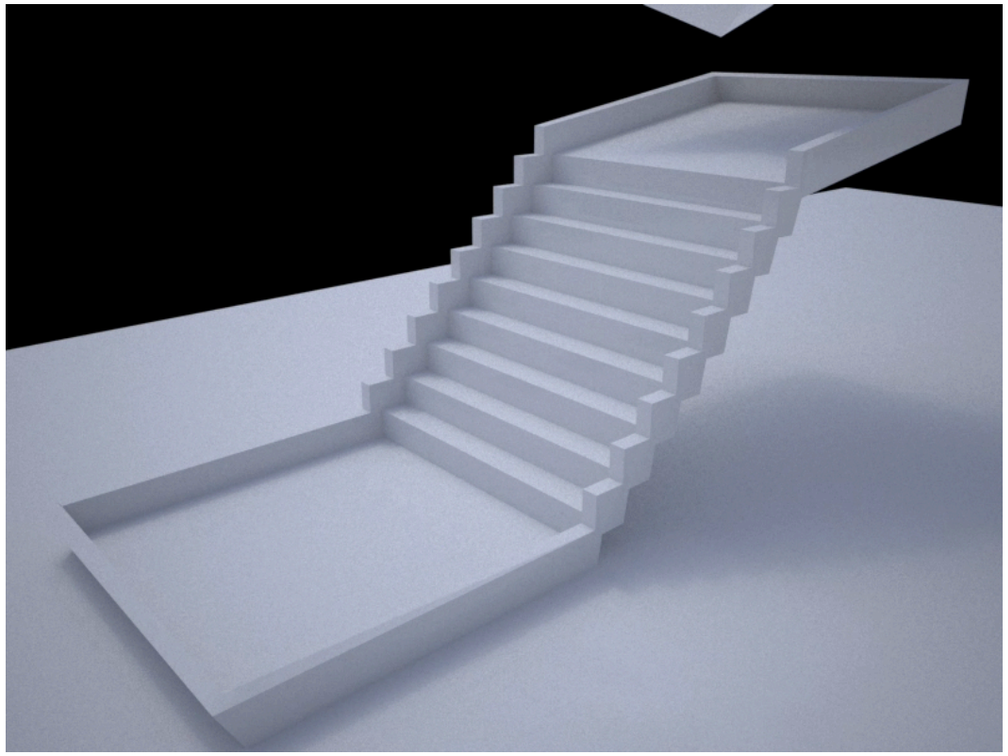


Plan:
How would I take these hats off the rack?

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

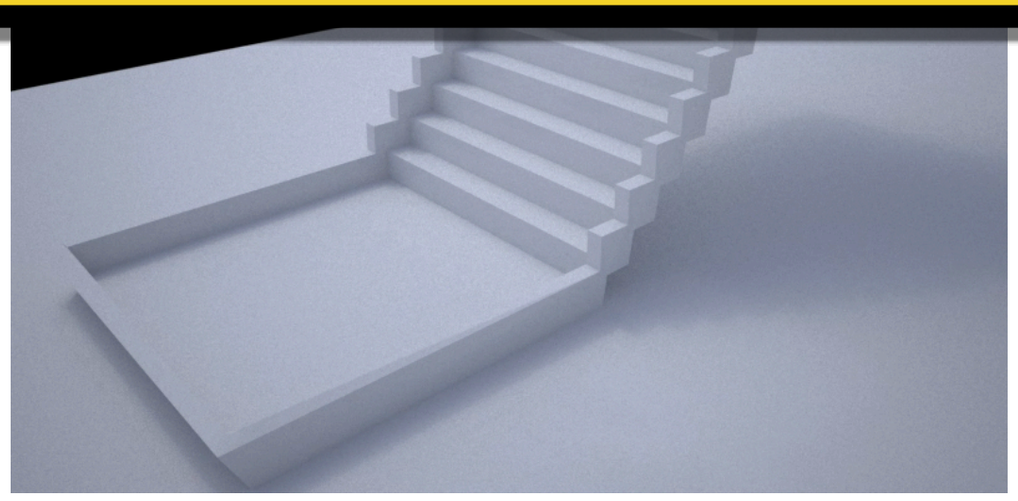


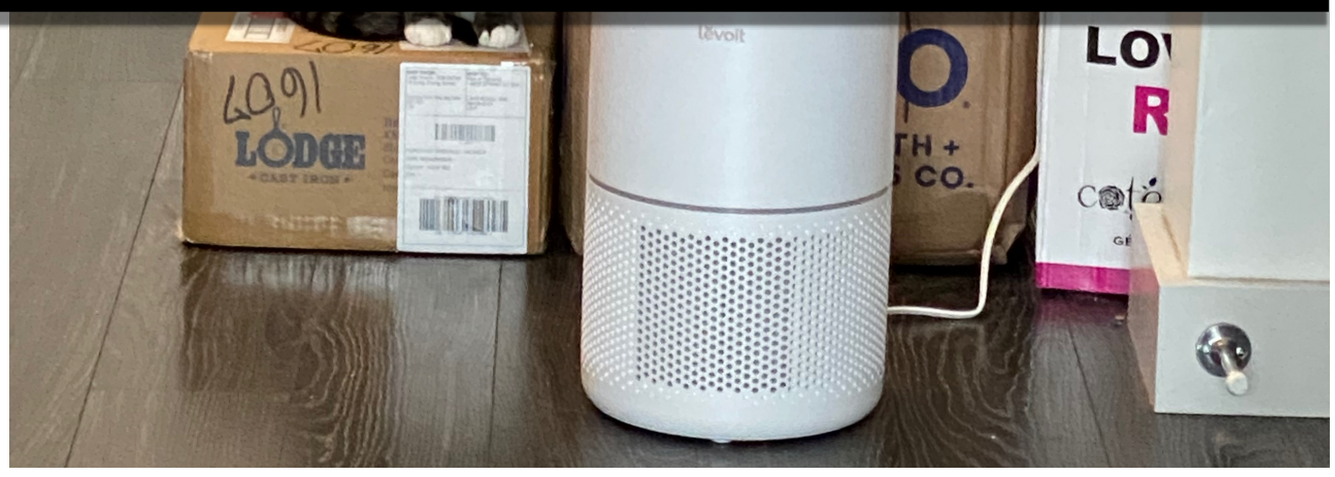
Plan:
How would I take these hats off the rack?

Predict:
Will this box support me?

Neurobiological Puzzle:

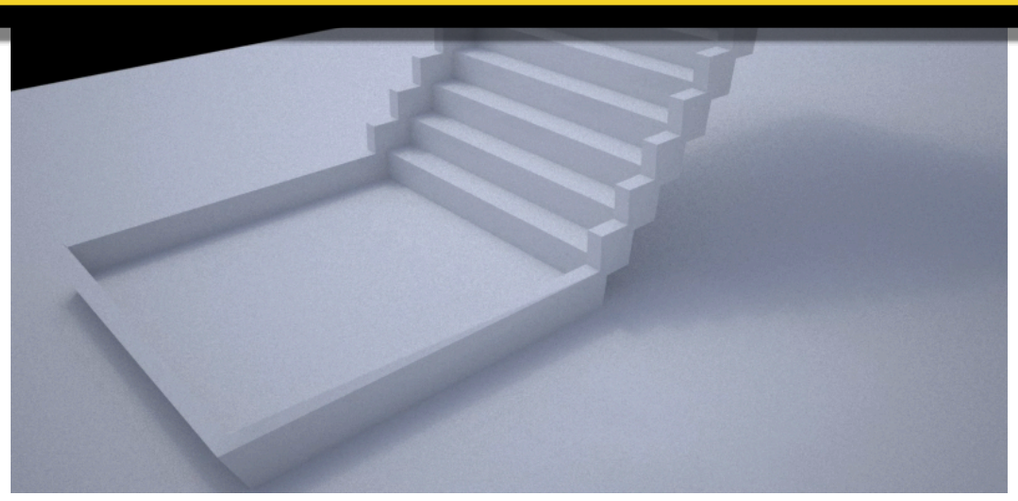
What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

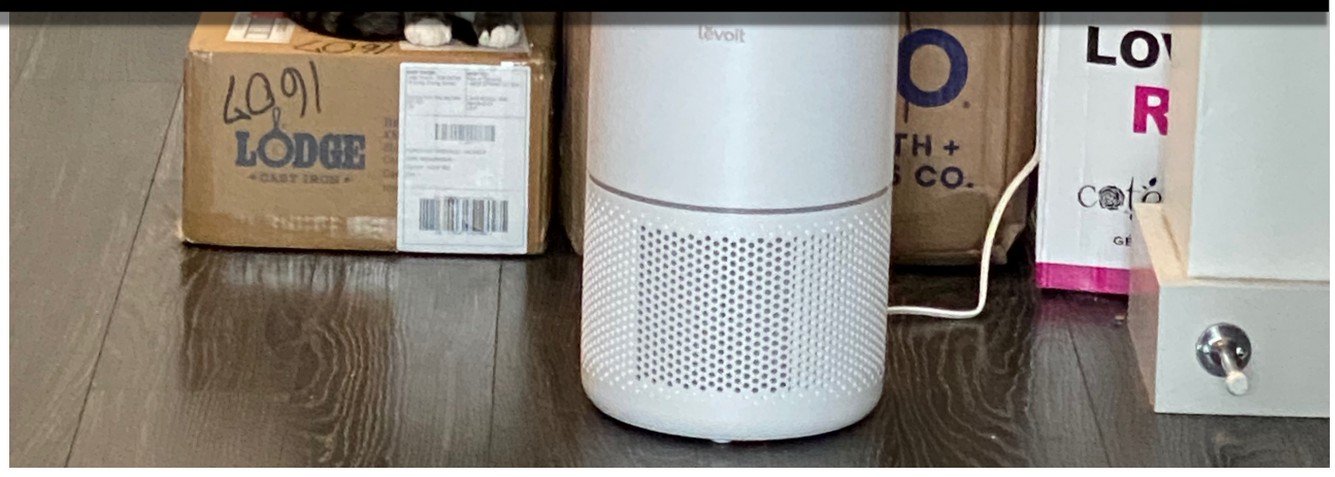




Neurobiological Puzzle:

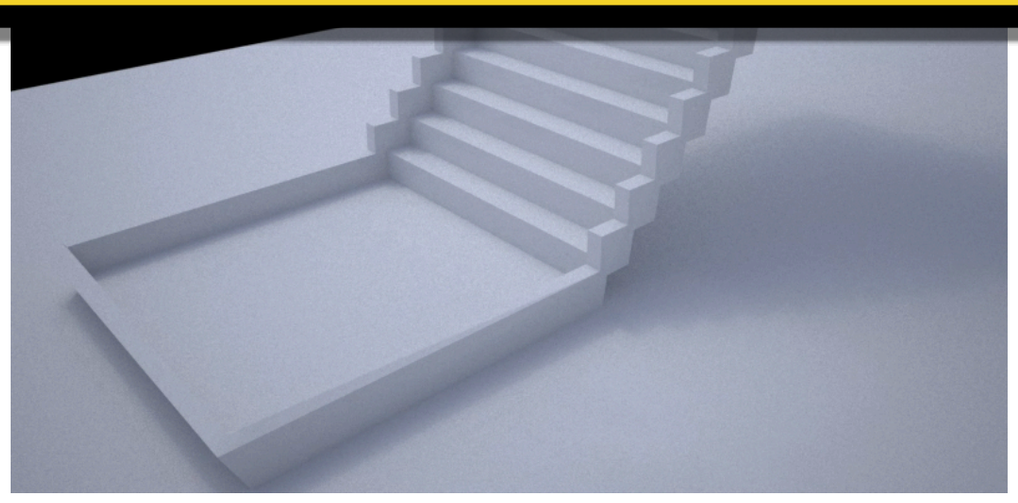
What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

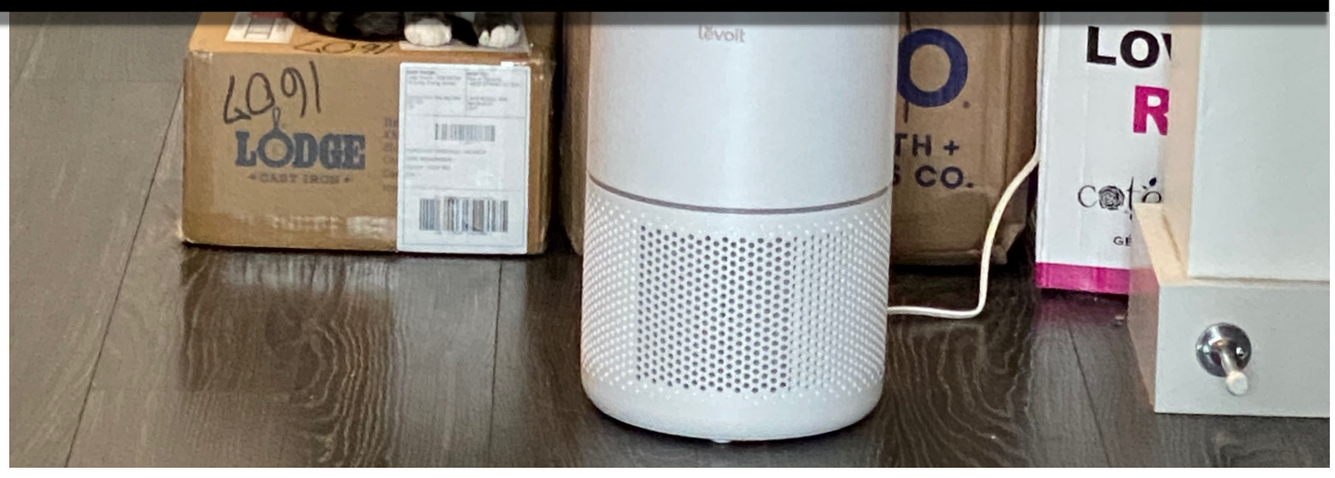




Neurobiological Puzzle:

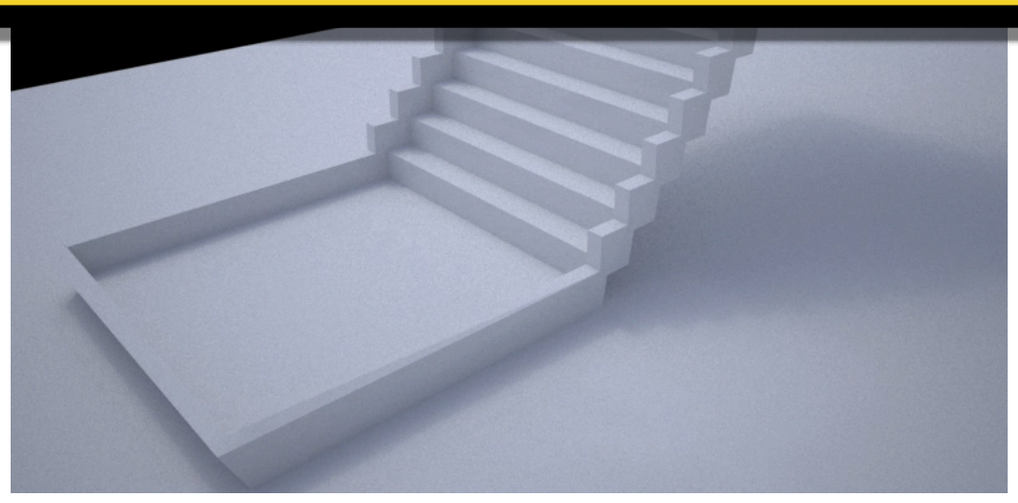
What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

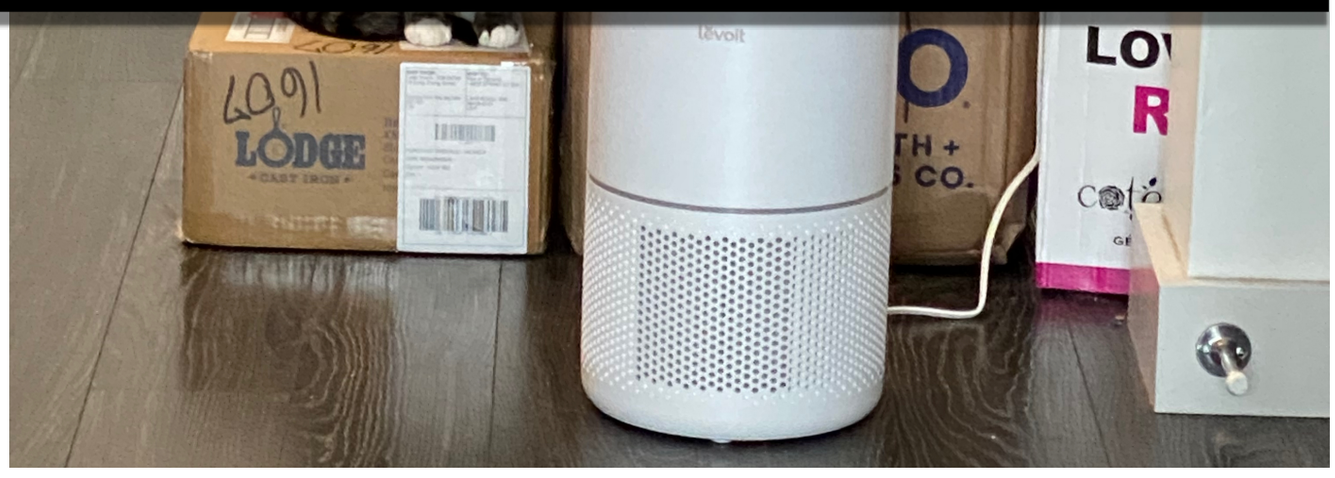




Neurobiological Puzzle:

What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?





Defining Hypotheses

Defining Hypotheses

"Sensory-Cognitive Networks"

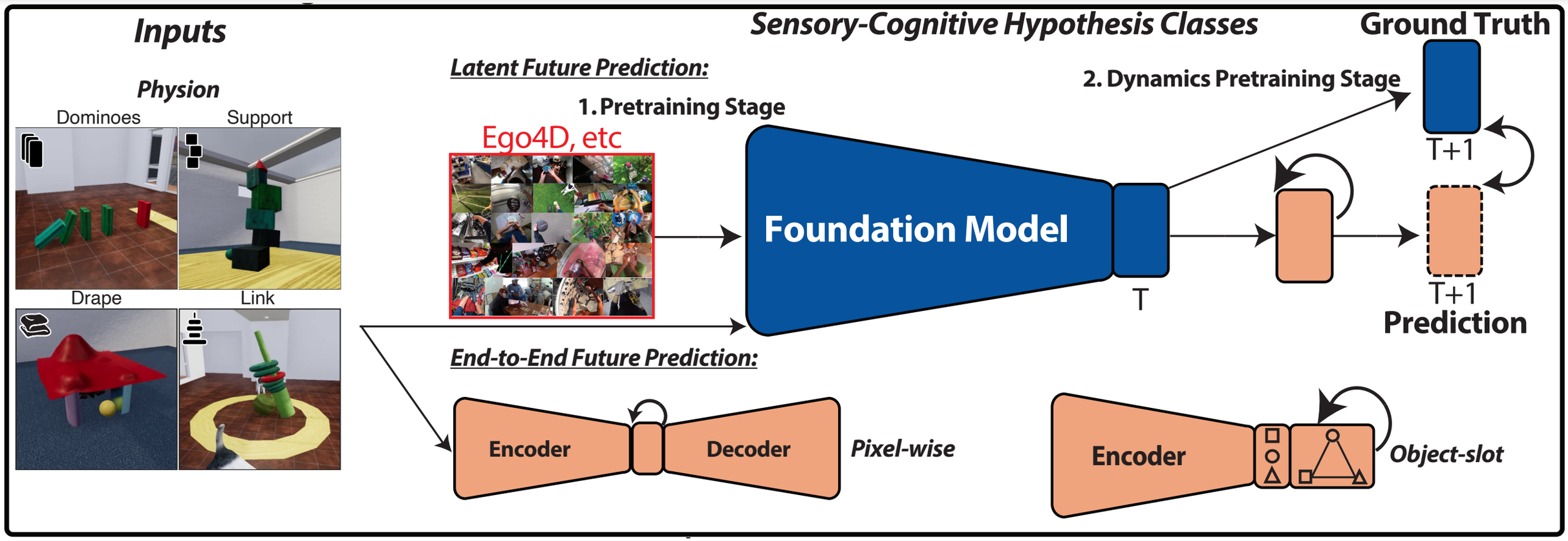
R1 (Input-Driven): Take in unstructured visual inputs across a range of physical phenomena.

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

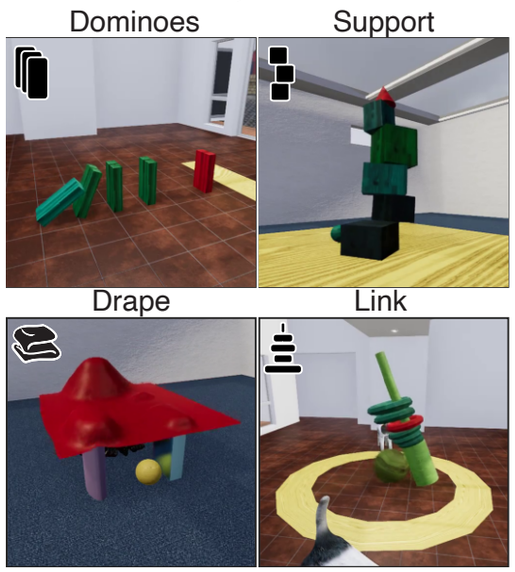
Overall Approach: Sensory-Cognitive Hypotheses



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

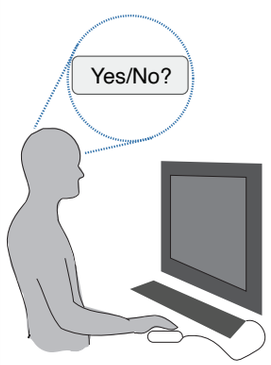
Ground Truth

End-to-End Future Prediction:

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

Encoder ***Object-slot***

Human Behavior: Physion Object Contact Prediction (OCP)



Observed Stimuli

Unobserved Outcome

Time

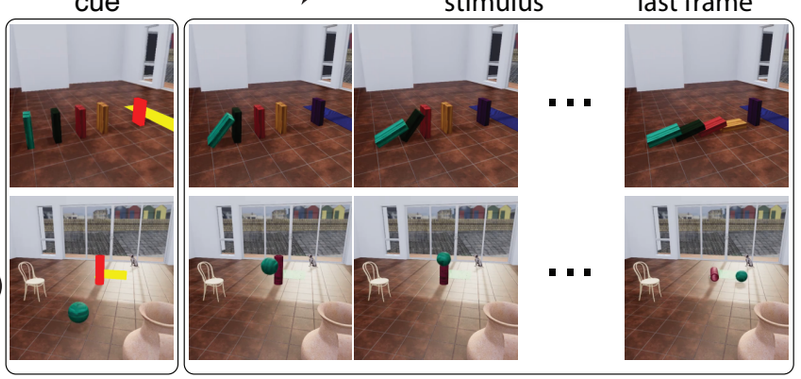
cue

stimulus

last frame

true label

Example Scenarios



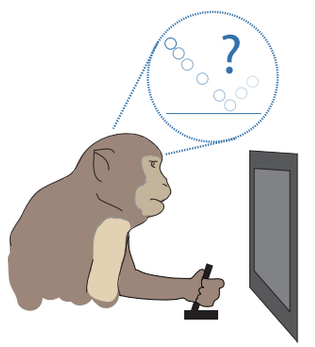
NO

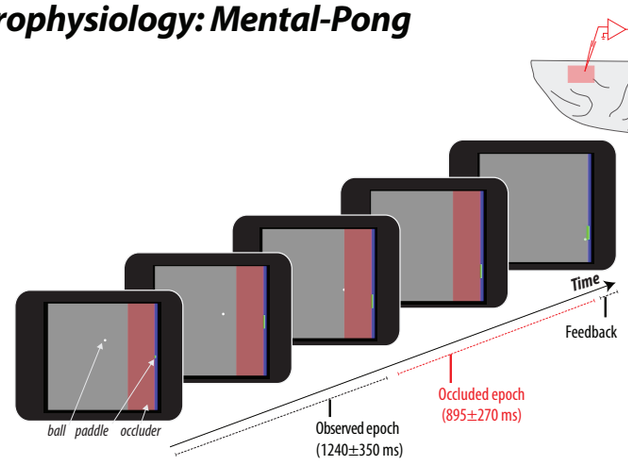
acc. = 0.89

YES

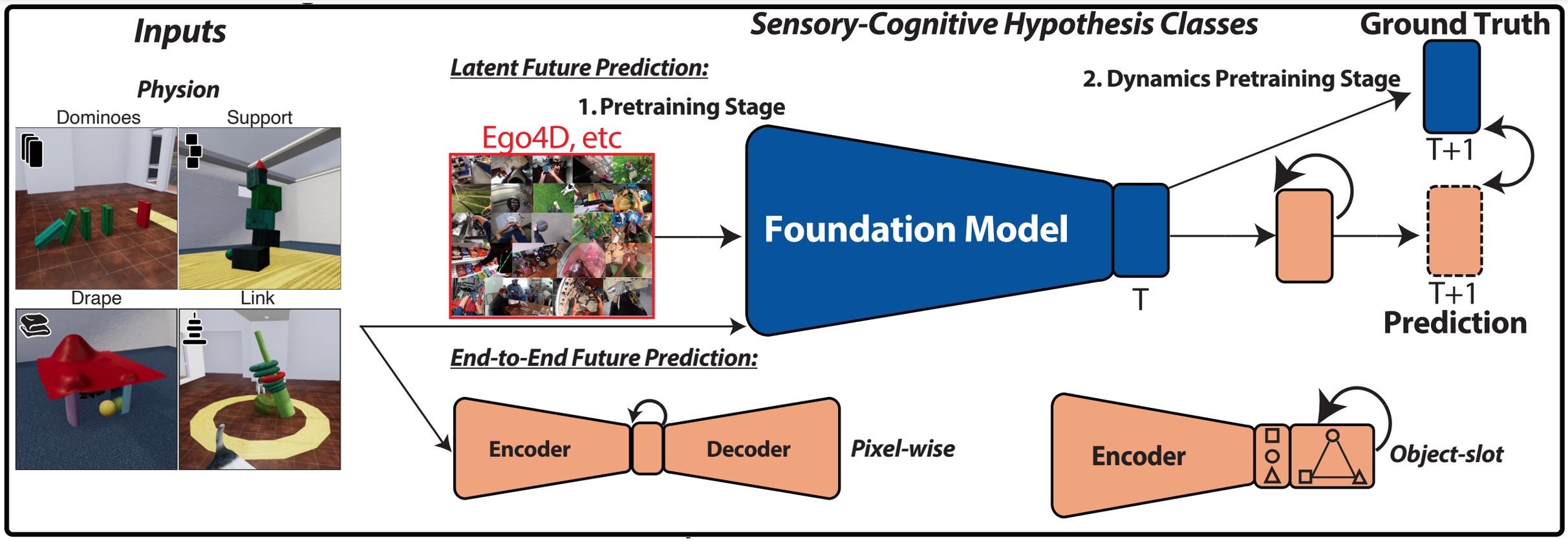
acc. = 0.96

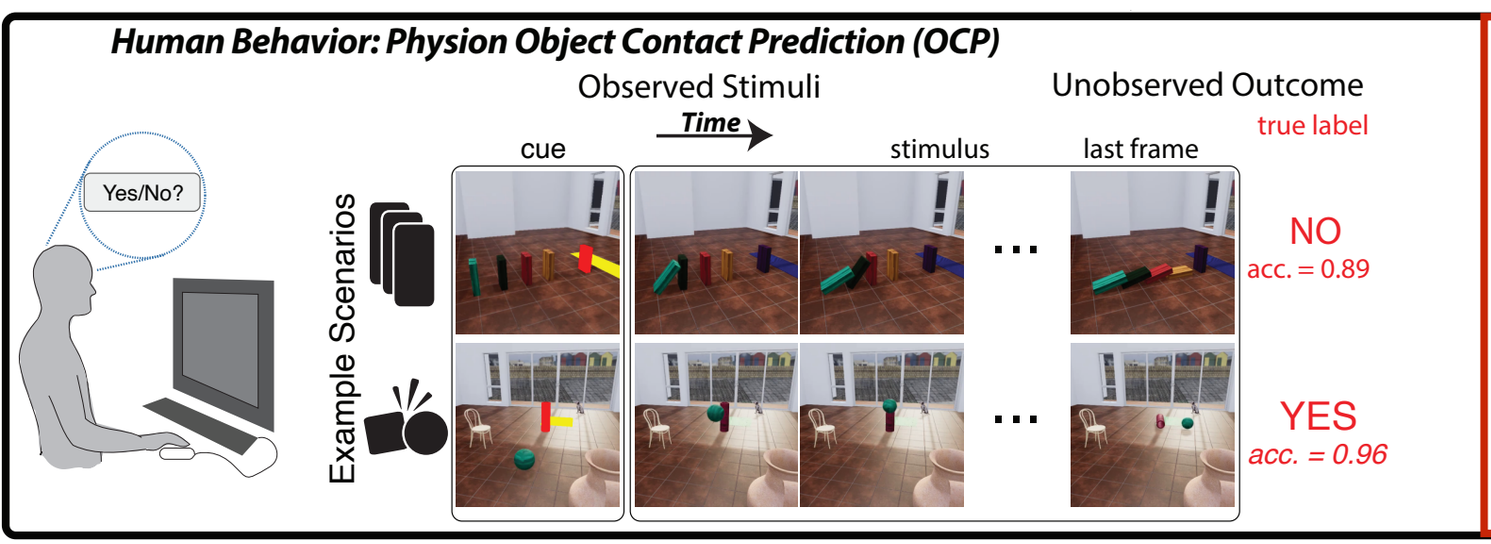
Macaque Neurophysiology: Mental-Pong

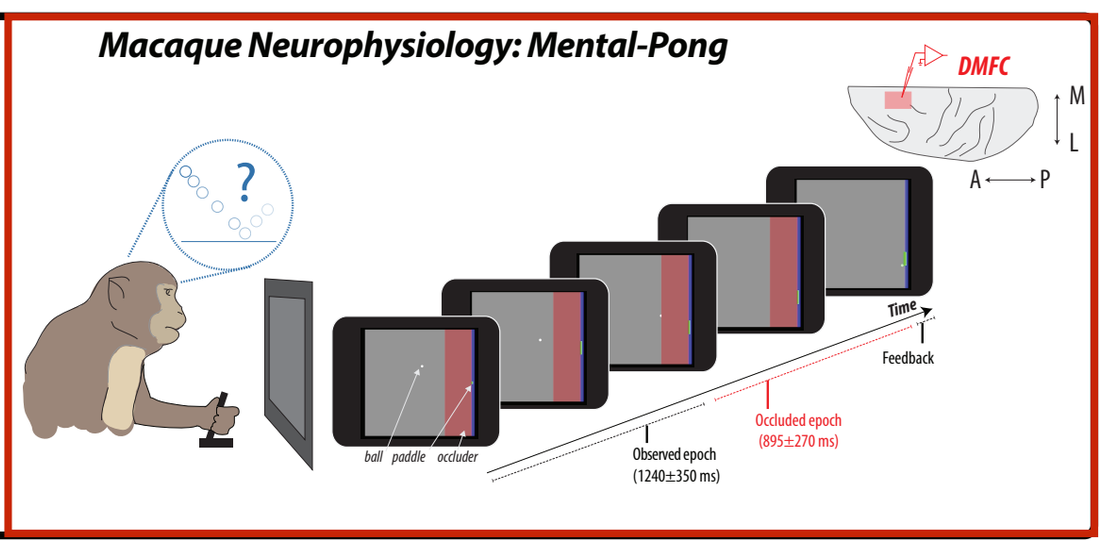




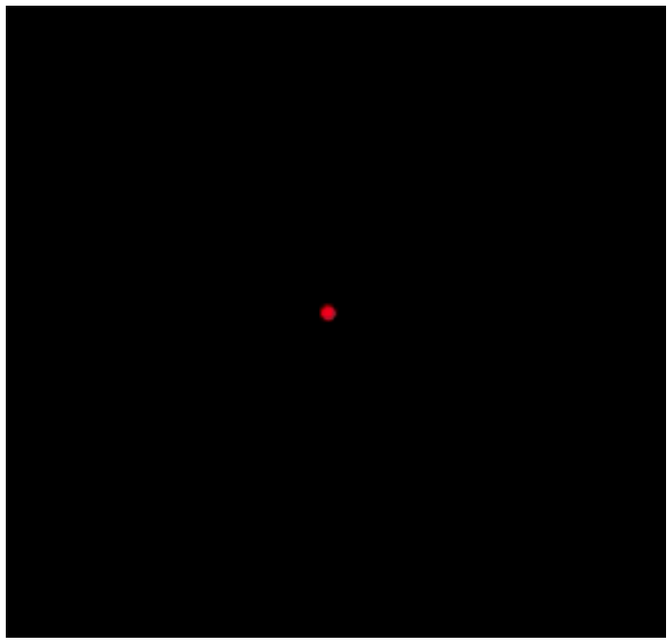
Macaque Neurophysiology: Mental Pong





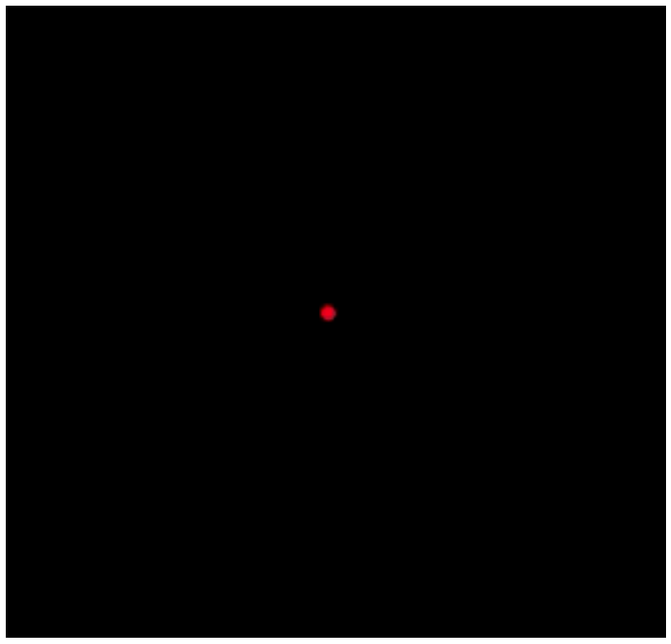


Macaque Neurophysiology: Mental Pong



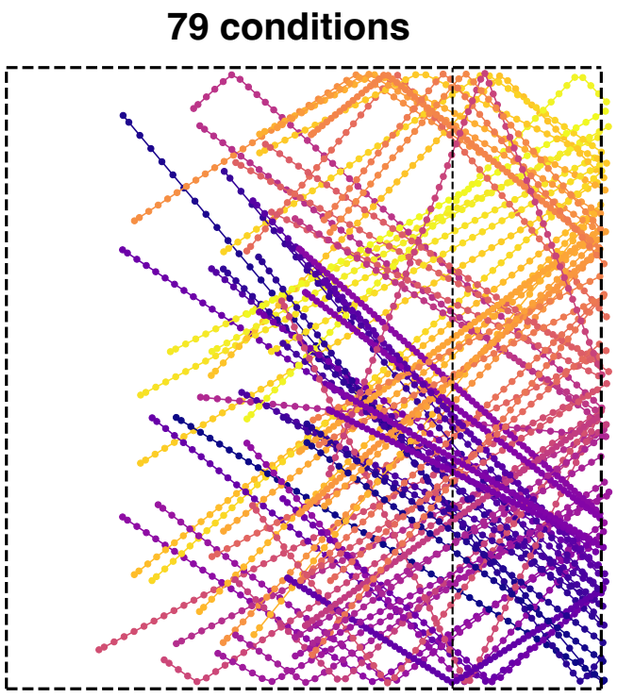
Rishi Rajalingham

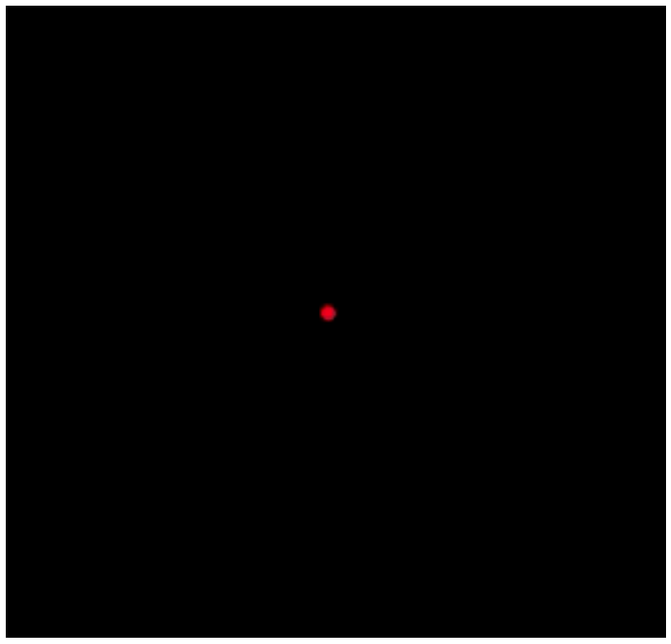
Macaque Neurophysiology: Mental Pong



Rishi Rajalingham

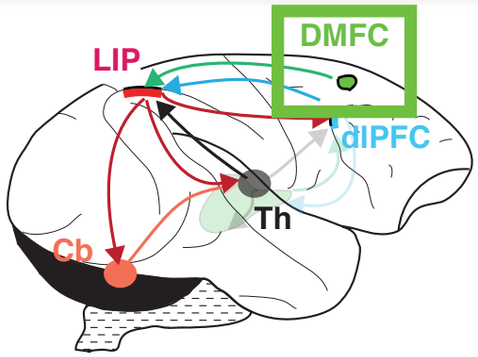
Macaque Neurophysiology: Mental Pong





Rishi Rajalingham

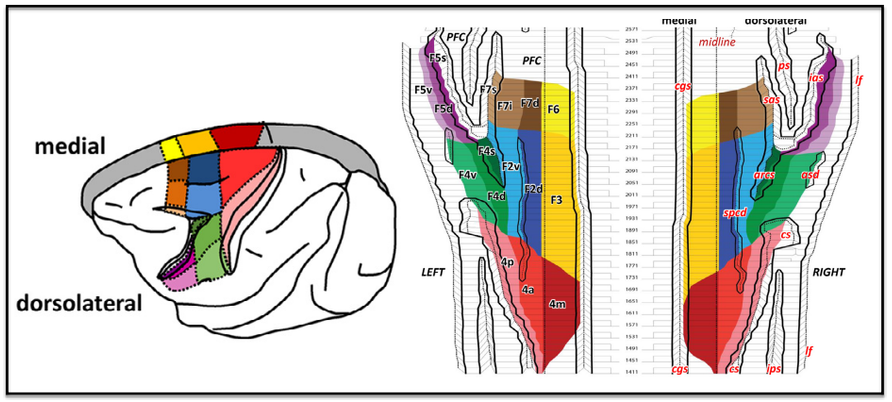
Macaque Neurophysiology: Mental Pong



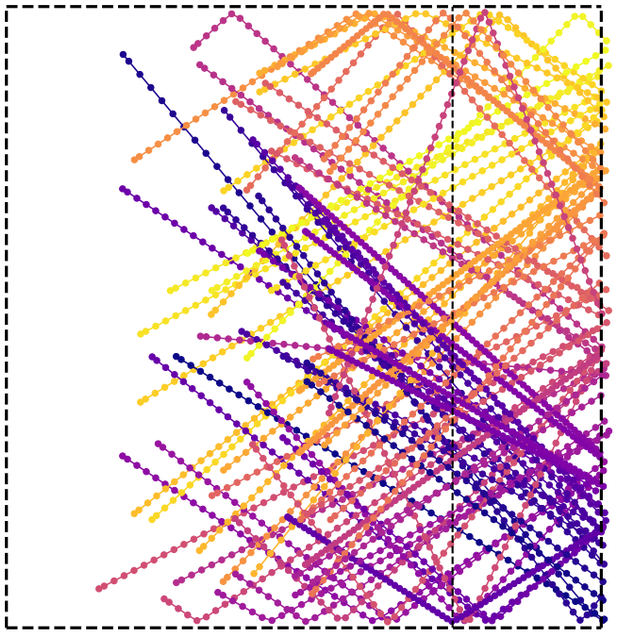
Fronto-Parietal Network

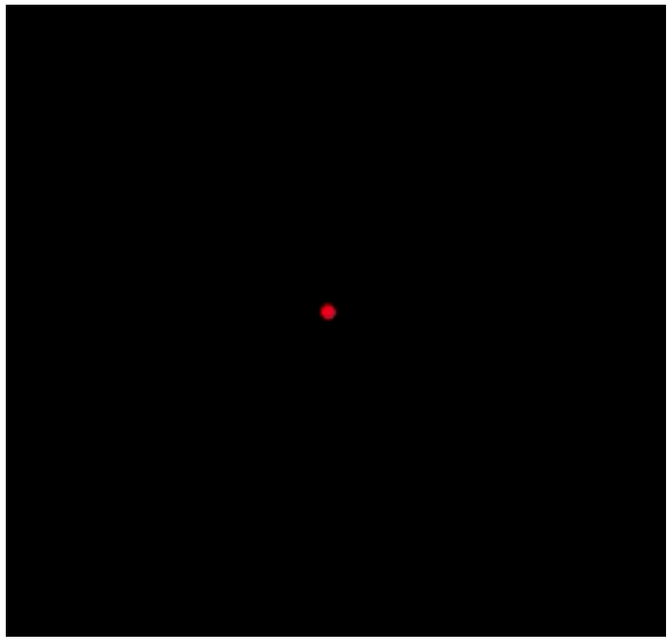
Dorsomedial frontal cortex (DMFC)

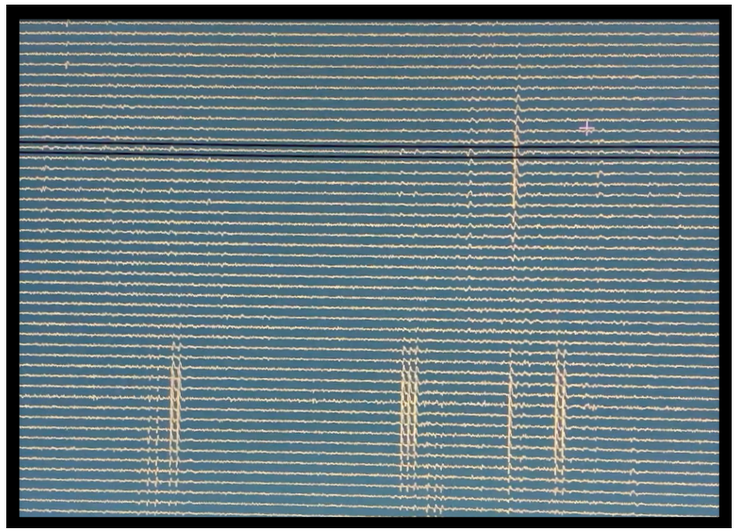
Monkey P *Monkey M*



79 conditions

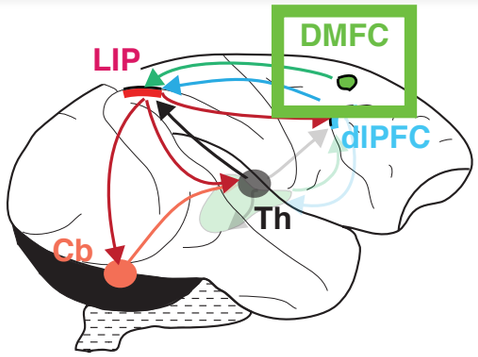






Rishi Rajalingham

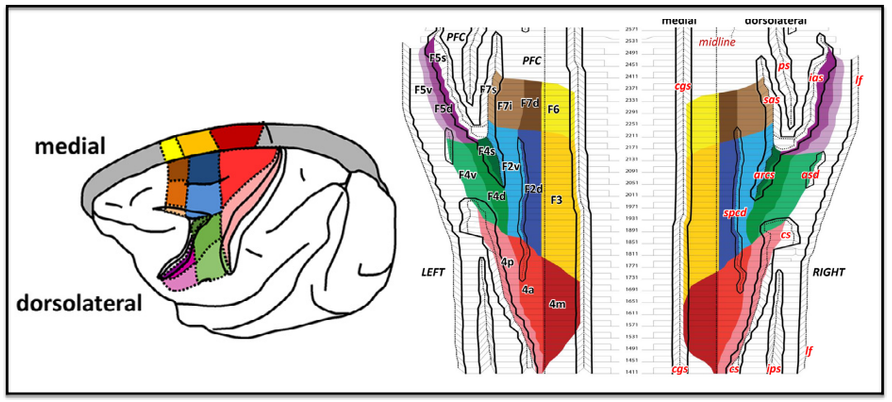
Macaque Neurophysiology: Mental Pong



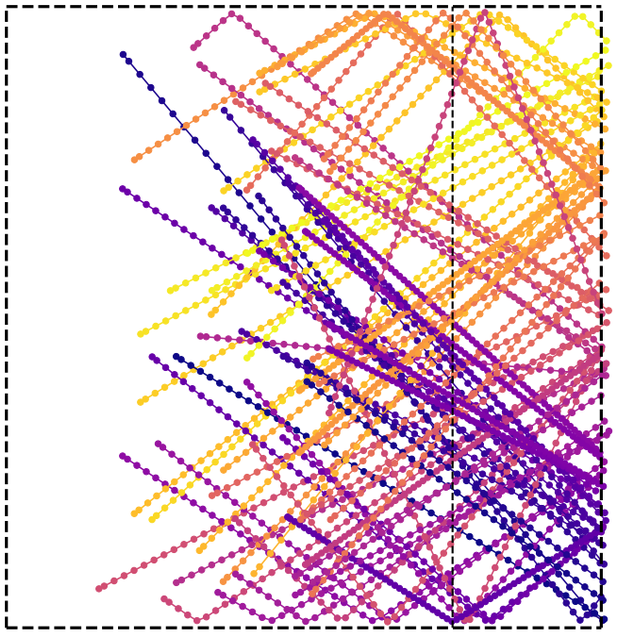
Fronto-Parietal Network

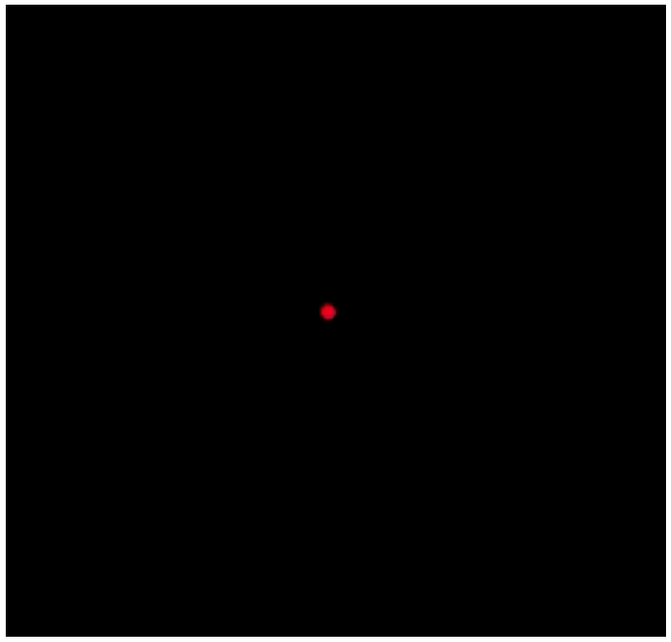
Dorsomedial frontal cortex (DMFC)

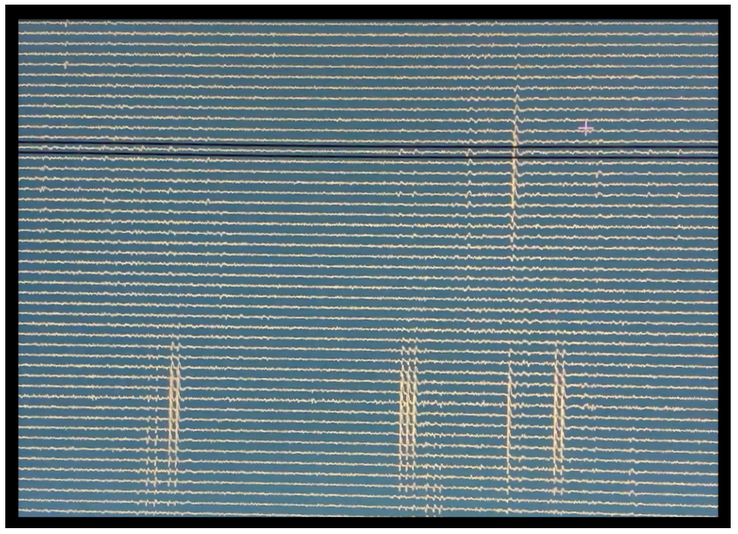
Monkey P *Monkey M*



79 conditions



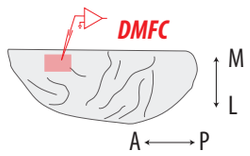


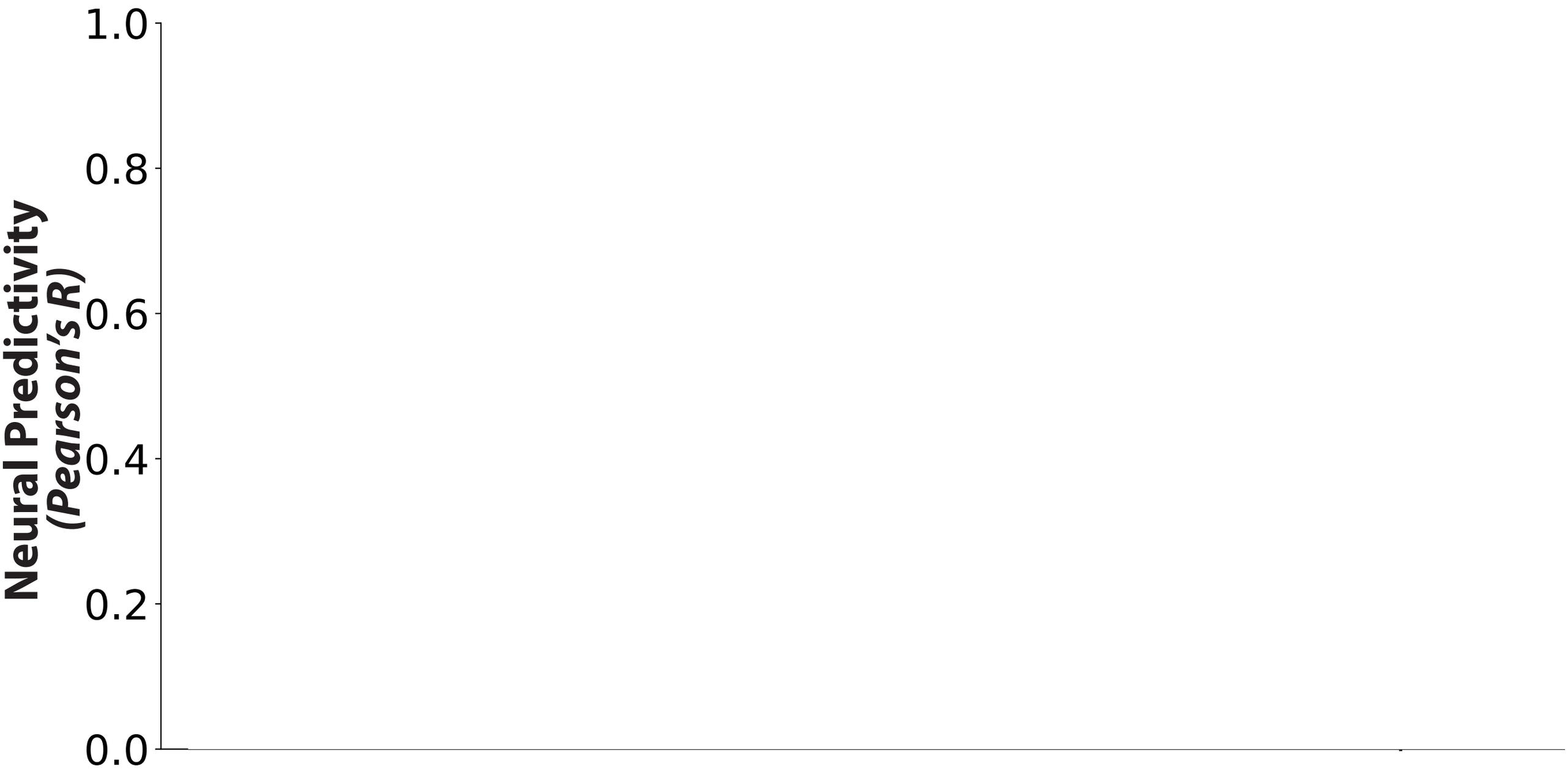


- *Data from two male adult monkeys*
- *79 subsampled M-Pong conditions*
- *64 channel v-probe (monkey P) and 384-channel Neuropixel probe (monkey M)*
- *Total of 1889 stable & reliable neurons recorded from DMFC*

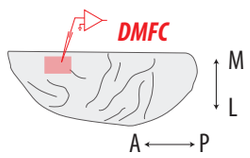
Rishi Rajalingham

Macaque Neurophysiology: Mental Pong





Macaque Neurophysiology: Mental Pong

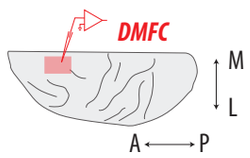


Inter-animal Consistency

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

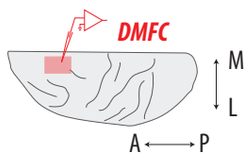
Perfect simulation oracle

Neural Predictivity (*Pearson's R*)

1.0
0.8
0.6
0.4
0.2
0.0

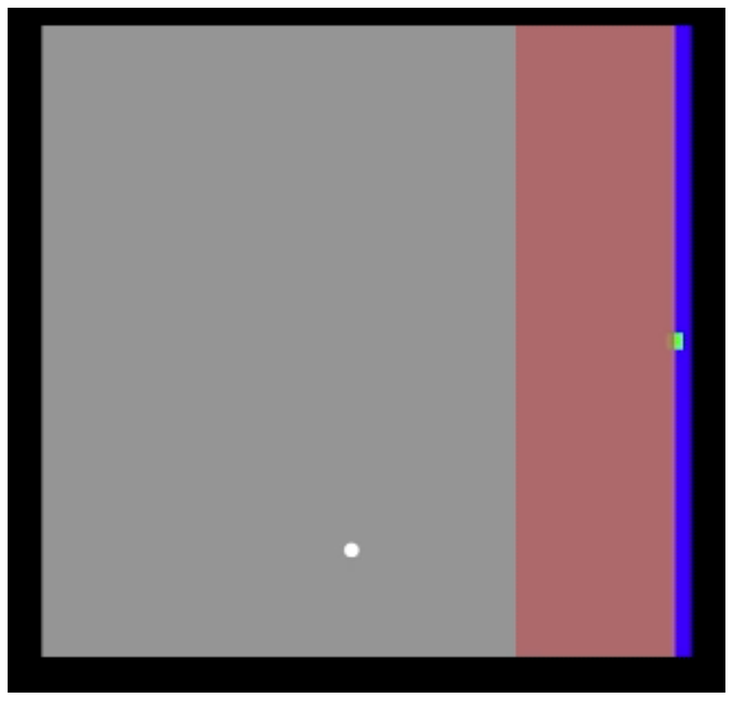
Ball Position + Velocity

Perfect Simulation Oracle Predicts Neural Data Well



Inter-animal Consistency

Perfect simulation oracle



Ball Position + Velocity

Neural Predictivity (*Pearson's R*)

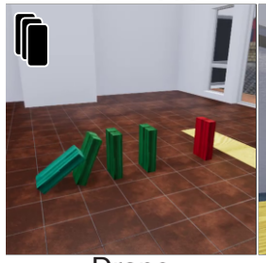
1.0
0.8
0.6
0.4
0.2
0.0

Functional Constraint Hypotheses

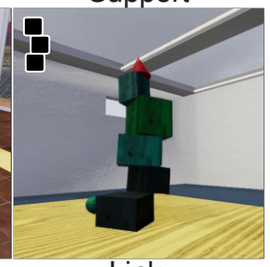
Inputs

Physion

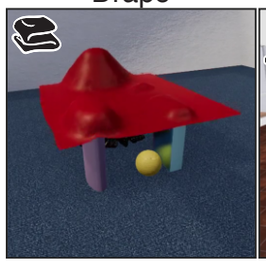
Dominoes



Support



Drape



Link

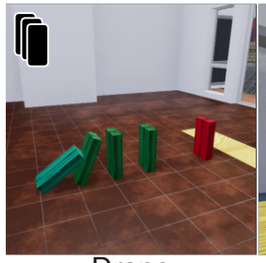
Sensory-Cognitive Hypothesis Classes

Hypothesis Class 1: Pixel-wise Future Prediction

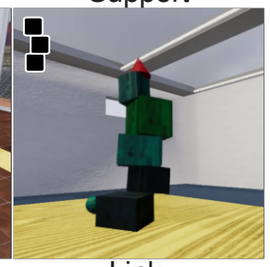
Inputs

Physion

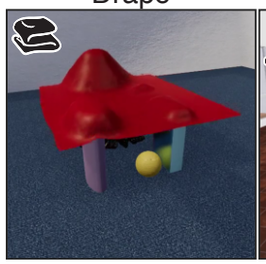
Dominoes



Support



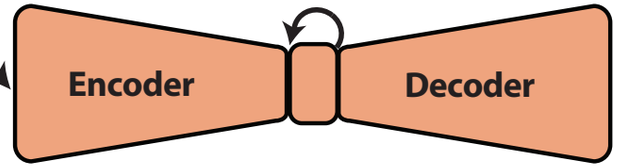
Drape



Link

Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



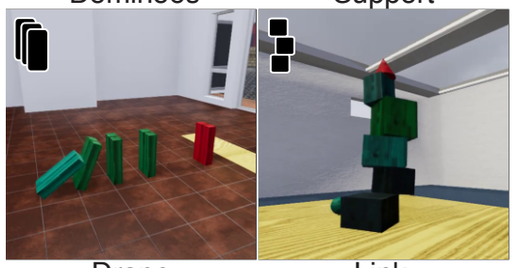
Pixel-wise

Hypothesis Class 1: Pixel-wise Future Prediction

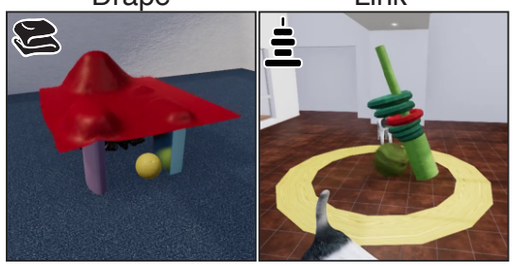
Inputs

Physion

Dominoes | Support

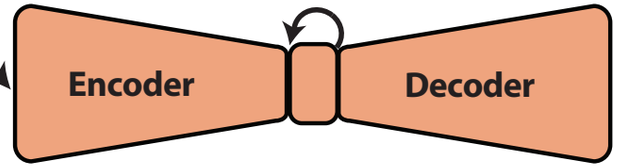


Drape | Link



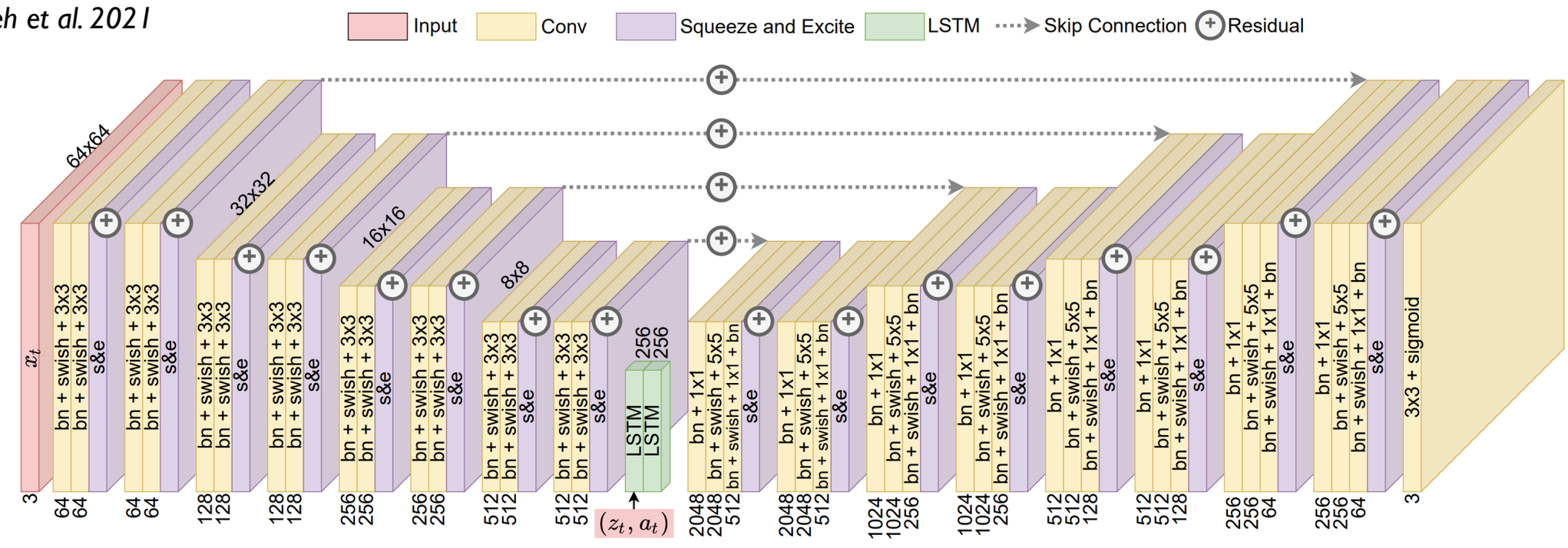
Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021

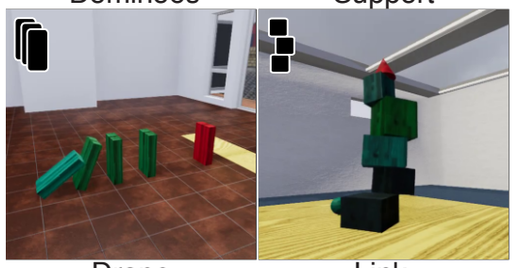


Hypothesis Class 1: Pixel-wise Future Prediction

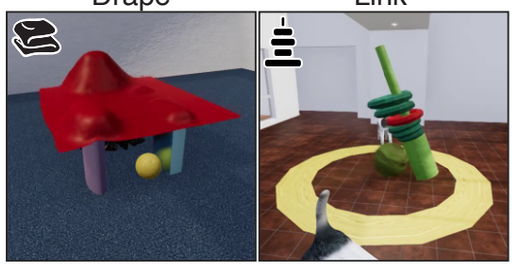
Inputs

Physion

Dominoes | Support

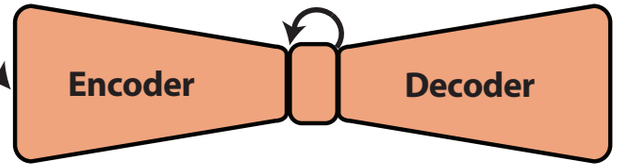


Drape | Link



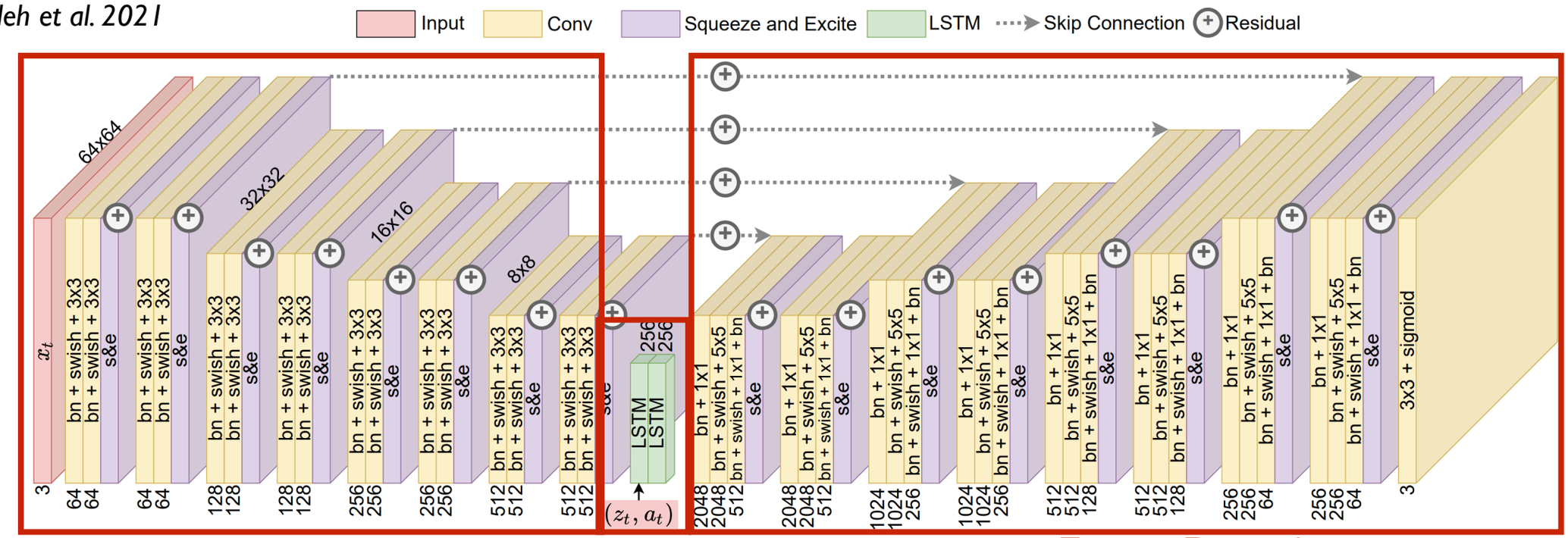
Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



Pixel-wise

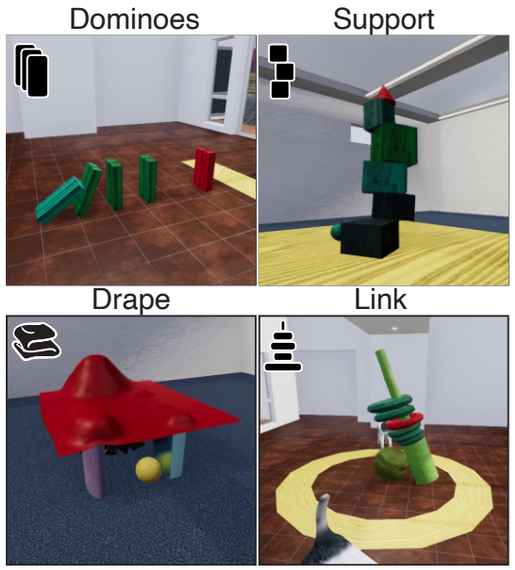
Babaeizadeh et al. 2021



Hypothesis Class 1: Pixel-wise Future Prediction

Inputs

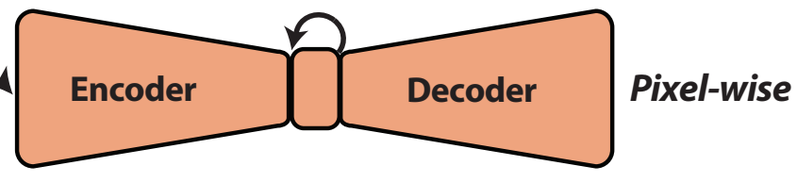
Physion



Sensory-Cognitive Hypothesis Classes

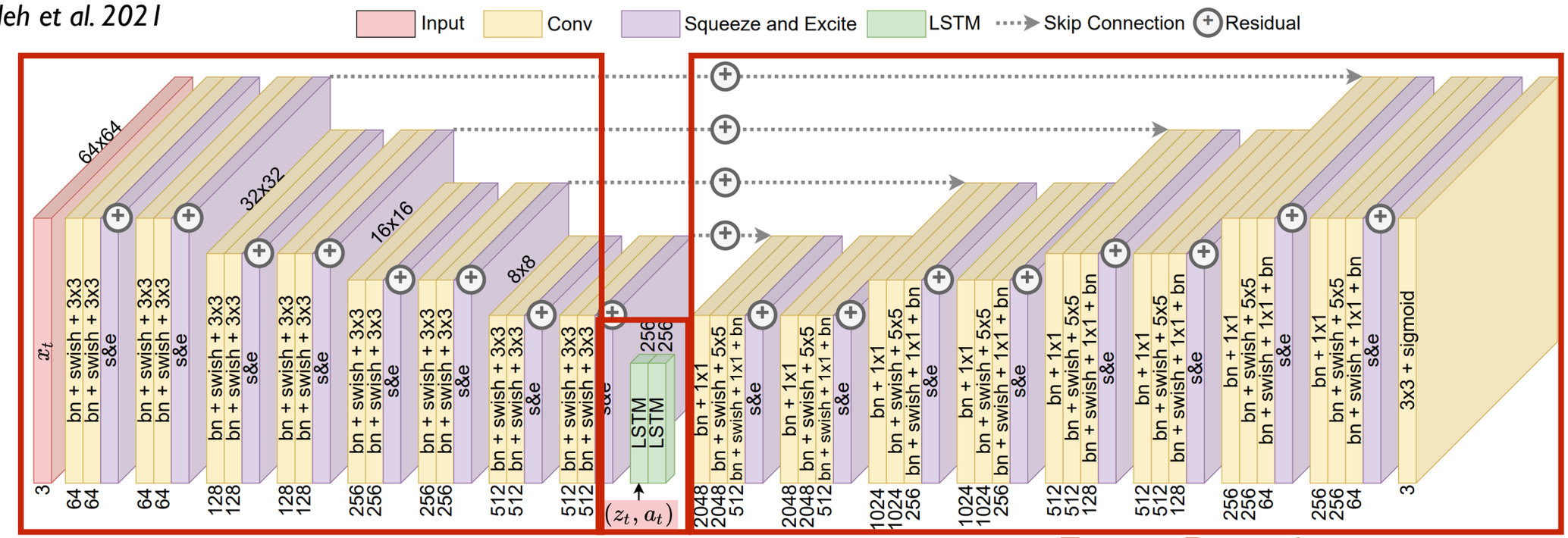
$$\mathcal{L} = \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\|_2^2 + \beta D_{KL}(\mathcal{N}(\mu, \sigma)\|\mathcal{N}(0, \mathbf{I}))$$

End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021



Visual Encoder ("Sensory")

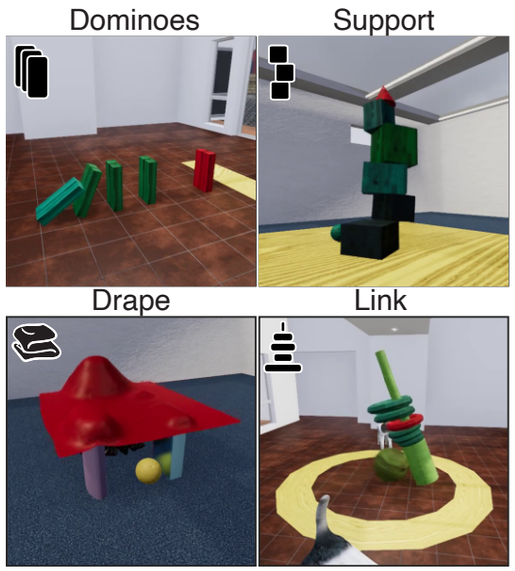
Dynamics Predictor ("Cognitive")

Frame Decoder ("Objective/Behavior")

Hypothesis Class 1: Pixel-wise Future Prediction

Inputs

Physion

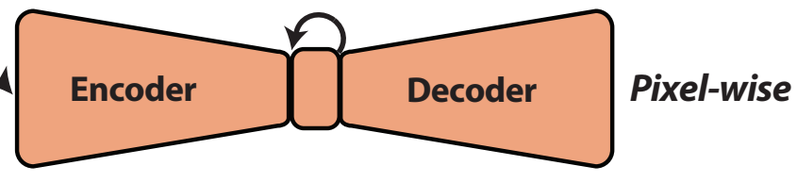


Sensory-Cognitive Hypothesis Classes

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

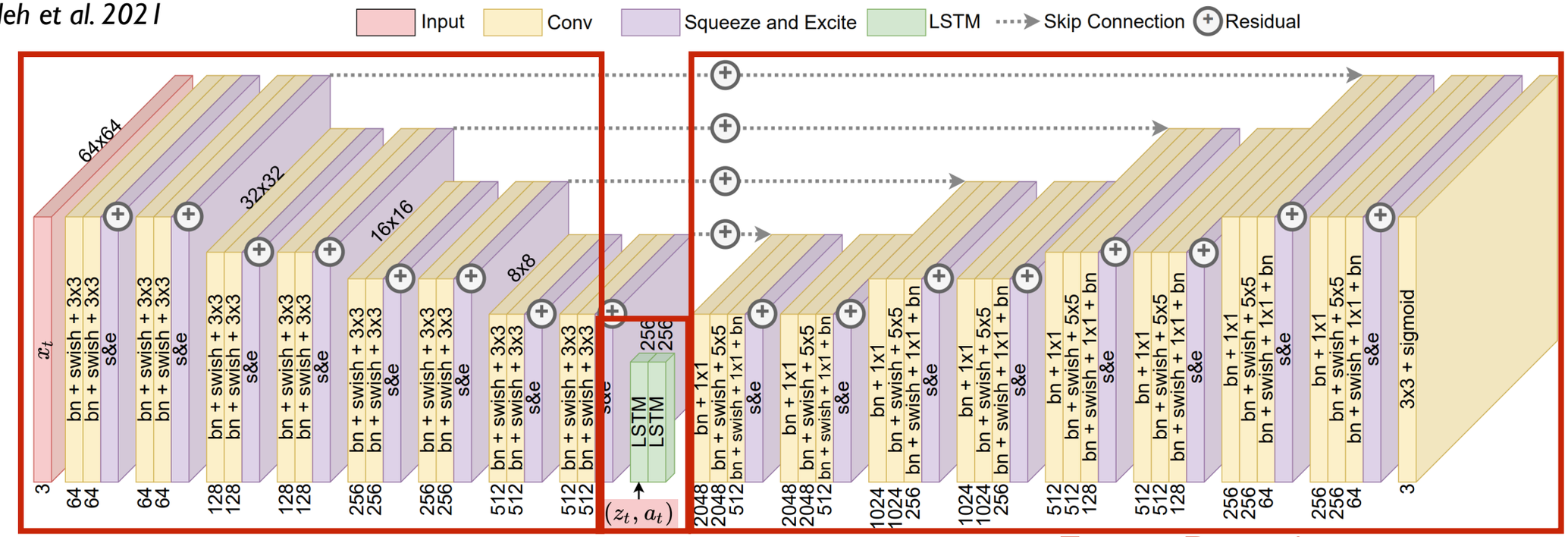
$$\mathcal{L} = \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\|_2^2 + \beta D_{KL}(\mathcal{N}(\mu, \sigma)\|\mathcal{N}(0, \mathbf{I}))$$

End-to-End Future Prediction:

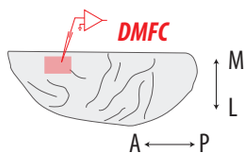


Pixel-wise

Babaeizadeh et al. 2021



Physical Simulation Oracles Predict Neural Data Well



Inter-animal Consistency

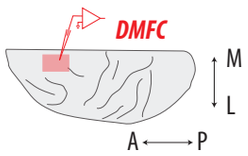
Perfect simulation oracle

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

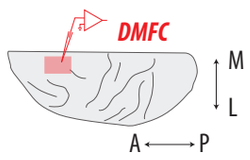
64x64
64x64+ RandAugment

SVG
FitVid

End-to-End

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

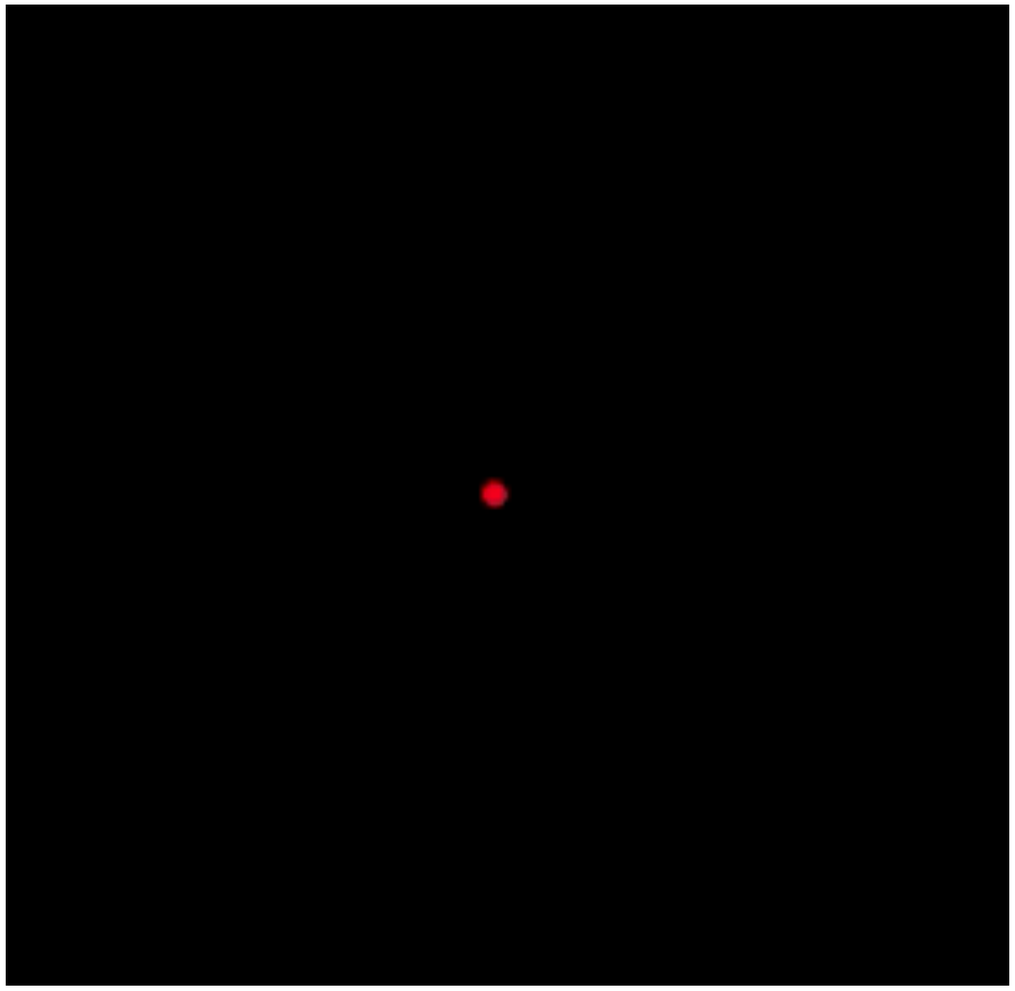
Ball Position + Velocity

…and they struggle to generalize to Pong

Input Frames

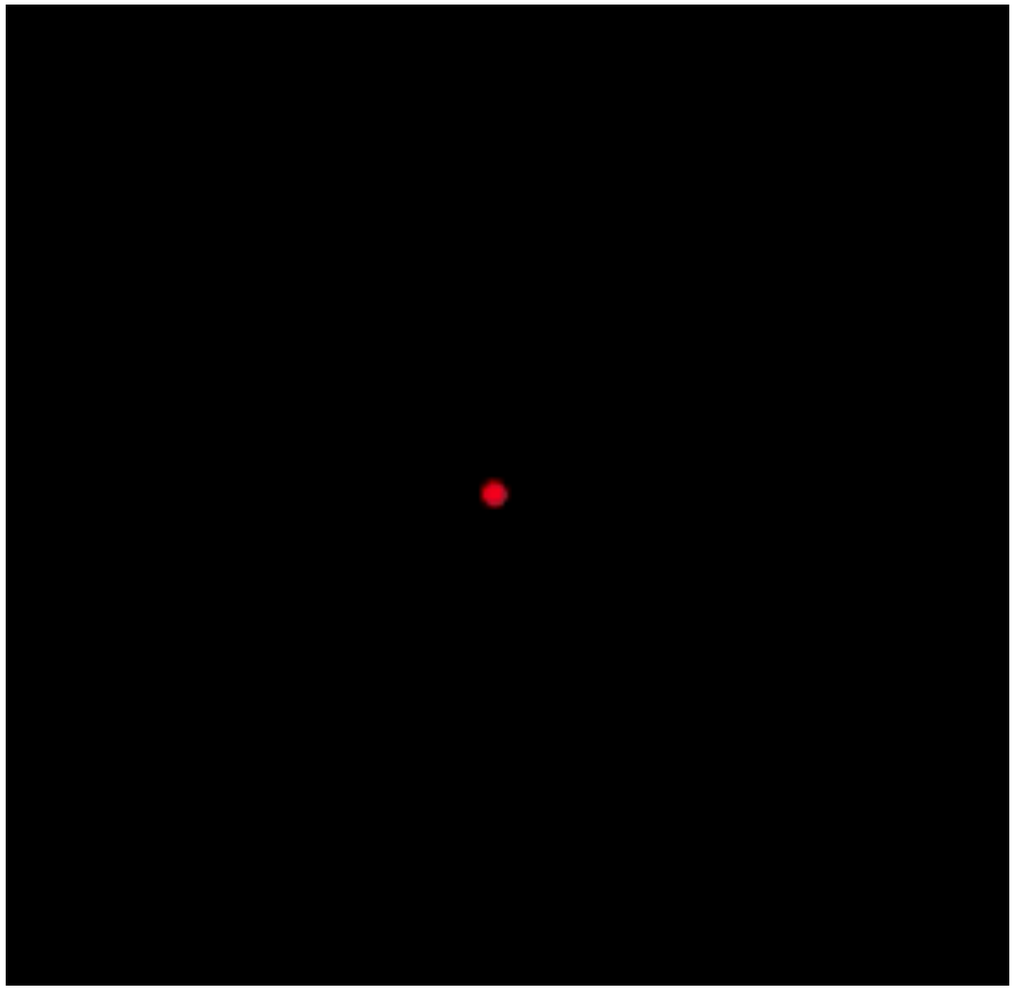
…and they struggle to generalize to Pong

Input Frames



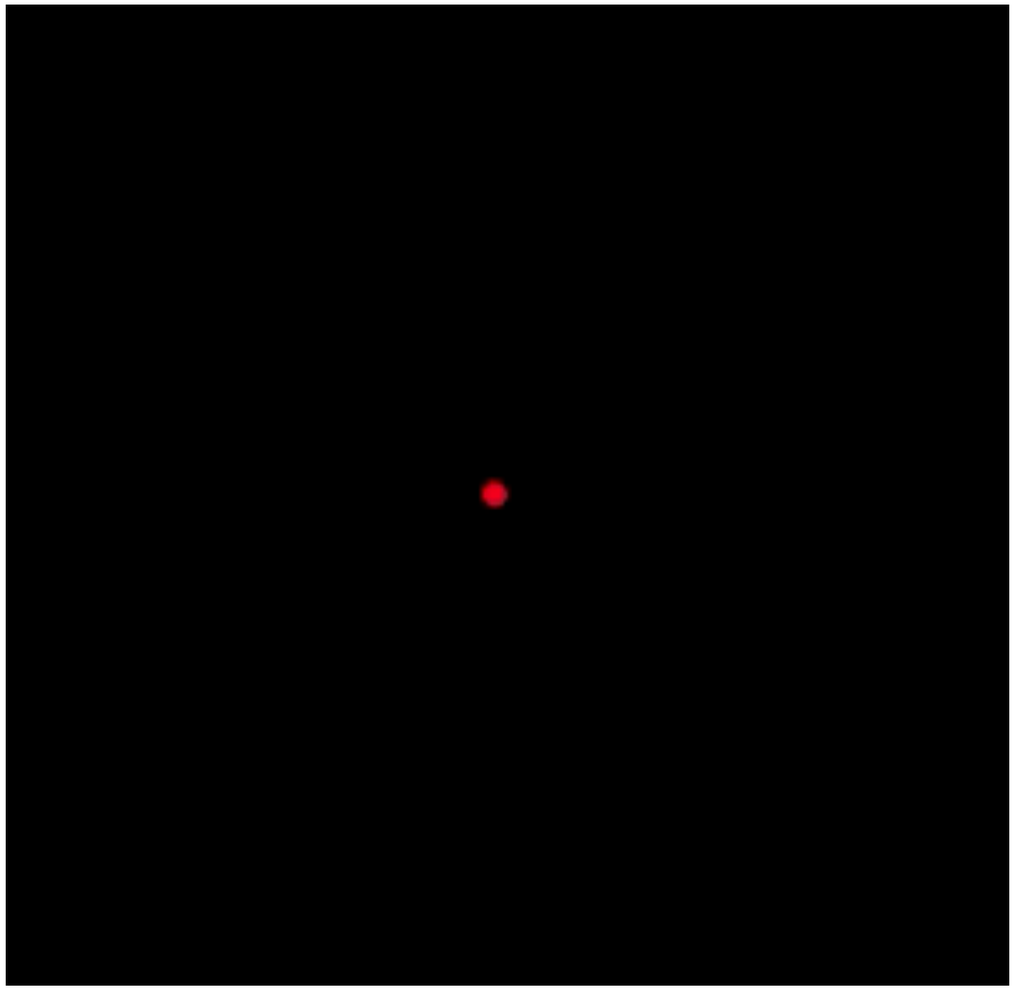
…and they struggle to generalize to Pong

Input Frames



…and they struggle to generalize to Pong

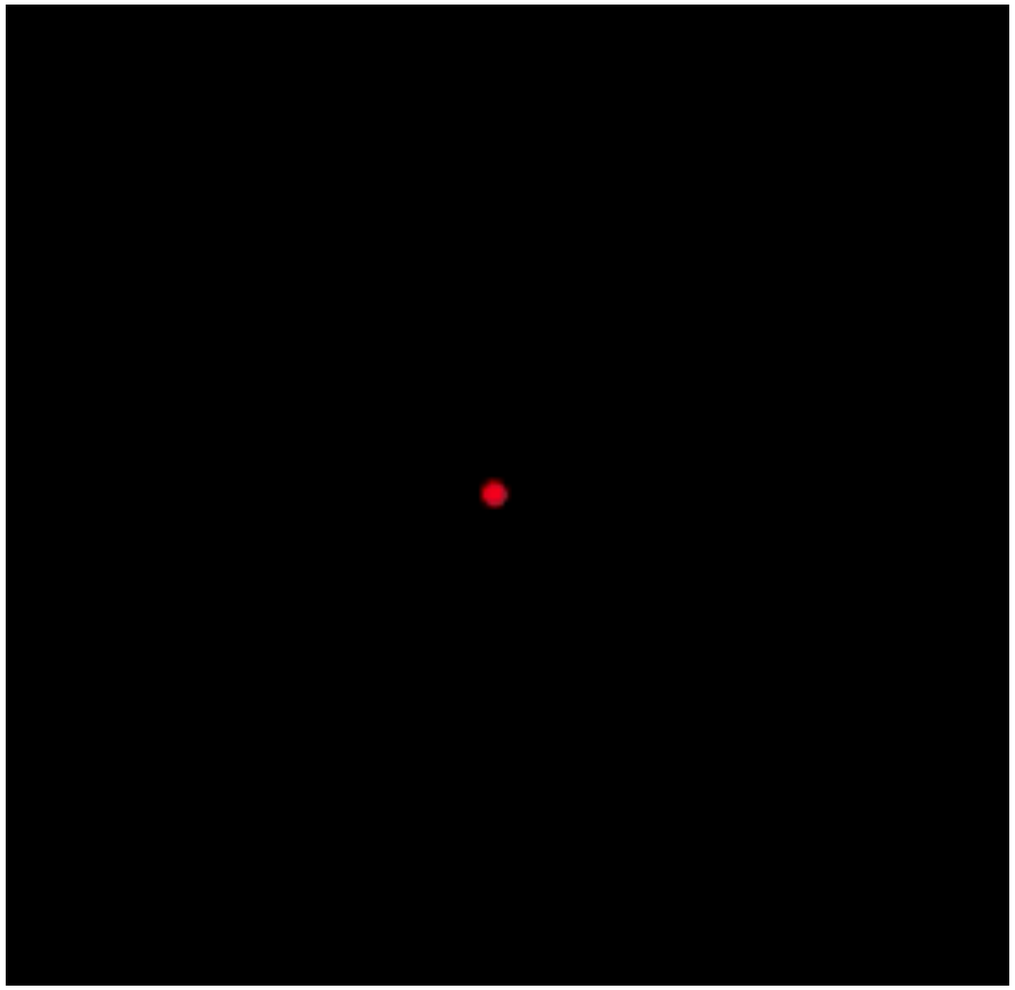
Input Frames



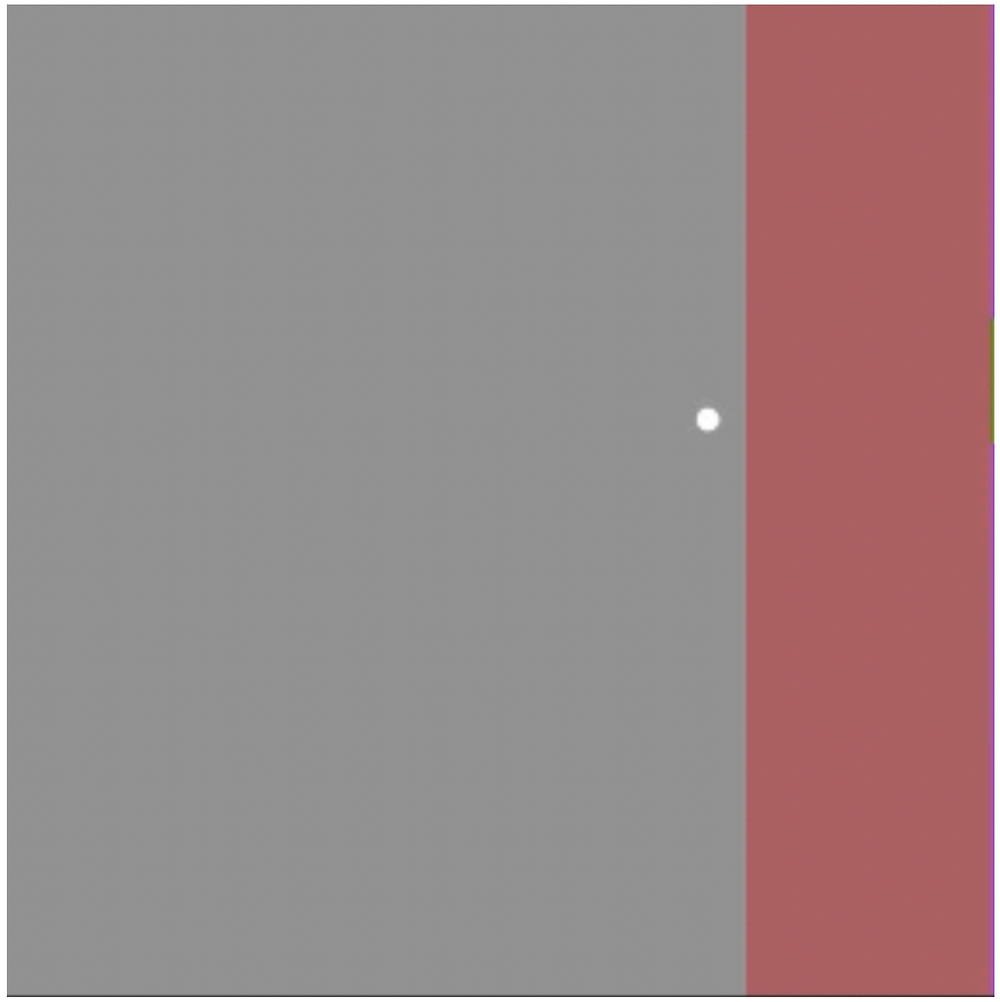
Predicted Frames

…and they struggle to generalize to Pong

Input Frames

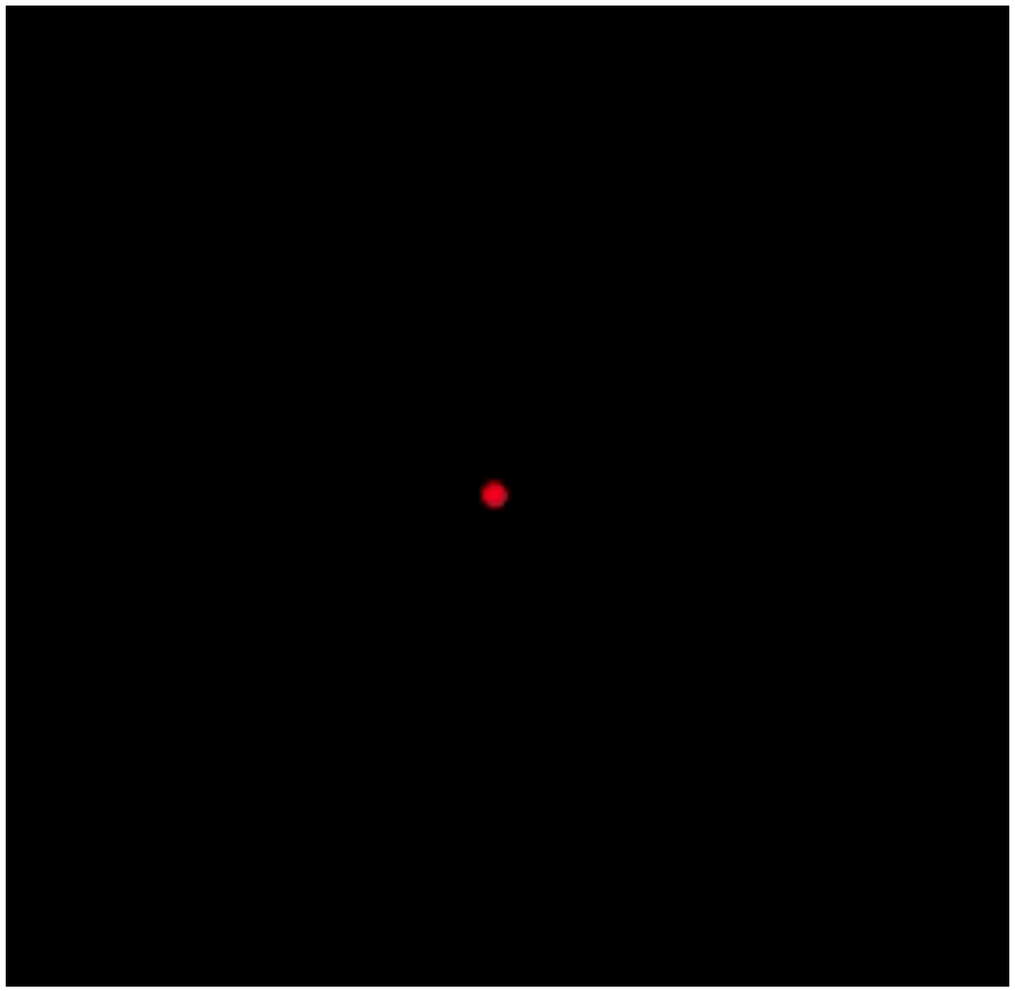


Predicted Frames

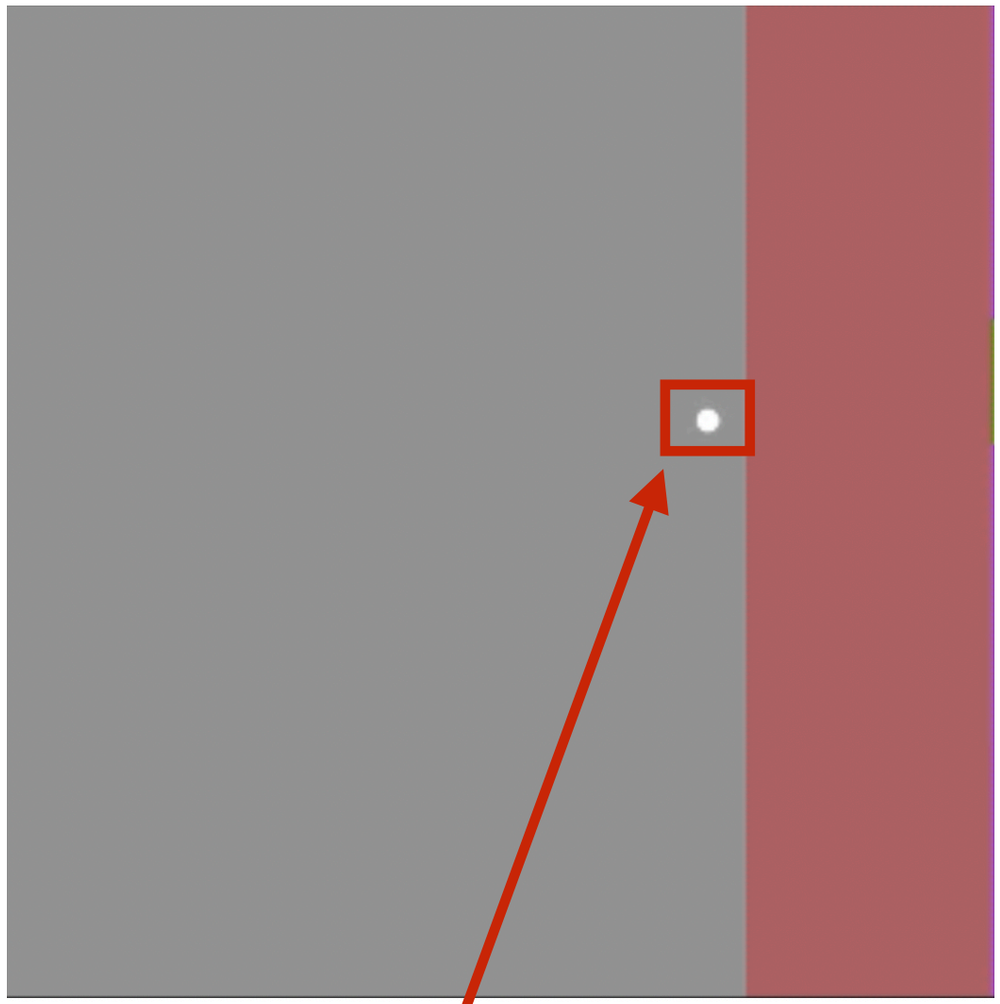


…and they struggle to generalize to Pong

Input Frames

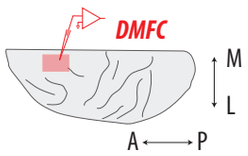


Predicted Frames



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

Pixel-wise Future Prediction Poorly Predicts Neurons



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

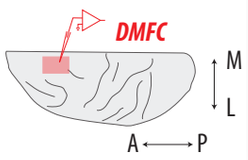
SVG
FitVid

End-to-End

Ball Position + Velocity

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

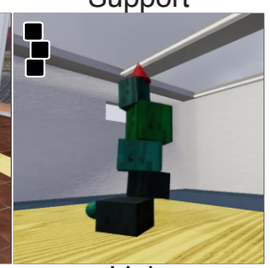
Hypothesis Class 2: Object Slots

Inputs

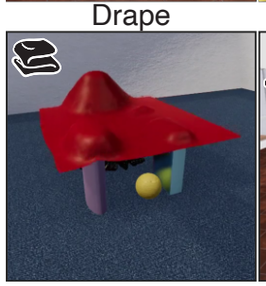
Physion

Dominoes

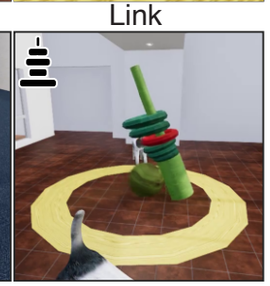
Support



Drape

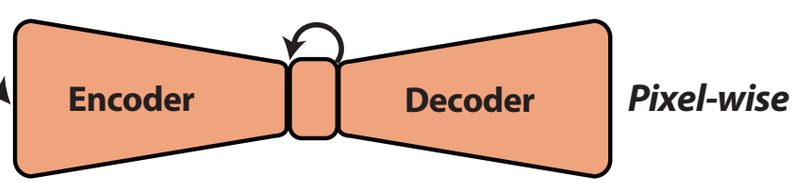


Link

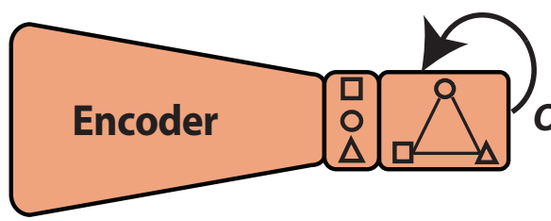


Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:

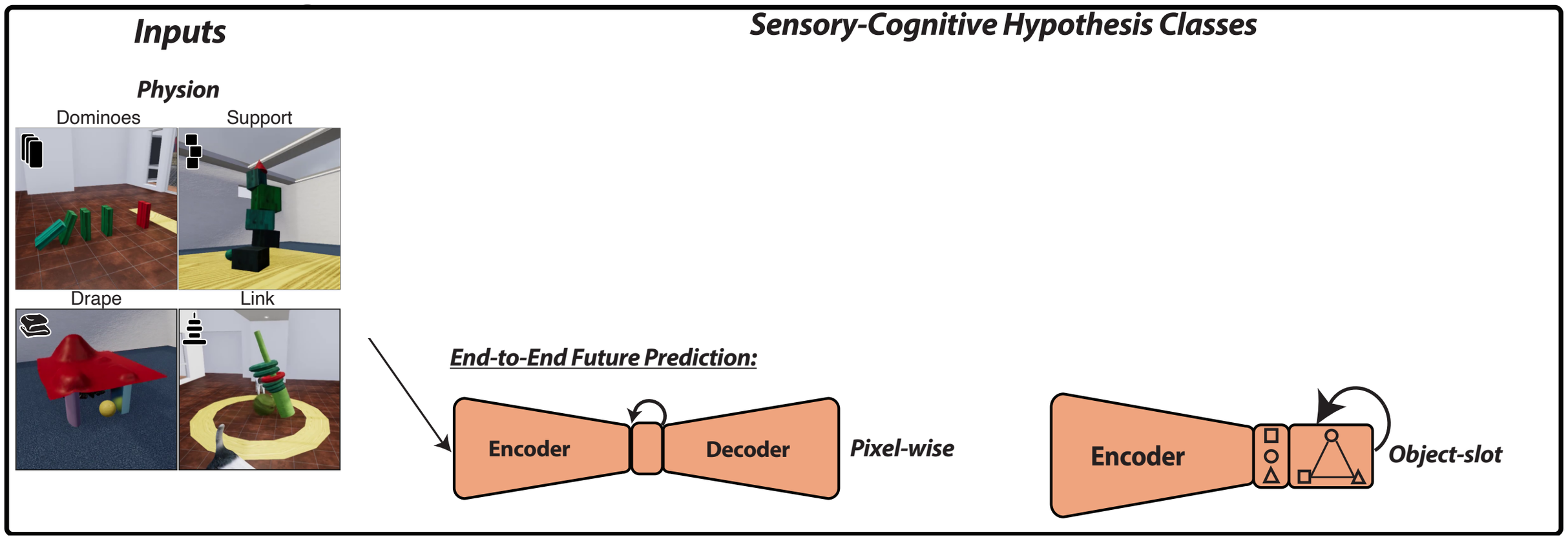


Pixel-wise



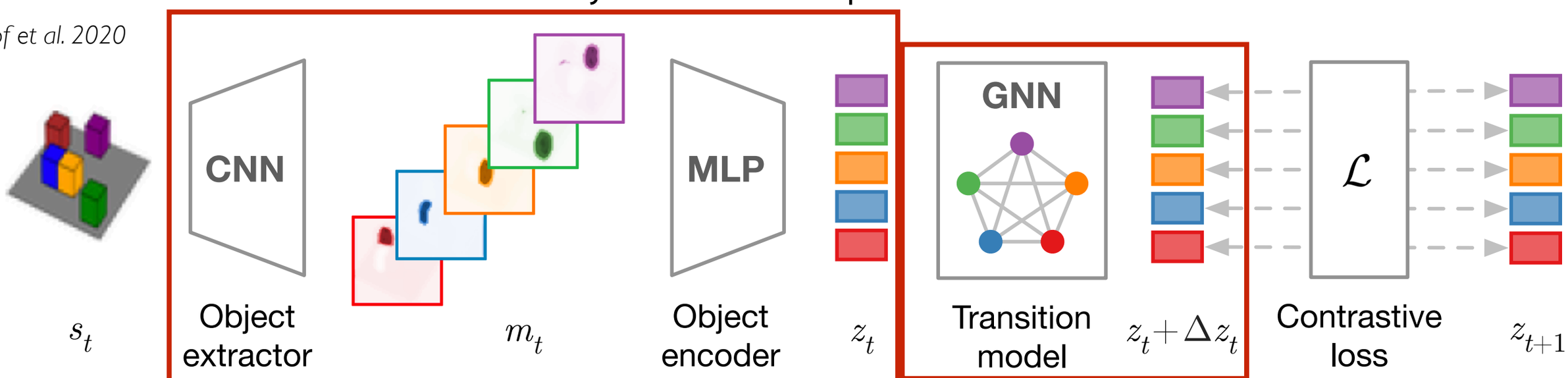
Object-slot

Hypothesis Class 2: Object Slots



Predicts at the level of object slot representations and their relations

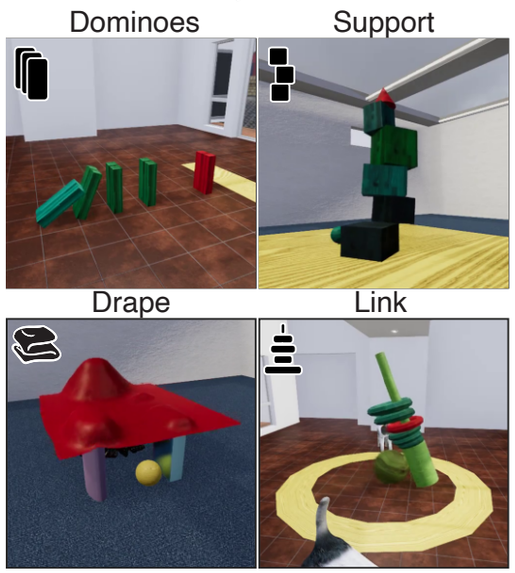
Kipf et al. 2020



Hypothesis Class 2: Object Slots

Inputs

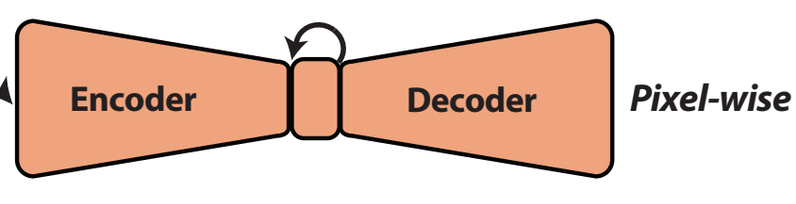
Physion



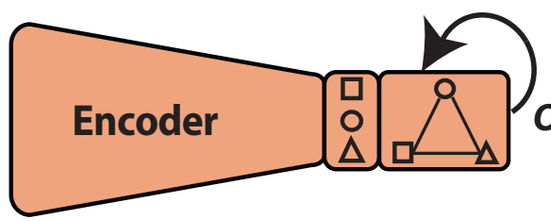
Sensory-Cognitive Hypothesis Classes

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

End-to-End Future Prediction:



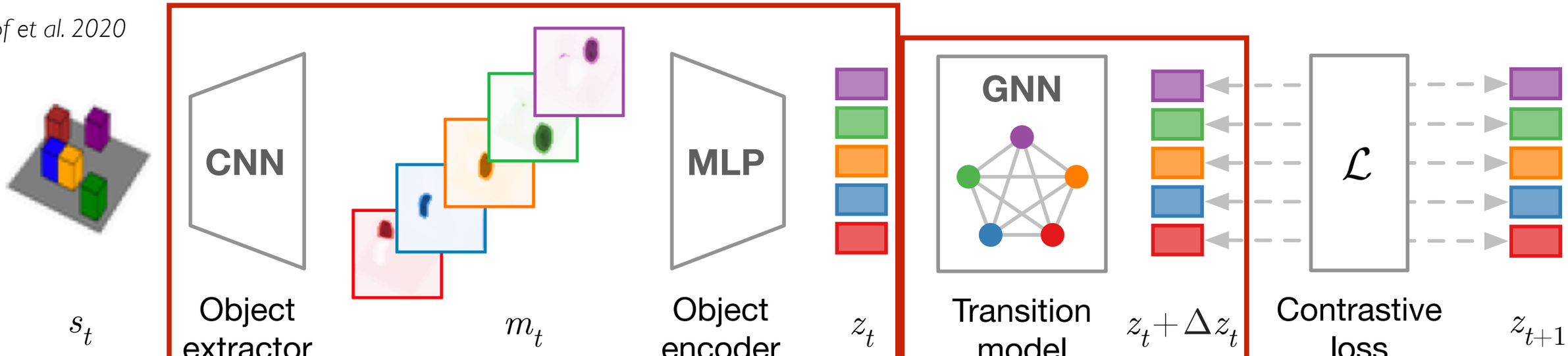
Pixel-wise



Object-slot

Predicts at the level of object slot representations and their relations

Kipf et al. 2020



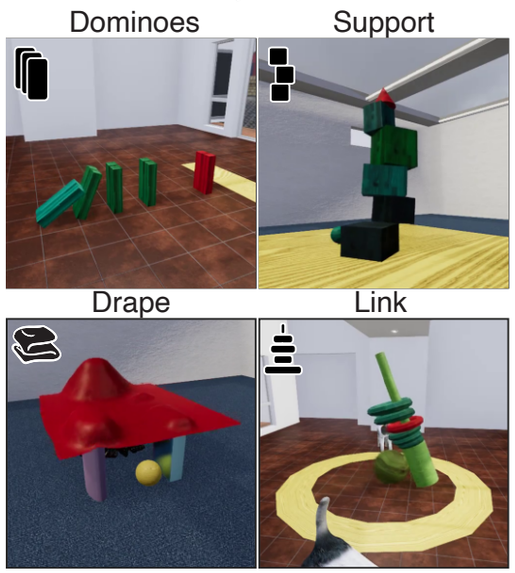
Visual Encoder ("Sensory")

Dynamics Predictor ("Cognitive")

Hypothesis Class 2: Object Slots

Inputs

Physion

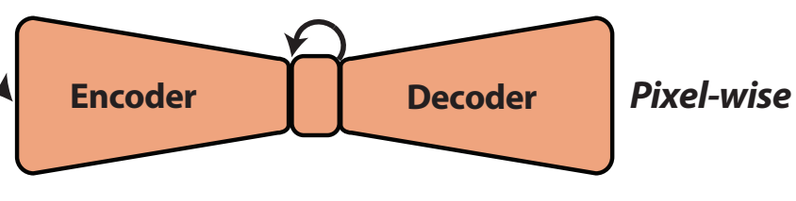


Sensory-Cognitive Hypothesis Classes

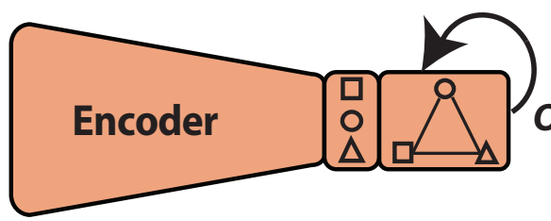
$$\mathcal{L} = \overbrace{\mathcal{H}}^{\text{attract}} + \max(0, 1 - \overbrace{\widetilde{\mathcal{H}}}^{\text{repel}})$$

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

End-to-End Future Prediction:



Pixel-wise

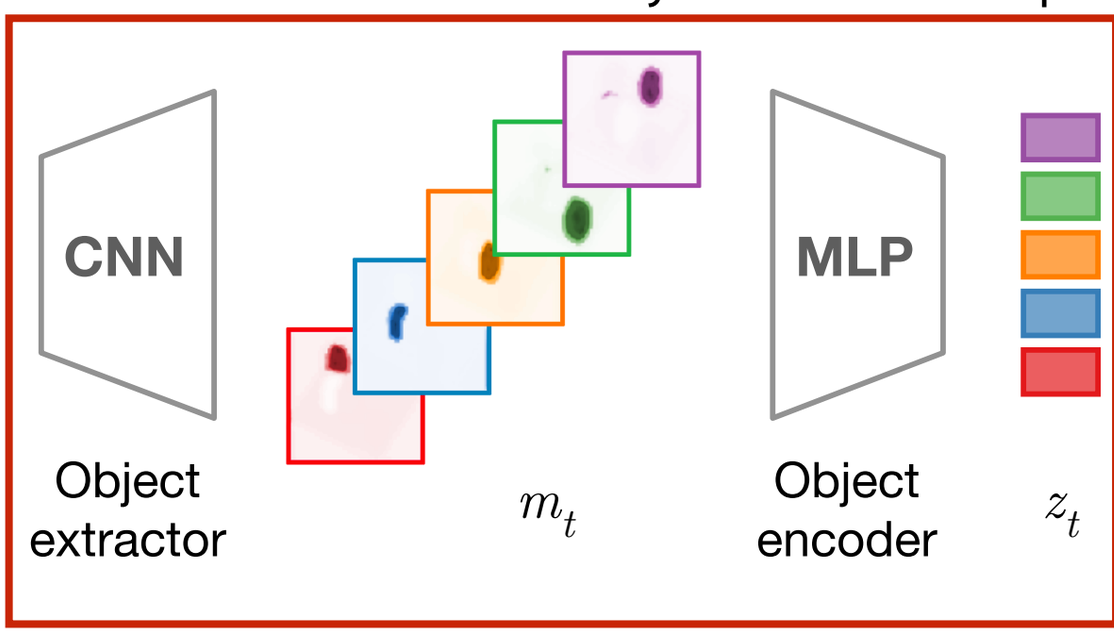


Object-slot

Predicts at the level of object slot representations and their relations

Kipf et al. 2020

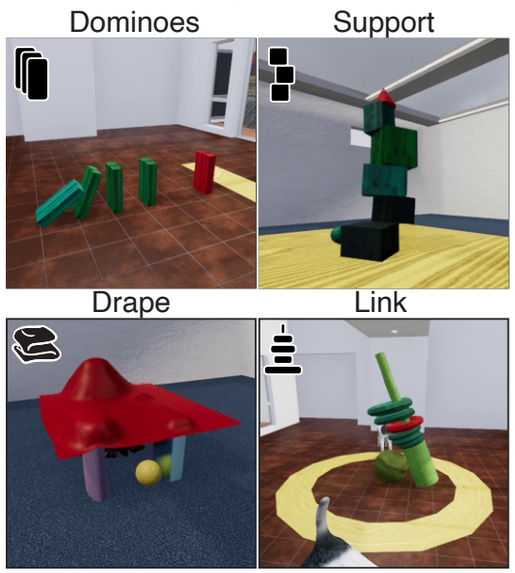
s_t



Hypothesis Class 2: Object Slots

Inputs

Physion

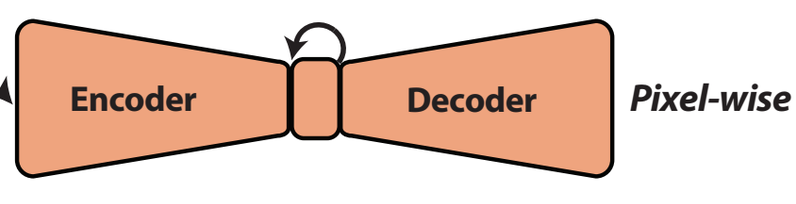


Sensory-Cognitive Hypothesis Classes

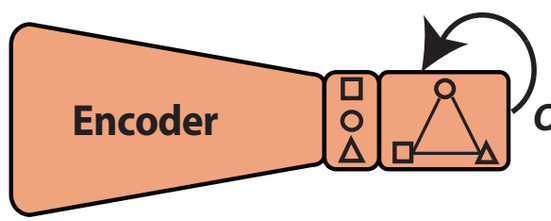
$$\mathcal{L} = \overbrace{\mathcal{H}}^{\text{attract}} + \max(0, 1 - \overbrace{\widetilde{\mathcal{H}}}^{\text{repel}})$$

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

End-to-End Future Prediction:



Pixel-wise

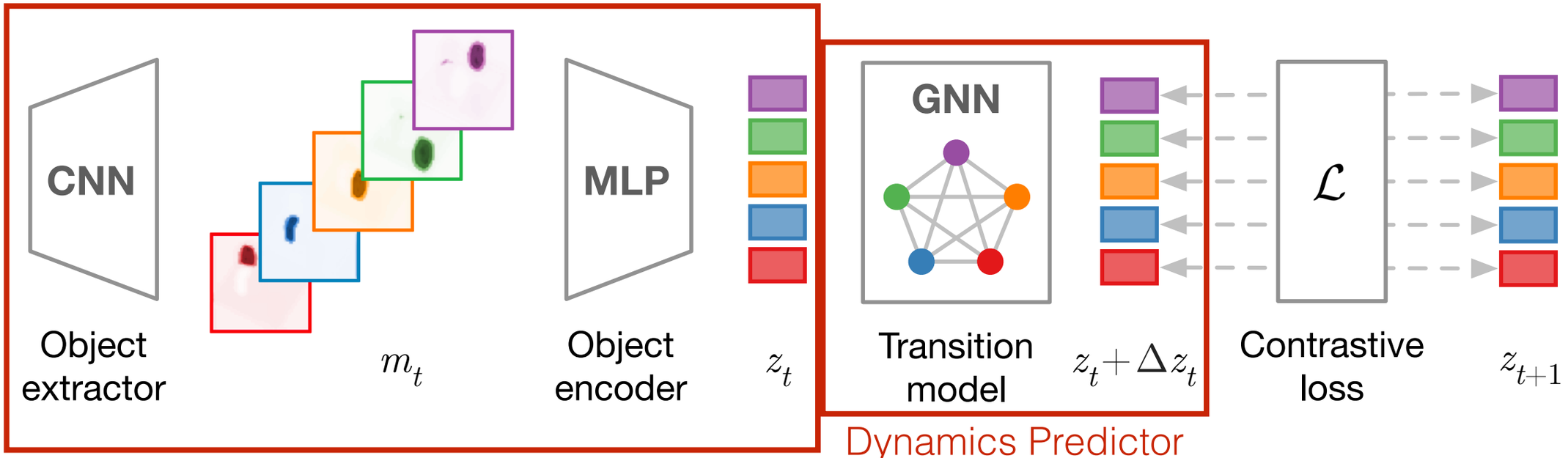


Object-slot

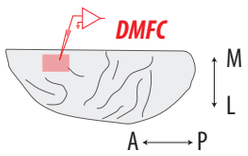
Predicts at the level of object slot representations and their relations

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

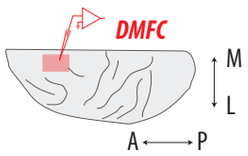
64x64
64x64+ RandAugment

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

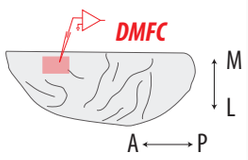
SVG
FitVid
C-SWM

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

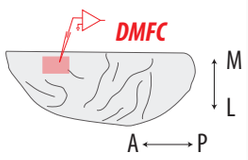
SVG
FitVid
C-SWM

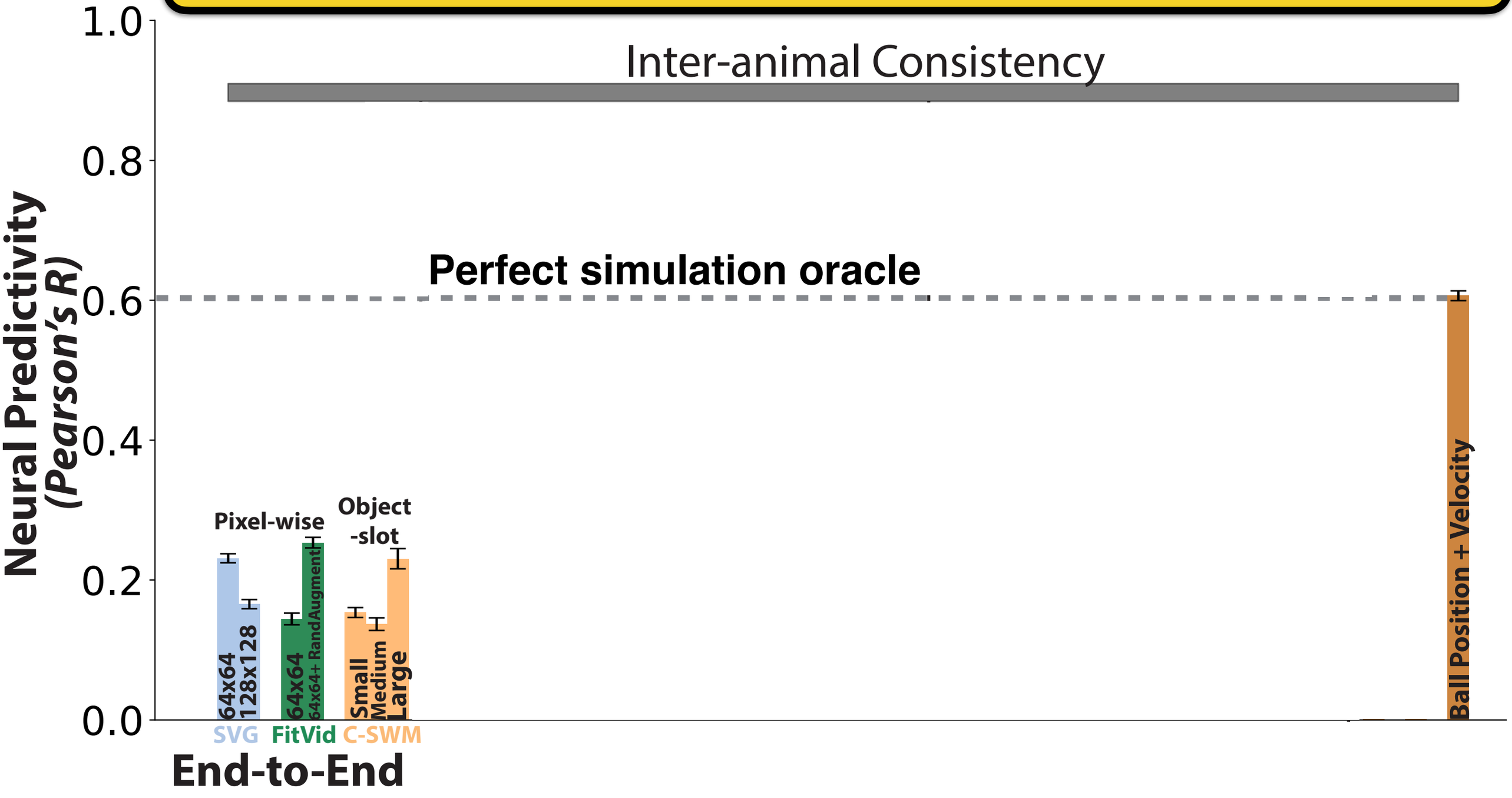
End-to-End

Ball Position + Velocity

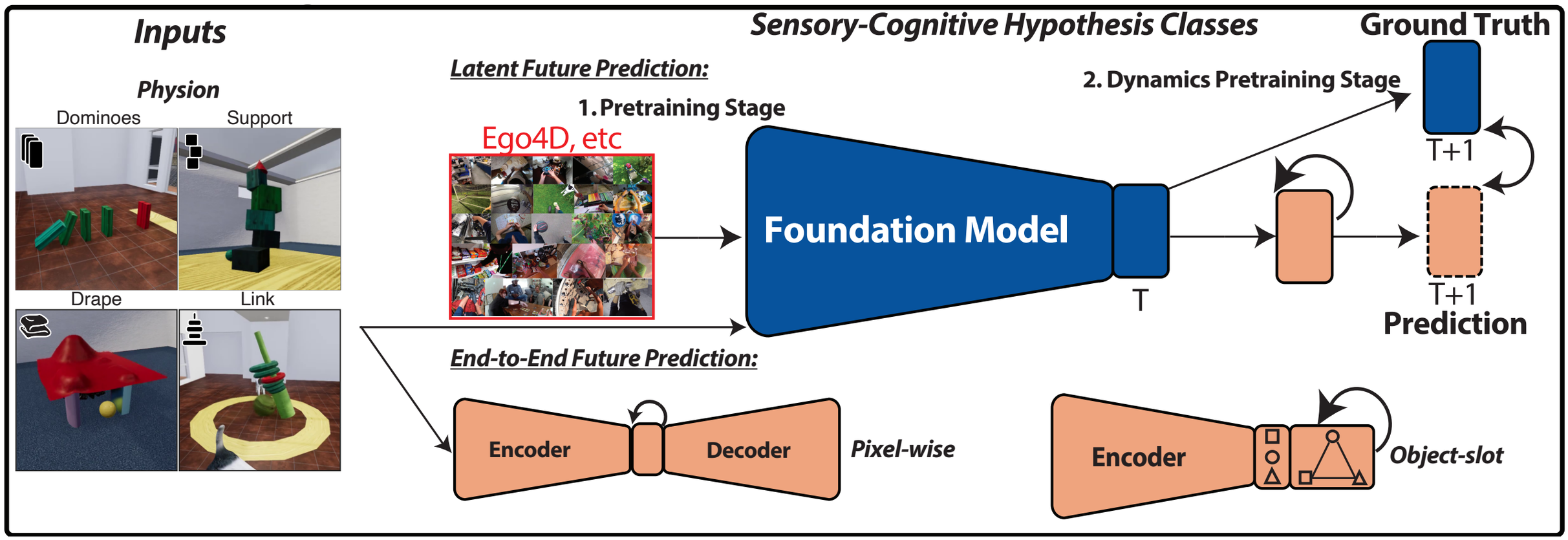
Object Slot Future Prediction Poorly Predicts Neurons

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

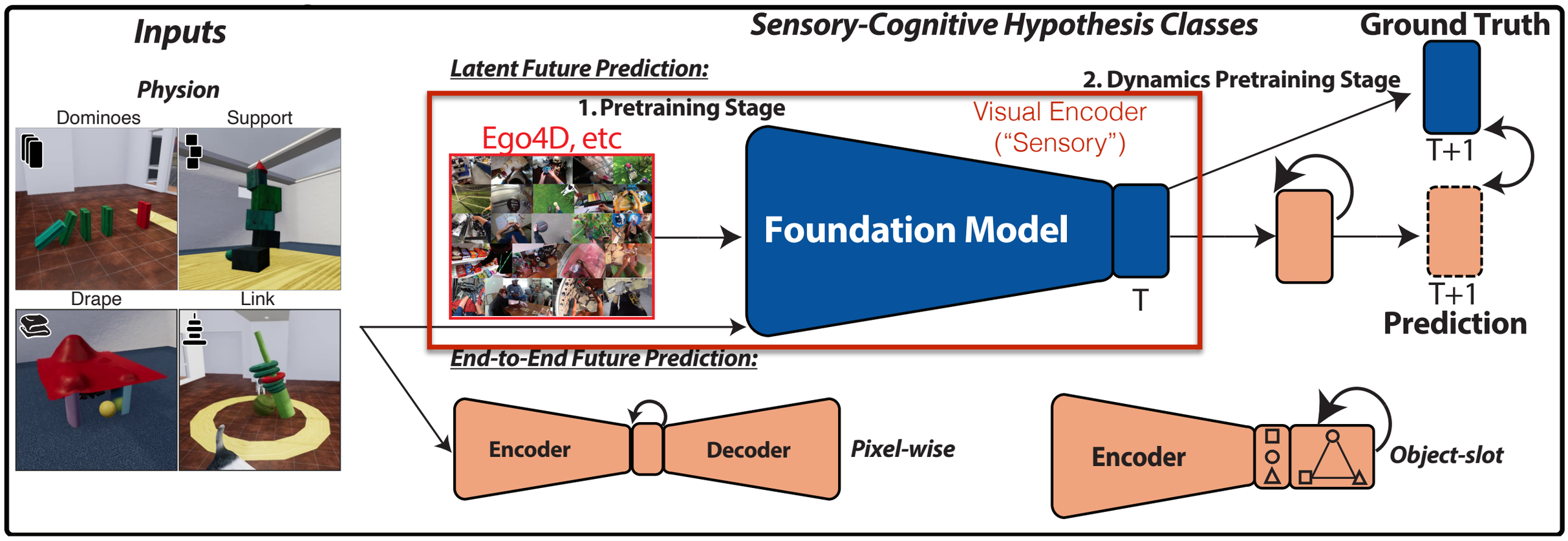




Hypothesis Class 3: Latent Future Prediction

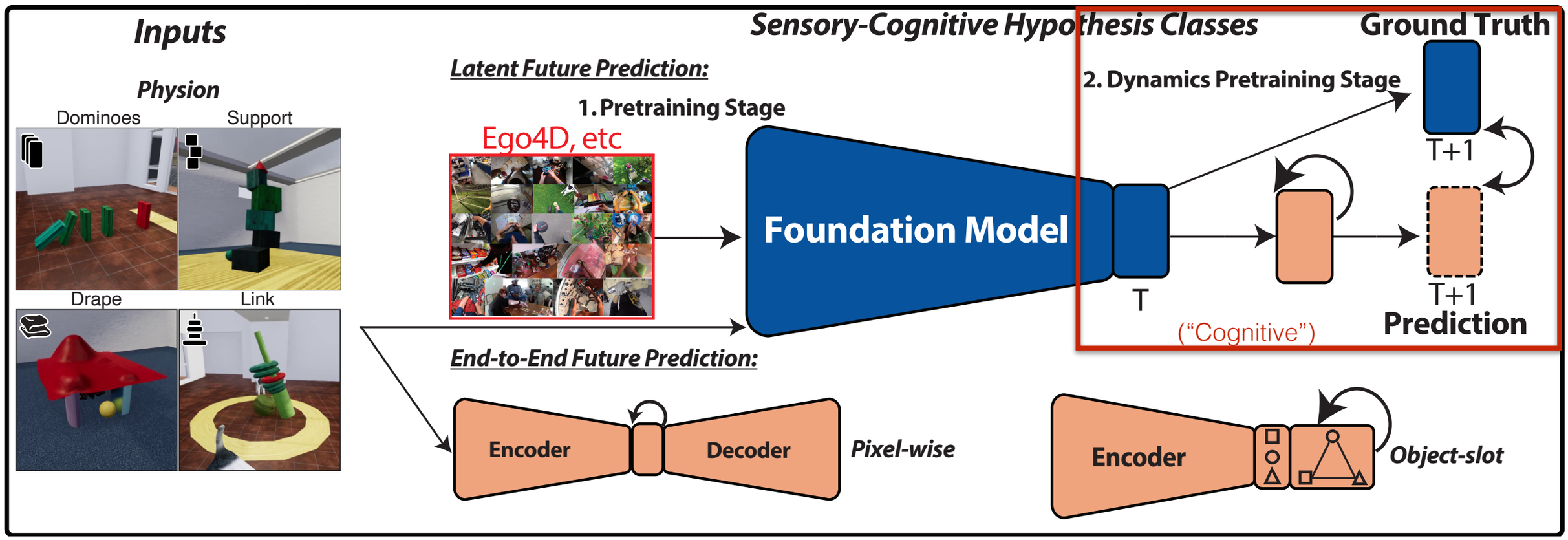


Hypothesis Class 3: Latent Future Prediction



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

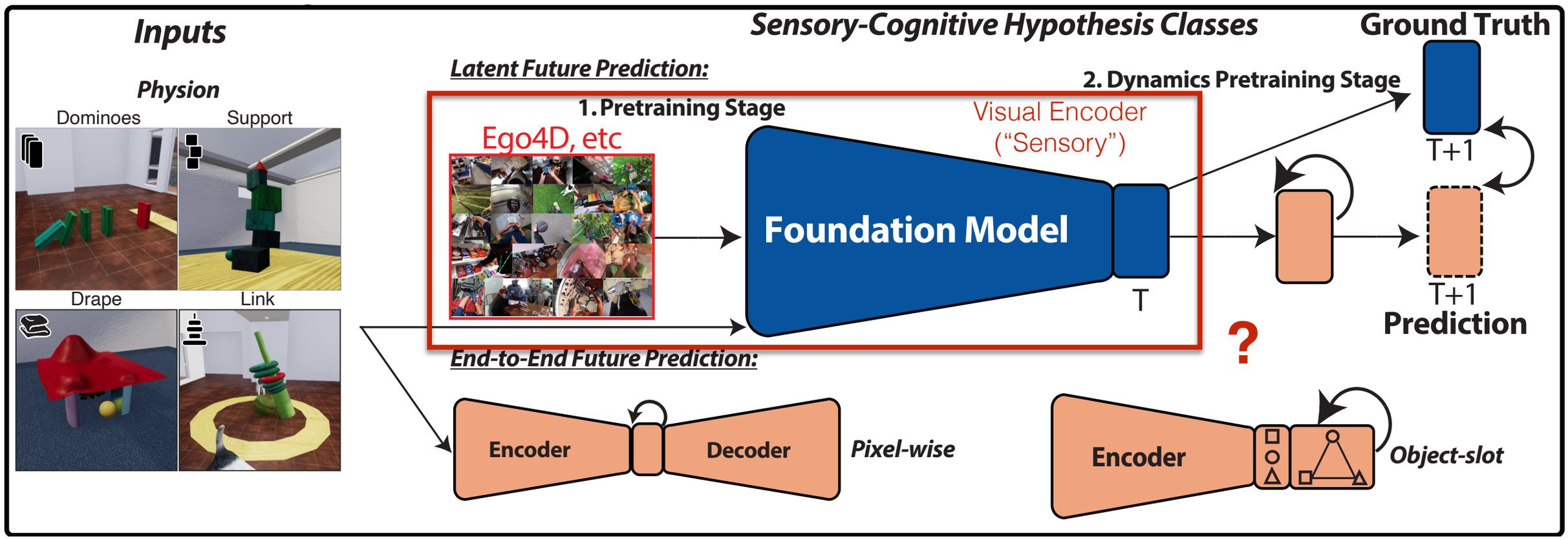
Hypothesis Class 3: Latent Future Prediction



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



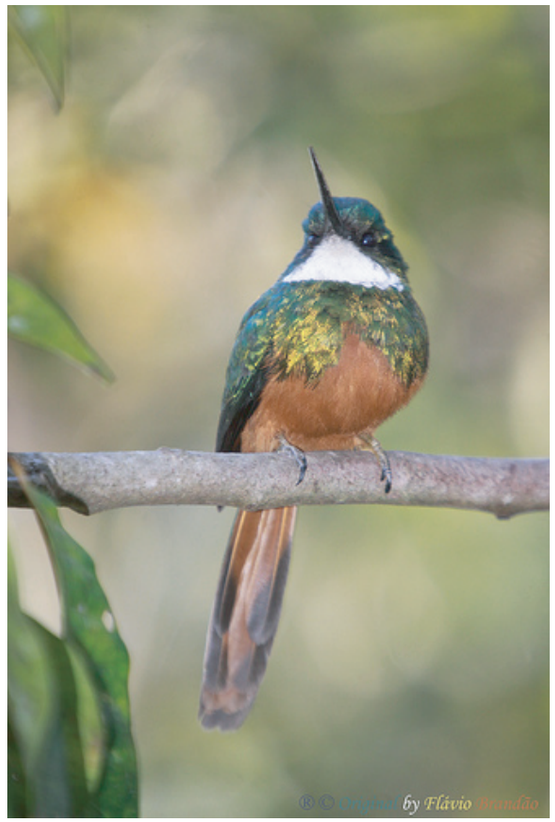
Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

What vision task?

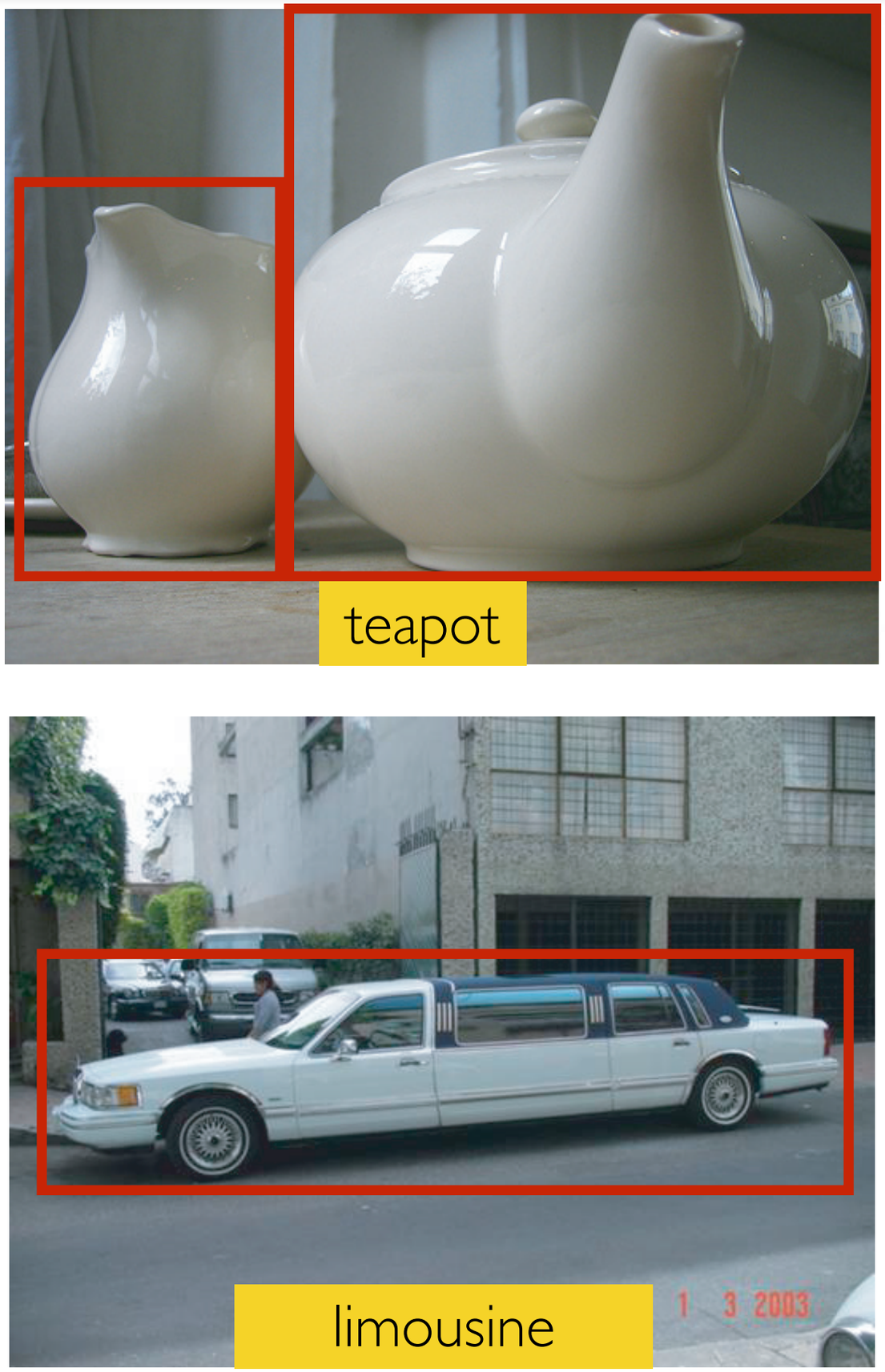
Leverage these dynamics to do explicit future prediction

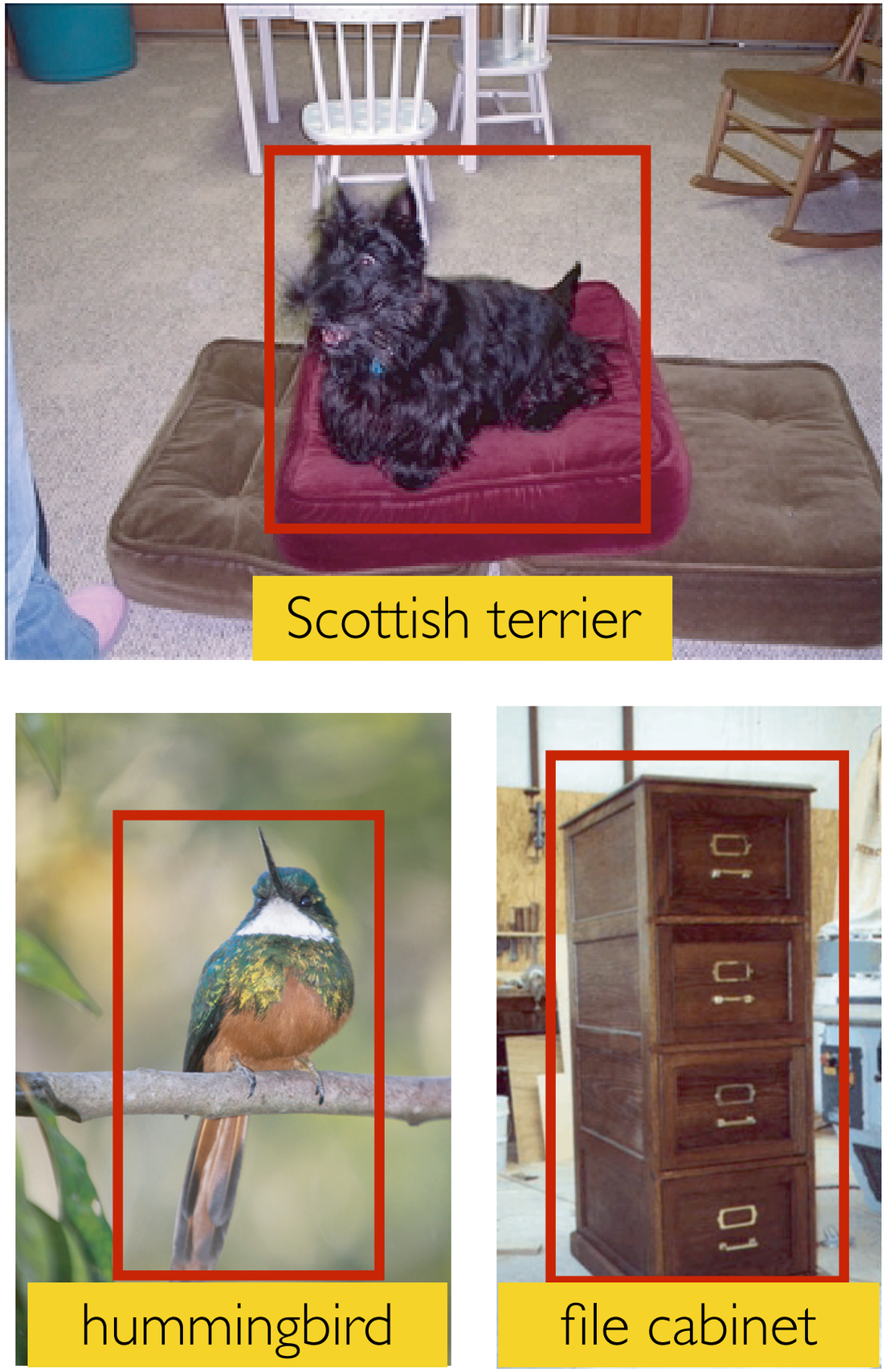
Hypothesis Class 3: Static Image Foundation Models

Hypothesis Class 3: Static Image Foundation Models

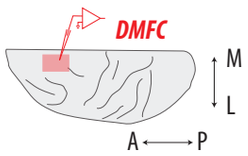


Hypothesis Class 3: Static Image Foundation Models





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

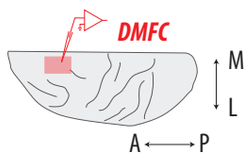
64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large

SVG
FitVid
C-SWM

End-to-End

Ball Position + Velocity

Static Image Foundation Future Prediction Poorly Predicts Neurons



Inter-animal Consistency

Perfect simulation oracle

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Image Foundation Models

Pixel-wise
64x64
128x128
64x64
64x64+ RandAugment

Object-slot
Small
Medium
Large

No Dynamics
LSTM Dynamics
CTRNN Dynamics

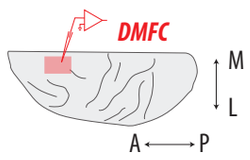
Ball Position + Velocity

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

End-to-End

Latent Future Prediction

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

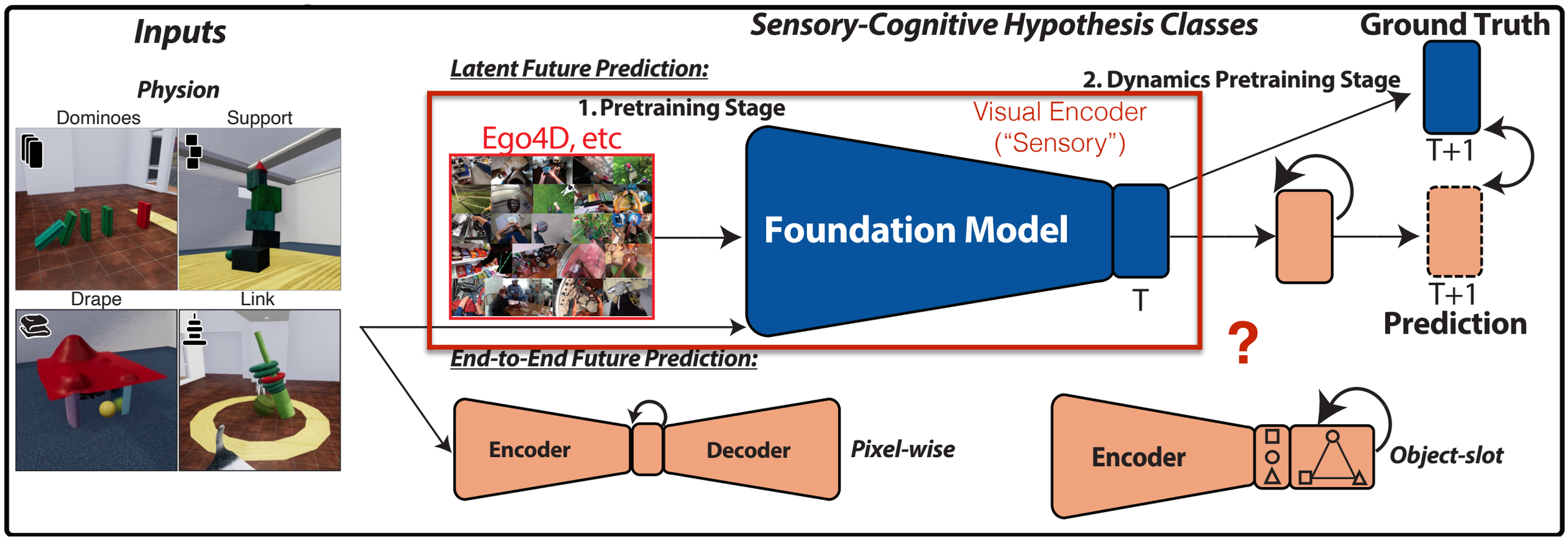
Pixel-wise
Object -slot
64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large
No Dynamics
LSTM Dynamics
CTRNN Dynamics
Ball Position + Velocity

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

End-to-End

Latent Future Prediction

Hypothesis Class 3: Foundation Models

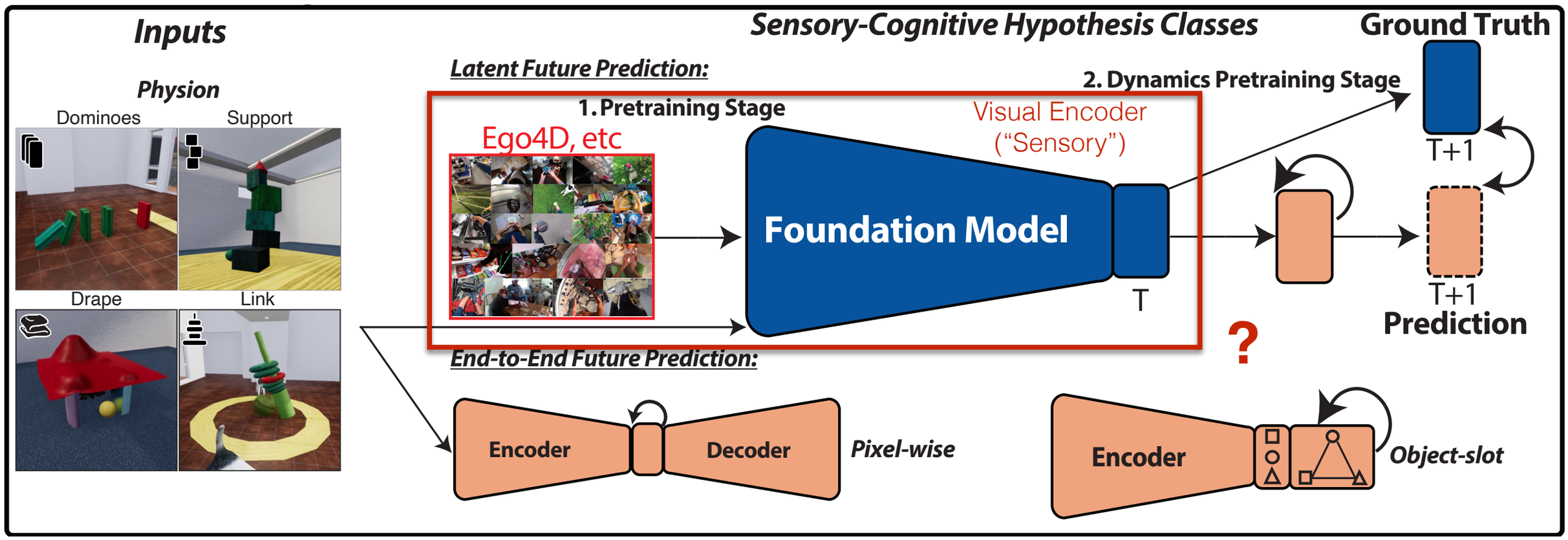


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

What vision task?

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")

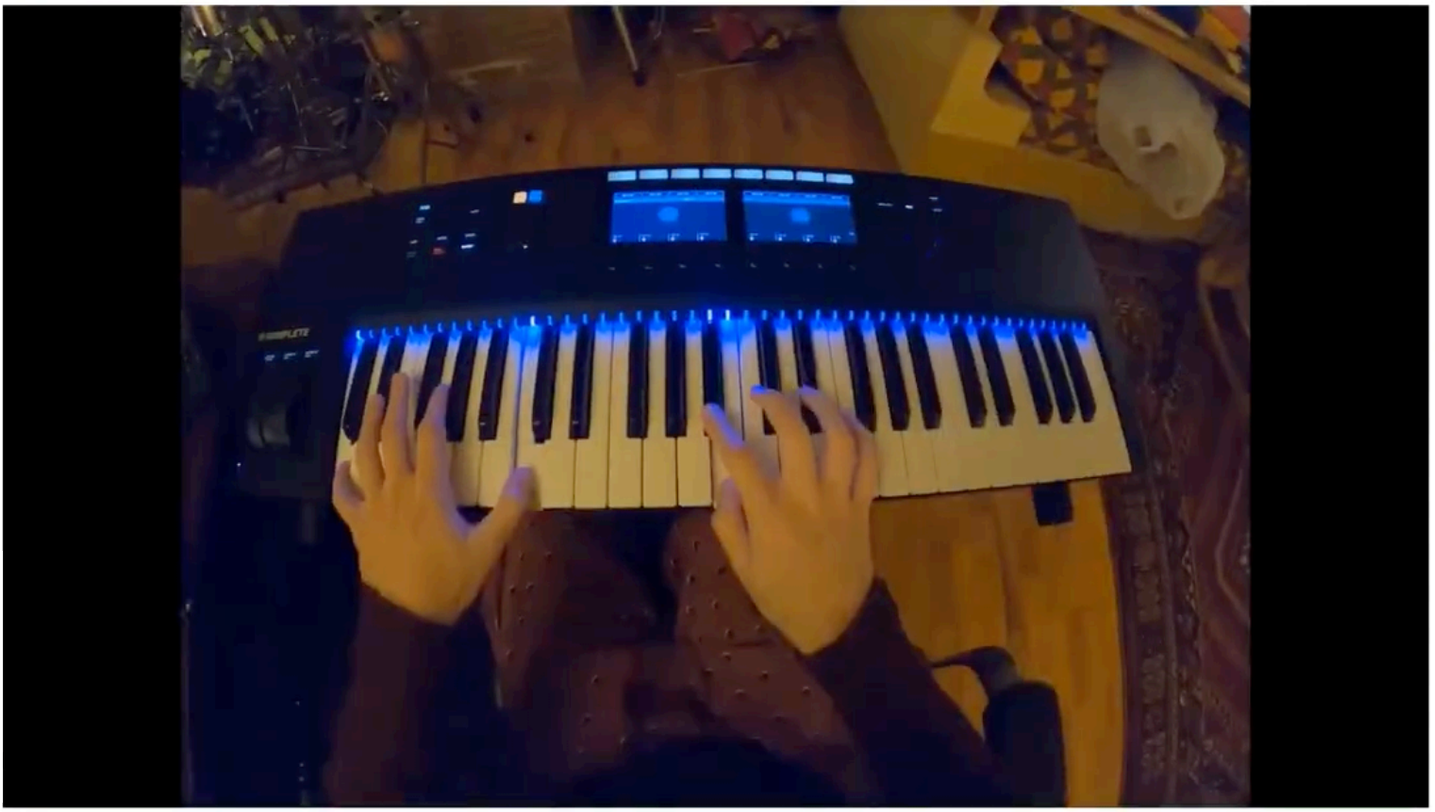
What vision task?

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



Ego4D: A massive-scale egocentric dataset

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

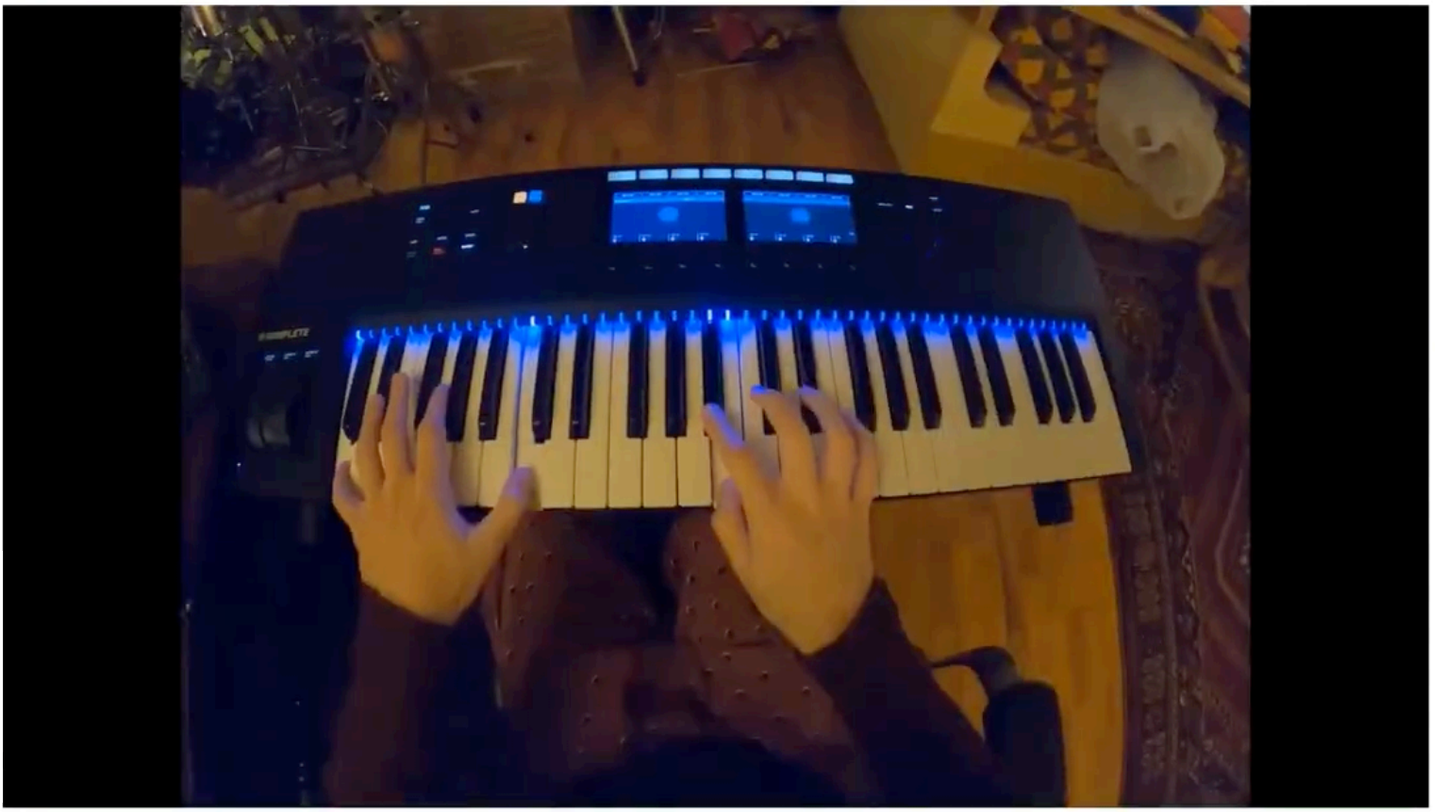
931 participants from 74 worldwide locations

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

Grauman et al. 2022

Hypothesis Class 3:Video Foundation Models

Ego4D: everyday activity around the world



Ego4D: A massive-scale egocentric dataset

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

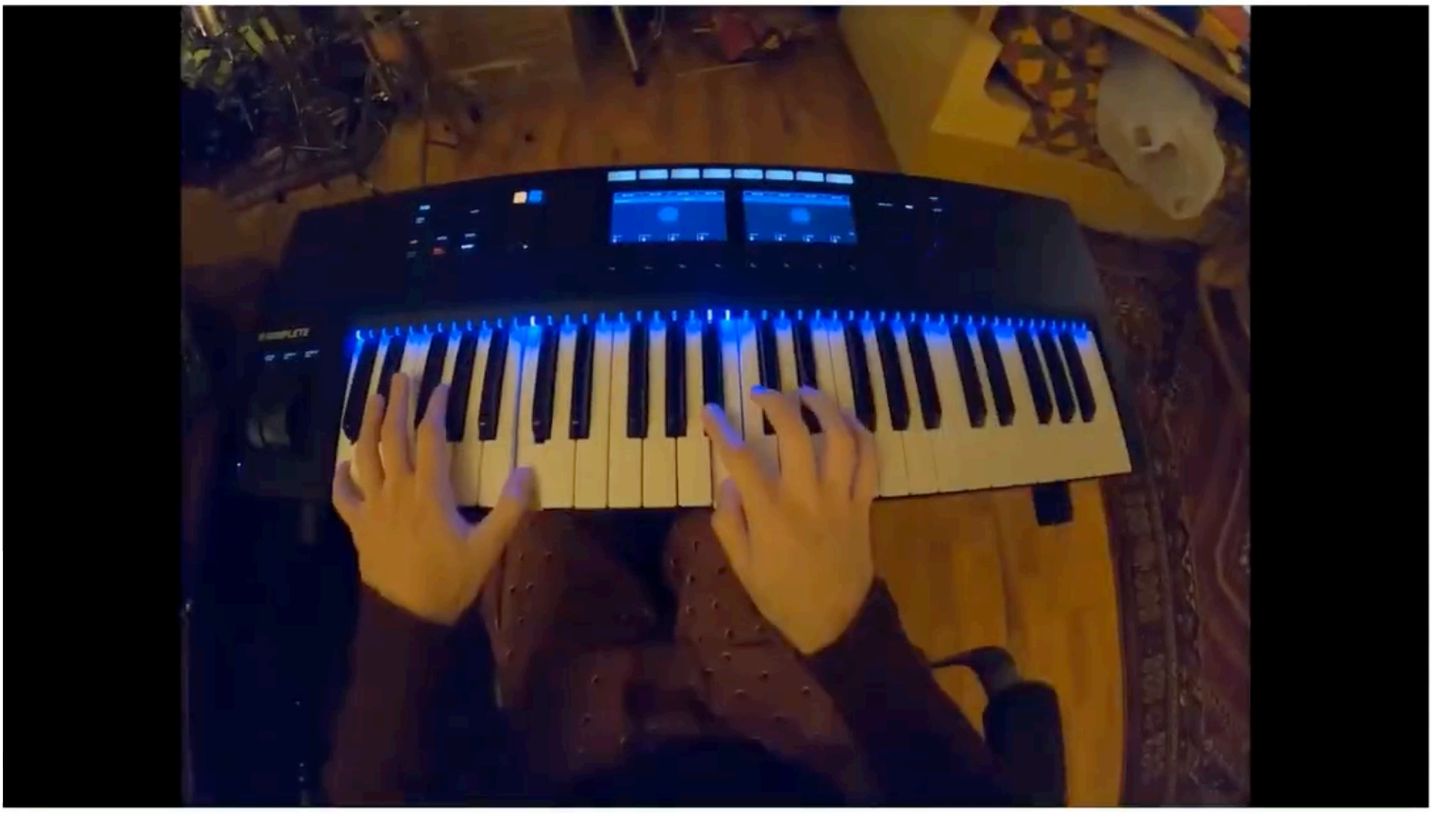
931 participants from 74 worldwide locations

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

Grauman et al. 2022

Hypothesis Class 3:Video Foundation Models

Ego4D: everyday activity around the world



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

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

Ego4D: A massive-scale egocentric dataset

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

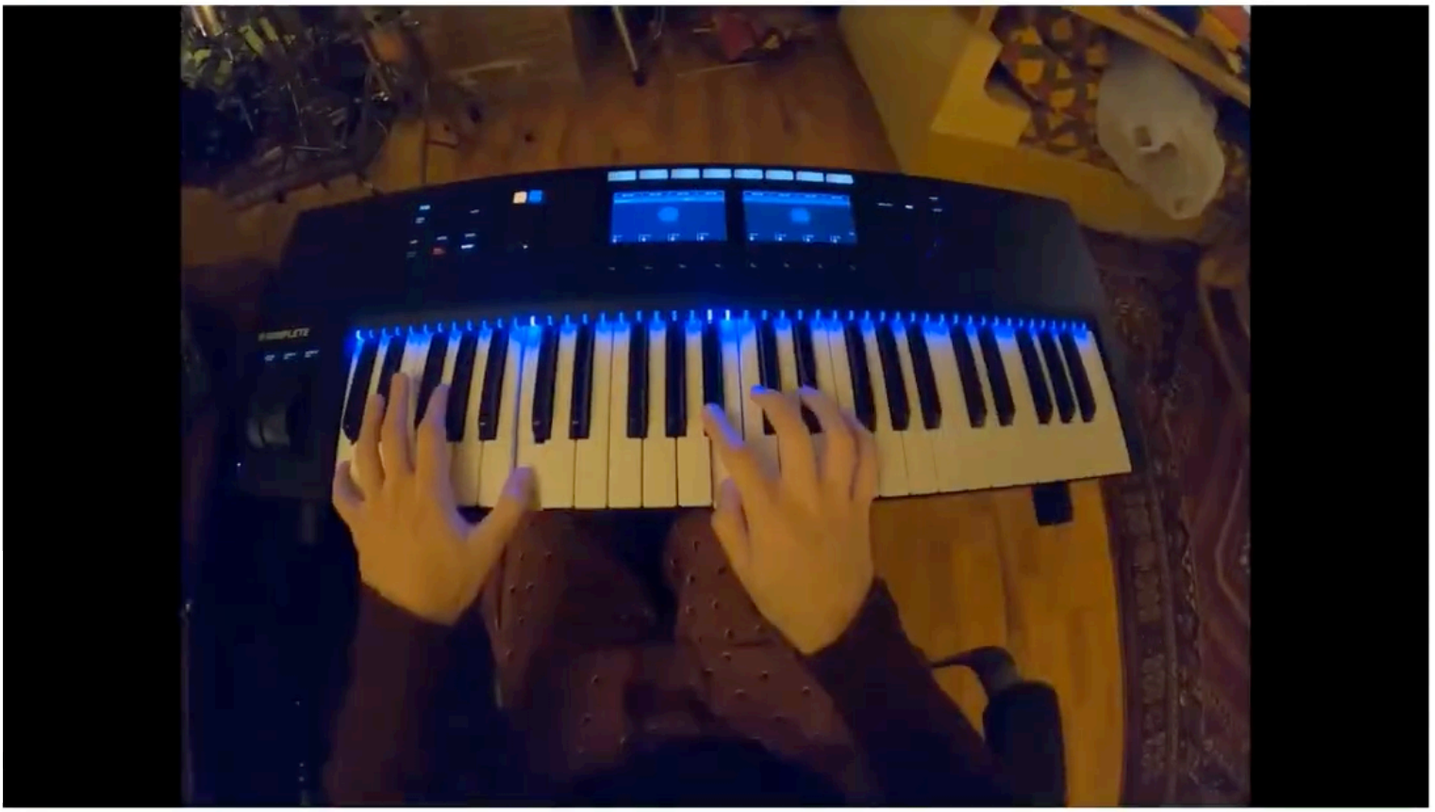
931 participants from 74 worldwide locations

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

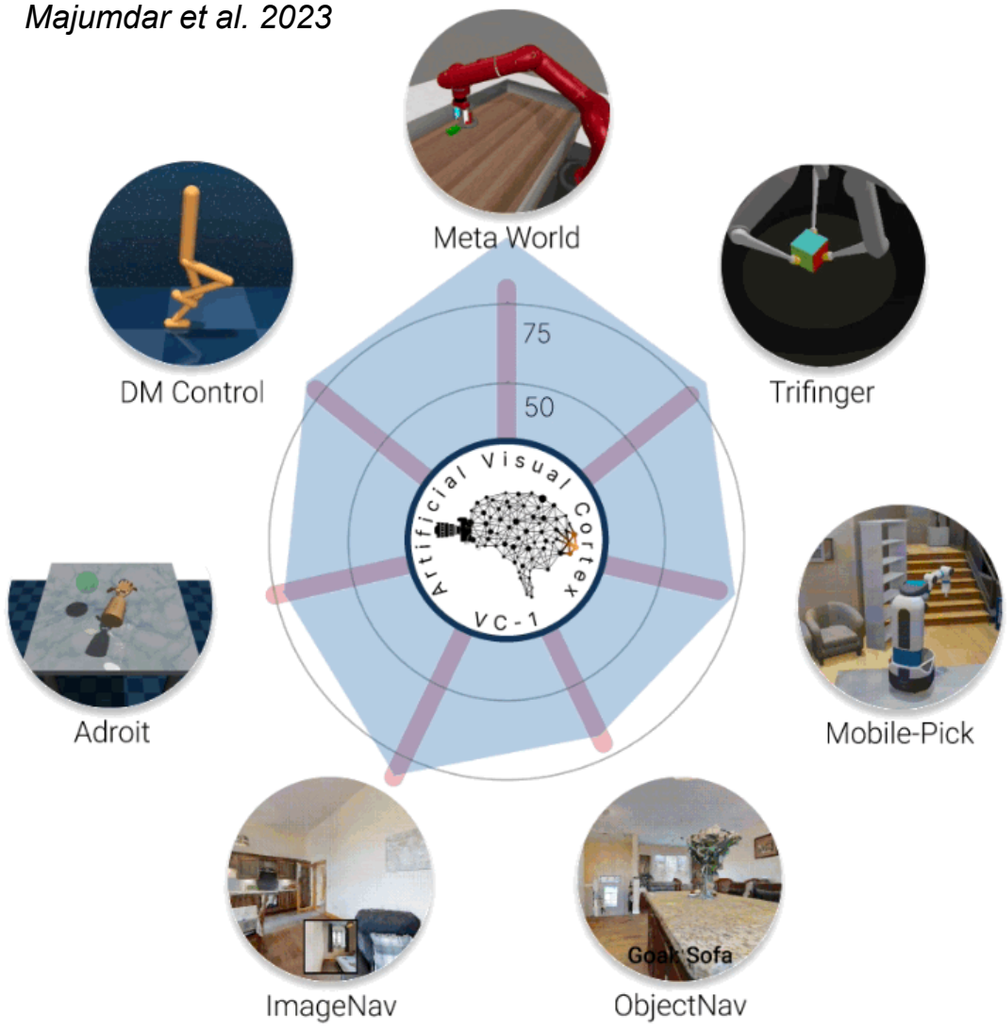
Grauman et al. 2022

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

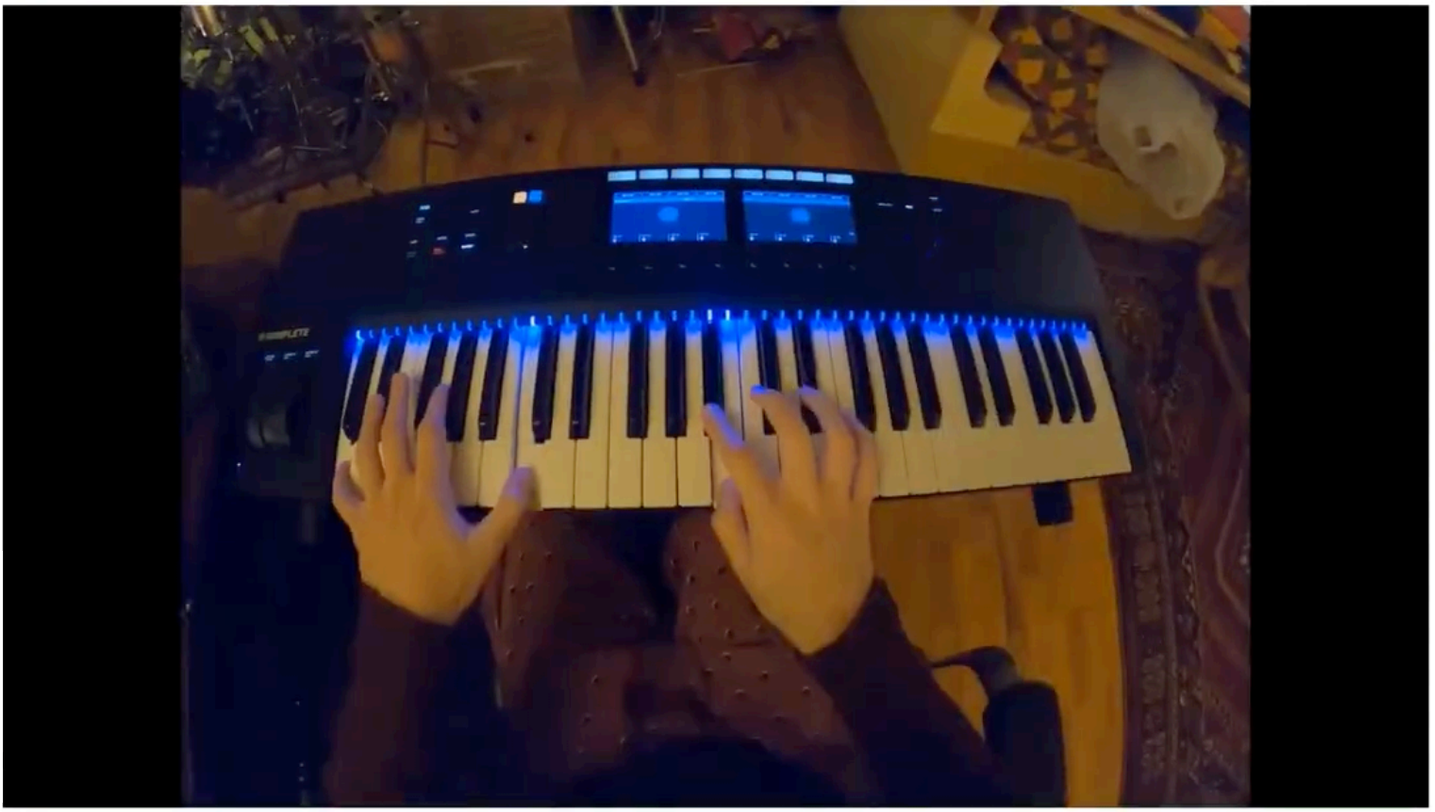
931 participants from 74 worldwide locations

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

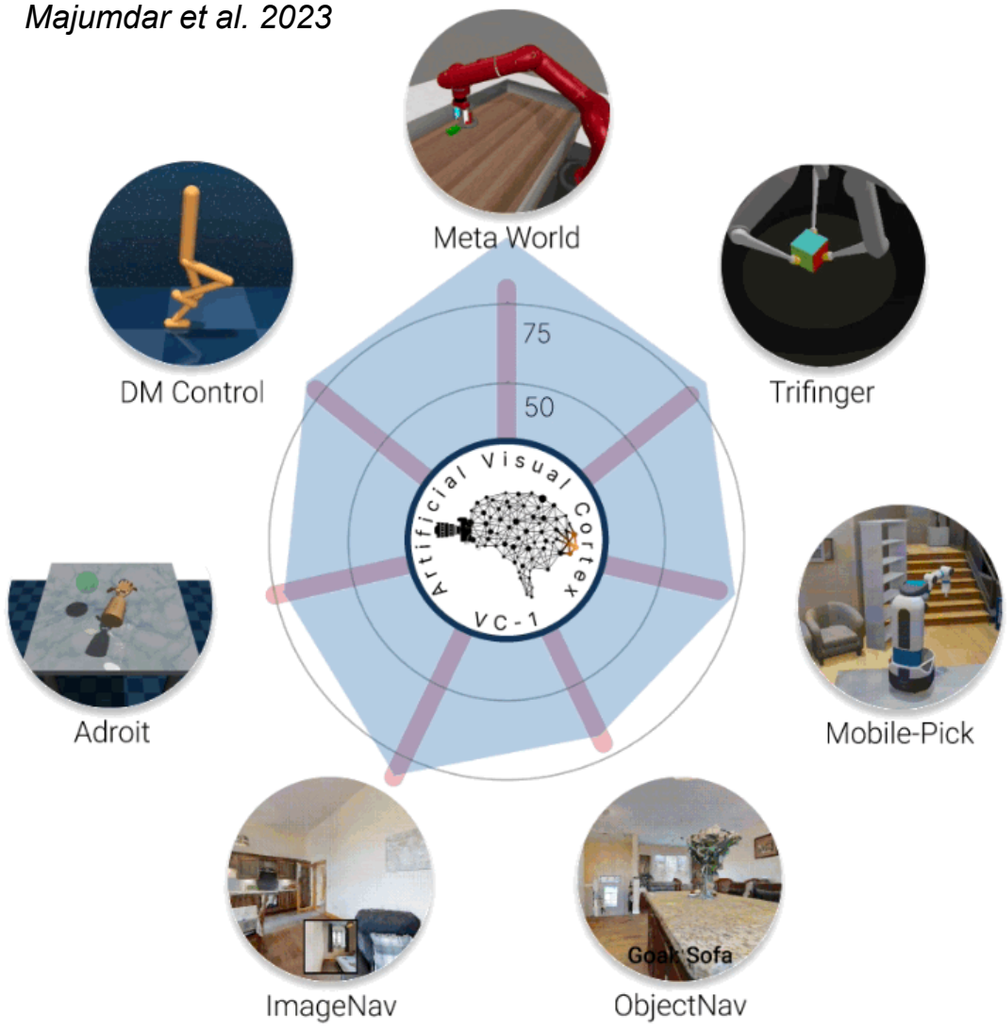
Grauman et al. 2022

Hypothesis Class 3: Video Foundation Models

Ego4D: everyday activity around the world



Majumdar et al. 2023



Ego4D: A massive-scale egocentric dataset

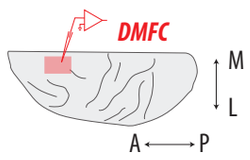
3,670 hours of in-the-wild daily life activity

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

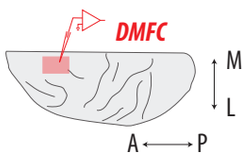
Pixel-wise
Object -slot
64x64
128x128
64x64
64x64+ RandAugment
Small
Medium
Large
No Dynamics
LSTM Dynamics
CTRNN Dynamics
Ball Position + Velocity

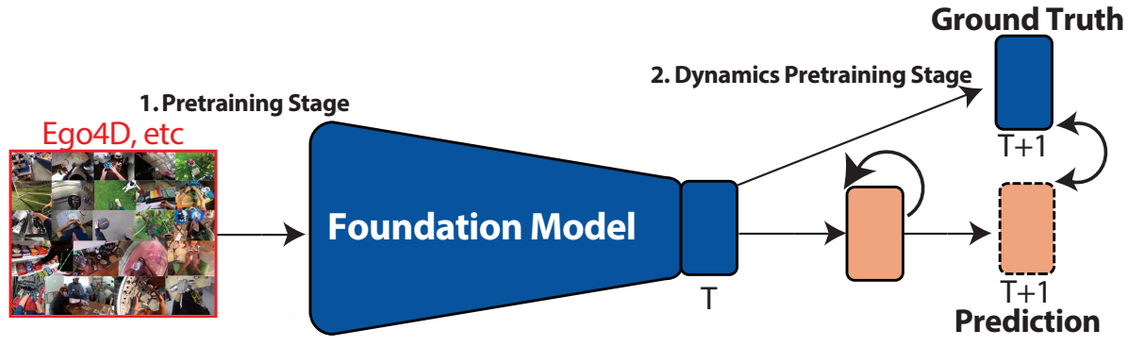
SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP

End-to-End

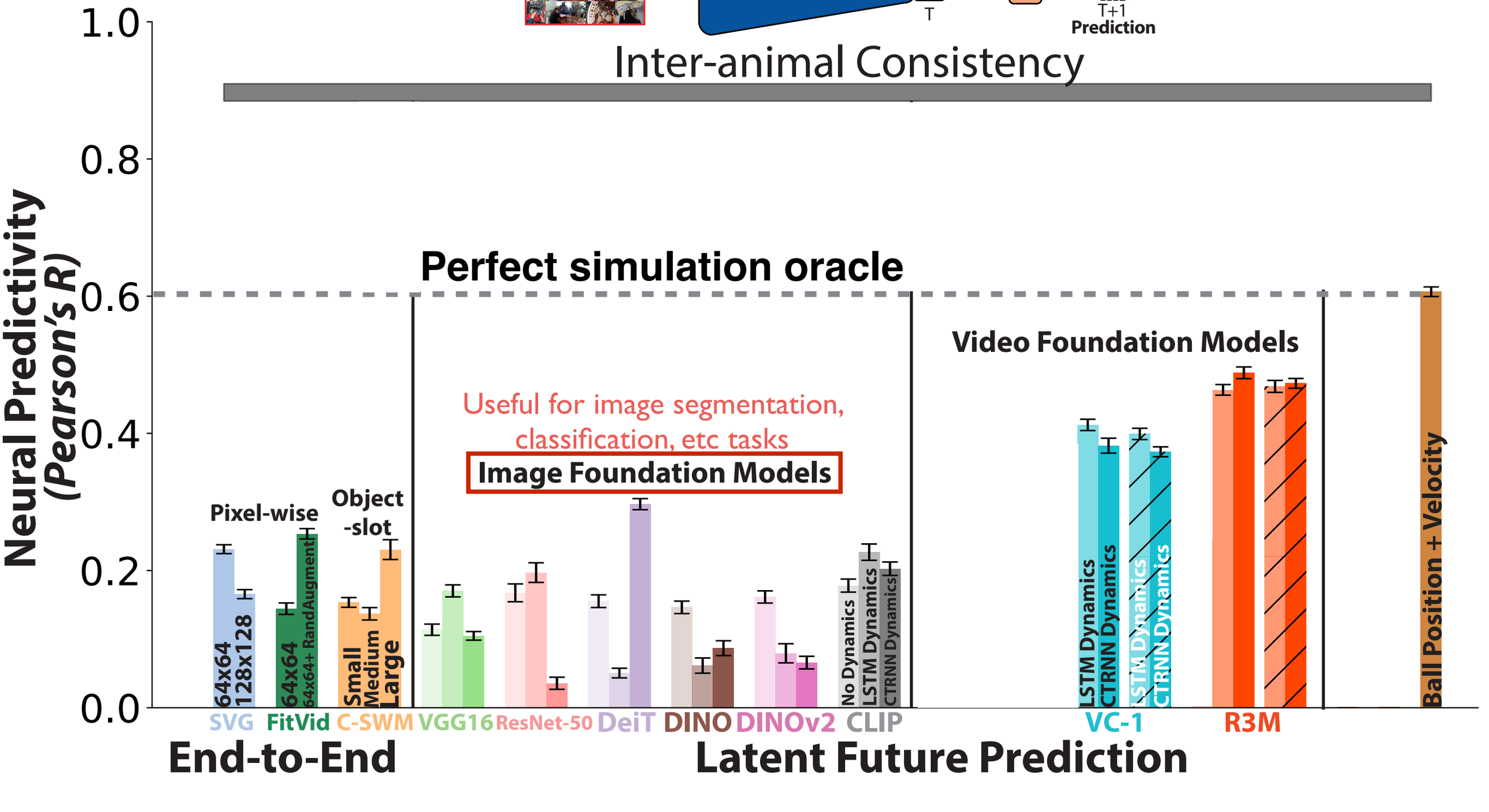
Latent Future Prediction

Video Foundation Future Prediction Best Predict Neurons

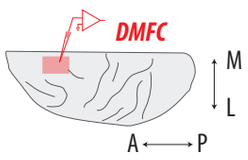


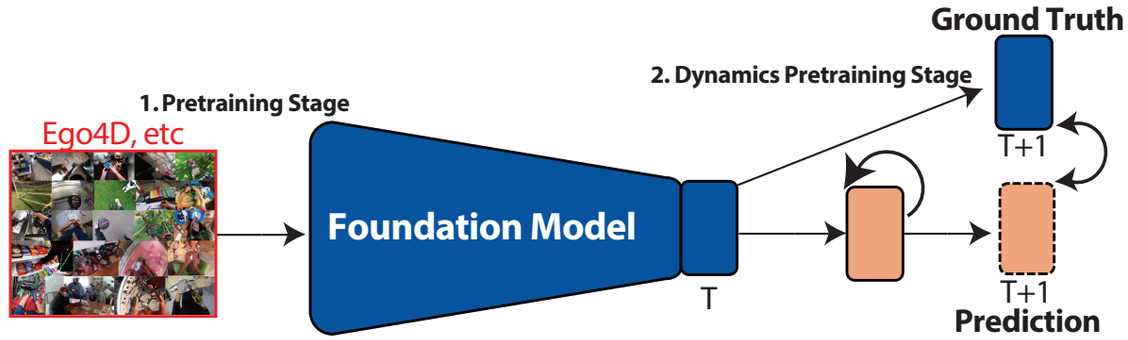


Inter-animal Consistency

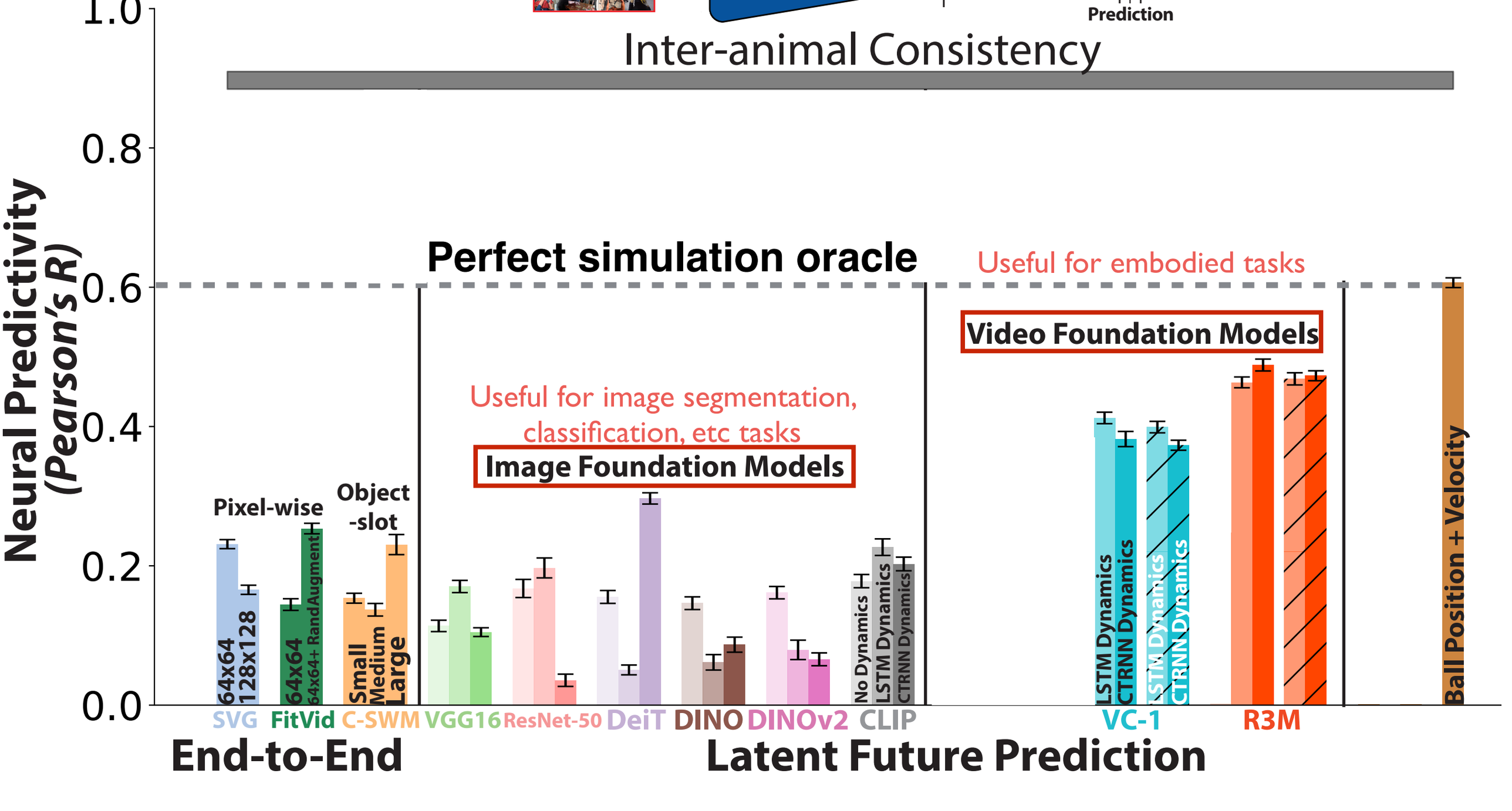


Video Foundation Future Prediction Best Predict Neurons

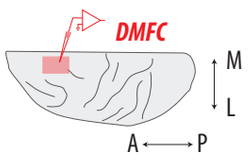


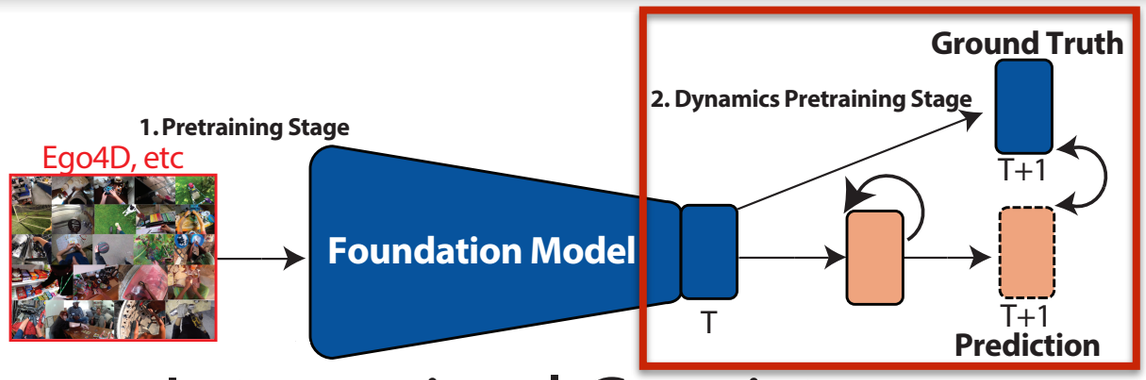


Inter-animal Consistency

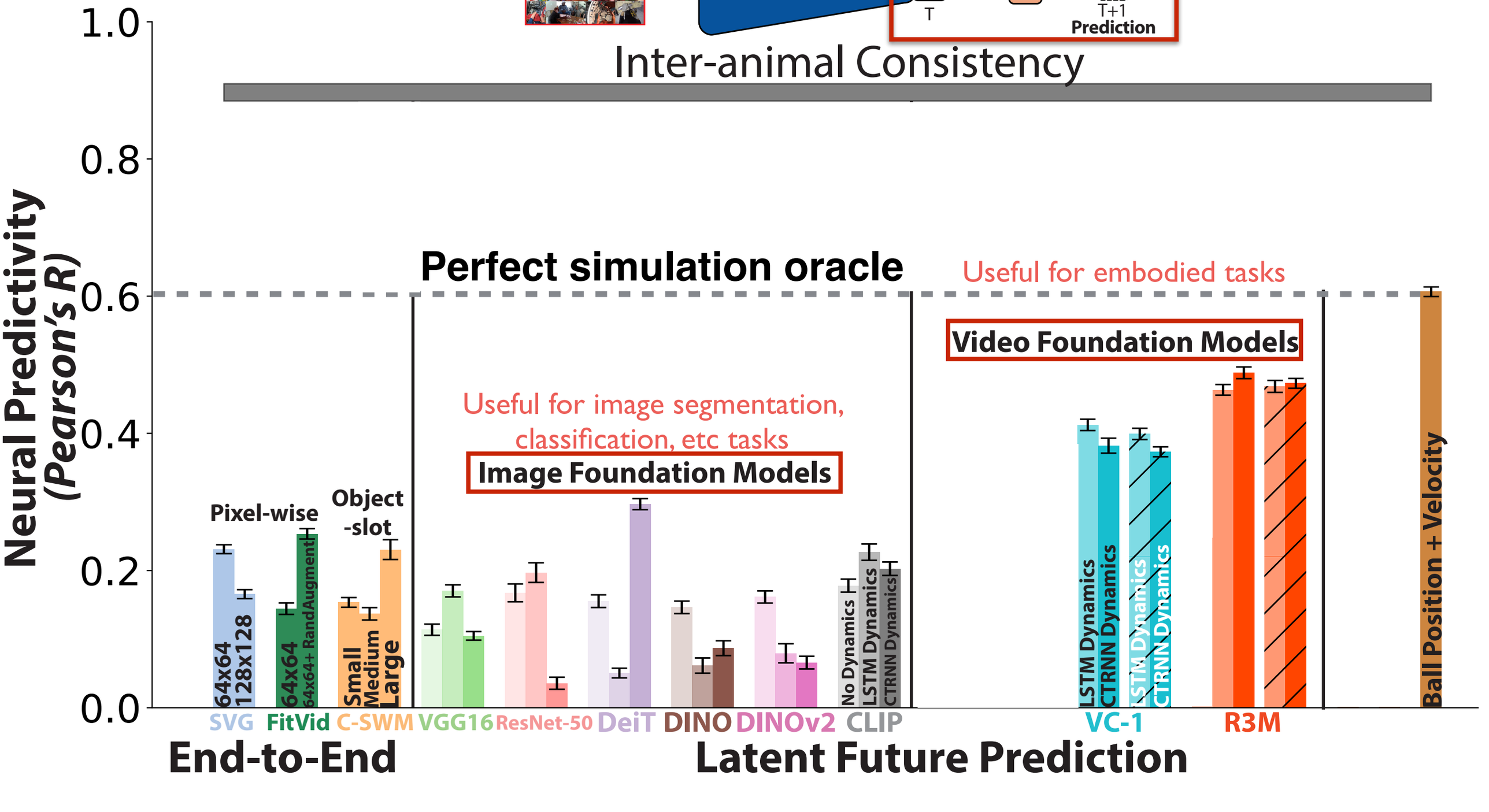


Video Foundation Future Prediction Best Predict Neurons

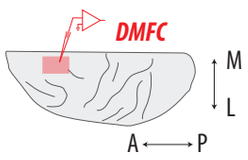


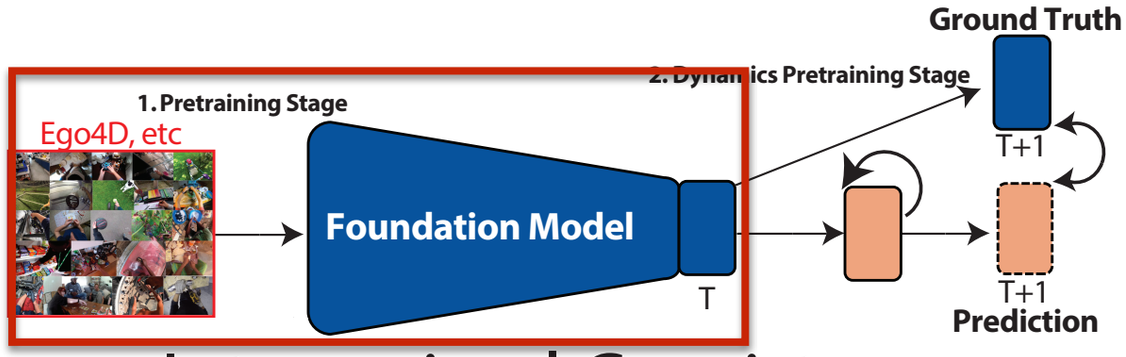


Inter-animal Consistency

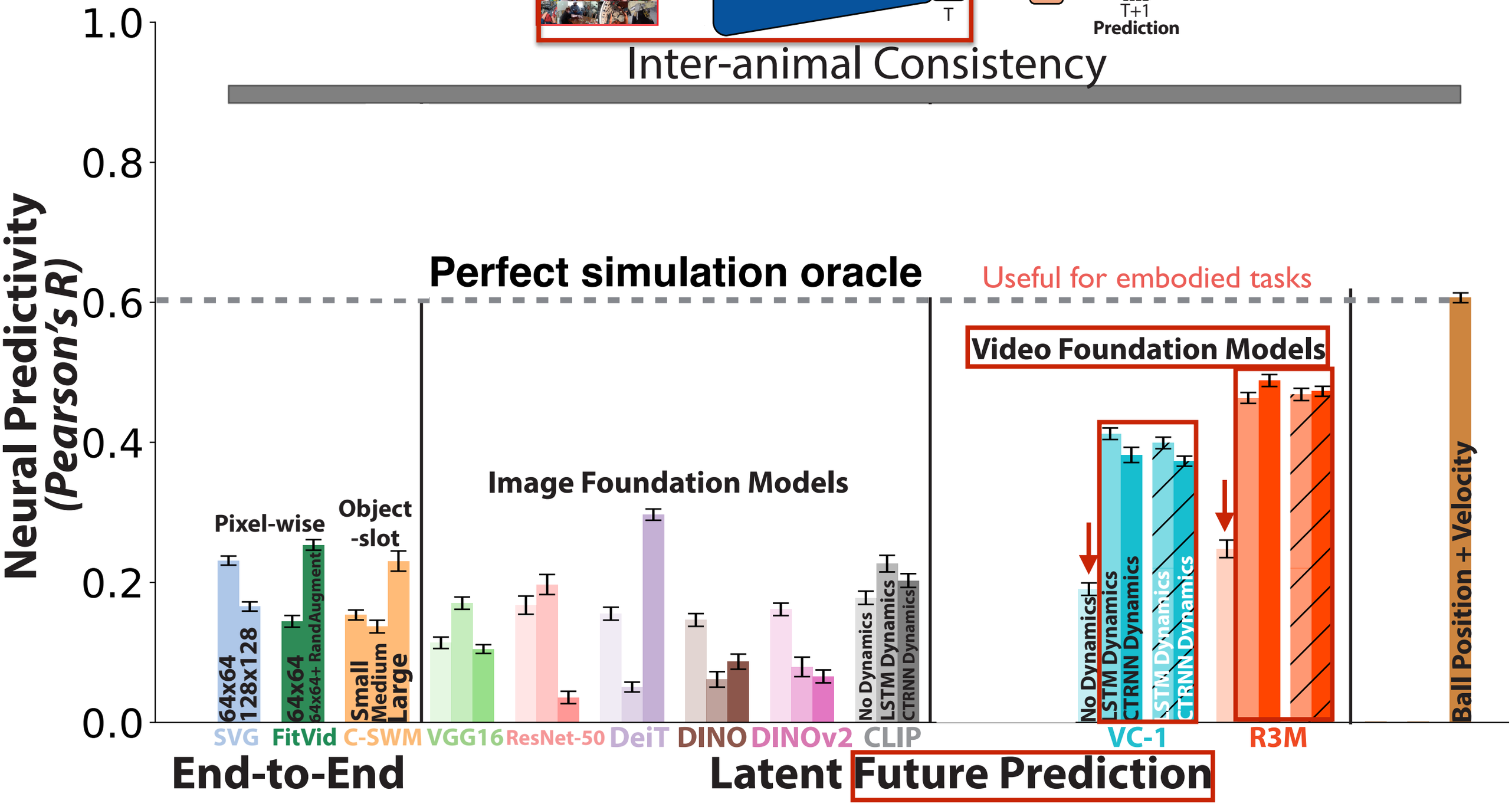


Video Foundation Future Prediction Best Predict Neurons

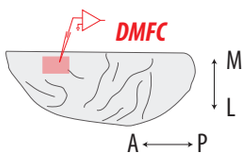


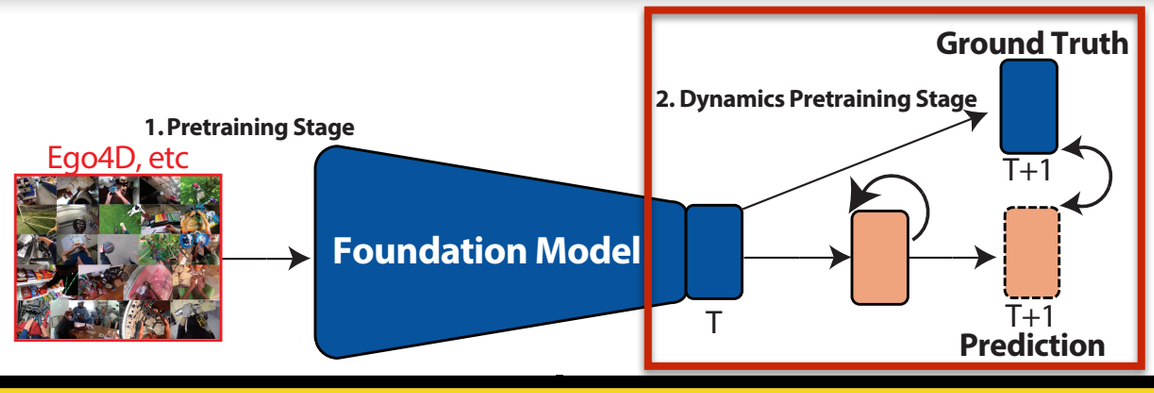


Inter-animal Consistency

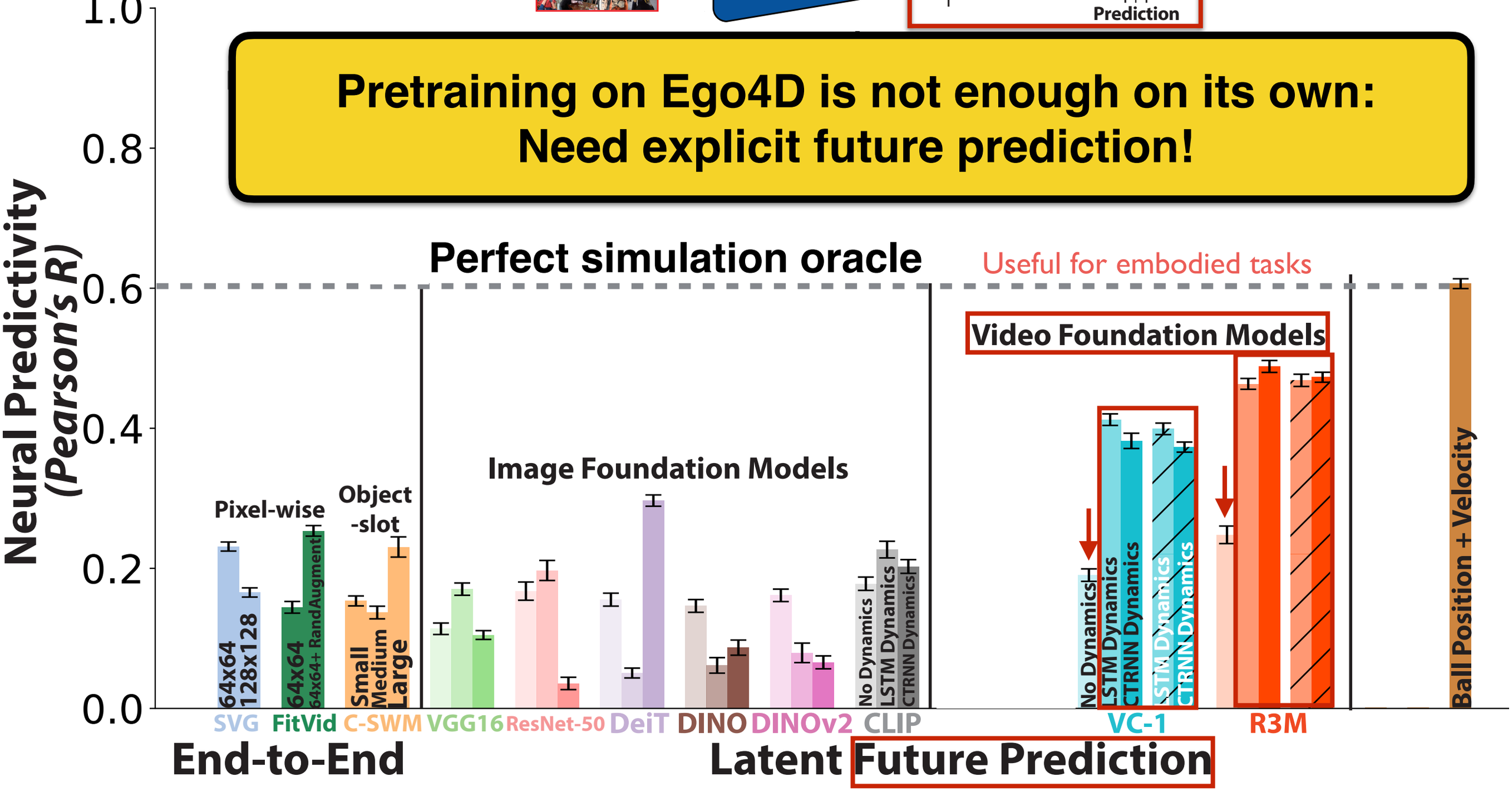


Video Foundation Future Prediction Best Predict Neurons

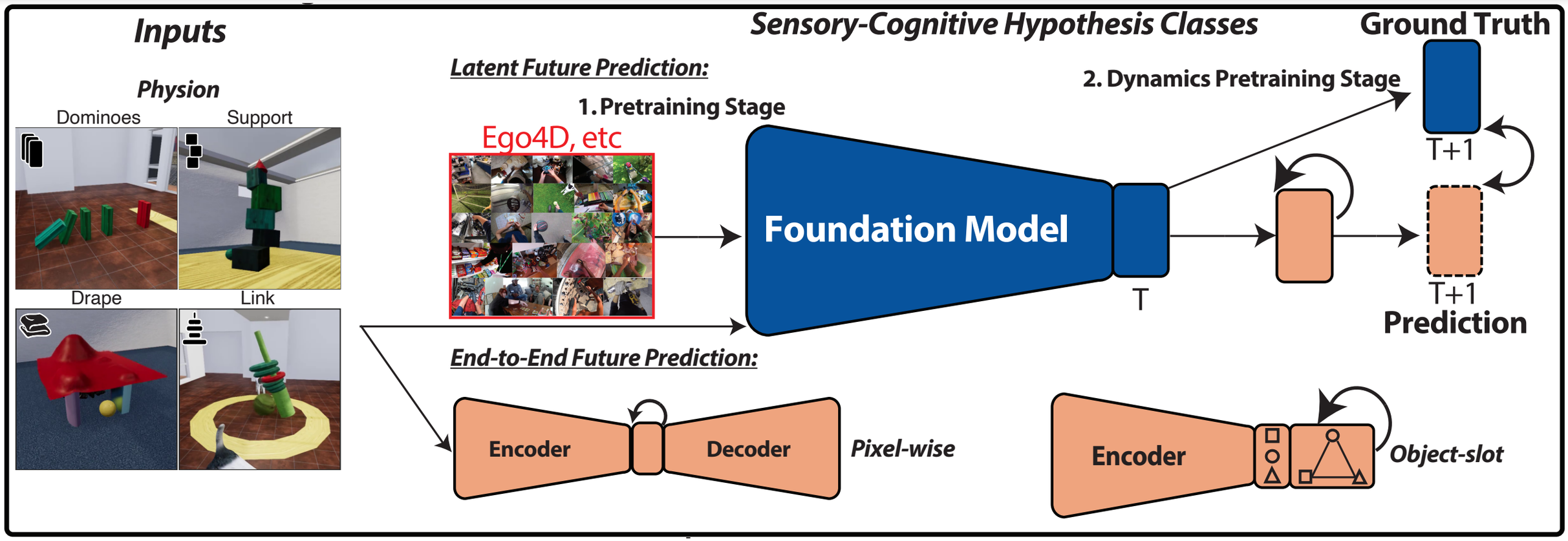


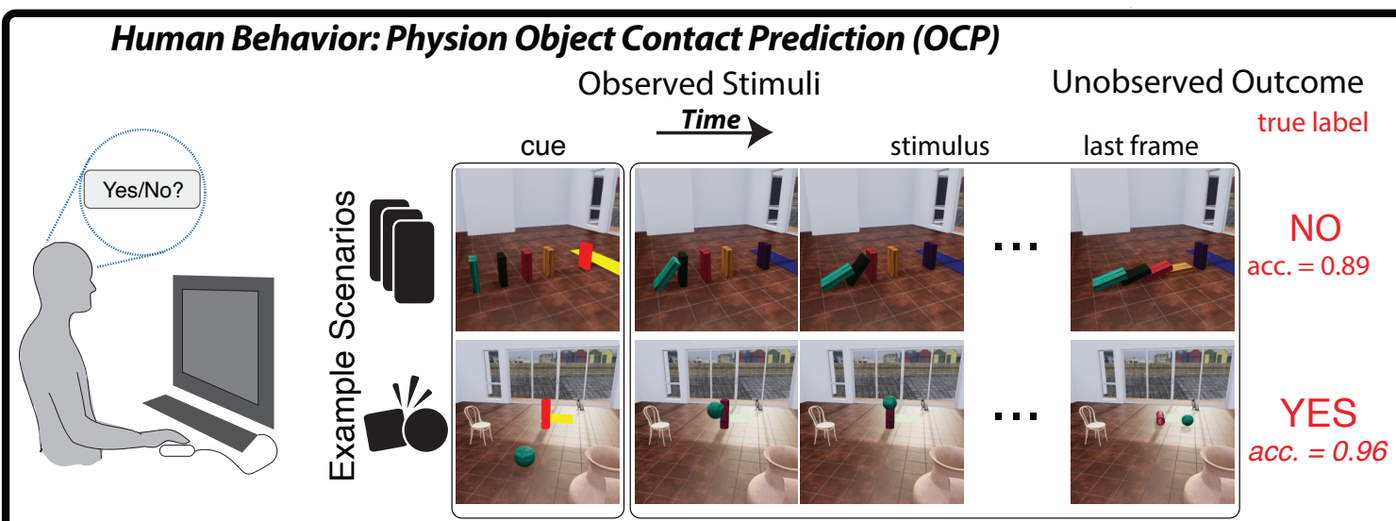


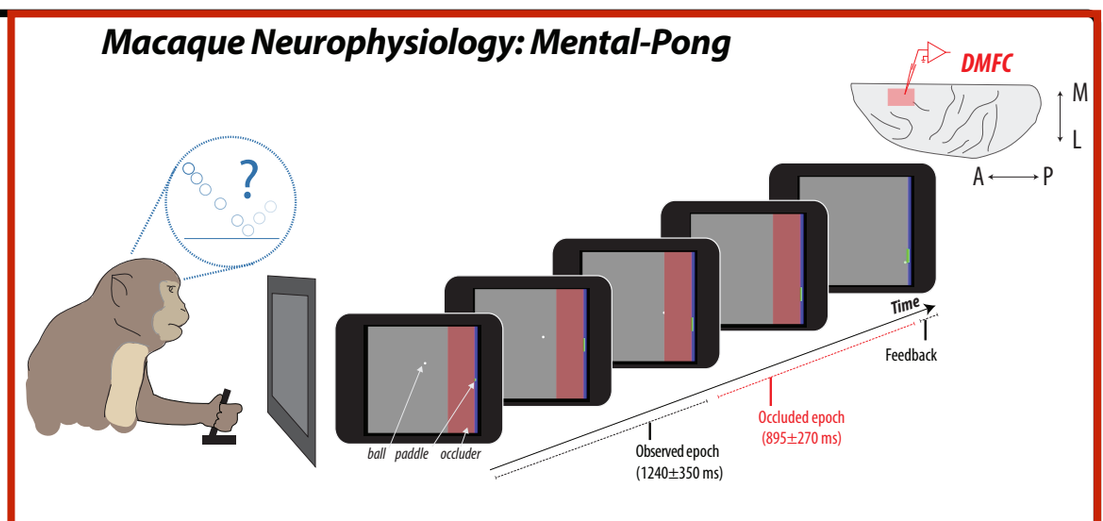
Pretraining on Ego4D is not enough on its own:
Need explicit future prediction!



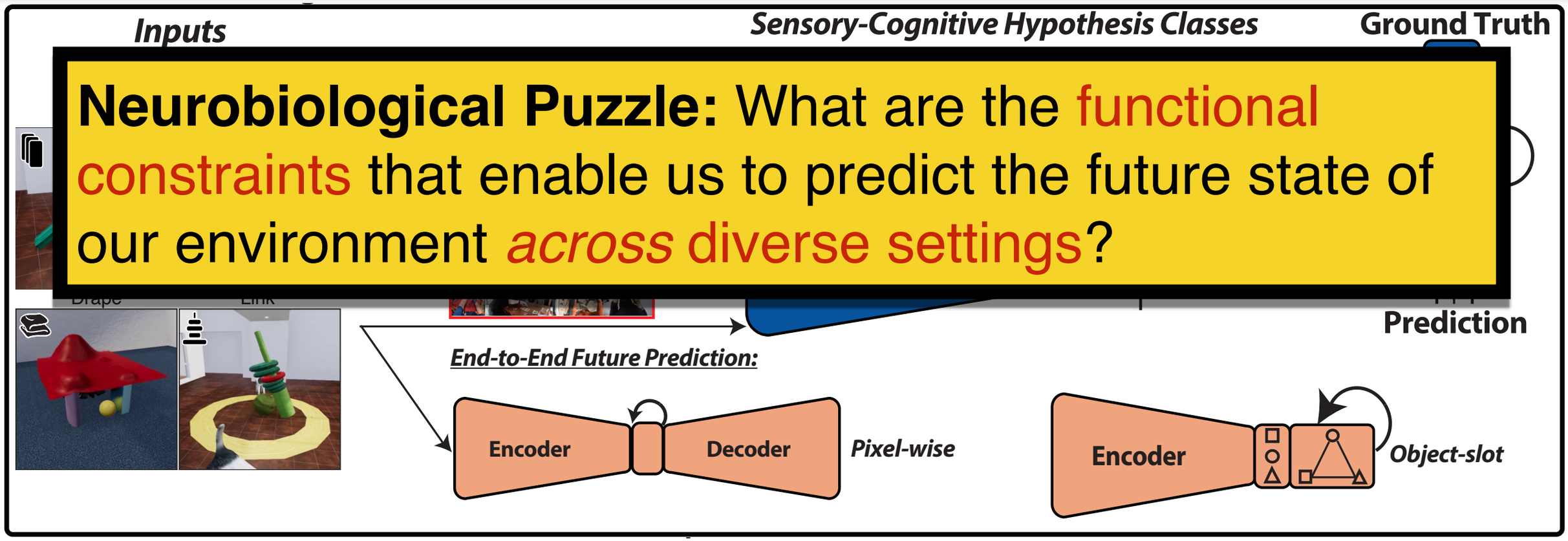
Macaque Neurophysiology: Mental Pong

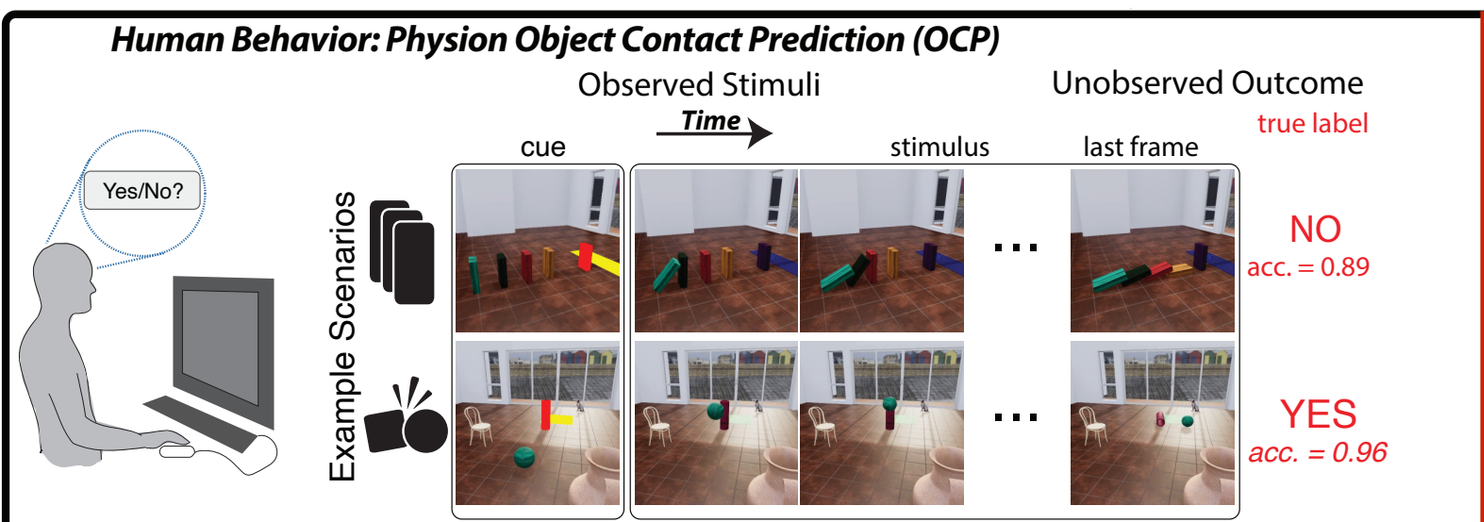


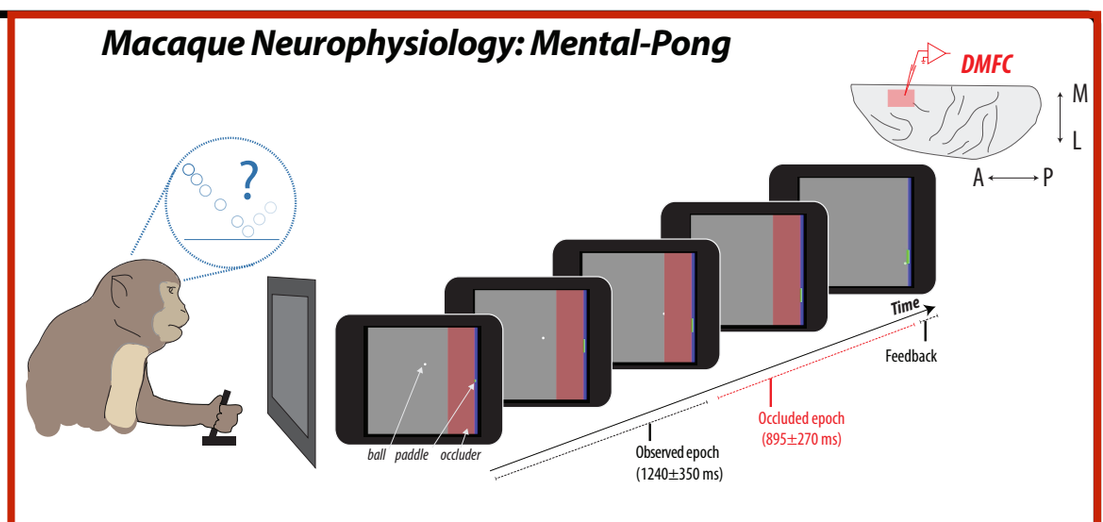




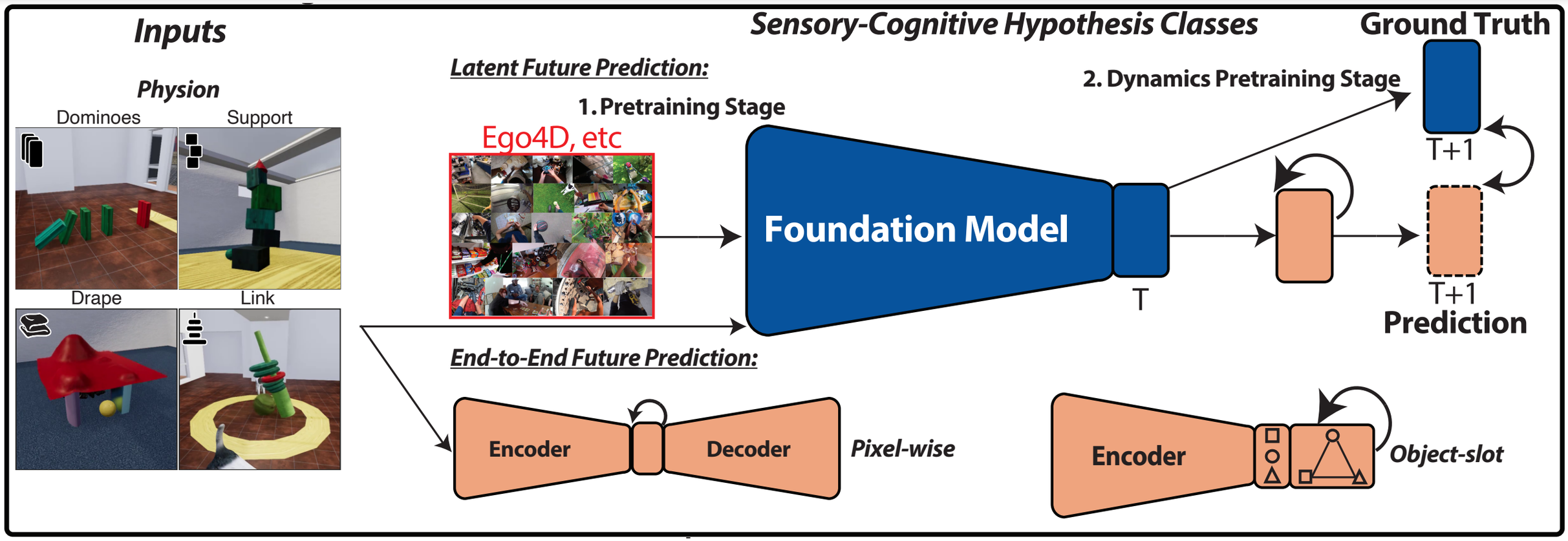
Macaque Neurophysiology: Mental Pong

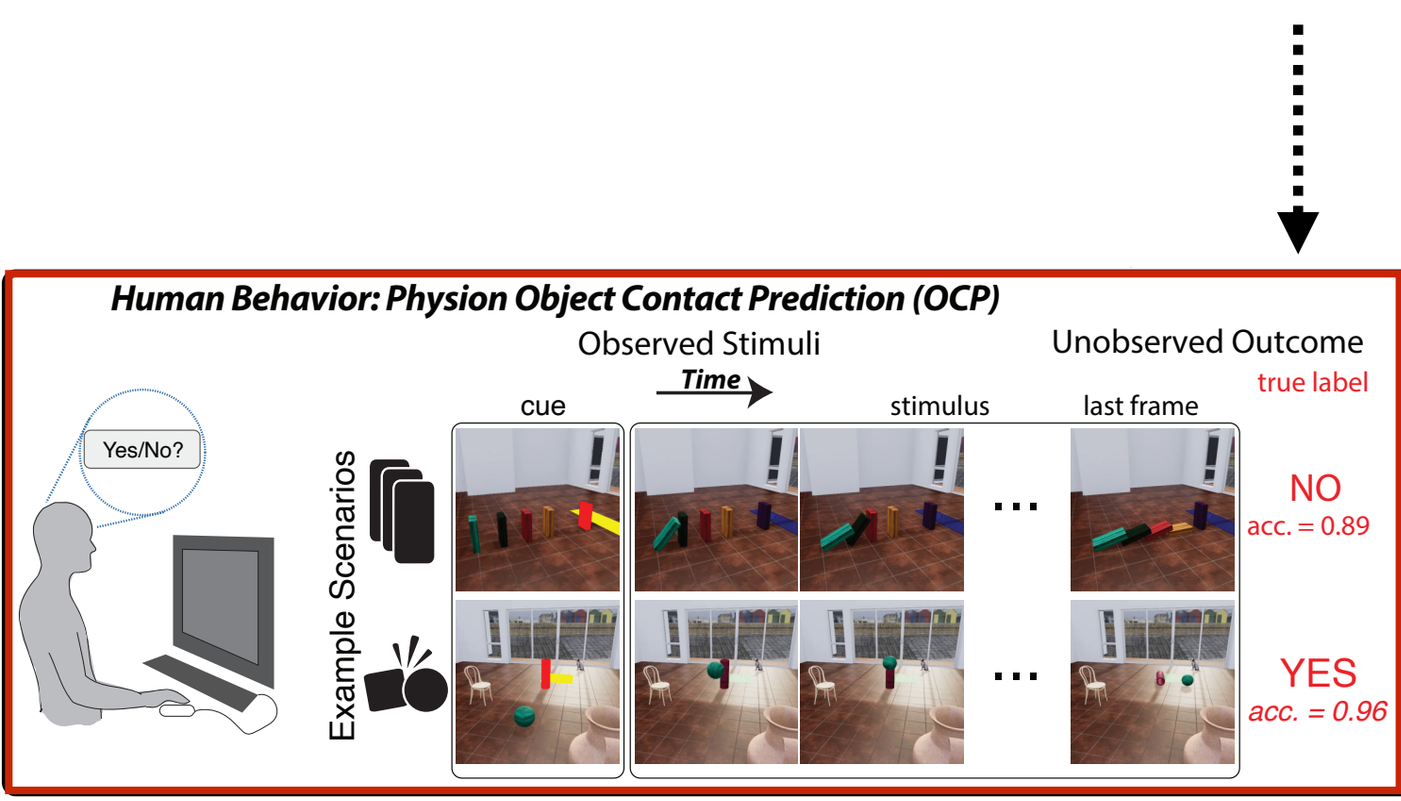


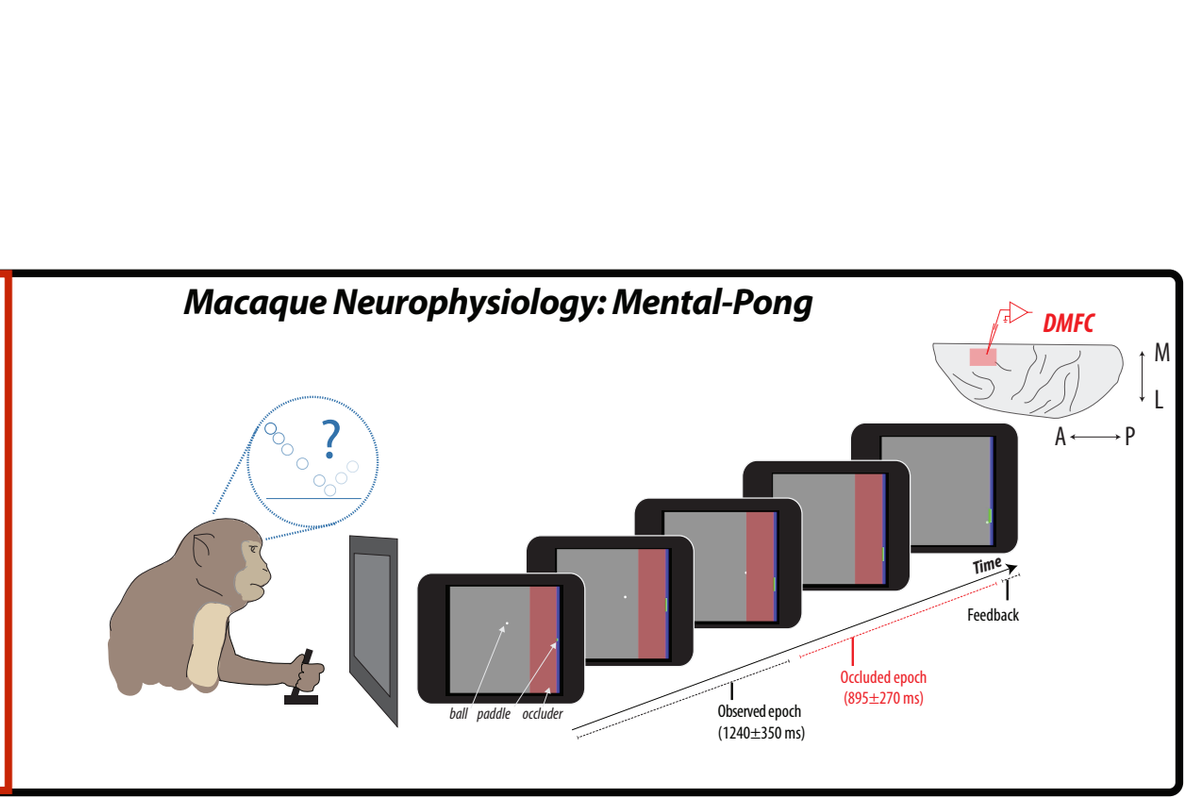




Human Behavior: Object Contact Prediction



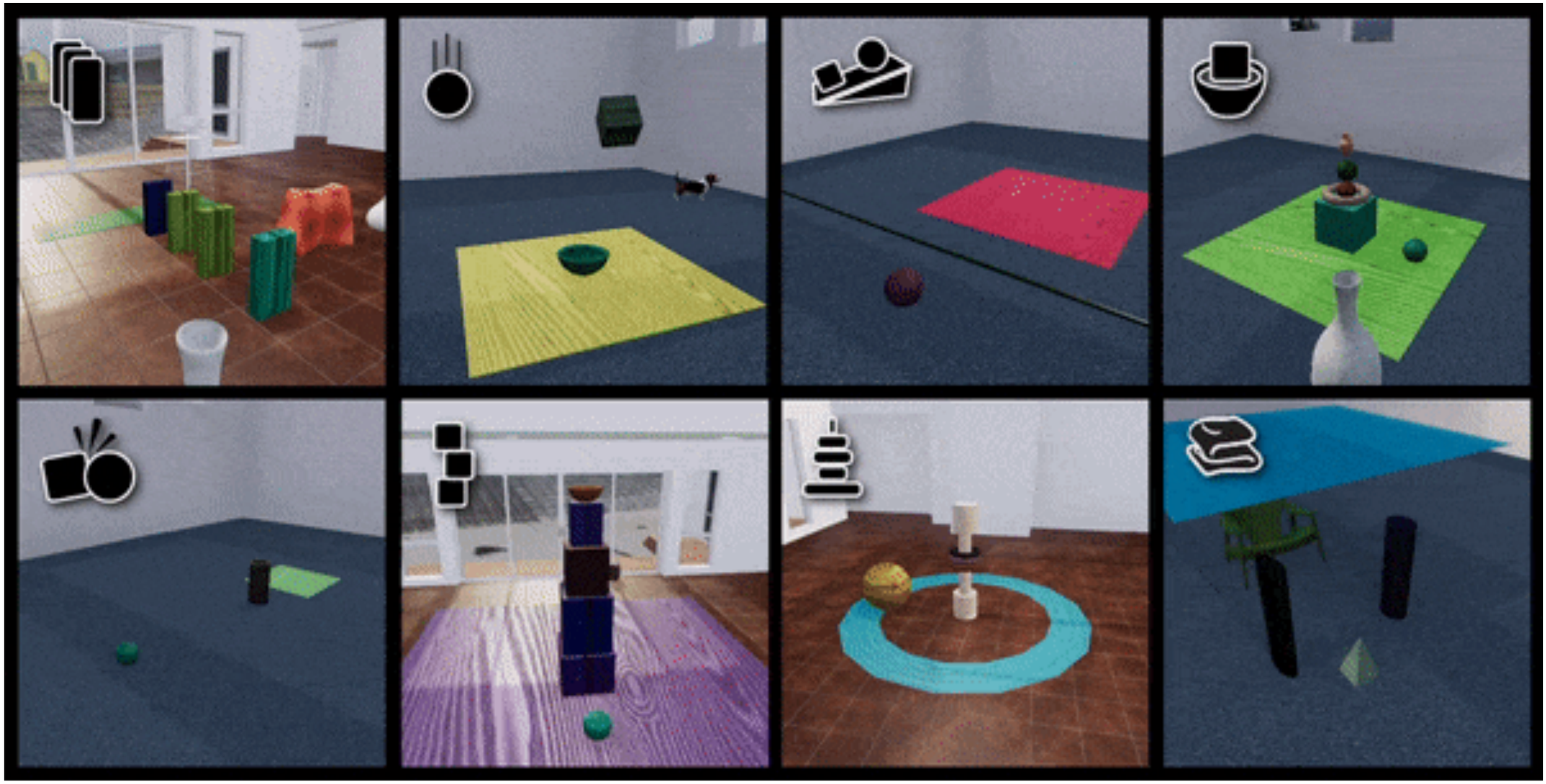




Object Contact Prediction Environment

Physion/ThreeD World (TDW)

Bear et al. 2021



Focus on everyday physical understanding


Daniel Bear


Joshua Tenenbaum

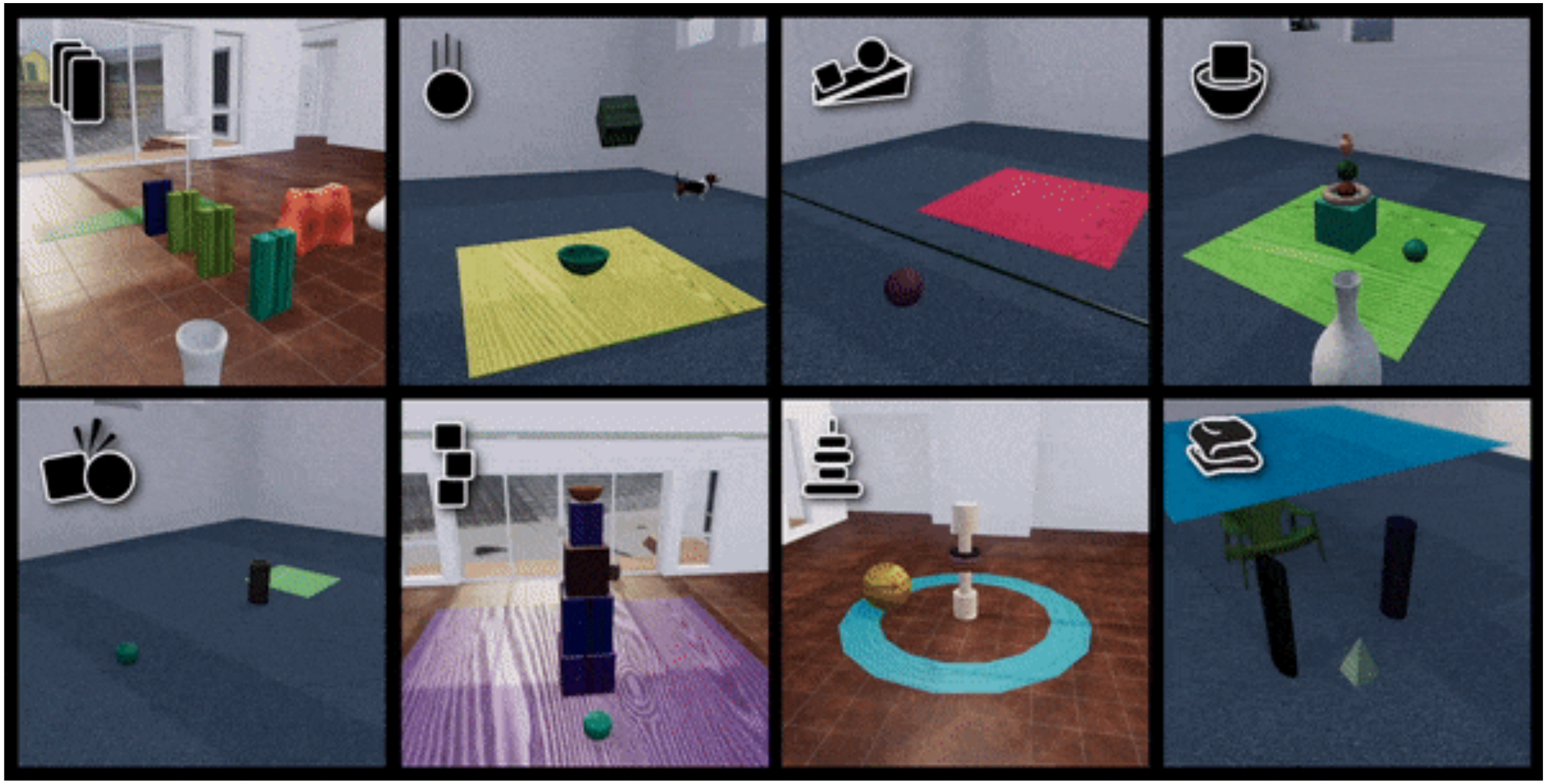

Daniel Yamins


Judith Fan

Object Contact Prediction Environment

Physion/ThreeD World (TDW)

Bear et al. 2021



Focus on everyday physical understanding


Daniel Bear


Joshua Tenenbaum


Daniel Yamins

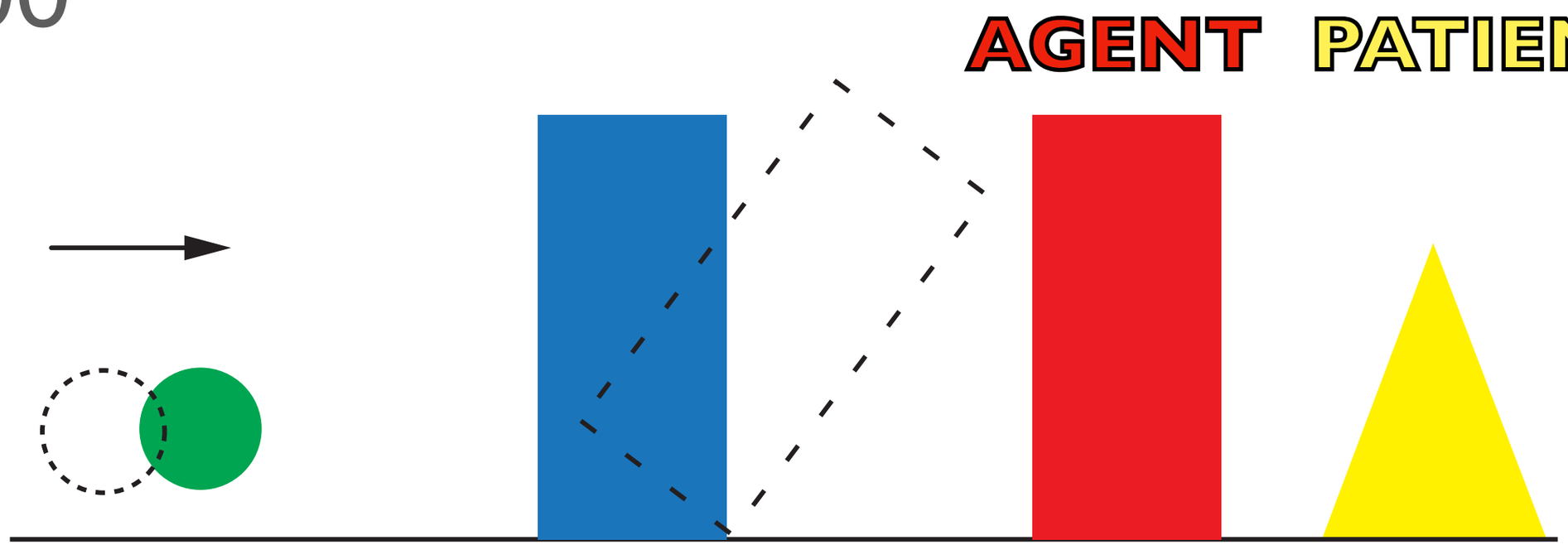

Judith Fan

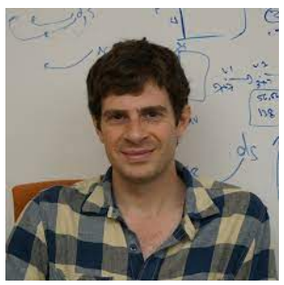
Human Behavior: Object Contact Prediction

Bear et al. 2021

n=100

"Will the agent object contact the patient object?"




Daniel Bear

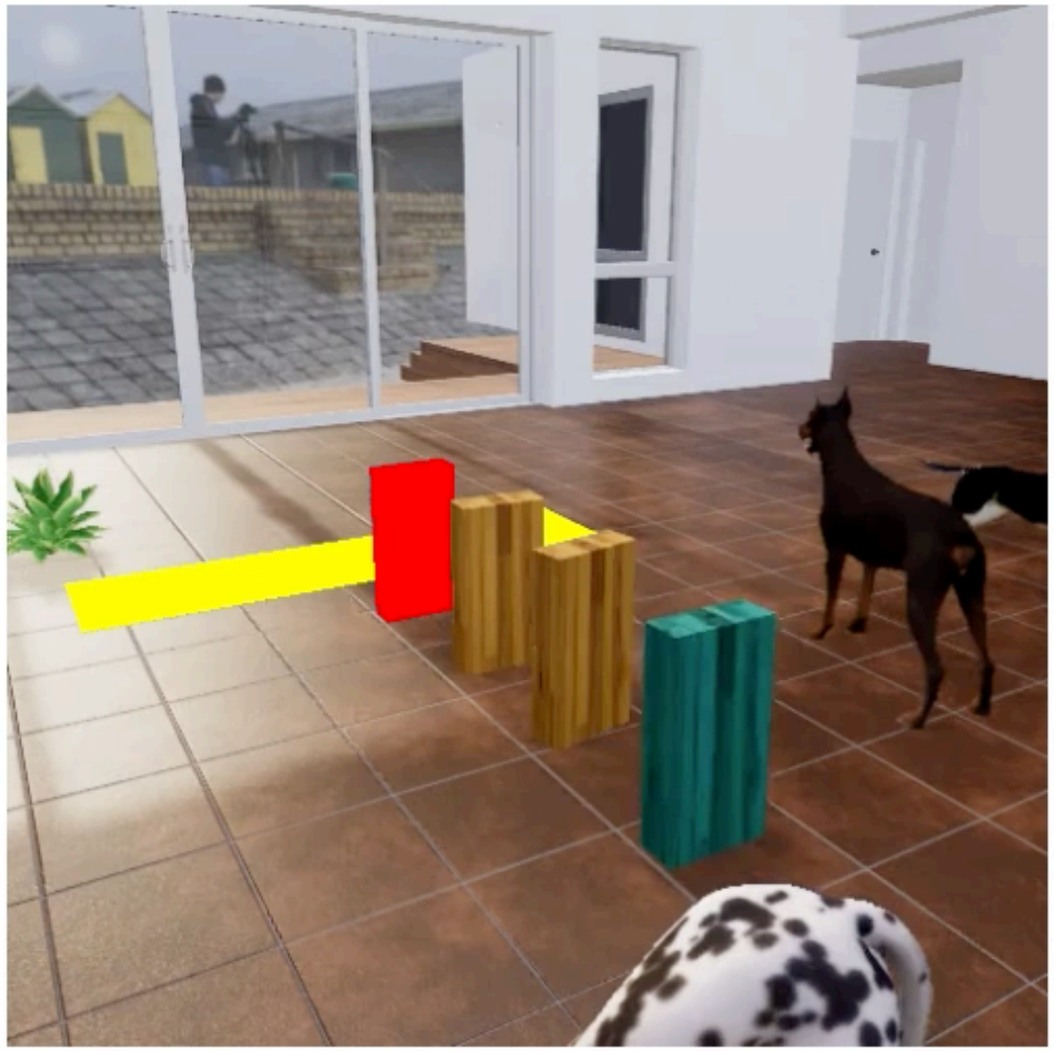

Joshua Tenenbaum


Daniel Yamins


Judith Fan

Completion Progress

Bear et al. 2021

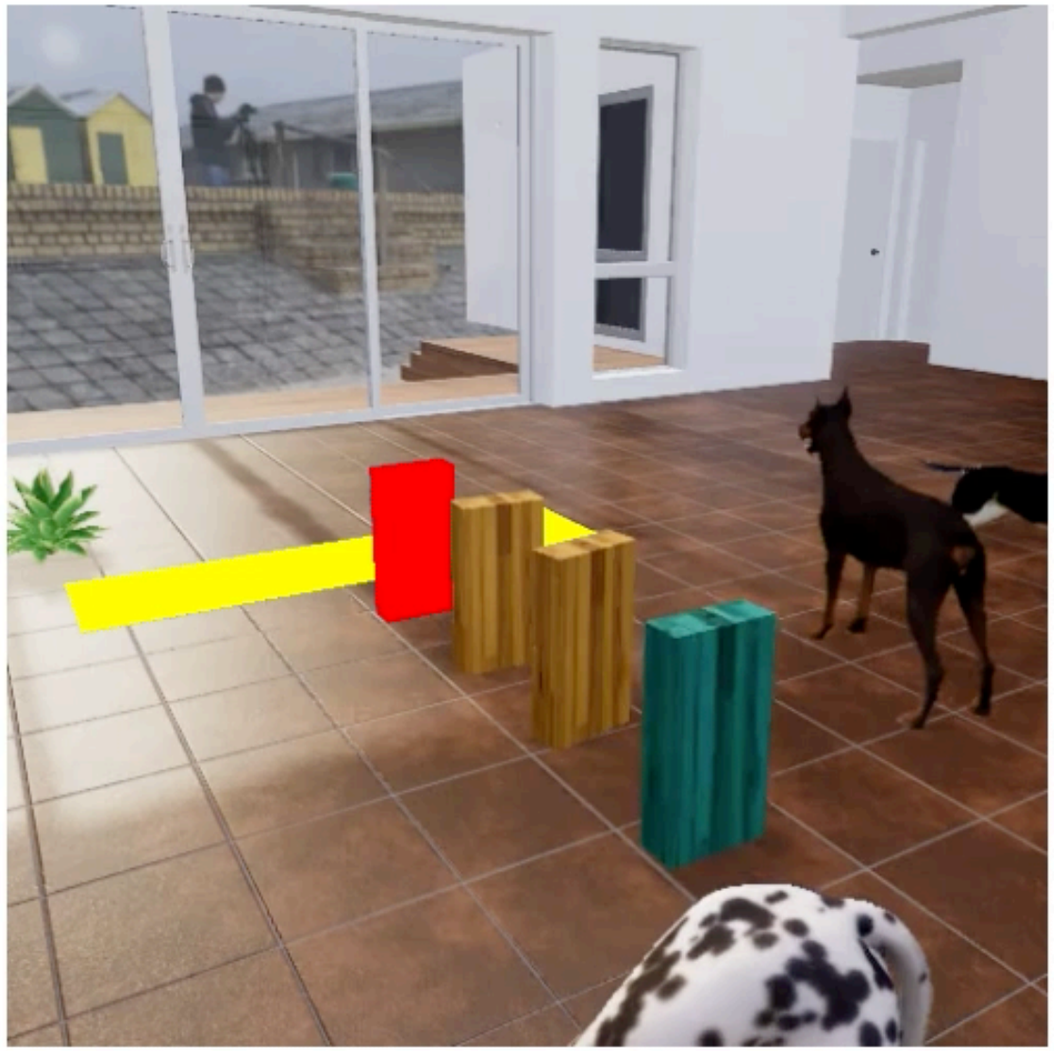


YES NO

Is the red object going to hit the yellow area?

Completion Progress

Bear et al. 2021



YES NO

Is the red object going to hit the yellow area?

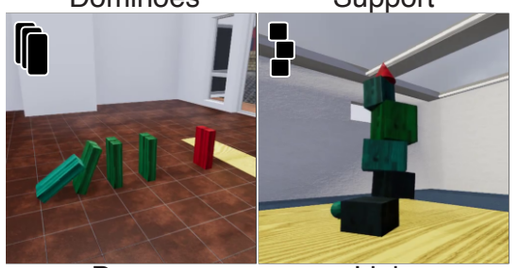
Model Evaluations: Both Metrics

(A) *Model Pretraining*

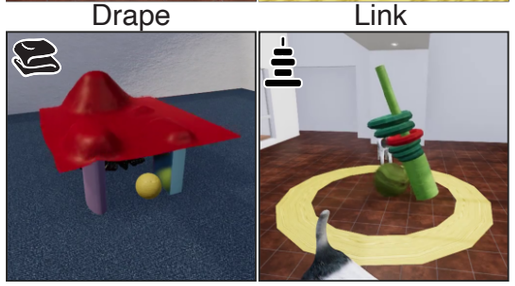
Inputs

Physion

Dominoes | Support



Drape | Link



Sensory-Cognitive Hypothesis Classes

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Pretraining Stage

T

T+1

T+1

Prediction

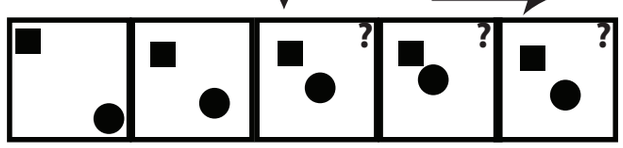
Ground Truth

End-to-End Future Prediction:

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

Encoder ***Object-slot***

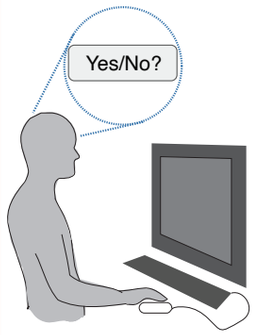
Time



Observed + Simulated

(B) *Model Evaluations*

1. *Human Behavior: Physion Object Contact Prediction (OCP)*



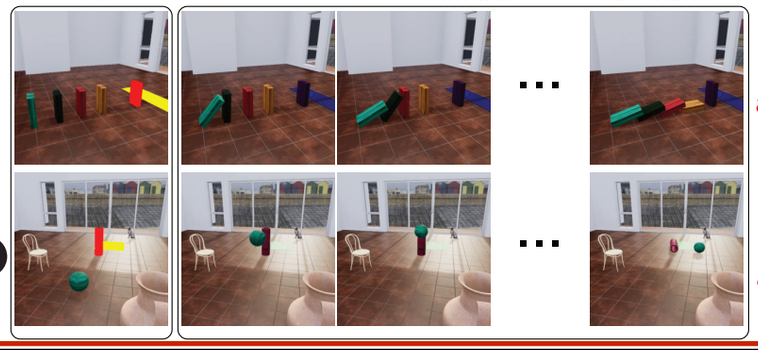
Observed Stimuli — Unobserved Outcome

Time

cue — stimulus — last frame

true label

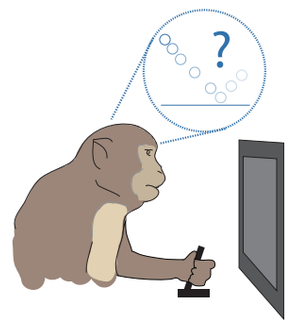
Example Scenarios

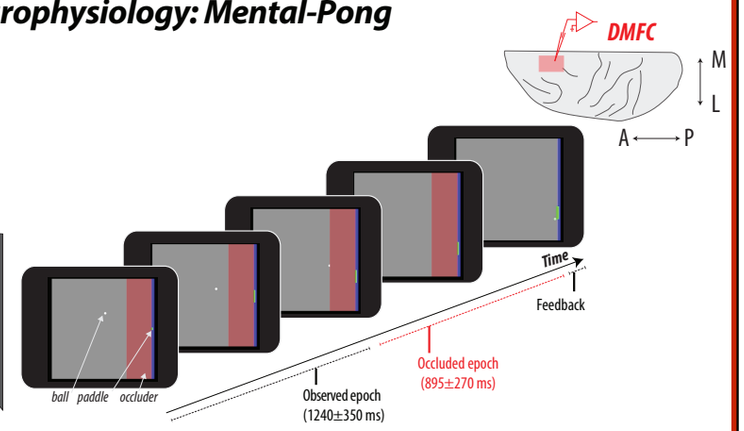


NO
acc. = 0.89

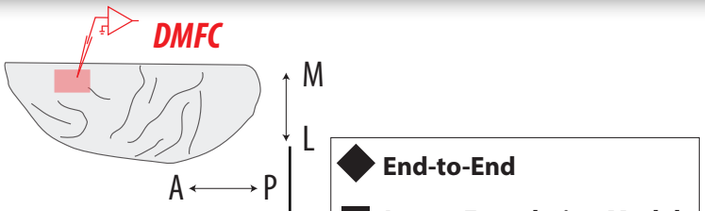
YES
acc. = 0.96

2. *Macaque Neurophysiology: Mental-Pong*





Dynamically-Equipped Video Foundation Models Can Match Both



End-to-End
Image Foundation Models
Video Foundation Models

Better Models

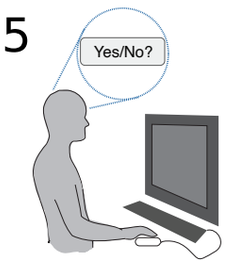
VC-1+LSTM

VC-1+CTRNN

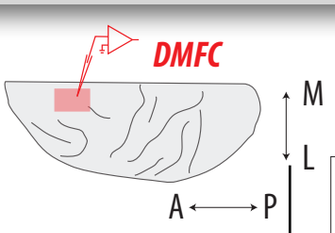
Neural Predictivity (Pearson's R)

0.0
0.1
0.2
0.3
0.4
0.5

Correlation to Average Human Response (Pearson's R)



Dynamically-Equipped Video Foundation Models Can Match Both



End-to-End
Image Foundation Models
Video Foundation Models

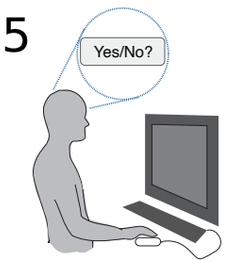
Better Models

VC-1+LSTM
VC-1+CTRNN

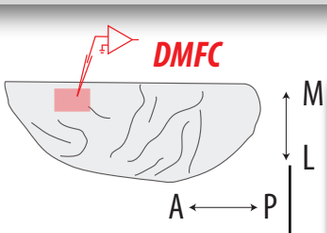
Neural Predictivity (Pearson's R)

0.0
0.1
0.2
0.3
0.4
0.5

Correlation to Average Human Response (Pearson's R)



Dynamically-Equipped Video Foundation Models Can Match Both



Exposed to the largest variety of egocentric video sources & transfers best across the widest range of embodied tasks.

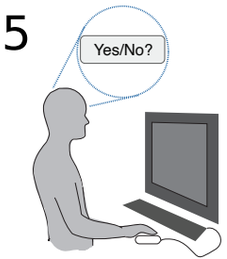
VC-1+LSTM

VC-1+CTRNN

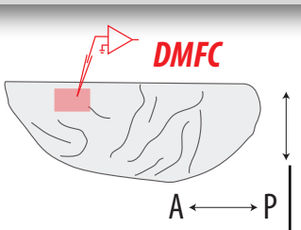
Neural Predictivity (Pearson's R)

0.0 0.1 0.2 0.3 0.4 0.5

Correlation to Average Human Response (Pearson's R)



Dynamically-Equipped Video Foundation Models Can Match Both



End-to-End
Image Foundation Models
Video Foundation Models

Better Models

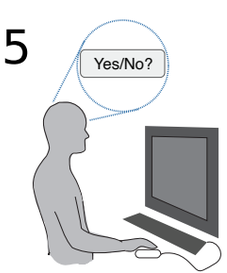
VC-1+LSTM
VC-1+CTRNN

But we have a ways to go to reach the NeuroAI Turing Test here!

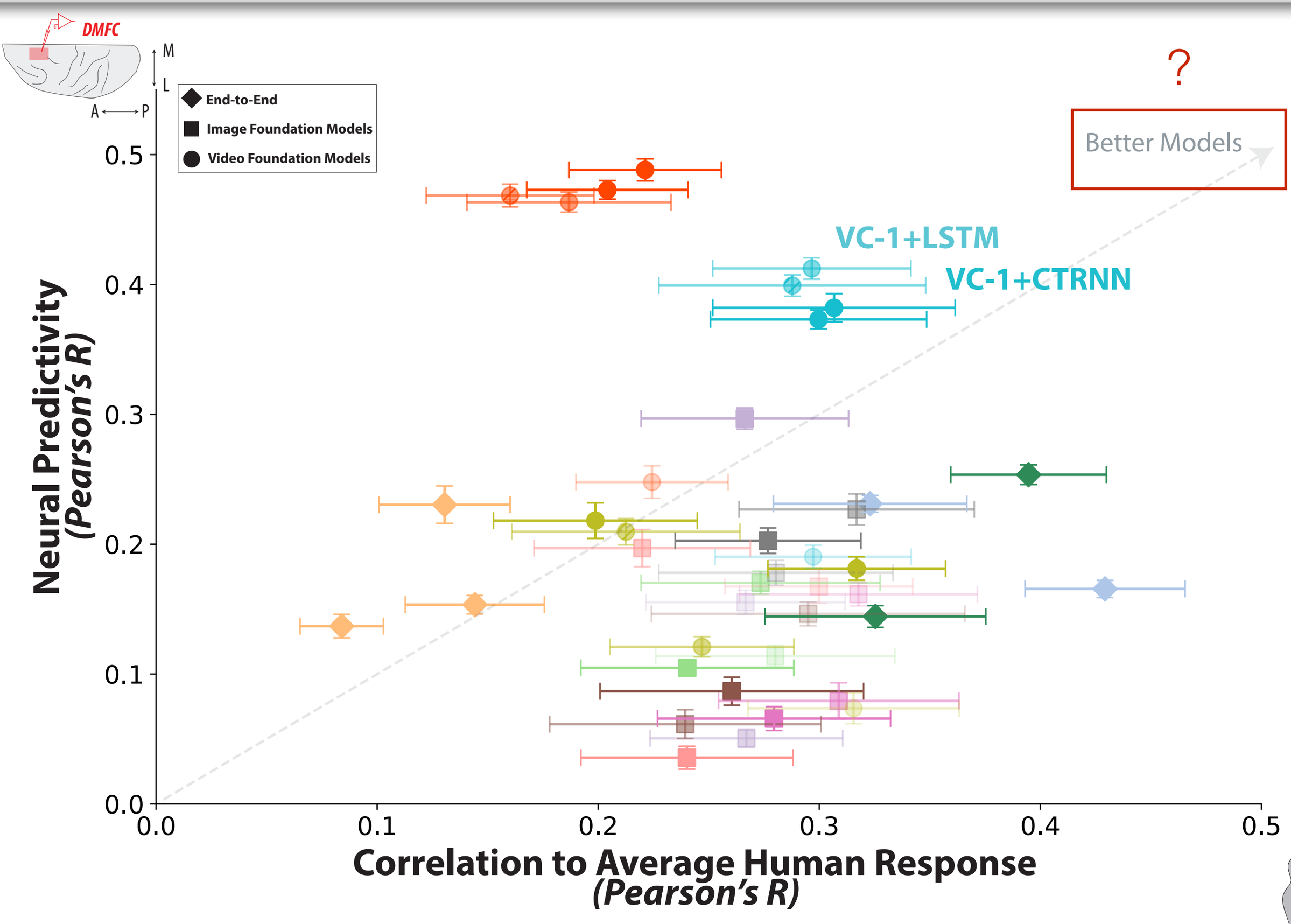
Neural Predictivity (Pearson's R)

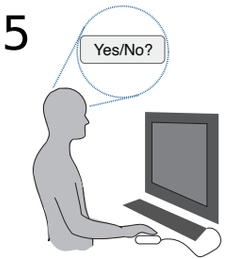
0.0
0.1
0.2
0.3
0.4
0.5

Correlation to Average Human Response (Pearson's R)



Future Directions





Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System

Mouse visual cortex (so far) is a low-acuity, shallow network that makes best use of the mouse's limited resources to create a general-purpose visual system, that can be deployed in novel environments and <u>embodied</u> contexts.

- Reusable Latent Representations for Primate Mental Simulation

- Heuristics for Interrogating Natural Intelligence

Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System

Mouse visual cortex (so far) is a low-acuity, shallow network that makes best use of the mouse's limited resources to create a general-purpose visual system, that can be deployed in novel environments and <u>embodied</u> contexts.

- Reusable Latent Representations for Primate Mental Simulation

Mental simulation crucially relies on explicit future prediction of a "factorized description" of visual scenes, where this "factorized description" is strongly constrained and must enable a wide range of dynamic <u>embodied</u> abilities.

- Heuristics for Interrogating Natural Intelligence

Outline

- Mouse Visual Cortex as a Task-General, Limited Resource System

Mouse visual cortex (so far) is a low-acuity, shallow network that makes best use of the mouse's limited resources to create a general-purpose visual system, that can be deployed in novel environments and <u>embodied</u> contexts.

- Reusable Latent Representations for Primate Mental Simulation

Mental simulation crucially relies on explicit future prediction of a "factorized description" of visual scenes, where this "factorized description" is strongly constrained and must enable a wide range of dynamic <u>embodied</u> abilities.

- Heuristics for Interrogating Natural Intelligence

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**

- **<u>Incorporating AI Insights:</u>**

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**
  - **Connectomics:**
  - **Ethology:**
- **<u>Incorporating AI Insights:</u>**

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**

  - **Connectomics:** Not usually a 1-1 mapping from a connectome to a functional model, and easy to get wrong. Rather, the best model often requires an *iterative balance* of functional optimization with macroscale structural constraints (e.g. shallow vs. deep cortex).

  - **Ethology:**

- **<u>Incorporating AI Insights:</u>**

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**

  - **Connectomics:** Not usually a 1-1 mapping from a connectome to a functional model, and easy to get wrong. Rather, the best model often requires an *iterative balance* of functional optimization with macroscale structural constraints (e.g. shallow vs. deep cortex).

  - **Ethology:** Ethology alone does not always give us the "correct" functional optimization target (e.g. for intermediate brain areas). Can be helpful to consider that the brain area is embodied in a larger cognitive agent.

- **<u>Incorporating AI Insights:</u>**

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**

  - **Connectomics:** Not usually a 1-1 mapping from a connectome to a functional model, and easy to get wrong. Rather, the best model often requires an *iterative balance* of functional optimization with macroscale structural constraints (e.g. shallow vs. deep cortex).

  - **Ethology:** Ethology alone does not always give us the "correct" functional optimization target (e.g. for intermediate brain areas). Can be helpful to consider that the brain area is embodied in a larger cognitive agent.

- **<u>Incorporating AI Insights:</u>**

  - End-to-end reinforcement learning (RL) does *not* seem to give us neurally-aligned visual systems in both rodents and primates.

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**

  - **Connectomics:** Not usually a 1-1 mapping from a connectome to a functional model, and easy to get wrong. Rather, the best model often requires an *iterative balance* of functional optimization with macroscale structural constraints (e.g. shallow vs. deep cortex).

  - **Ethology:** Ethology alone does not always give us the "correct" functional optimization target (e.g. for intermediate brain areas). Can be helpful to consider that the brain area is embodied in a larger cognitive agent.

- **<u>Incorporating AI Insights:</u>**

  - End-to-end reinforcement learning (RL) does *not* seem to give us neurally-aligned visual systems in both rodents and primates.

  - Suggests a possible functional modularization of optimization targets, with reusable SSL representations best matching visual areas overall.

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**

  - **Connectomics:** Not usually a 1-1 mapping from a connectome to a functional model, and easy to get wrong. Rather, the best model often requires an *iterative balance* of functional optimization with macroscale structural constraints (e.g. shallow vs. deep cortex).

  - **Ethology:** Ethology alone does not always give us the "correct" functional optimization target (e.g. for intermediate brain areas). Can be helpful to consider that the brain area is embodied in a larger cognitive agent.

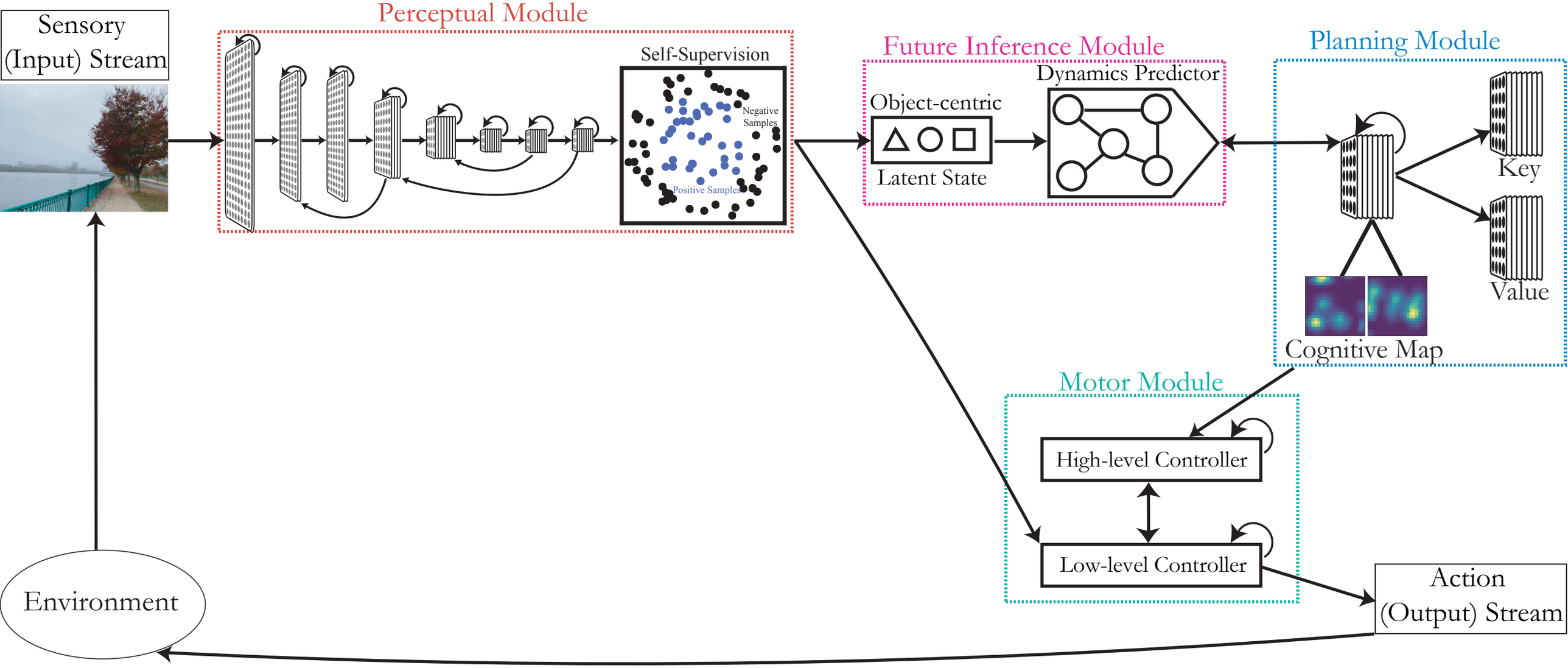
- **<u>Incorporating AI Insights:</u>**

  - End-to-end reinforcement learning (RL) does *not* seem to give us neurally-aligned visual systems in both rodents and primates.

  - Suggests a possible functional modularization of optimization targets, with reusable SSL representations best matching visual areas overall.

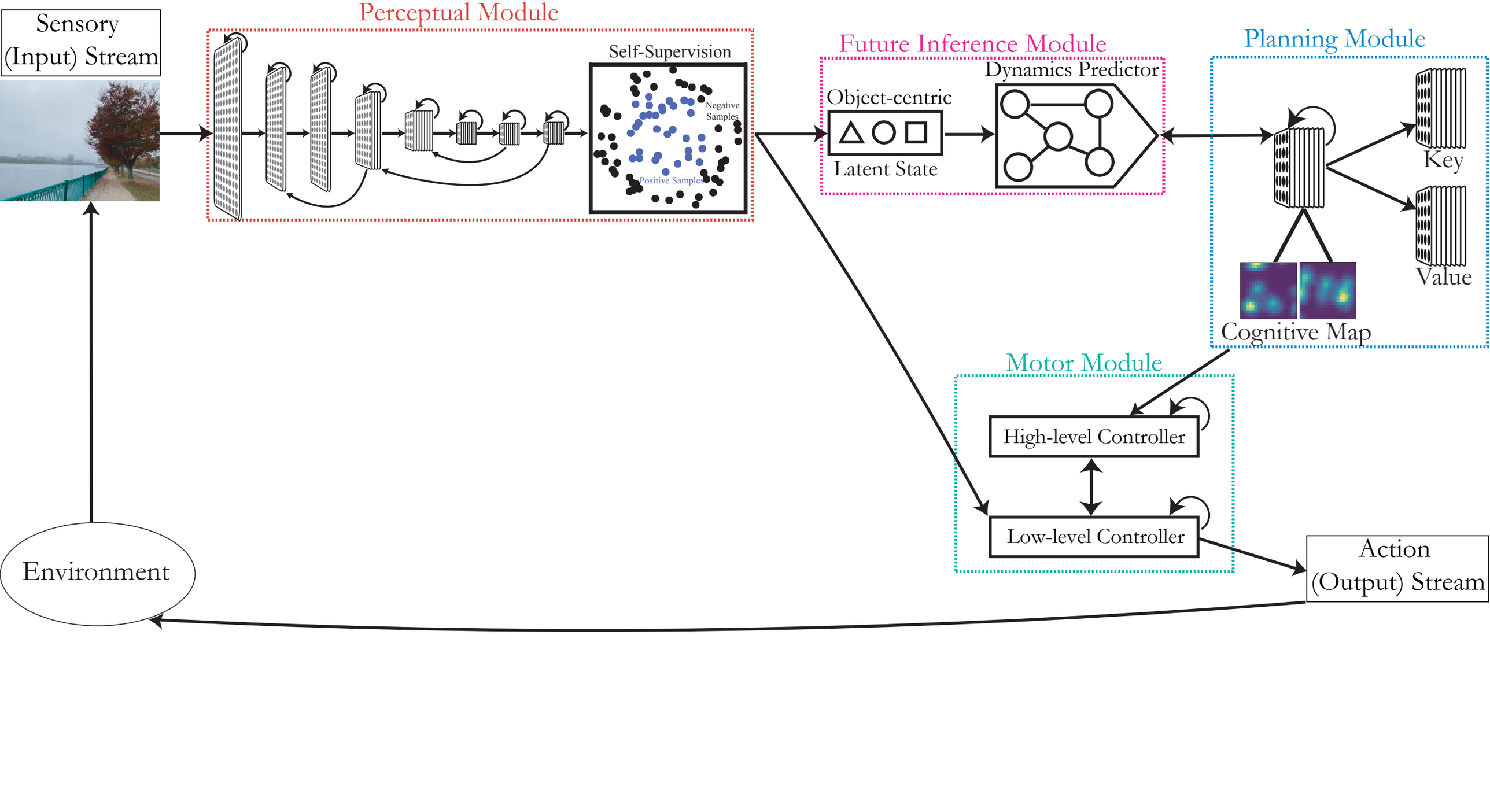
Next Steps: Modularized, Embodied Agents?

Next Steps: Modularized, Embodied Agents?



Next Steps: Modularized, Embodied Agents?

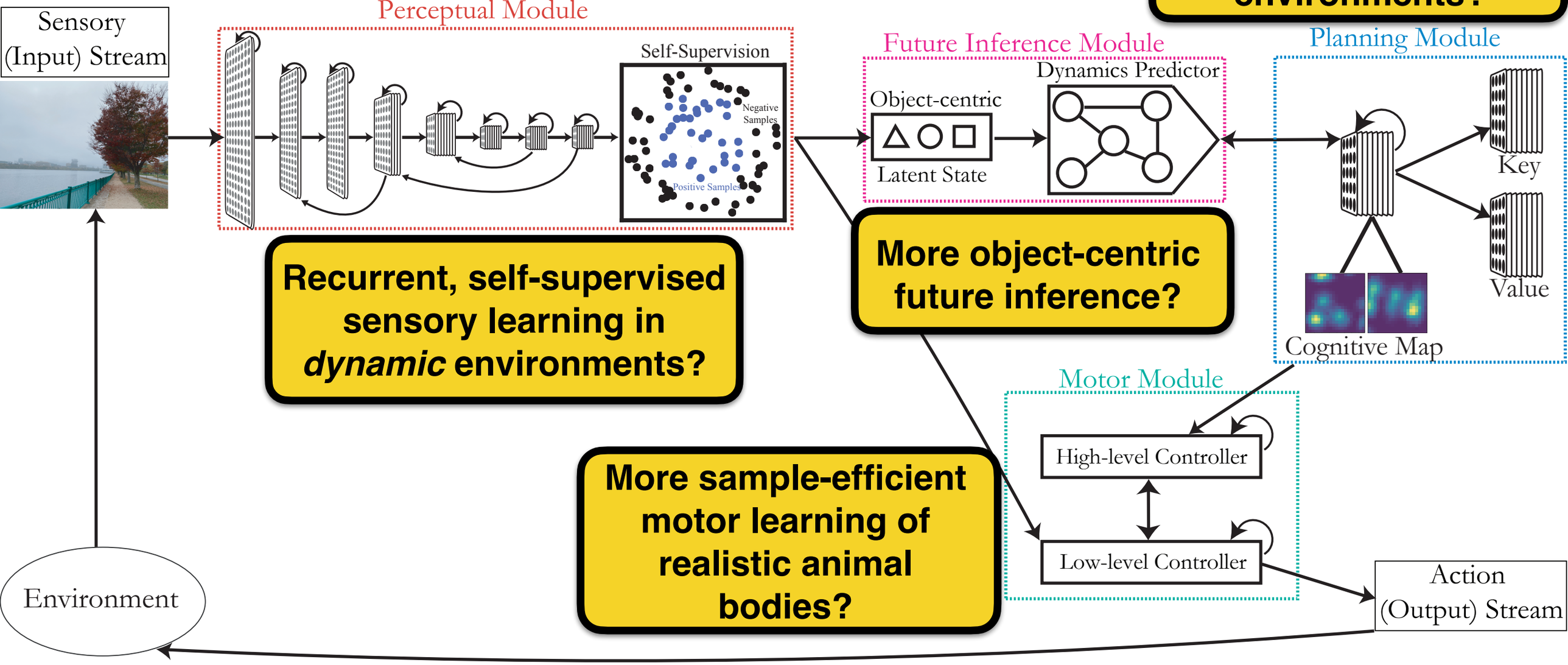
How does the brain *represent*, *predict*, *plan*, and enable *action*?



Next Steps: Modularized, Embodied Agents?

How does the brain *represent*, *predict*, *plan*, and enable *action*?

Longer-range planning in more complex environments?



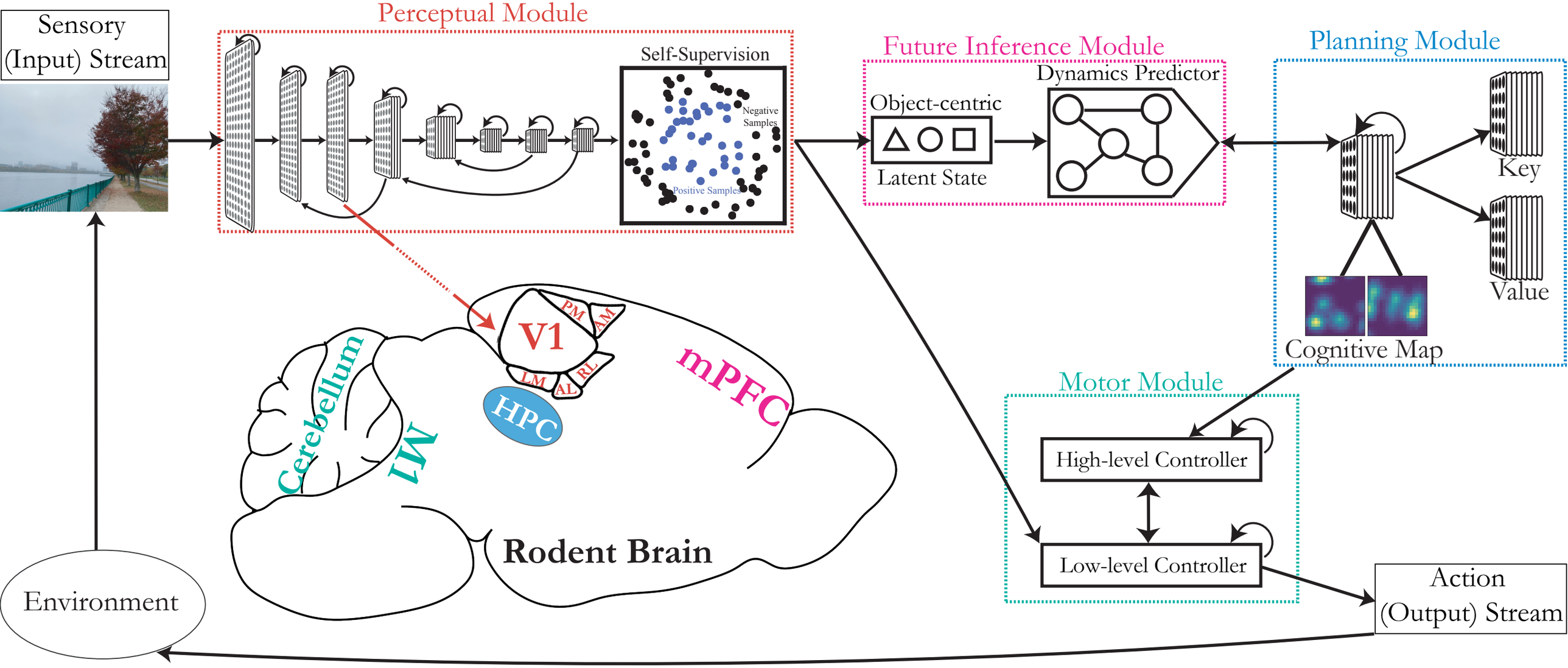
Recurrent, self-supervised sensory learning in *dynamic* environments?

More object-centric future inference?

More sample-efficient motor learning of realistic animal bodies?

Next Steps: Modularized, Embodied Agents?

How does the brain *represent*, *predict*, *plan*, and enable *action*?



What Happens When We Get There?

What Happens When We Get There?

Barriers and Pathways to Human-AI Alignment: A Game-Theoretic Approach

ARAN NAYEBI, Carnegie Mellon University, USA

https://arxiv.org/abs/2502.05934

What Happens When We Get There?

Barriers and Pathways to Human-AI Alignment: A Game-Theoretic Approach

ARAN NAYEBI, Carnegie Mellon University, USA

https://arxiv.org/abs/2502.05934

PROPOSITION 2.6 (LOWER BOUND). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}_j^i\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(M N^2 \log(1/\varepsilon)\right)$ bits[6] to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

What Happens When We Get There?

Barriers and Pathways to Human-AI Alignment: A Game-Theoretic Approach

ARAN NAYEBI, Carnegie Mellon University, USA

https://arxiv.org/abs/2502.05934

PROPOSITION 2.6 (LOWER BOUND). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}_j^i\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(M N^2 \log(1/\varepsilon)\right)$ bits[6] to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

There will always be a task that is *impossible* to align efficiently over, even if the agents are computationally unbounded

What Happens When We Get There?

Barriers and Pathways to Human-AI Alignment: A Game-Theoretic Approach

ARAN NAYEBI, Carnegie Mellon University, USA

https://arxiv.org/abs/2502.05934

PROPOSITION 2.6 (LOWER BOUND). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}_j^i\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(M N^2 \log(1/\varepsilon)\right)$ bits[6] to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

There will always be a task that is *impossible* to align efficiently over, even if the agents are computationally unbounded

$$O\left(\frac{(1100)^{\frac{2097152}{(1/4)^6}}}{(1/2)^{\frac{256}{(1/4)^2}}}\right) = O\left(1.31 \times 10^{26125365467}\right)$$

If the agents are computationally *bounded*, this can currently take more time than the number of atoms in the universe! (~4.8 x 10^{79})

Acknowledgements


Nathan C.L. Kong


Chengxu Zhuang


Justin L. Gardner


Anthony M. Norcia


Daniel Yamins


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

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