Using Embodied AI to help answer "why" questions in systems neuroscience

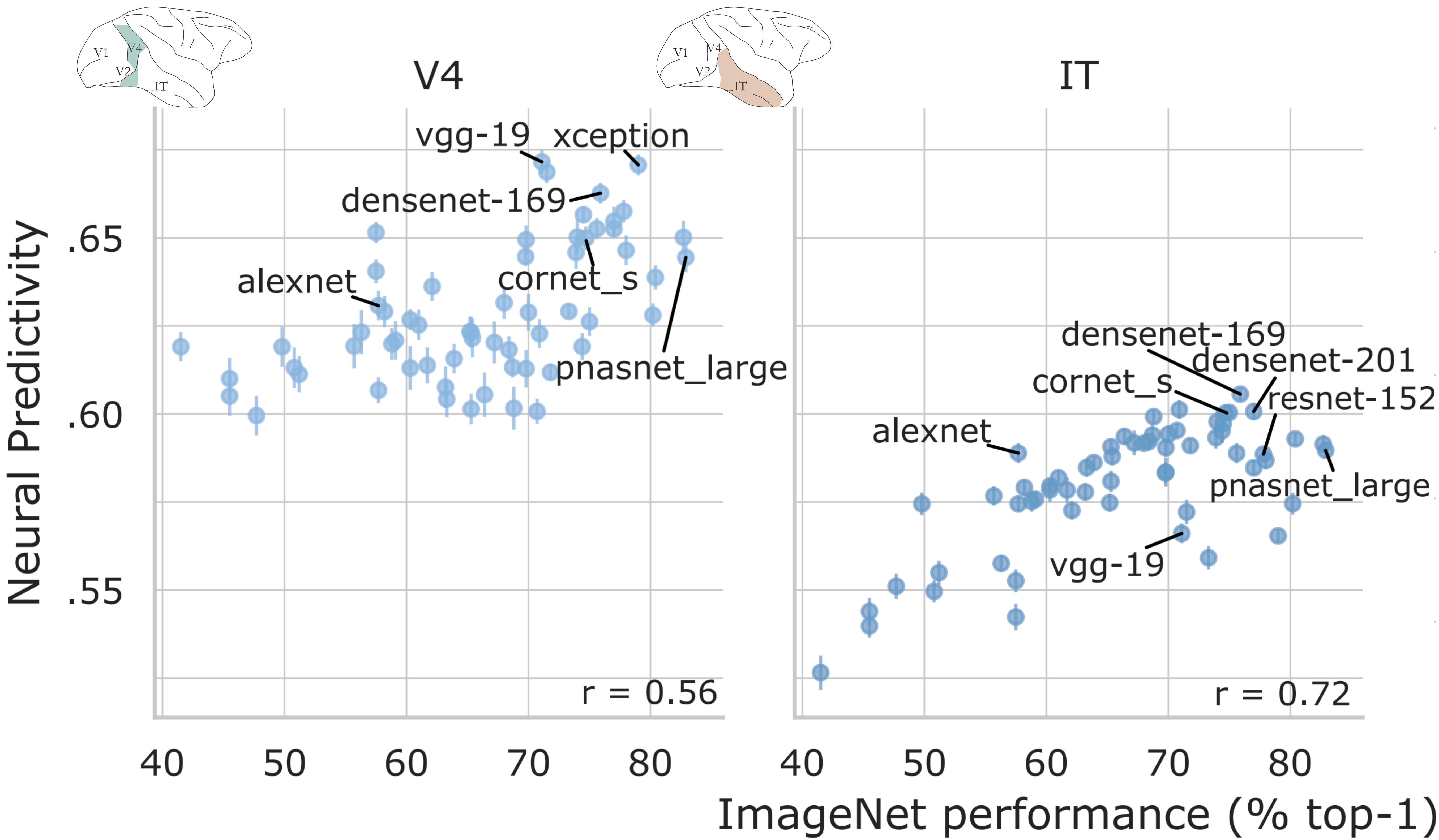
Aran Nayebi
K. Lisa Yang ICoN Postdoctoral Fellow
McGovern Institute

CBMM Research Meeting
2023.09.19

CENTER FOR
Brains
Minds+
Machines

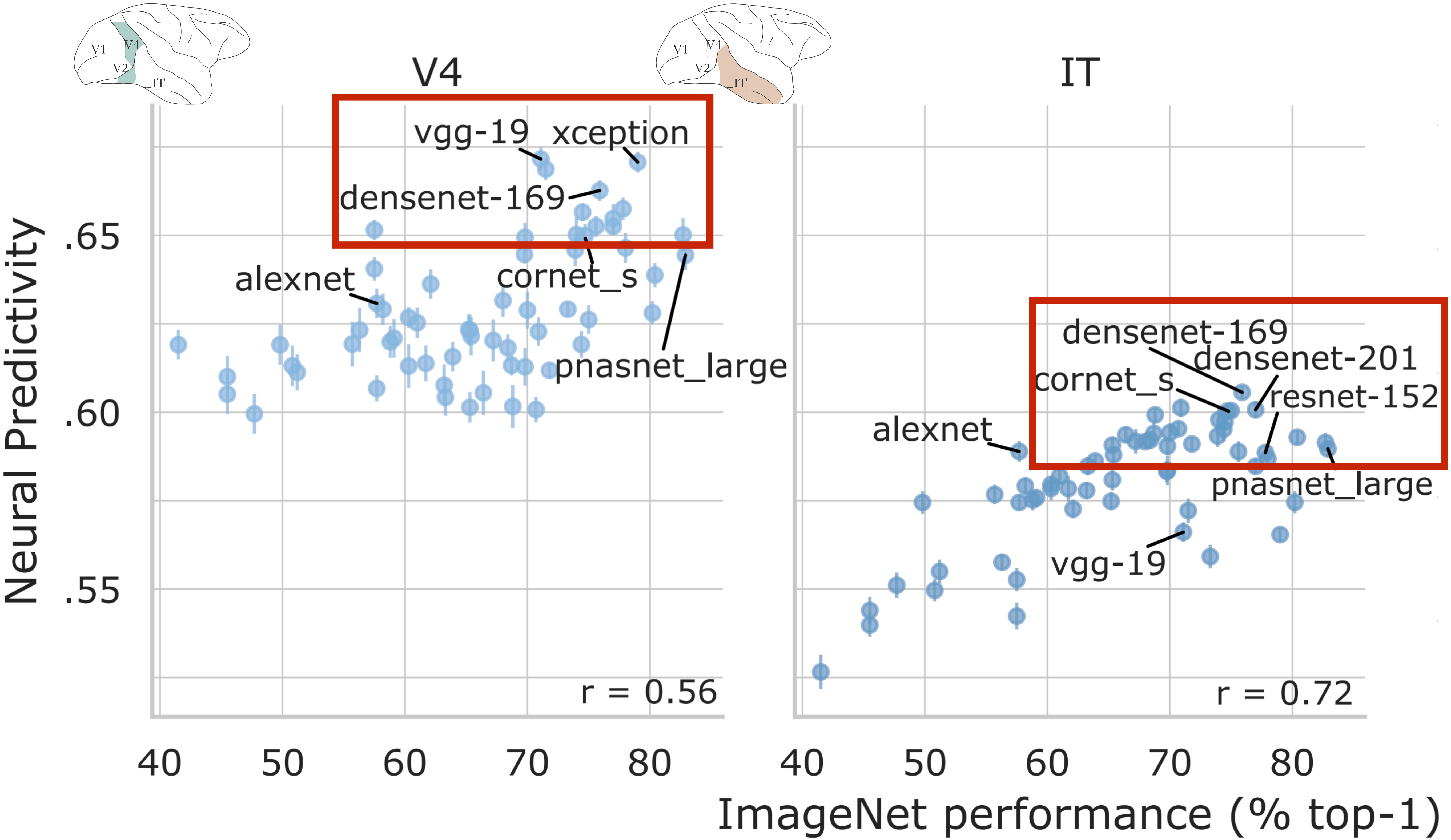
K. LISA YANG
ICoN CENTER
Integrative Computational Neuroscience

Similar predictivities among very different CNN architectures



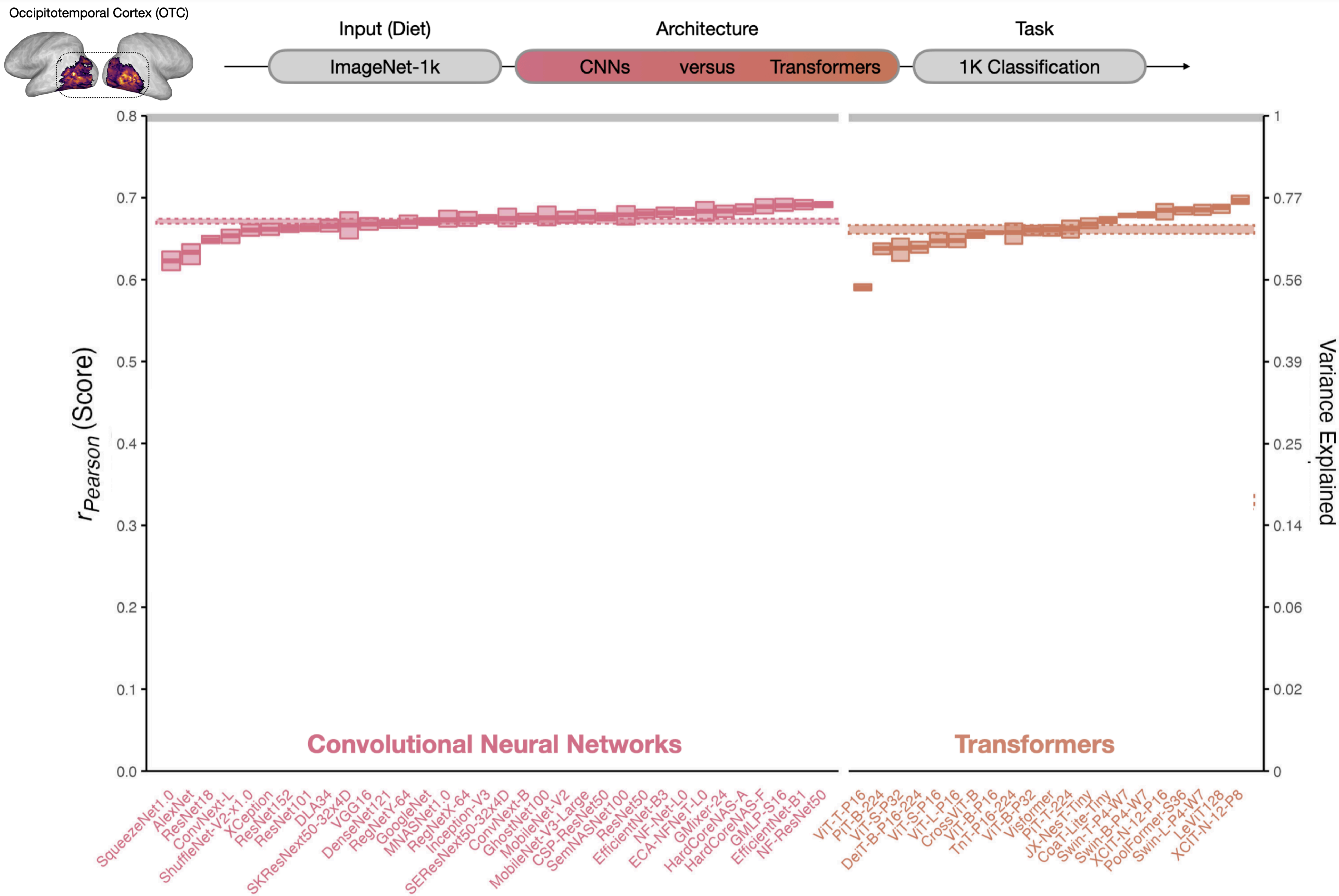
Schrimpf, Kubilius* et al. 2018*

Similar predictivities among very different CNN architectures



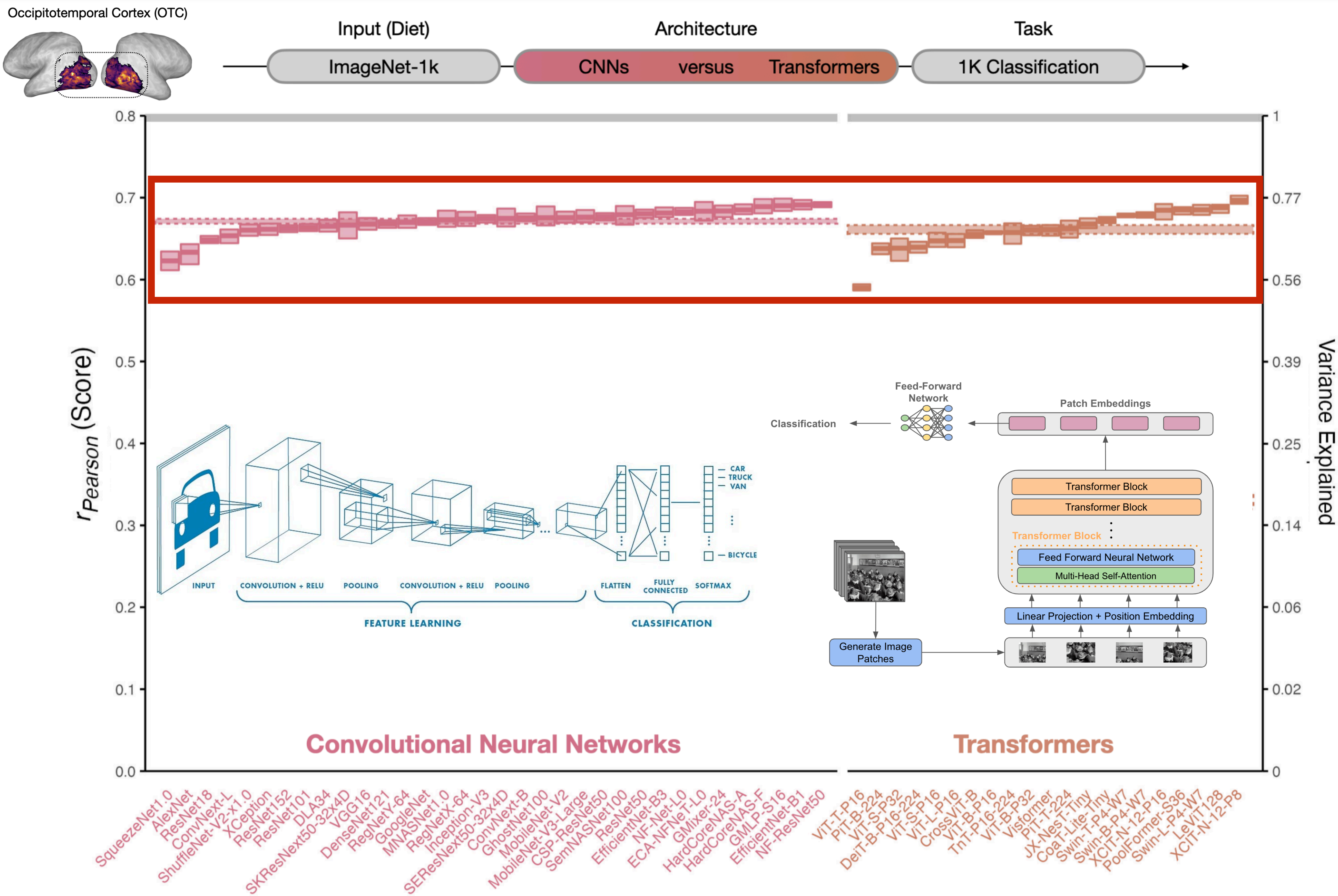
Schrimpf, Kubilius* et al. 2018*

Similar predictivities between CNNs vs. Transformers



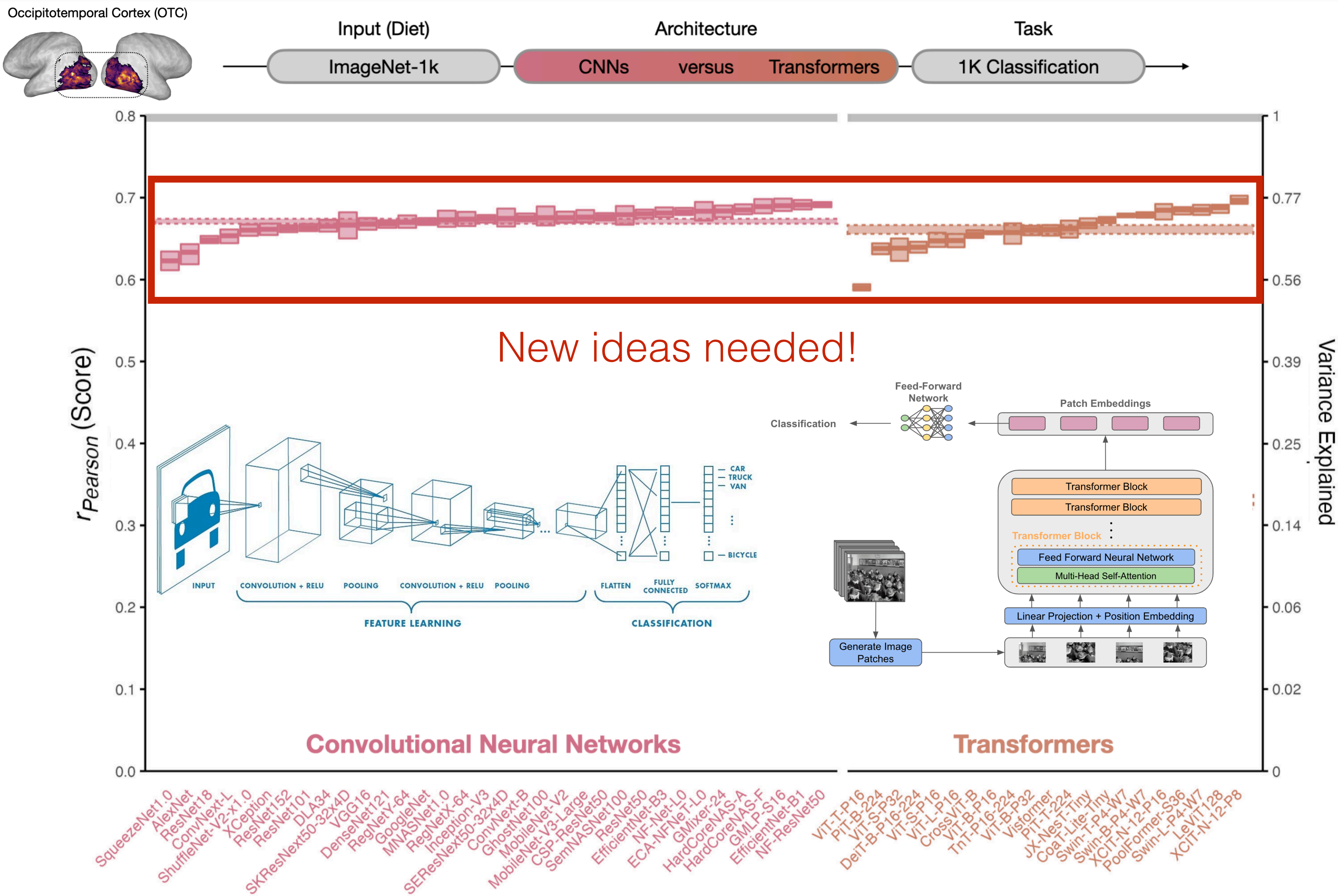
Conwell et al. 2023

Similar predictivities between CNNs vs. Transformers



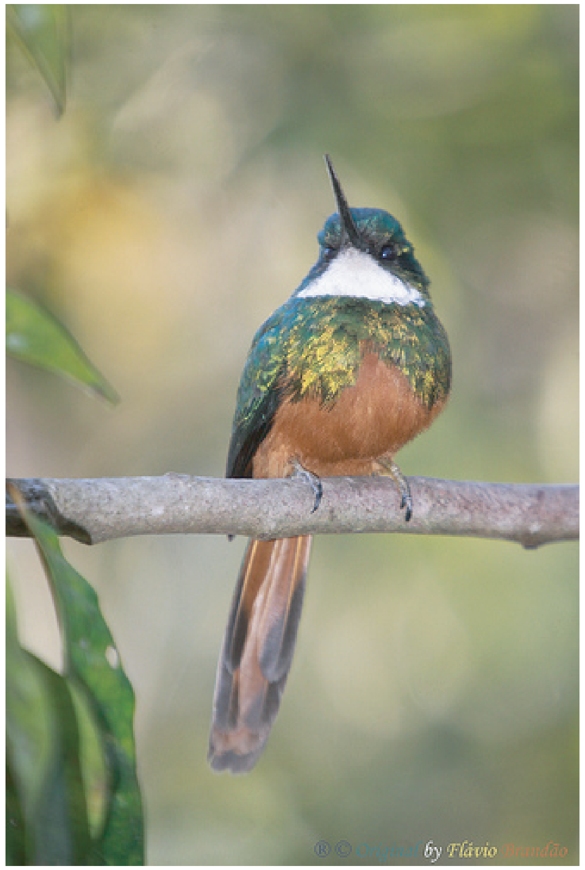
Conwell et al. 2023

Similar predictivities between CNNs vs. Transformers



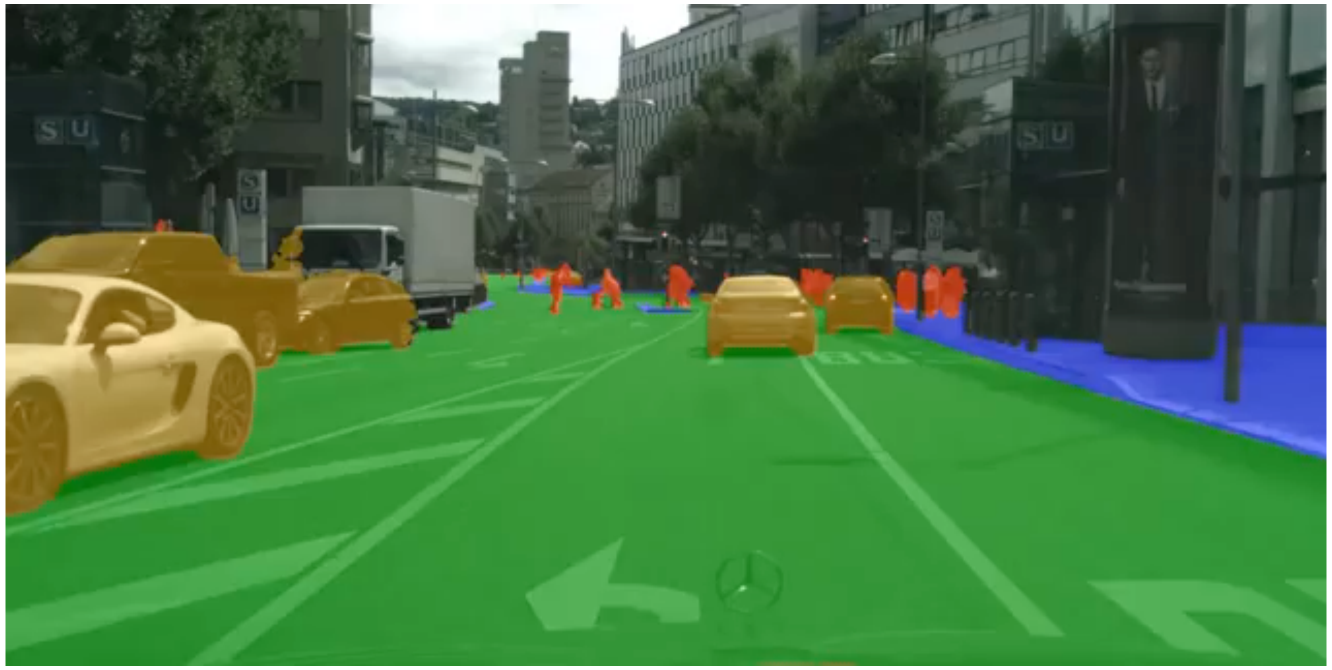
Conwell et al. 2023

We do a lot more than passive viewing…



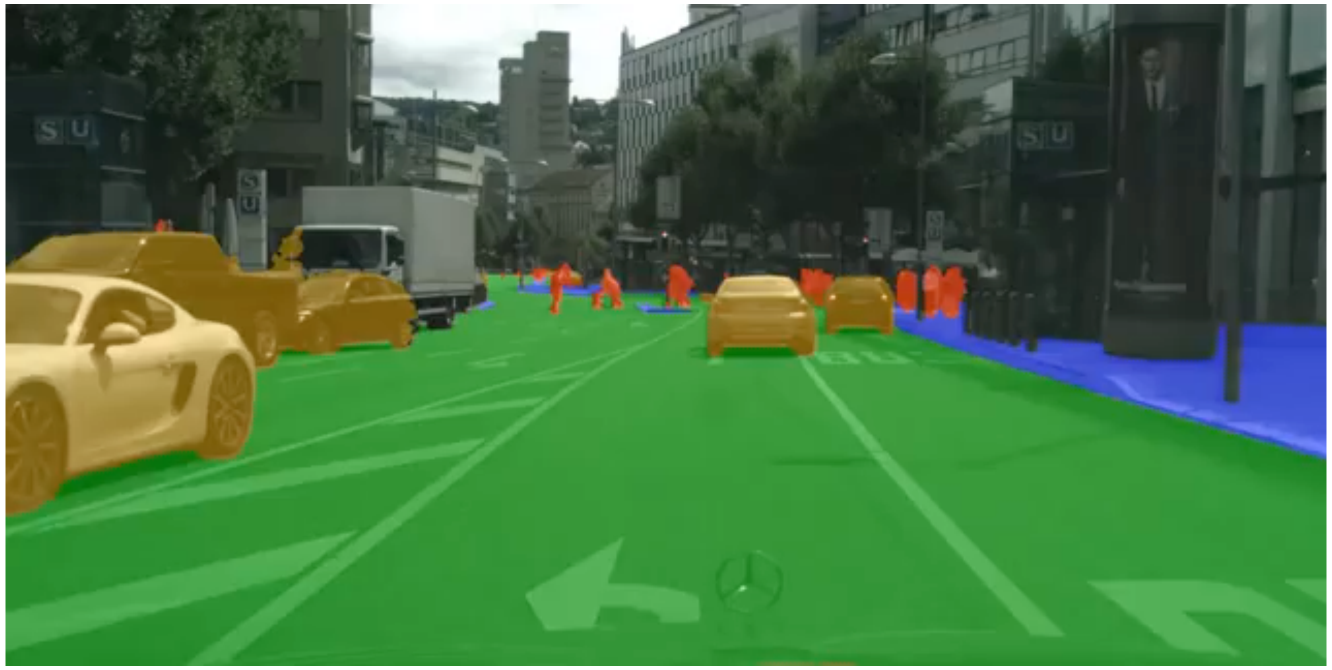
We do a lot more than passive viewing…

Scene Understanding



We do a lot more than passive viewing…

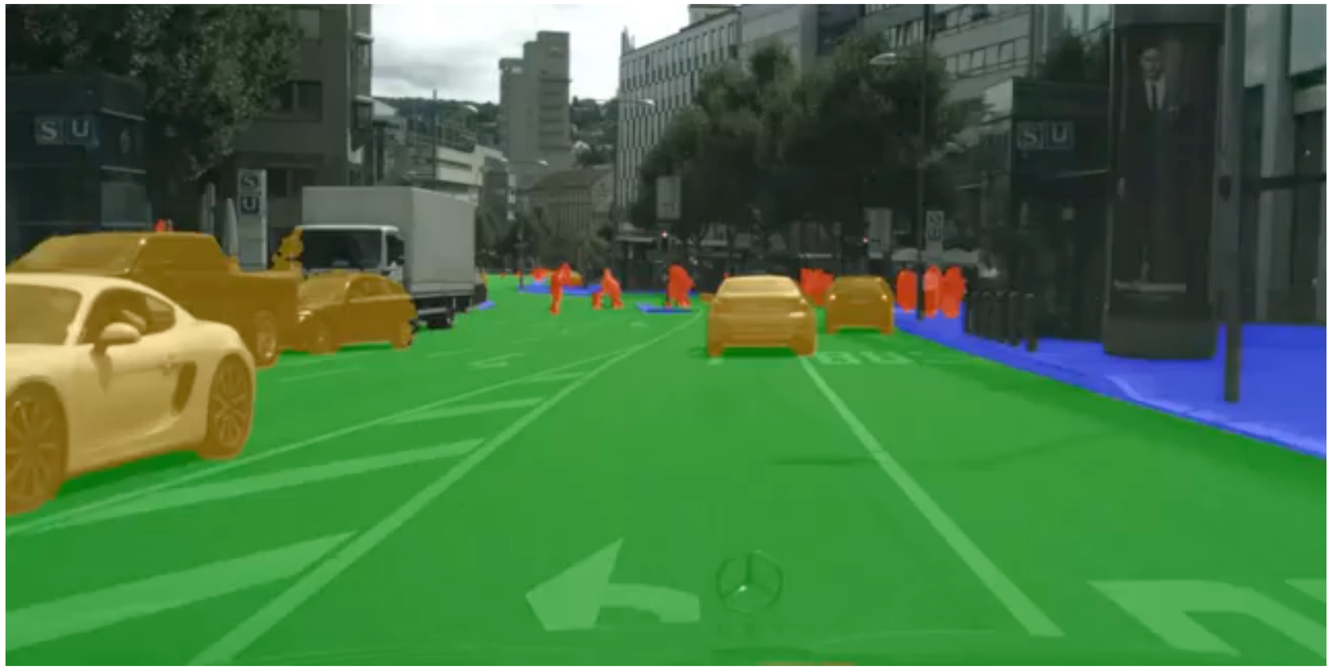
Scene Understanding



Multi-Step Planning

We do a lot more than passive viewing…

Scene Understanding

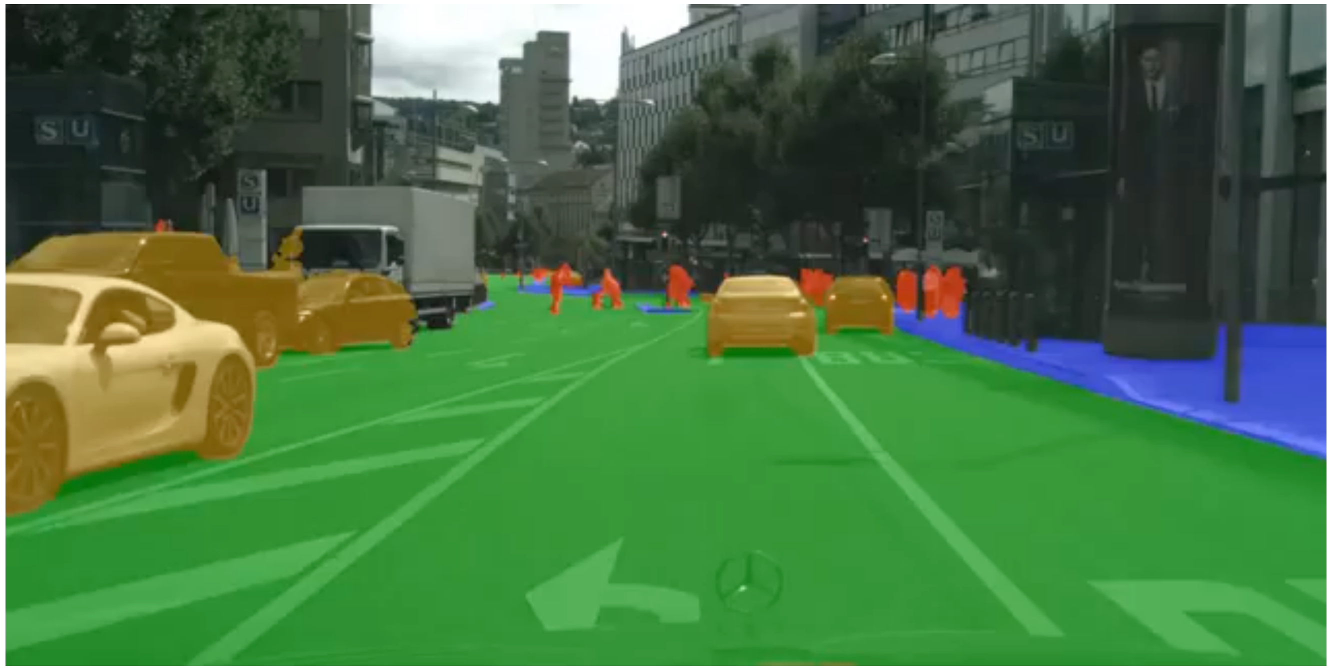


Multi-Step Planning

Navigation

We do a lot more than passive viewing…

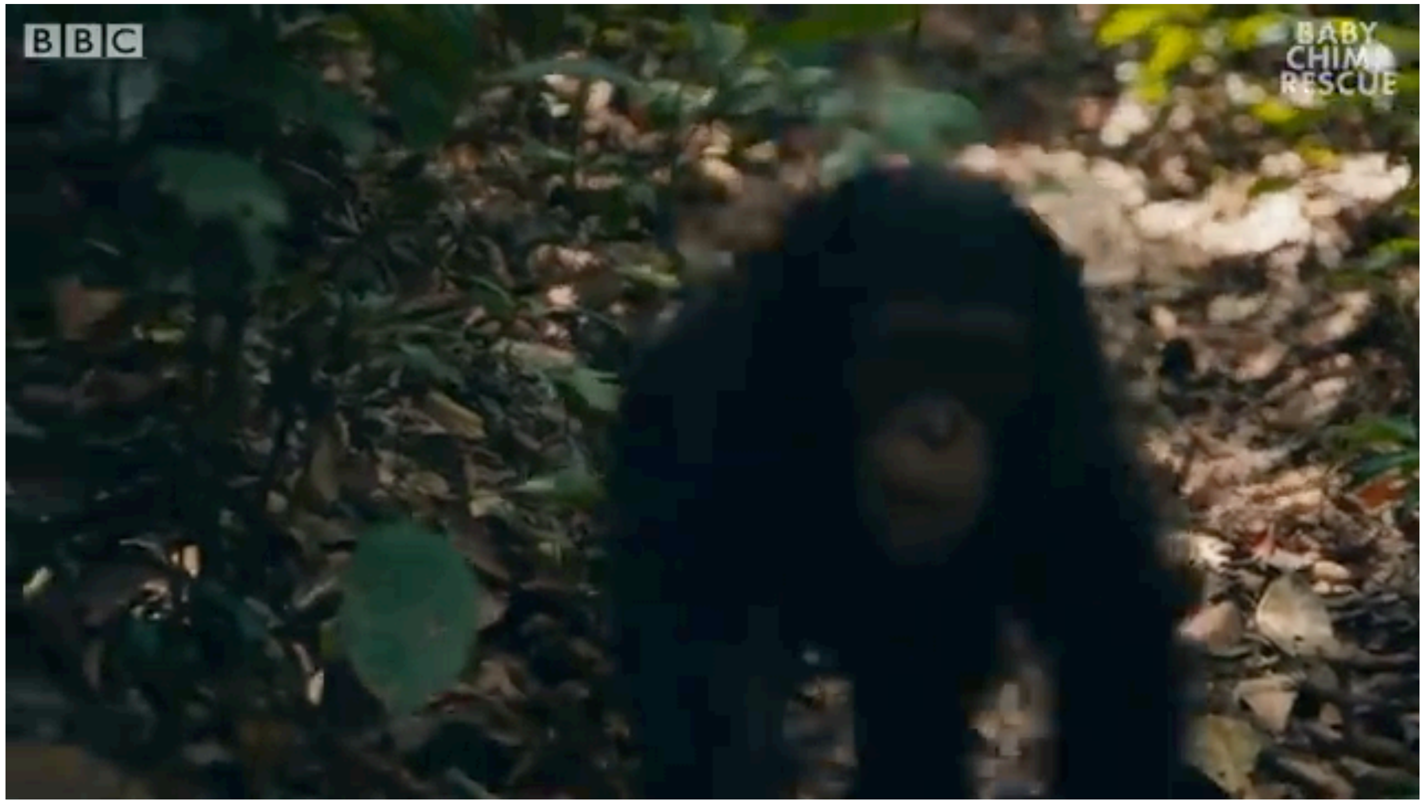
Scene Understanding



Multi-Step Planning

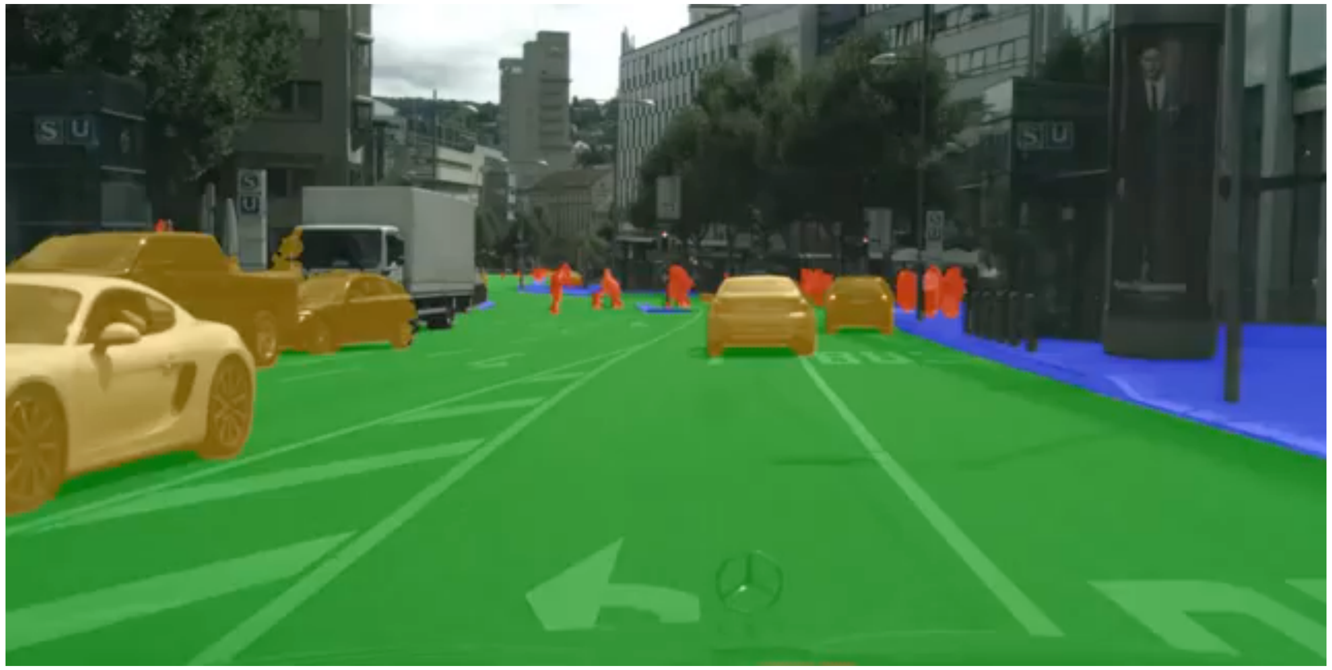
Navigation

Flexible Embodiment



From Neurons to Behavior

Scene Understanding

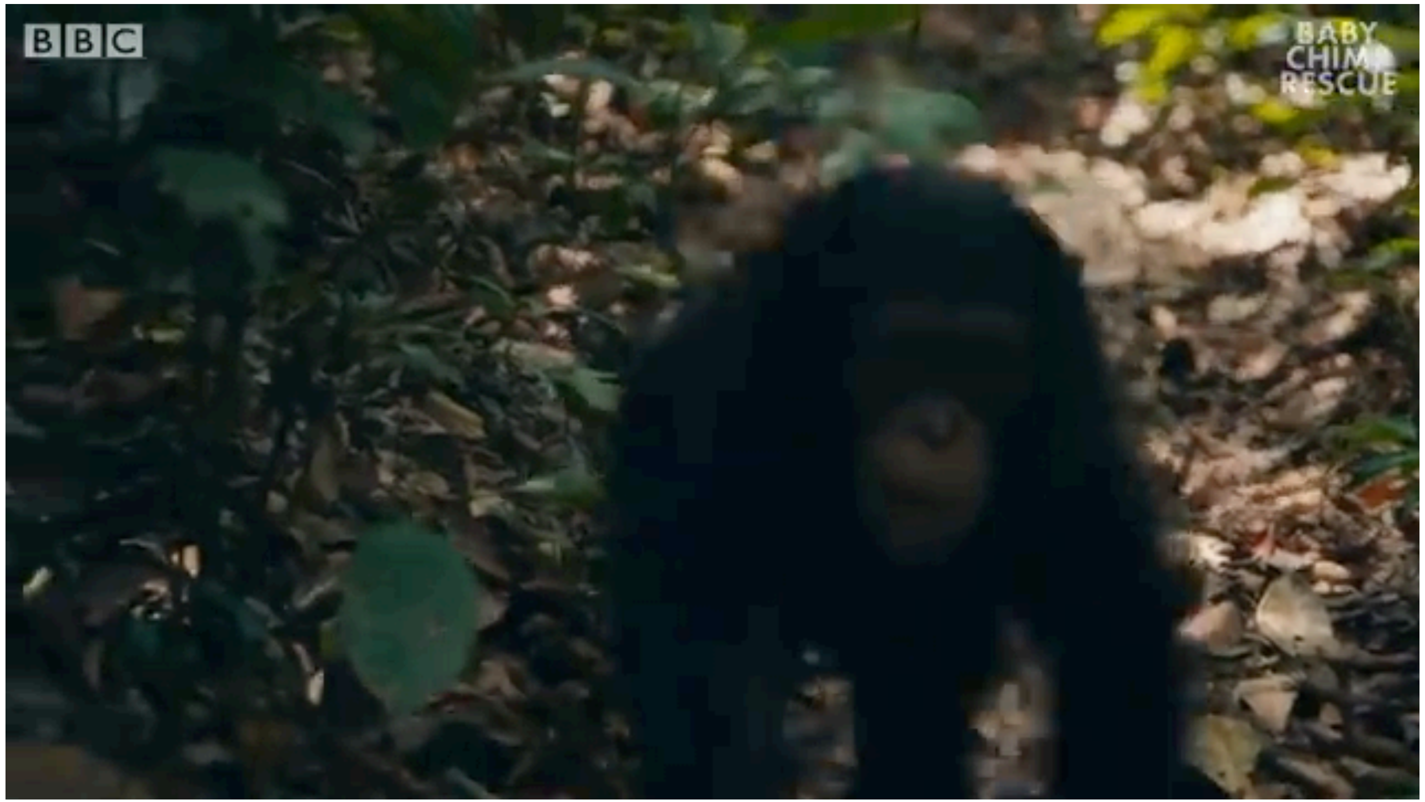


Multi-Step Planning

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

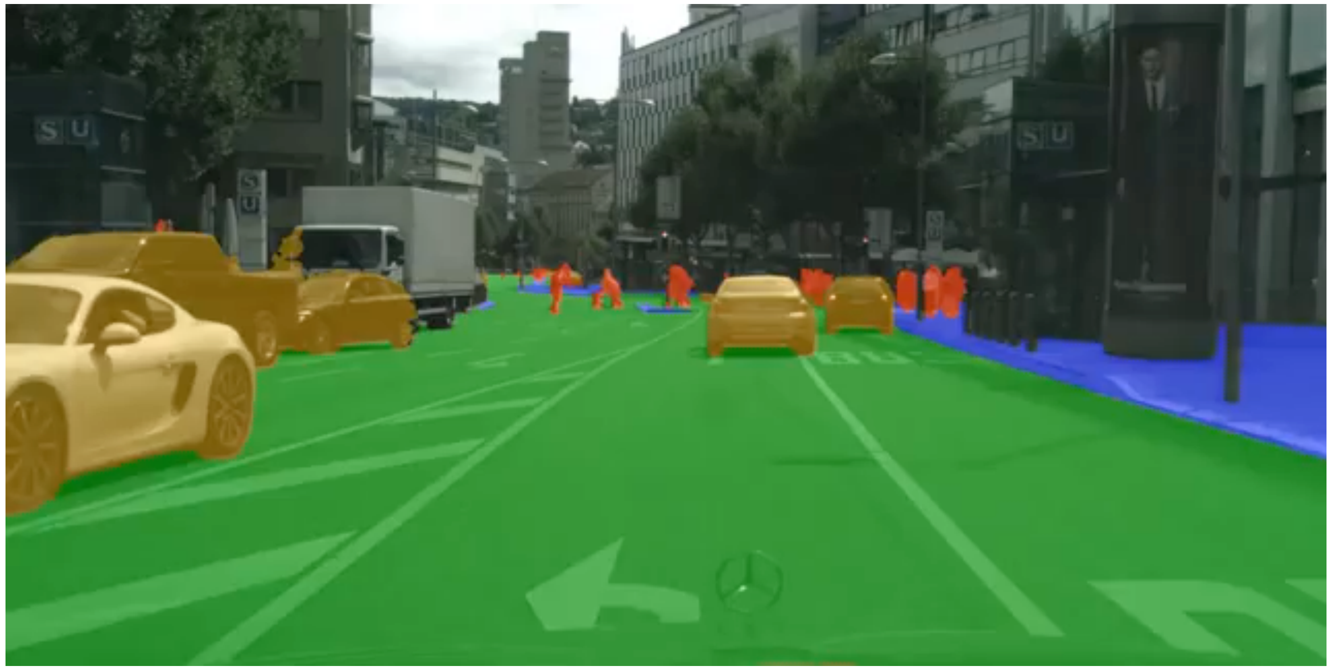
Navigation

Flexible Embodiment



From Neurons to Behavior

Scene Understanding

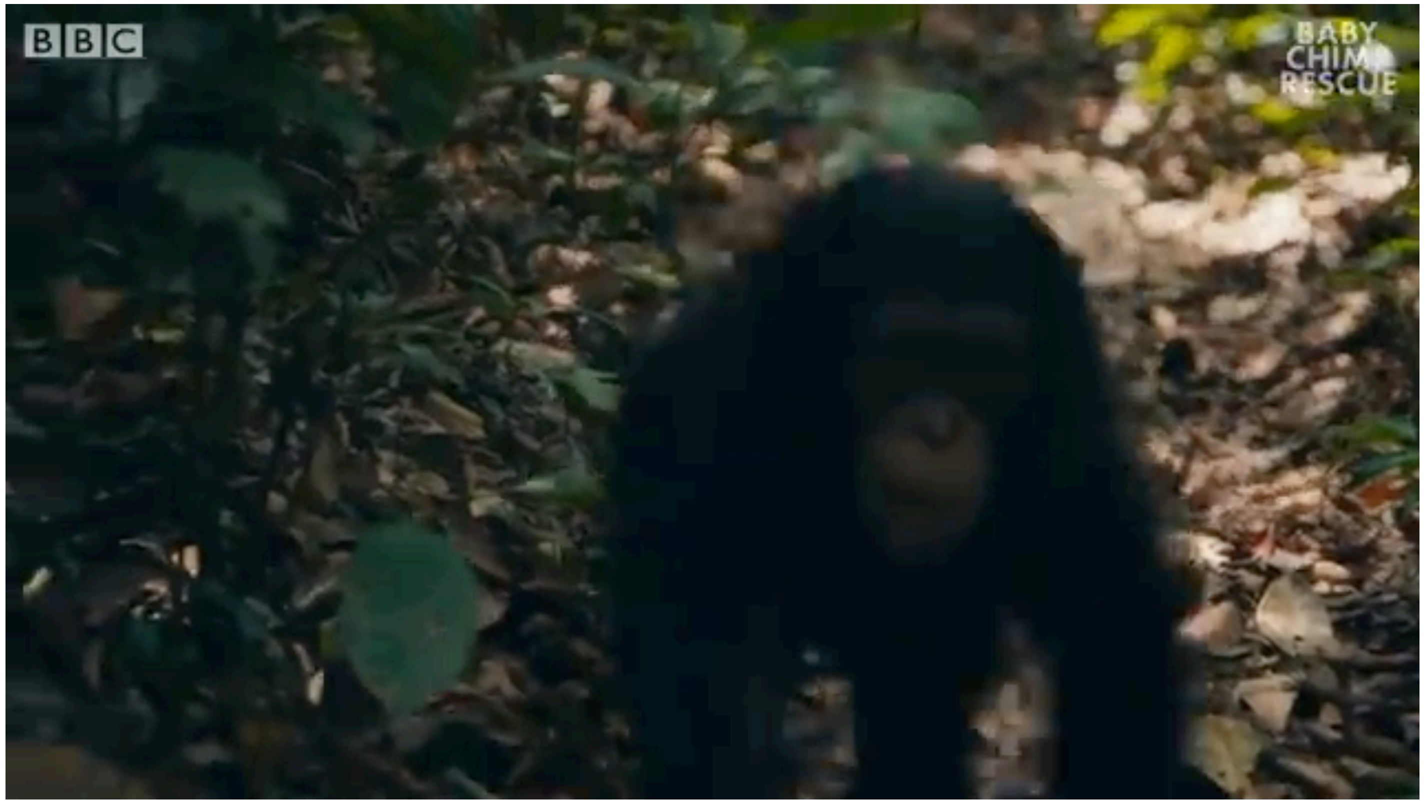


Multi-Step Planning

How do we bridge the gap from neurons to behavior?

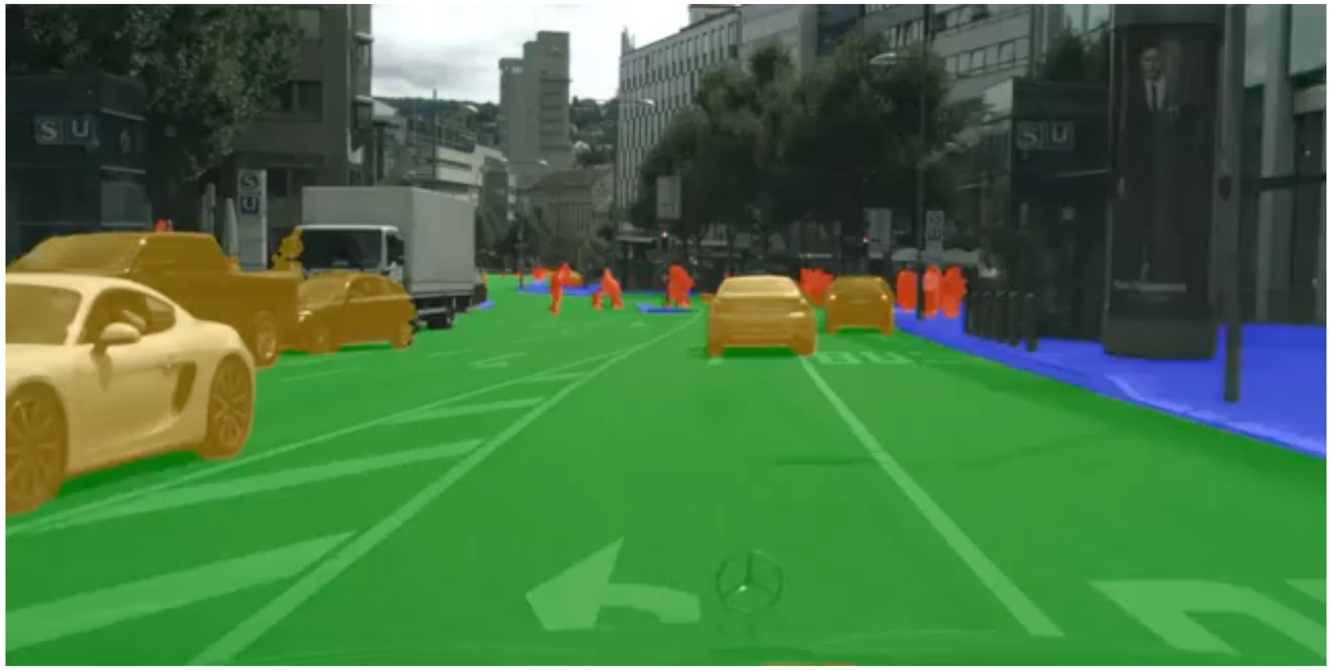
Navigation

Flexible Embodiment



From Neurons to Behavior

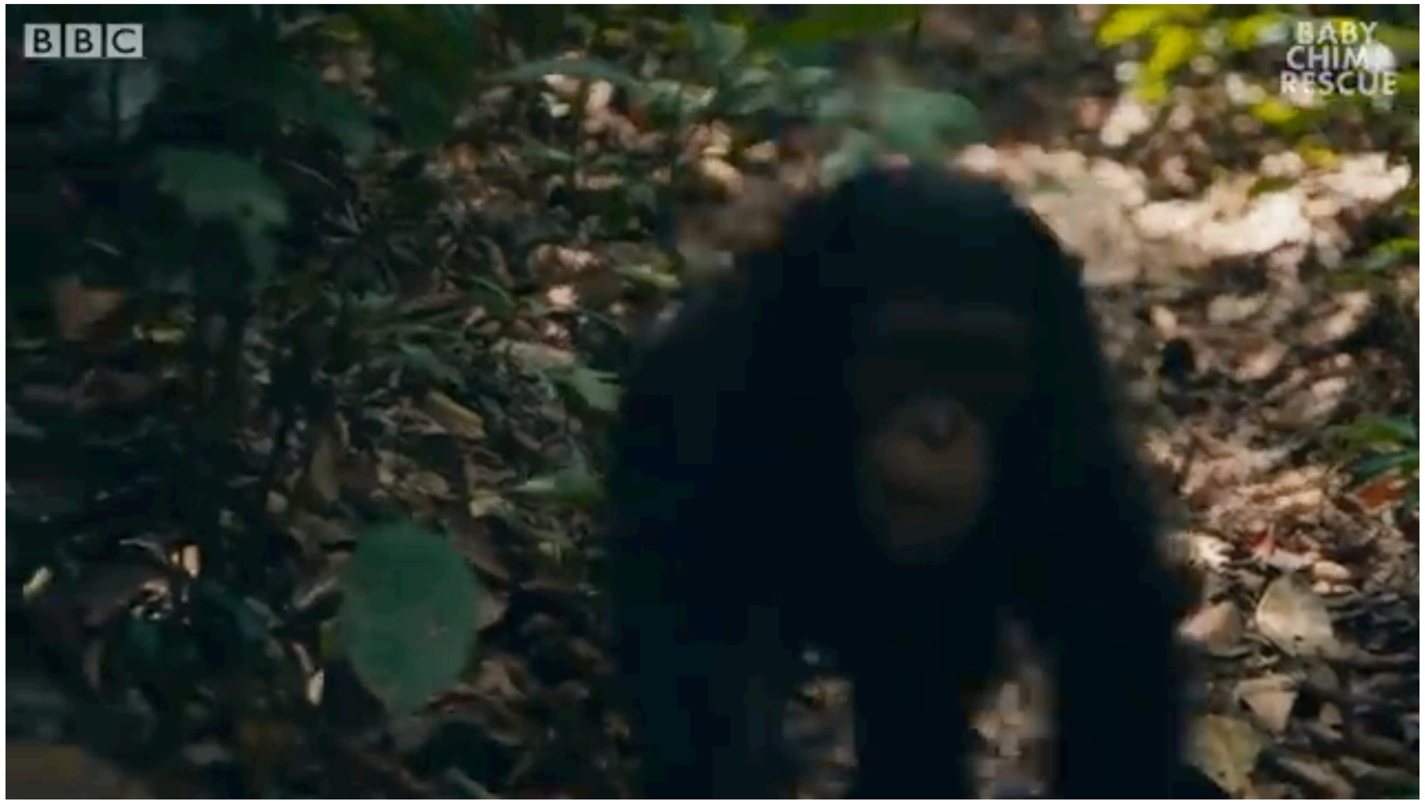
Scene Understanding



Multi-Step Planning

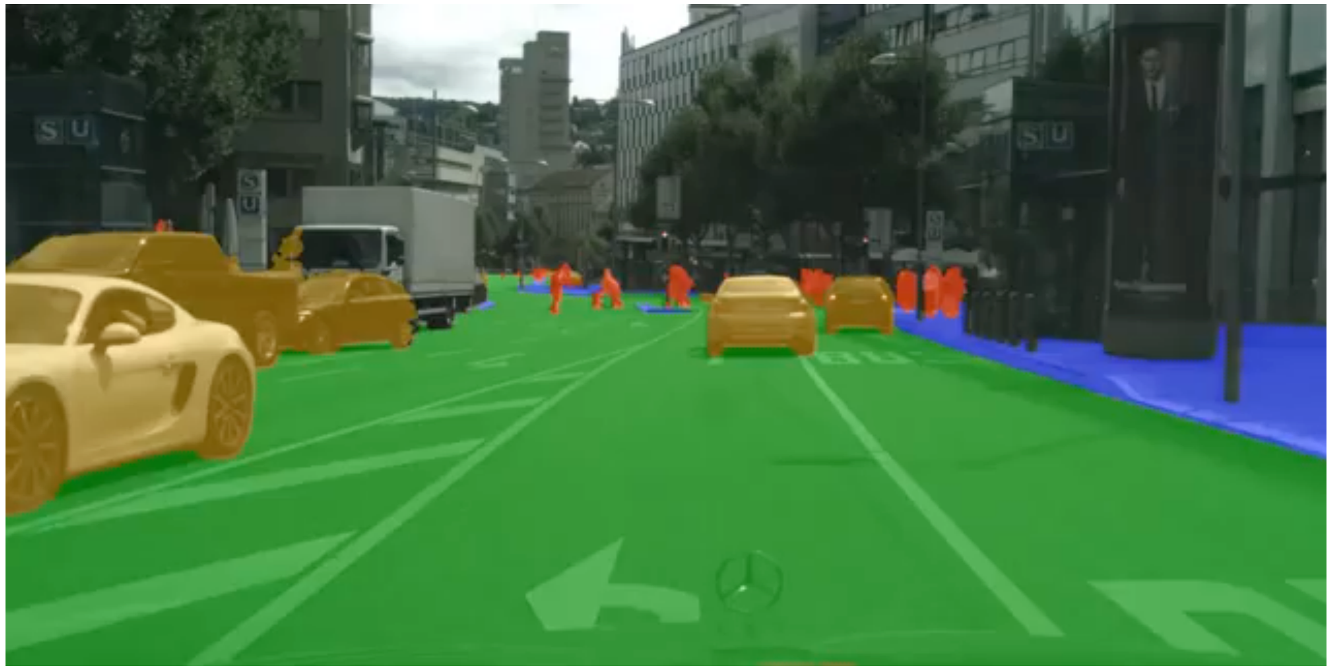
Navigation

Flexible Embodiment



From Neurons to Behavior

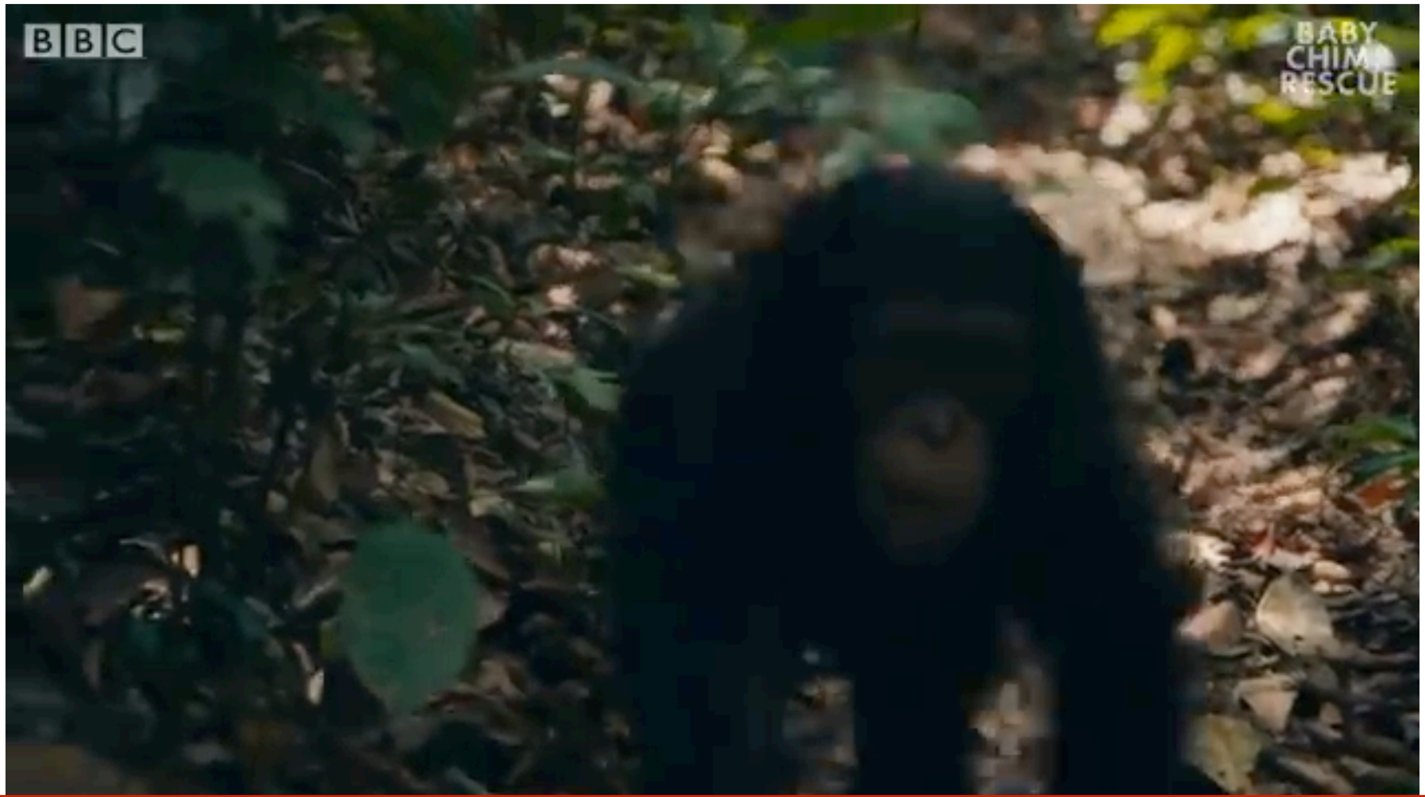
Scene Understanding



Multi-Step Planning

Navigation

Flexible Embodiment

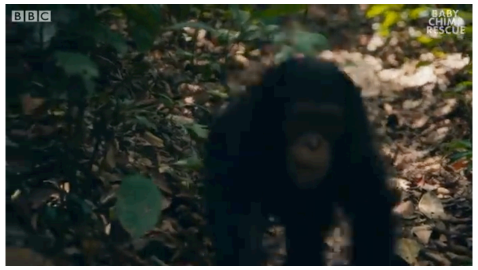


Outline

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

Reverse-Engineering Strategy: Bridging Neurons to Behavior

Reverse-Engineering Strategy: Bridging Neurons to Behavior



Flexible Embodiment

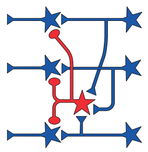
1. Operationalize a behavioral domain of interest.
Ways in which brains surpass current engineered systems? (e.g. performance, size, energy, etc)

Core object recognition

Reverse-Engineering Strategy: Bridging Neurons to Behavior

A = *architecture class*

1. *"Circuit"*

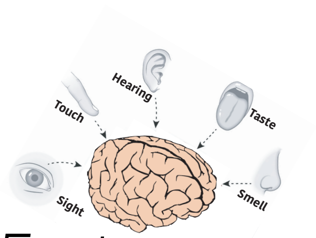


T = *task loss*

3. *"Ecological niche/behavior"*

1. Operationalize a behavioral domain of interest. Ways in which brains surpass current engineered systems? (e.g. performance, size, energy, etc)

2. Hypothesize architectures and tasks (loss functions).



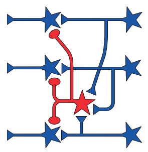
2. *"Environment"*

D = *data stream*

Reverse-Engineering Strategy: Bridging Neurons to Behavior

A *= architecture class*

1. *"Circuit"*



Neuroscience constrained

T *= task loss*

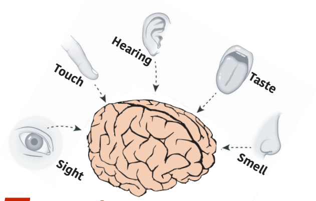
3. *"Ecological niche/behavior"*

Cognitive science constrained

1. Operationalize a behavioral domain of interest. Ways in which brains surpass current engineered systems? (e.g. performance, size, energy, etc)

2. Hypothesize architectures and tasks (loss functions).

Neuroscience constrained



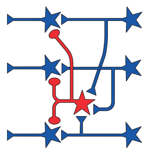
2. *"Environment"*

D *= data stream*

Reverse-Engineering Strategy: Bridging Neurons to Behavior

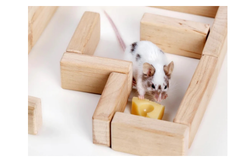
A = *architecture class*

1. *"Circuit"*



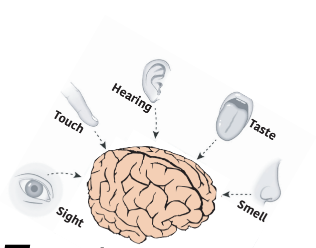
T = *task loss*

3. *"Ecological niche/behavior"*



1. Operationalize a behavioral domain of interest. Ways in which brains surpass current engineered systems? (e.g. performance, size, energy, etc)

2. Hypothesize architectures and tasks (loss functions). Optimize for task (evolution *in silico*).



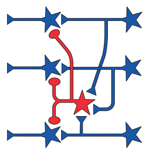
2. *"Environment"*

D = *data stream*

Reverse-Engineering Strategy: Bridging Neurons to Behavior

A = *architecture class*

1. *"Circuit"*

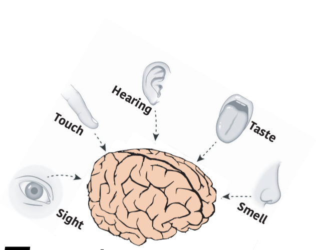


T = *task loss*

3. *"Ecological niche/behavior"*

1. Operationalize a behavioral domain of interest.
Ways in which brains surpass current engineered systems? (e.g. performance, size, energy, etc)

2. Hypothesize architectures and tasks (loss functions).
Optimize for task (evolution *in silico*).
Predict held out neural & behavioral data.



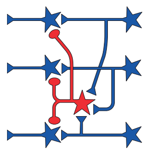
2. *"Environment"*

D = *data stream*

Reverse-Engineering Strategy: Bridging Neurons to Behavior

A = *architecture class*

1. *"Circuit"*



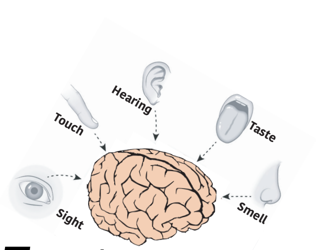
T = *task loss*

3. *"Ecological niche/behavior"*

1. Operationalize a behavioral domain of interest.
Ways in which brains surpass current engineered systems? (e.g. performance, size, energy, etc)

2. Hypothesize architectures and tasks (loss functions).
Optimize for task (evolution *in silico*).
Predict held out neural & behavioral data.

3. Core Conceptual Insights:
Identify the patterns of the best architectures & tasks.



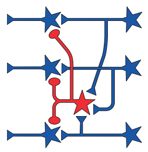
2. *"Environment"*

D = *data stream*

Reverse-Engineering Strategy: Bridging Neurons to Behavior

A *= architecture class*

1. *"Circuit"*



T *= task loss*

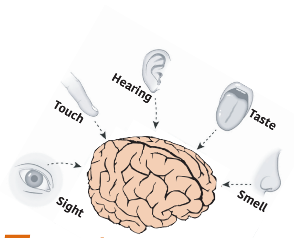
3. *"Ecological niche/behavior"*

1. Operationalize a behavioral domain of interest. Ways in which brains surpass current engineered systems? (e.g. performance, size, energy, etc)

2. Hypothesize architectures and tasks (loss functions). Optimize for task (evolution *in silico*). *Predict* held out neural & behavioral data.

3. Core Conceptual Insights: Identify the patterns of the best architectures & tasks.

A structural & functional *normative* understanding of the evolutionary pressures of the biological system to produce the behavior in (1).



2. *"Environment"*

D *= data stream*

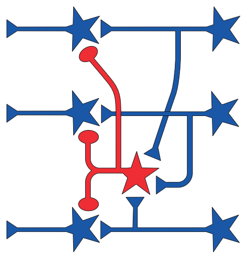
Outline

- 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 *= architecture class*

1. *"Circuit"*

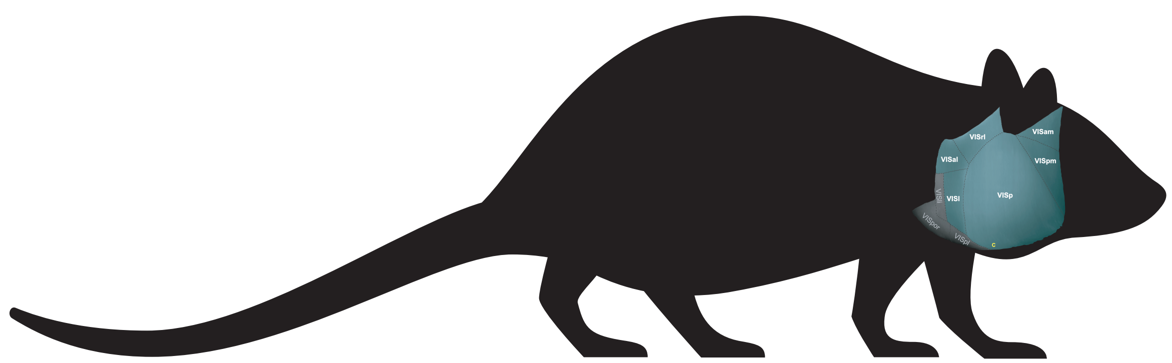


T *= task loss*

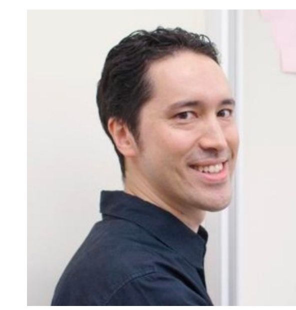
3. *"Ecological niche/behavior"*

A. Nayebi*, N. 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 (in press)

Nathan C.L. Kong*



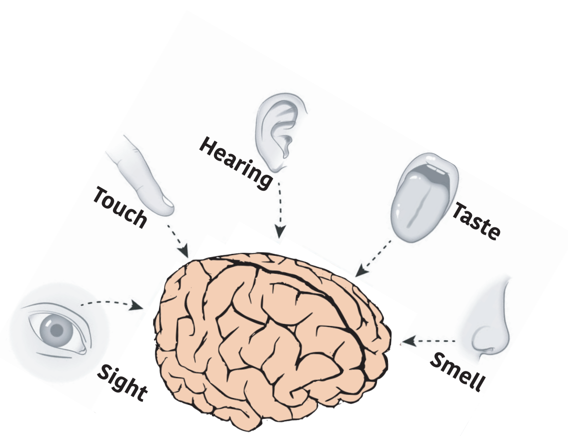
Chengxu Zhuang



Justin L. Gardner

Anthony M. Norcia

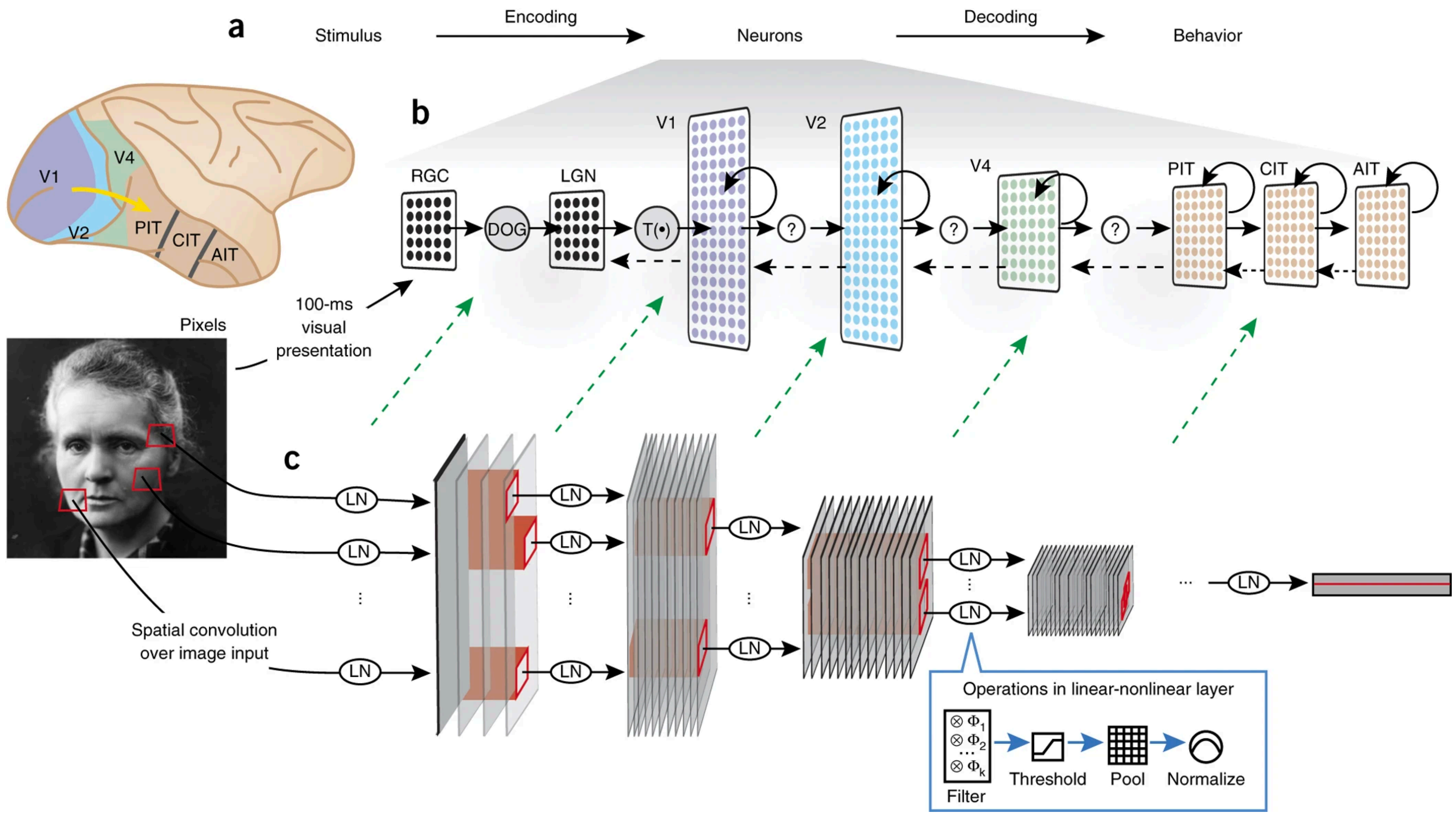
Daniel Yamins



2. *"Environment"*

D *= data stream*

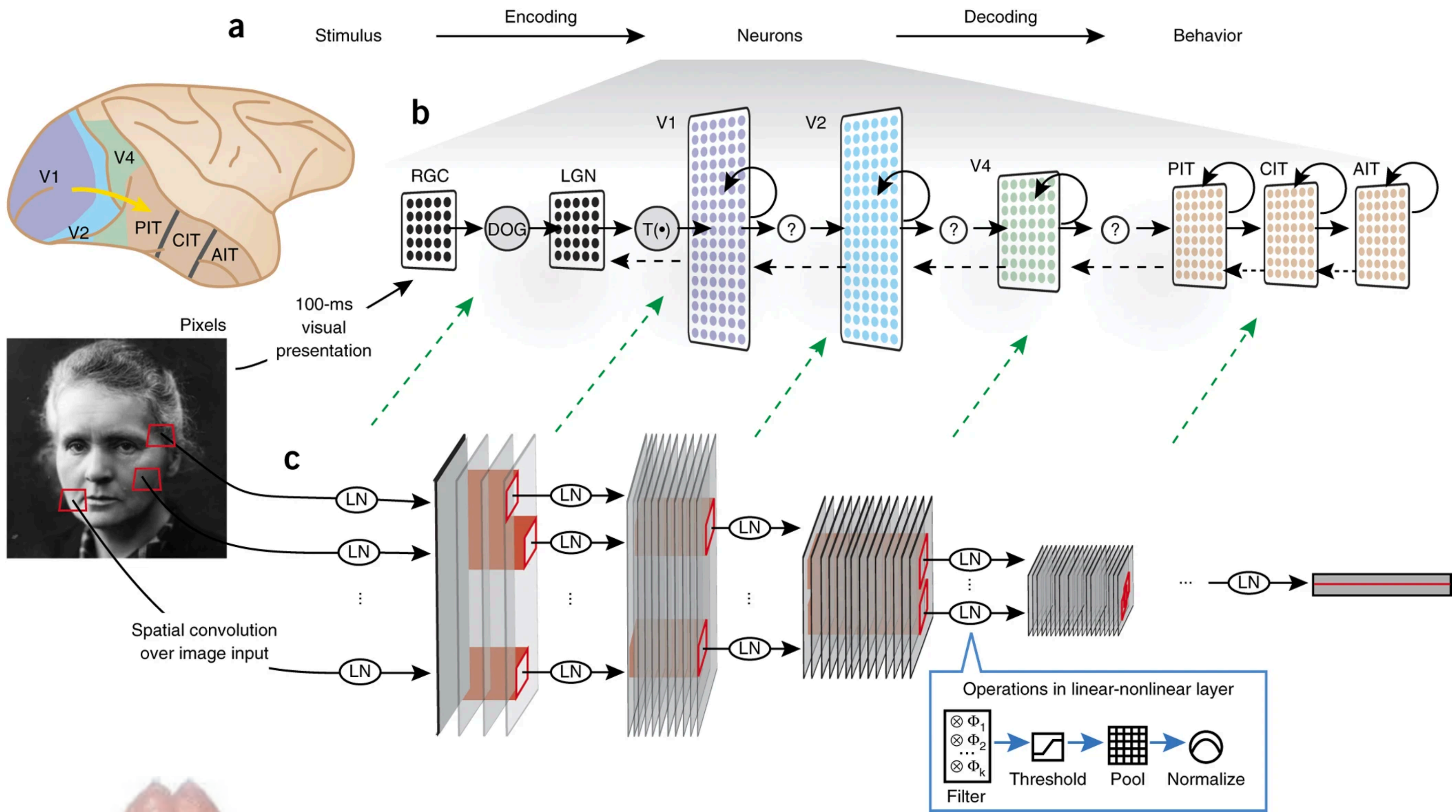
Anatomical differences between mouse and primate visual systems



Primate Ventral Stream

Hierarchical CNN

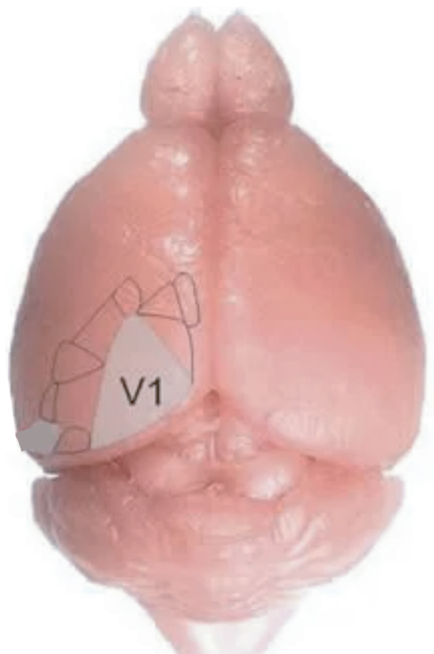
Anatomical differences between mouse and primate visual systems

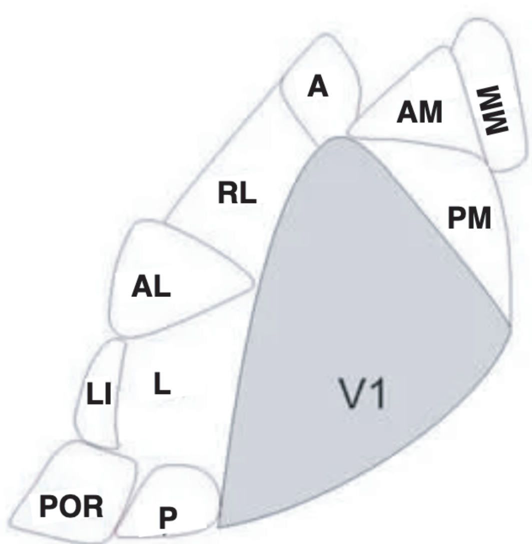


Primate Ventral Stream

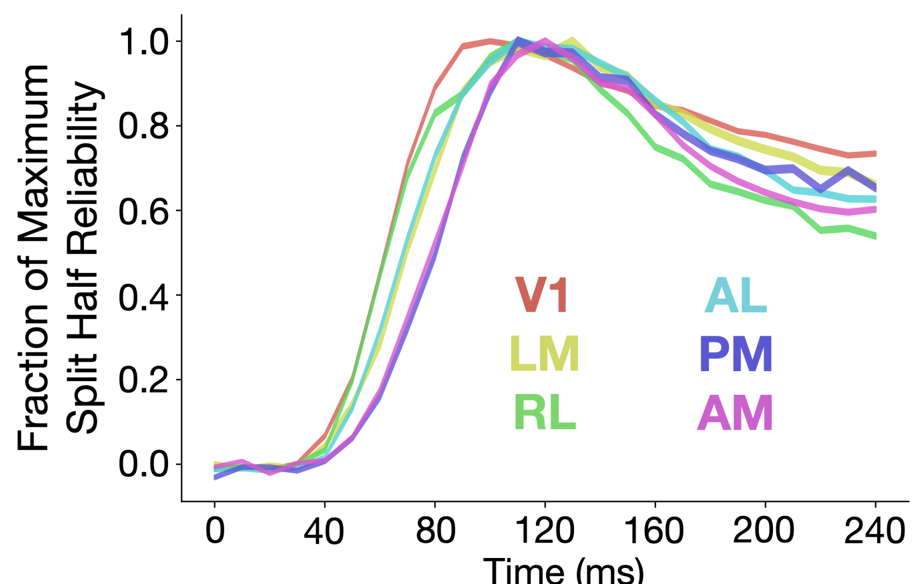
Hierarchical CNN

Mouse Visual Cortex

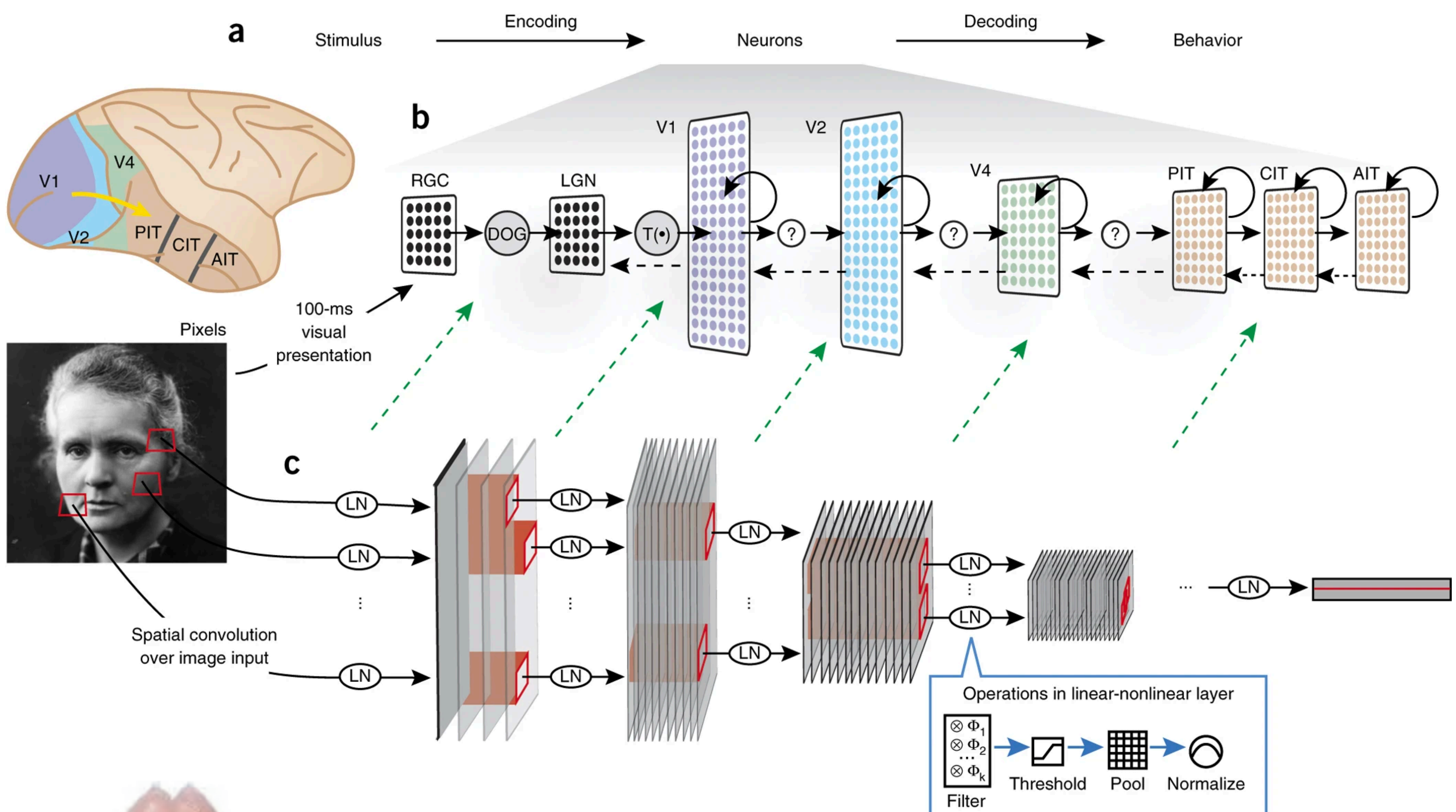




Temporal Hierarchy



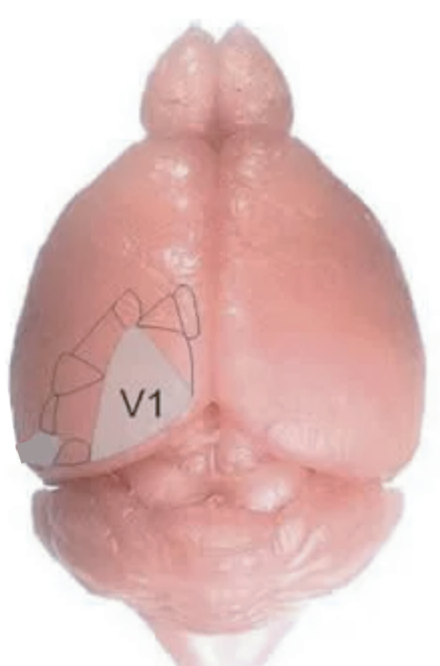
Anatomical differences between mouse and primate visual systems

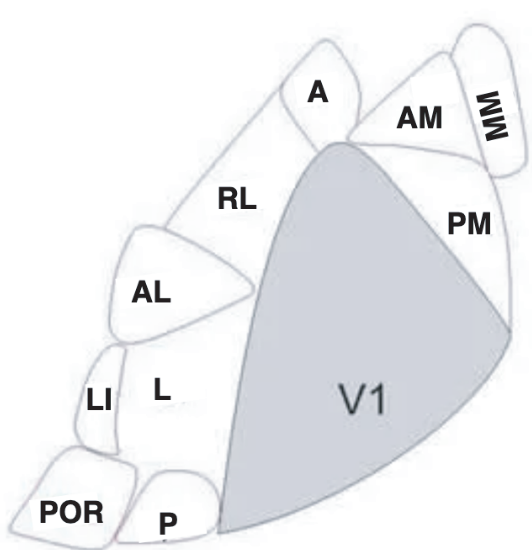


Primate Ventral Stream

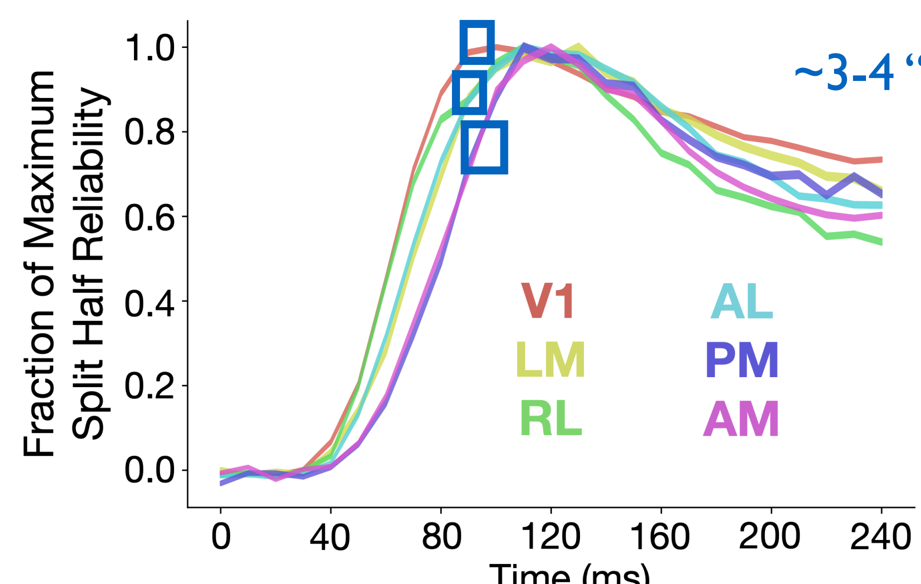
Hierarchical CNN

Mouse Visual Cortex



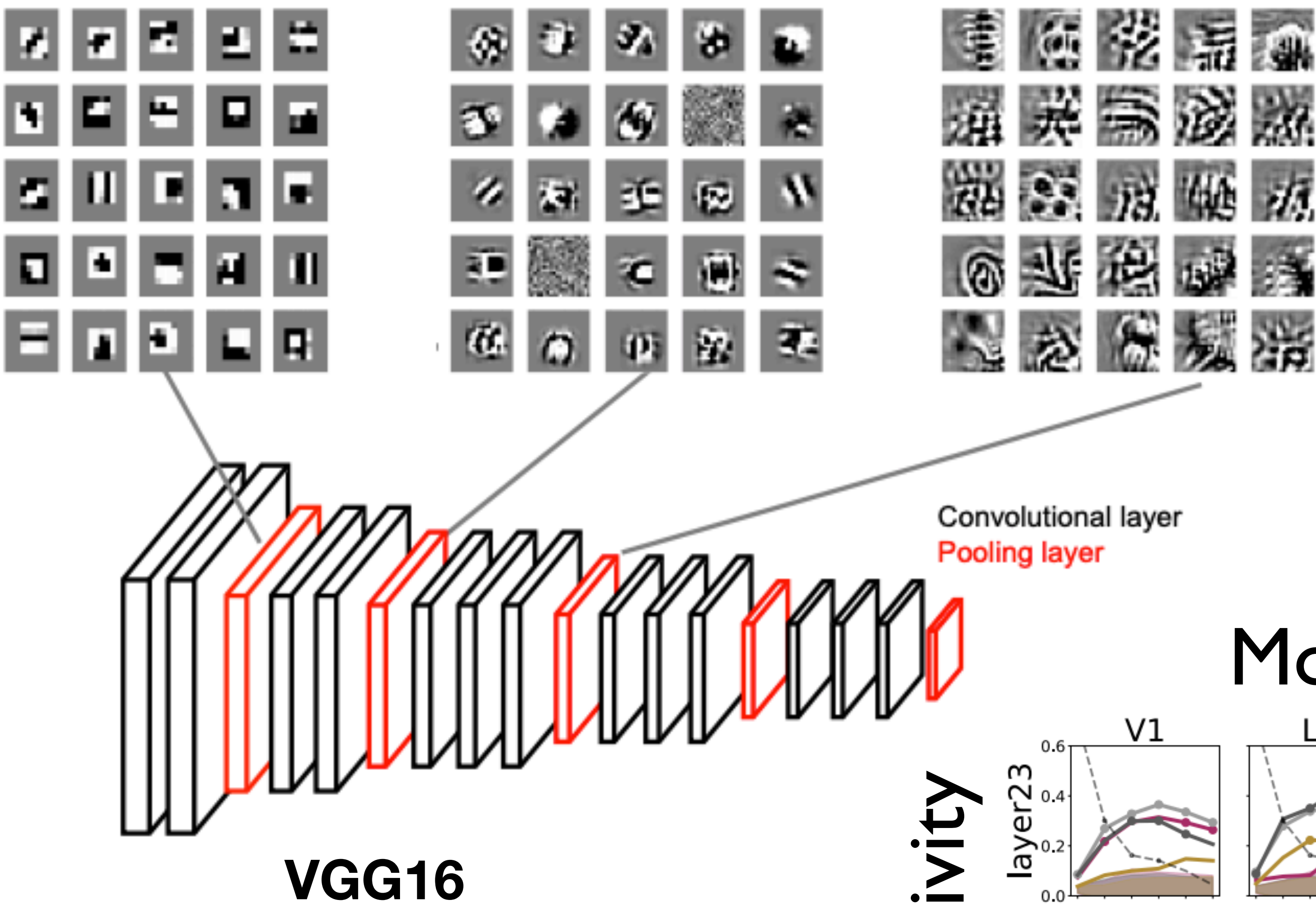


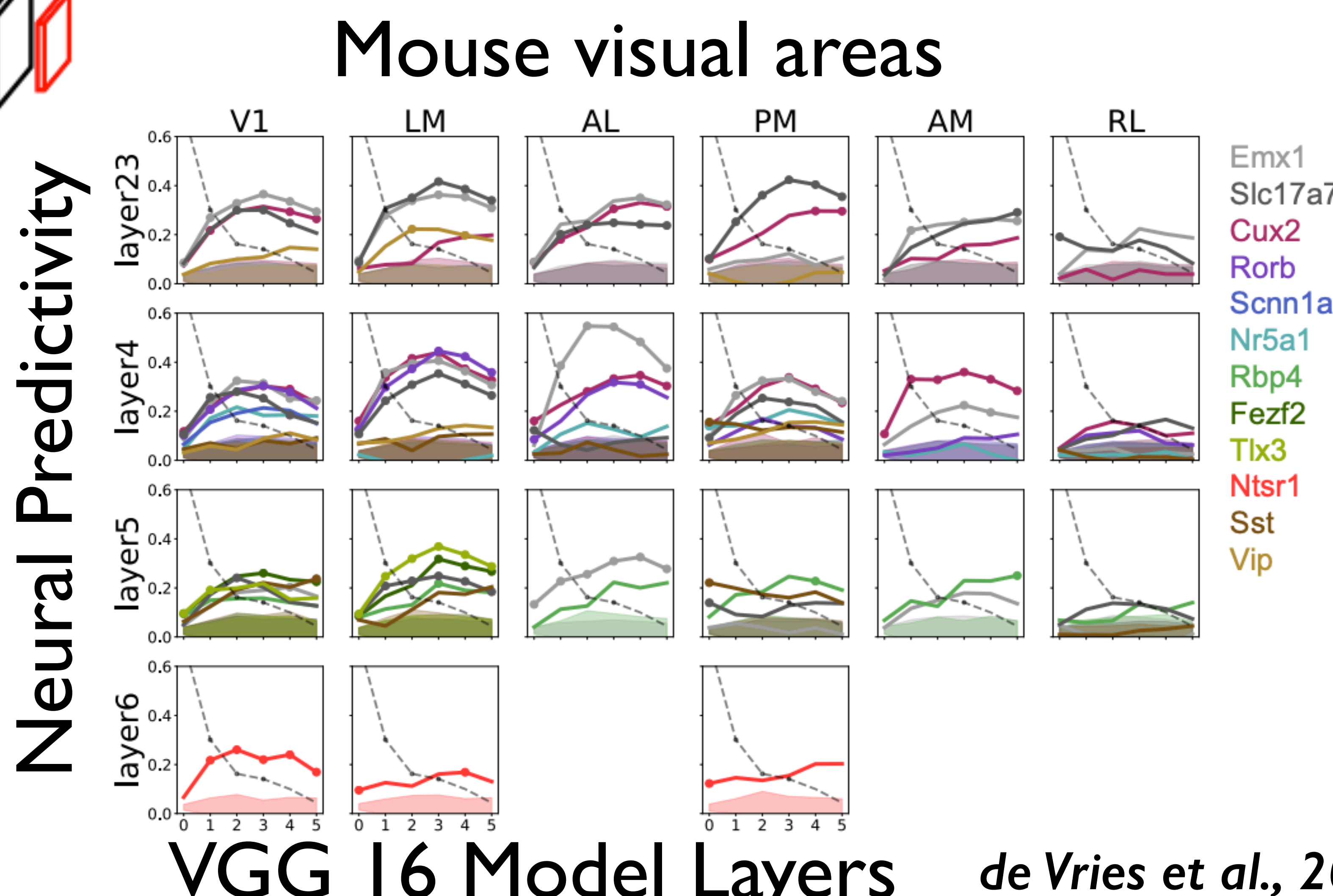
Temporal Hierarchy



~3-4 "groups" based on timing

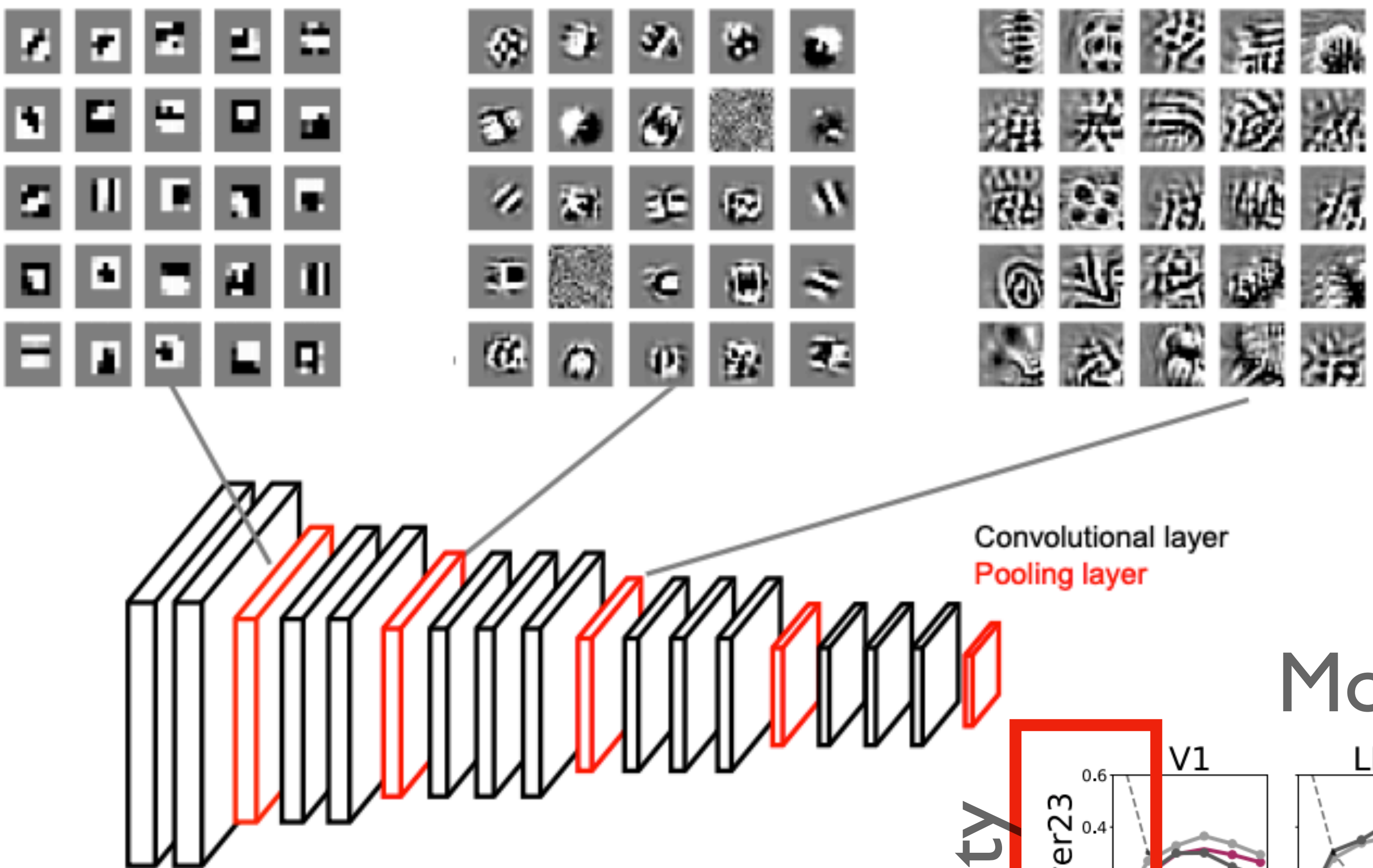
Initial deep neural network models of mouse visual cortex



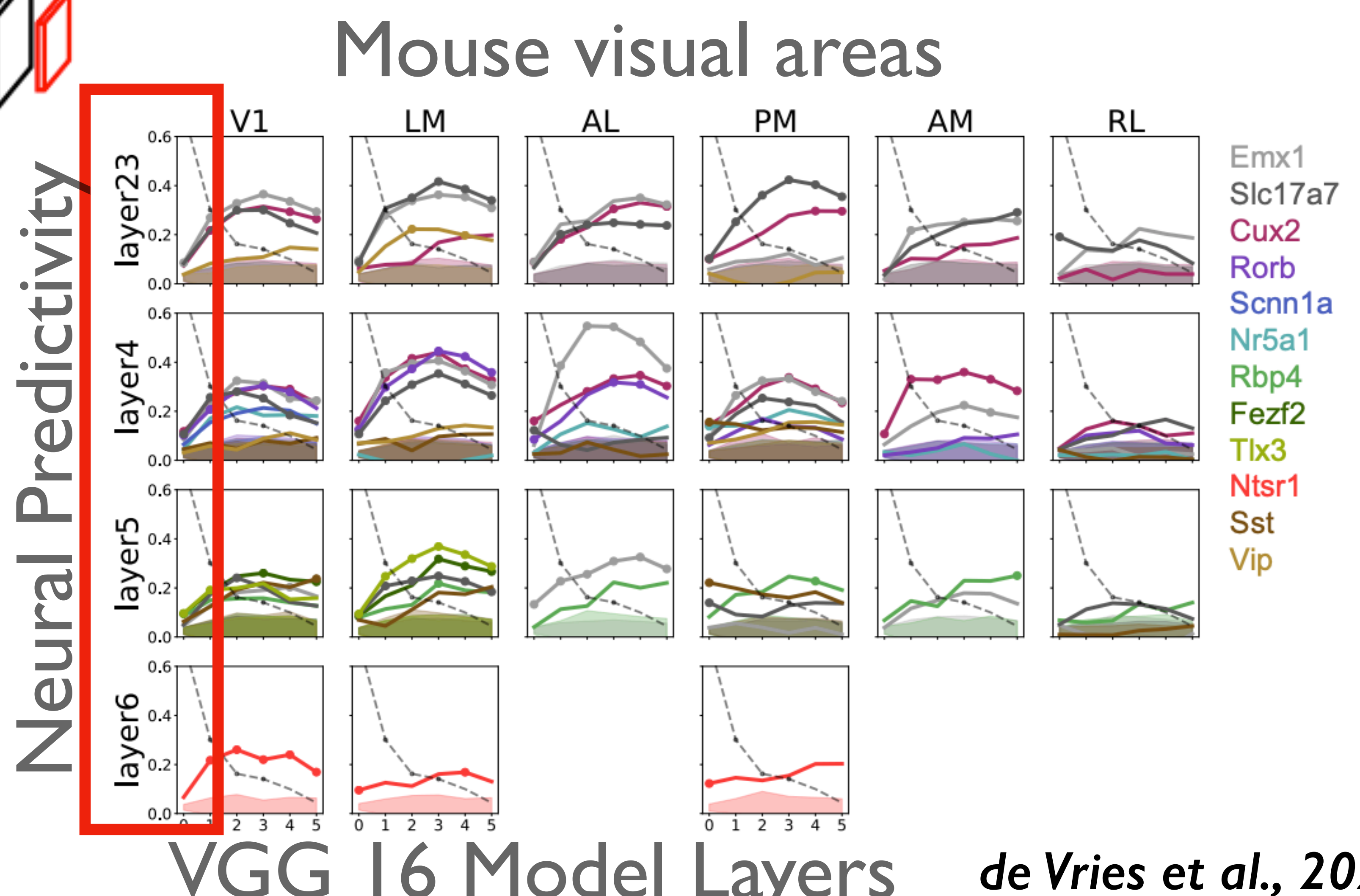


de Vries et al., 2020

Deep models are a poor match to responses

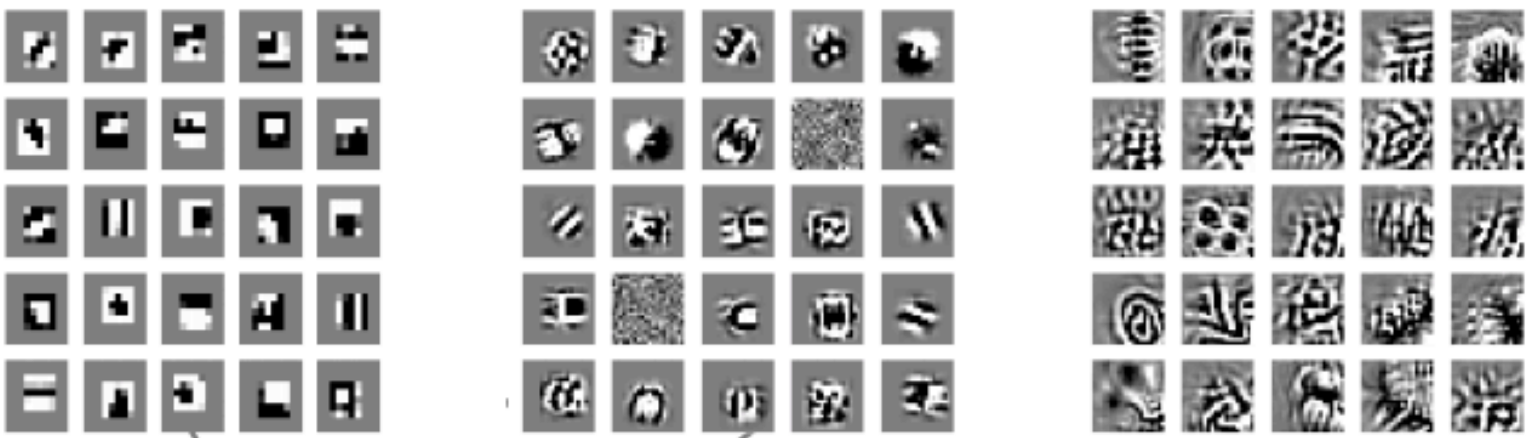


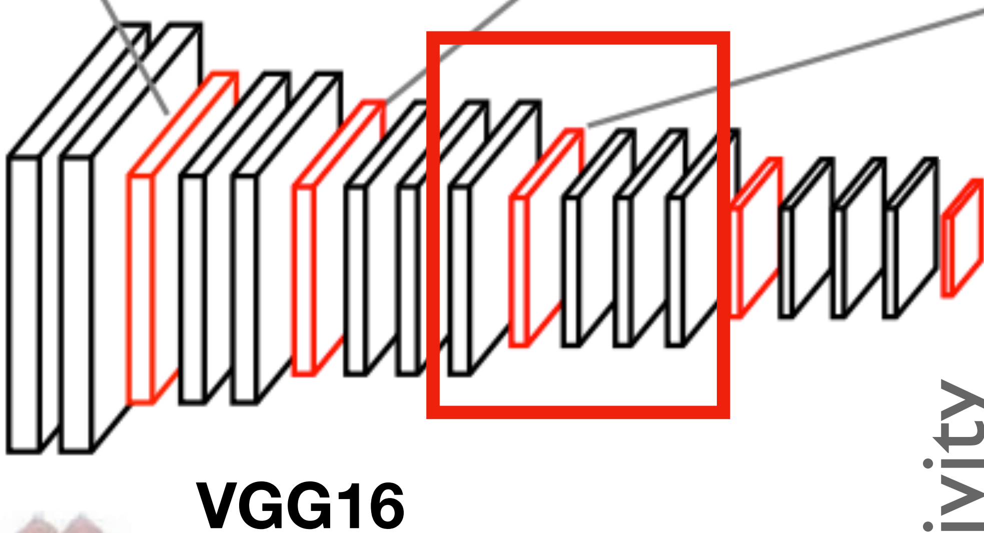
Low predictivity between model's layers and mouse visual areas!



de Vries et al., 2020

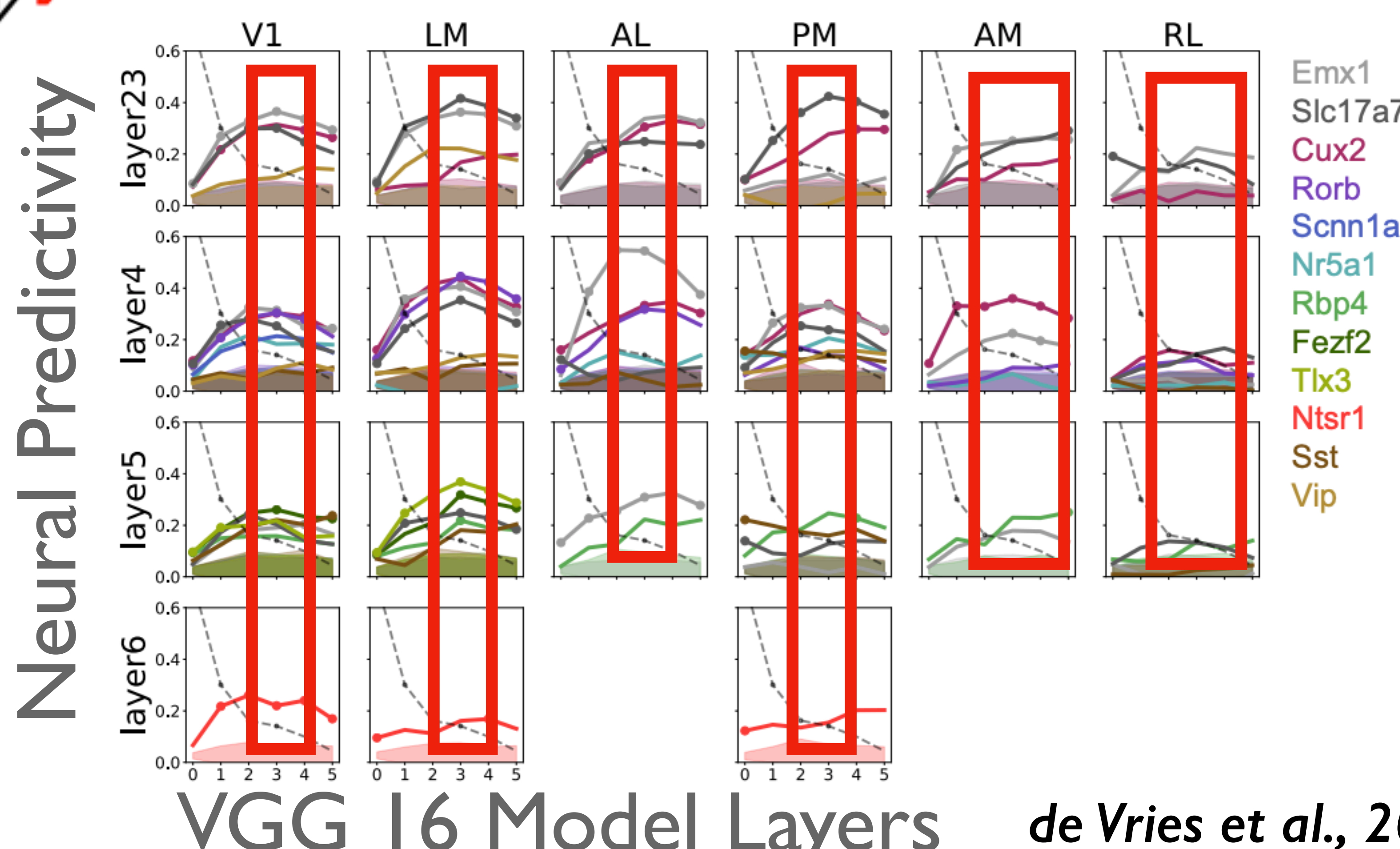
Deep models suggest mouse visual cortex is representationally deep

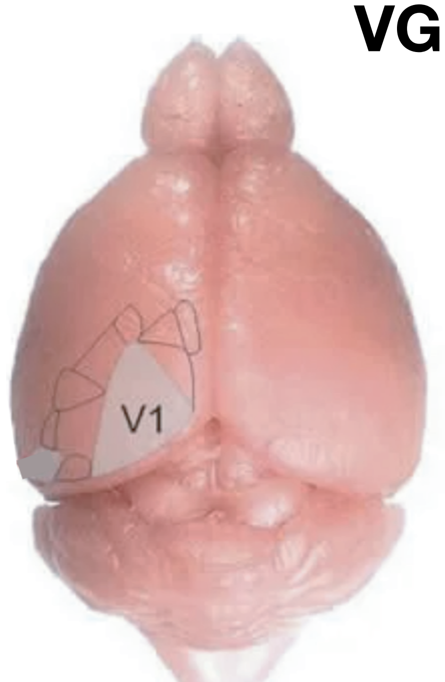


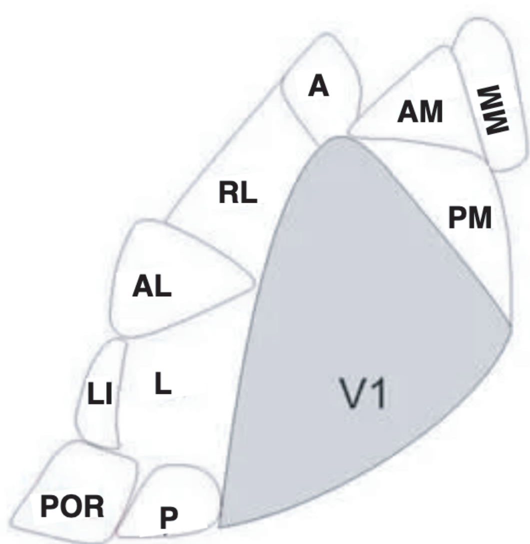


Mid-levels best predict visual areas

Mouse visual areas







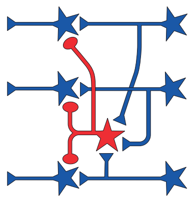
Scanziani Lab

de Vries et al., 2020

Goal-Driven Models of Mouse Visual Cortex

A *= architecture class*

1. *"Circuit"*

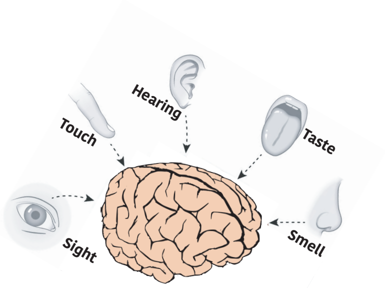


L *= loss function*

3. *"Ecological niche/behavior"*

<u>Neurobiological Puzzle:</u>

What are the core differences between mouse visual cortex and the primate ventral stream?



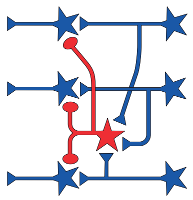
2. *"Environment"*

D *= data stream*

Putting it all together: Circuit, Inputs, Behavior

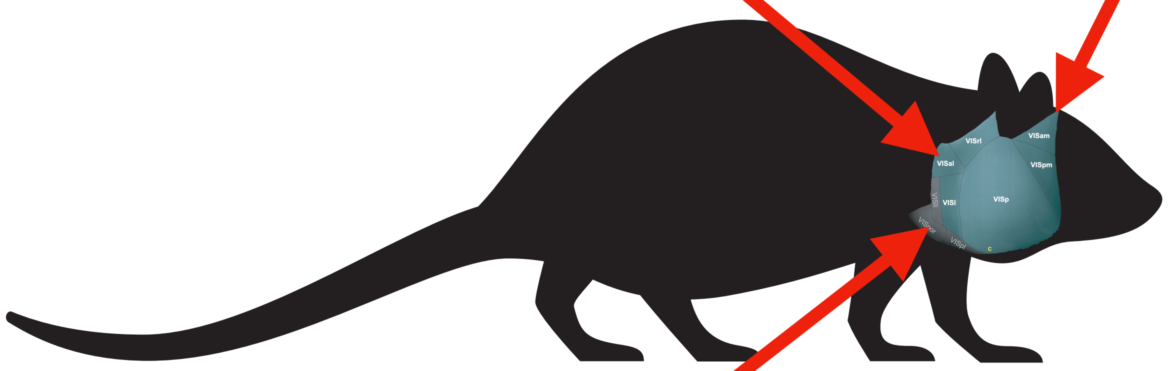
A = *architecture class*

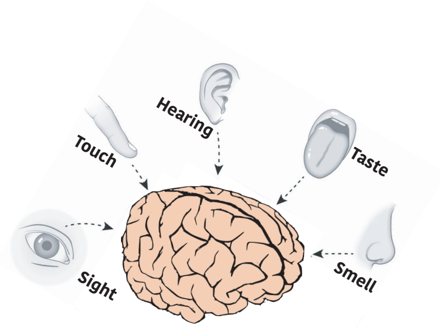
1. *"Circuit"*



T = *task loss*

3. *"Ecological niche/behavior"*





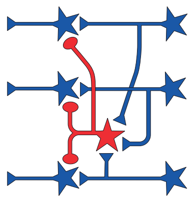
2. *"Environment"*

D = *data stream*

Putting it all together: Circuit, Inputs, Behavior

A = *architecture class*

1. *"Circuit"*

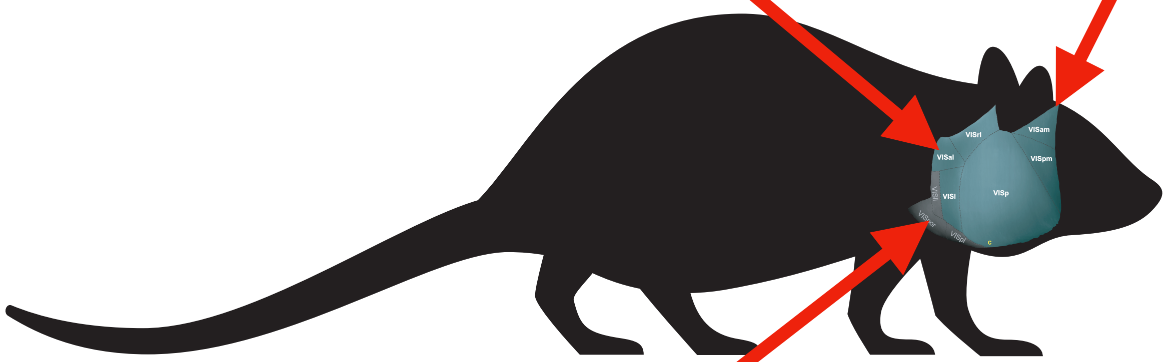


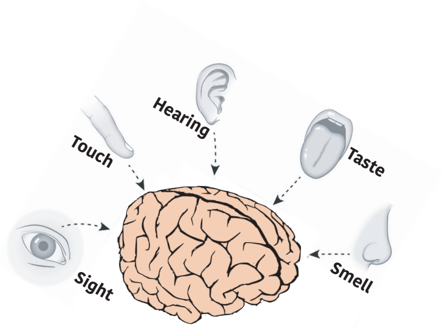
shallow

~~deep~~

T = *task loss*

3. *"Ecological niche/behavior"*





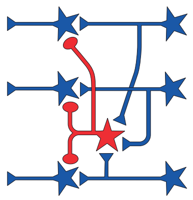
2. *"Environment"*

D = *data stream*

Putting it all together: Circuit, Inputs, Behavior

A = *architecture class*

1. *"Circuit"*



shallow

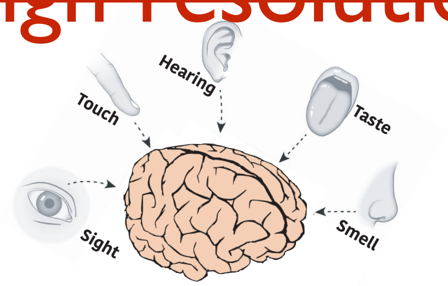
~~deep~~

T = *task loss*

3. *"Ecological niche/behavior"*

low resolution

~~high resolution~~



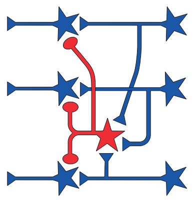
2. *"Environment"*

D = *data stream*

Putting it all together: Circuit, Inputs, Behavior

A *= architecture class*

1. *"Circuit"*



shallow

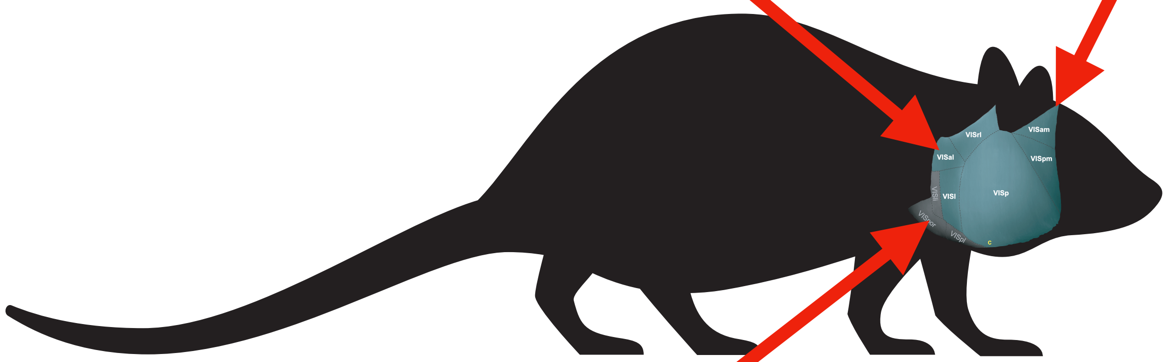
~~deep~~

T *= task loss*

3. *"Ecological niche/behavior"*

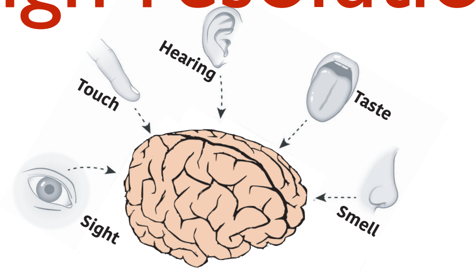
self-supervised

~~supervised~~



low resolution

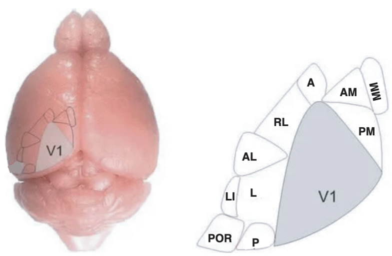
~~high resolution~~

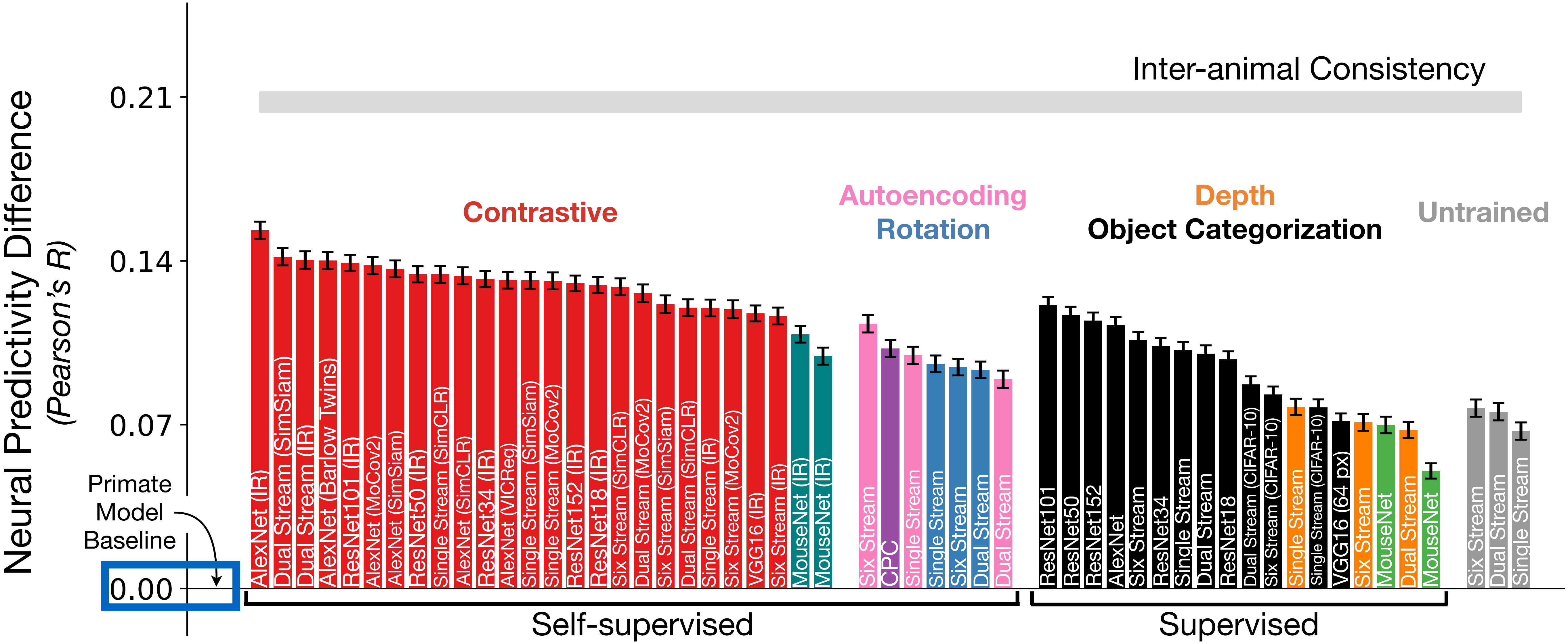


2. *"Environment"*

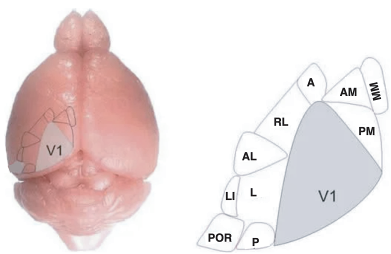
D *= data stream*

Substantially improving neural response predictivity of models of mouse visual cortex

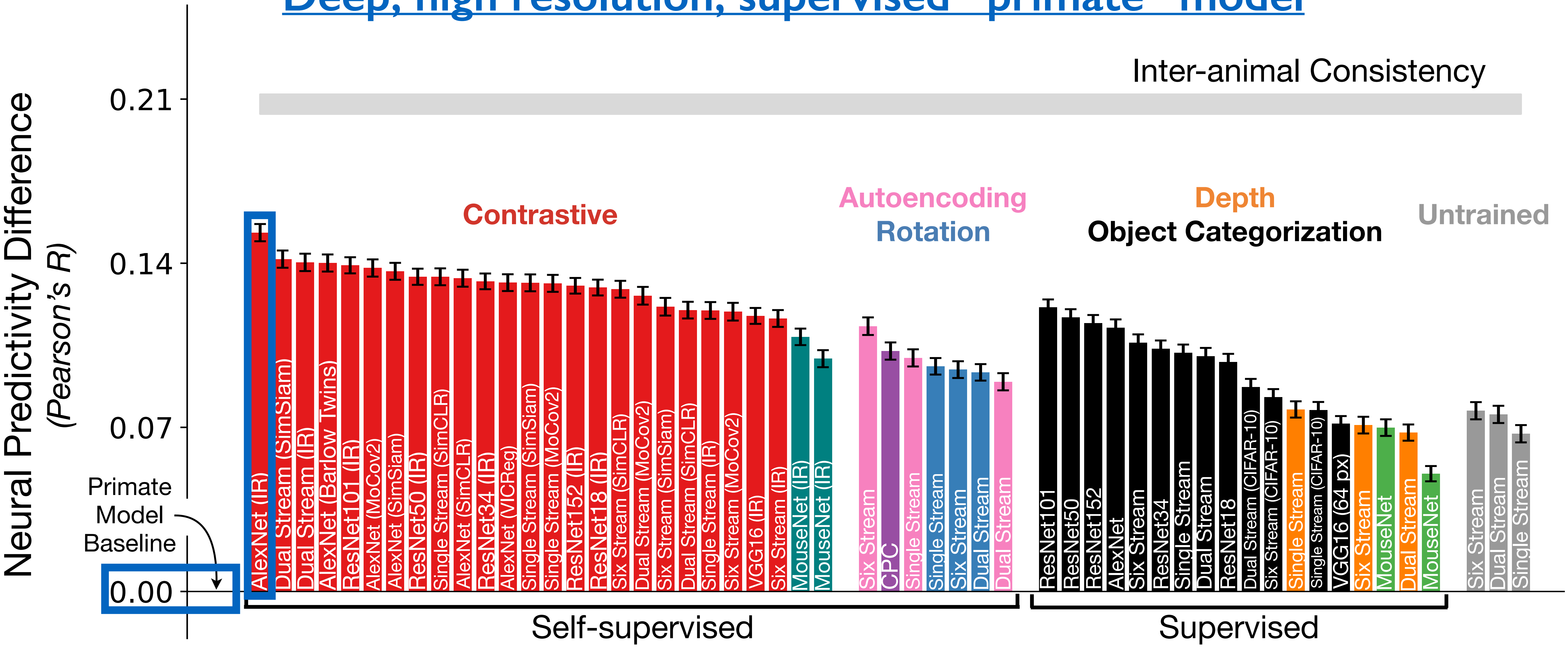




Substantially improving neural response predictivity of models of mouse visual cortex



Shallow, low resolution, self-supervised model
>
Deep, high resolution, supervised "primate" model

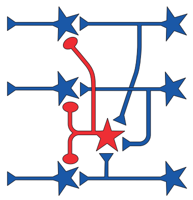


Improves neural predictivity from 56% to 90% of the data noise ceiling

Distilling Constraints: Circuit

A *= architecture class*

1. *"Circuit"*



shallow

~~deep~~

T *= task loss*

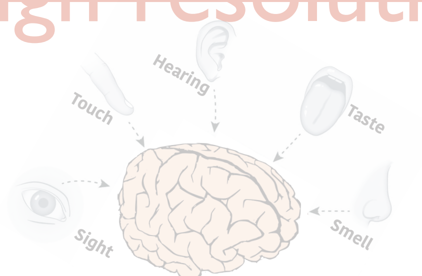
3. *"Ecological niche/behavior"*

self-supervised

~~supervised~~

low resolution

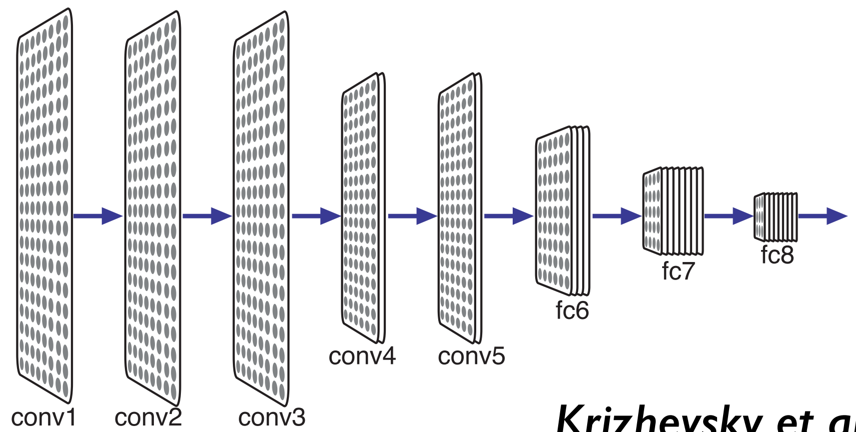
~~high resolution~~



2. *"Environment"*

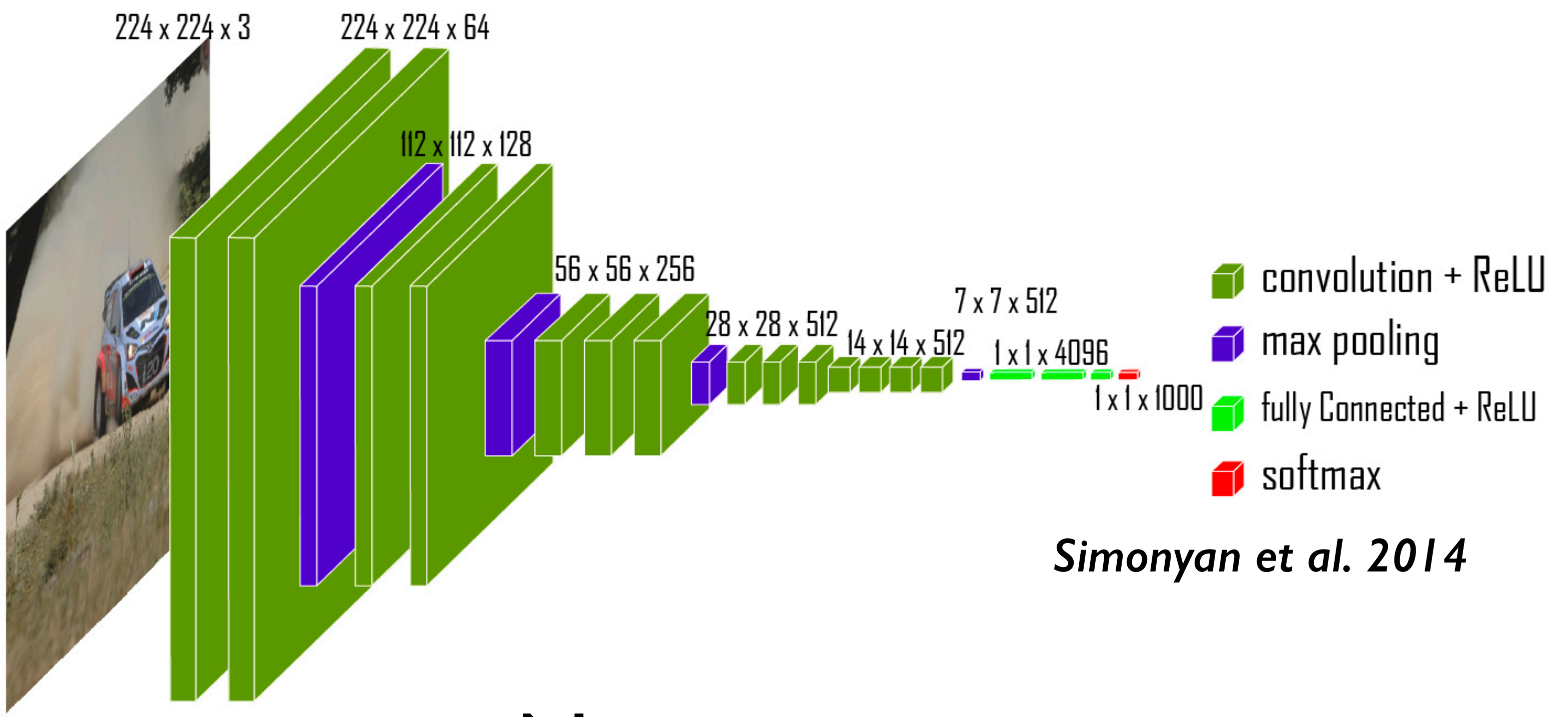
D *= data stream*

Architectural Choices



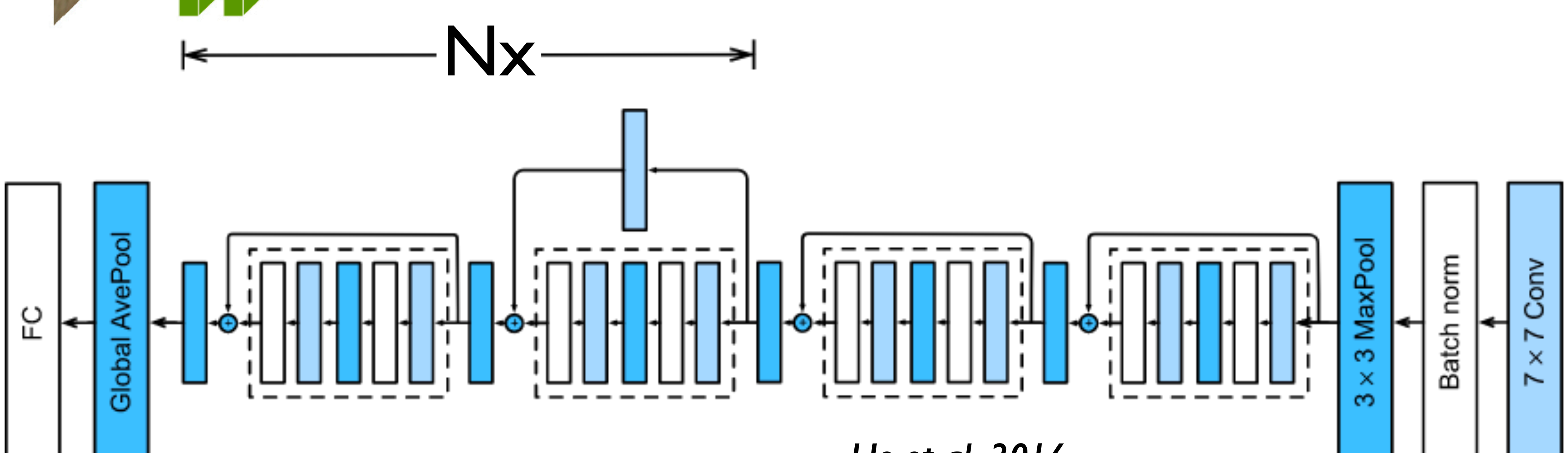
AlexNet

Krizhevsky et al. 2012



Simonyan et al. 2014

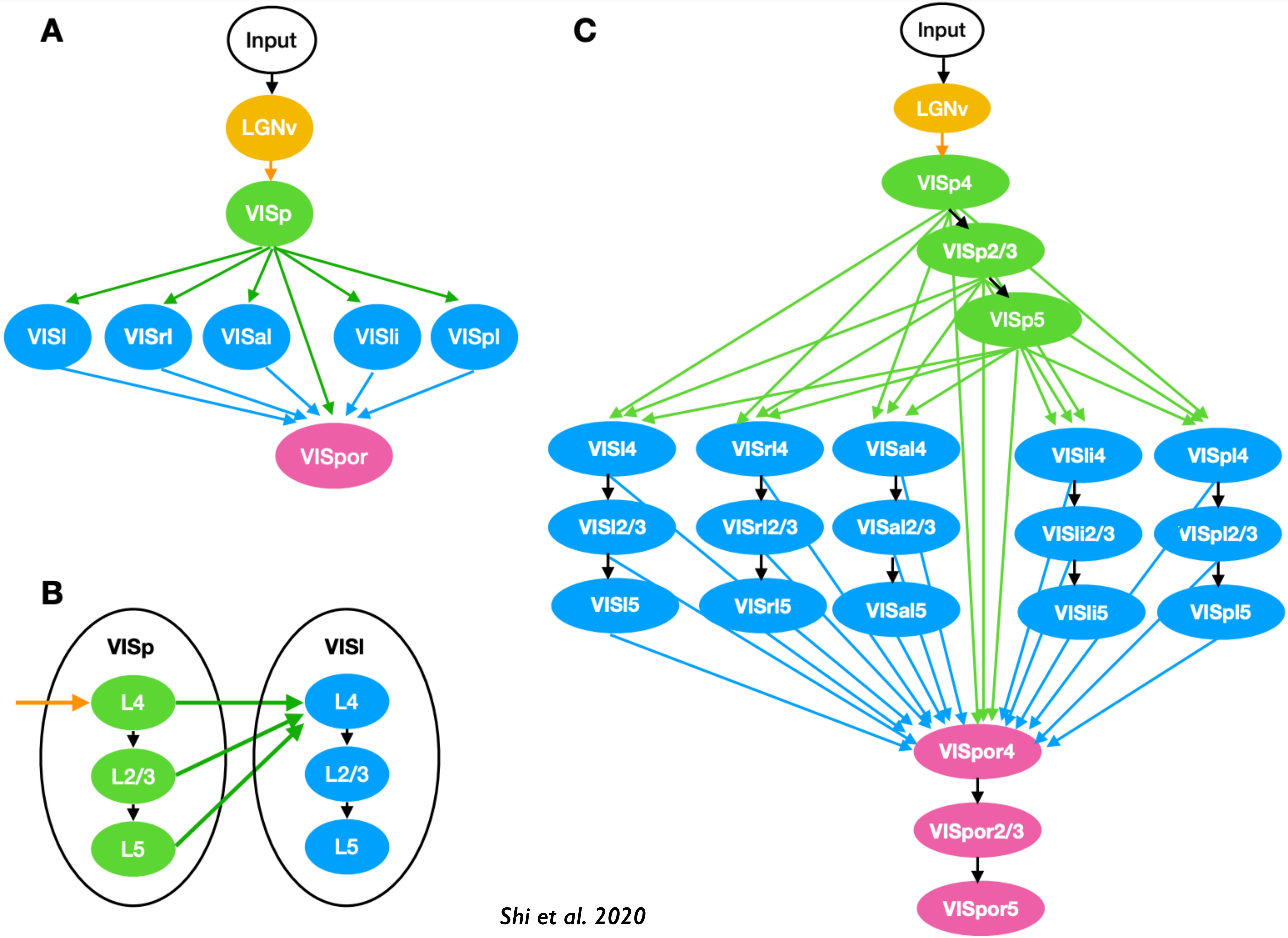
VGG-16



He et al. 2016

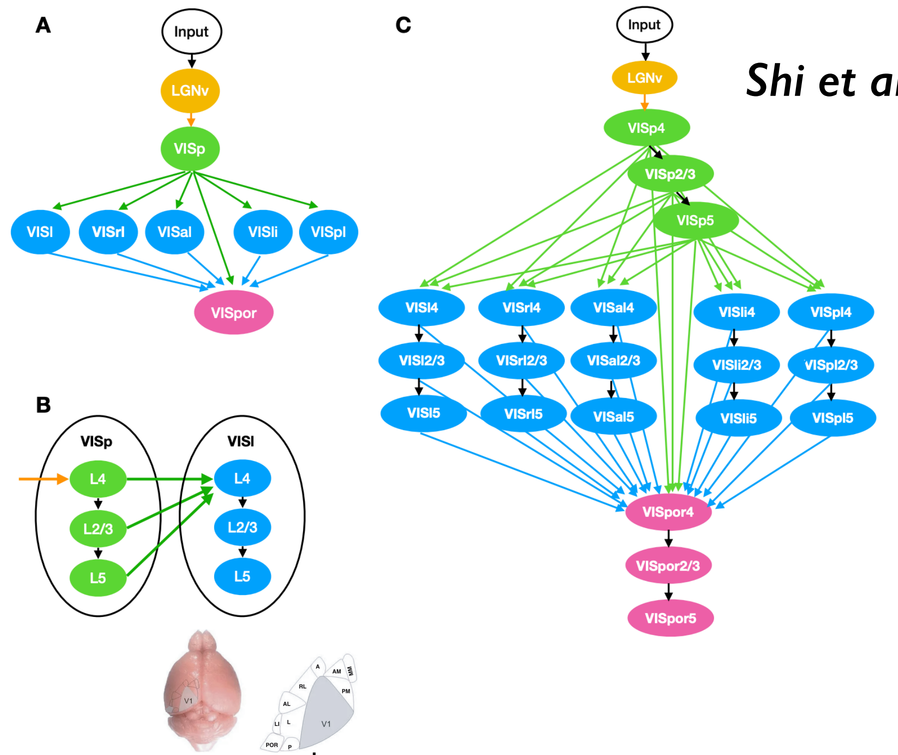
ResNet Family: ResNet-18, 34, 101, 152

More Connectome-Inspired Architectures?



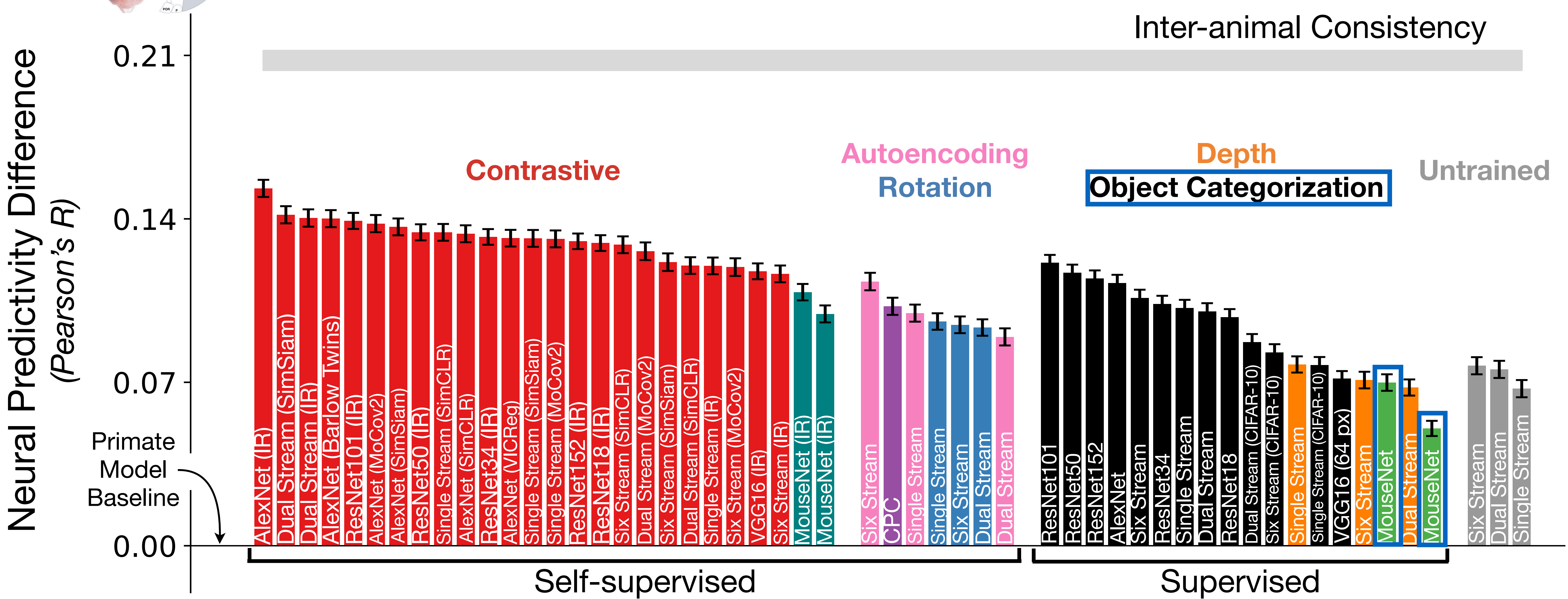
Shi et al. 2020

More Connectome-Inspired Architectures?

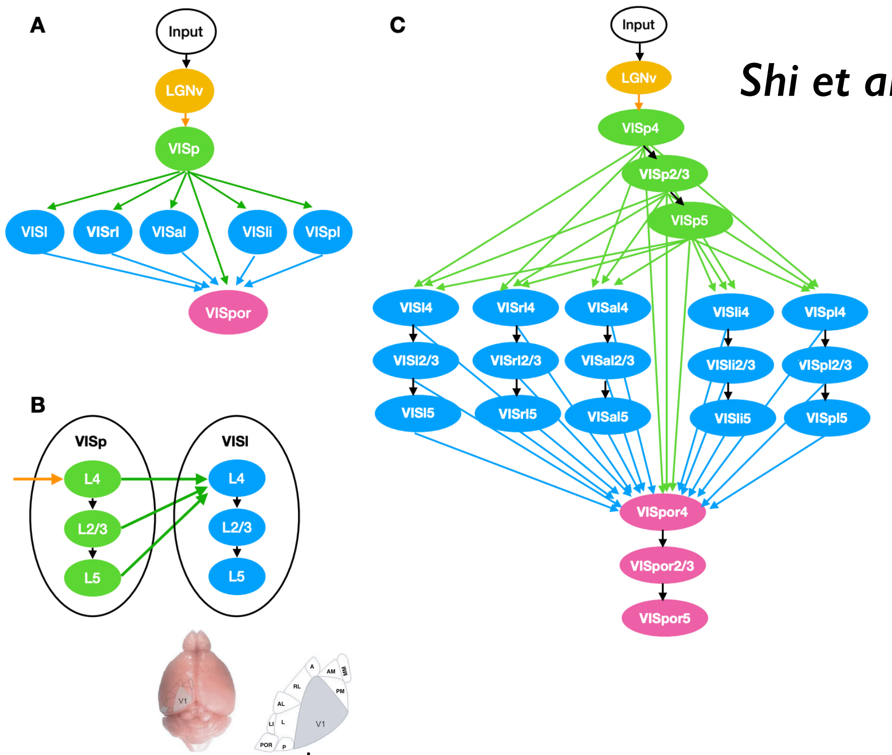


Shi et al. 2020

Model ends up *assuming* too many biological details to best match neural responses



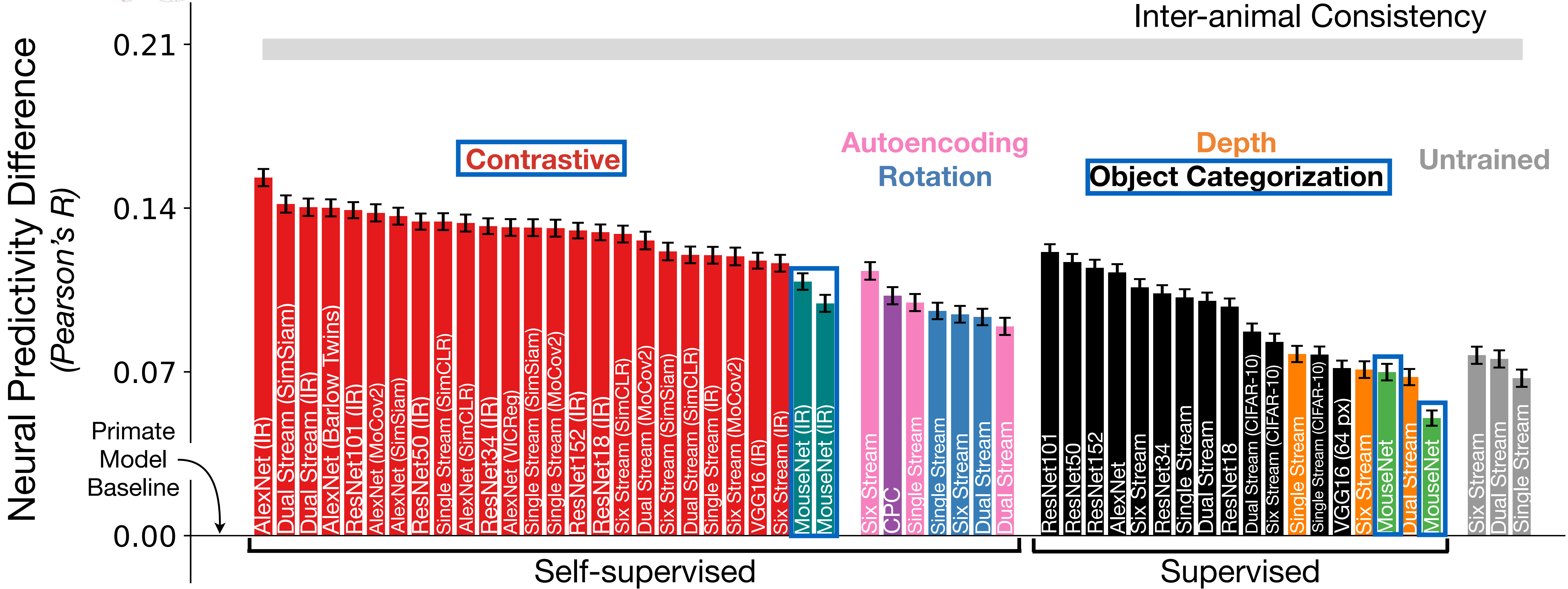
More Connectome-Inspired Architectures?



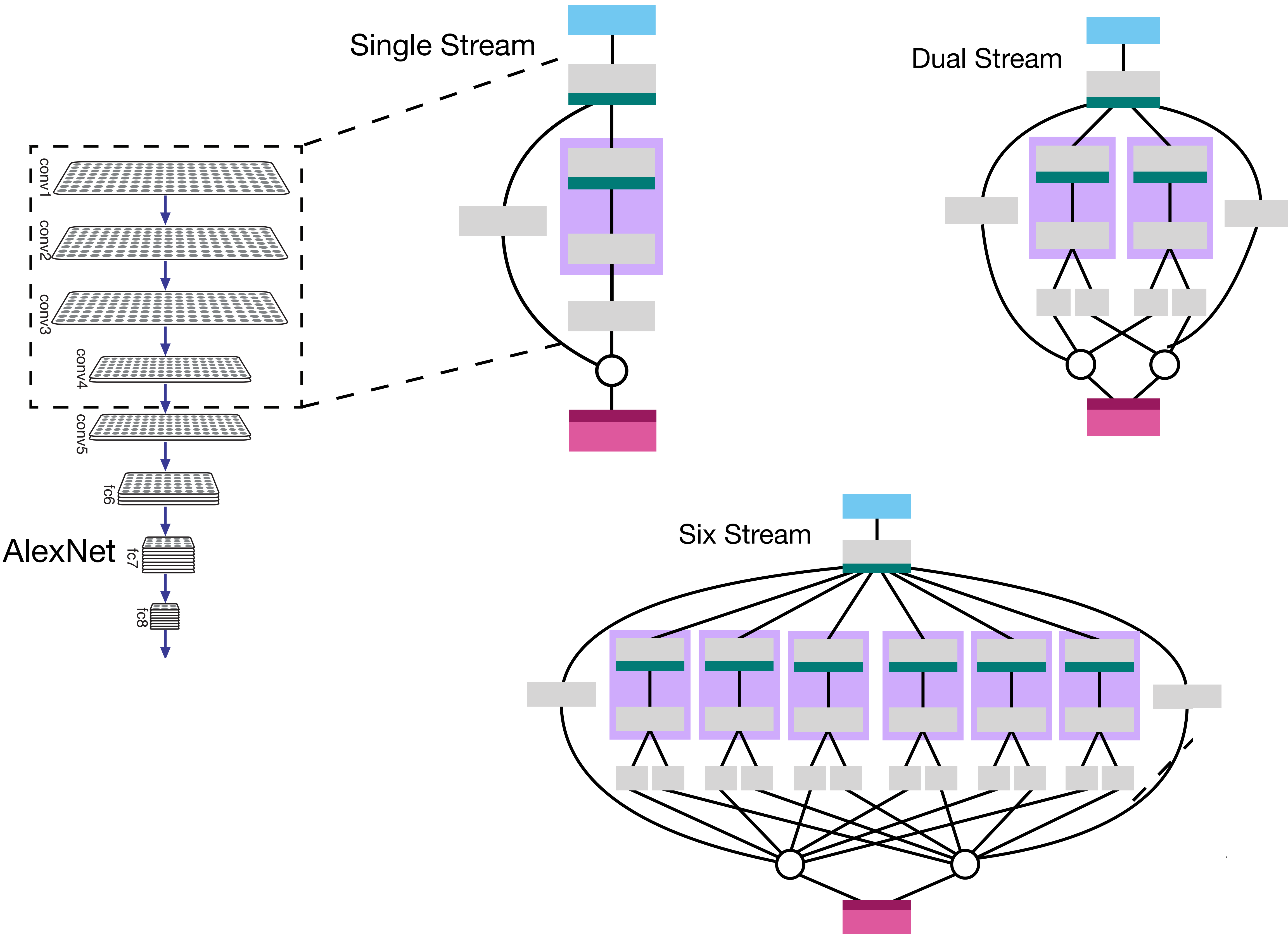
Shi et al. 2020

Model ends up *assuming* too many biological details to best match neural responses

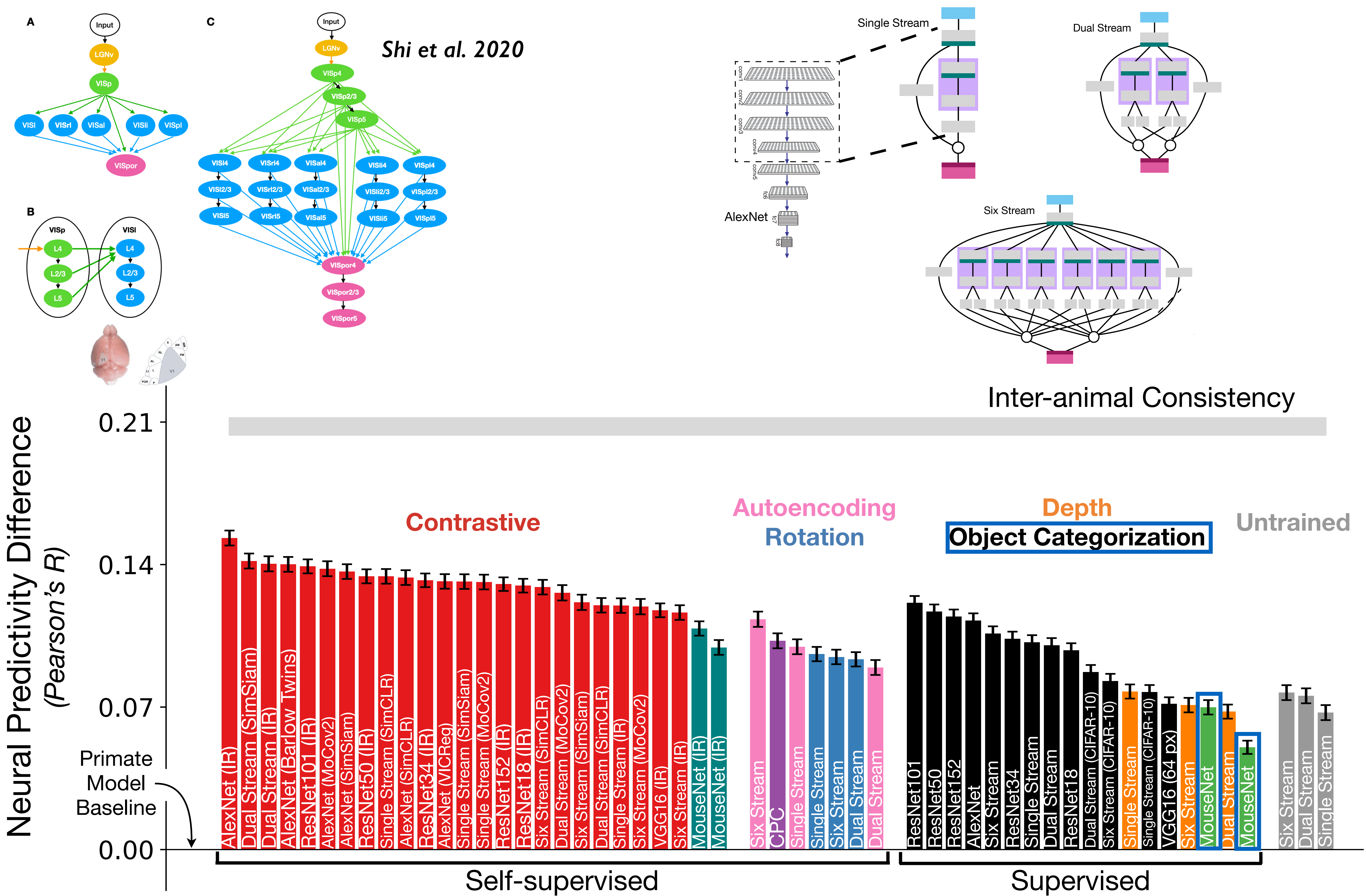
Even across loss functions



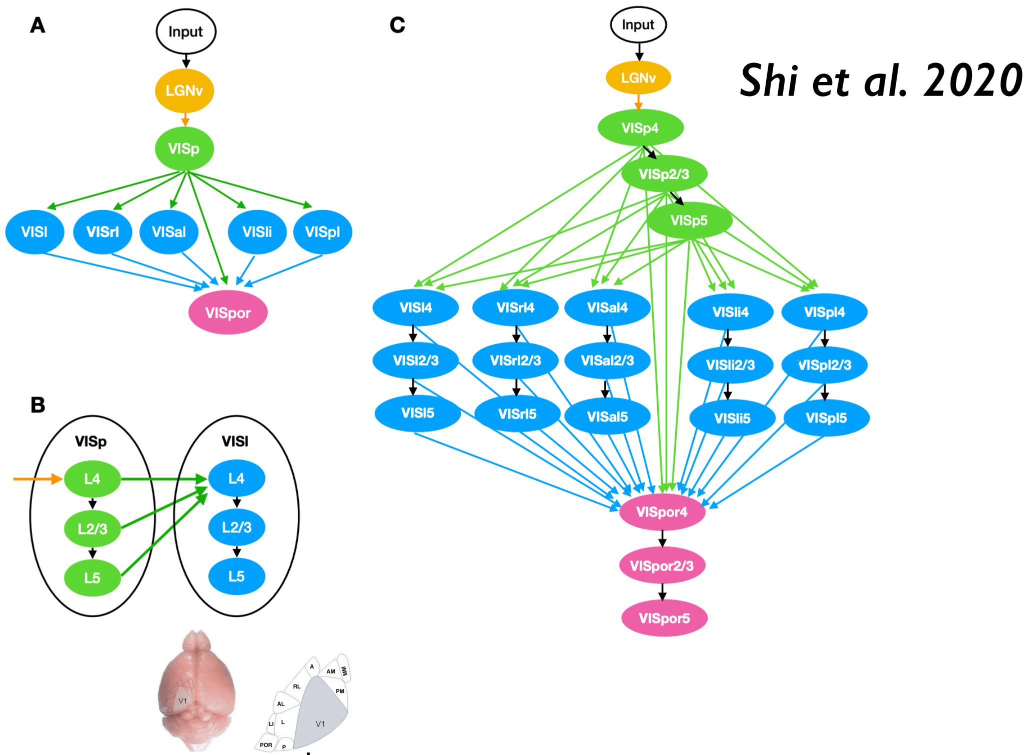
Towards Shallower Models

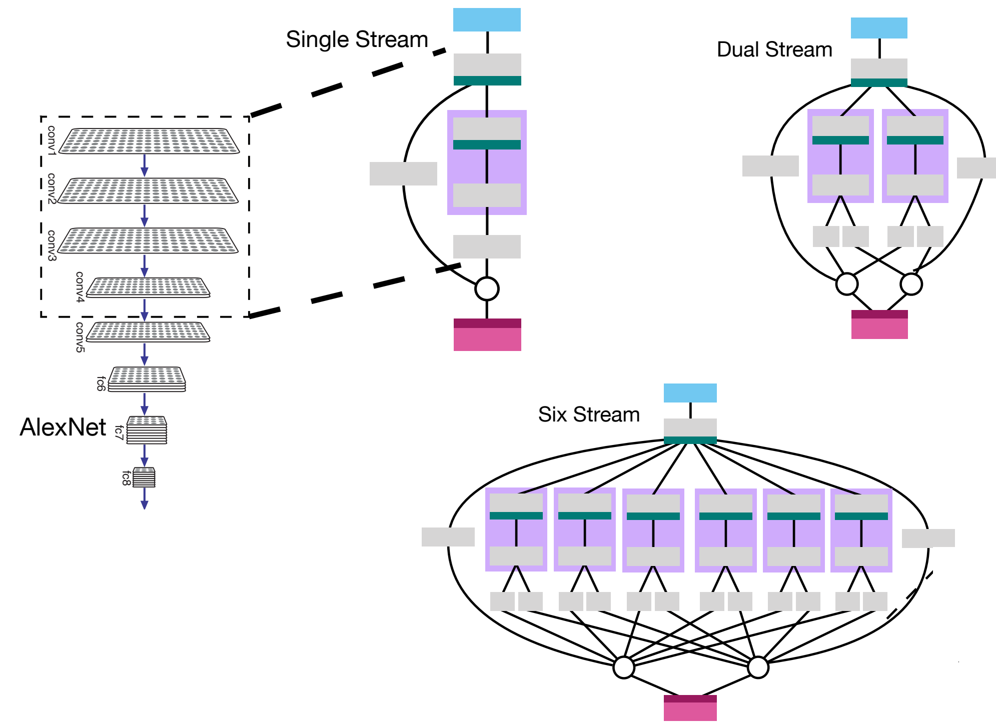


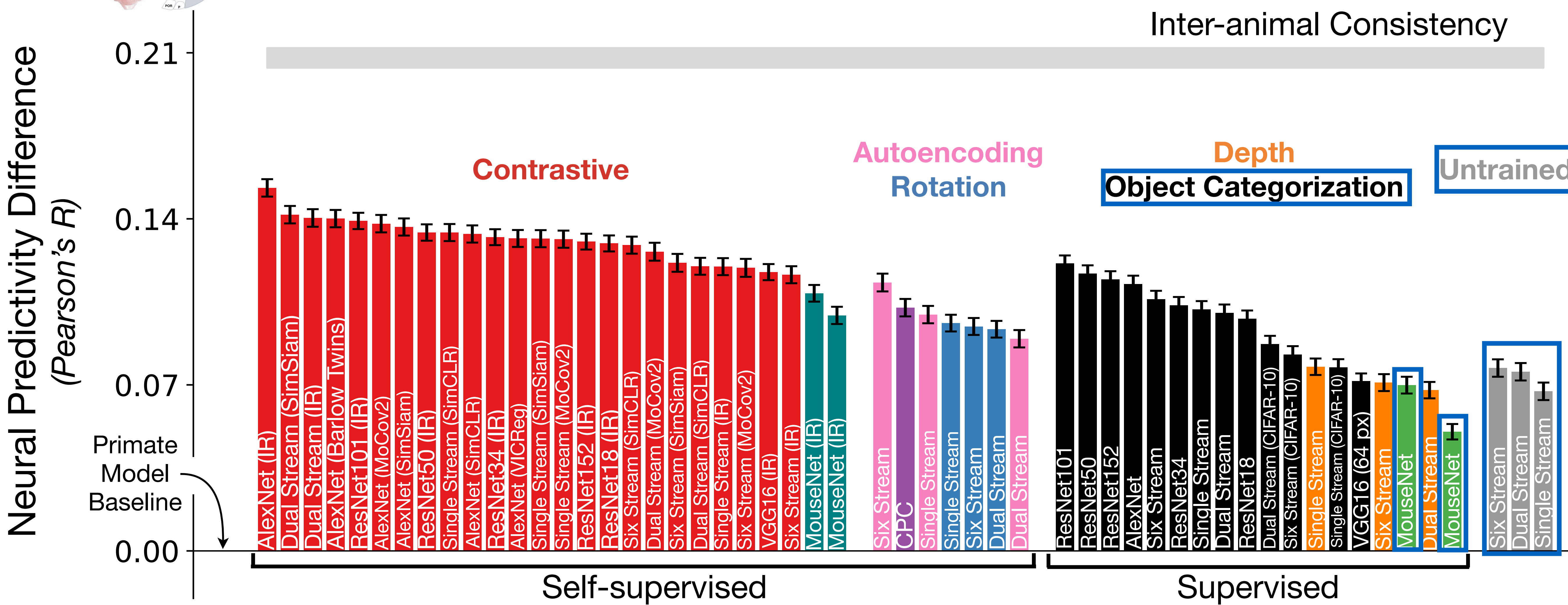
Towards Shallower Models



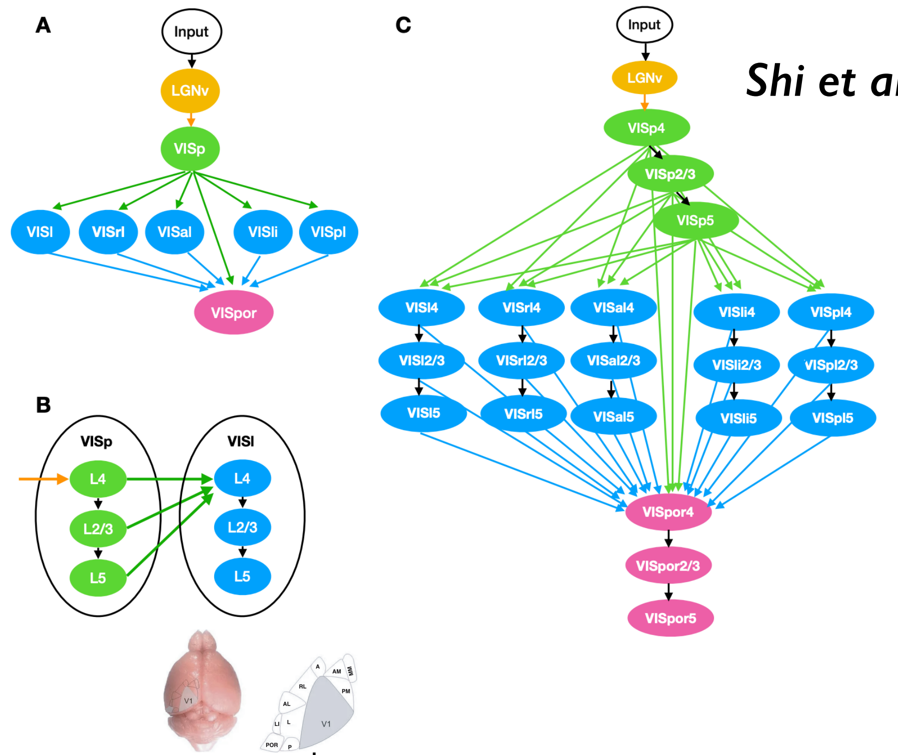
Towards Shallower Models



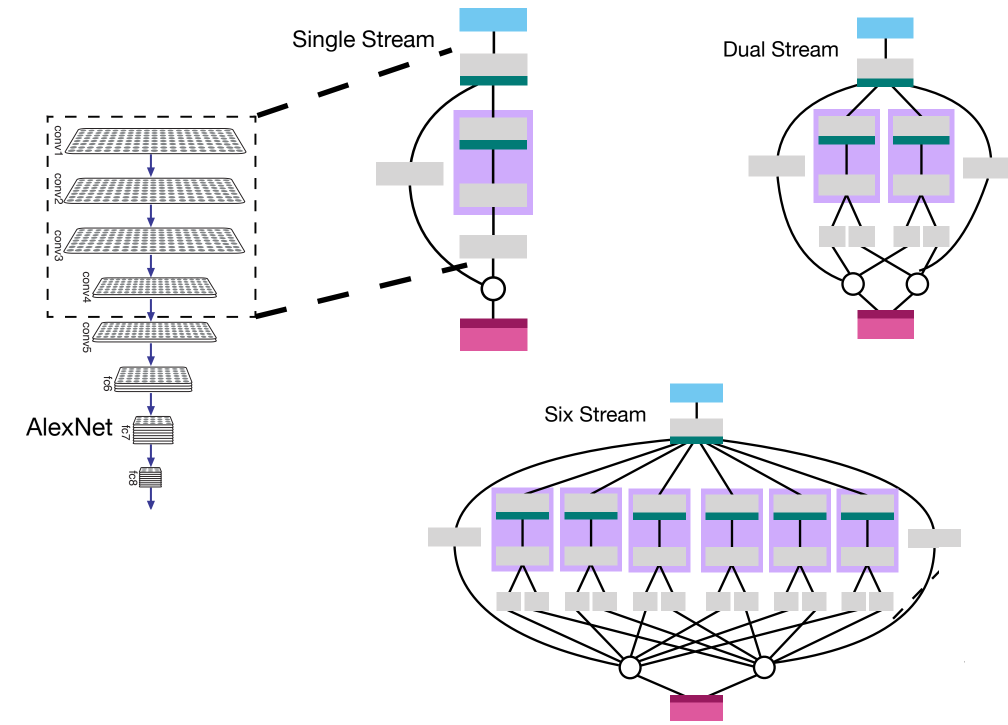


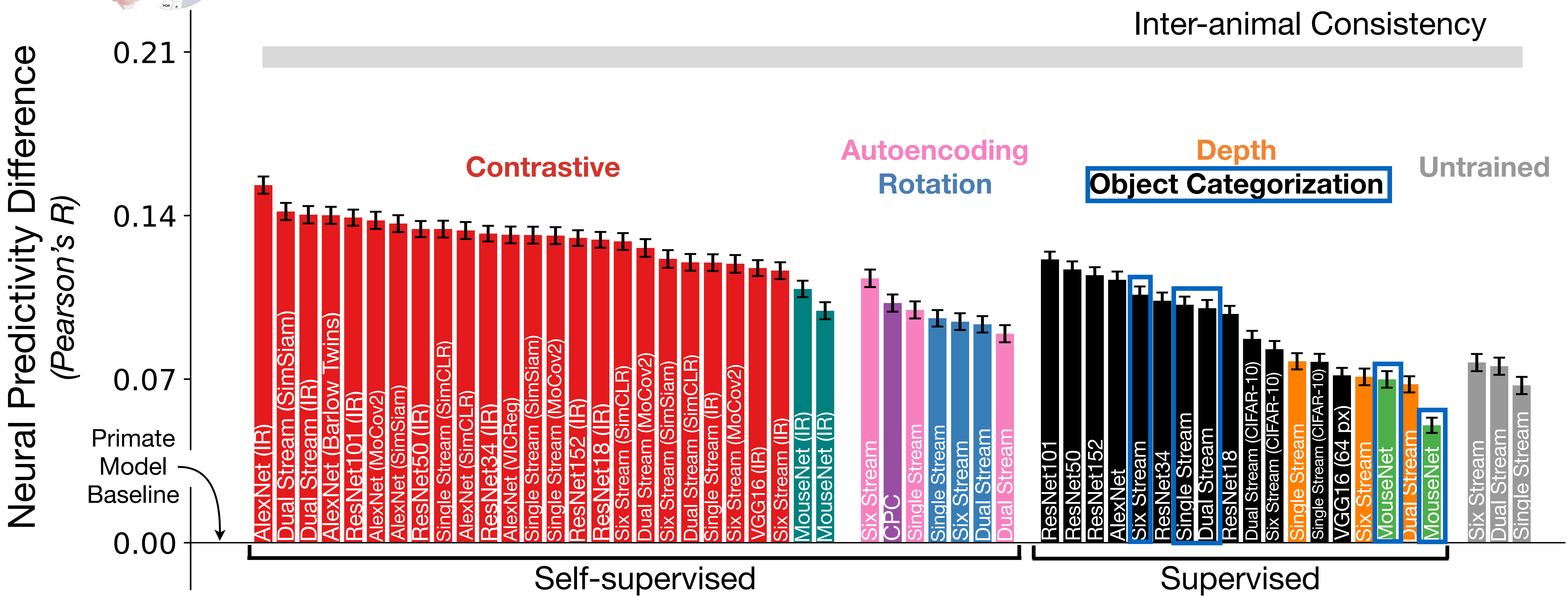


Towards Shallower Models

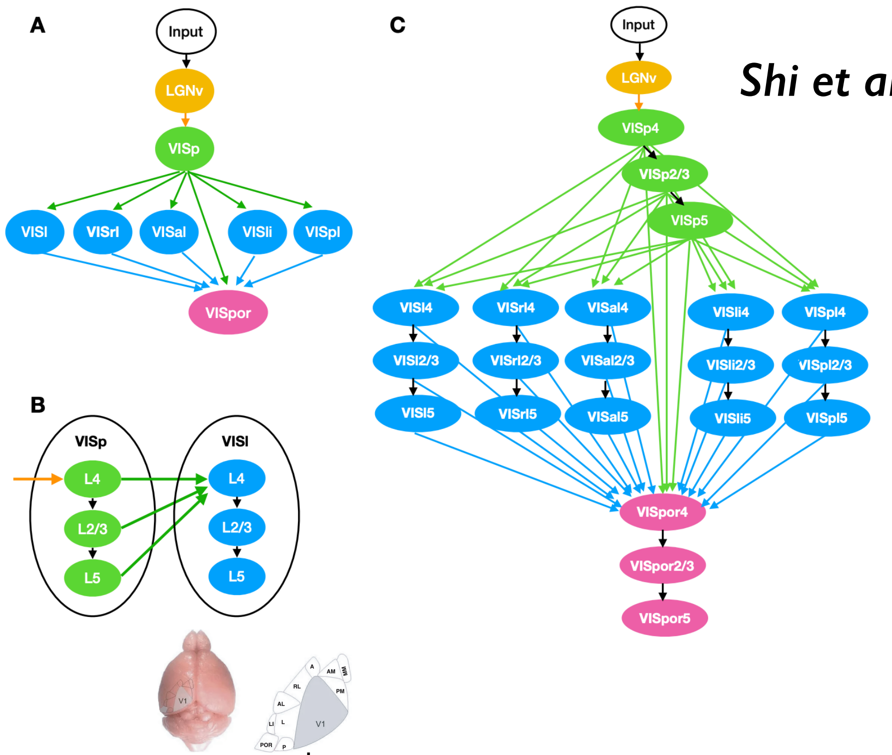


Shi et al. 2020

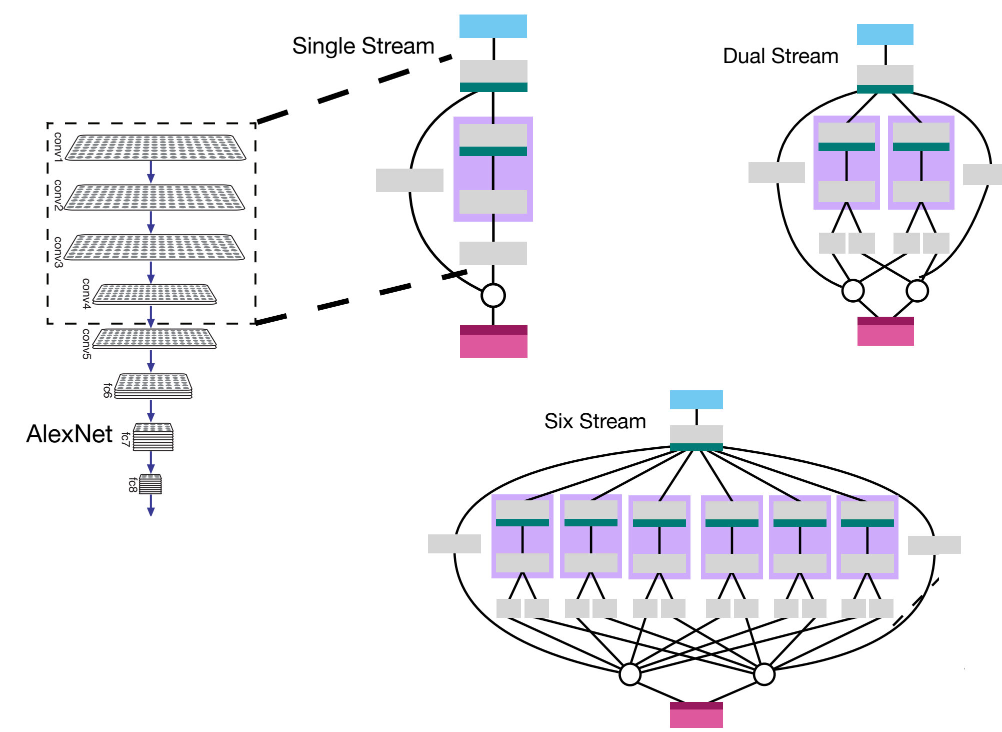


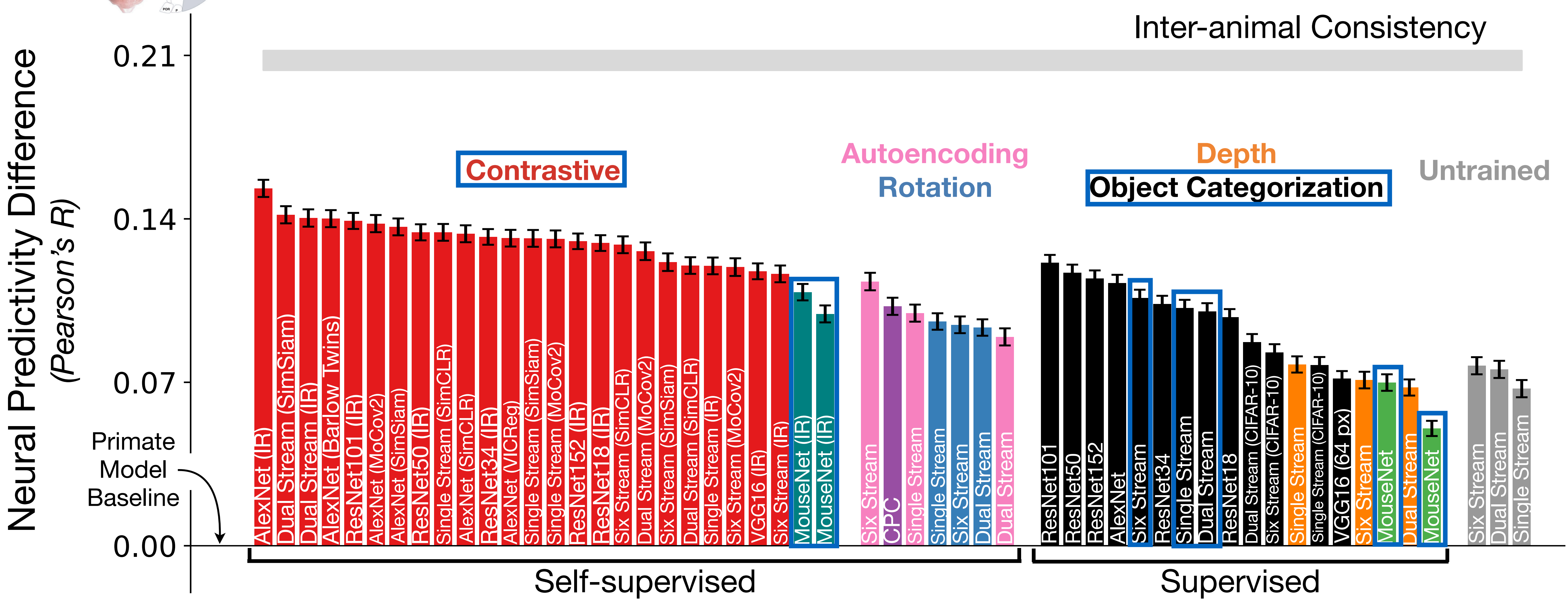


Towards Shallower Models

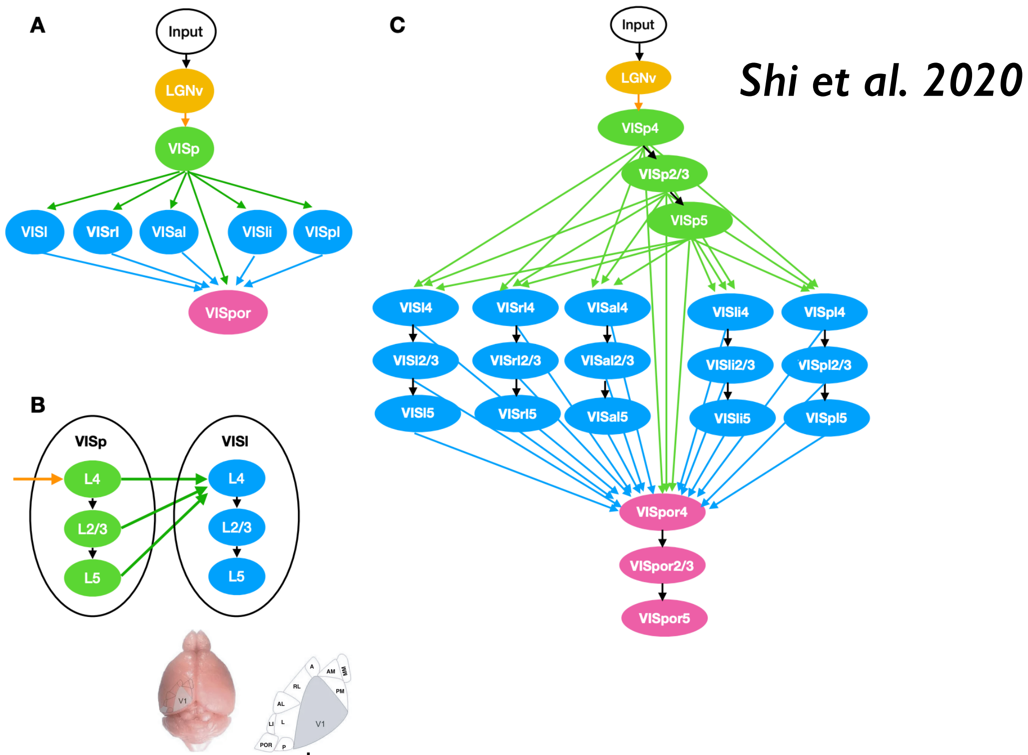


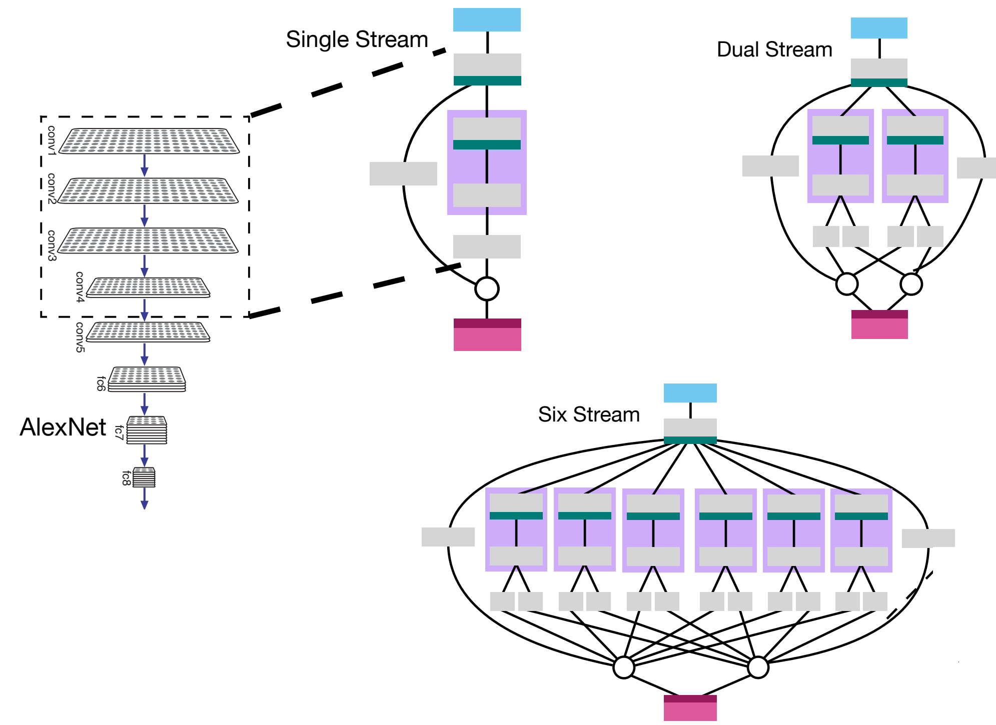
Shi et al. 2020

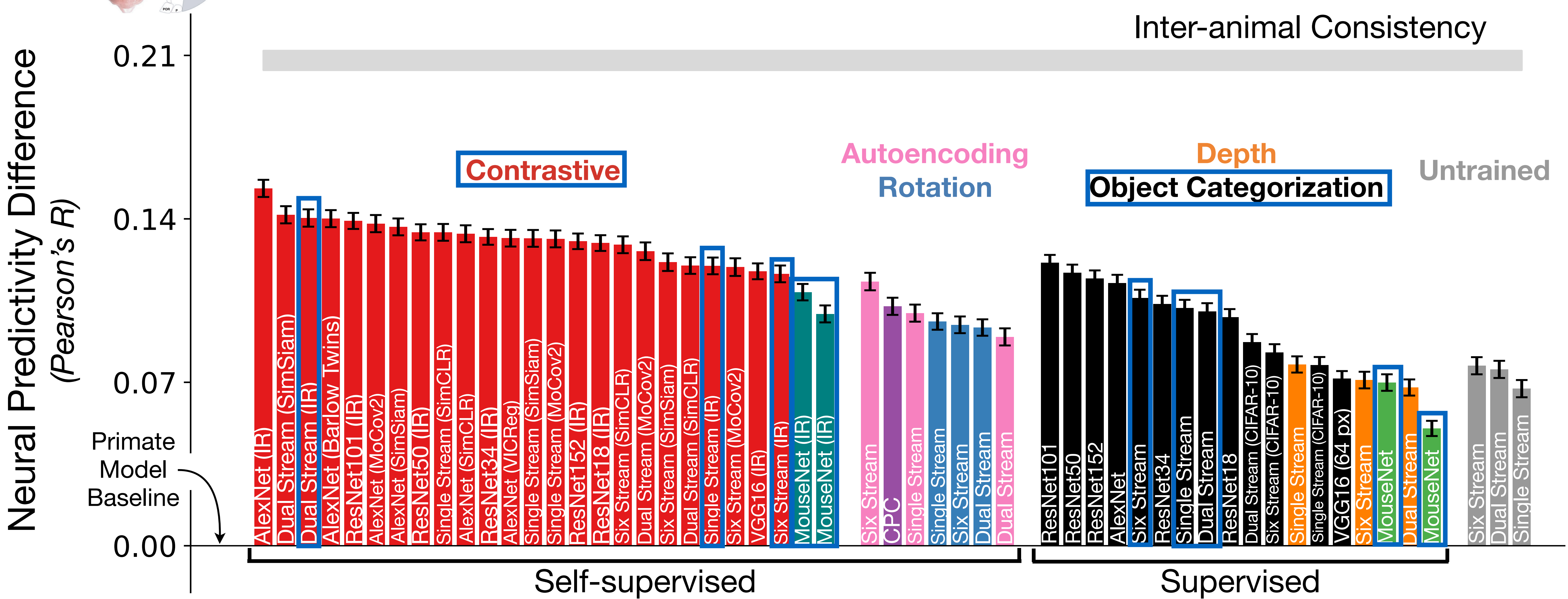




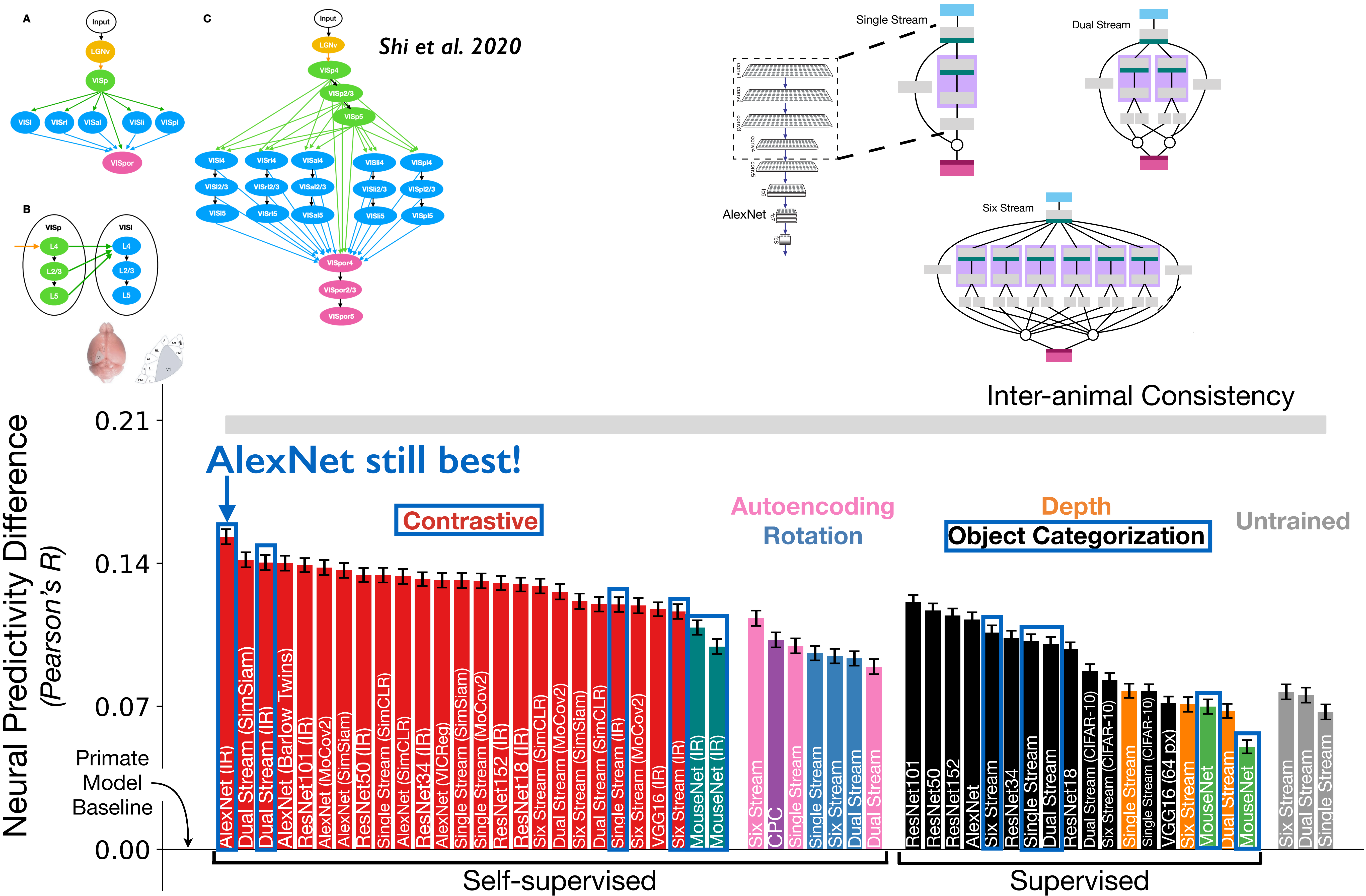
Towards Shallower Models



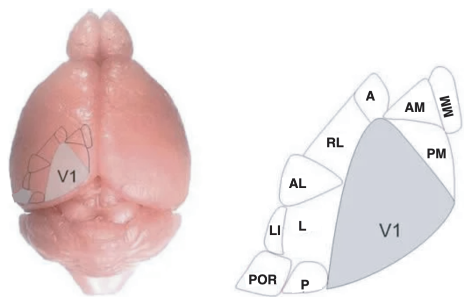


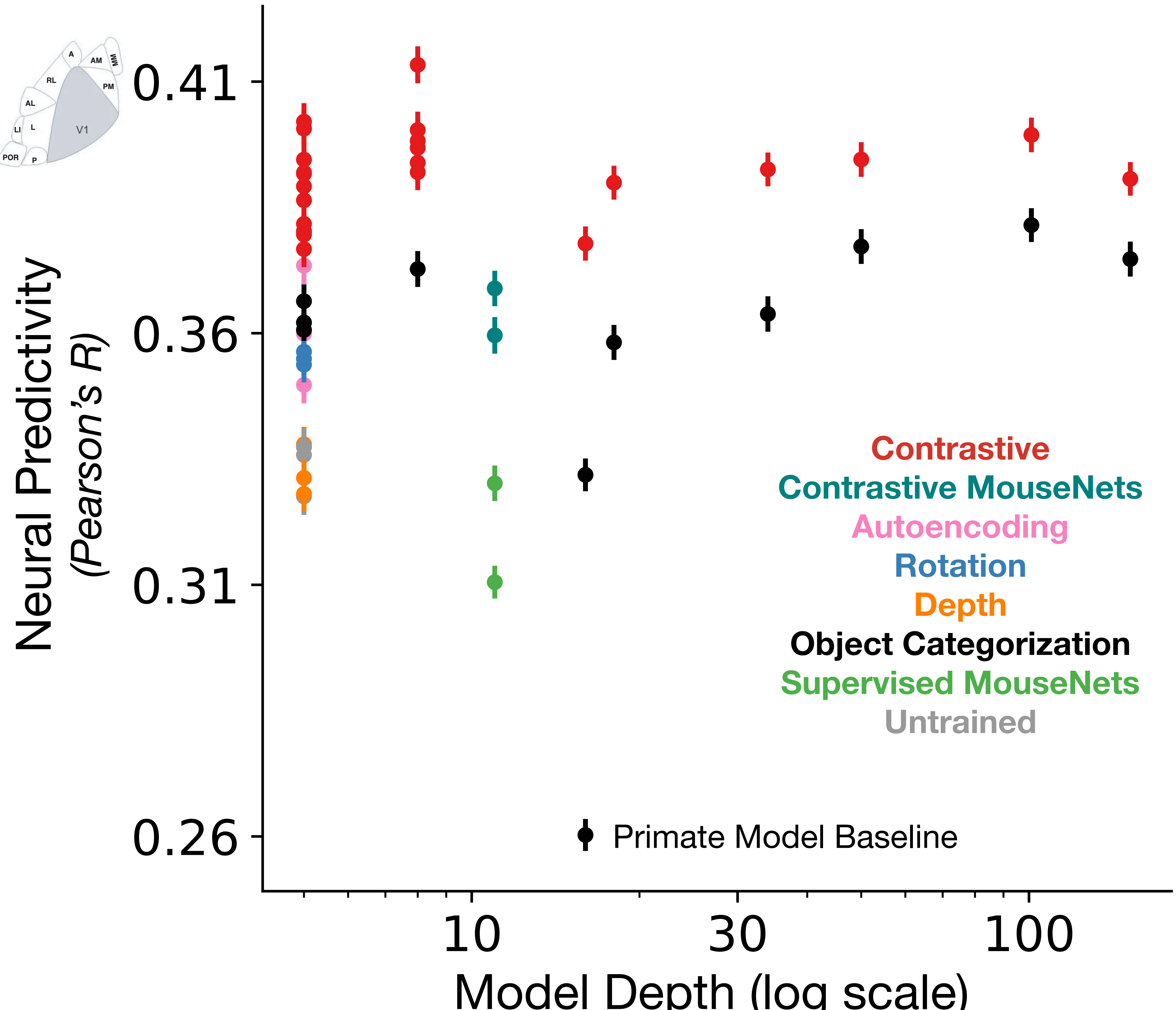


Towards Shallower Models

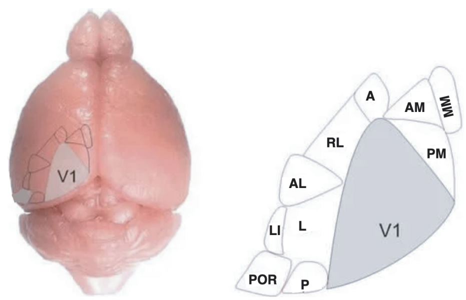


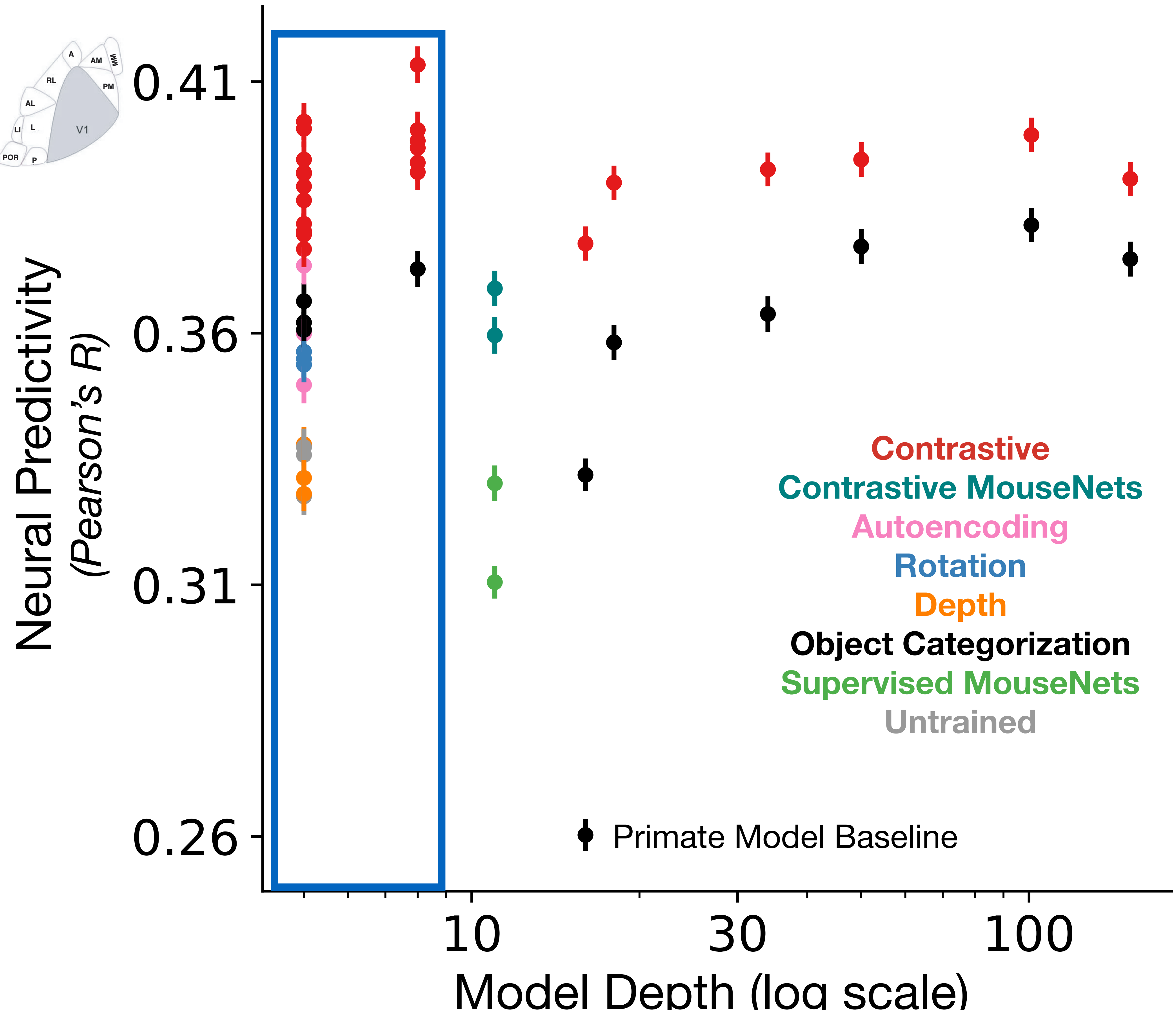
Depth as a Macroscale Architectural Constraint



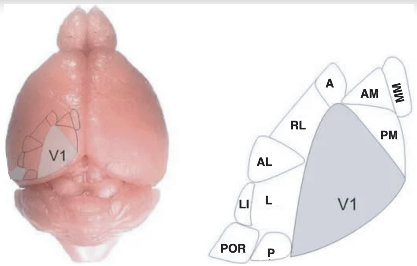


Depth as a Macroscale Architectural Constraint: Shallower models suffice

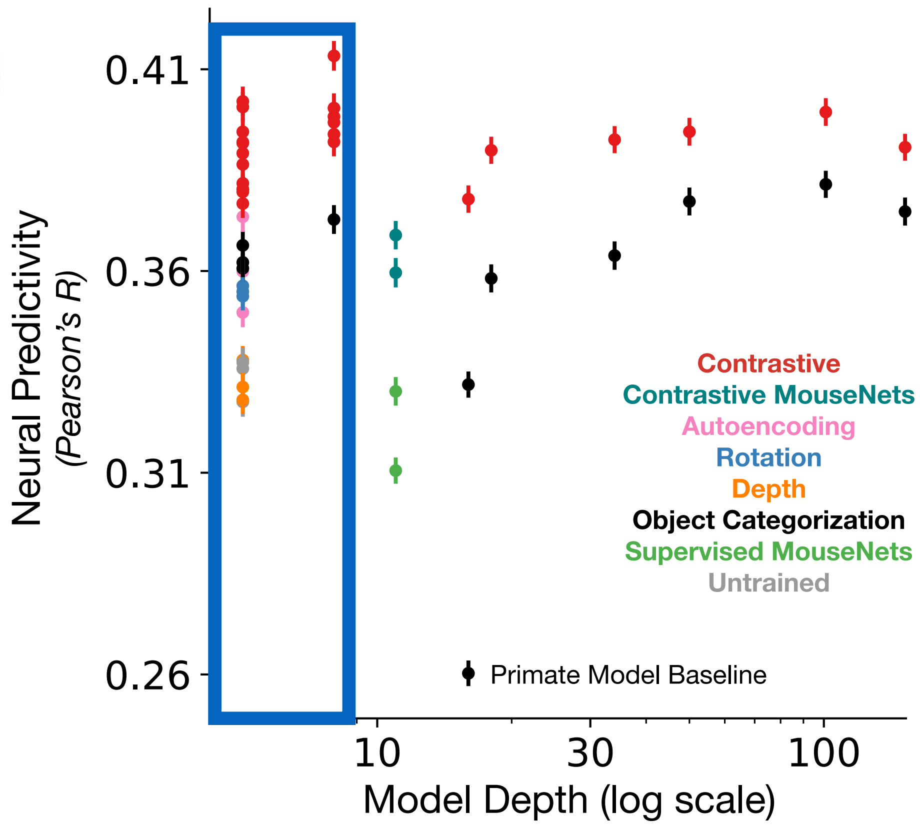


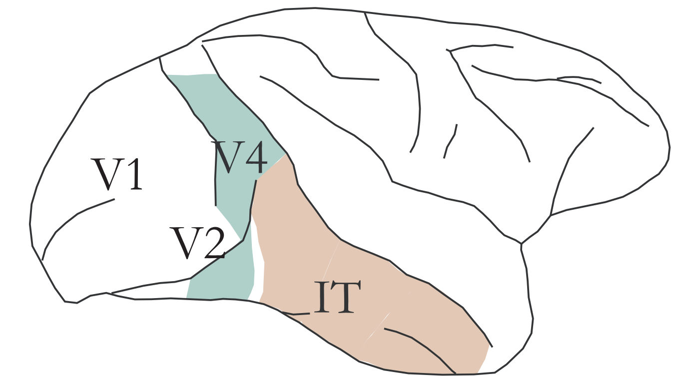


…unlike in primates!



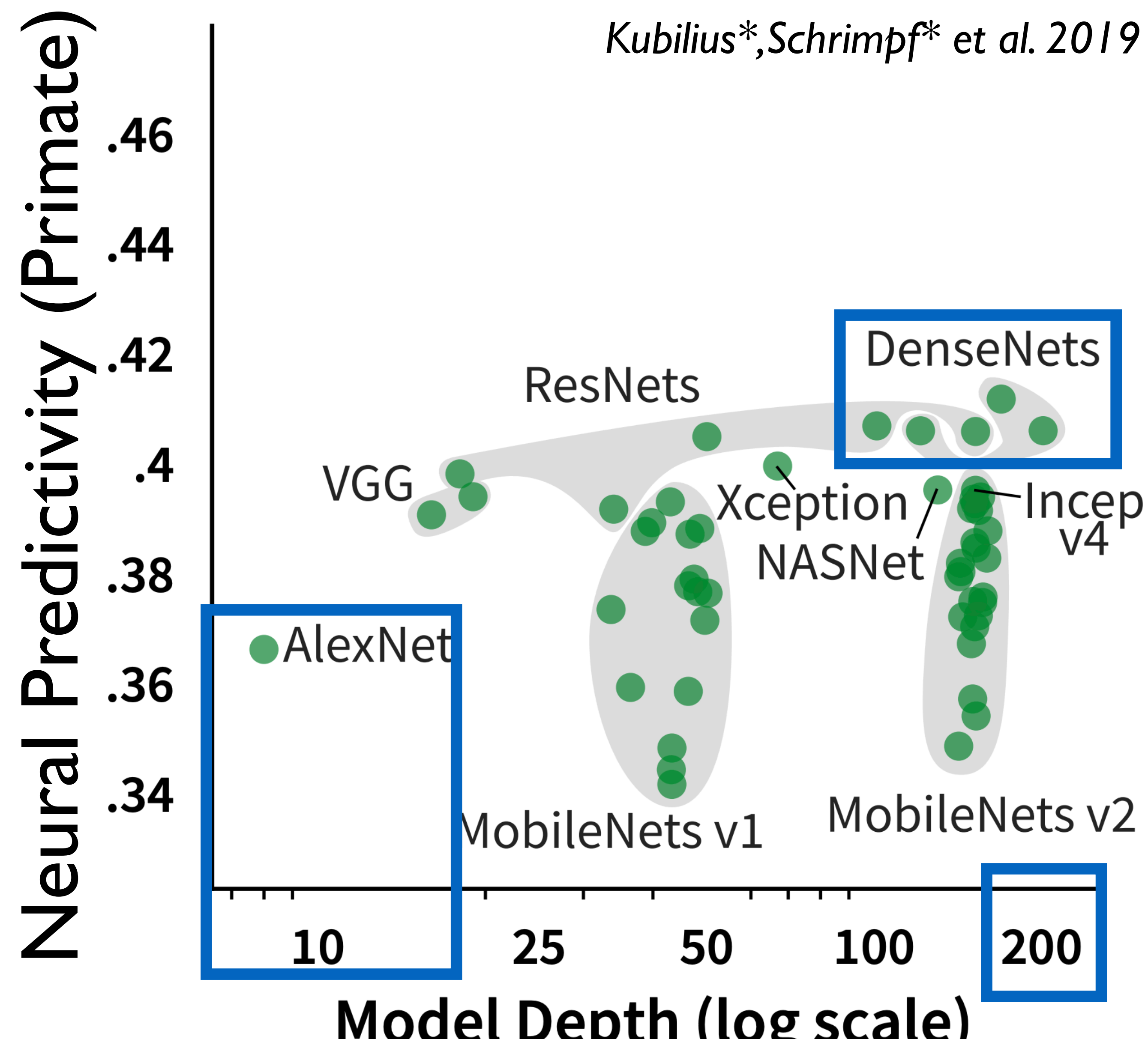
Mouse





Unlike in primates!

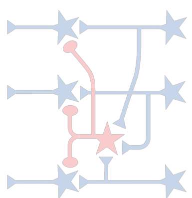
Kubilius,Schrimpf* et al. 2019*



Distilling Constraints: Environment

A *= architecture class*

1. *"Circuit"*



shallow

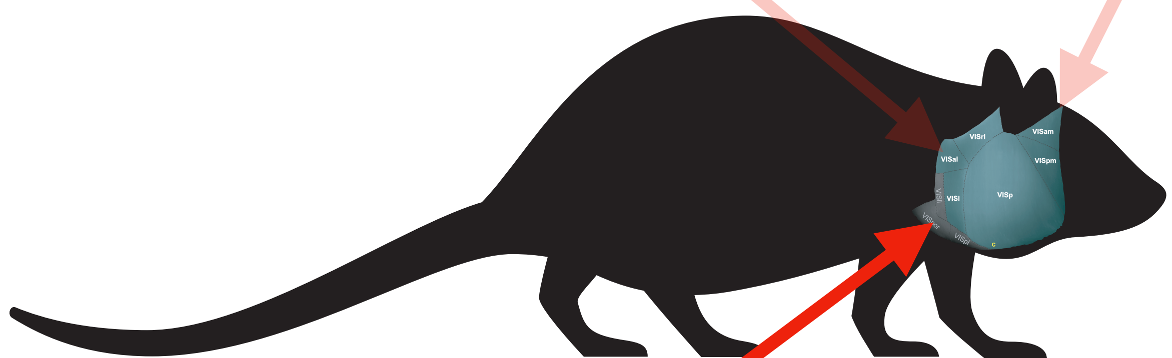
~~deep~~

T *= task loss*

3. *"Ecological niche/behavior"*

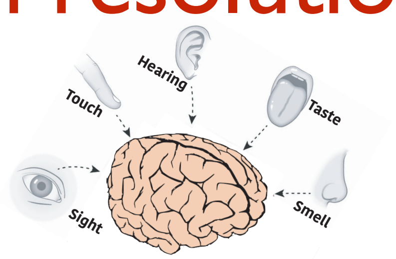
self-supervised

~~supervised~~



low resolution

~~high resolution~~

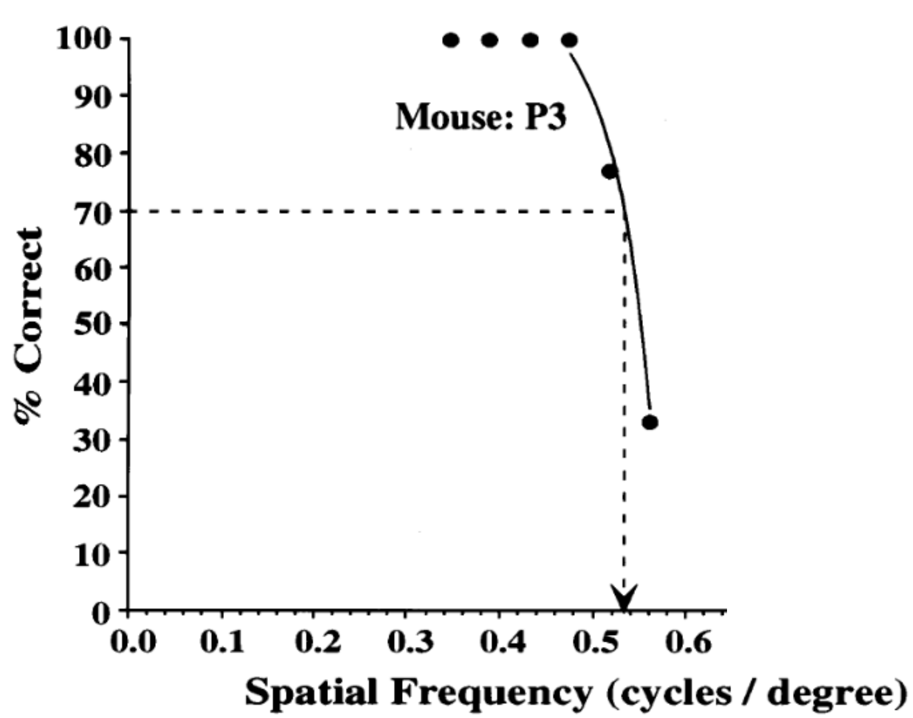


2. *"Environment"*

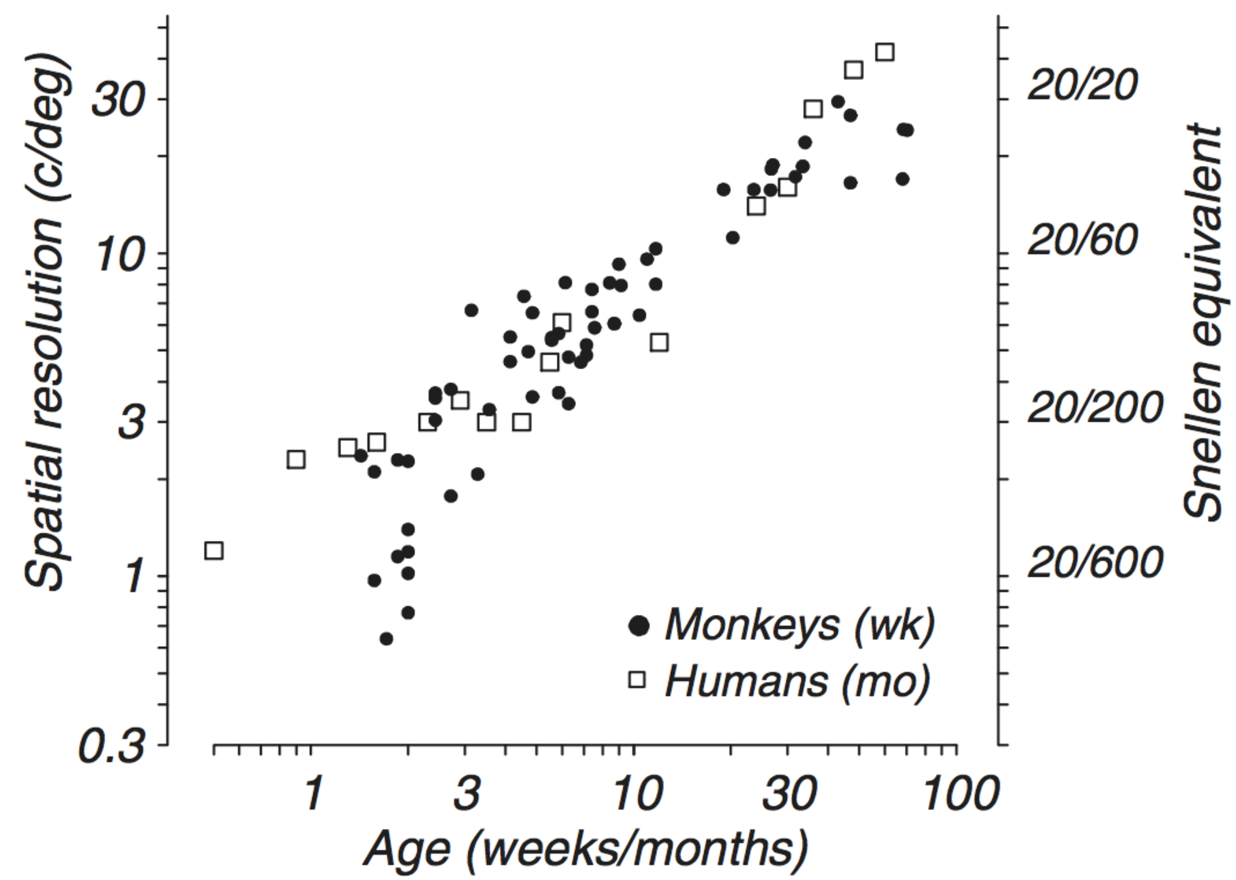
D *= data stream*

Distilling Constraints: Inputs

Mice



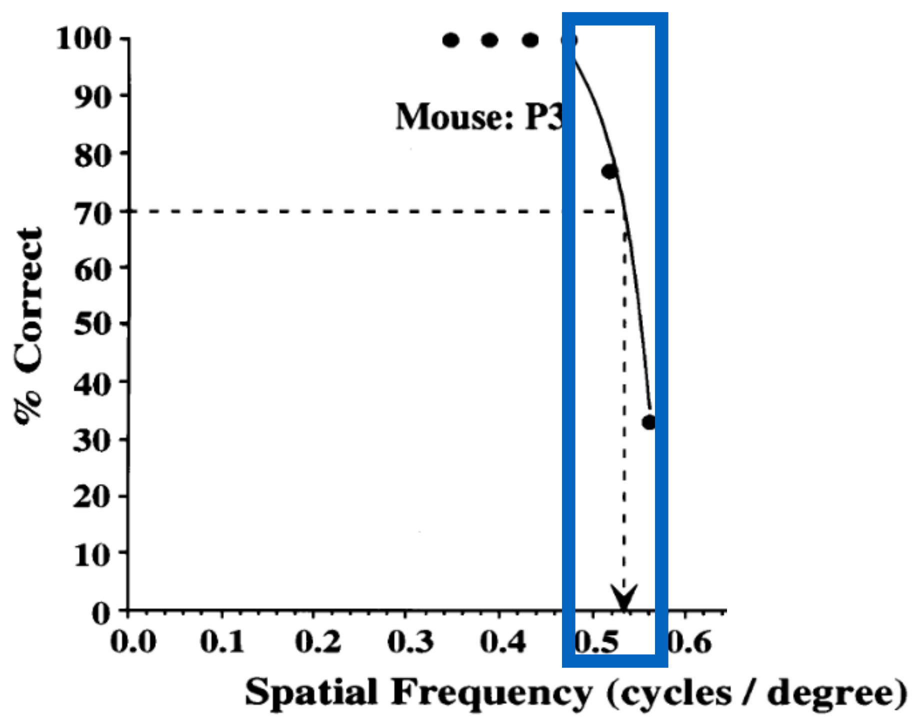
Primates



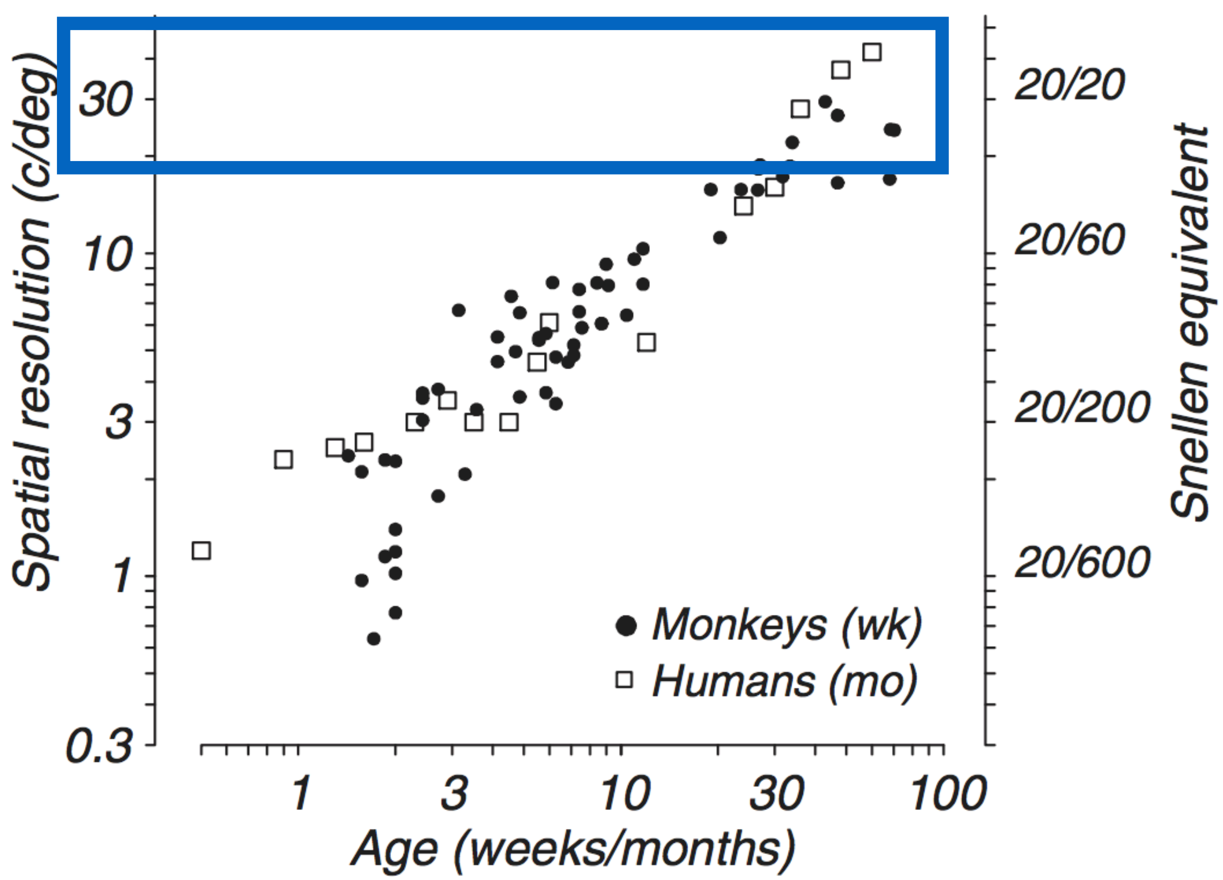
Prusky et al., 2000; Kiorpes, 2019

Distilling Constraints: Inputs

Mice

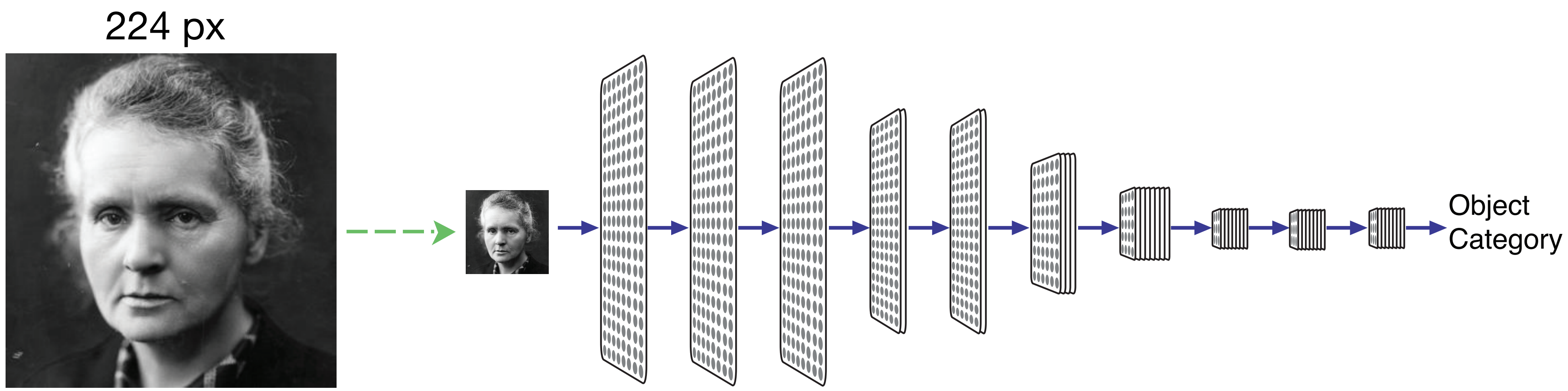


Primates

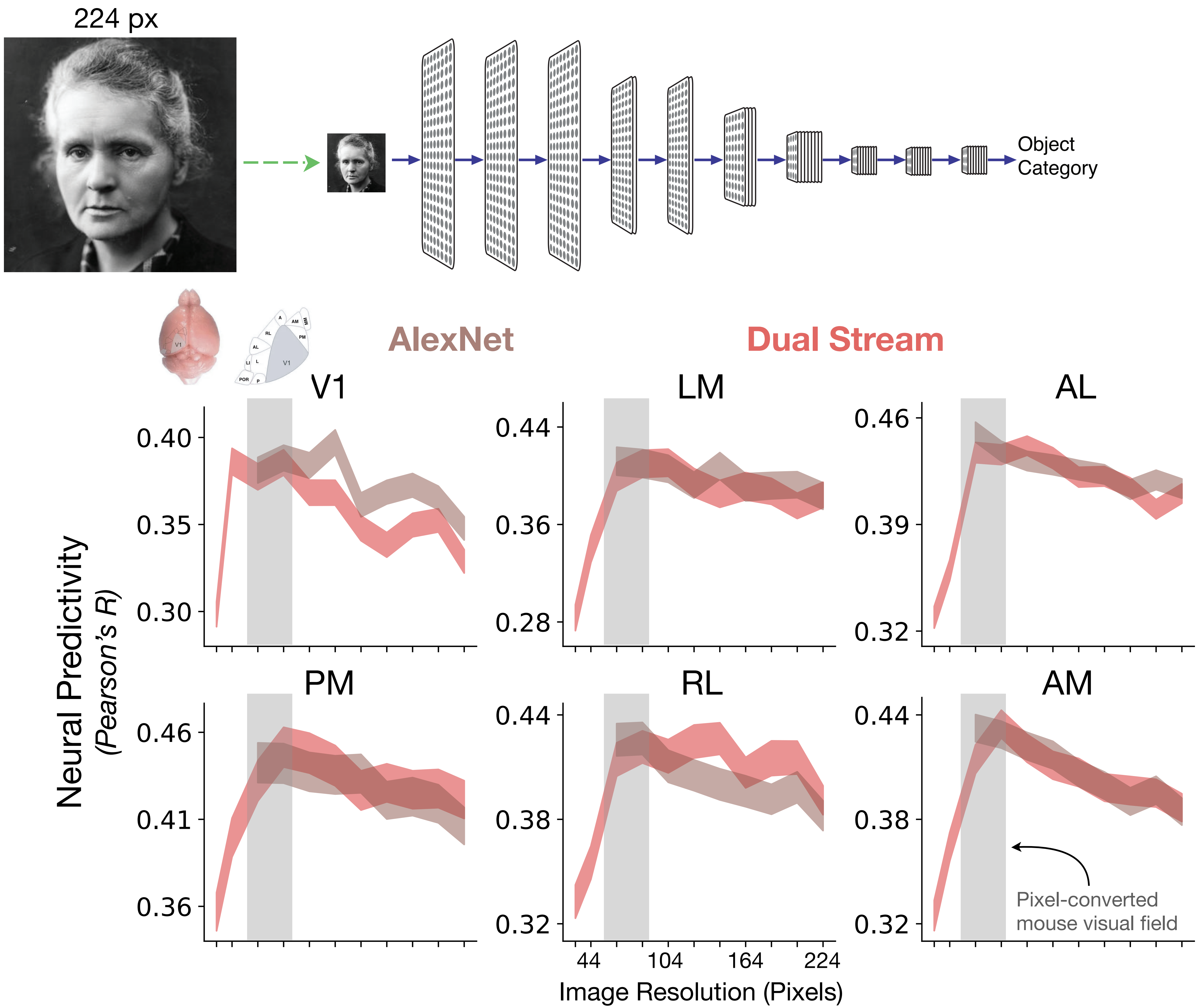


Prusky et al., 2000; Kiorpes, 2019

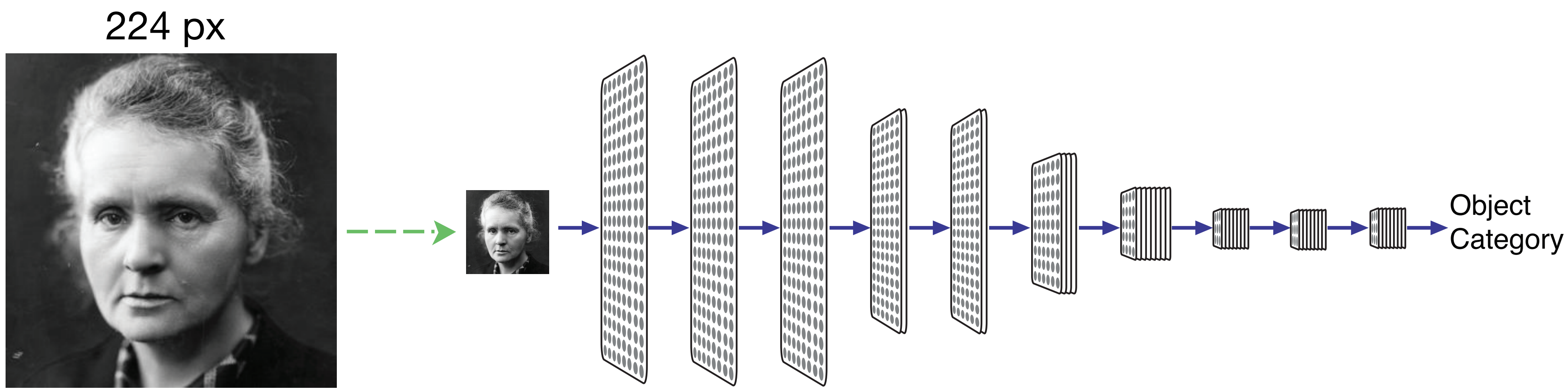
Distilling Constraints: Inputs



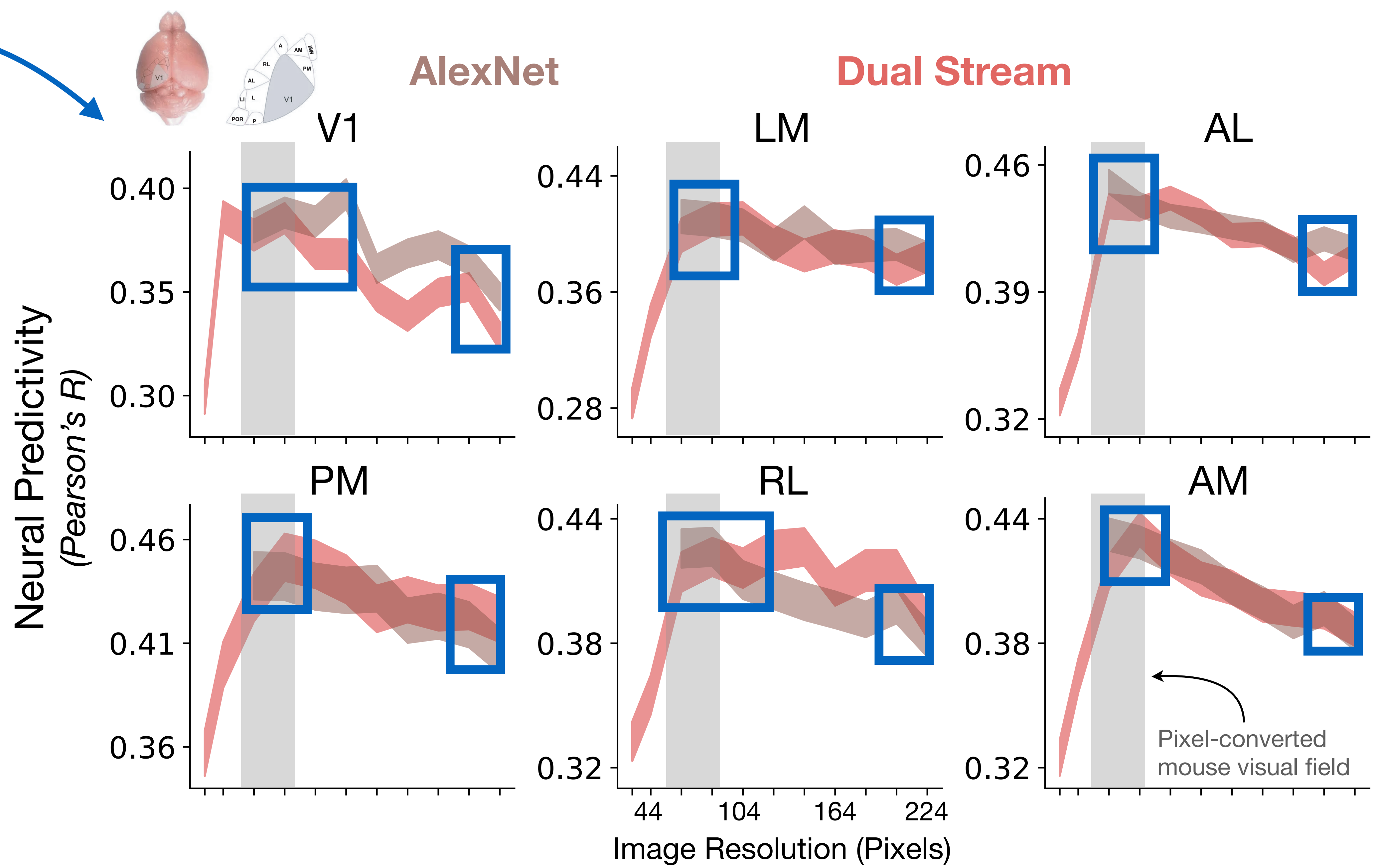
Distilling Constraints: Inputs



Lower resolution improves neural predictivity



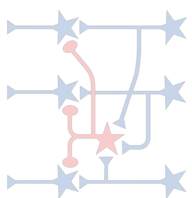
Reducing image resolution during training improves neural predictivity across visual areas



Distilling Constraints: Behavioral Goals

A = *architecture class*

1. *"Circuit"*



shallow

~~deep~~

T = *task loss*

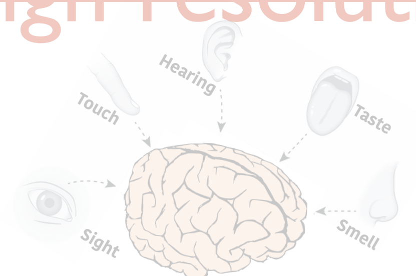
3. *"Ecological niche/behavior"*

self-supervised

~~supervised~~

low resolution

~~high resolution~~



2. *"Environment"*

D = *data stream*

Distilling Constraints: Behavioral Goals

Supervised Losses

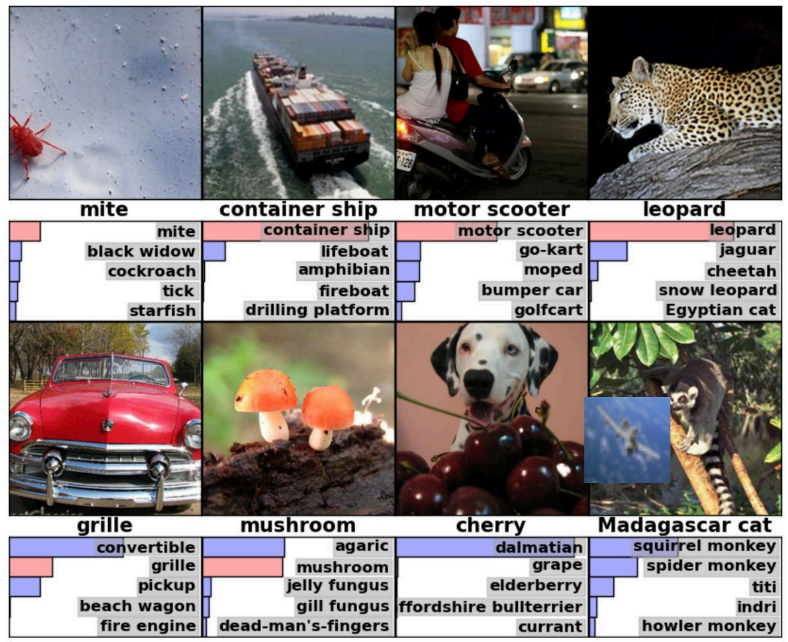
Distilling Constraints: Behavioral Goals

Supervised Losses

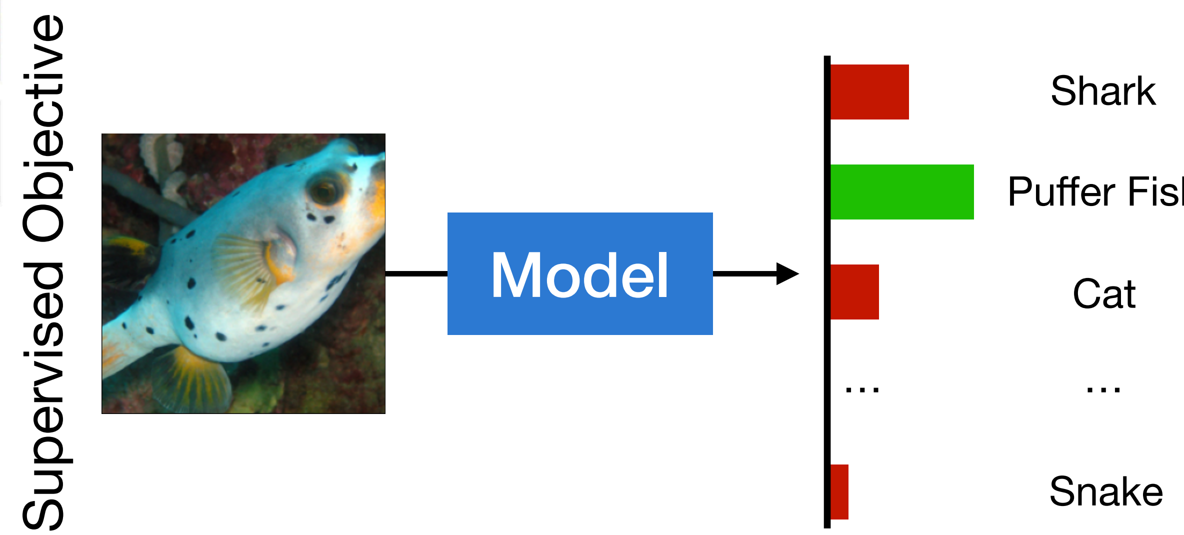
ImageNet Challenge

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



Most Supervision (1000 classes)



Typical setting: supervision with (1000) category labels ...but is very "unnatural" for mice!

Both the type and number of categories is unrealistic for mice

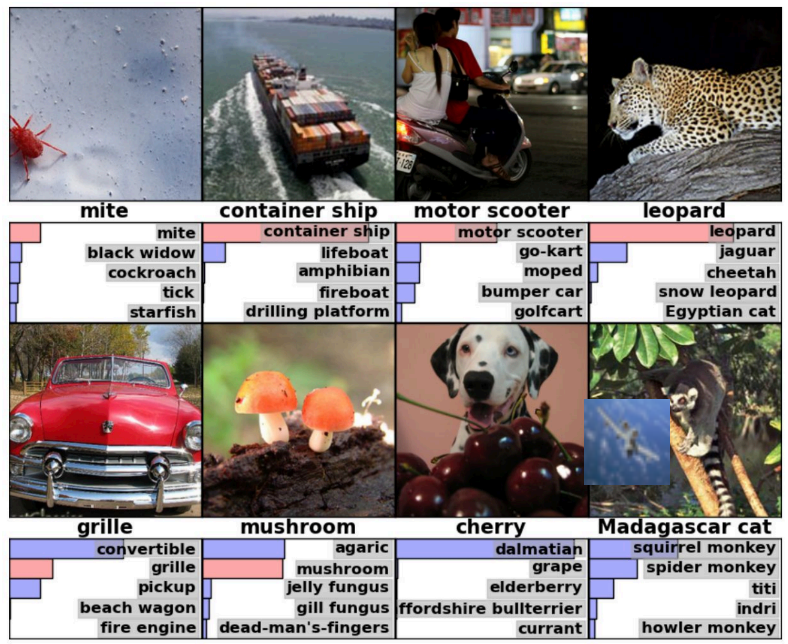
Distilling Constraints: Behavioral Goals

Supervised Losses

ImageNet Challenge

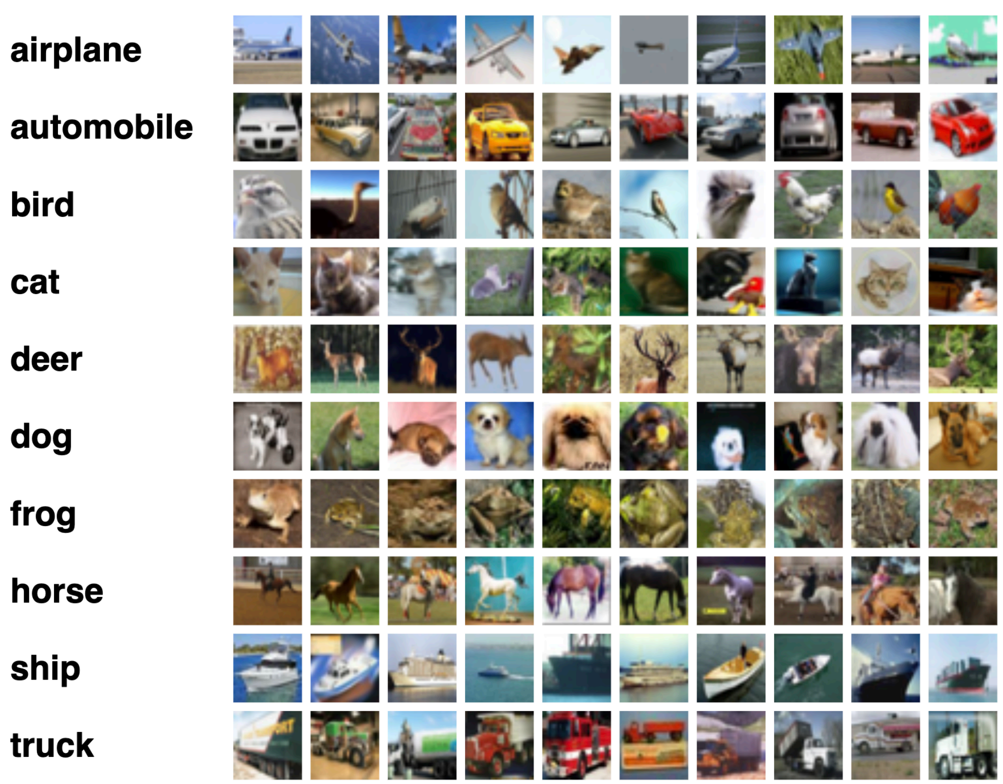
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



Most Supervision
(1000 classes)

CIFAR-10



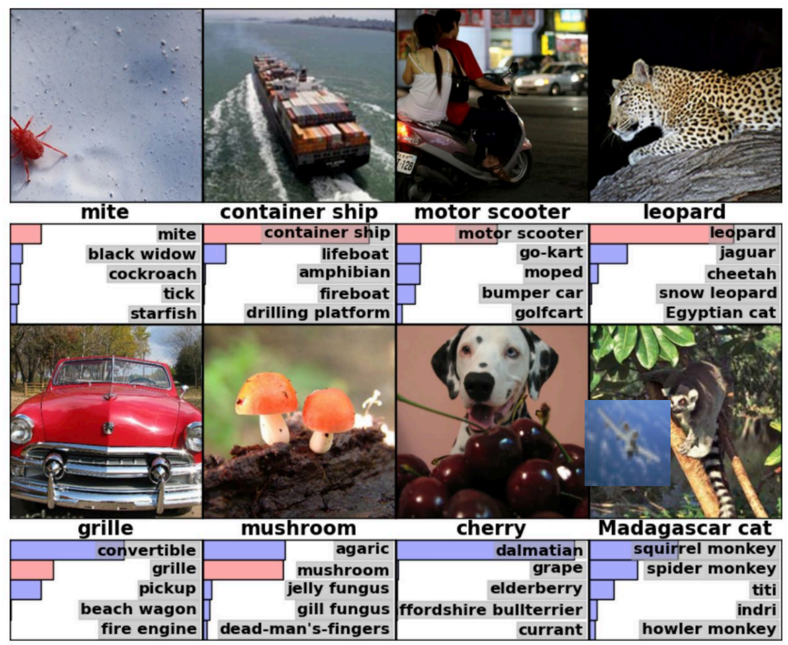
Less Supervision
(10 object classes)

Supervised Losses

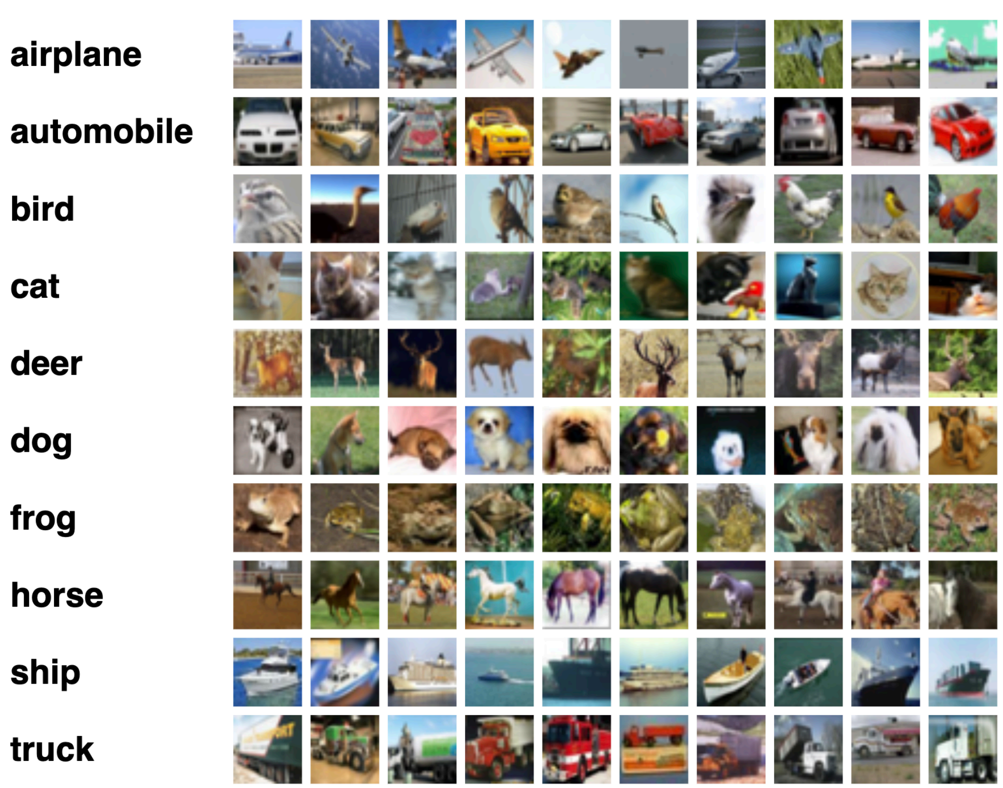
ImageNet Challenge

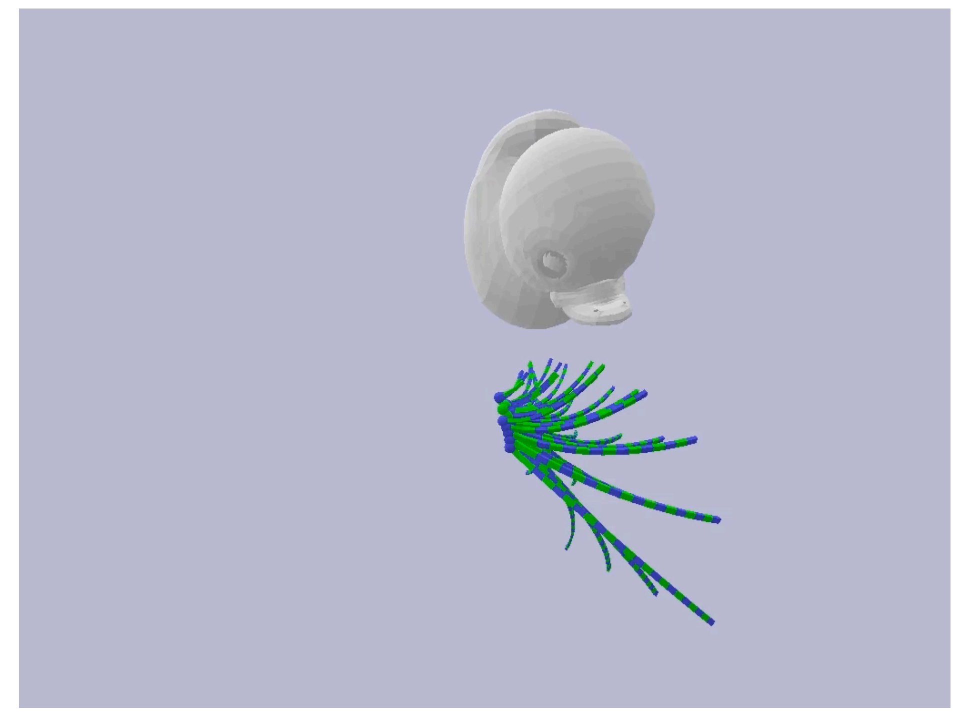
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



CIFAR-10





Depth Map Prediction
(Visual proxy for whisking)

PBRNet (500K images)

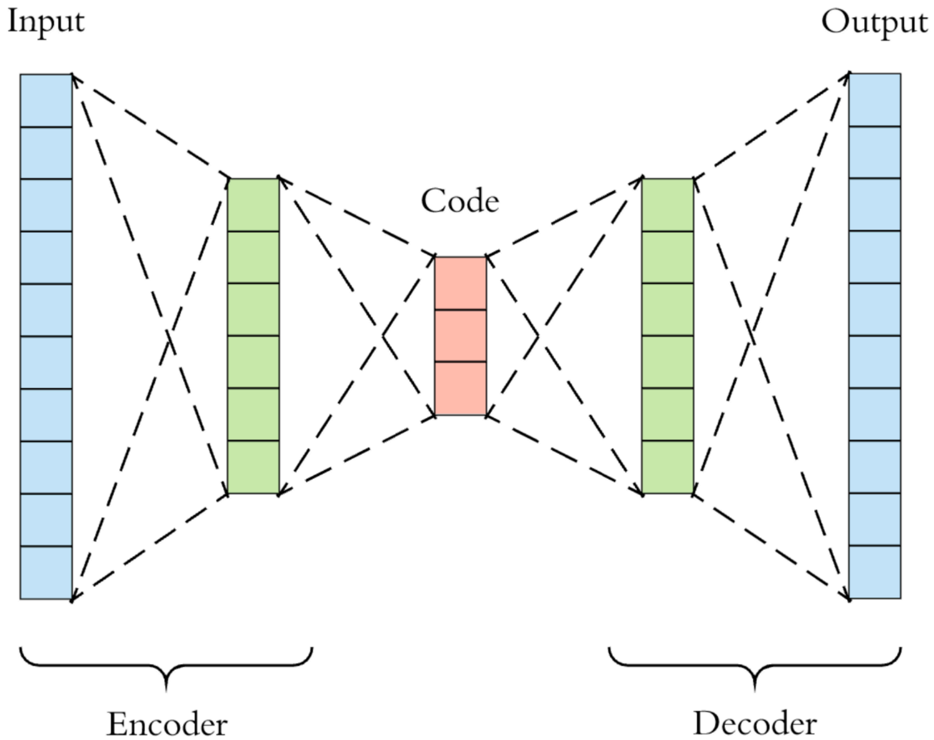
Distilling Constraints: Behavioral Goals

Self-supervised Losses

Distilling Constraints: Behavioral Goals

Self-supervised Losses

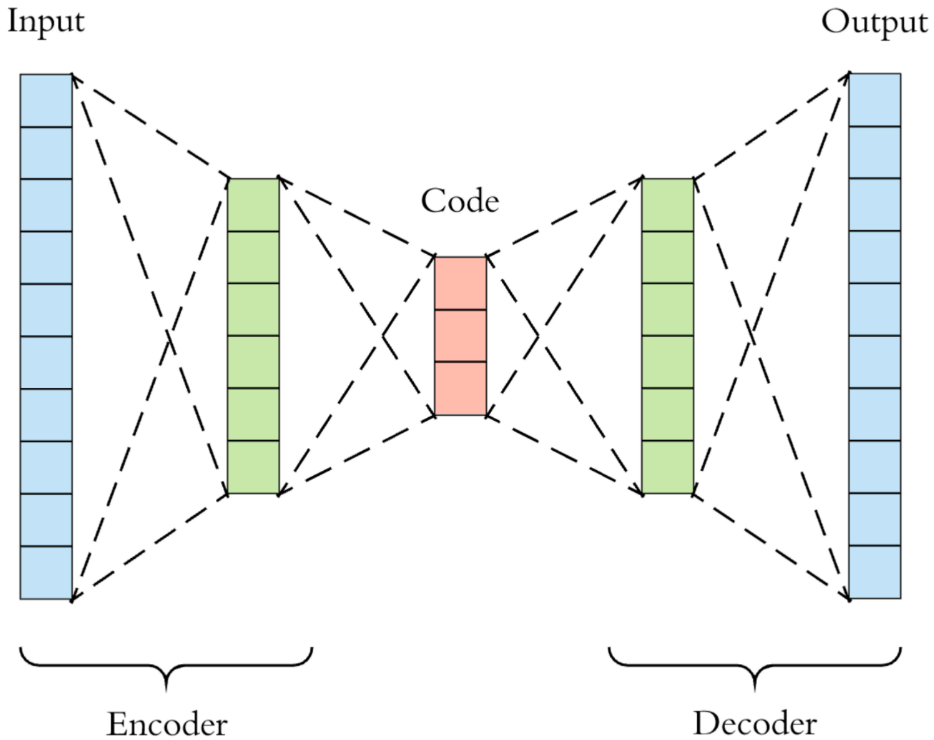
<u>Sparse Autoencoding</u>

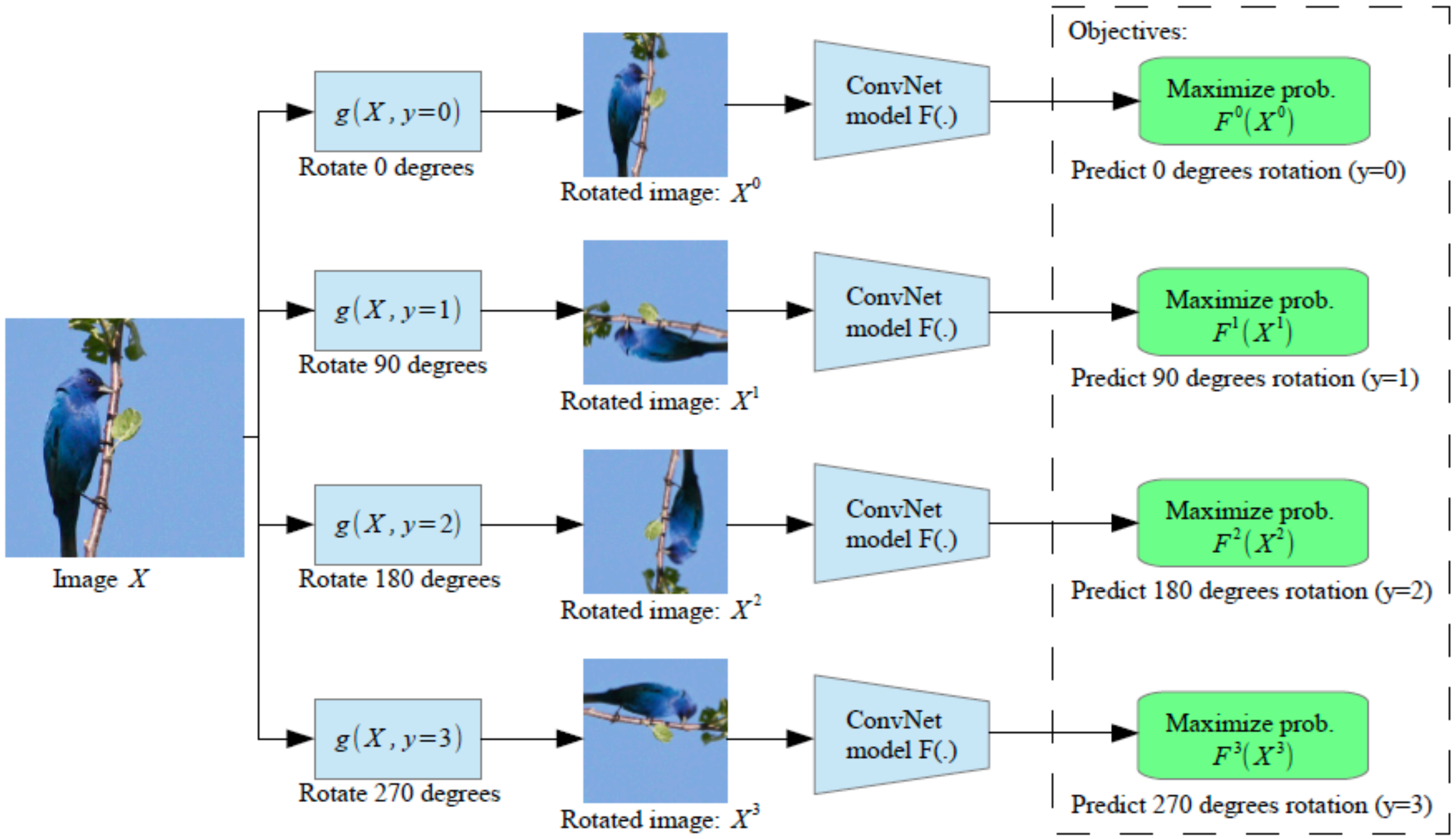


Distilling Constraints: Behavioral Goals

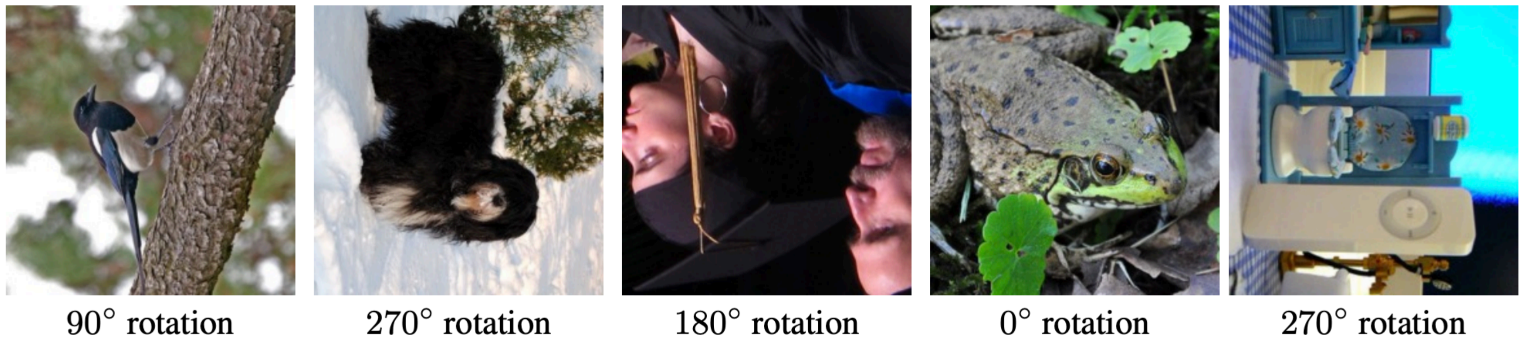
Self-supervised Losses

Sparse Autoencoding





Predict Image Rotations (RotNet, Gidaris et al. 2018)

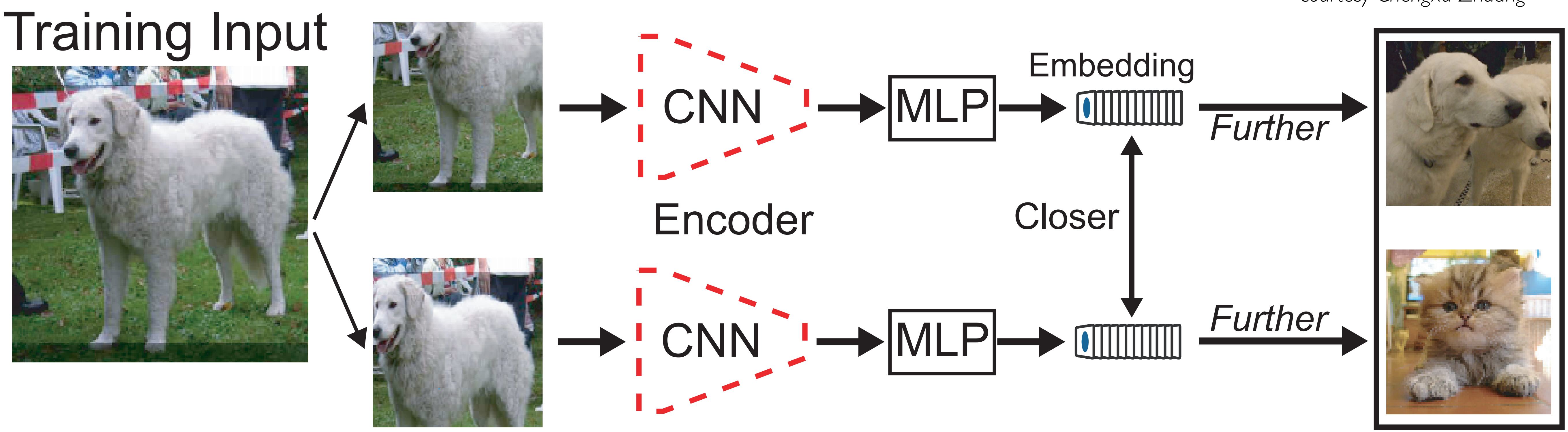


Distilling Constraints: Behavioral Goals

Self-supervised Losses

Contrastive Objectives

courtesy Chengxu Zhuang



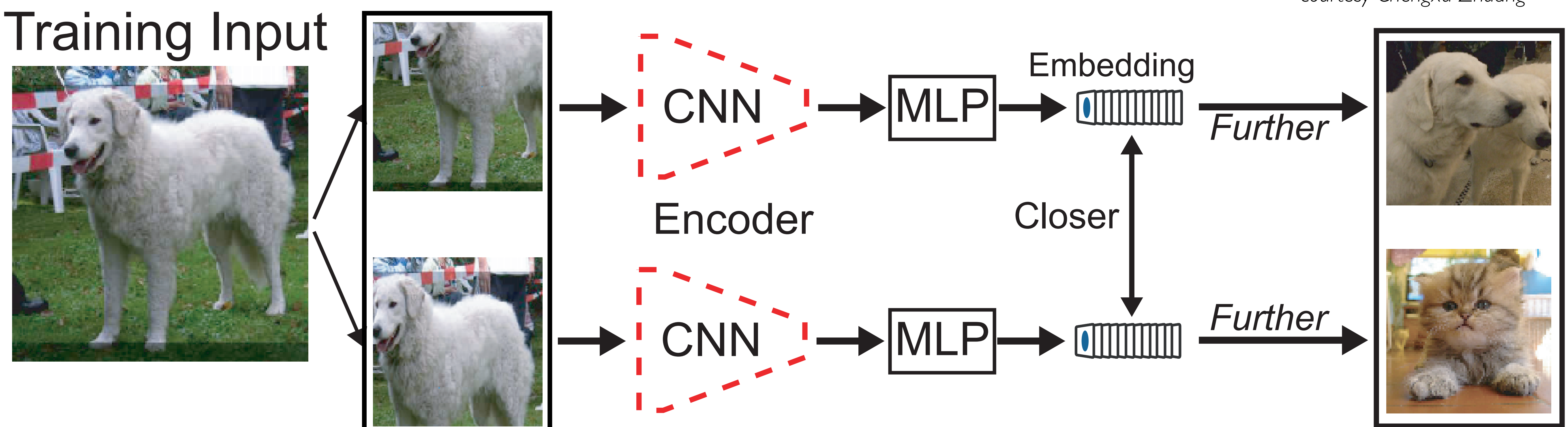
CNN: Convolutional Neural Network, MLP: Multi-Layer Perceptron

Distilling Constraints: Behavioral Goals

Self-supervised Losses

<u>Contrastive Objectives</u>

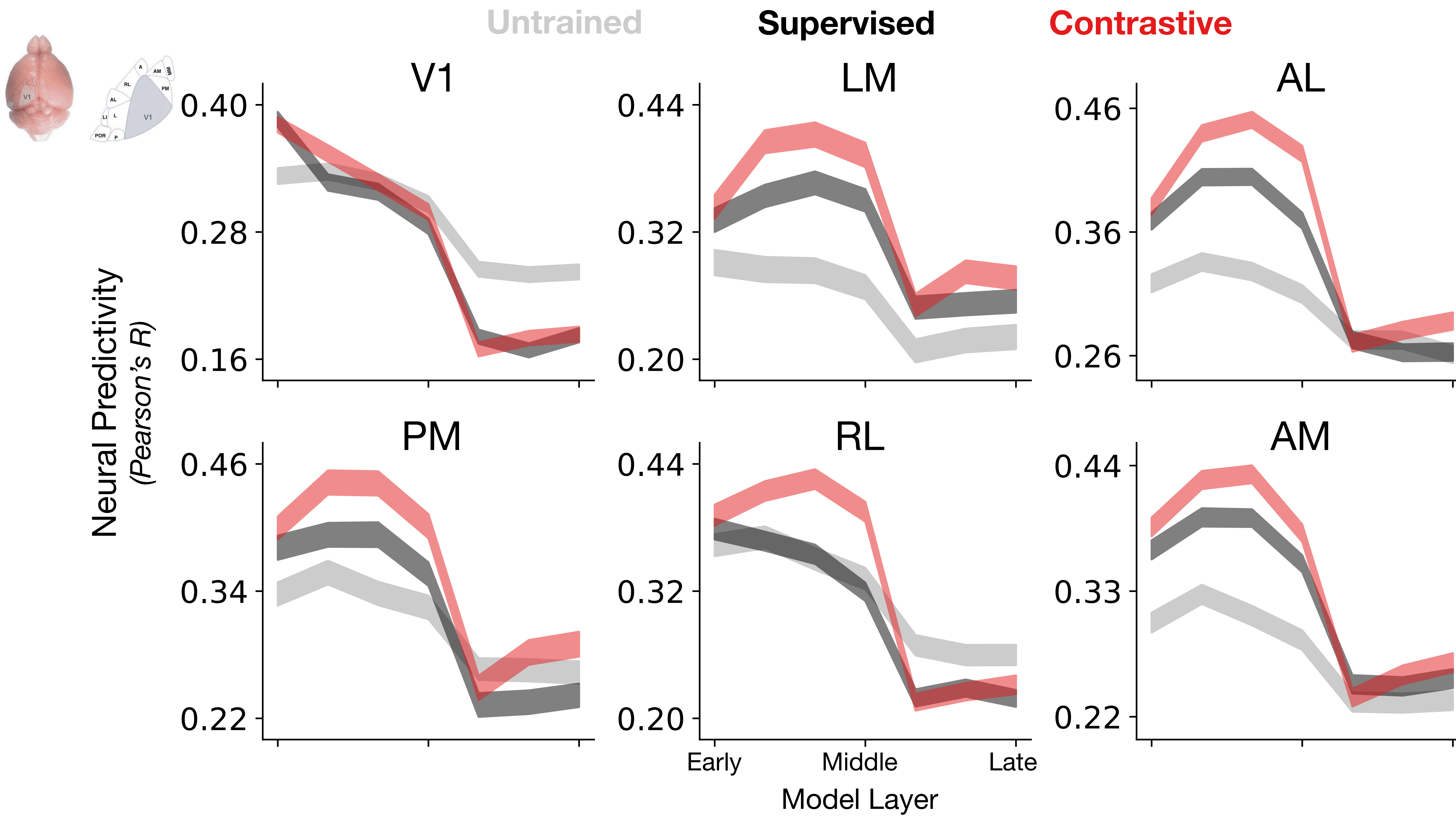
courtesy Chengxu Zhuang



CNN: Convolutional Neural Network, MLP: Multi-Layer Perceptron

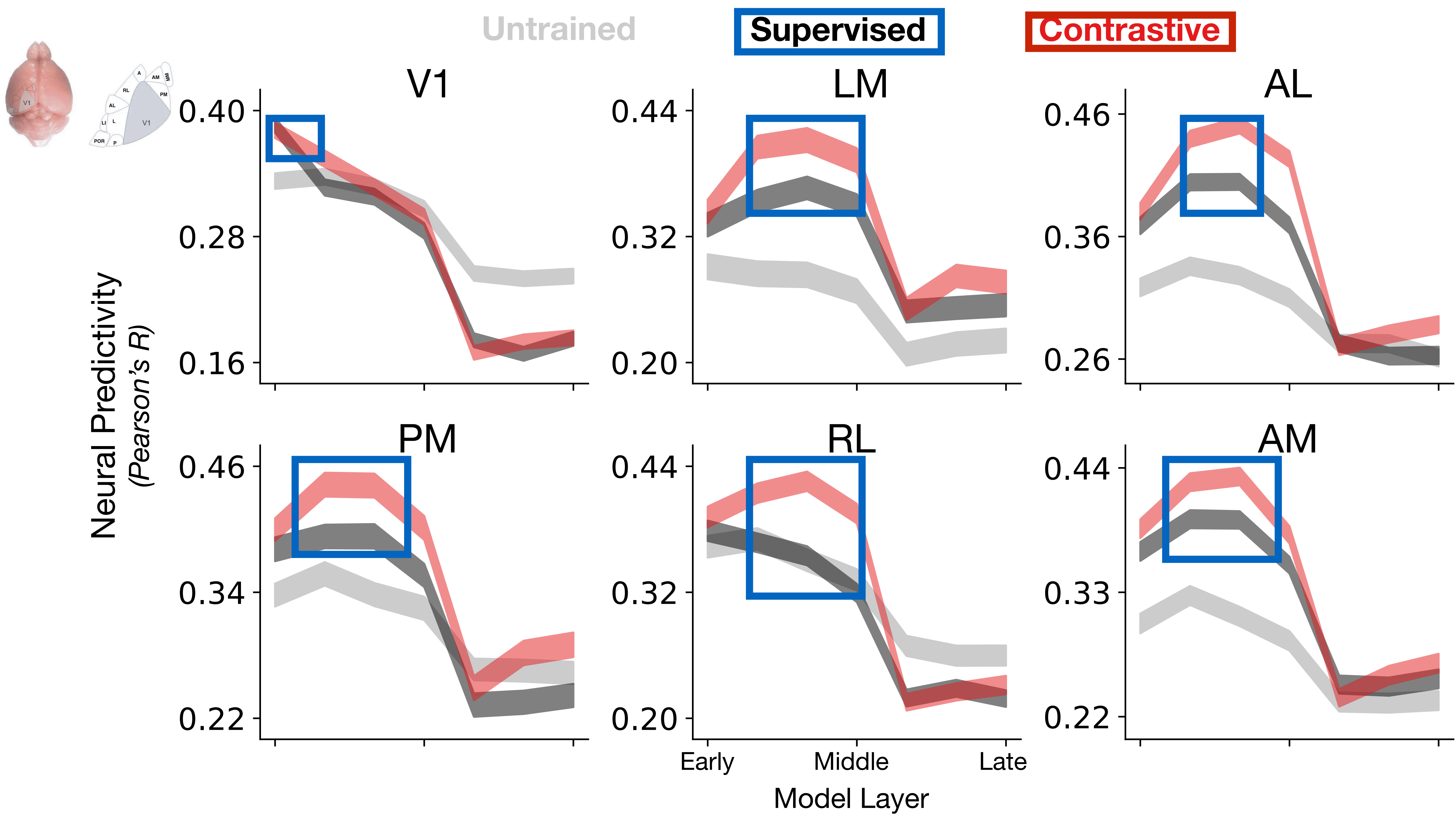
Make the embeddings of <u>different</u> views of the *same* image to be similar, while pushing them apart from different images

Self-supervised > Supervised Losses



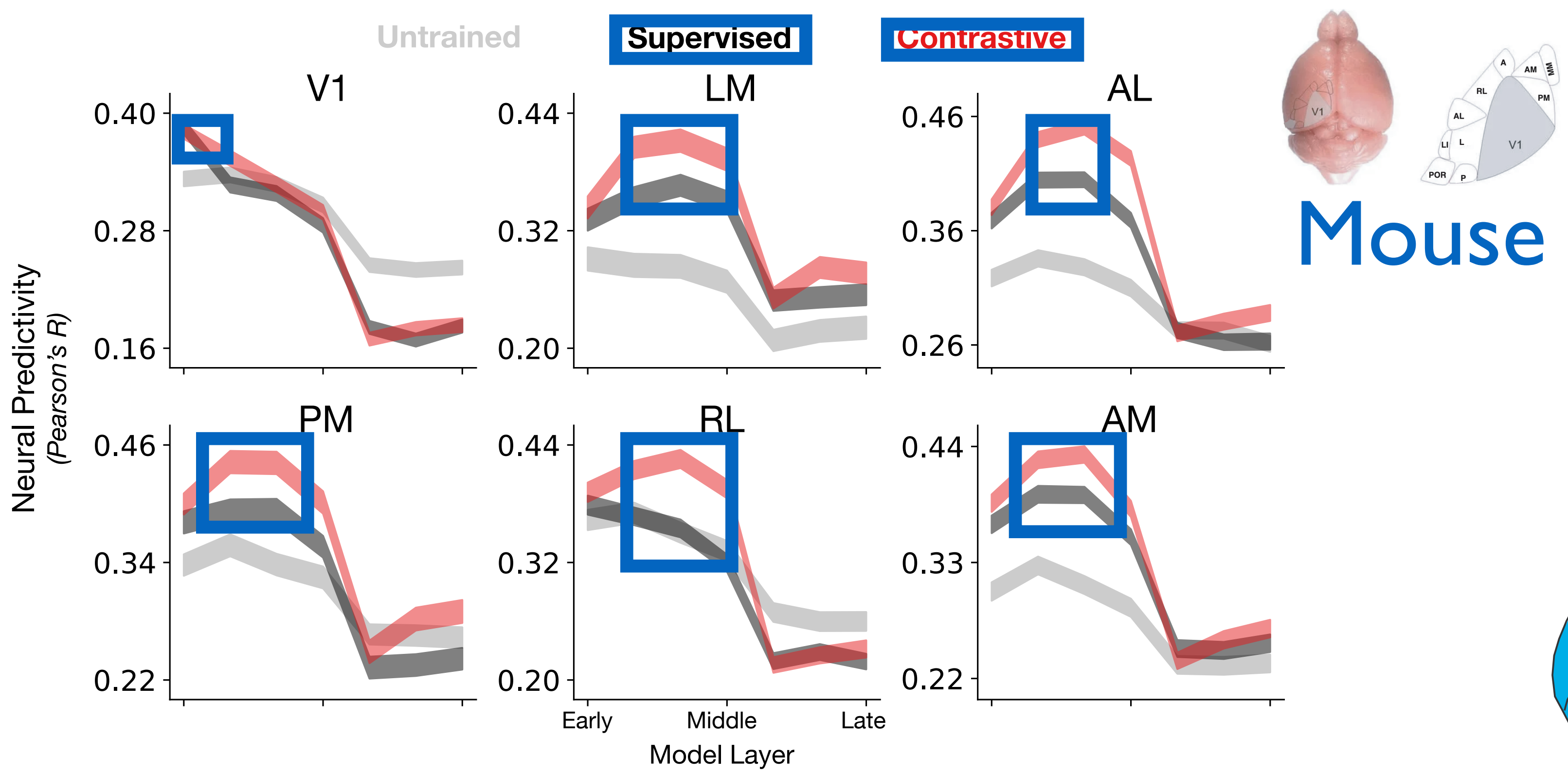
Self-supervised > Supervised Losses

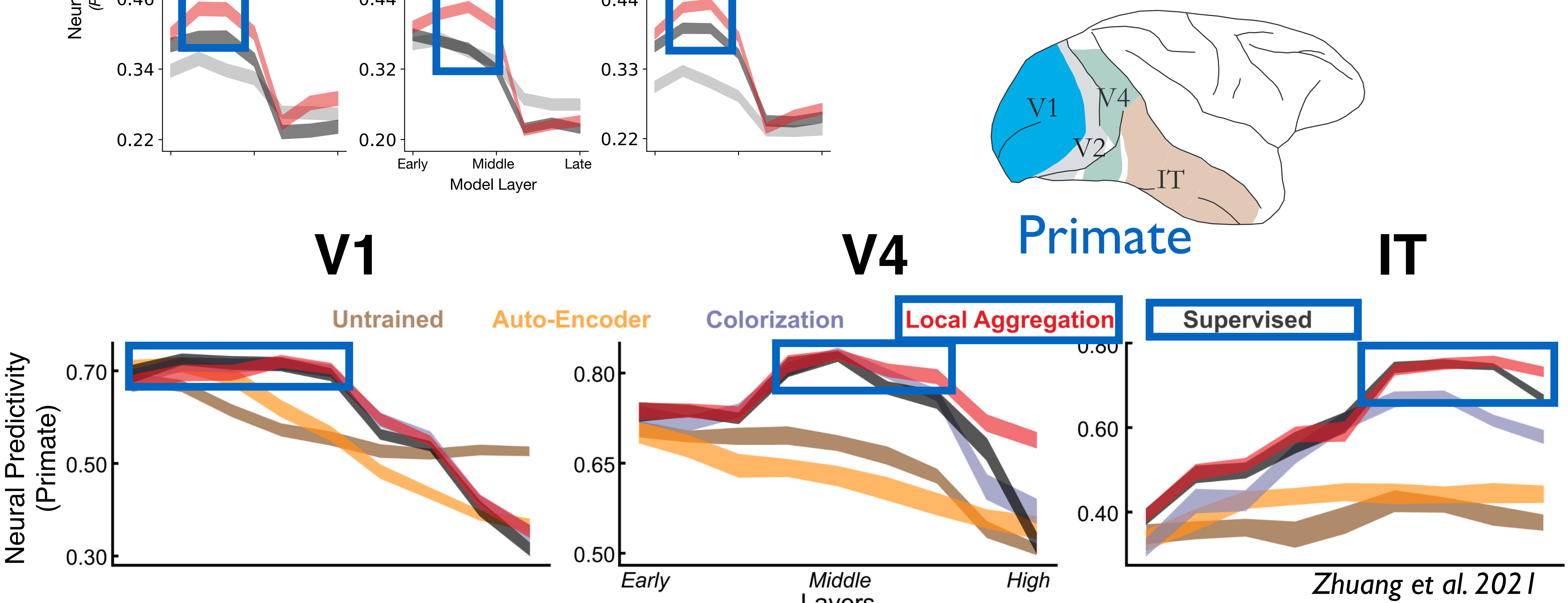
Using a self-supervised, contrastive objective function improves neural predictivity (*best* architecture & data stream fixed)



Self-supervised > Supervised Losses

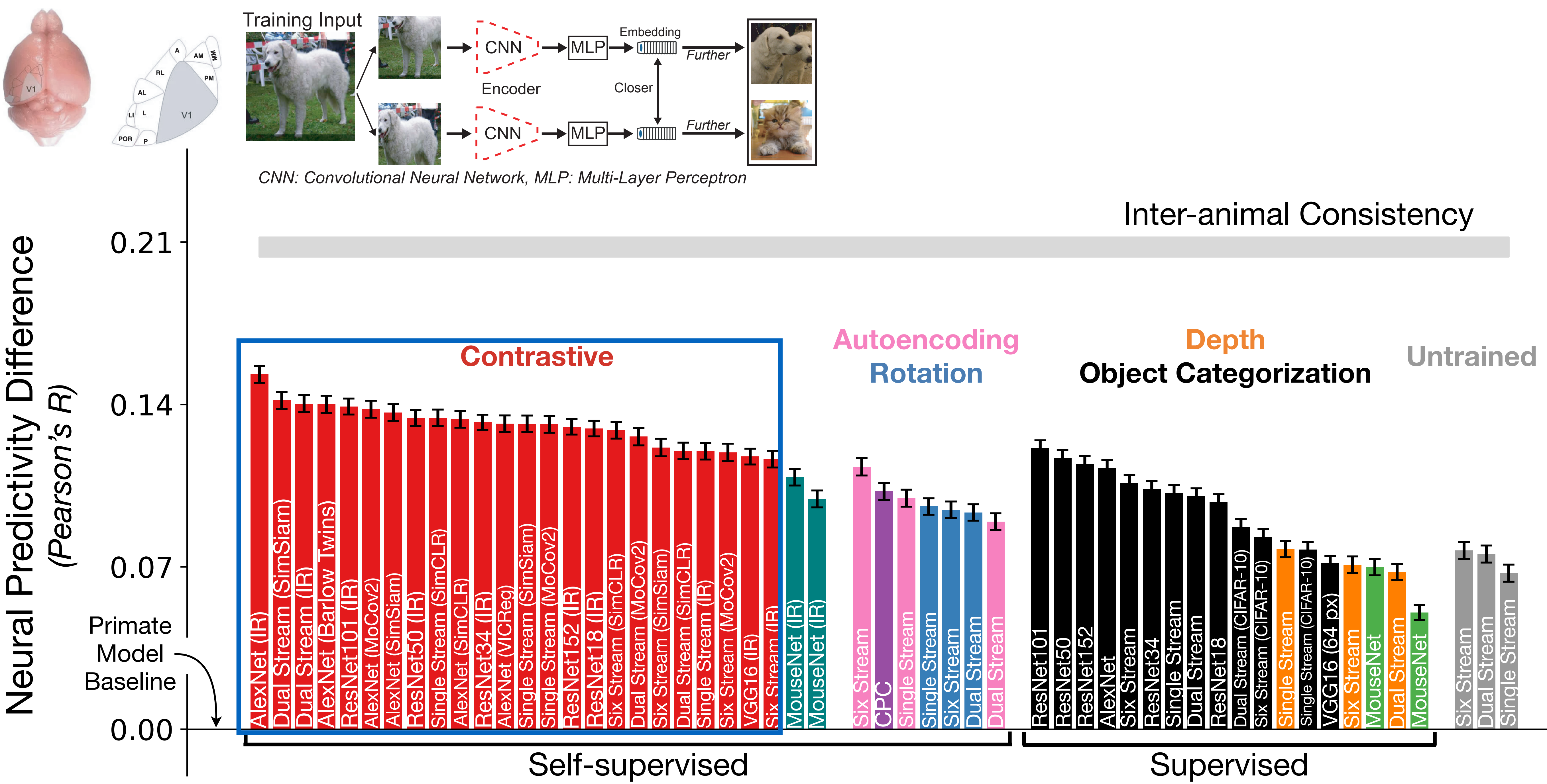
Unlike in primates where contrastive matches supervised



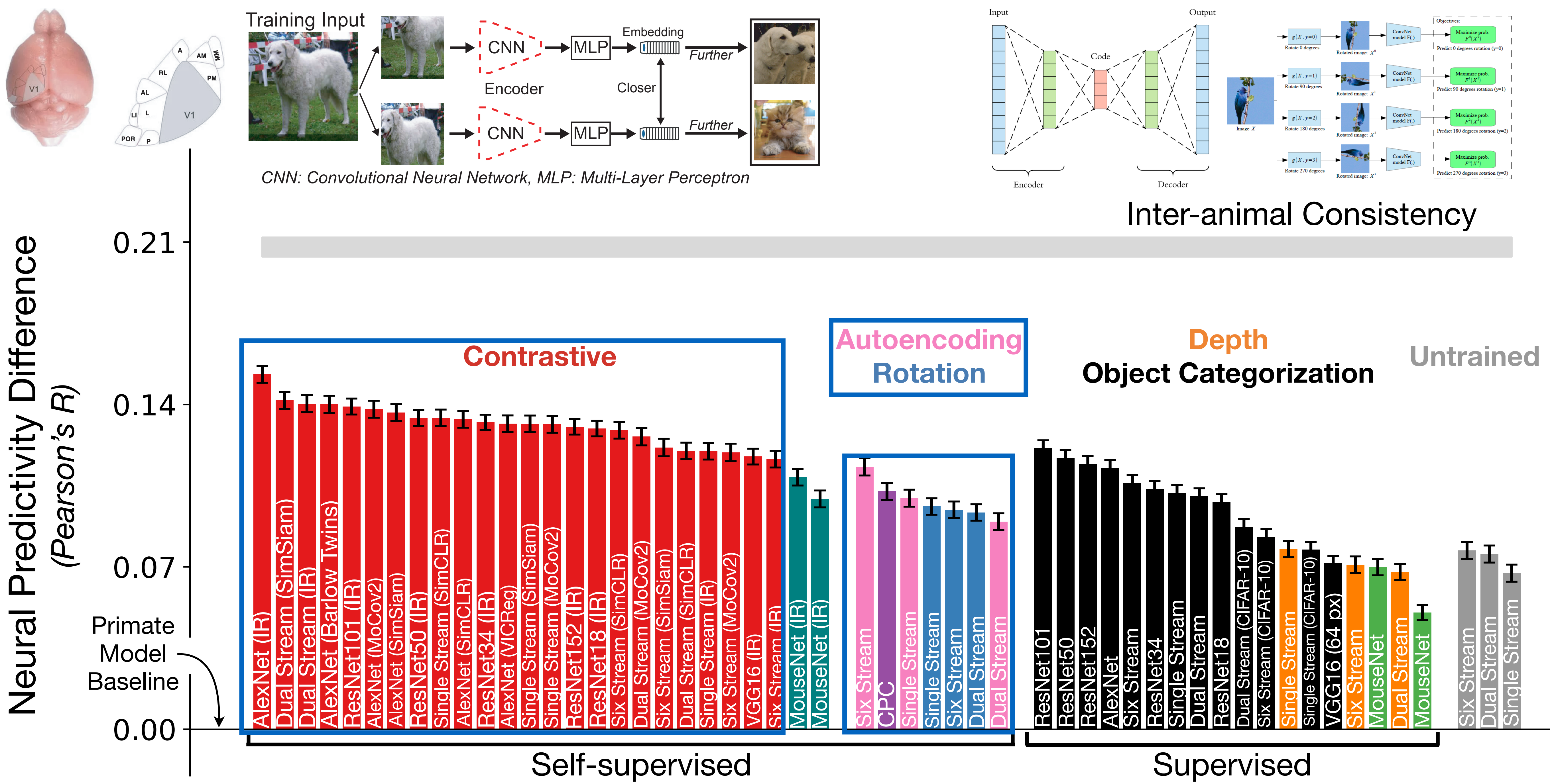


Zhuang et al. 2021

Contrastive losses are overall the best *self-supervised* loss

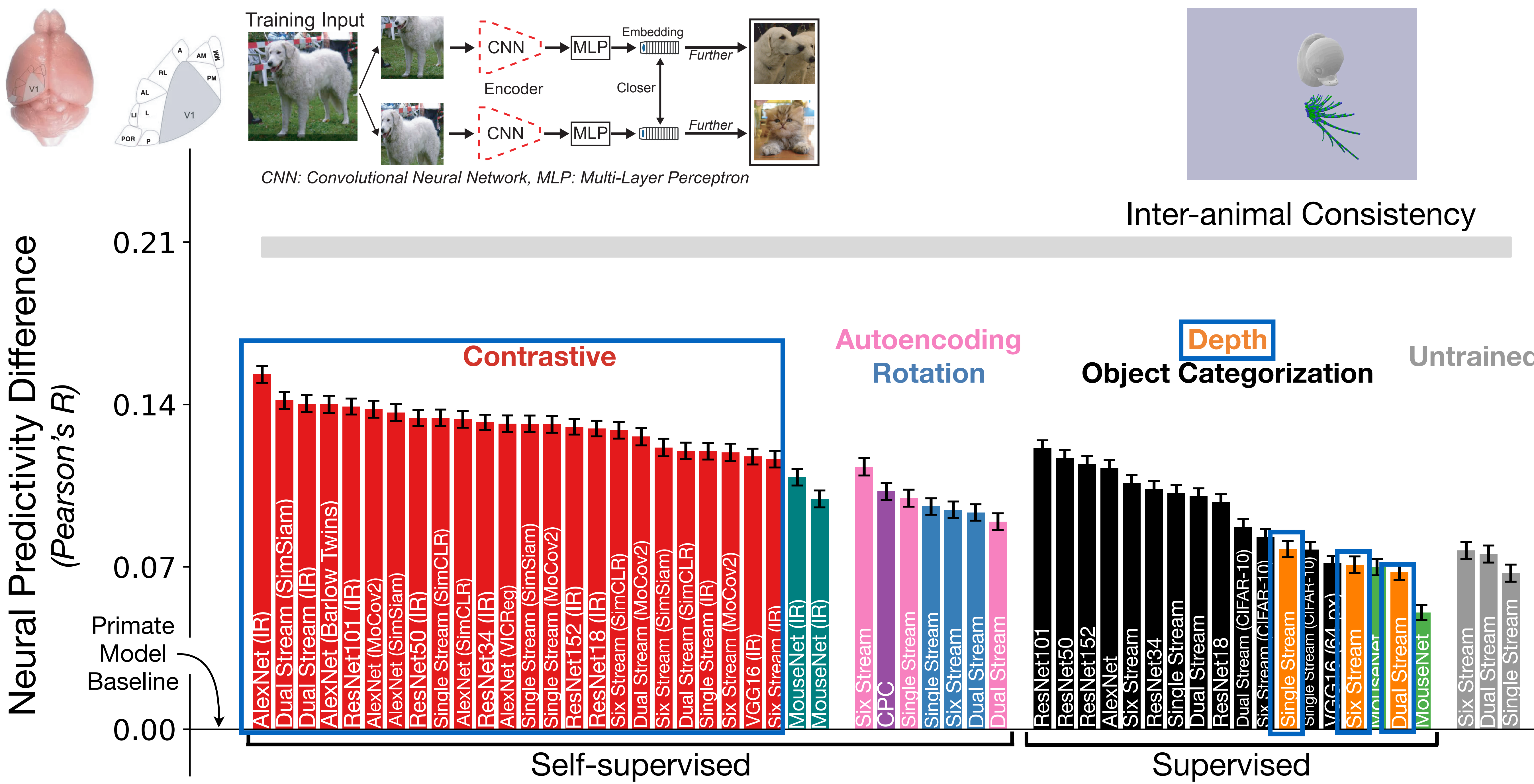


Contrastive losses are overall the best *self-supervised* loss



Contrastive objectives are overall the best compared to other *self-supervised* alternatives

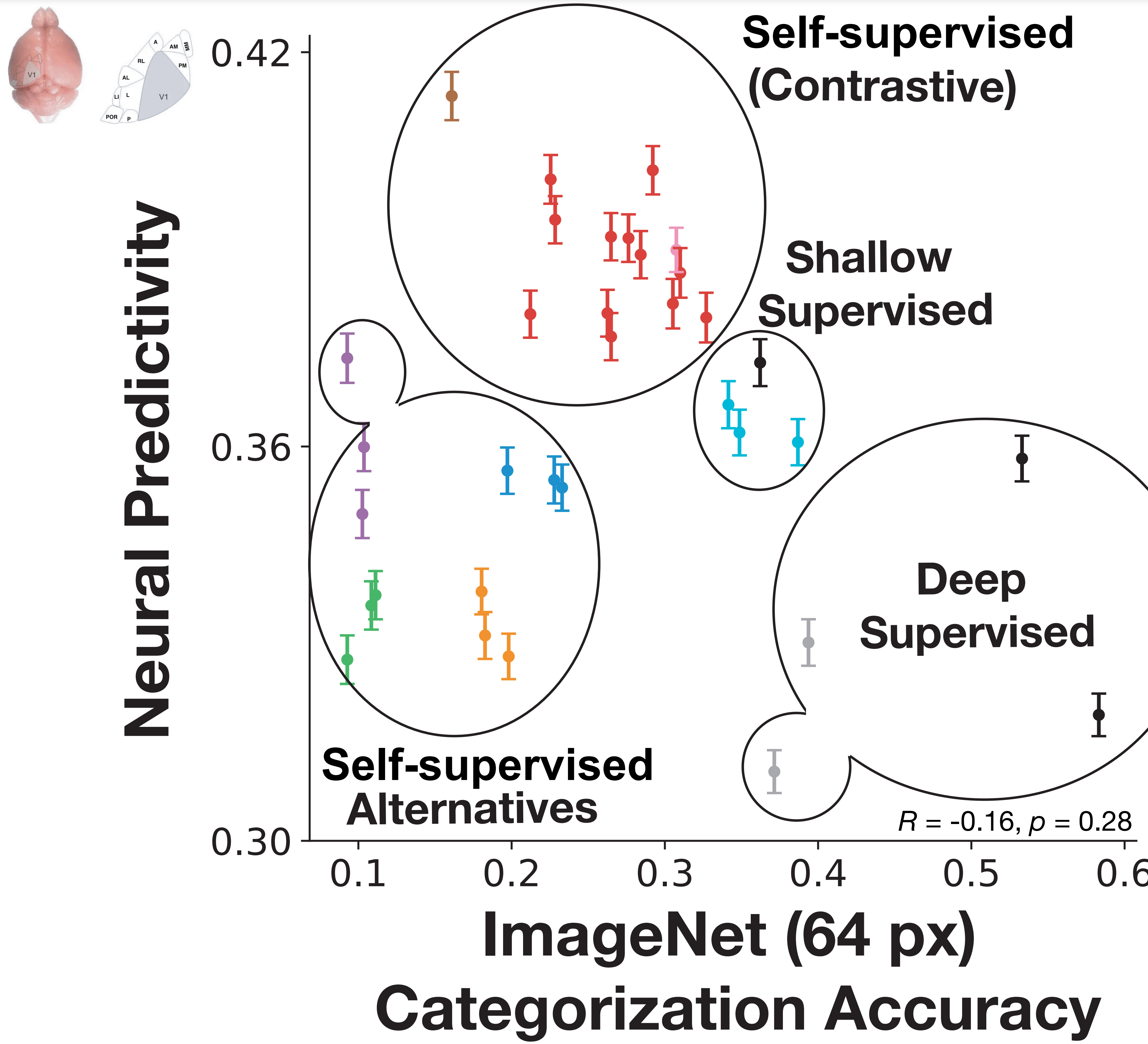
Contrastive losses outperform purely behaviorally-driven loss



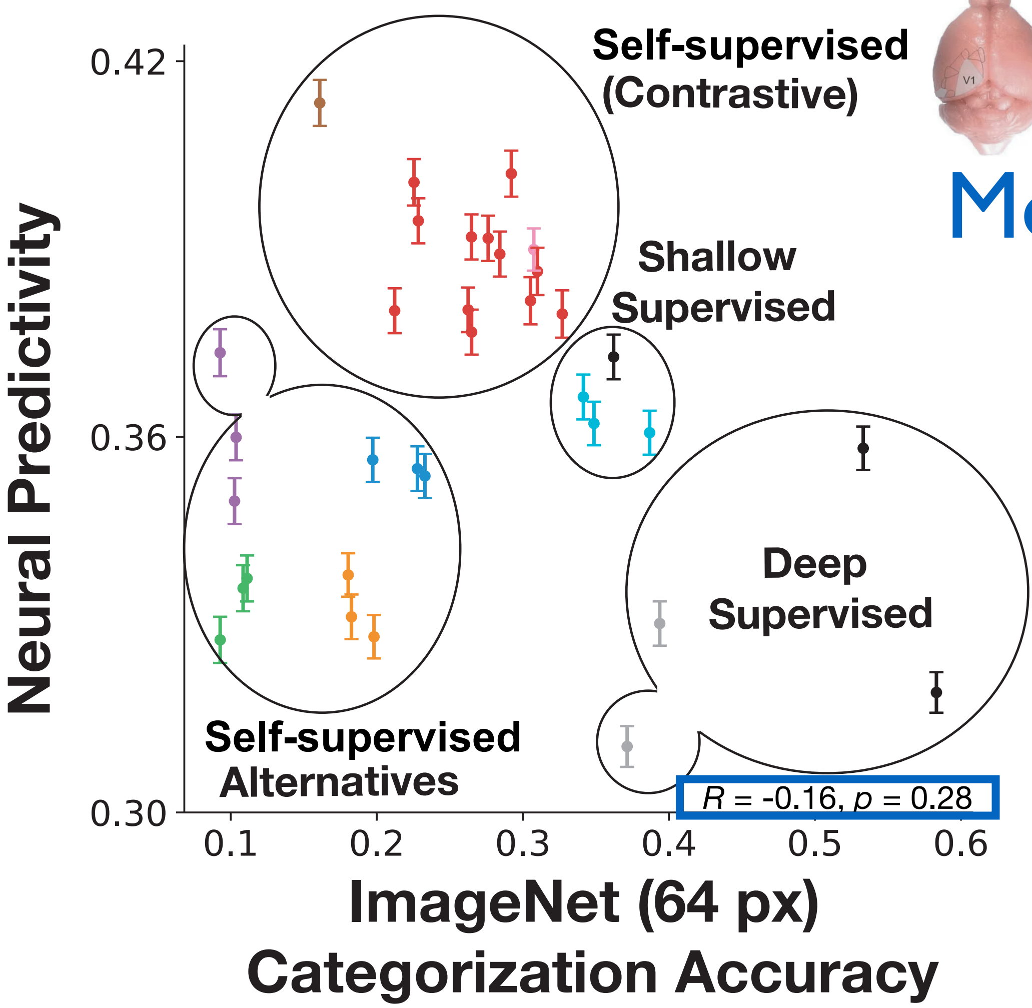
Contrastive objectives outperform "whisking" visual proxy loss

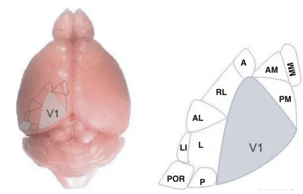
What is the ecological reason to prefer self-supervision?

ImageNet categorization performance **NOT** correlated with neural predictivity

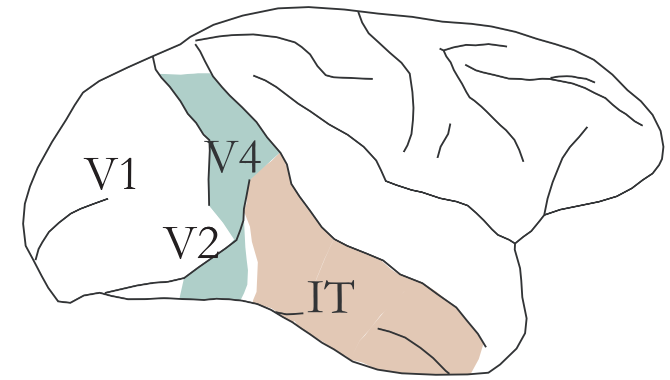


ImageNet categorization performance **NOT** correlated with neural predictivity



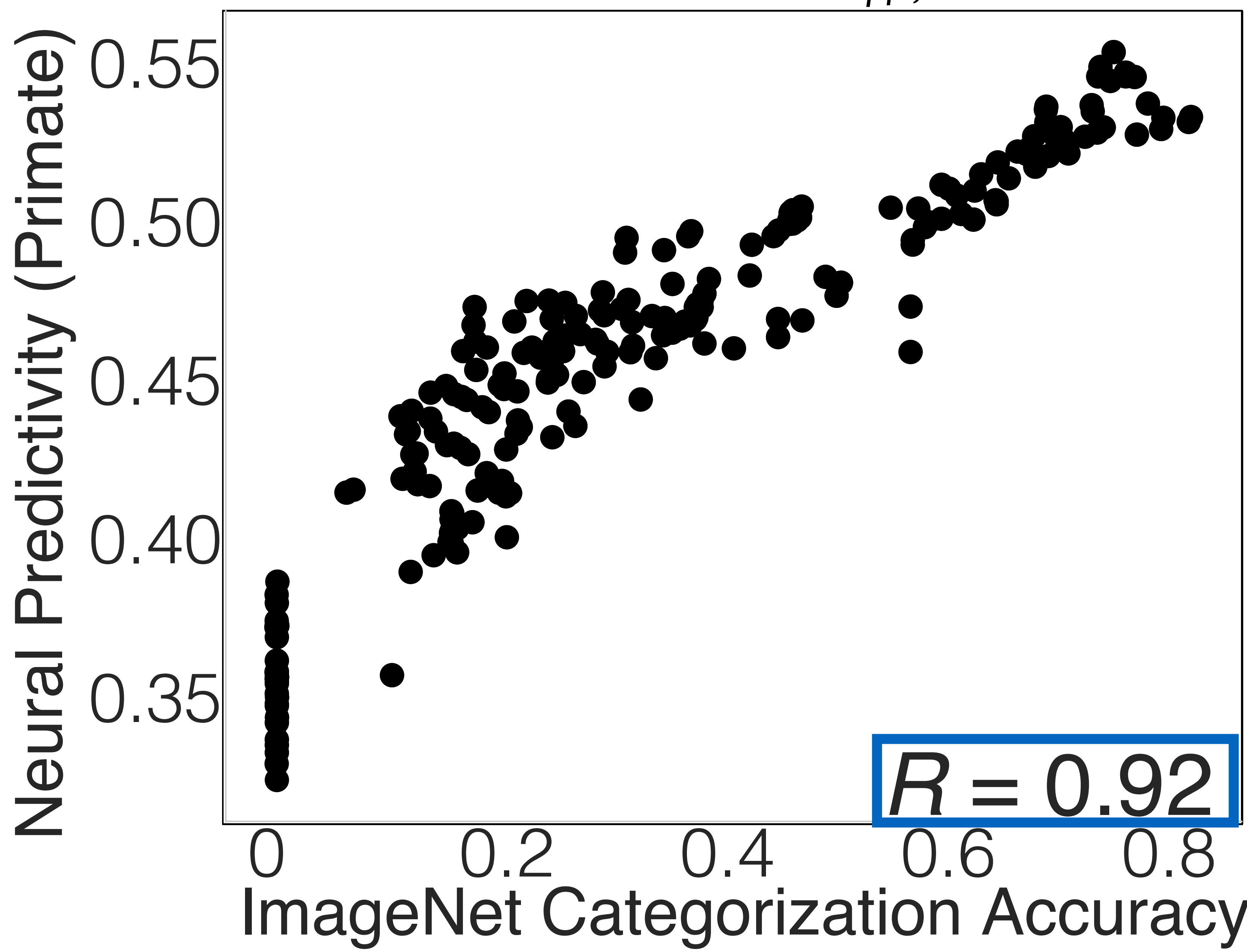


Mouse



Unlike in primates!

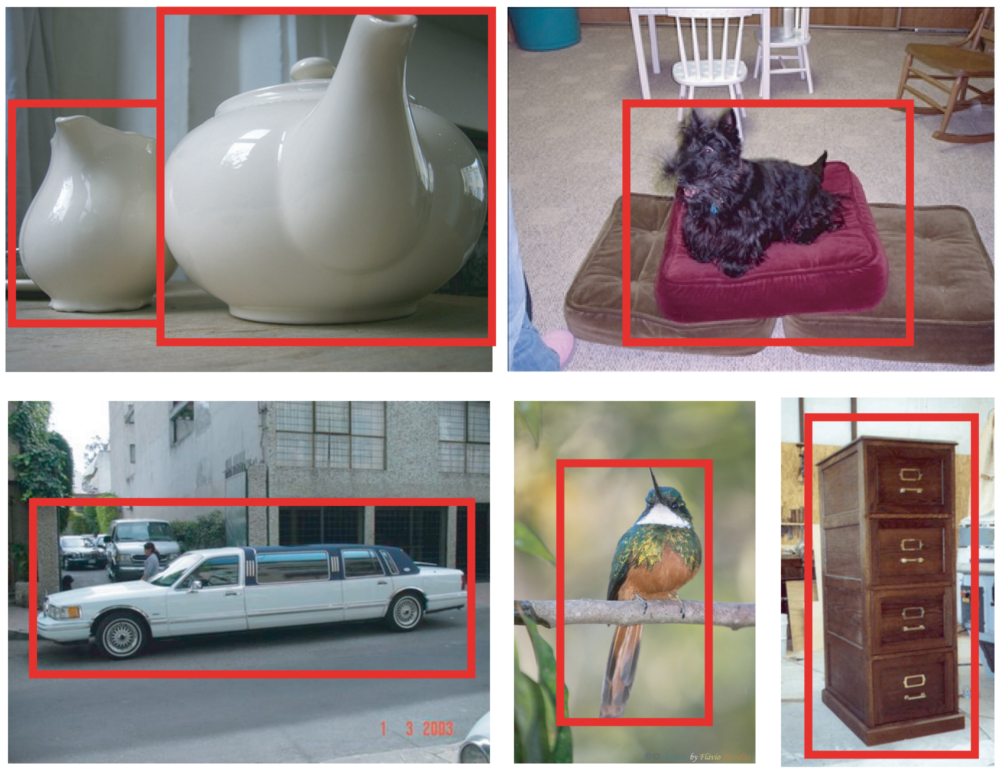
Schrimpf, Kubilius* et al. 2018*



Assessing Task-Generality

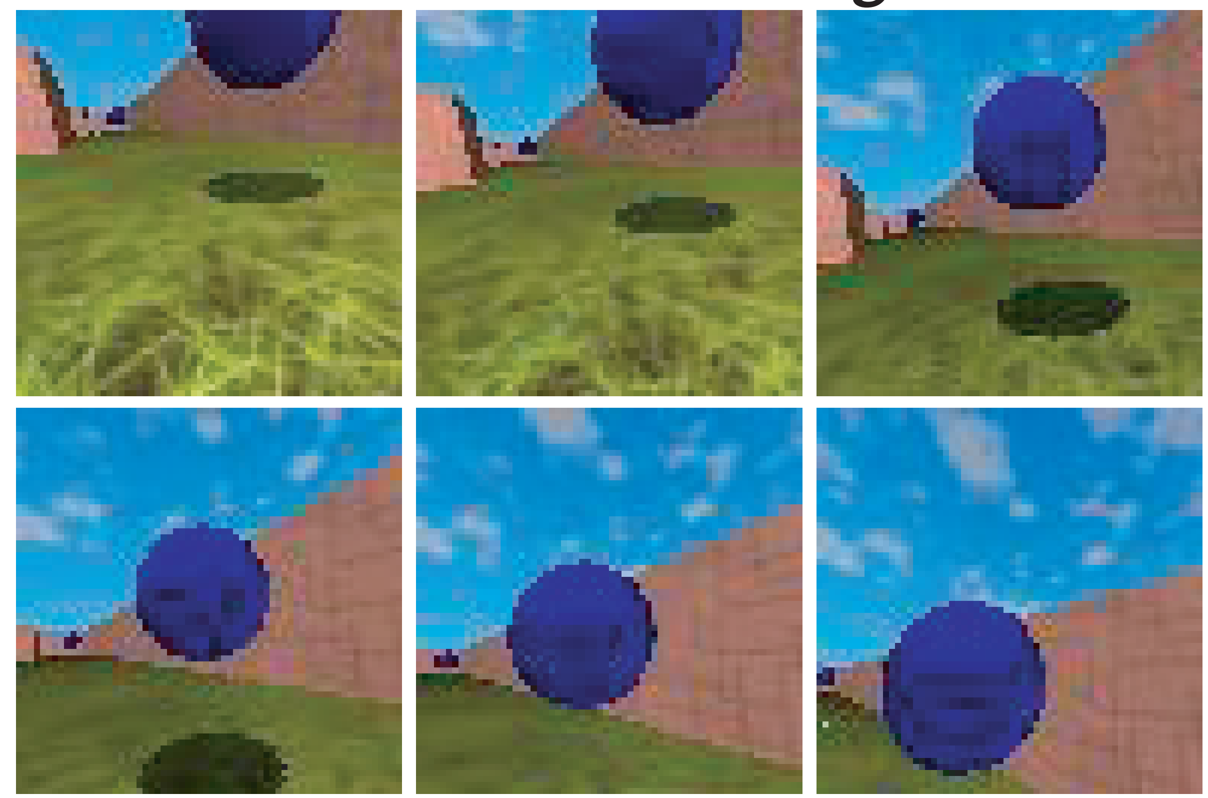
Train

ImageNet

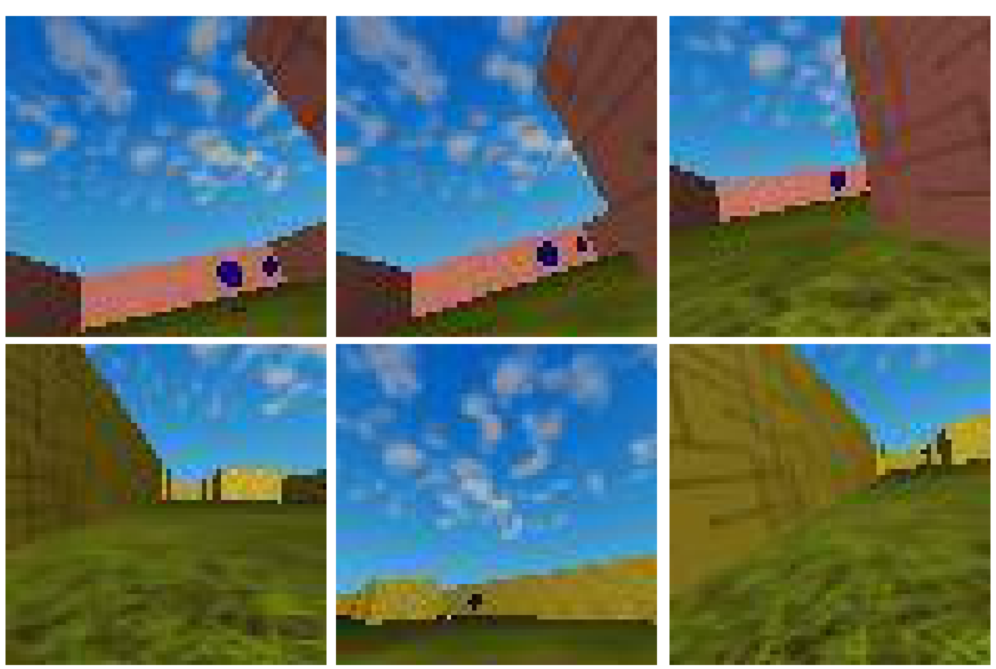


Evaluate

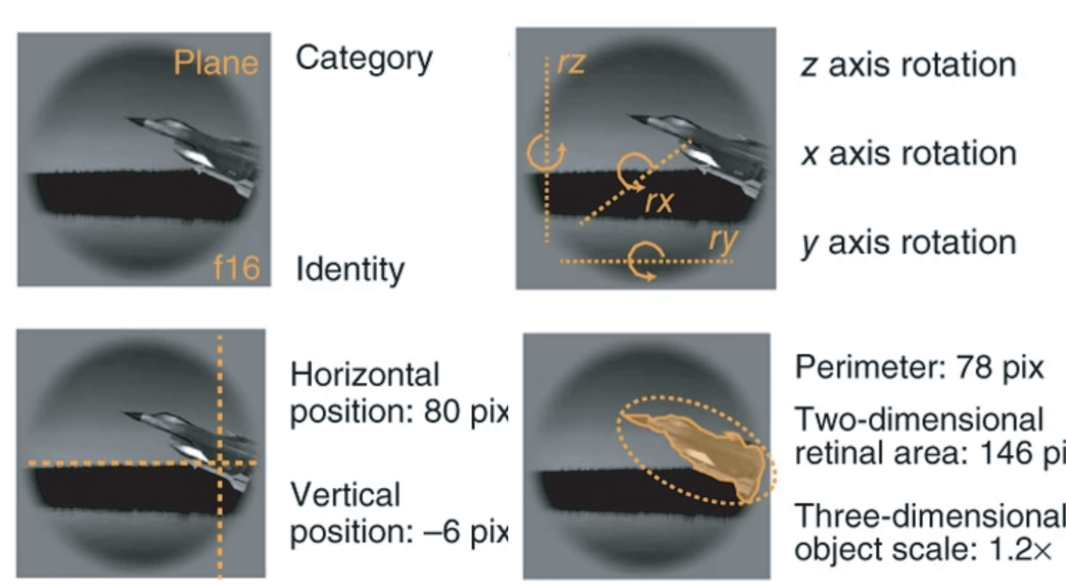
Reward-Based Navigation



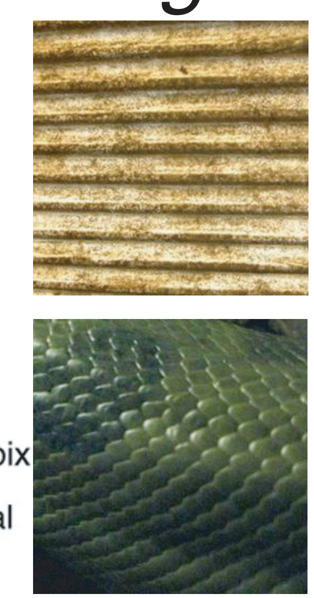
Maze Environment



Visual Scene Understanding

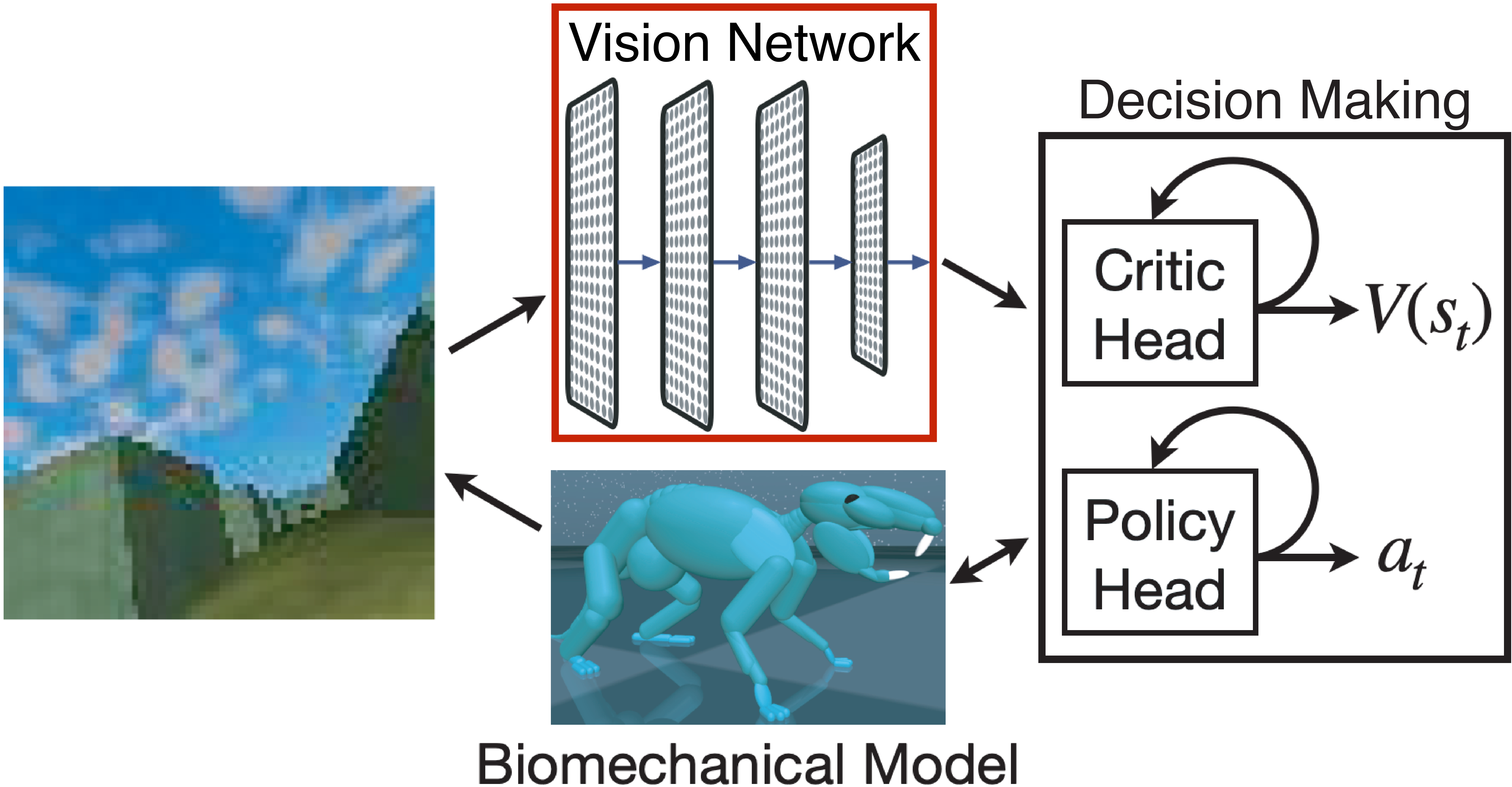


Object properties



Texture

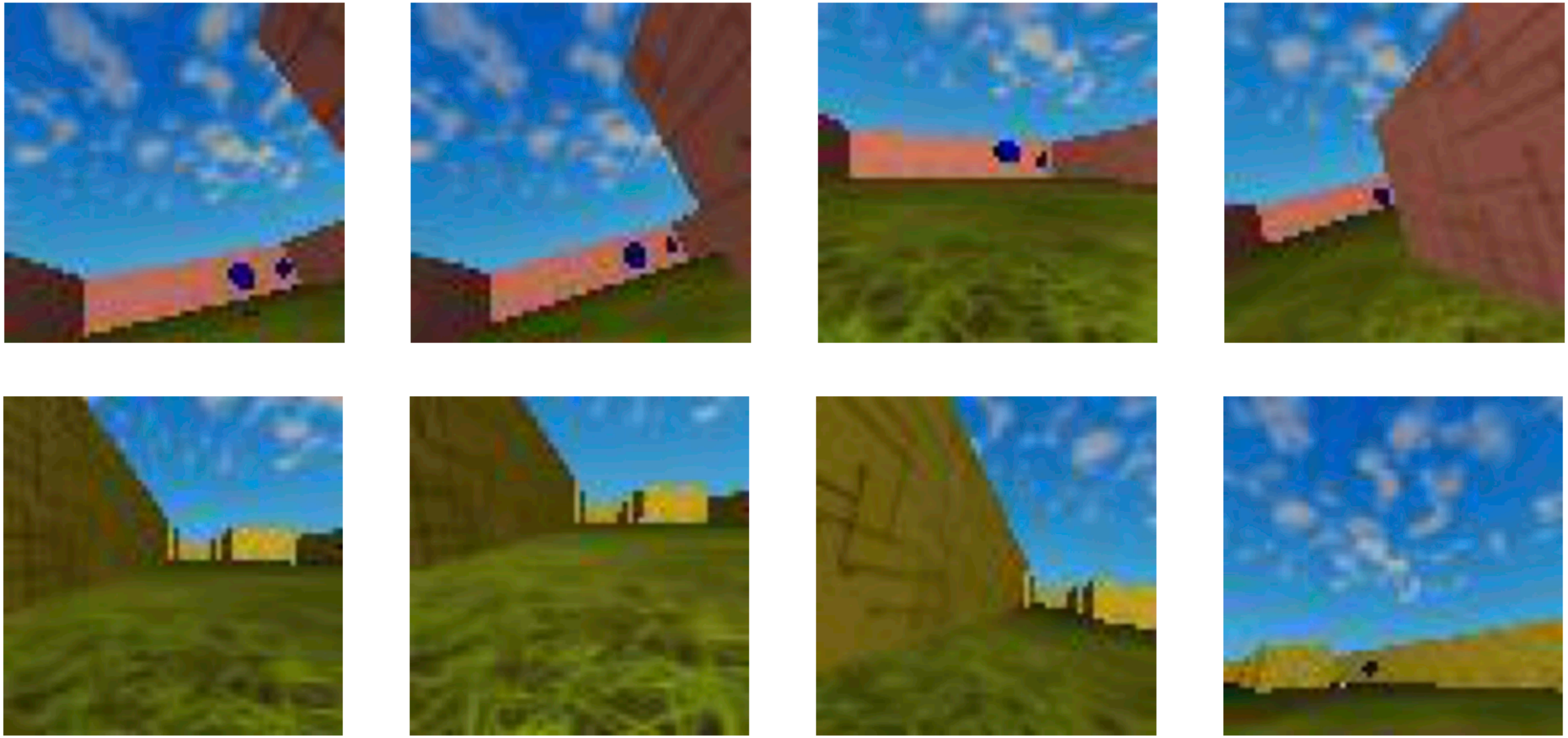
Schematic of Virtual Rodent



(*Merel**, *Aldorando**, *Marshal** *et al. 2020*)

Navigation task in "Rodent Mazes" environment

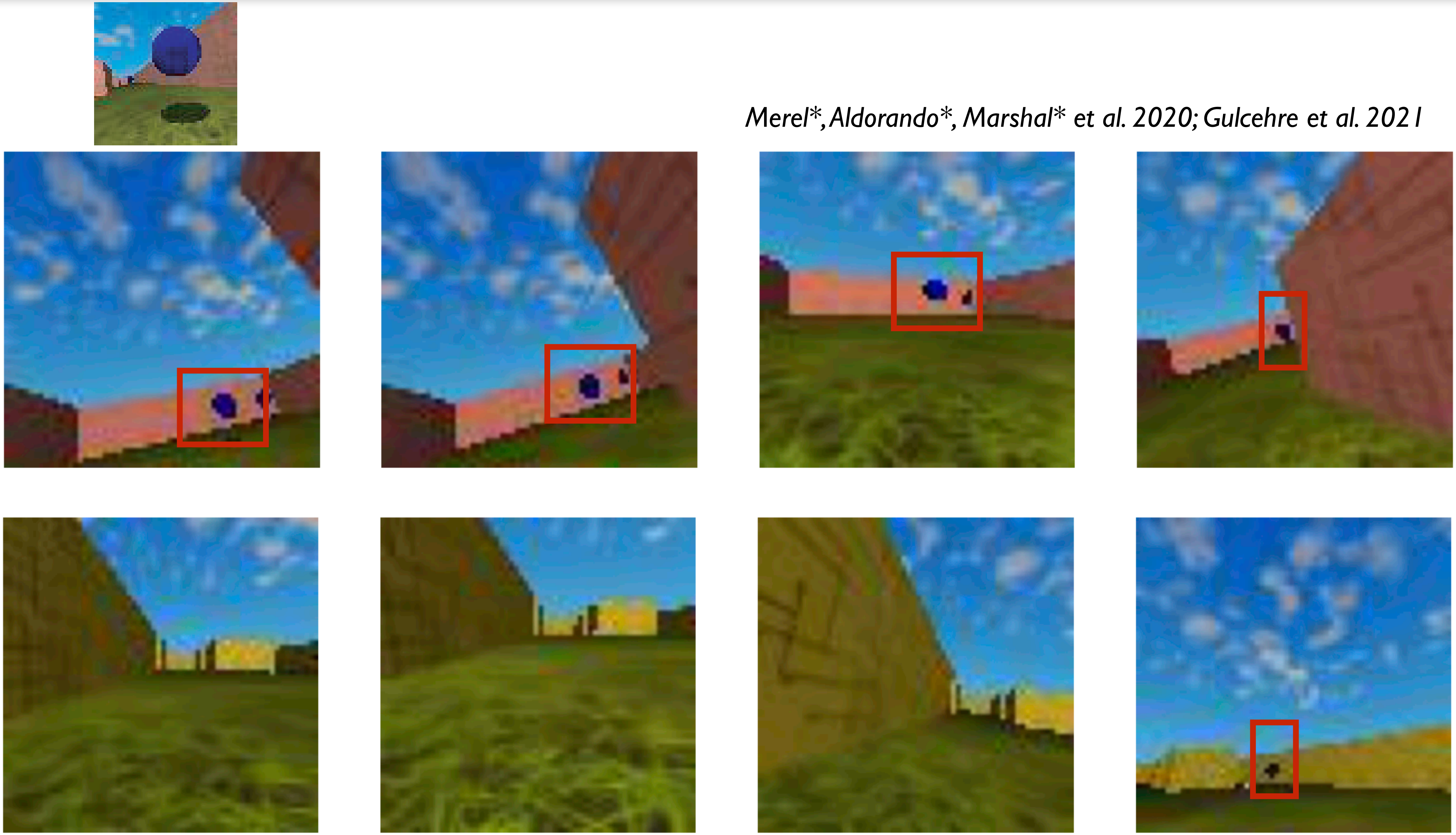
Merel, Aldorando*, Marshal* et al. 2020; Gulcehre et al. 2021*



Requires keeping track of history over long timescales with high-dimensional, continuous inputs

Navigation task in "Rodent Mazes" environment

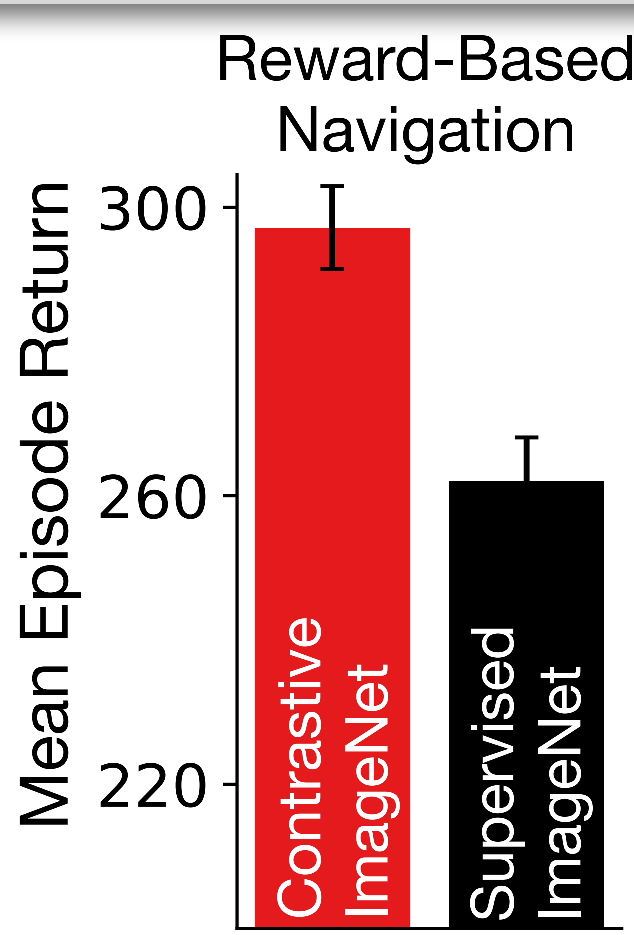
Merel, Aldorando*, Marshal* et al. 2020; Gulcehre et al. 2021*

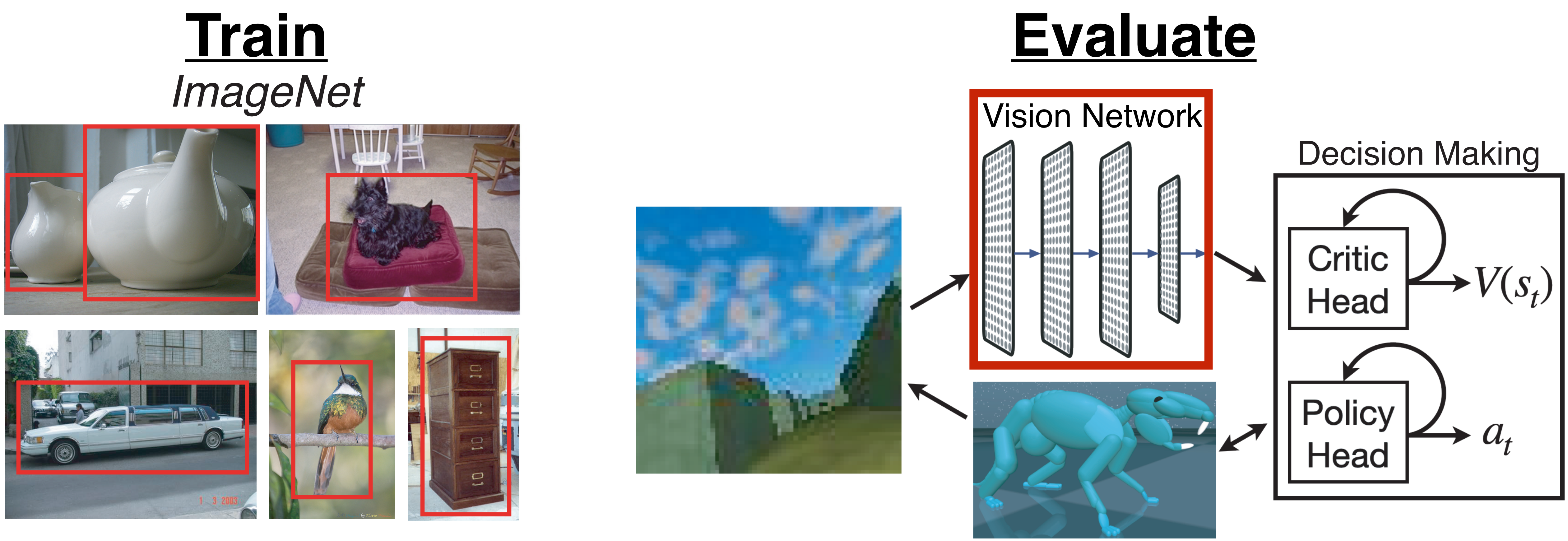


Requires keeping track of history over long timescales with high-dimensional, continuous inputs

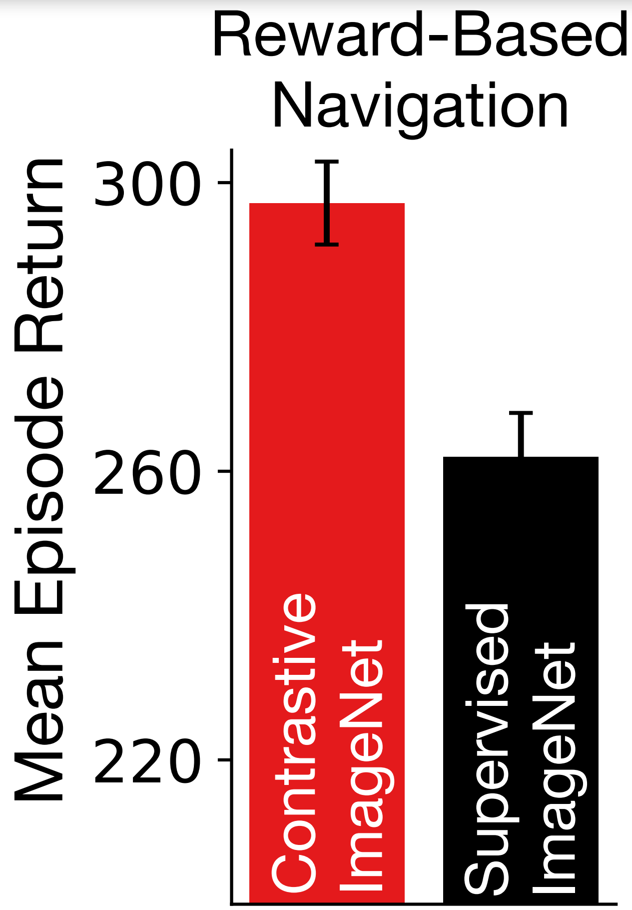
Contrastive models yield better transfer performance

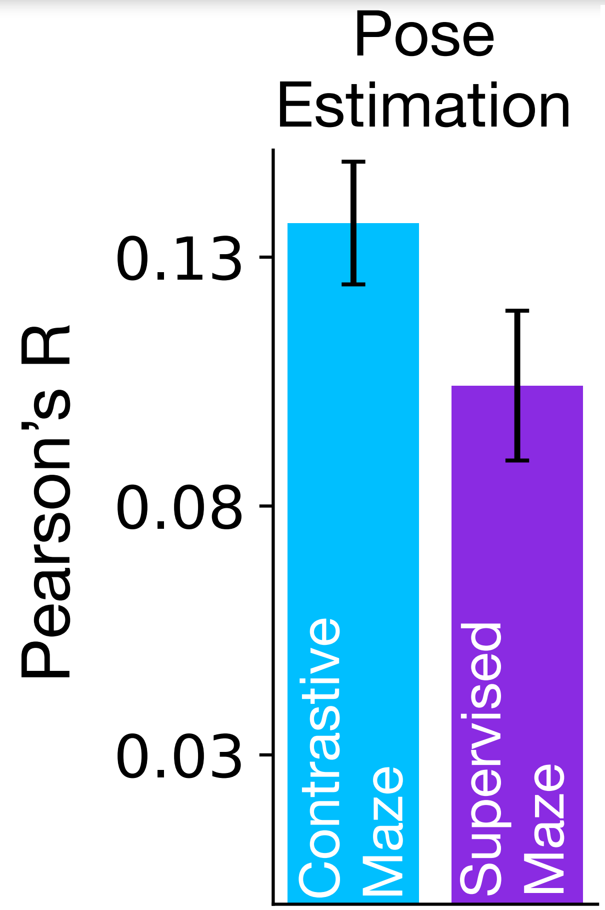
Contrastive models yield better transfer performance





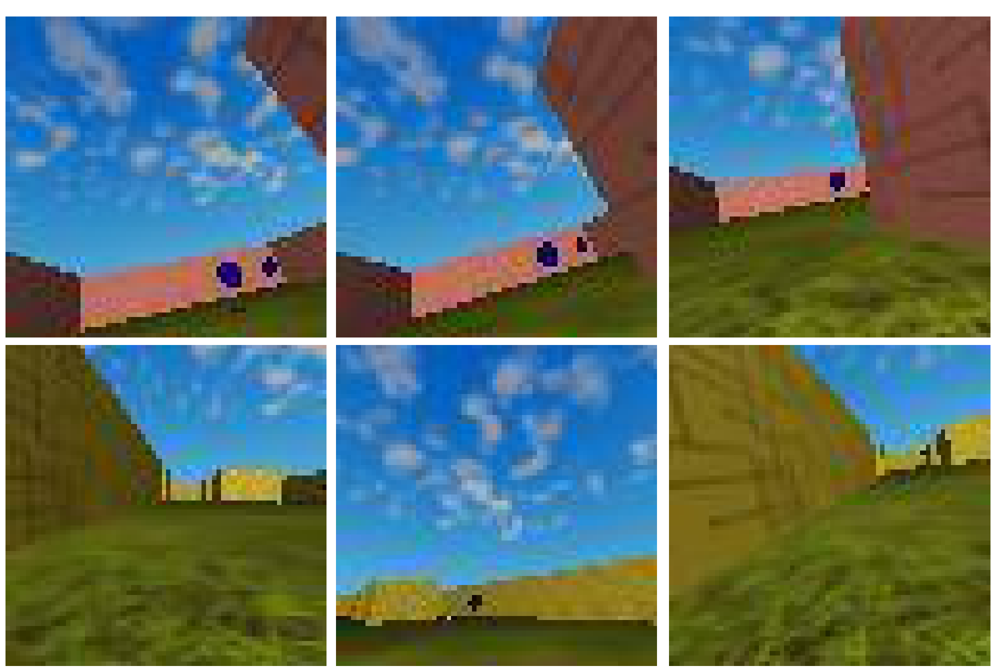
Contrastive models yield better transfer performance





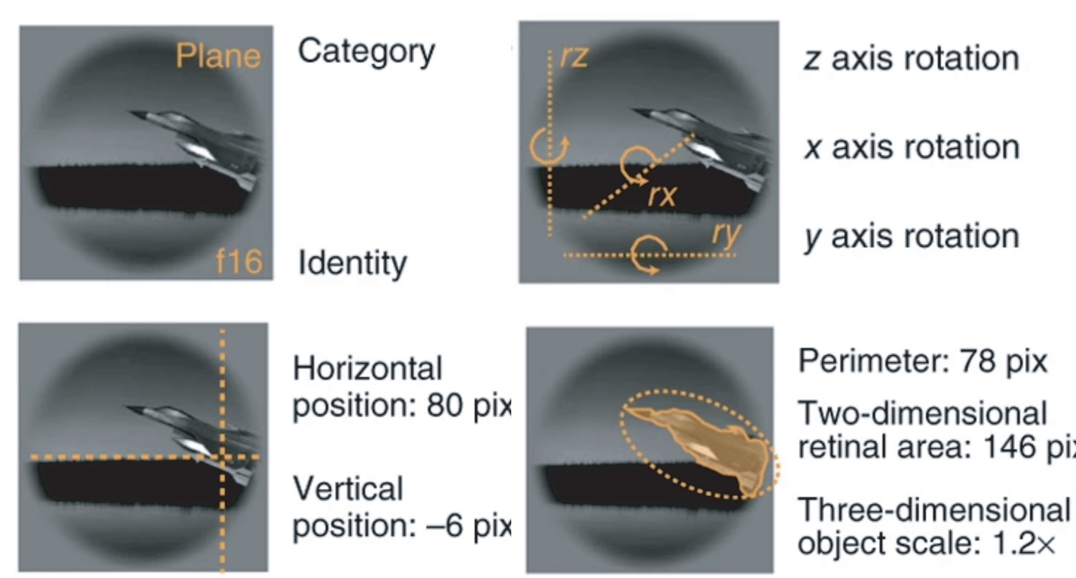
Train

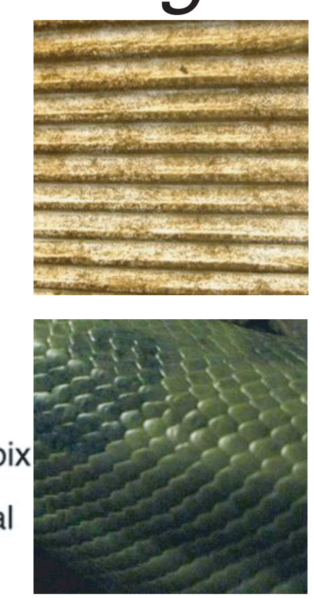
Maze Environment



Evaluate

Visual Scene Understanding

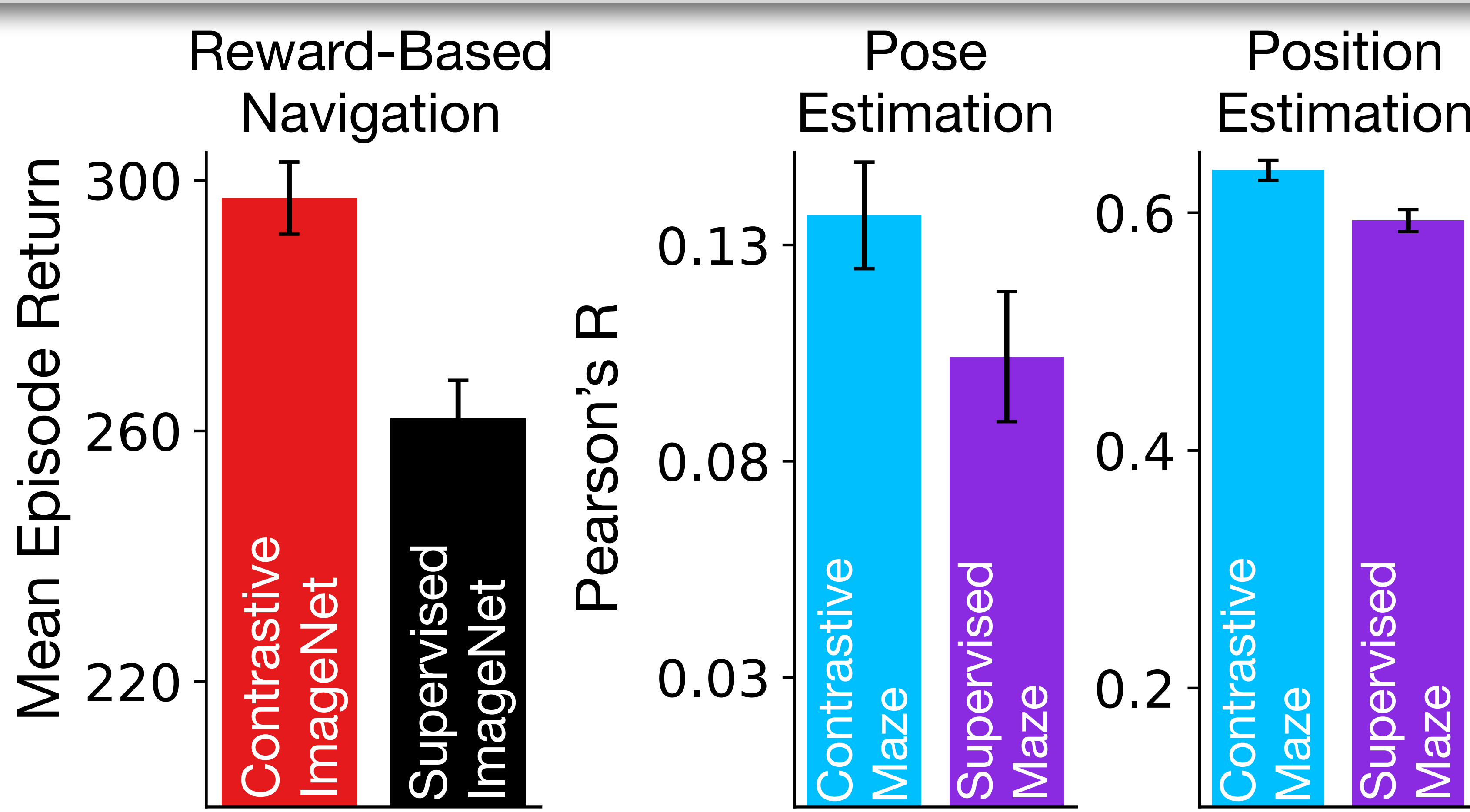




Object properties

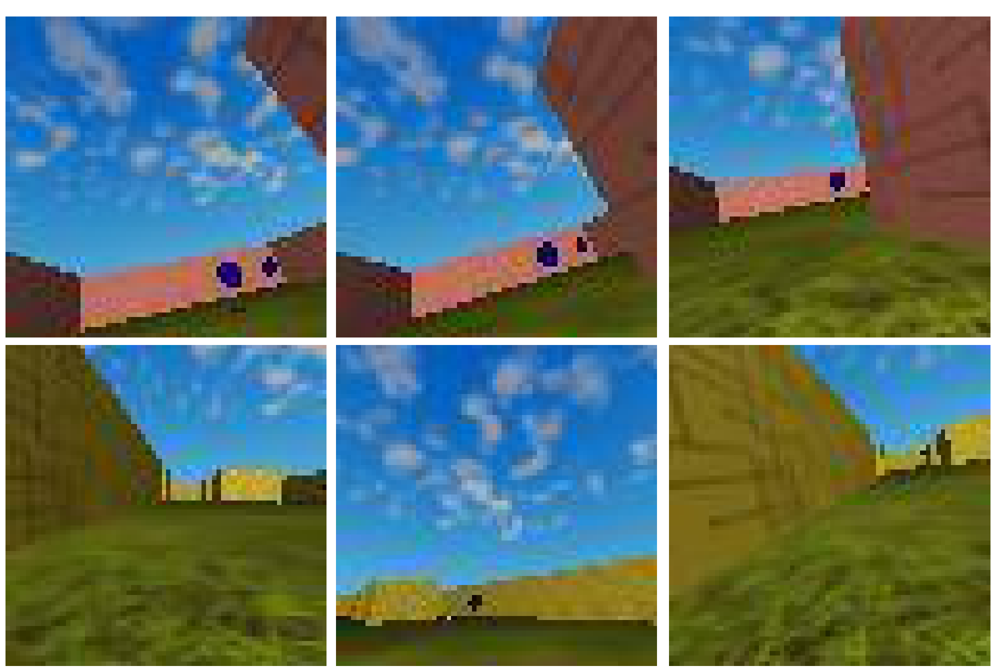
Texture

Contrastive models yield better transfer performance



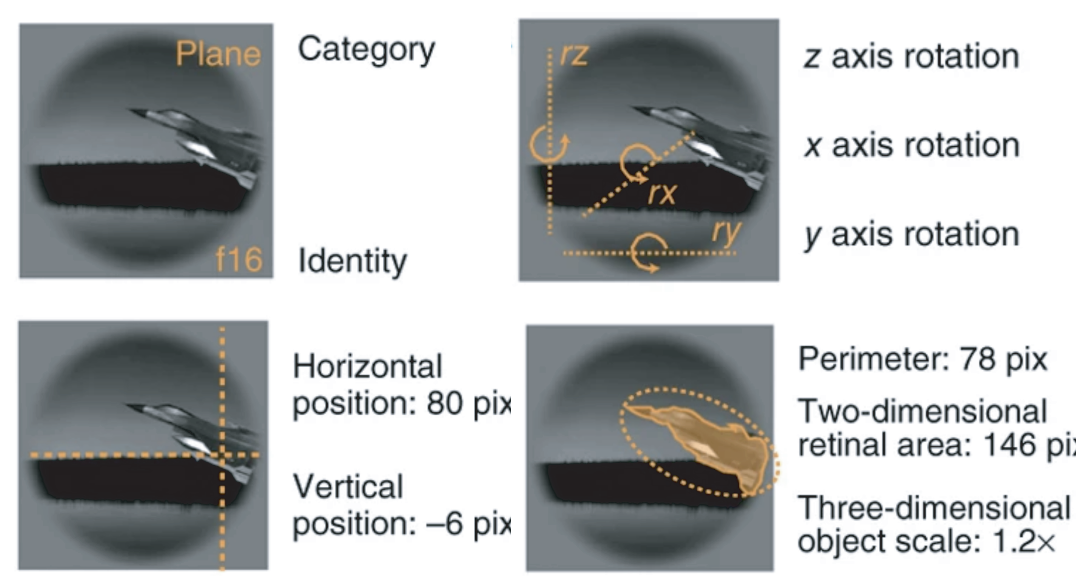
Train

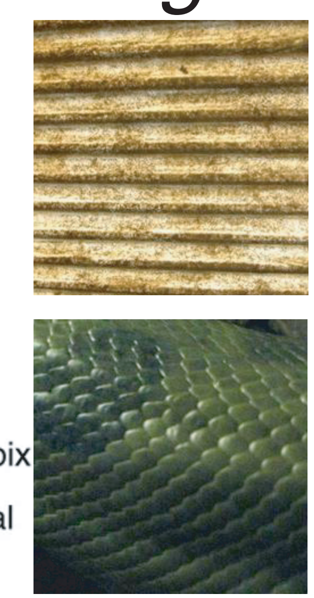
Maze Environment



Evaluate

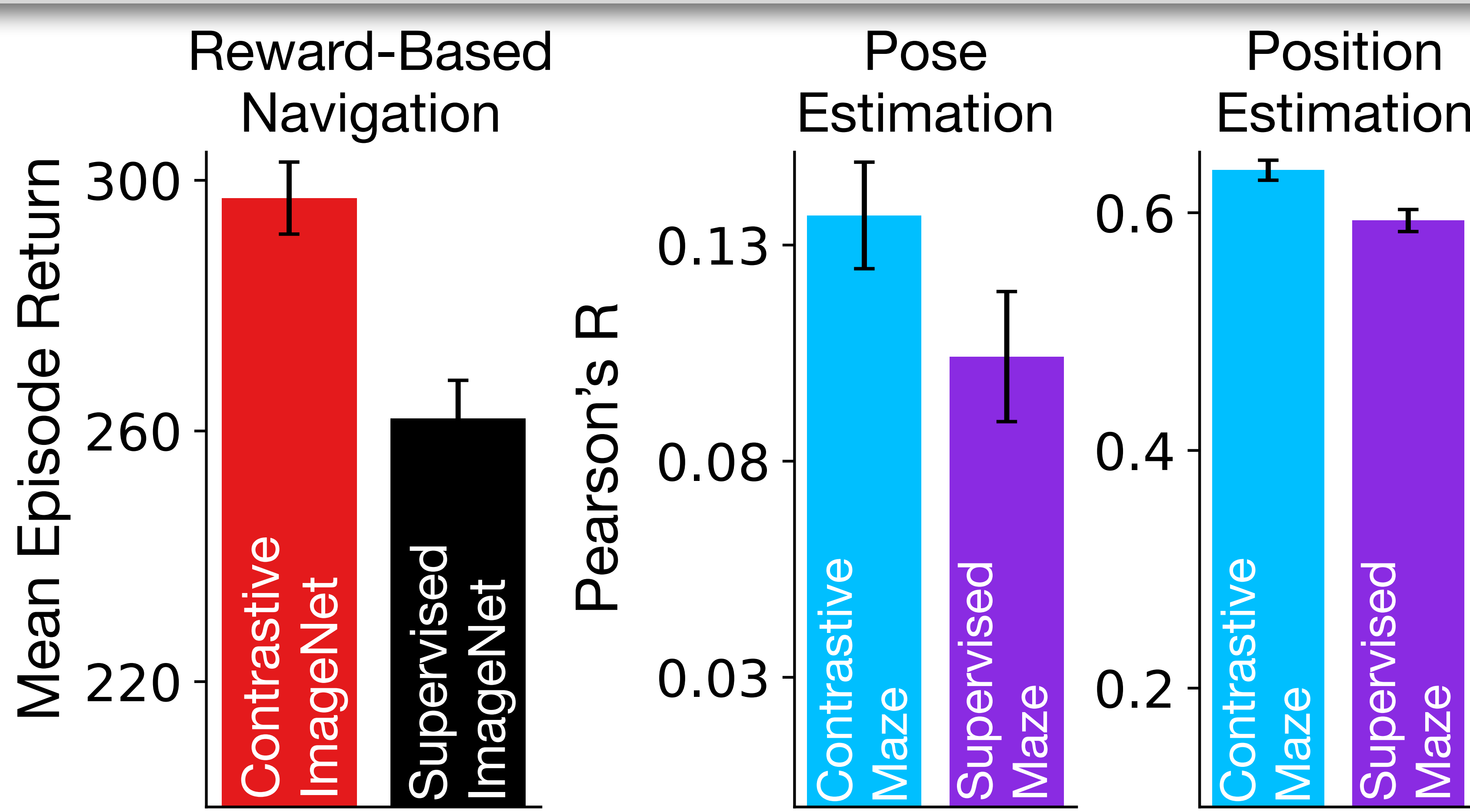
Visual Scene Understanding

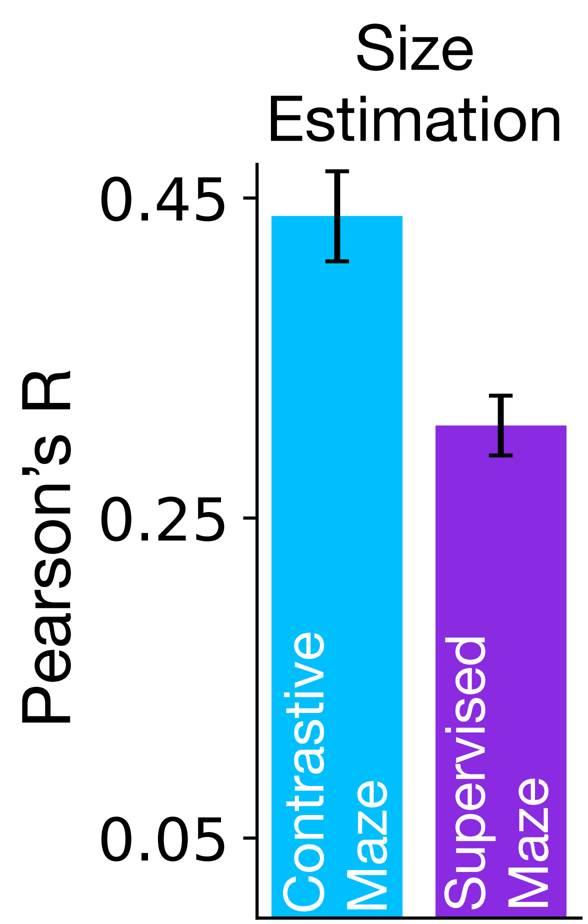




Object properties *Texture*

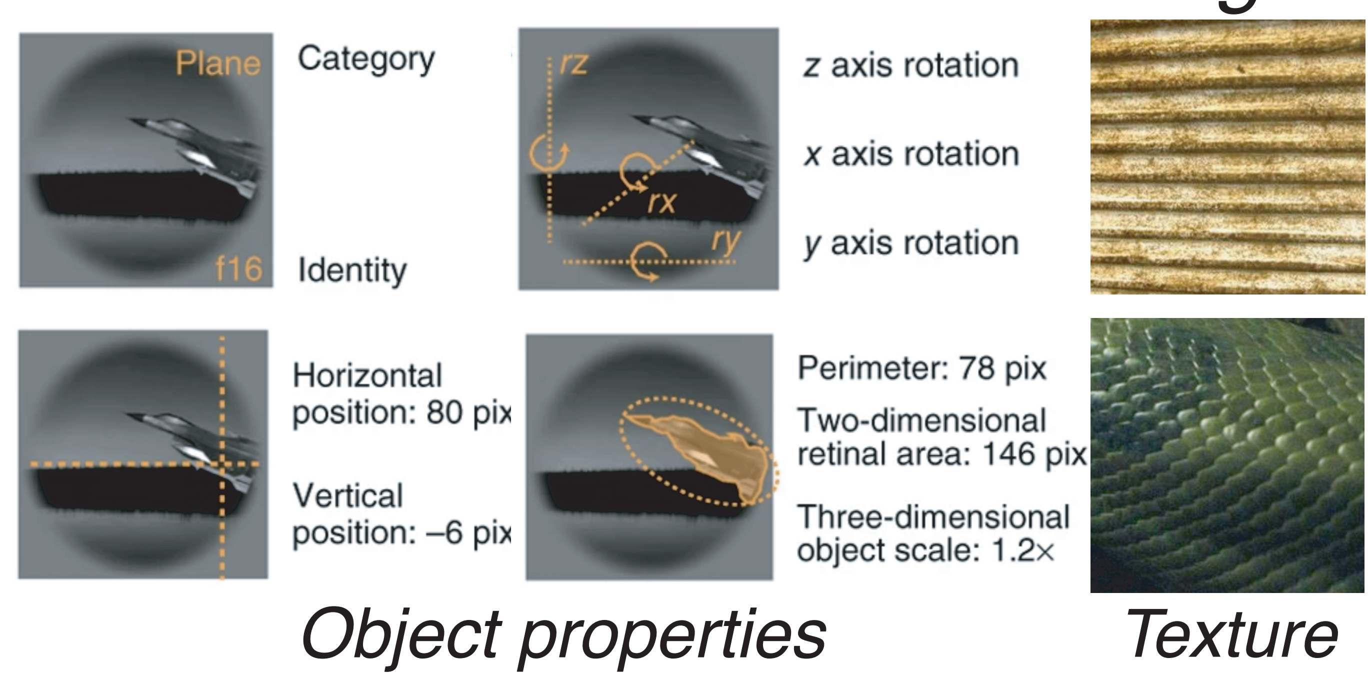
Contrastive models yield better transfer performance



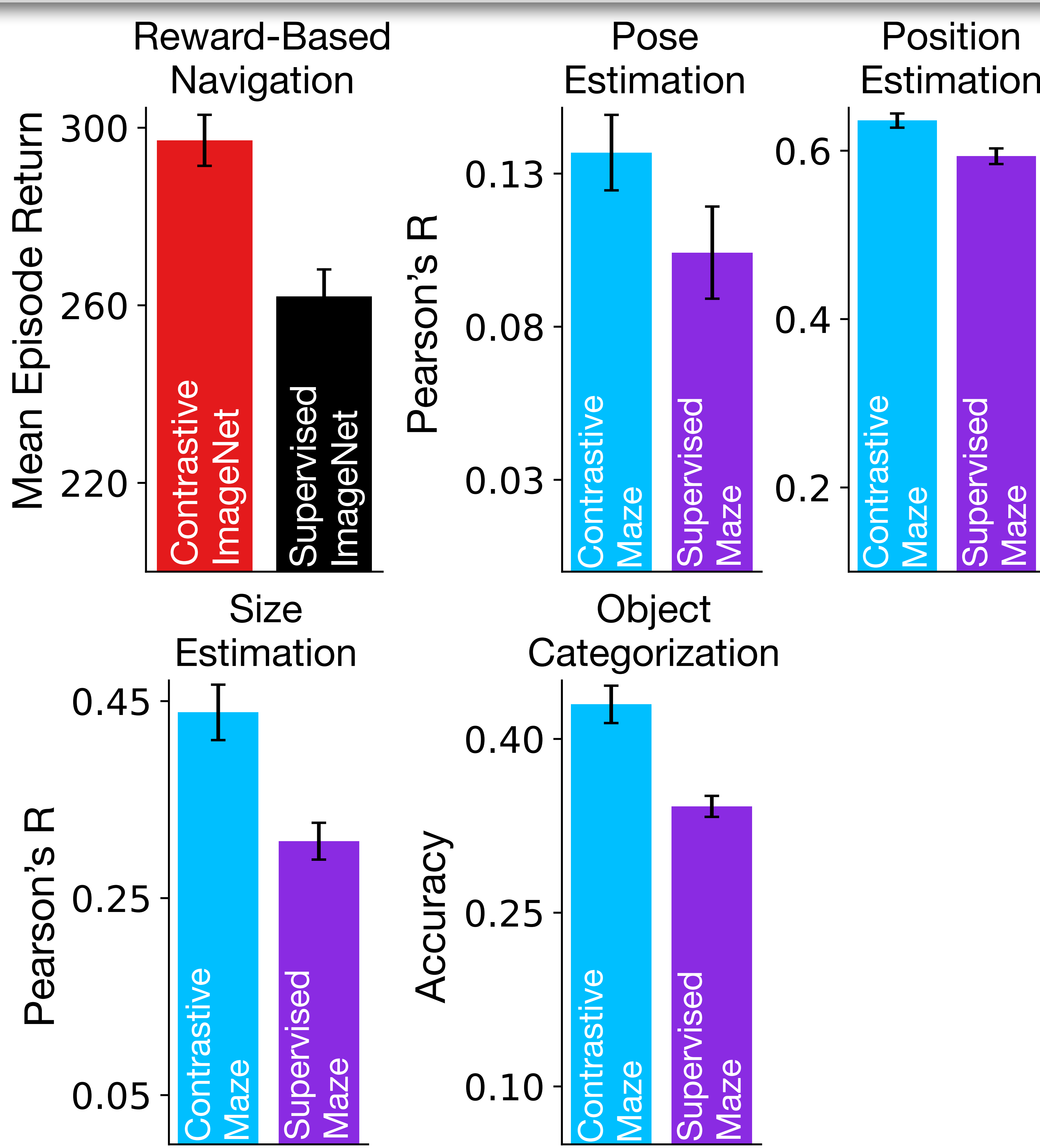


Evaluate

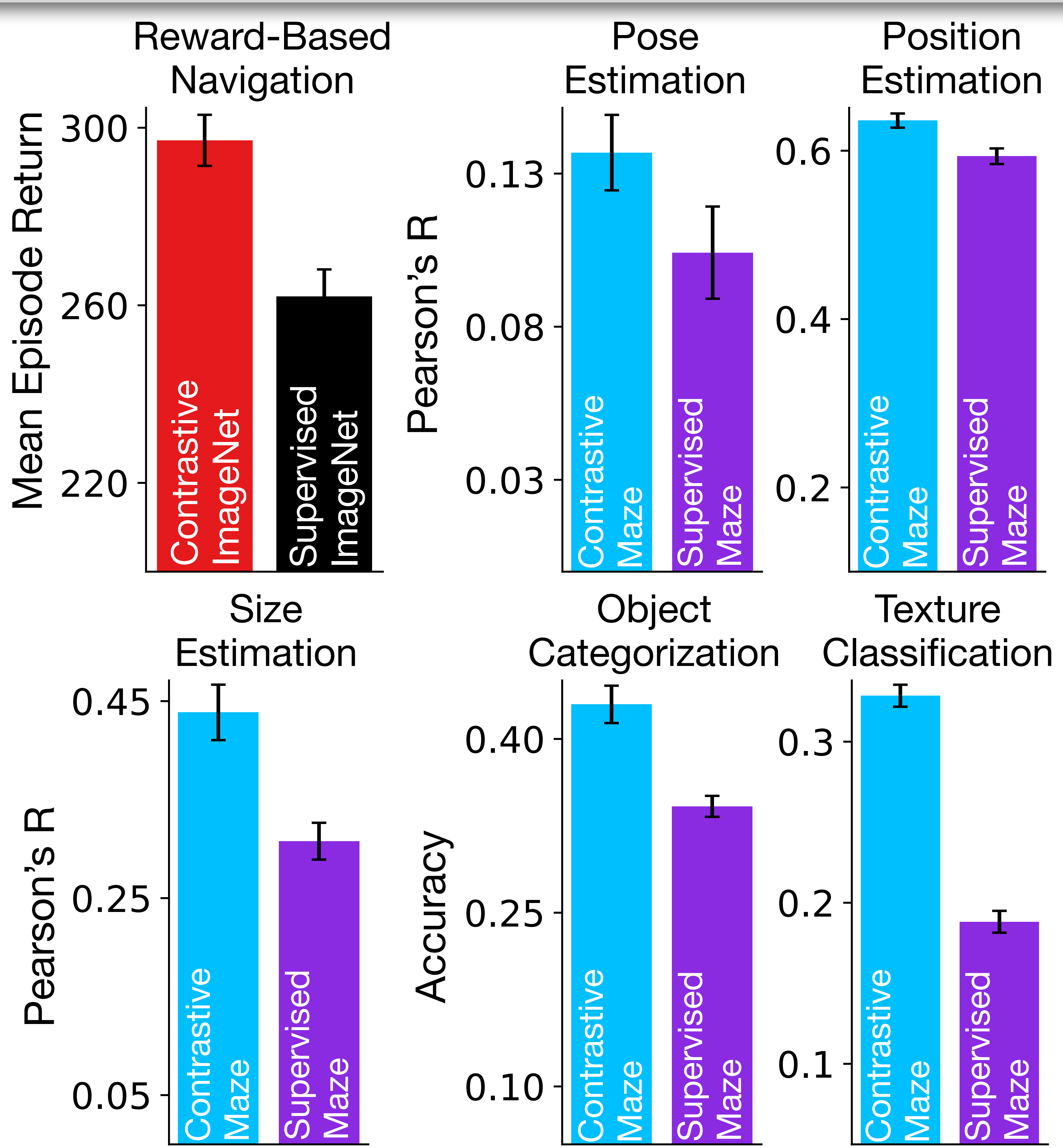
Visual Scene Understanding



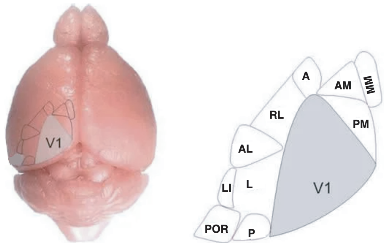
Contrastive models yield better transfer performance

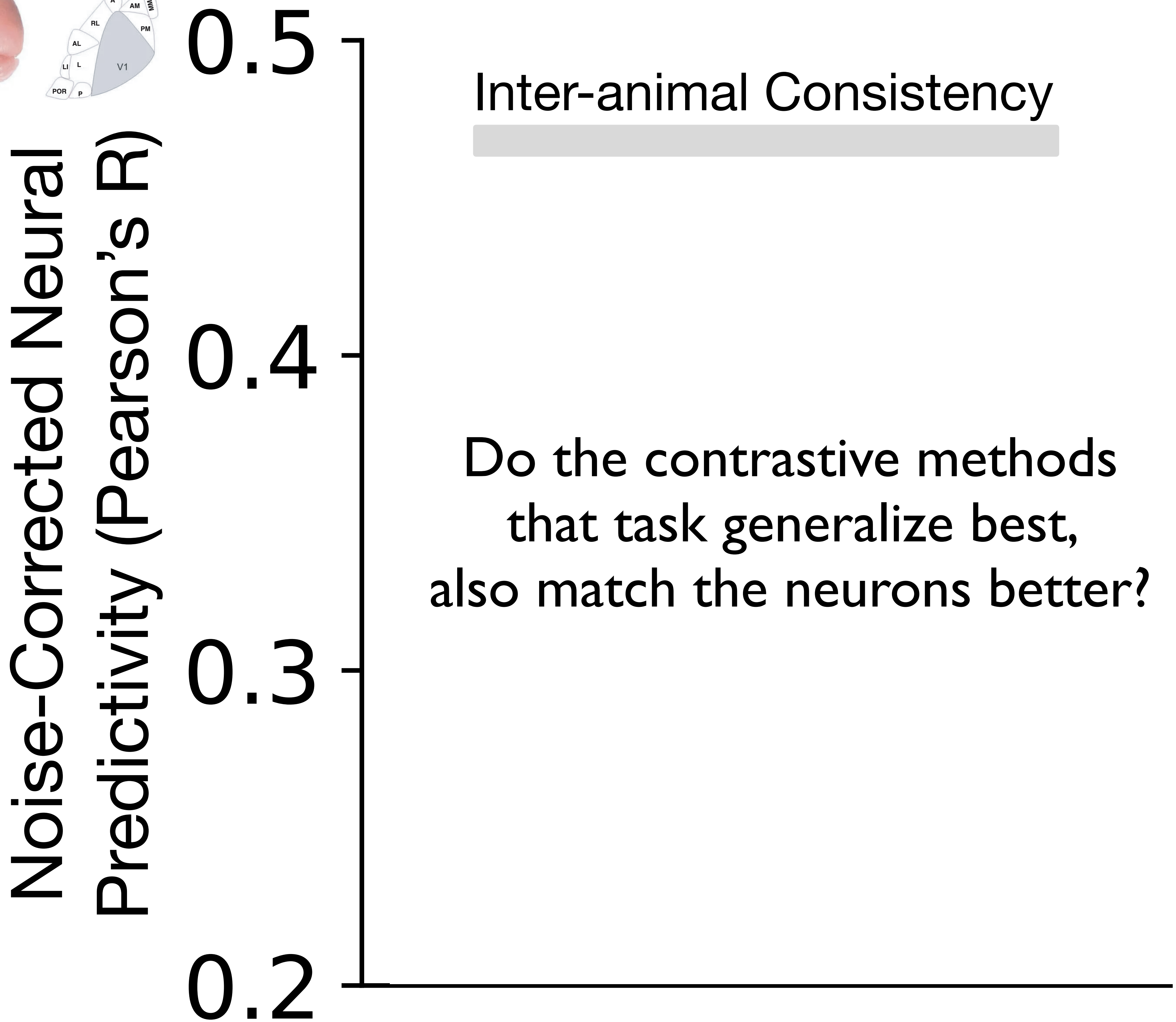


Contrastive models yield better transfer performance



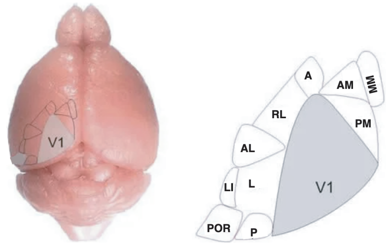
Best models that match the neurons have the best transfer

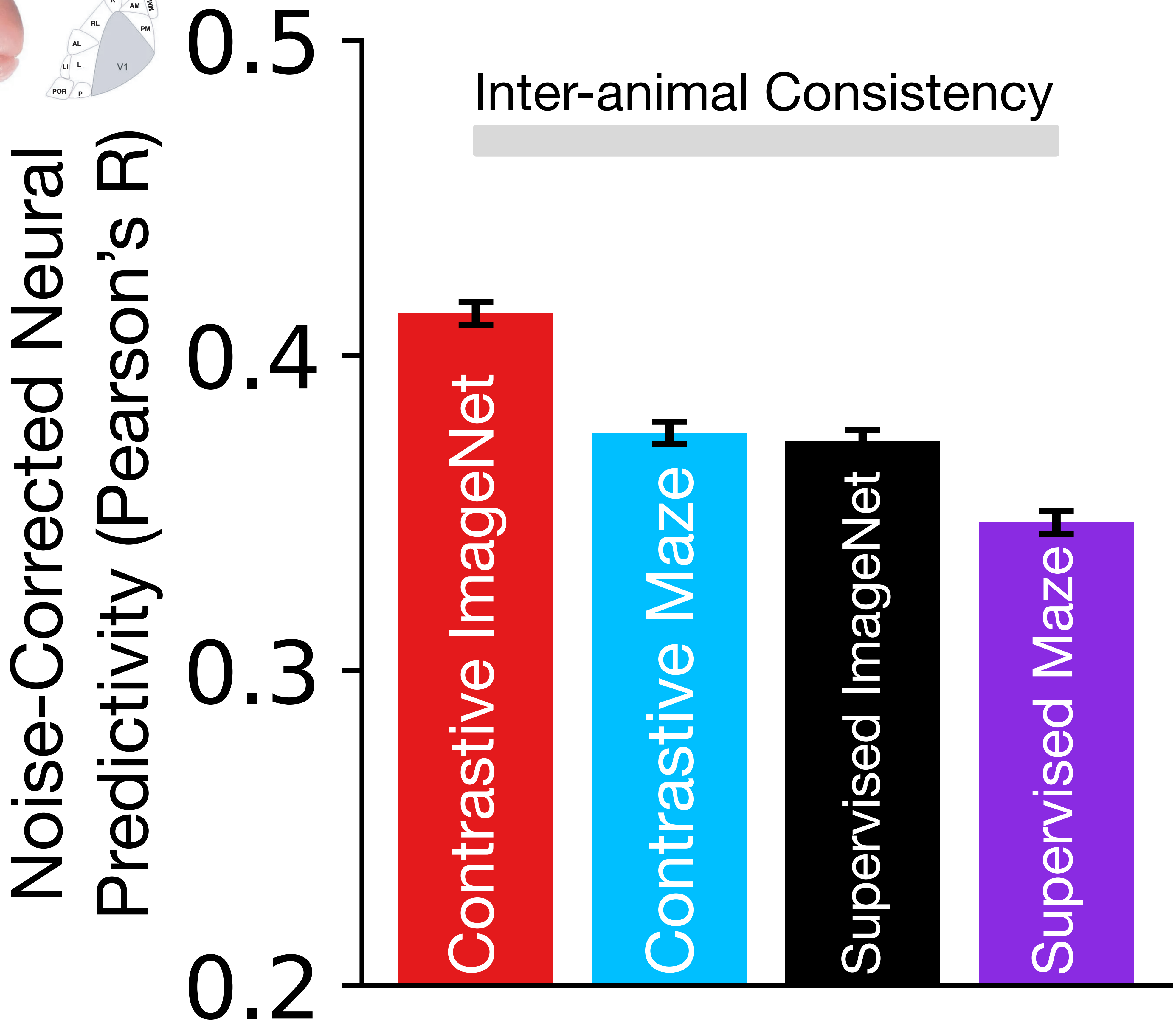




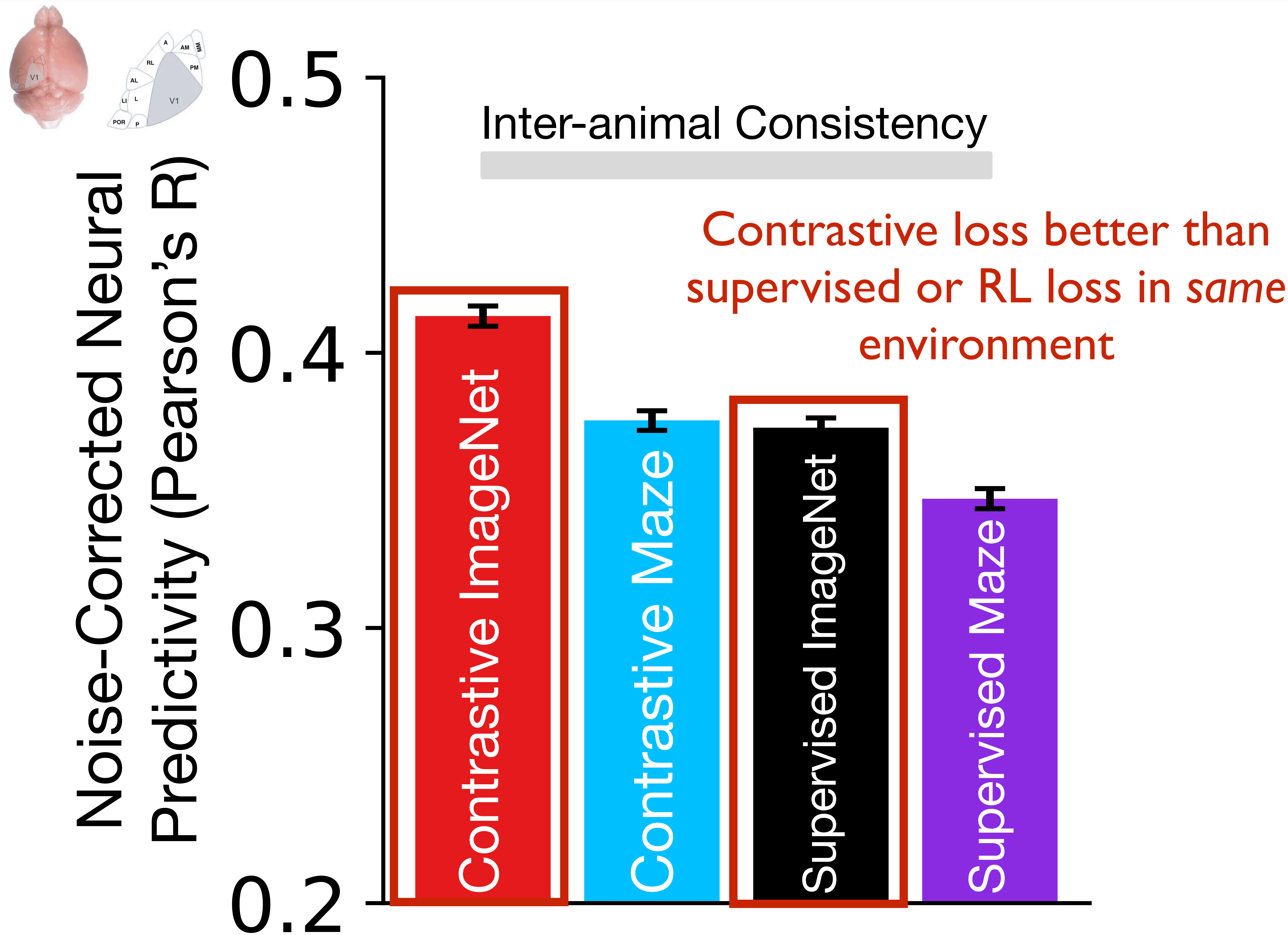
Do the contrastive methods that task generalize best, also match the neurons better?

Best models that match the neurons have the best transfer

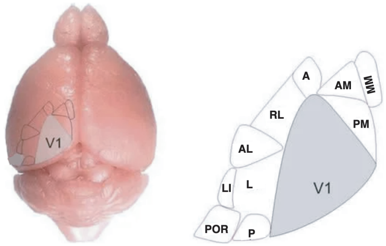


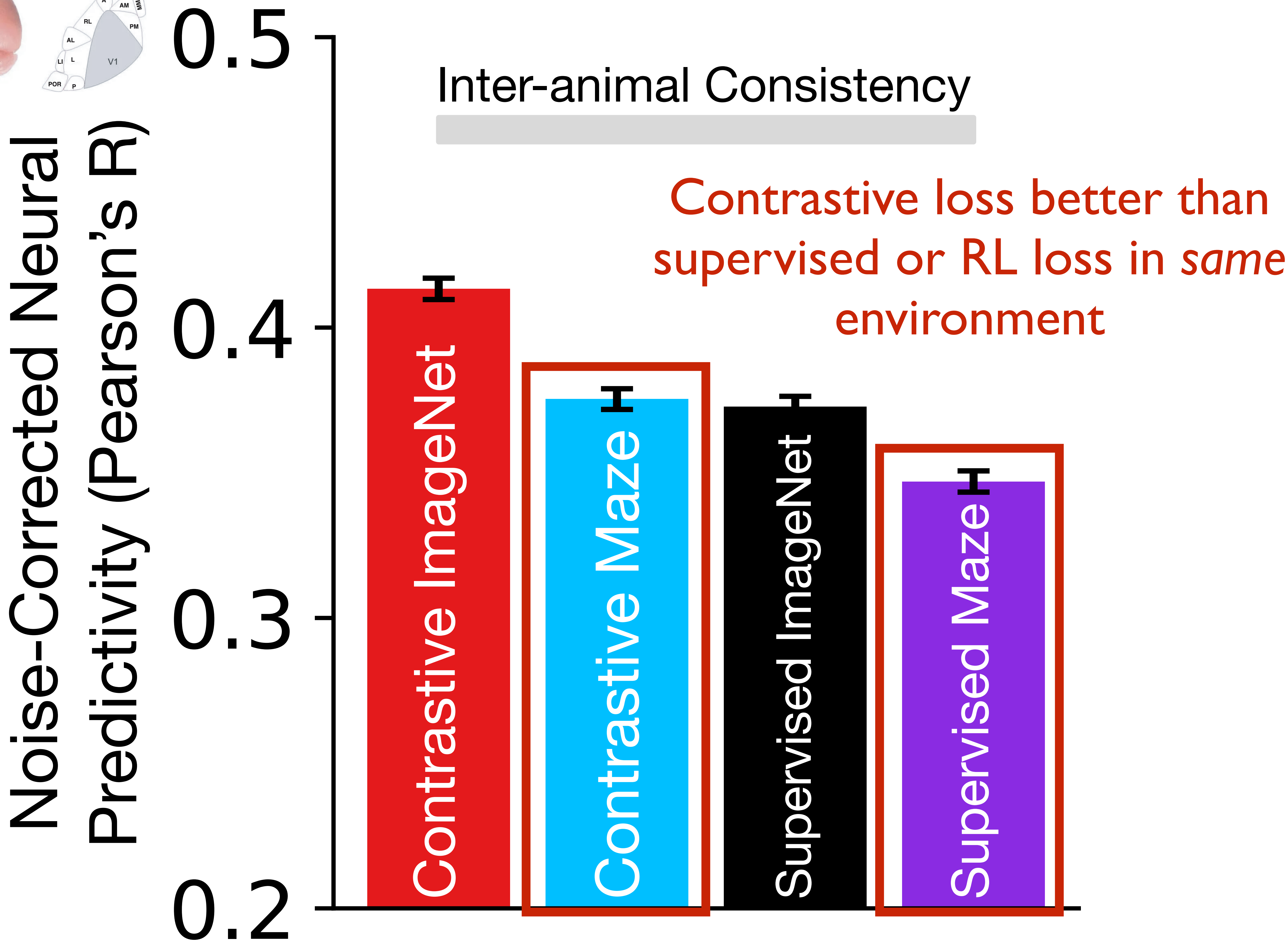


Best models that match the neurons have the best transfer

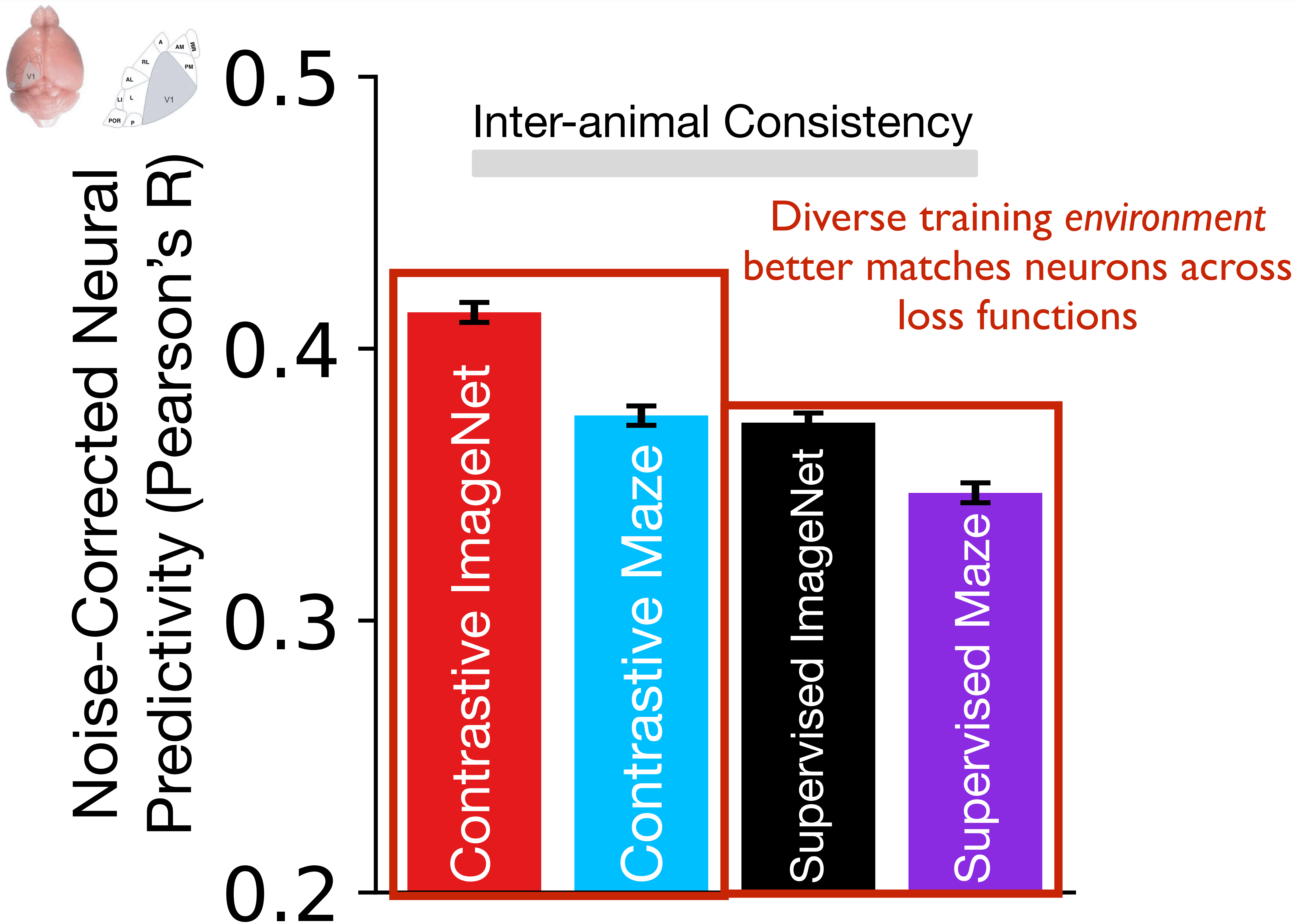


Best models that match the neurons have the best transfer





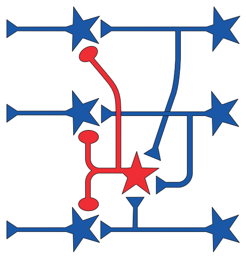
Best models that match the neurons have the best transfer



Takeaways

A = *architecture class*

1. *"Circuit"*

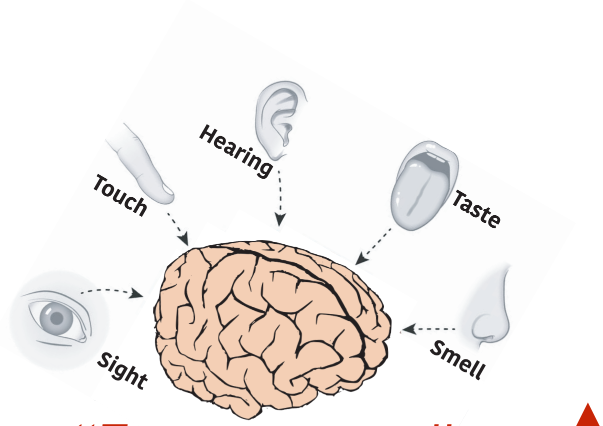


T = *task loss*

3. *"Ecological niche/behavior"*

Neurobiological Puzzle:

What are the core differences between mouse visual cortex and the primate ventral stream?



2. *"Environment"*

D = *data stream*

Takeaways

A = *architecture class*

1. *"Circuit"*

shallow

T = *task loss*

3. *"Ecological niche/behavior"*

self-supervised

Neurobiological Puzzle:

What are the core differences between mouse visual cortex and the primate ventral stream?

Findings:

Mouse visual cortex is a general-purpose system utilizing its limited resources to perform a variety of visual tasks.

In contrast to the deep, high-resolution, and more categorization-dominated primate ventral stream.

low resolution

2. *"Environment"*

D = *data stream*

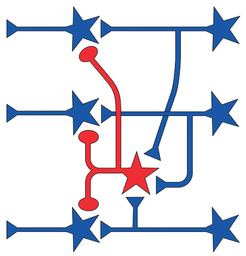
Outline

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

Reusable Latent Representations for Primate Mental Simulation

A = *architecture class*

1. *"Circuit"*



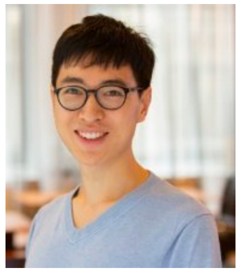
T = *task loss*

3. *"Ecological niche/behavior"*

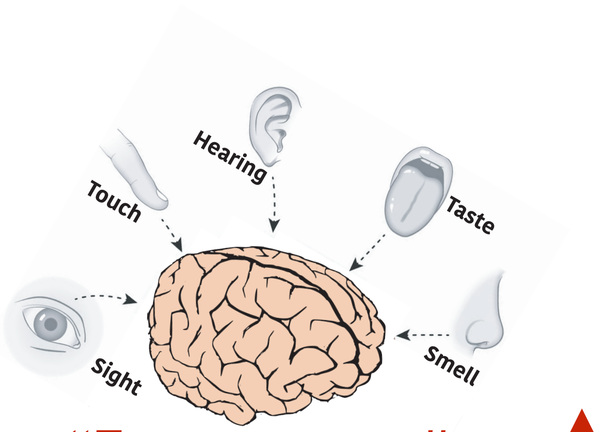
A. Nayebi, R. Rajalingham, M. Jazayeri, G.R. Yang
Neural foundations of mental simulation: future prediction of latent representations on dynamic scenes.
arXiv:2305.11772

Rishi Rajalingham

Mehrdad Jazayeri



Guangyu Robert Yang



2. *"Environment"*

D = *data stream*

Motivation


氷みくじ
氷みくじ


NEW
Specialty Drinks
CHOKING
NO SMOKING
5094M
F 32803


WHOLE WHEAT
Mueller
PRO-SERIES
MULTI
LODGE
CAST IRON
levoit

Motivation

Infer:
Has this ice block been out longer?

Motivation

Infer:
Has this ice block been out longer?

Predict:
Are these stacks stable?

Motivation

Infer:
Has this ice block been out longer?

Predict:
Are these stacks stable?

Predict:
Will this box support me?

Motivation

Infer:
Has this ice block been out longer?

Predict:
Are these stacks stable?

Predict:
Will this box support me?

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

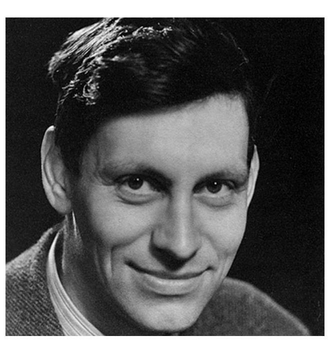
The Mental Simulation Hypothesis

The Nature of Explanation

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

Craik (1943)



Kenneth Craik

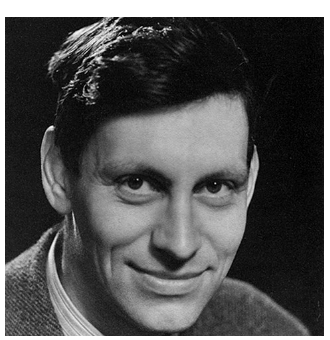
The Mental Simulation Hypothesis

The Nature of Explanation

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

<u>Craik (1943):</u> The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



Kenneth Craik

The Mental Simulation Hypothesis

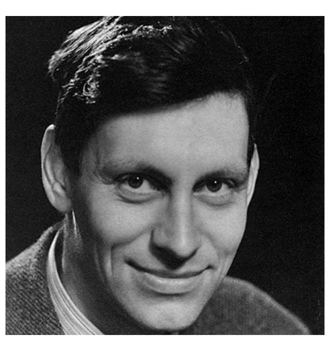
The Nature of Explanation

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

Pre-dates the modern computer!

Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



Kenneth Craik

The Mental Simulation Hypothesis: Behavioral Evidence

The Nature of Explanation

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

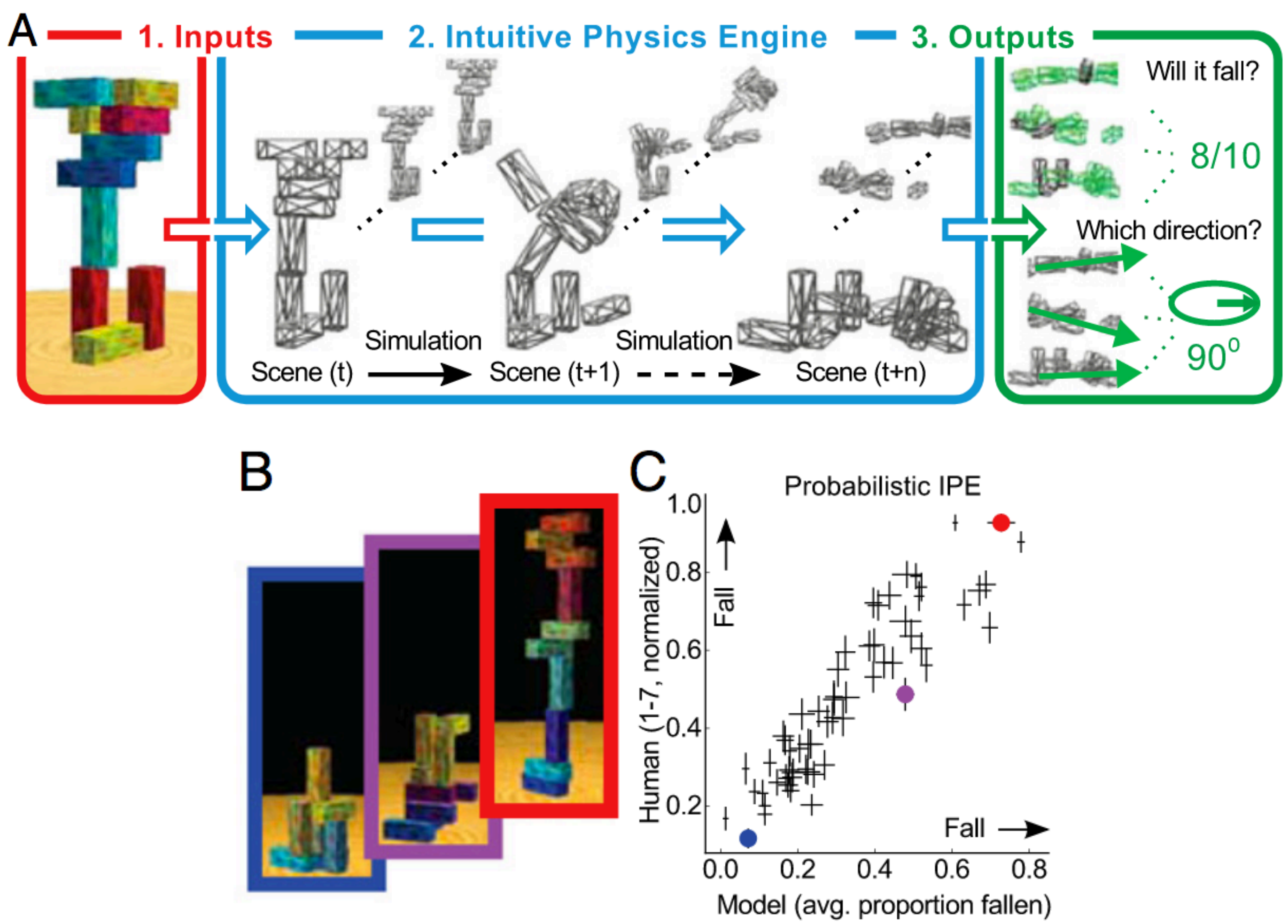
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**

The Mental Simulation Hypothesis: Behavioral Evidence

The Nature of Explanation

> My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .
>
> If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



Battaglia, Hamrick, Tenenbaum 2013

Intuitive Physics Engine (IPE) can match human physical judgements

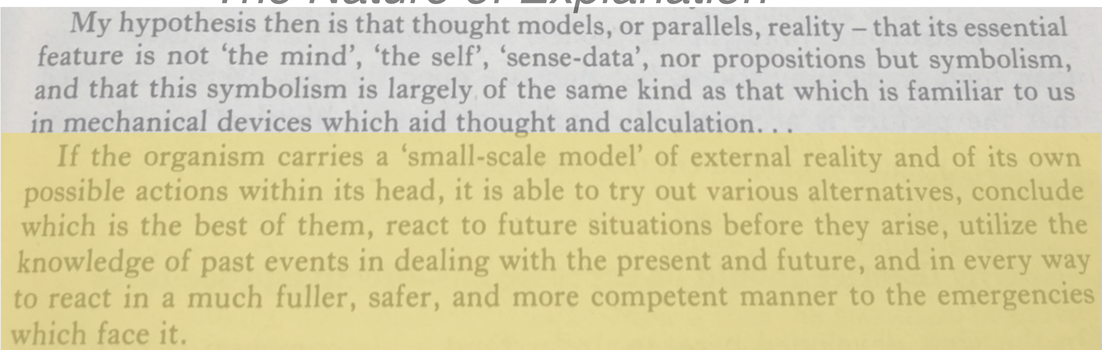
Peter Battaglia

Jessica Hamrick

Joshua Tenenbaum

The Mental Simulation Hypothesis: Human Neuroimaging Evidence

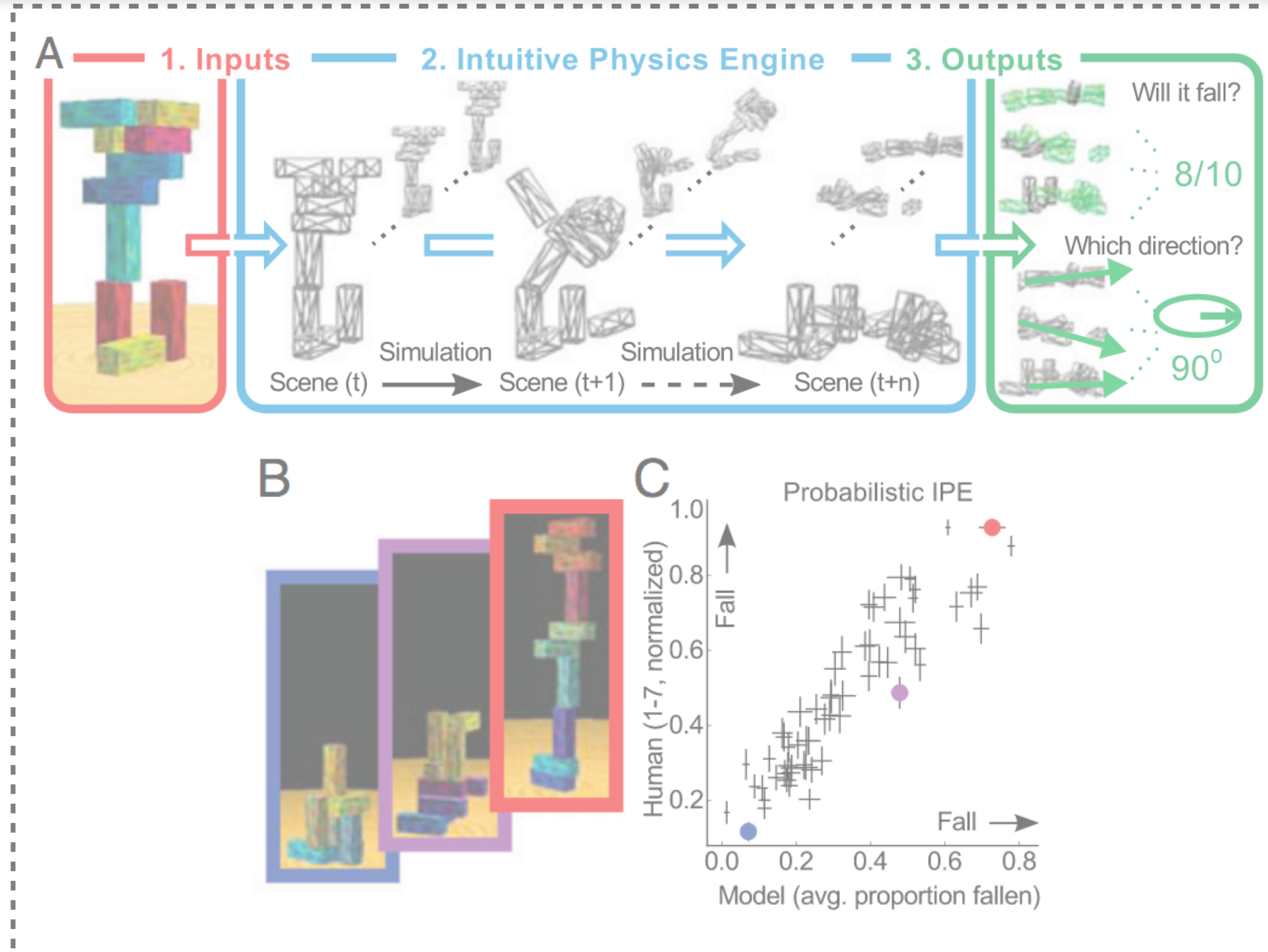
The Nature of Explanation



My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

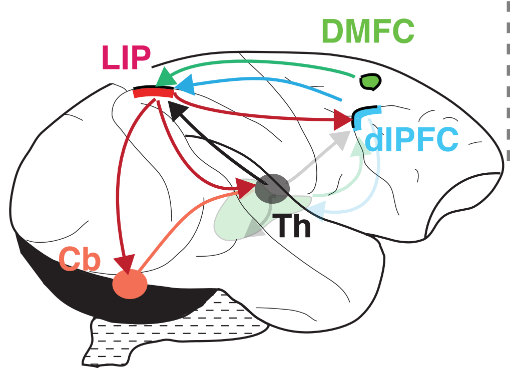
If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**

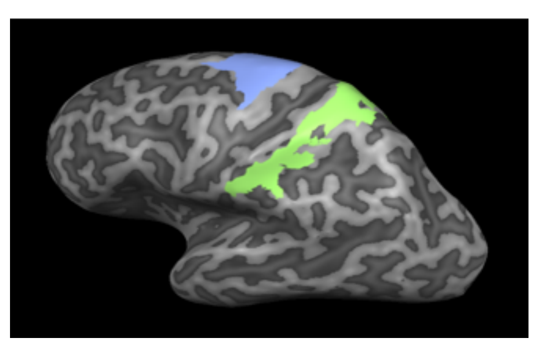


Battaglia, Hamrick, Tenenbaum 2013

The Brain's "Physics Engine"



Fronto-Parietal Network



Joshua Tenenbaum

Nancy Kanwisher

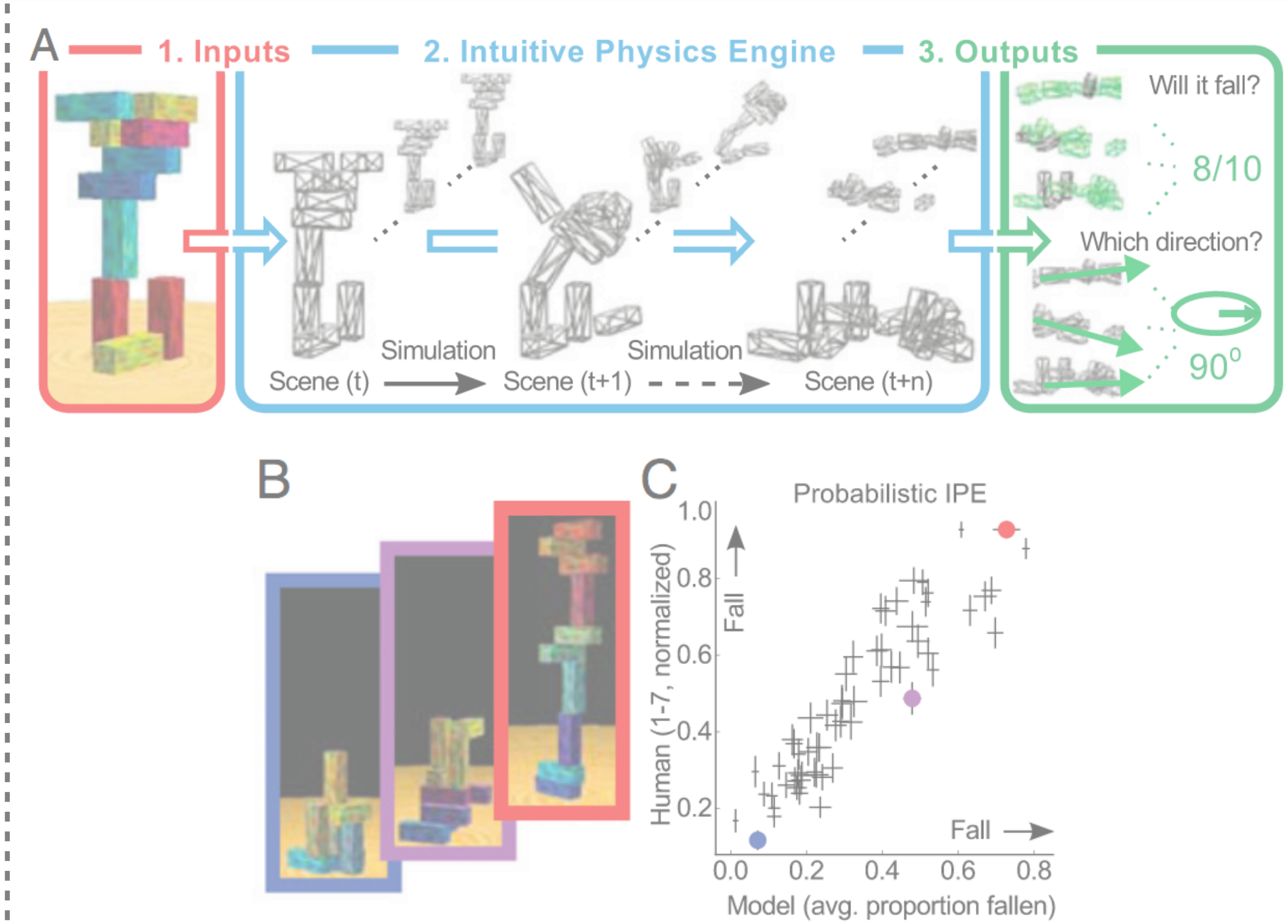
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

The Nature of Explanation

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

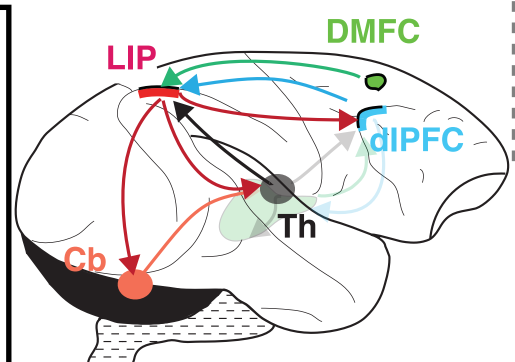
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



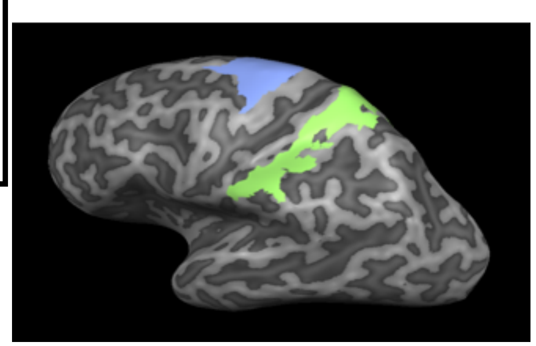
Battaglia, Hamrick, Tenenbaum 2013

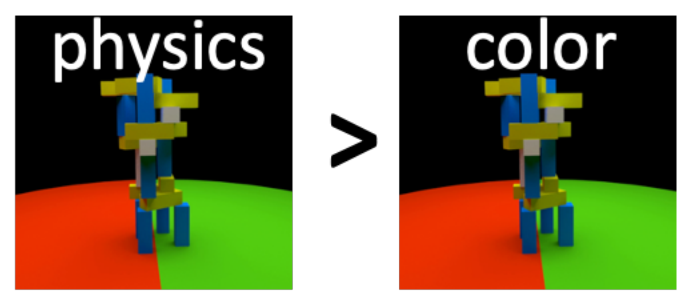
The Brain's "Physics Engine"

- A network of brain regions recruited by physical inferences (*Fischer et al. 2016*)



Fronto-Parietal Network





Fischer et al. 2016

Joshua Tenenbaum

Nancy Kanwisher

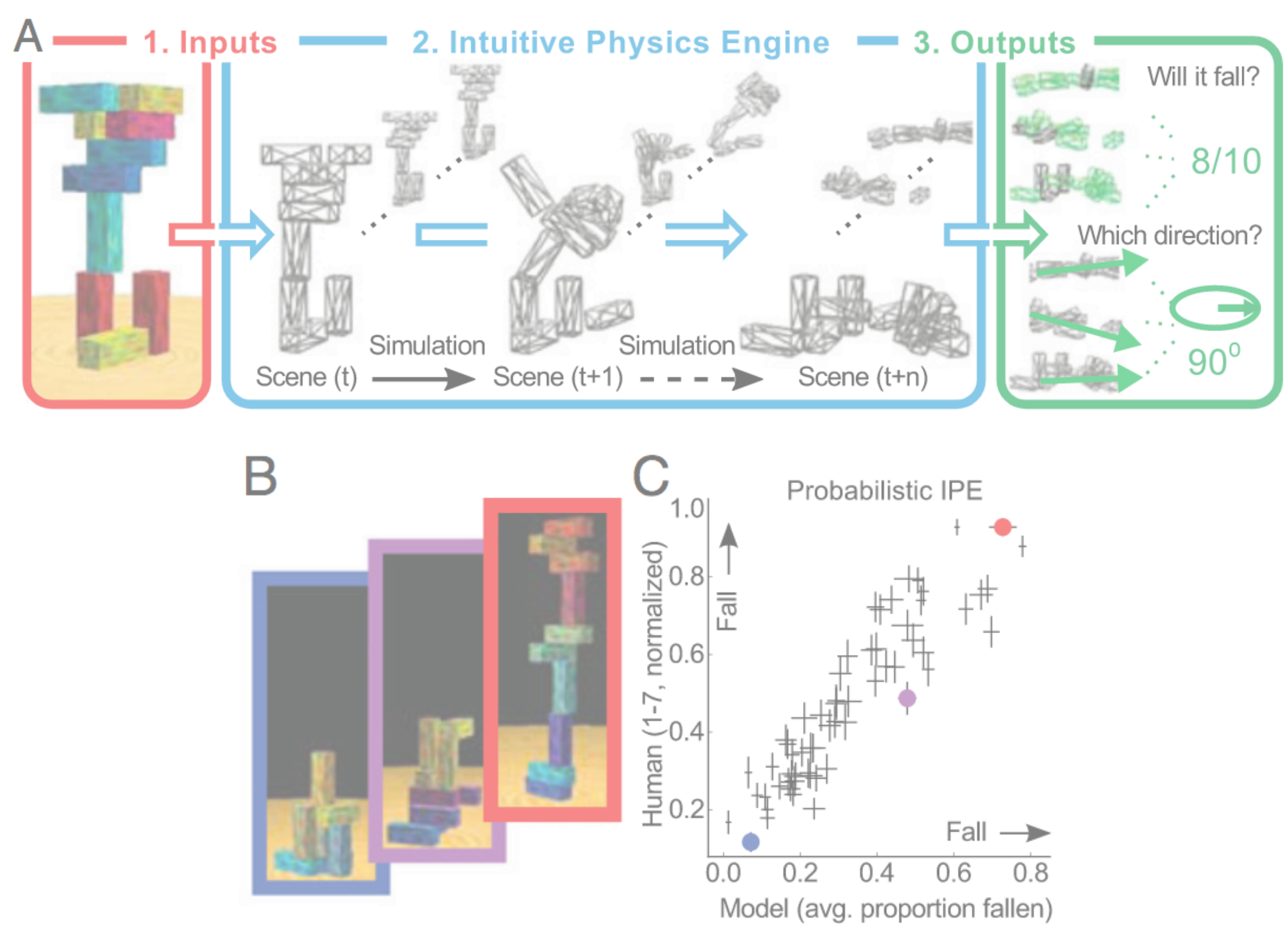
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

The Nature of Explanation

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

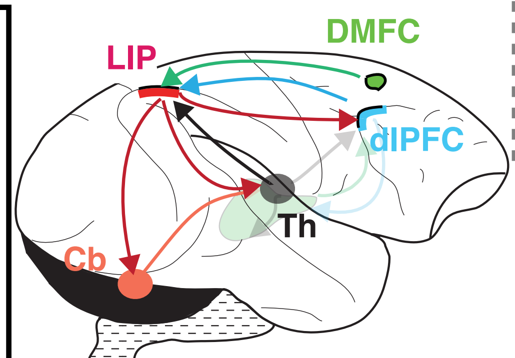
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



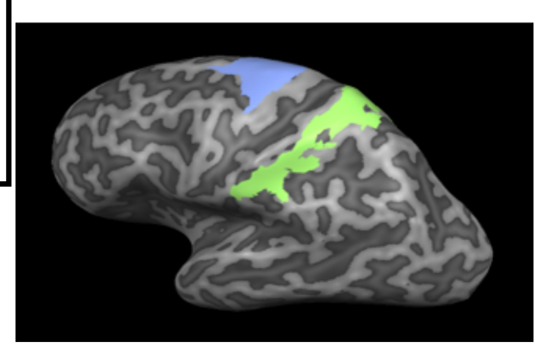
Battaglia, Hamrick, Tenenbaum 2013

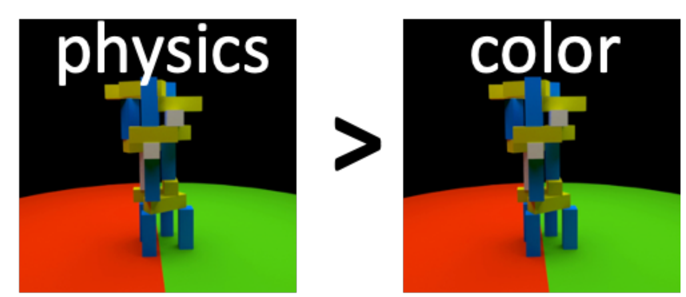
The Brain's "Physics Engine"

- A network of brain regions recruited by physical inferences (*Fischer et al. 2016*)
- Contains information about mass (*Schwettmann et al. 2019*)

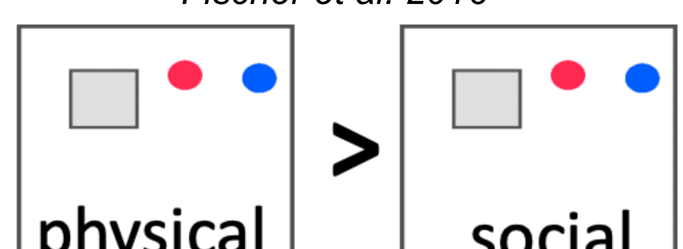


Fronto-Parietal Network





Fischer et al. 2016



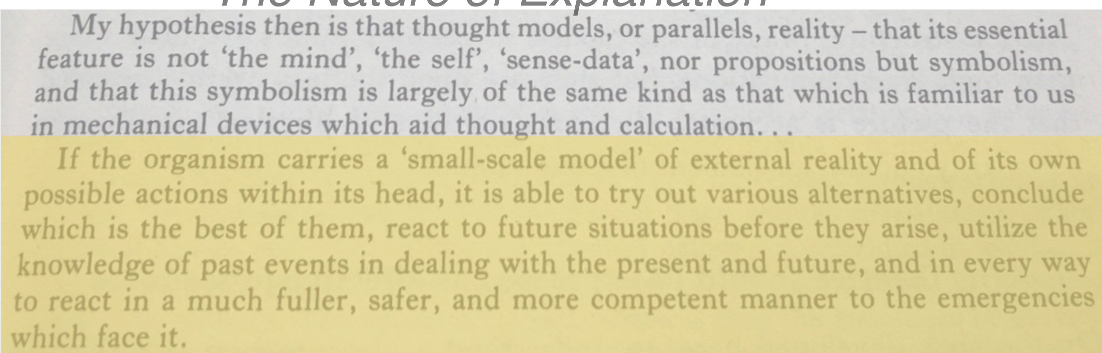
Schwettmann et al. 2019

Joshua Tenenbaum

Nancy Kanwisher

The Mental Simulation Hypothesis: Human Neuroimaging Evidence

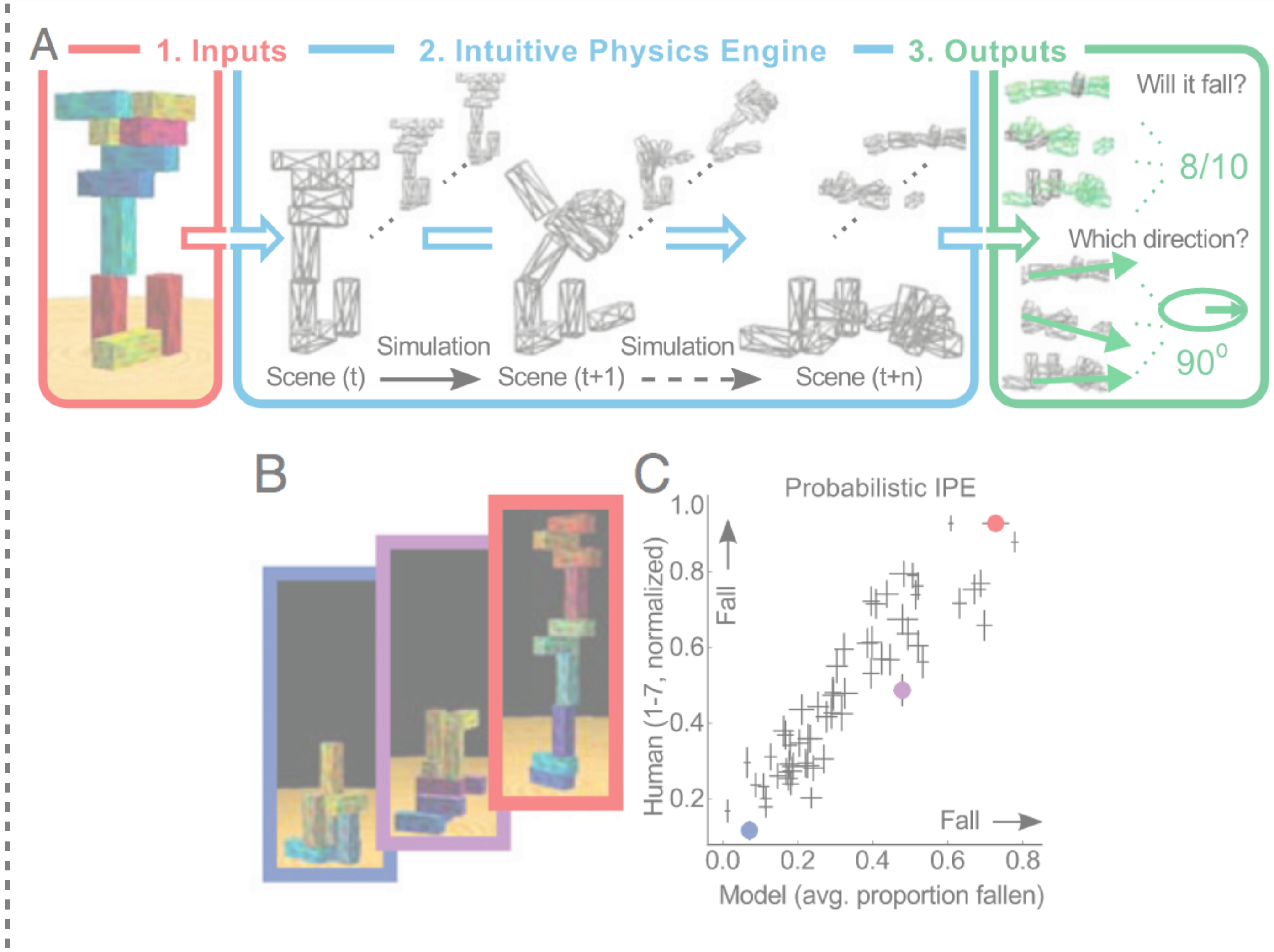
The Nature of Explanation



My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

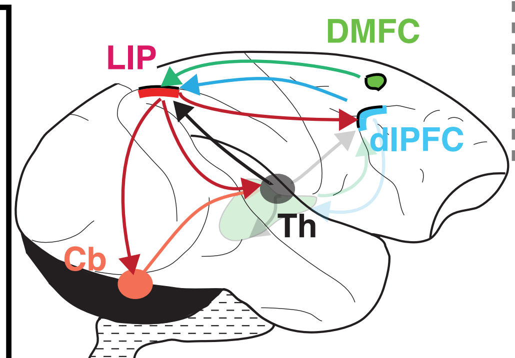
<u>Craik (1943):</u> The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



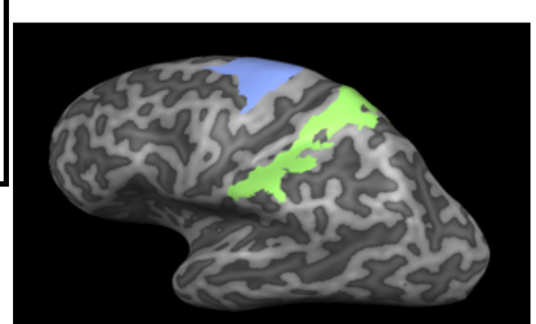
Battaglia, Hamrick, Tenenbaum 2013

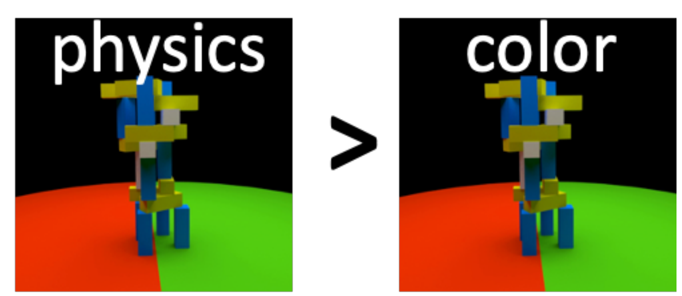
The Brain's "Physics Engine"

- A network of brain regions recruited by physical inferences (*Fischer et al. 2016*)
- Contains information about mass (*Schwettmann et al. 2019*)
- Contains information about physical stability (*Pramod et al. 2022*)

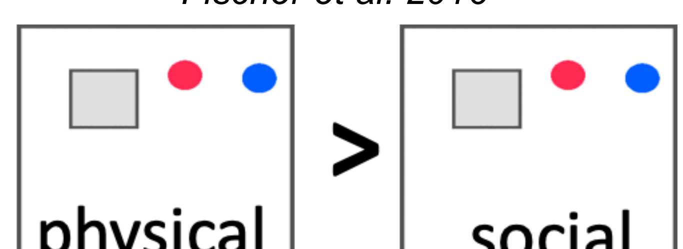


Fronto-Parietal Network





Fischer et al. 2016



Schwettmann et al. 2019

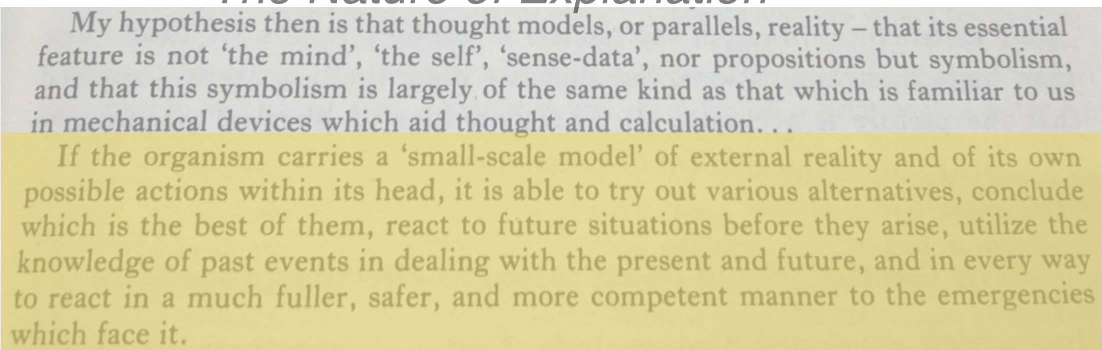
Pramod et al. 2022

Joshua Tenenbaum

Nancy Kanwisher

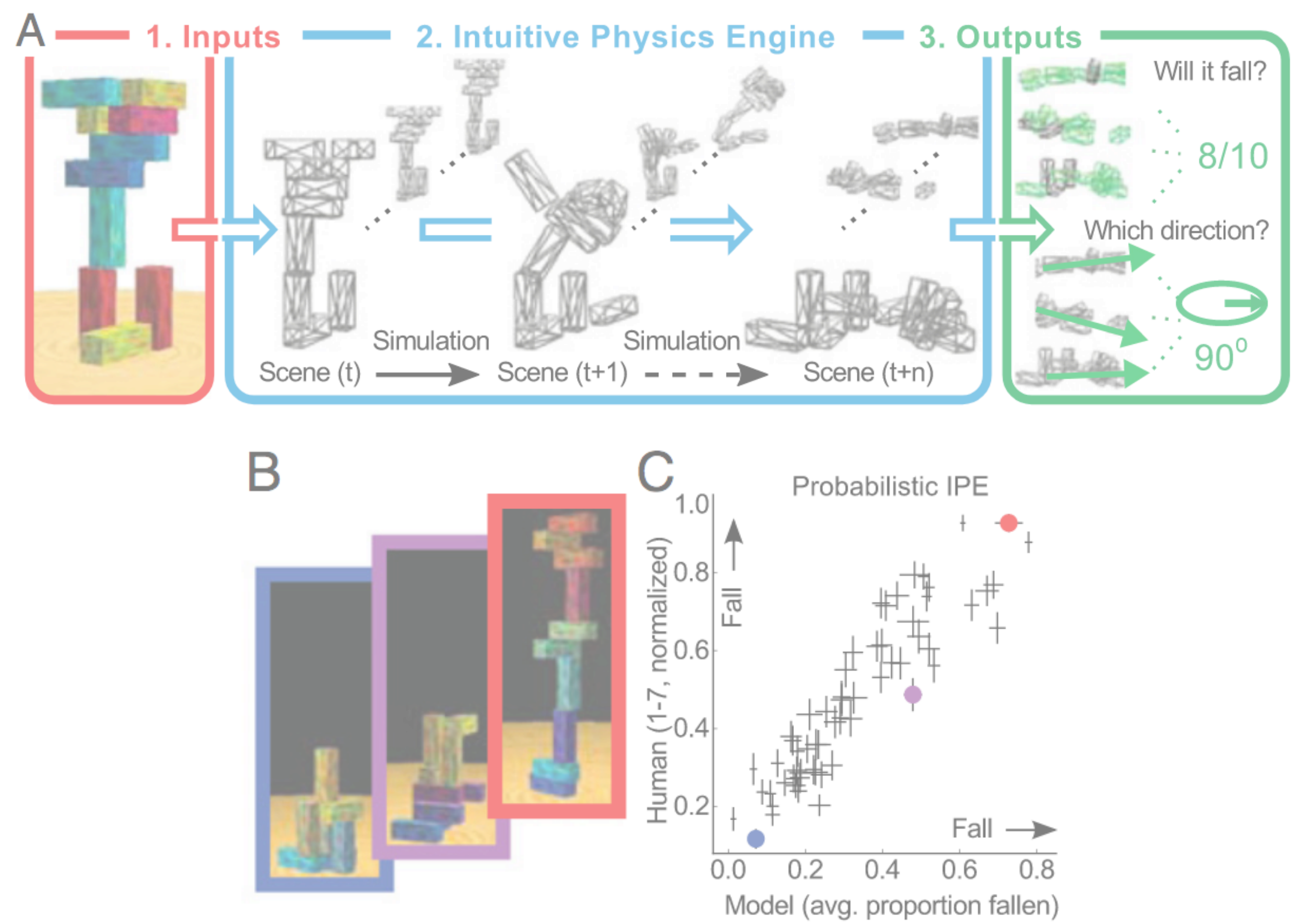
The Mental Simulation Hypothesis: Primate Electrophysiological Evidence

The Nature of Explanation

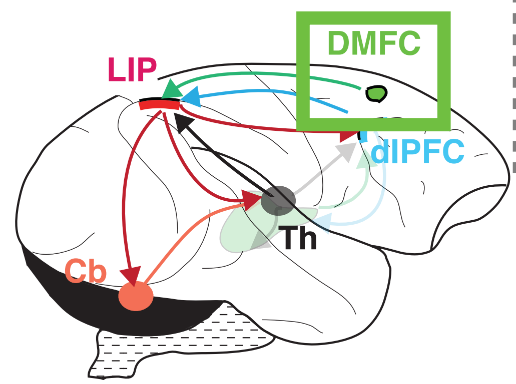

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. . .

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

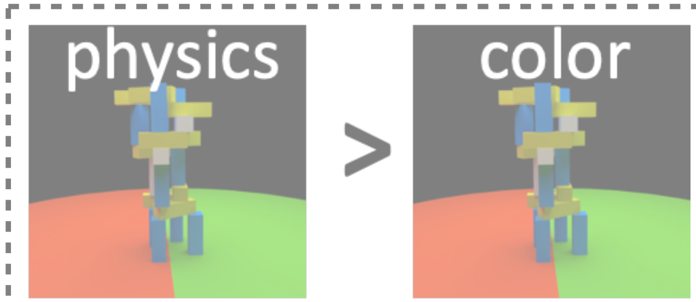
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



Battaglia, Hamrick, Tenenbaum 2013



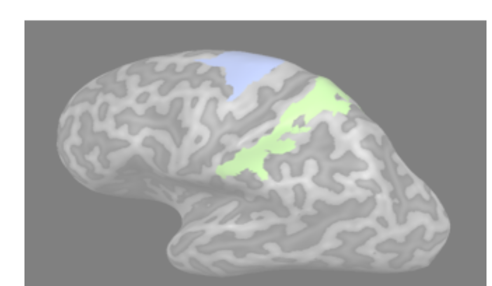
Fronto-Parietal Network

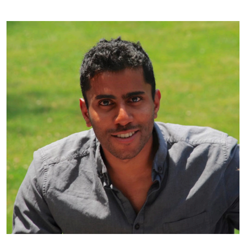


Fischer et al. 2016

Pramod et al. 2022

Schwettmann et al. 2019

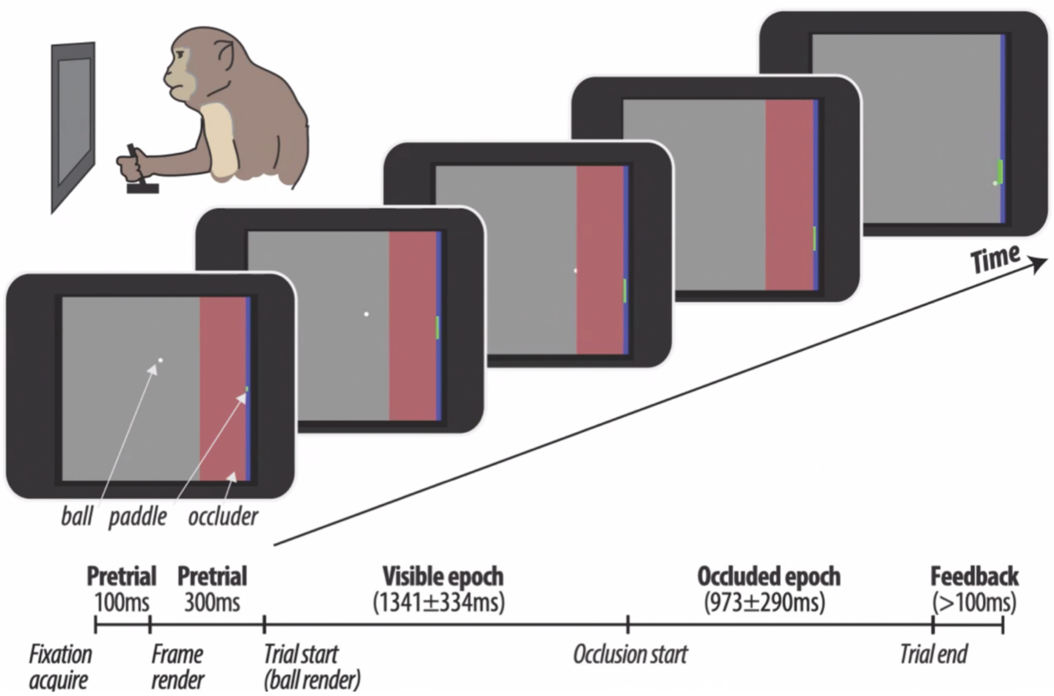




Rishi Rajalingham

Mehrdad Jazayeri

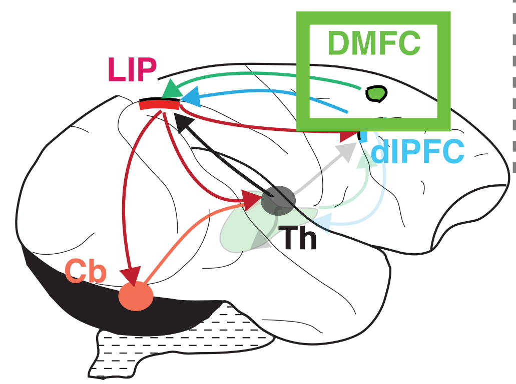
The Mental Simulation Hypothesis: Primate Electrophysiological Evidence



The role of mental simulation in primate physical inference abilities

Rishi Rajalingham, Aida Piccato, Mehrdad Jazayeri

doi: https://doi.org/10.1101/2021.01.14.426741



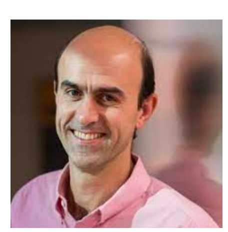
Fronto-Parietal Network

Dynamic tracking of objects in the macaque dorsomedial frontal cortex

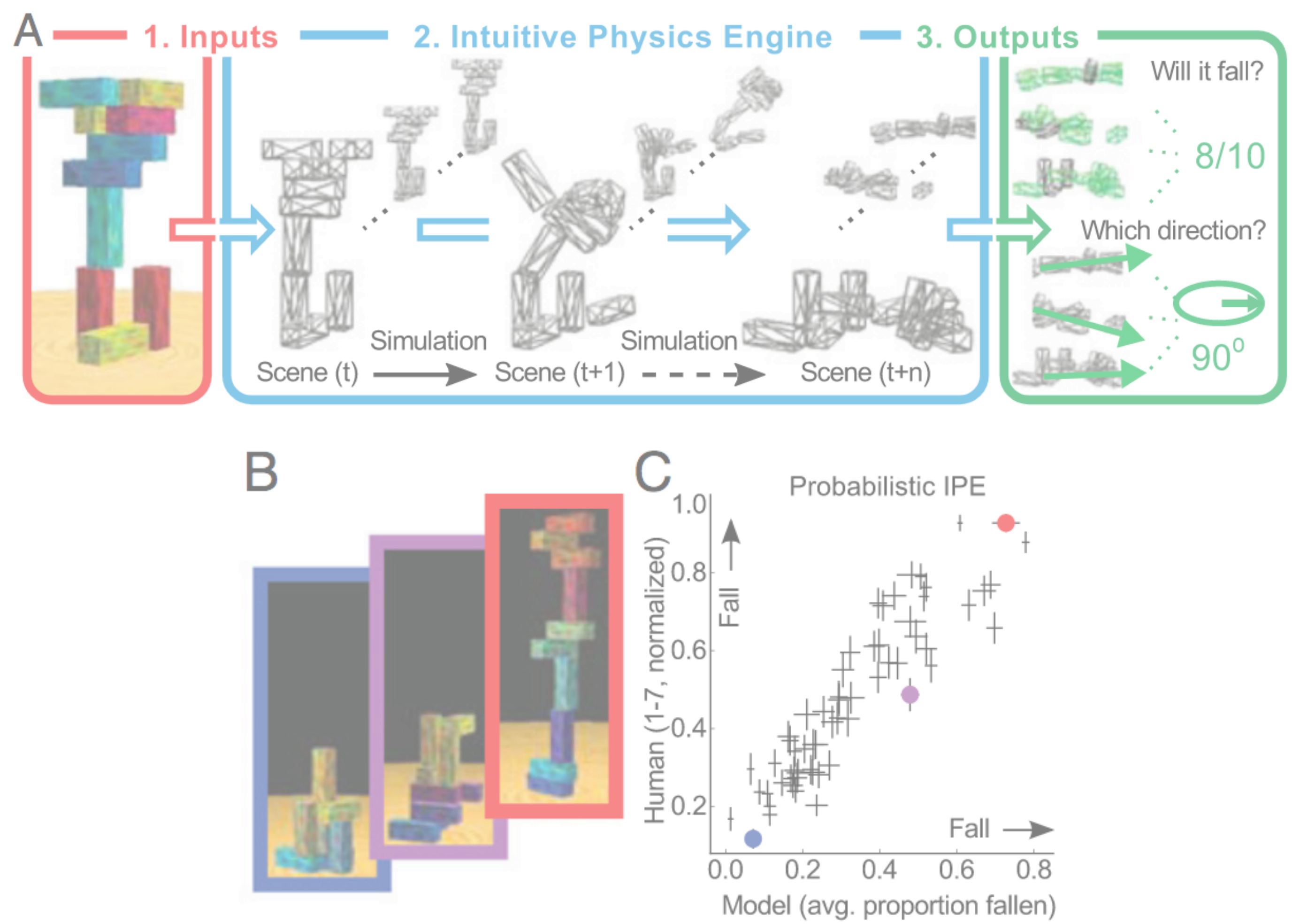
Rishi Rajalingham, Hansem Sohn, Mehrdad Jazayeri

doi: https://doi.org/10.1101/2022.06.24.497529

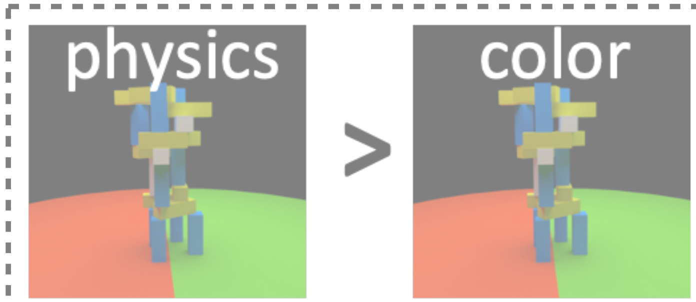
Rishi Rajalingham



Mehrdad Jazayeri



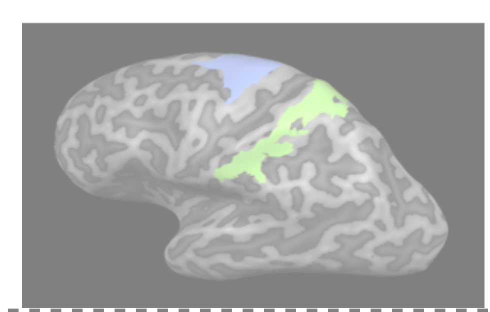
Battaglia, Hamrick, Tenenbaum 2013



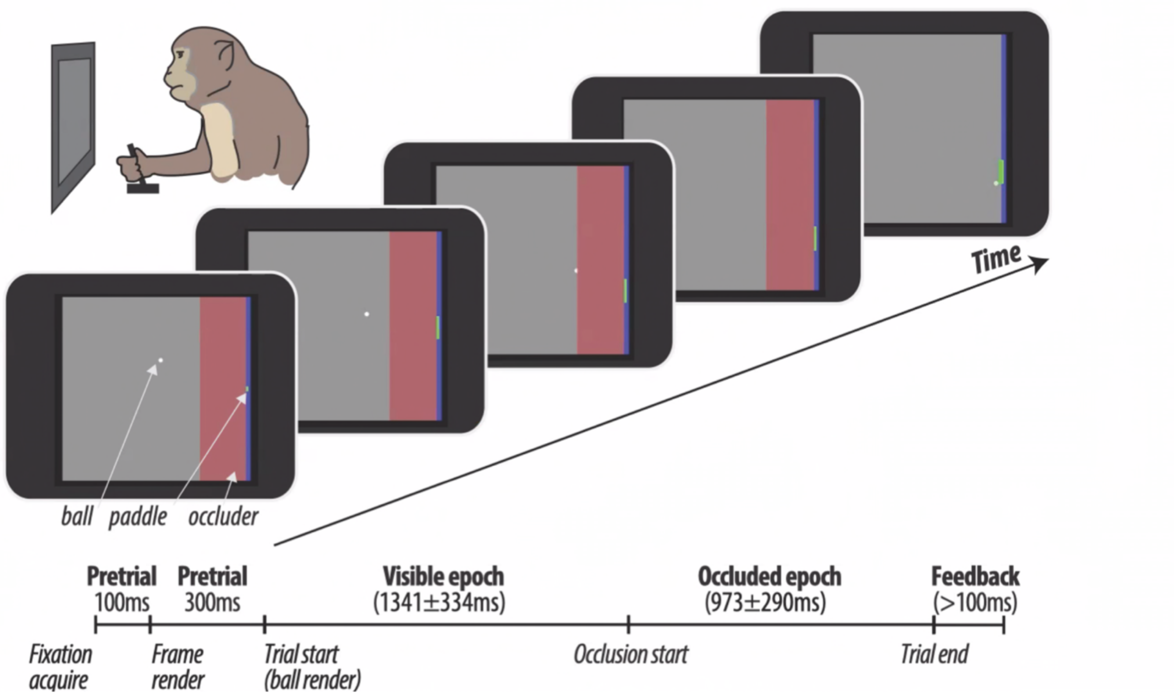
Fischer et al. 2016

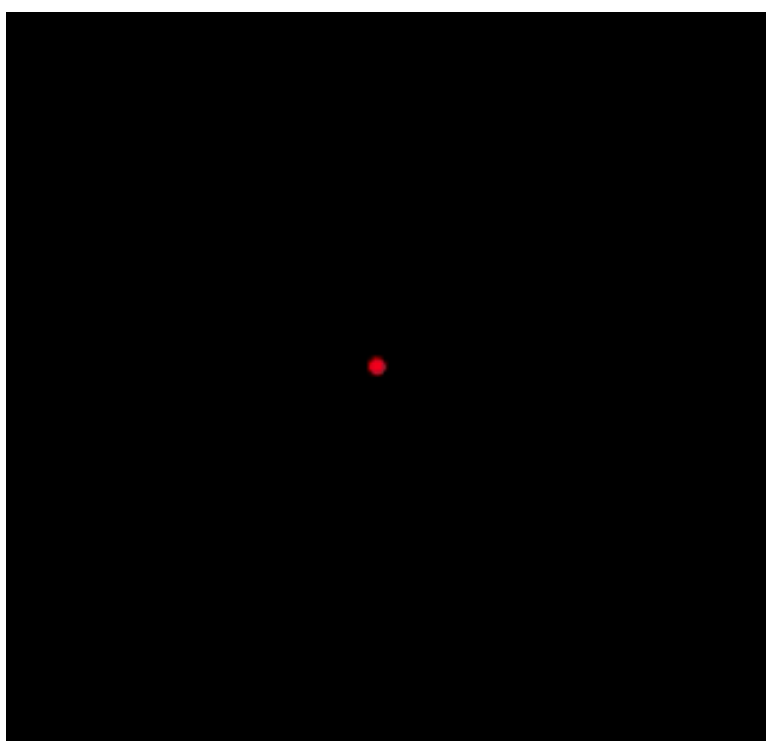
Pramod et al. 2022

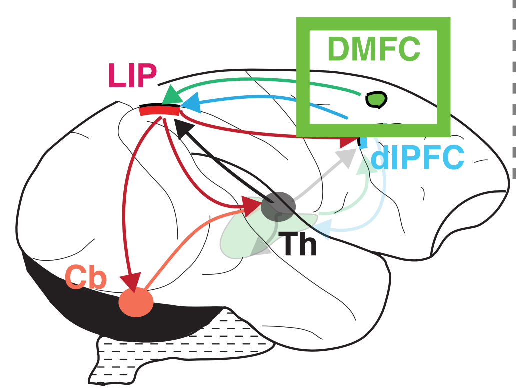
Schwettmann et al. 2019



The Mental Simulation Hypothesis: Primate Electrophysiological Evidence

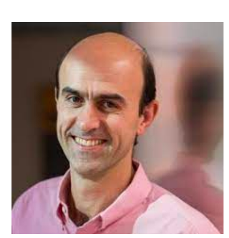




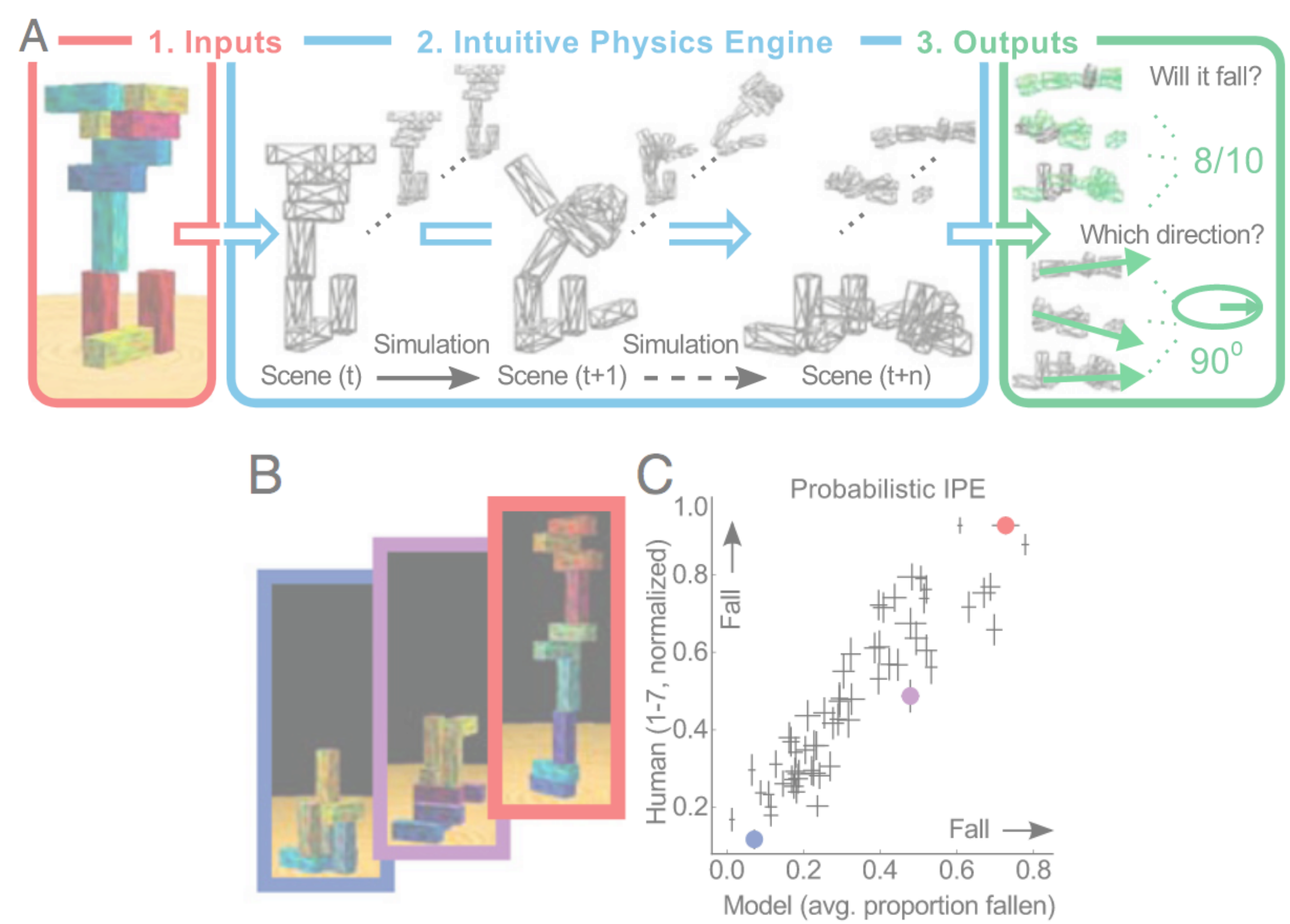


Fronto-Parietal Network

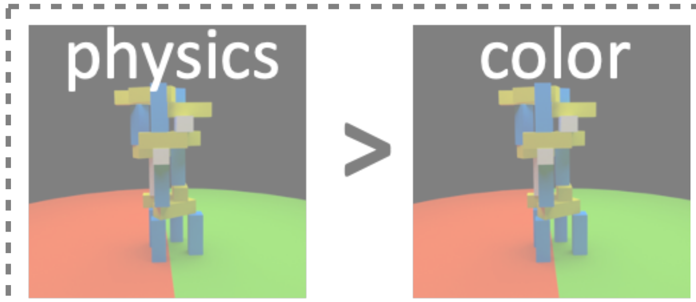
Rishi Rajalingham



Mehrdad Jazayeri



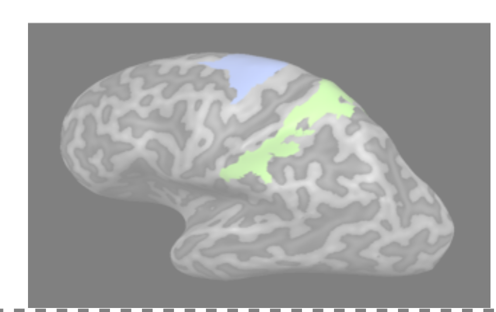
Battaglia, Hamrick, Tenenbaum 2013



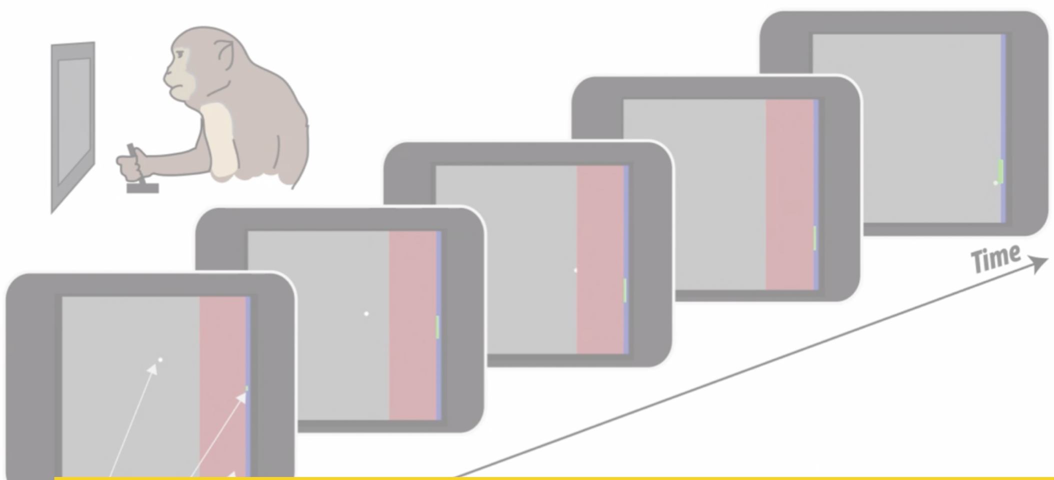
Fischer et al. 2016

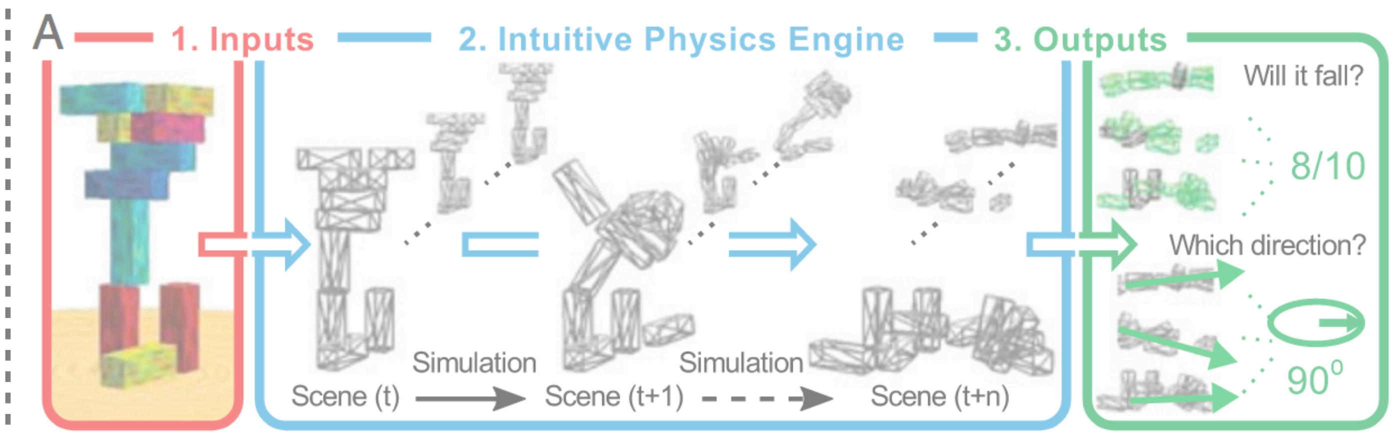
Pramod et al. 2022

Schwettmann et al. 2019



Neural Mechanisms of Mental Simulation





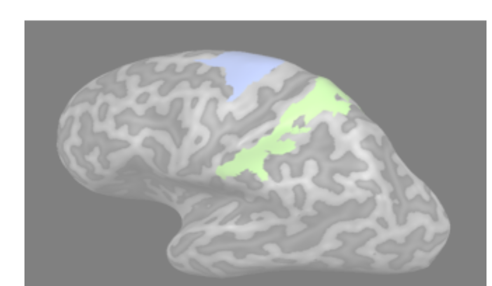
...um 2013

Neurobiological Puzzle: What are the neural mechanisms that enable us to predict the future state of our environment *across* diverse settings?

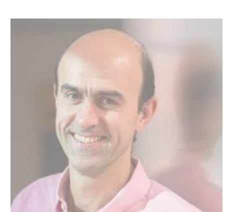
Fischer et al. 2016

Pramod et al. 2022

Schwettmann et al. 2019



Rishi Rajalingham



Mehrdad Jazayeri

Defining Hypotheses

Defining Hypotheses

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

Defining Hypotheses

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

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

Defining Hypotheses

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

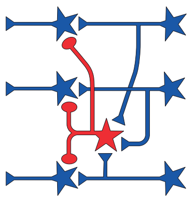
R2 (Behavioral Outputs): Generate physical predictions for each scenario ("behavior").

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

Defining Hypotheses

A = *architecture class*

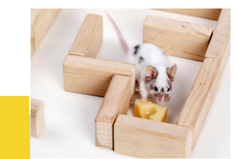
1. *"Circuit"*



"Sensory-Cognitive Networks"

3. *"Ecological niche/behavior"*

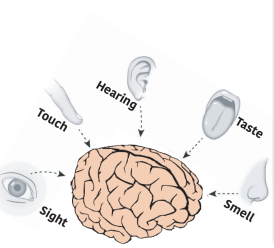
T = *task loss*



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").



2. *"Environment"*

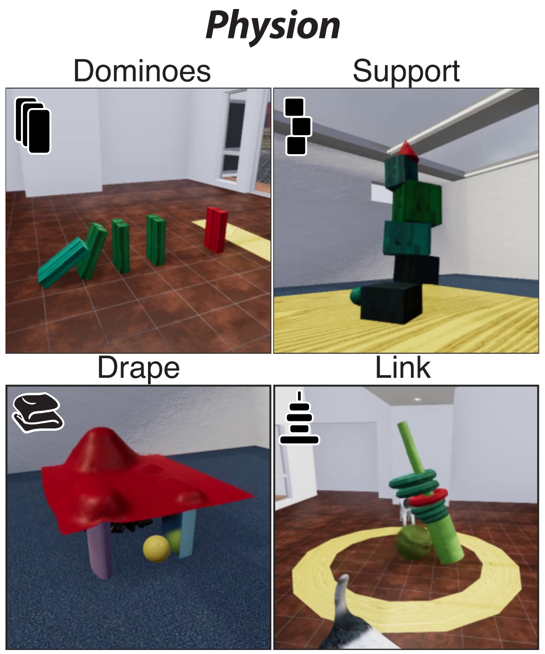
D = *data stream*

Overall Approach

Overall Approach: Training Datasets

(A) *Model Pretraining*

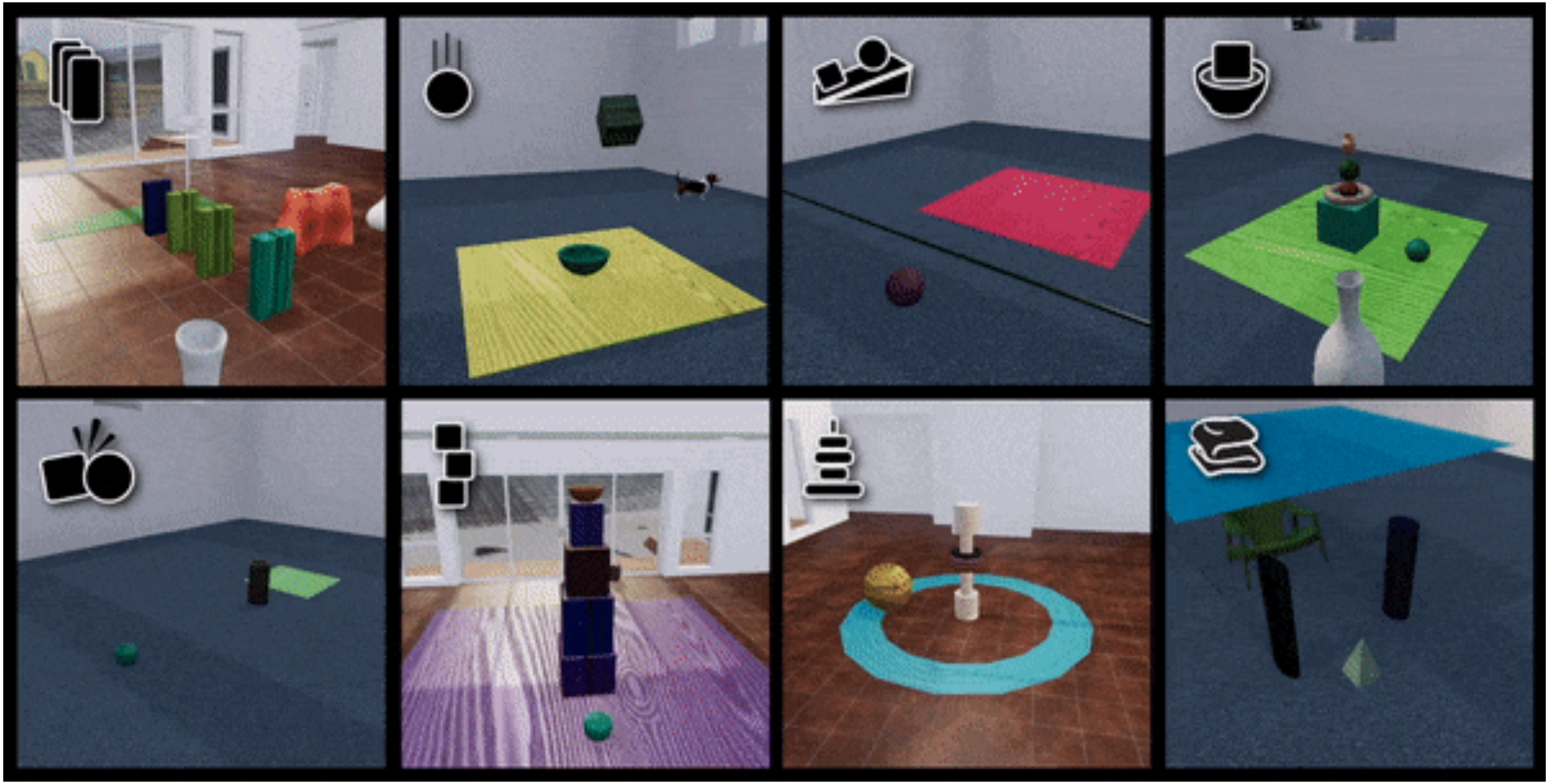
Inputs



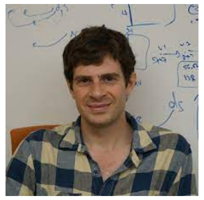
Overall Approach: Training Datasets

Physion/ThreeD World (TDW)

Bear et al. 2021



Focus on everyday physical understanding



Daniel Bear

Joshua Tenenbaum

Daniel Yamins

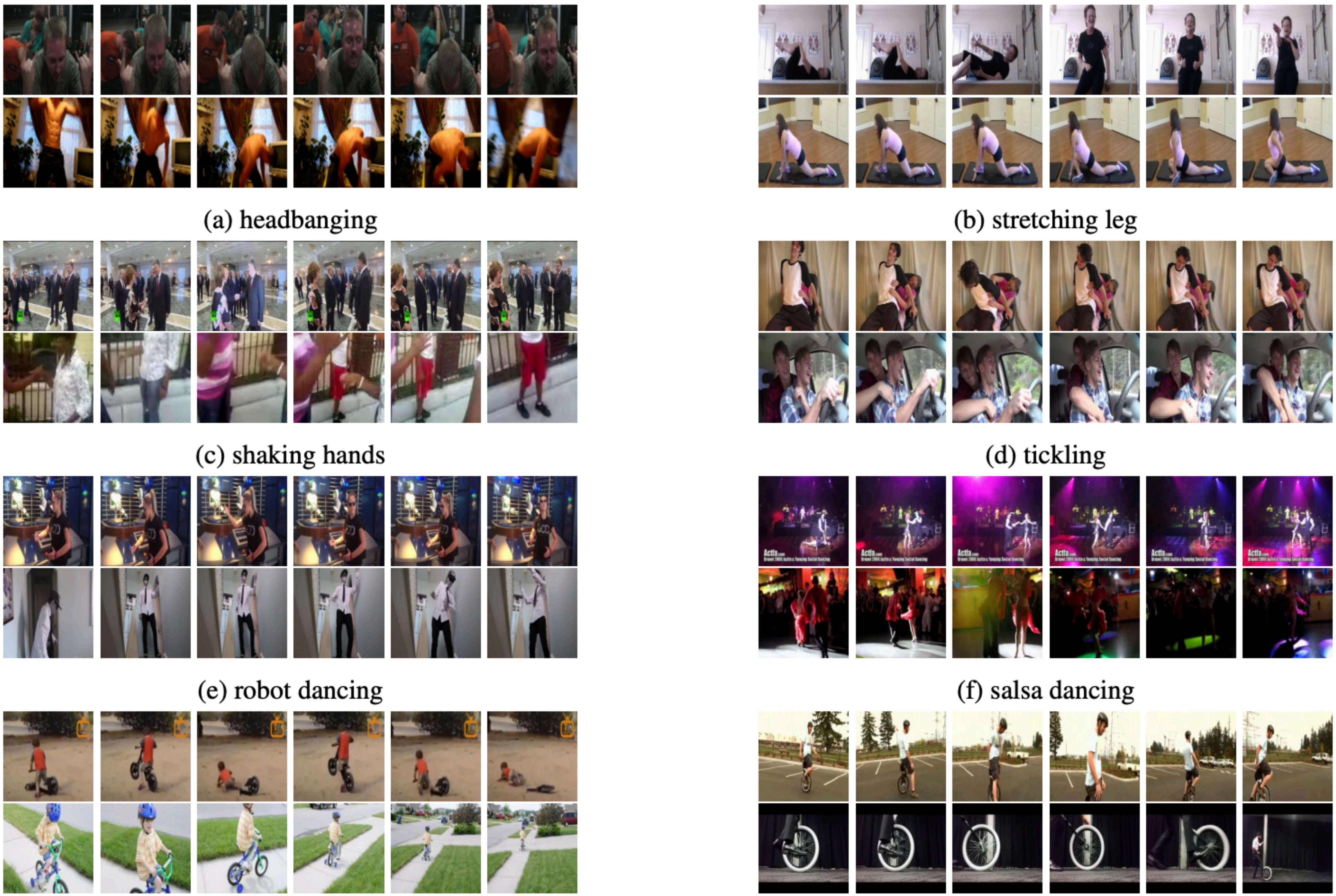
Judith Fan

Overall Approach: Training Datasets

Kinetics 700

Focus on human actions

Carreira et al. 2019, Kay et al. 2017



(a) headbanging

(b) stretching leg

(c) shaking hands

(d) tickling

(e) robot dancing

(f) salsa dancing

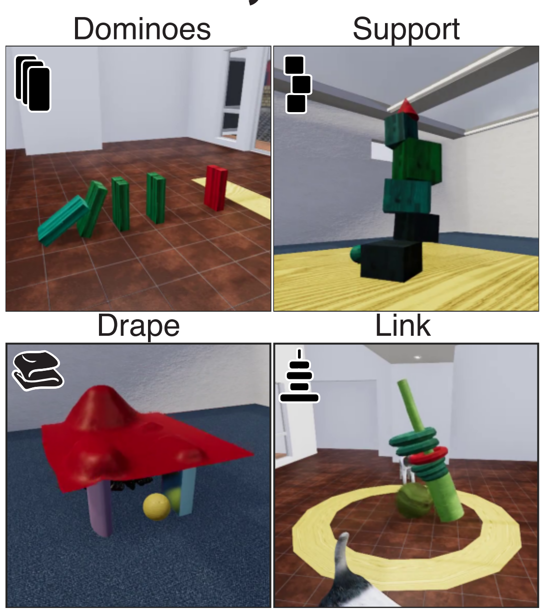
Overall Approach: Sensory-Cognitive Hypotheses

(A) *Model Pretraining*

Inputs

Sensory-Cognitive Hypothesis Classes

Physion

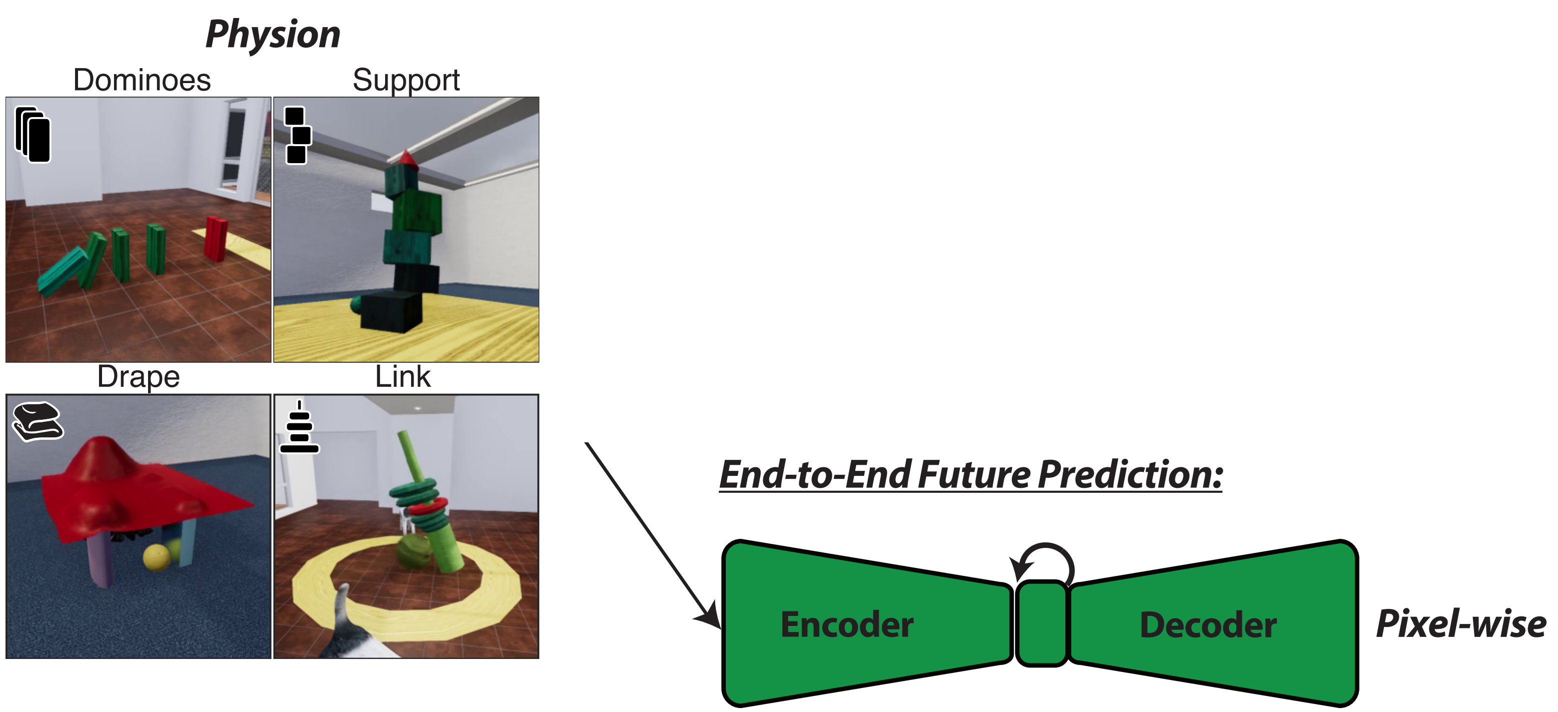


Overall Approach: Pixel-wise Future Prediction

(A) ***Model Pretraining***

Inputs

Sensory-Cognitive Hypothesis Classes



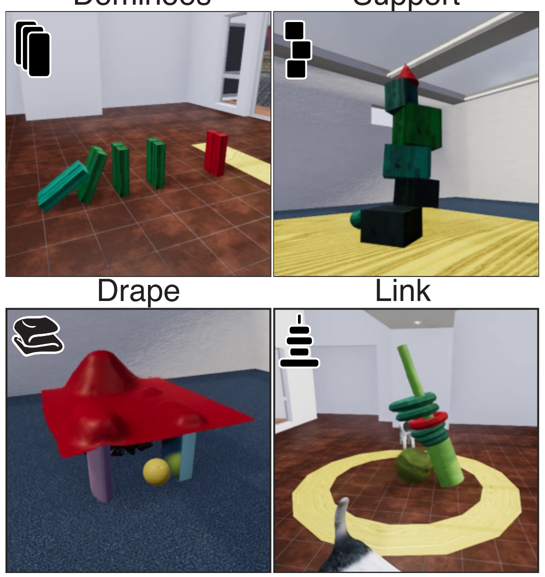
Overall Approach: Pixel-wise Future Prediction

(A) *Model Pretraining*

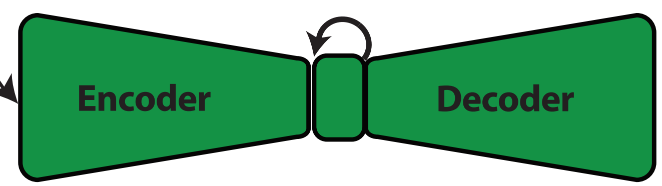
Inputs

Sensory-Cognitive Hypothesis Classes

Physion

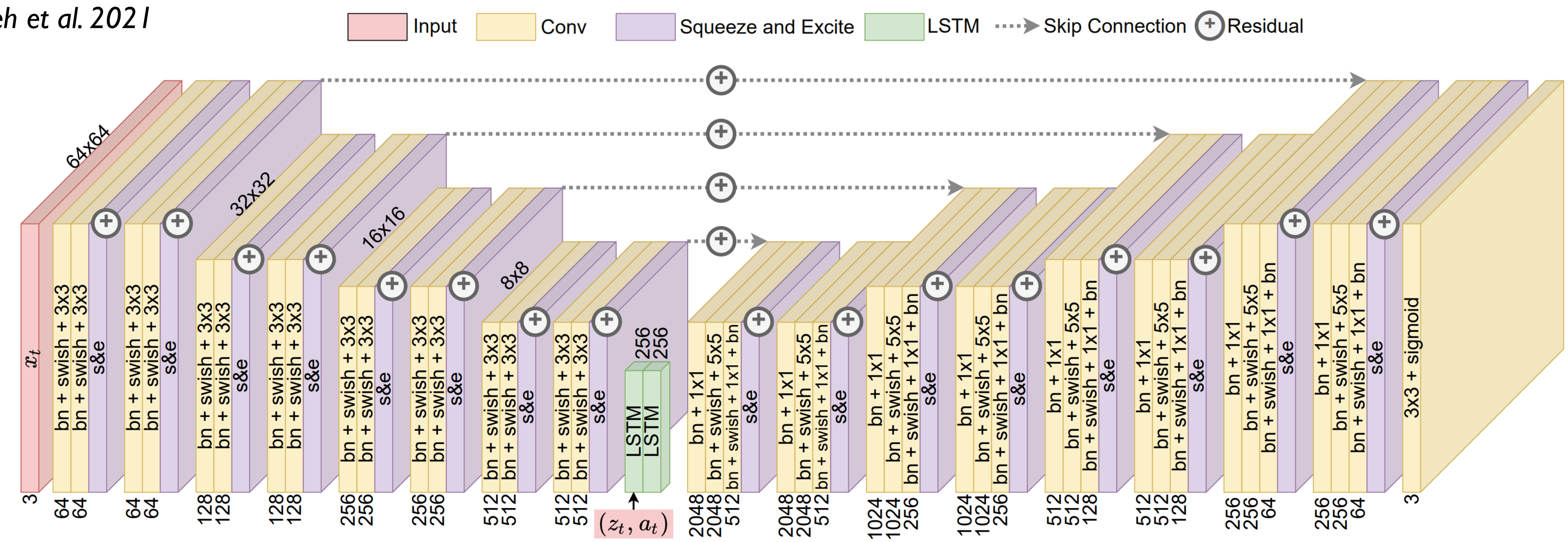


End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021



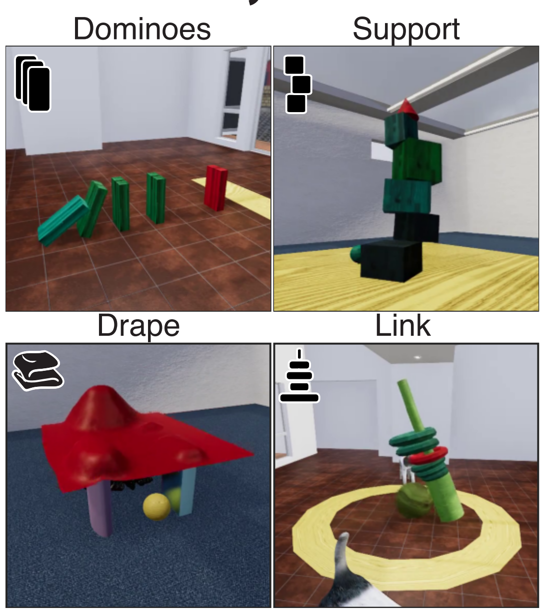
Overall Approach: Pixel-wise Future Prediction

(A) *Model Pretraining*

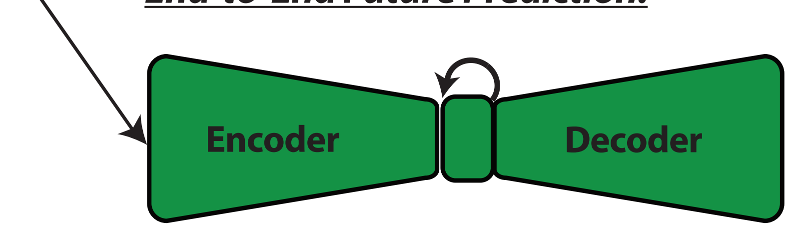
Inputs

Sensory-Cognitive Hypothesis Classes

Physion

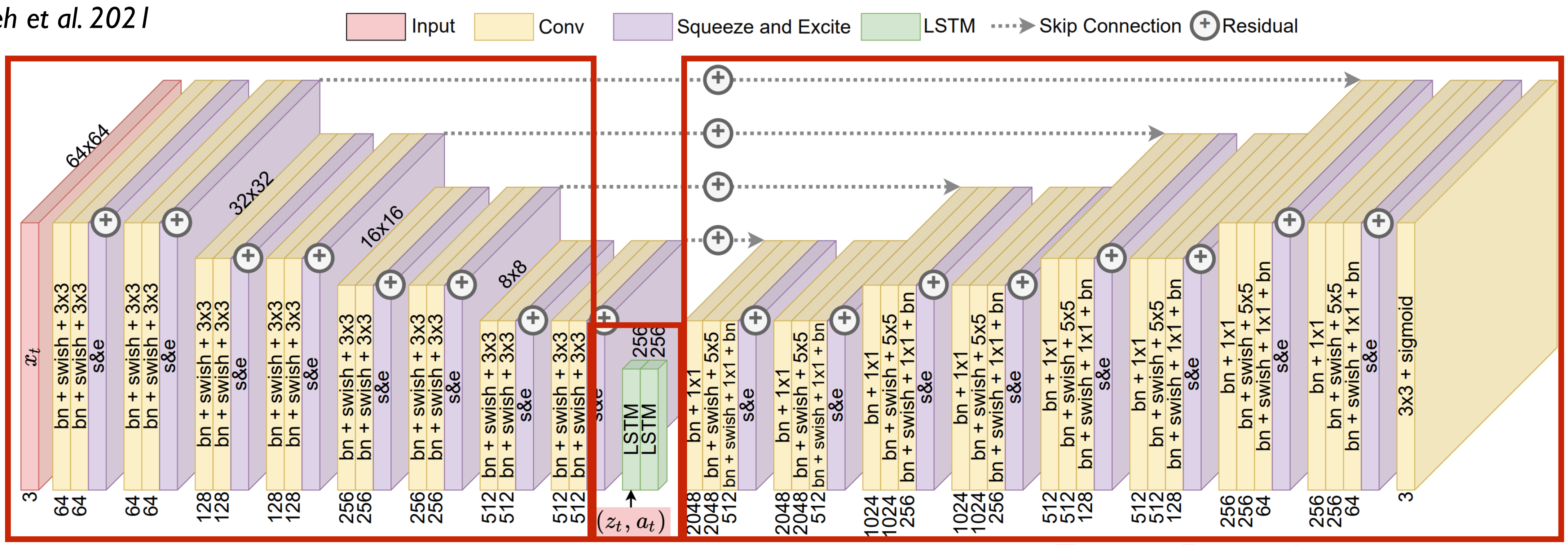


End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021



Visual Encoder ("Sensory")

Dynamics Predictor ("Cognitive")

Frame Decoder ("Objective/Behavior")

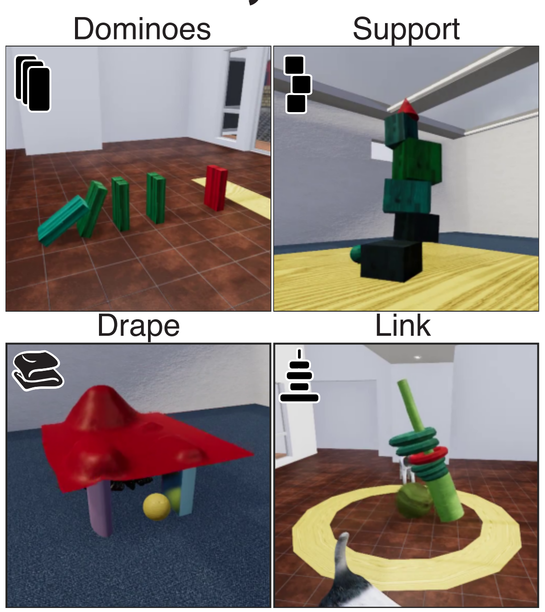
Overall Approach: Pixel-wise Future Prediction

(A) ***Model Pretraining***

Inputs

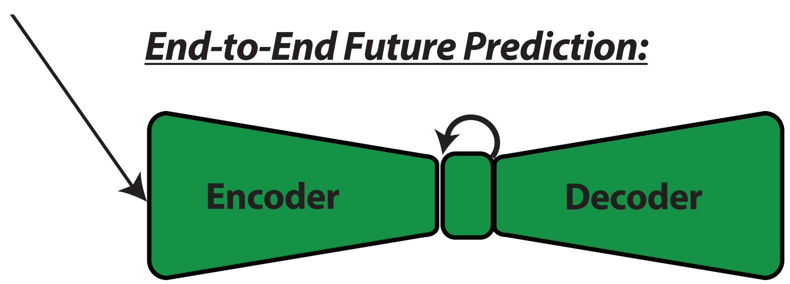
Sensory-Cognitive Hypothesis Classes

Physion



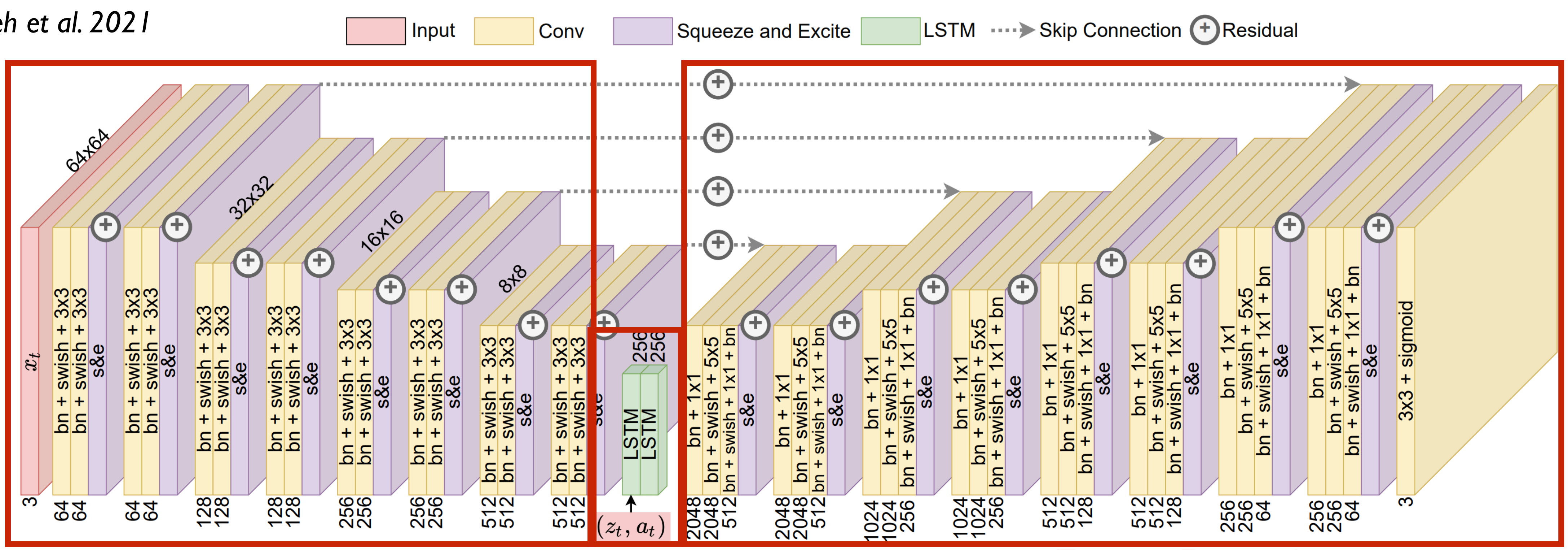
Self-supervised on future frame prediction — can be readily applied to large-scale, real-world video datasets

End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021



Visual Encoder ("Sensory")

Dynamics Predictor ("Cognitive")

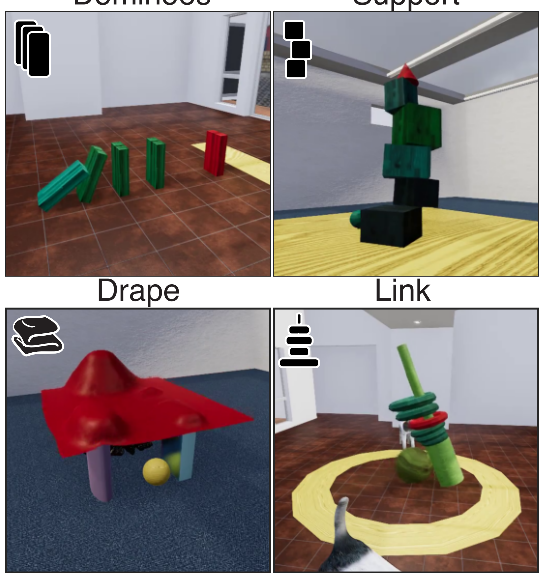
Frame Decoder ("Objective/Behavior")

Overall Approach: Pixel-wise Future Prediction

(A) *Model Pretraining*

Inputs

Physion

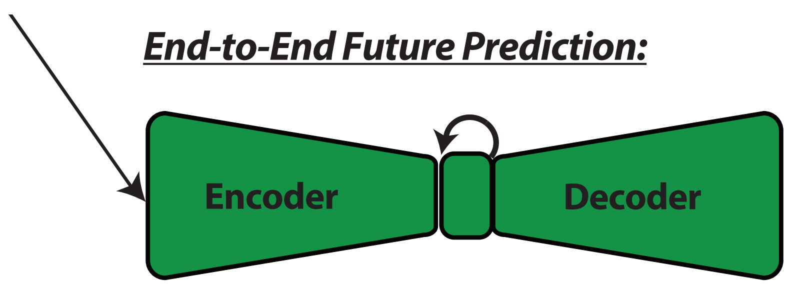


Sensory-Cognitive Hypothesis Classes

Self-supervised on future frame prediction — can be readily applied to large-scale, real-world video datasets

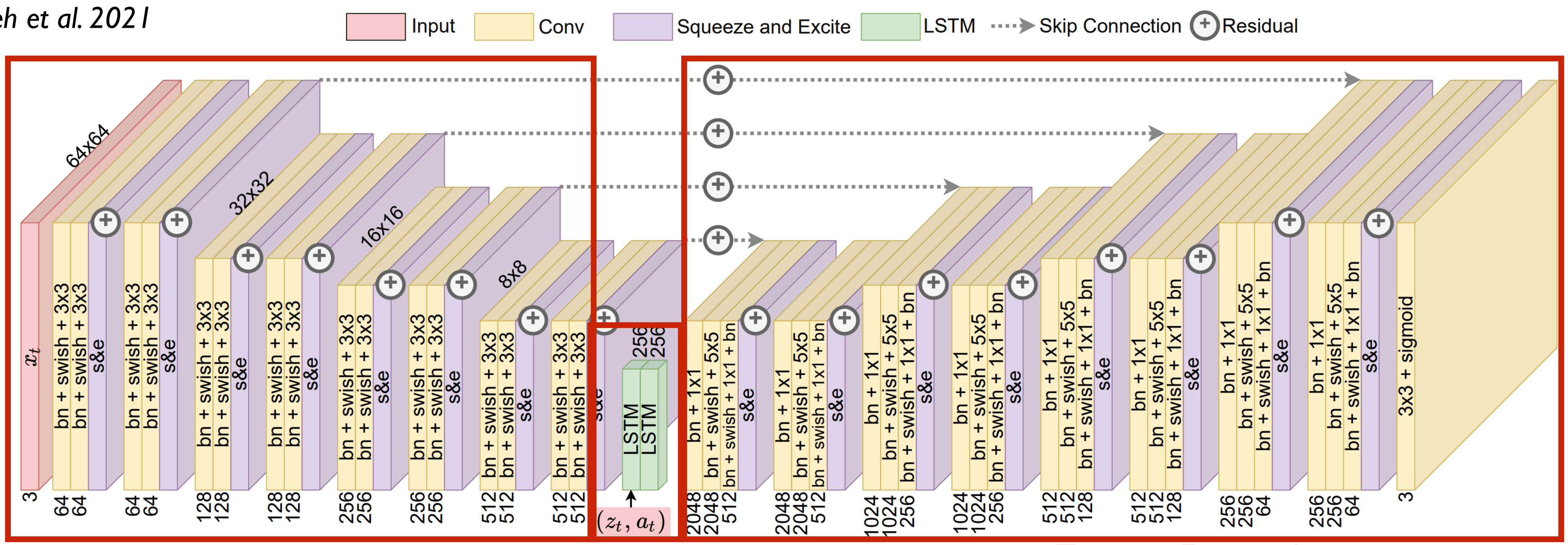
Predicts the future at the resolution of the sensory input (very detailed!)

End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021



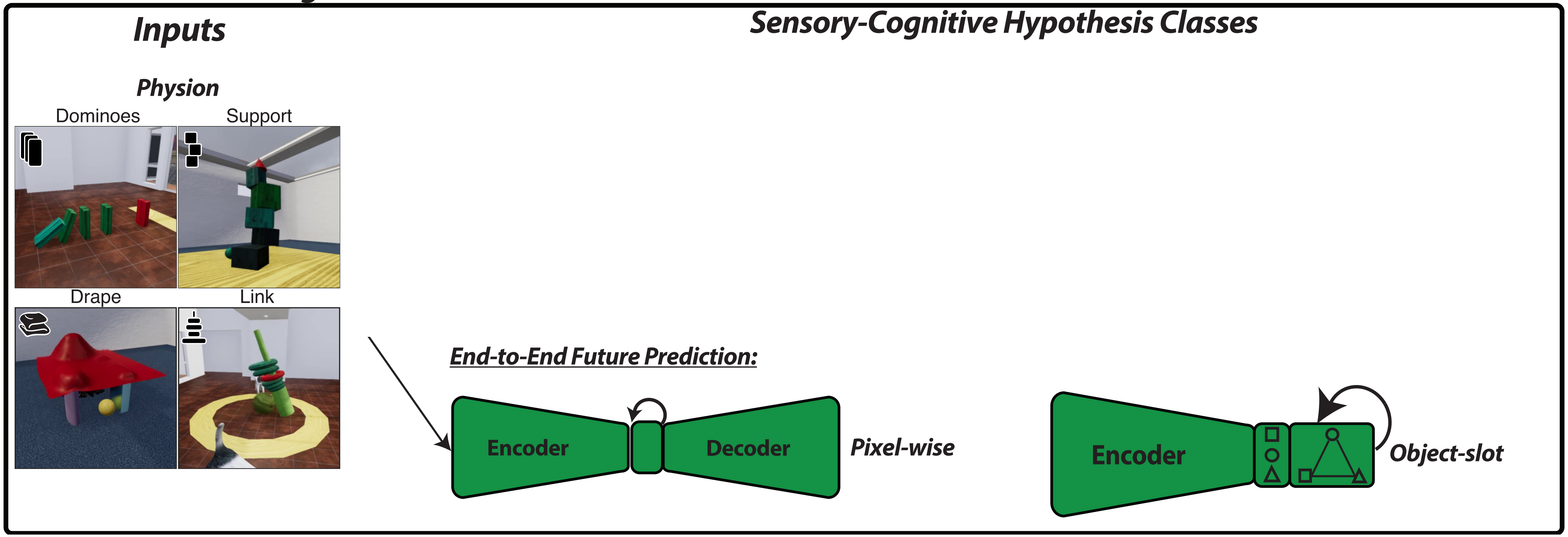
Visual Encoder (“Sensory”)

Dynamics Predictor (“Cognitive”)

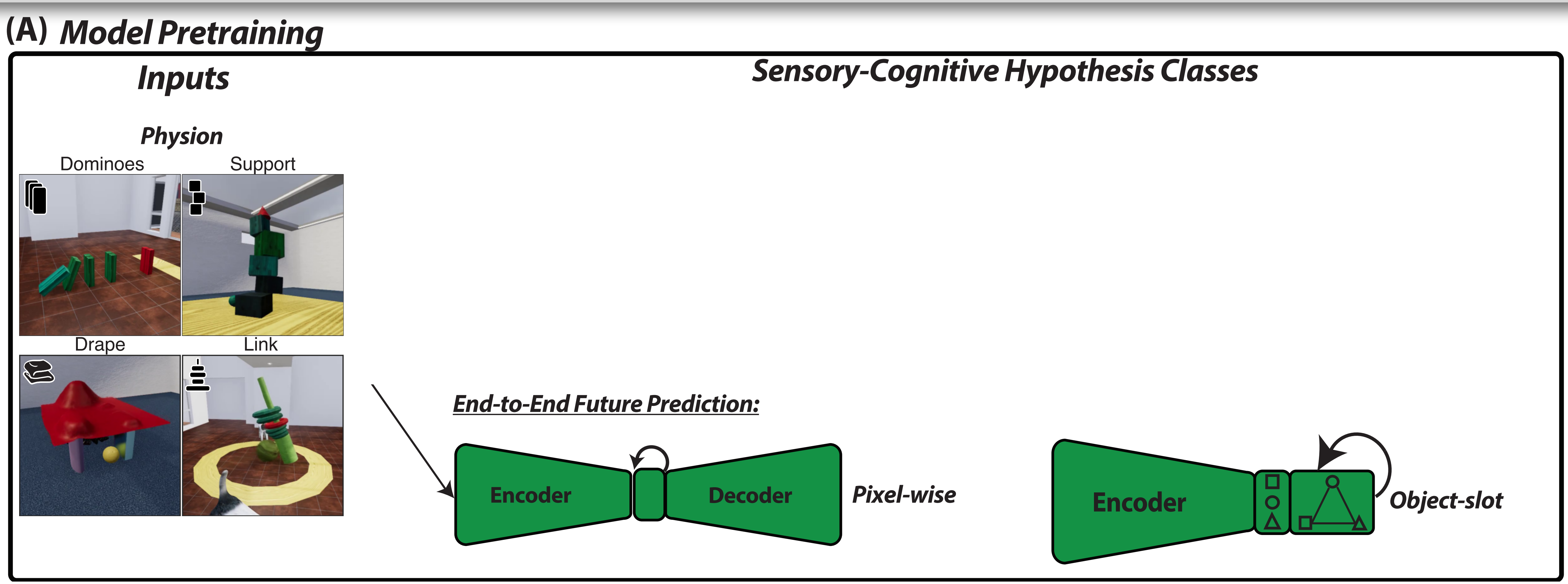
Frame Decoder (“Objective/Behavior”)

Overall Approach: Structured World Models

(A) ***Model Pretraining***

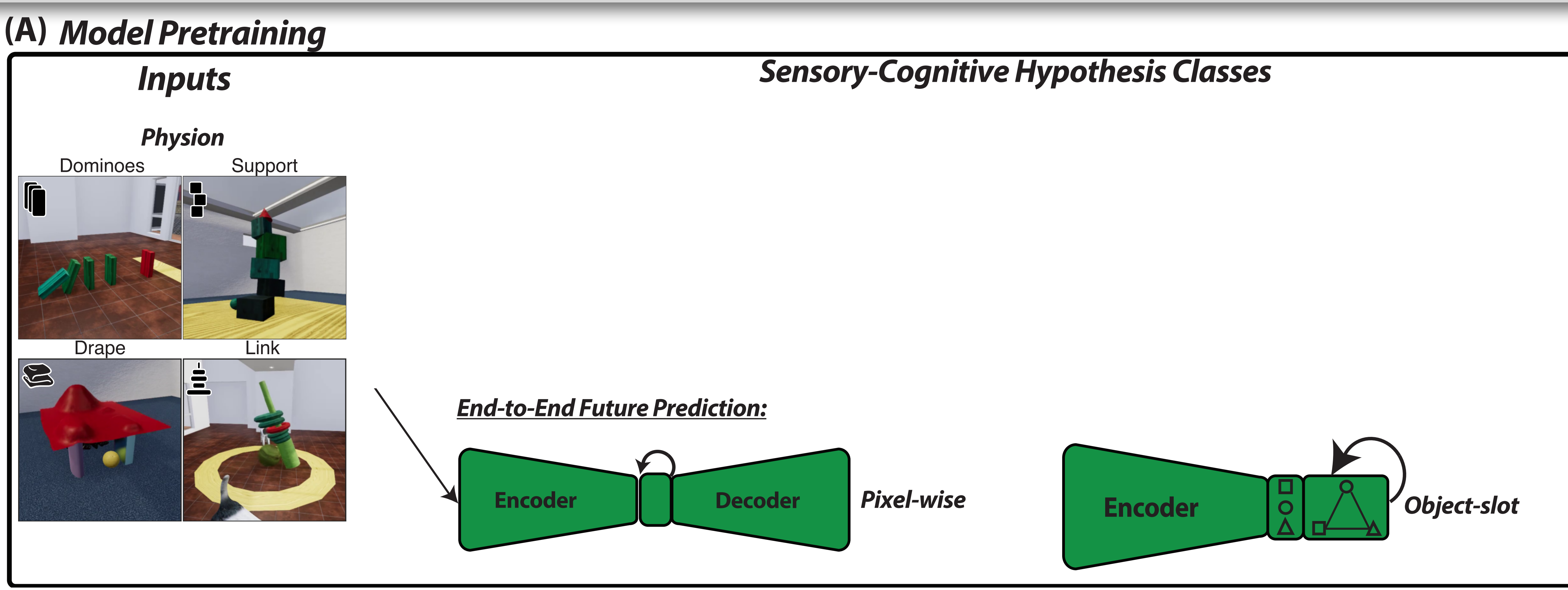


Overall Approach: Structured World Models

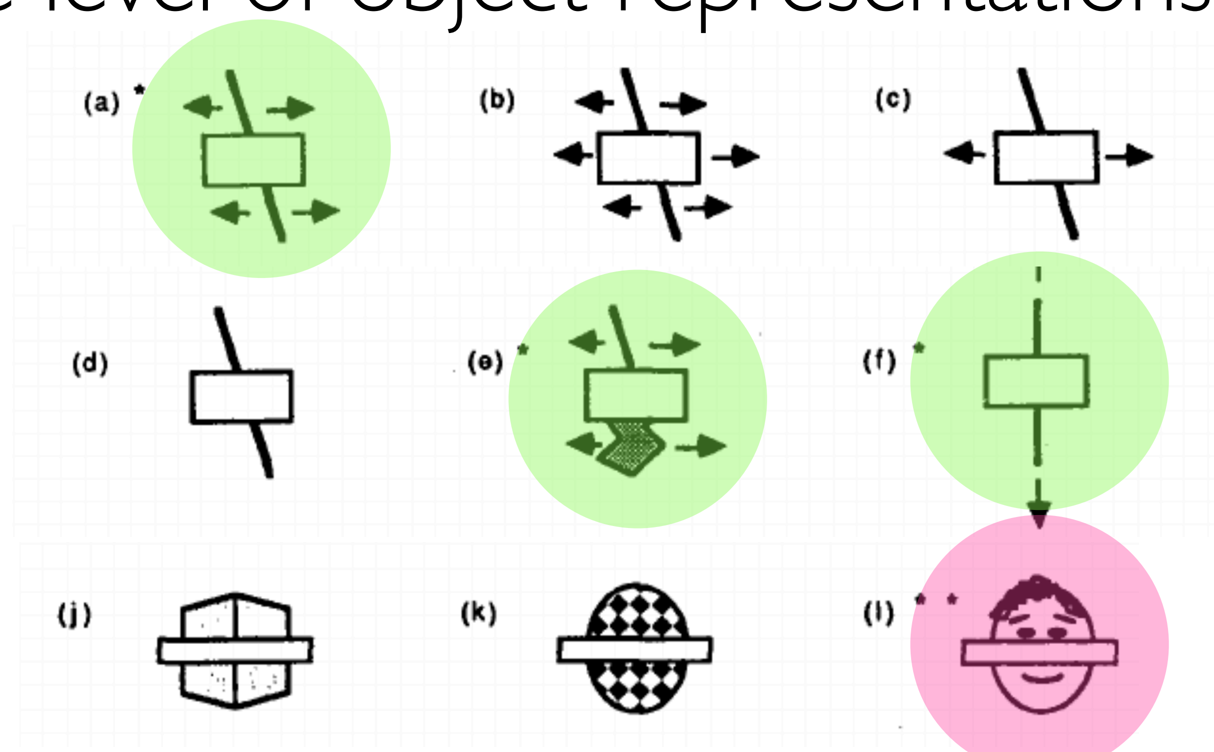


Predicts at the level of object representations and their relations

Overall Approach: Structured World Models



Predicts at the level of object representations and their relations

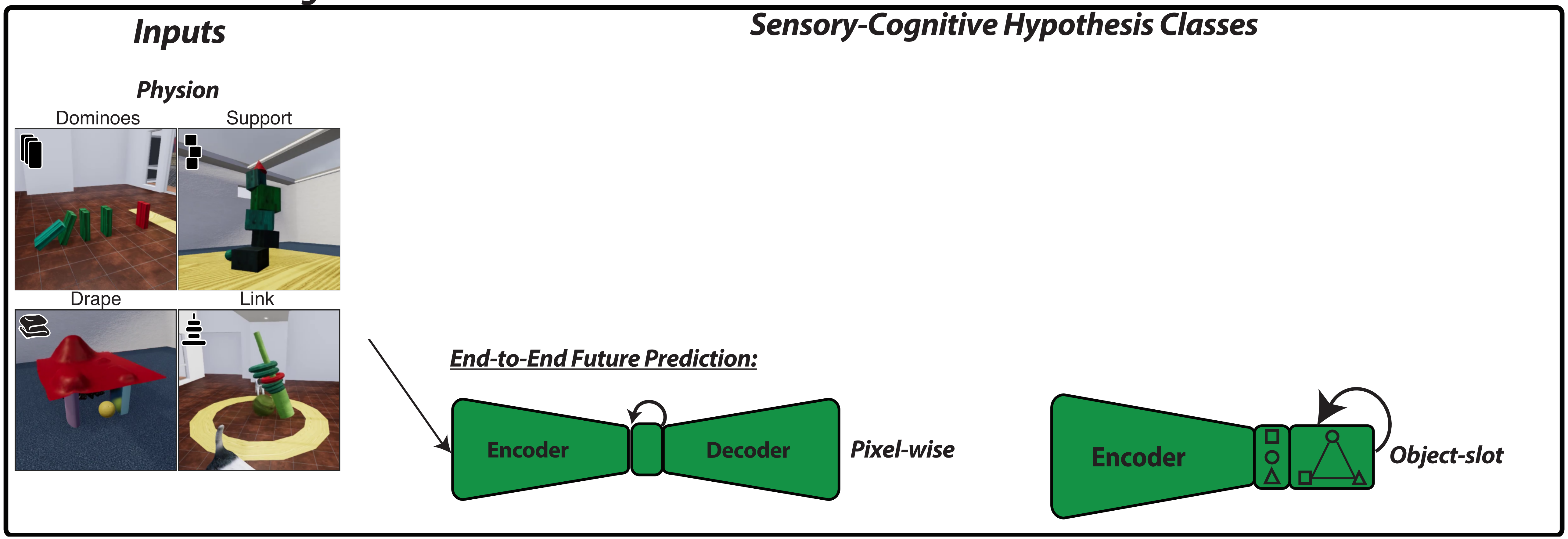


Principles of Object Perception Elizabeth Spelke, 1990

Elizabeth Spelke

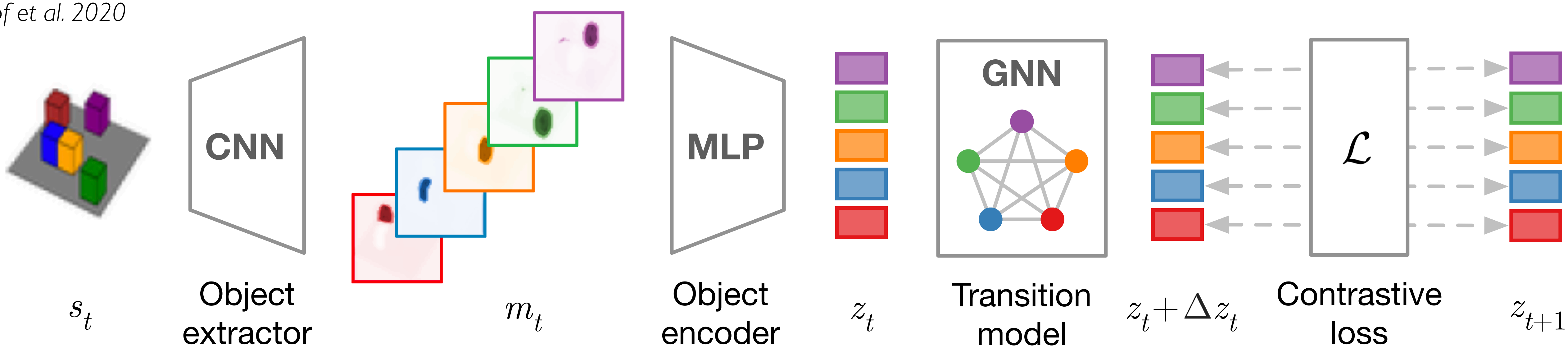
Overall Approach: Structured World Models

(A) *Model Pretraining*



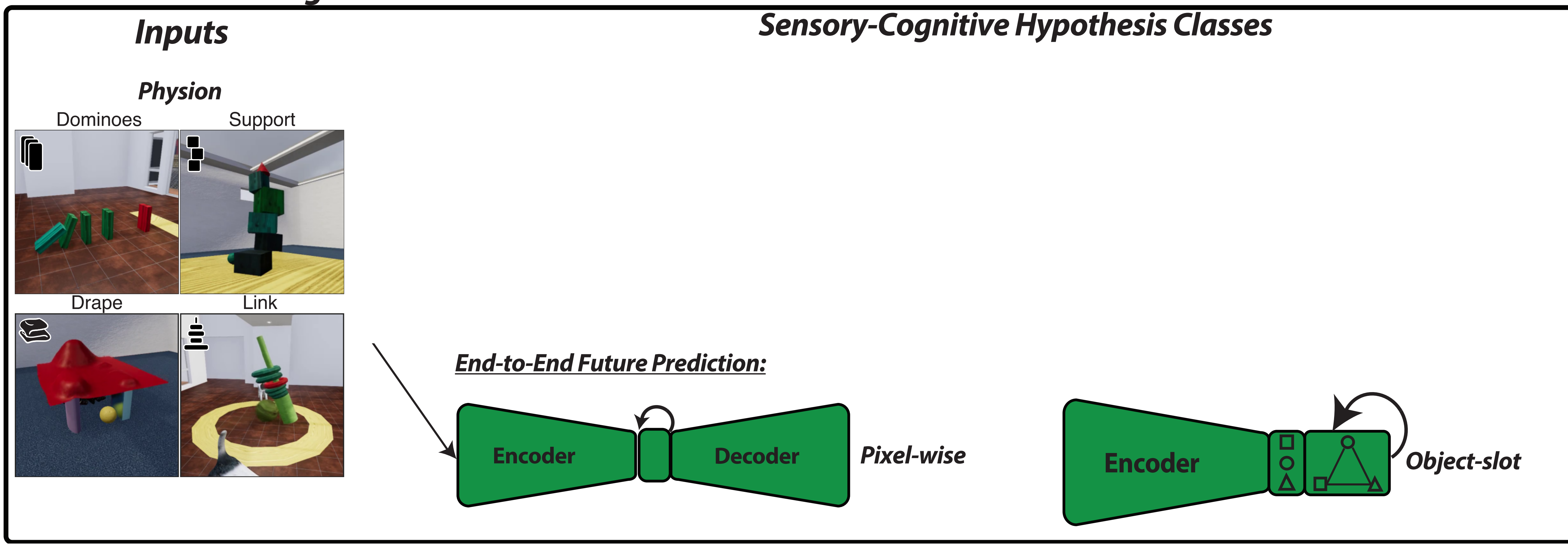
Predicts at the level of object representations and their relations

Kipf et al. 2020

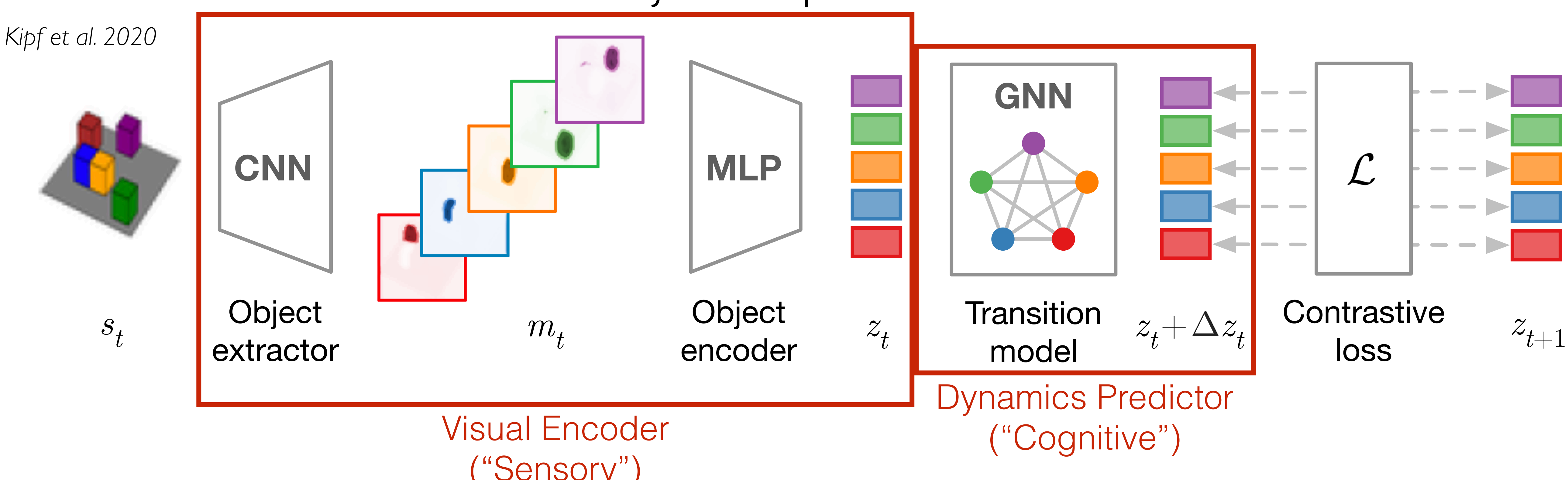


Overall Approach: Structured World Models

(A) *Model Pretraining*

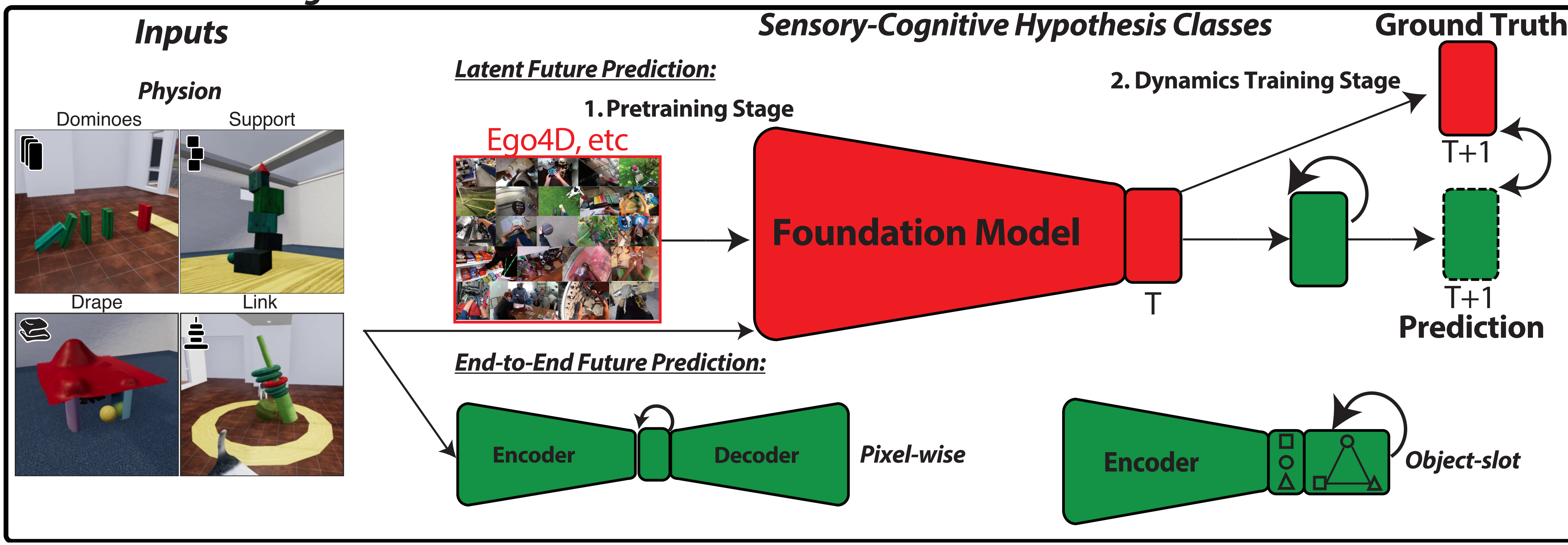


Predicts at the level of object representations and their relations

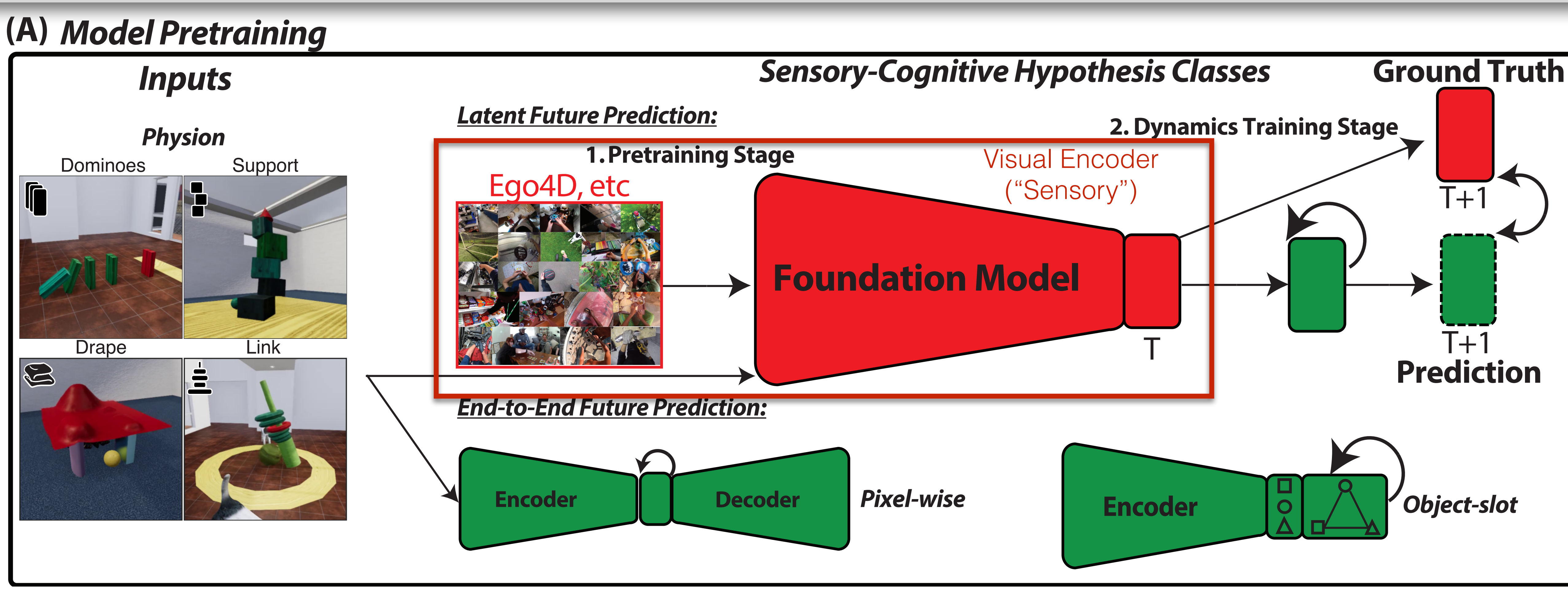


Overall Approach: Foundation Models

(A) ***Model Pretraining***

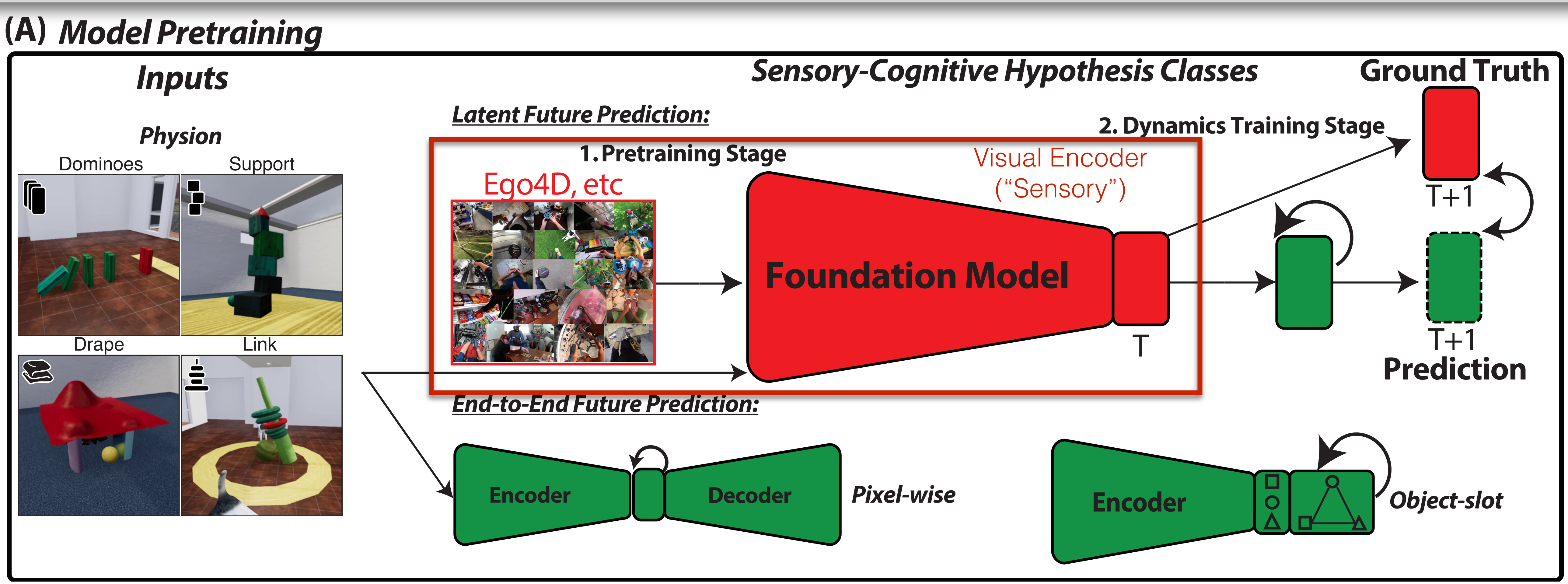


Overall Approach: Foundation Models



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

Overall Approach: Foundation Models

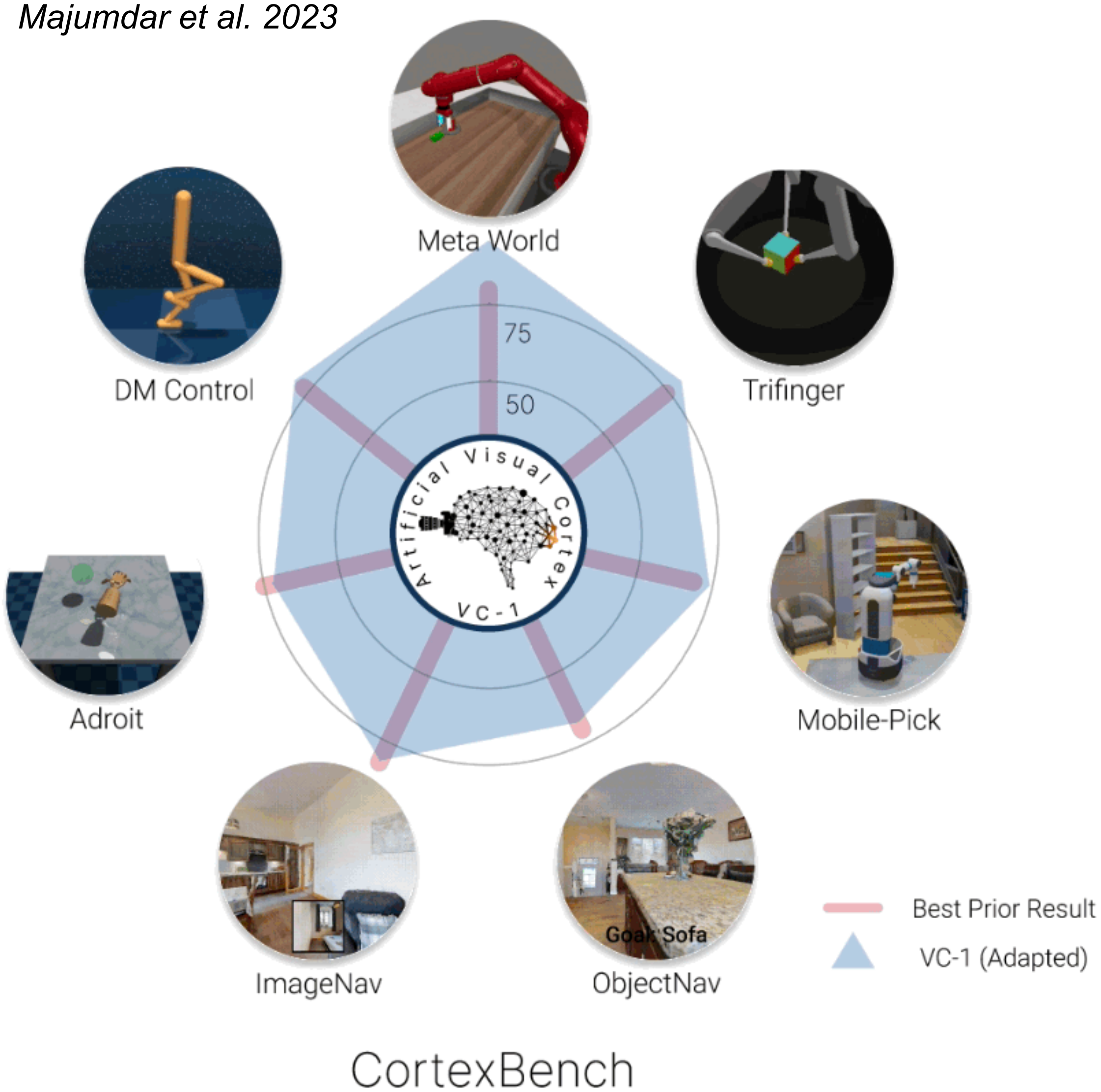


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

Emphasis on *reusability*!

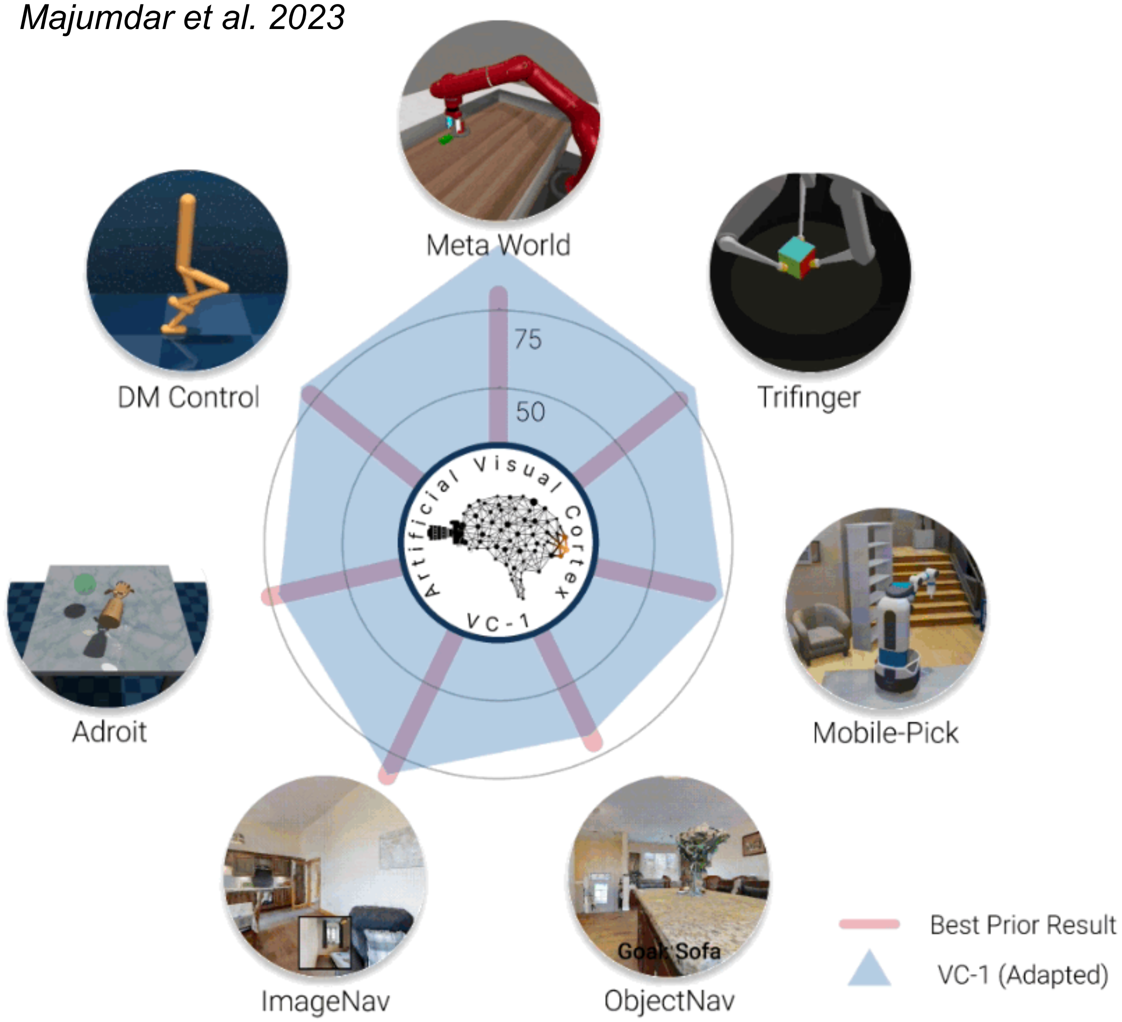
Overall Approach: Foundation Models

Majumdar et al. 2023



Overall Approach: Foundation Models

Majumdar et al. 2023



CortexBench

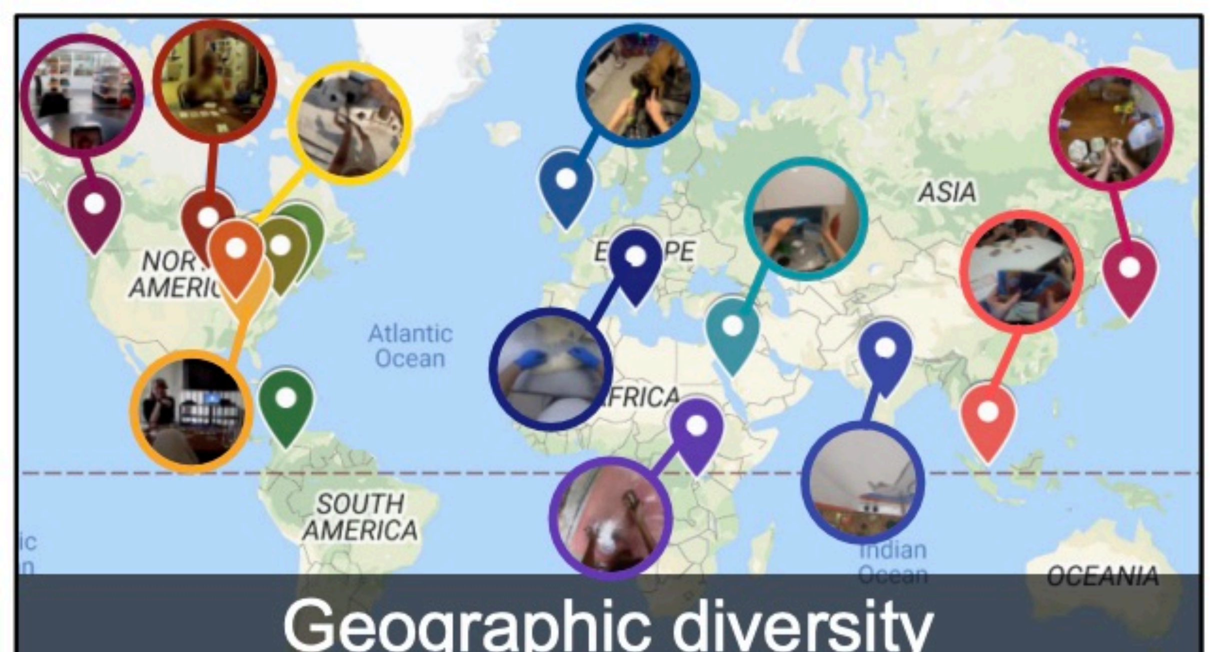
Ego4D: everyday activity around the world

Ego4D: A massive-scale egocentric dataset

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

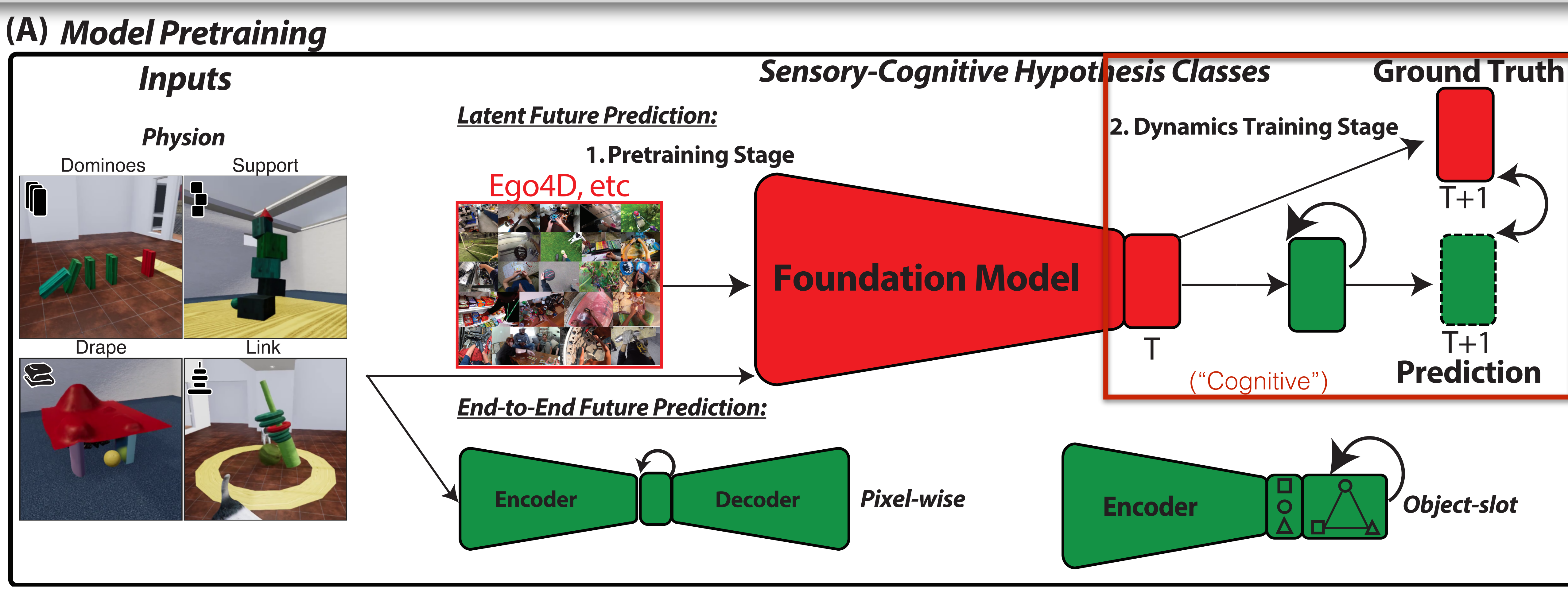
931 participants from 74 worldwide locations

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



Grauman et al. 2022

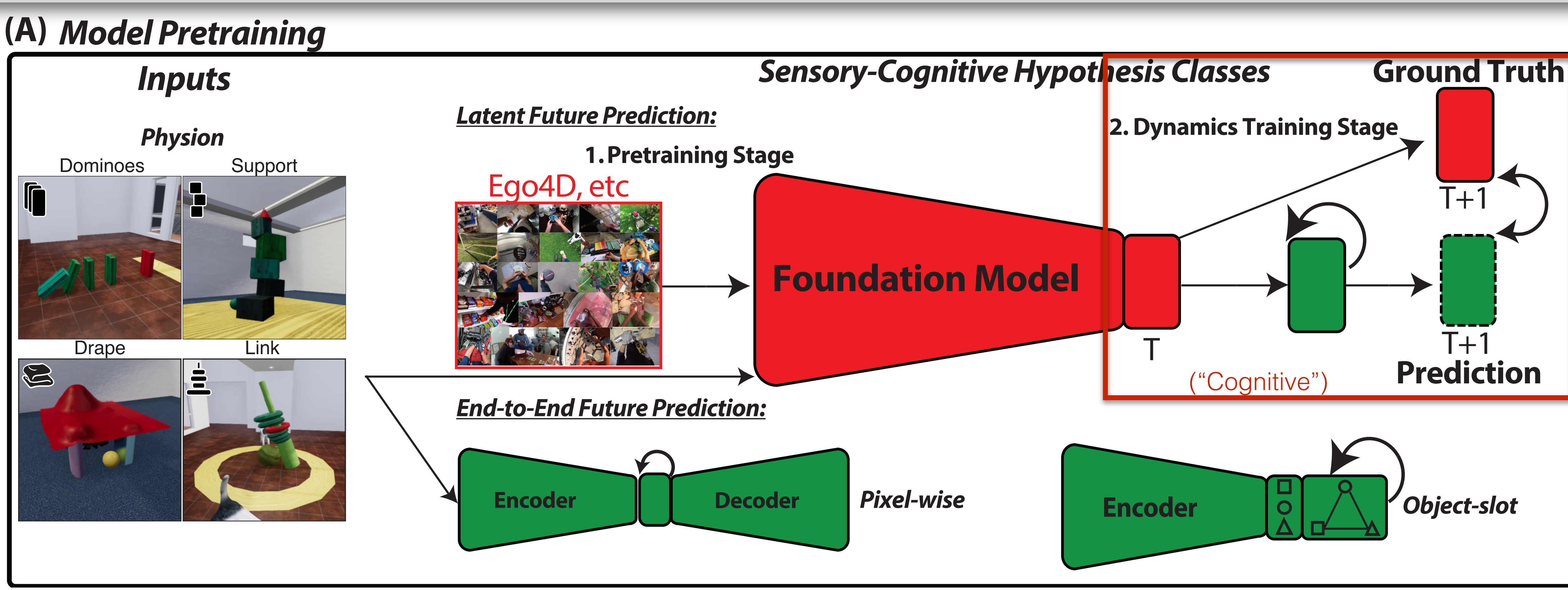
Overall Approach: Foundation Models



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

Emphasis on *reusability*!

Overall Approach: Foundation Models + Dynamics

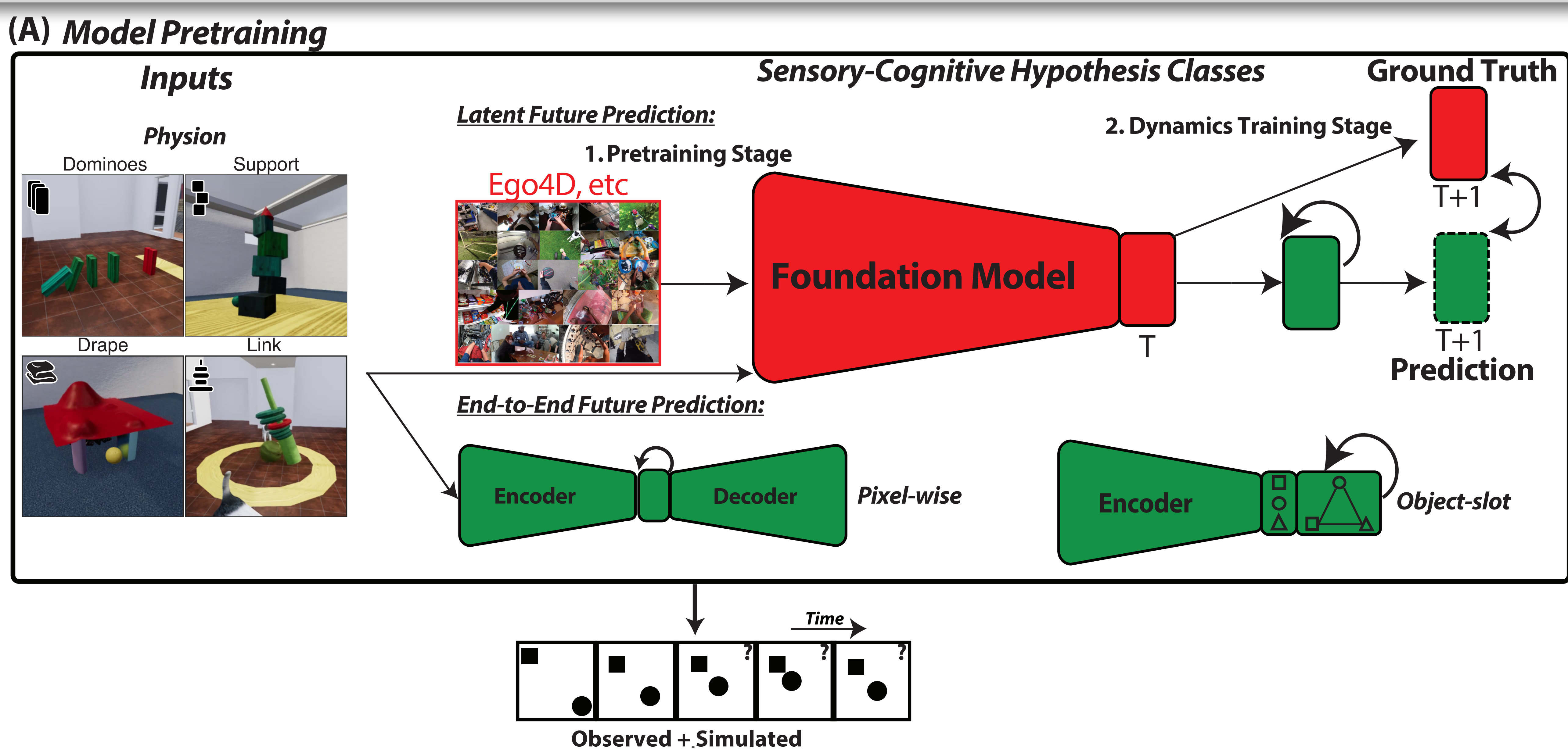


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

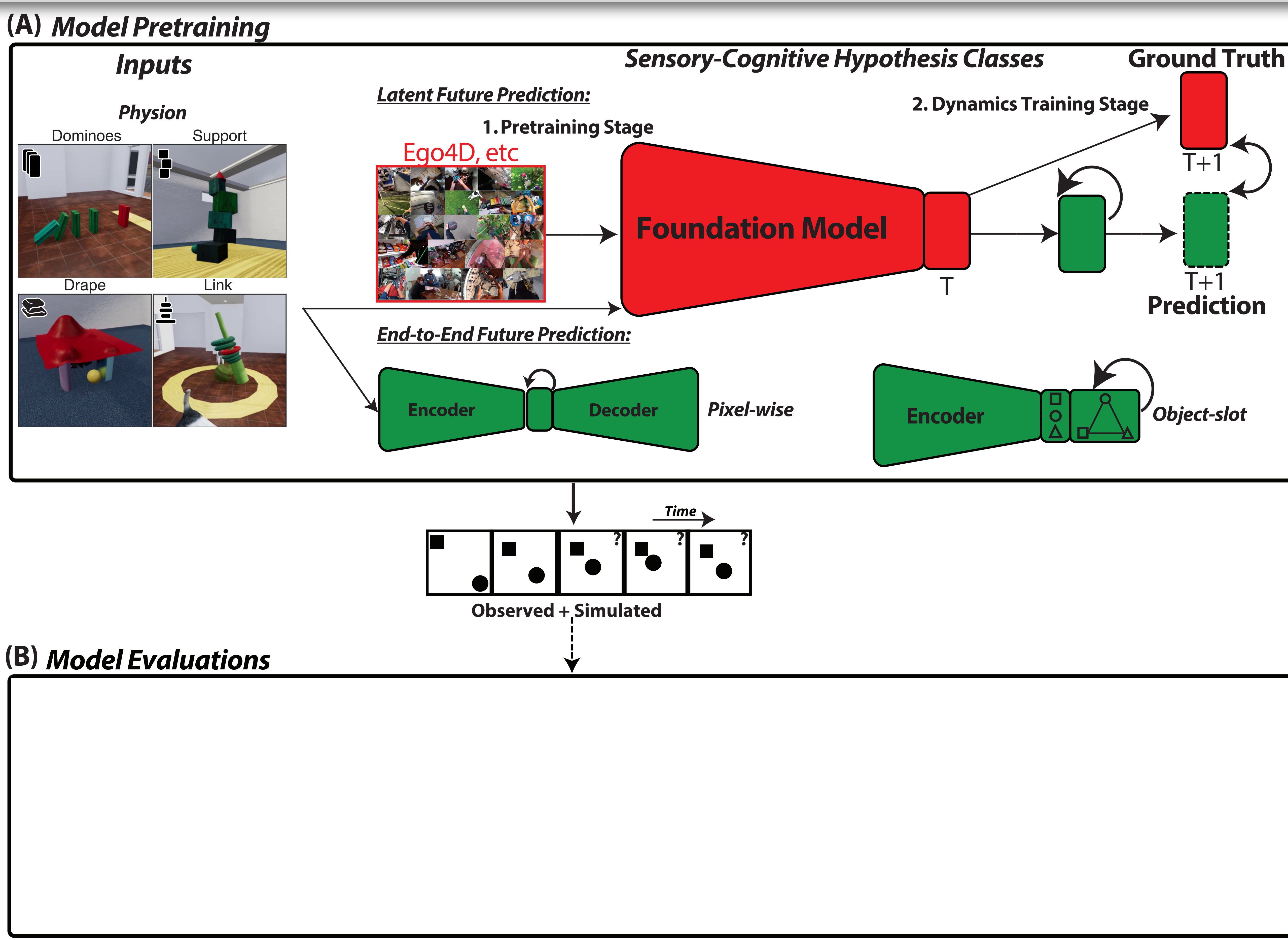
Emphasis on *reusability*!

Leverage these dynamics to do explicit physical simulation

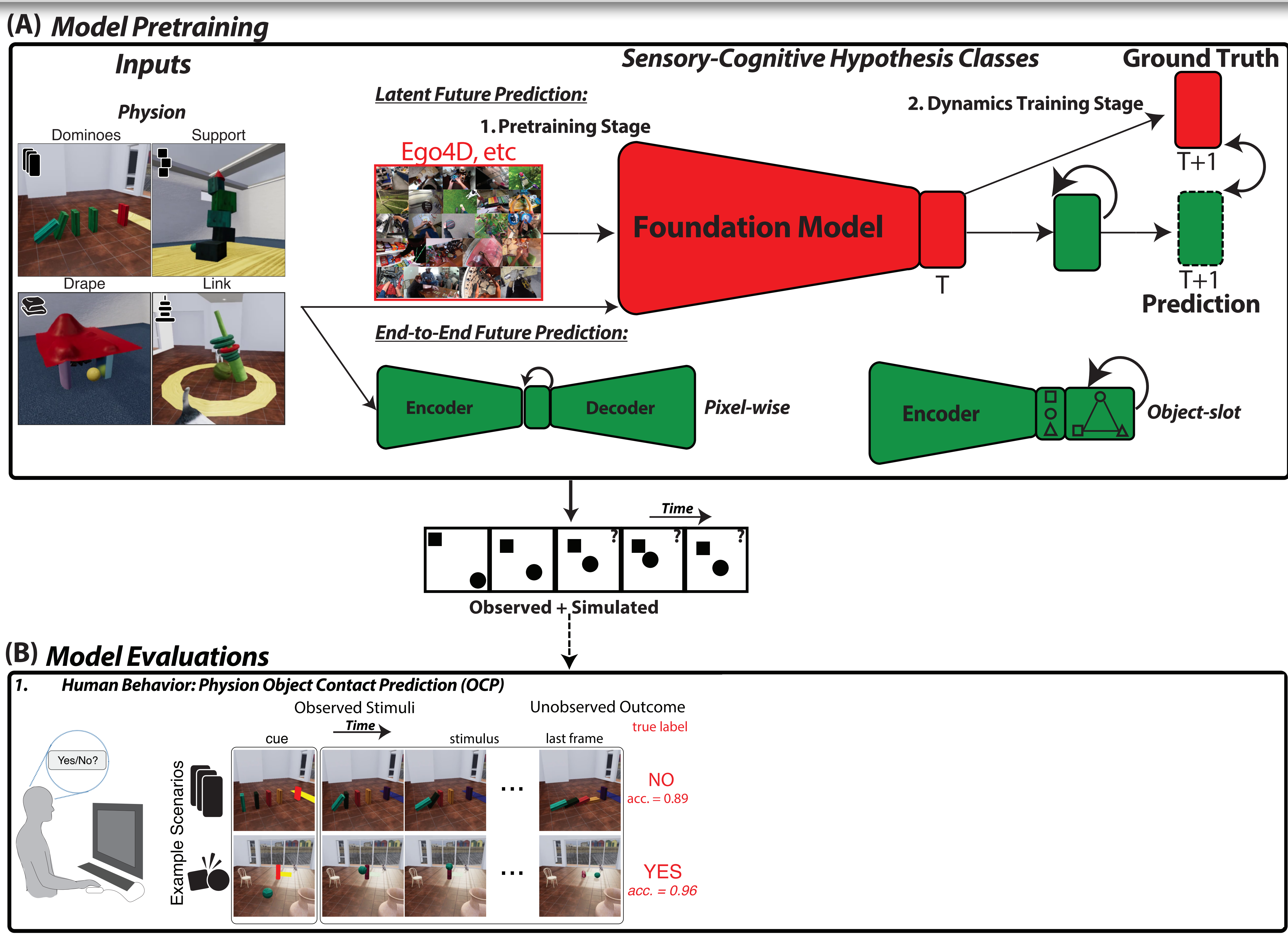
Overall Approach



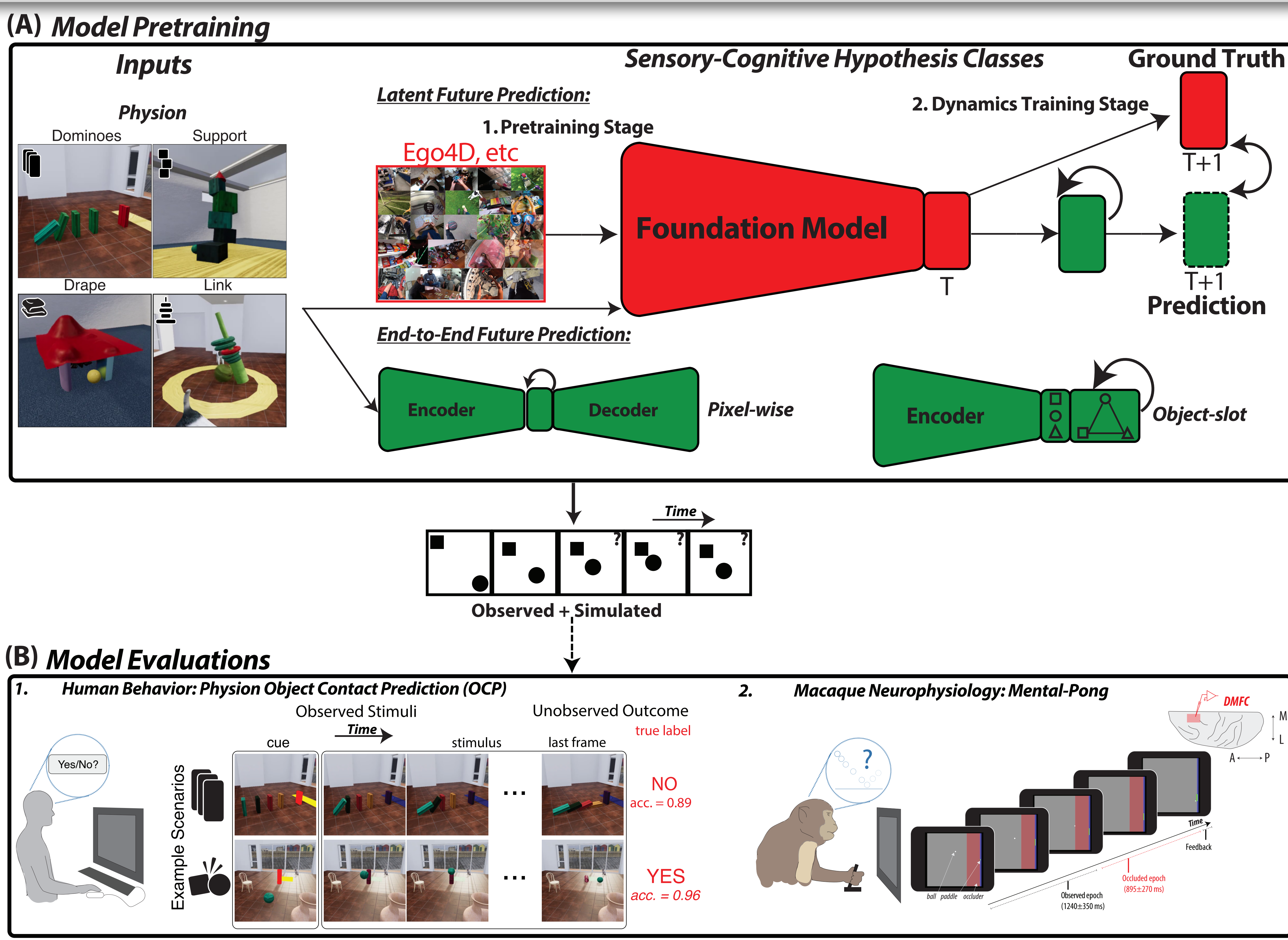
Overall Approach: Model Evaluations



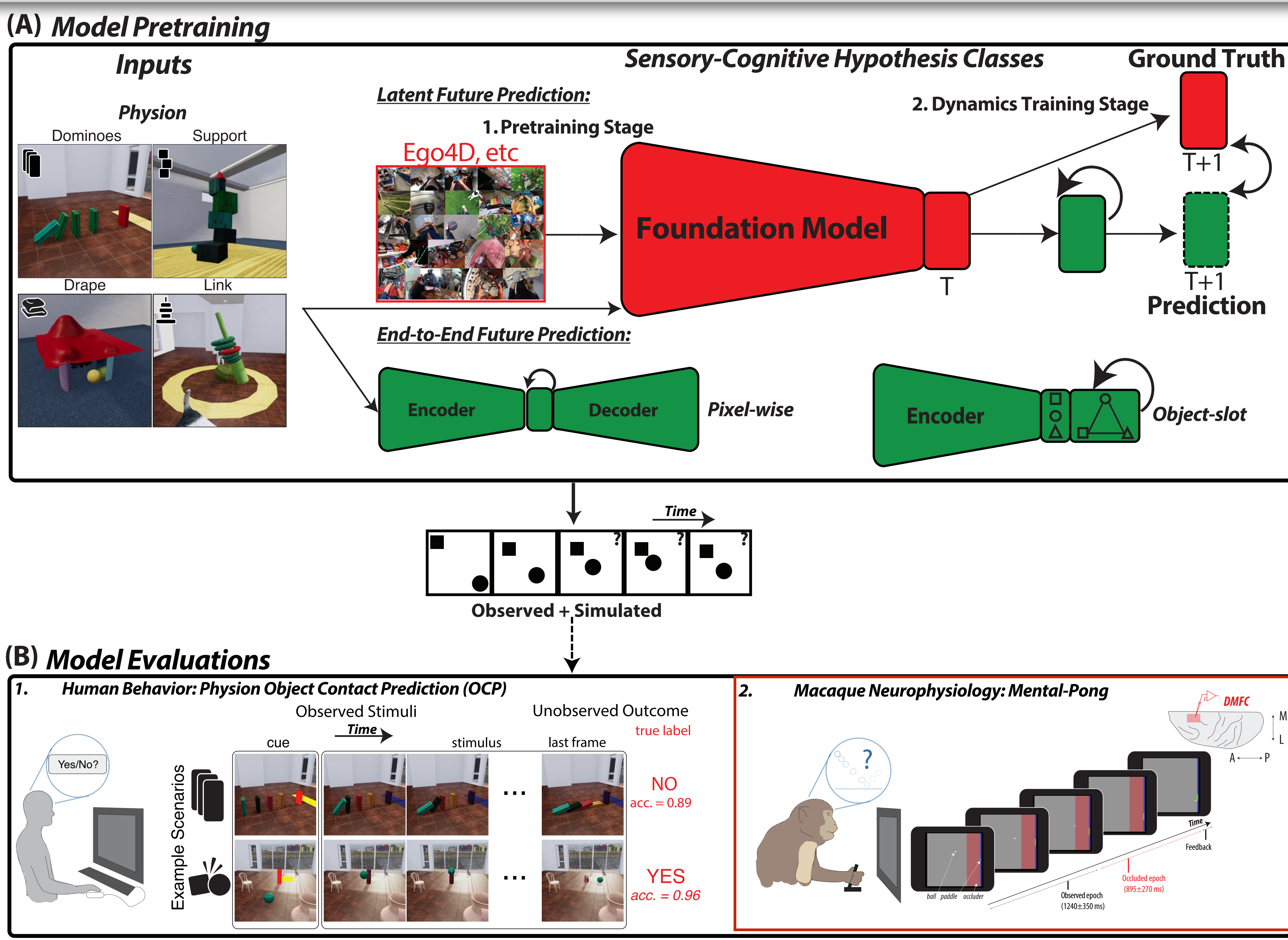
Overall Approach: Model Evaluations (Human Behavior)



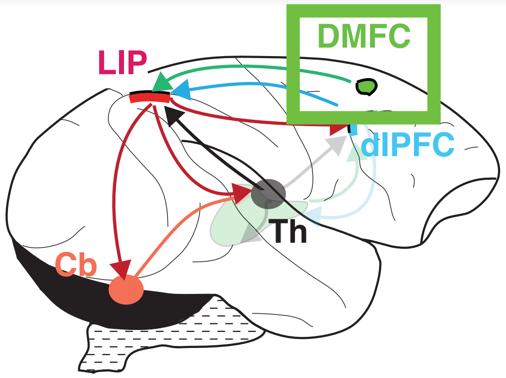
Overall Approach: Model Evaluations (Macaque Physiology)



Model Evaluations: Macaque Neurophysiology



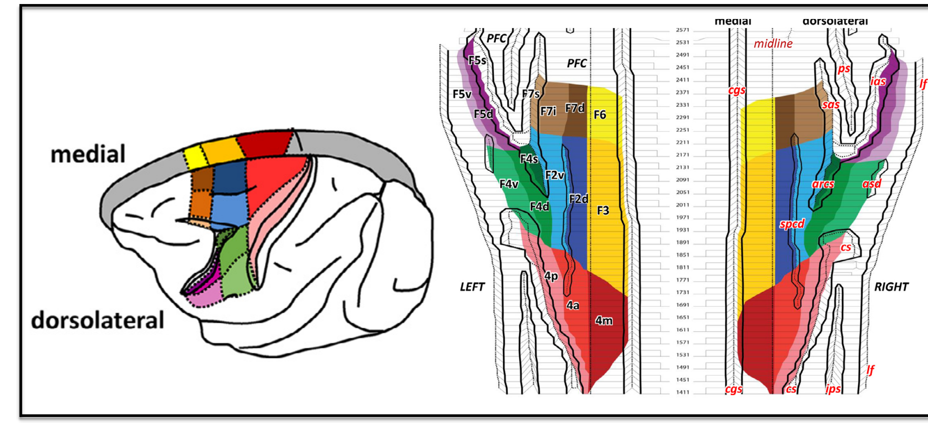
Model Evaluations: Macaque Neurophysiology



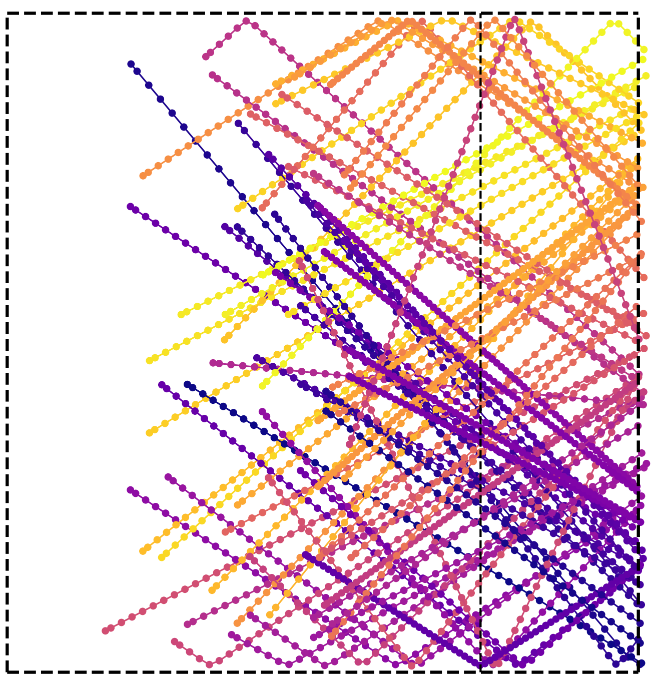
Fronto-Parietal Network

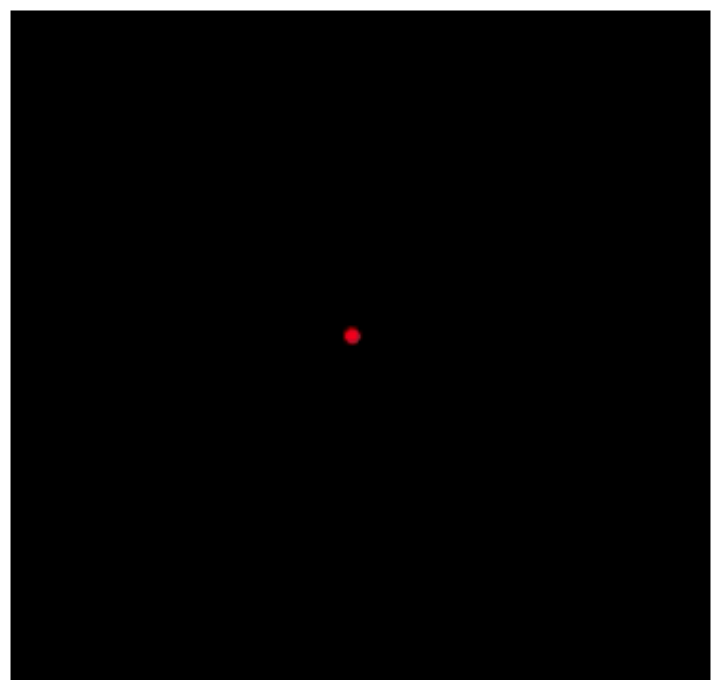
Dorsomedial frontal cortex (DMFC)

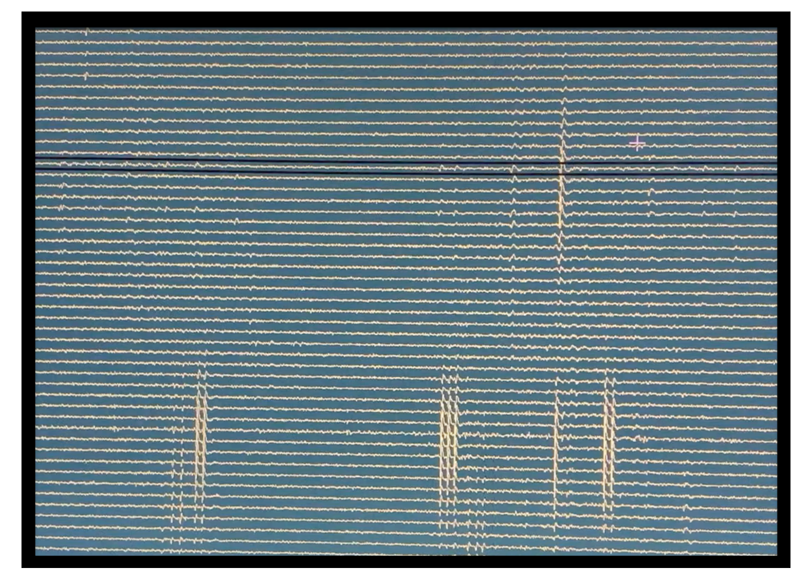
Monkey P *Monkey M*



79 conditions



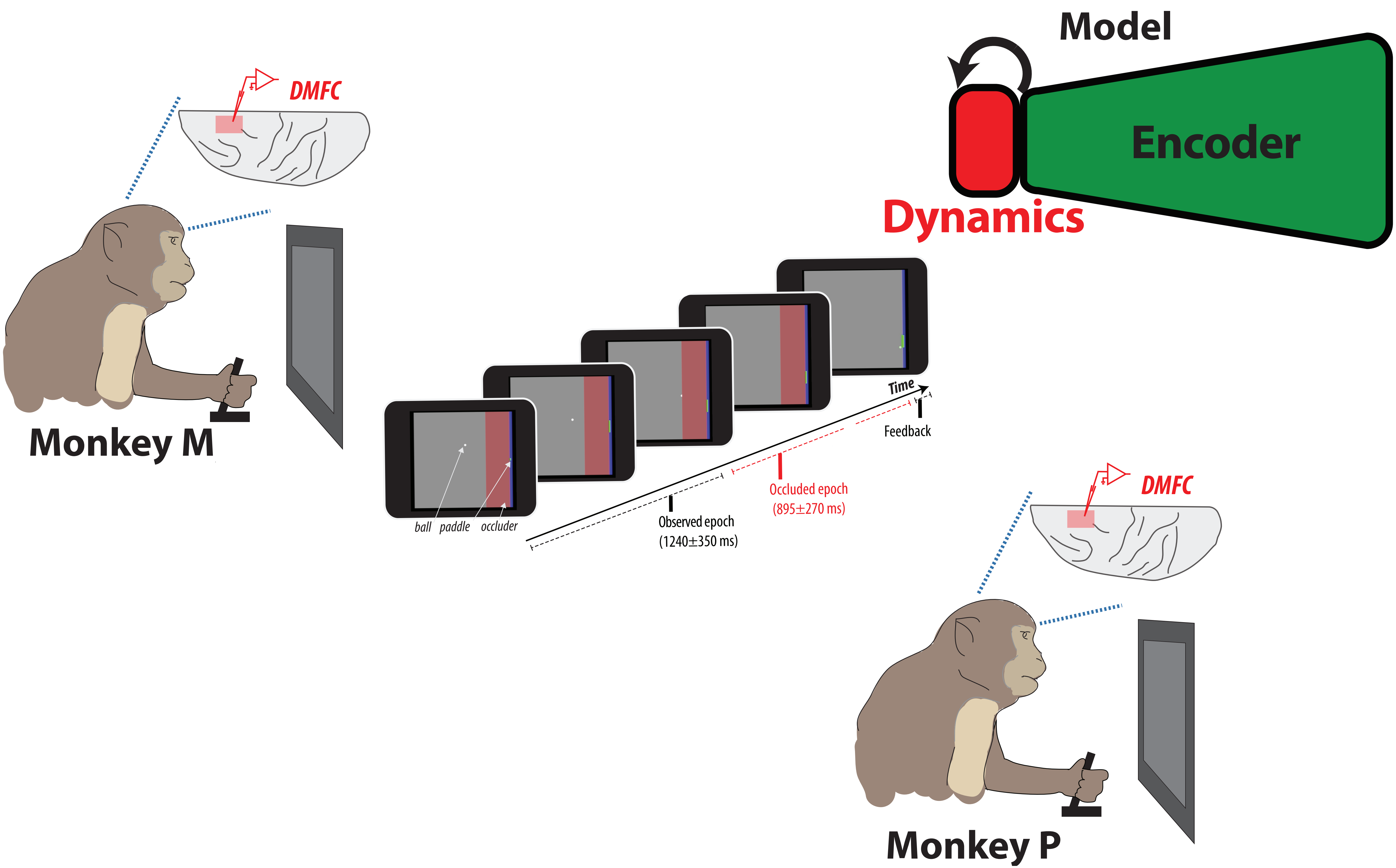




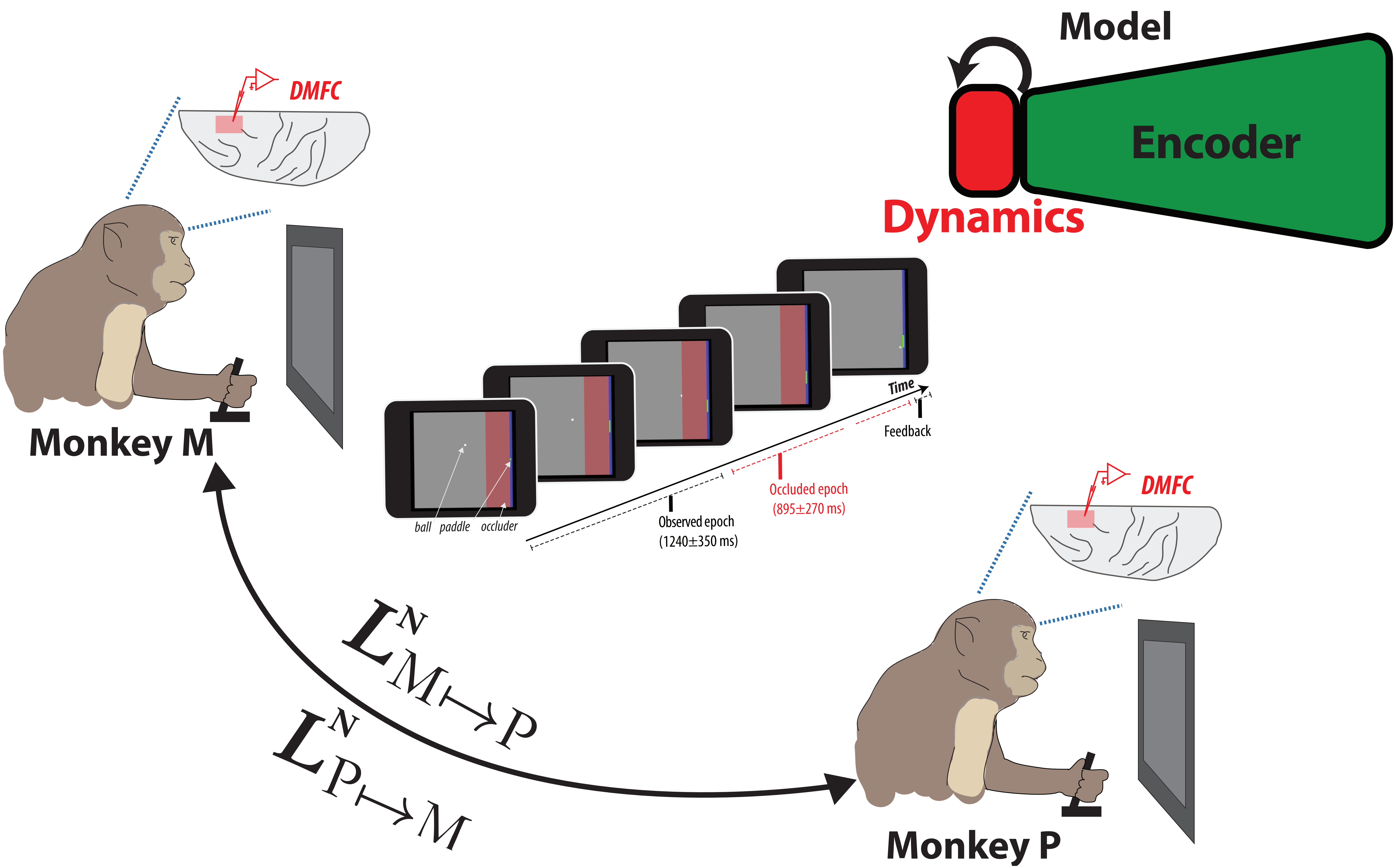
- *Data from two male adult monkeys*
- *79 subsampled M-Pong conditions*
- *64 channel v-probe (monkey P) and 384-channel Neuropixel probe (monkey M)*
- *Total of 1889 stable & reliable neurons recorded from DMFC*

Rishi Rajalingham

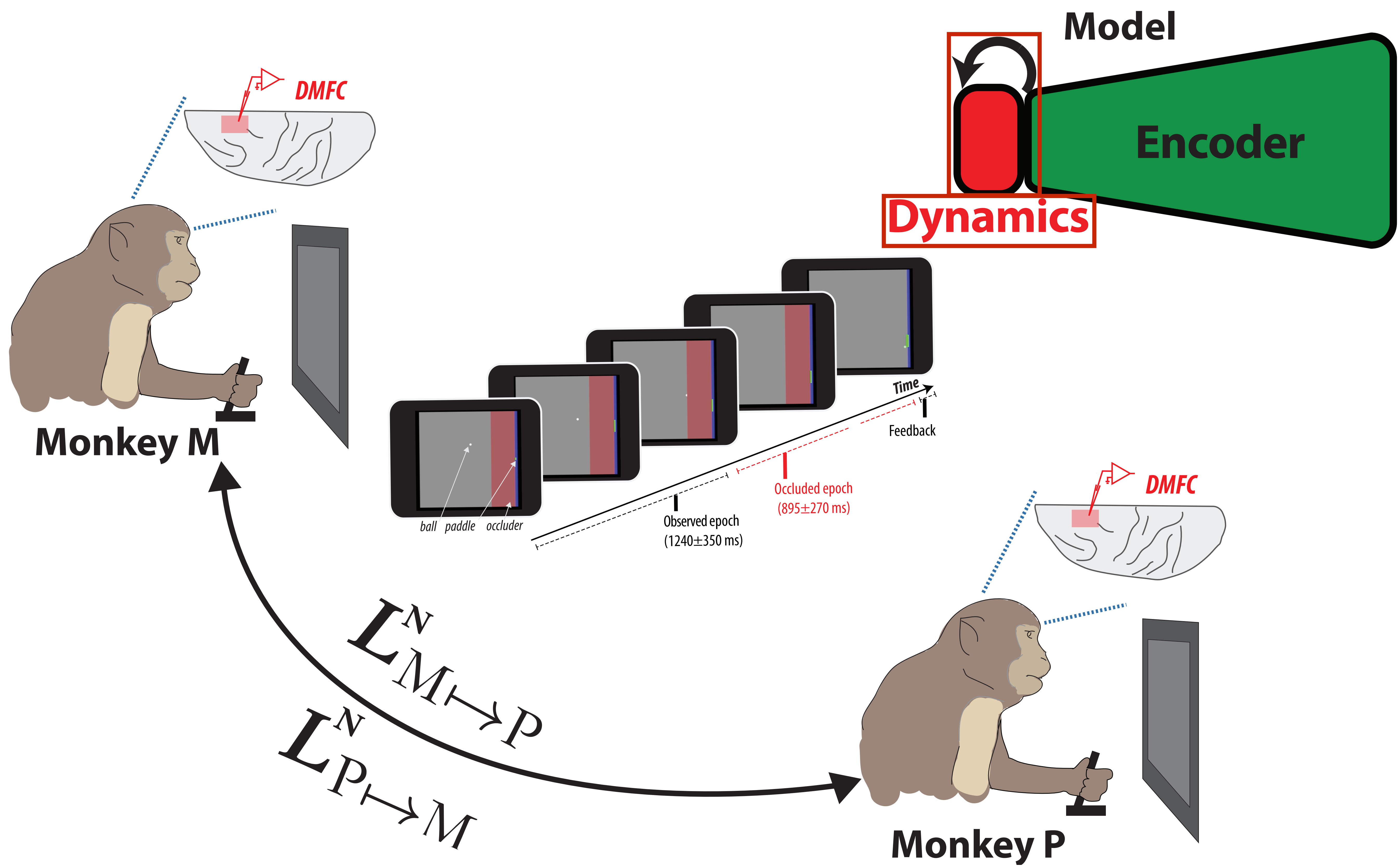
Assessing Model Similarity: Neural Response Predictivity



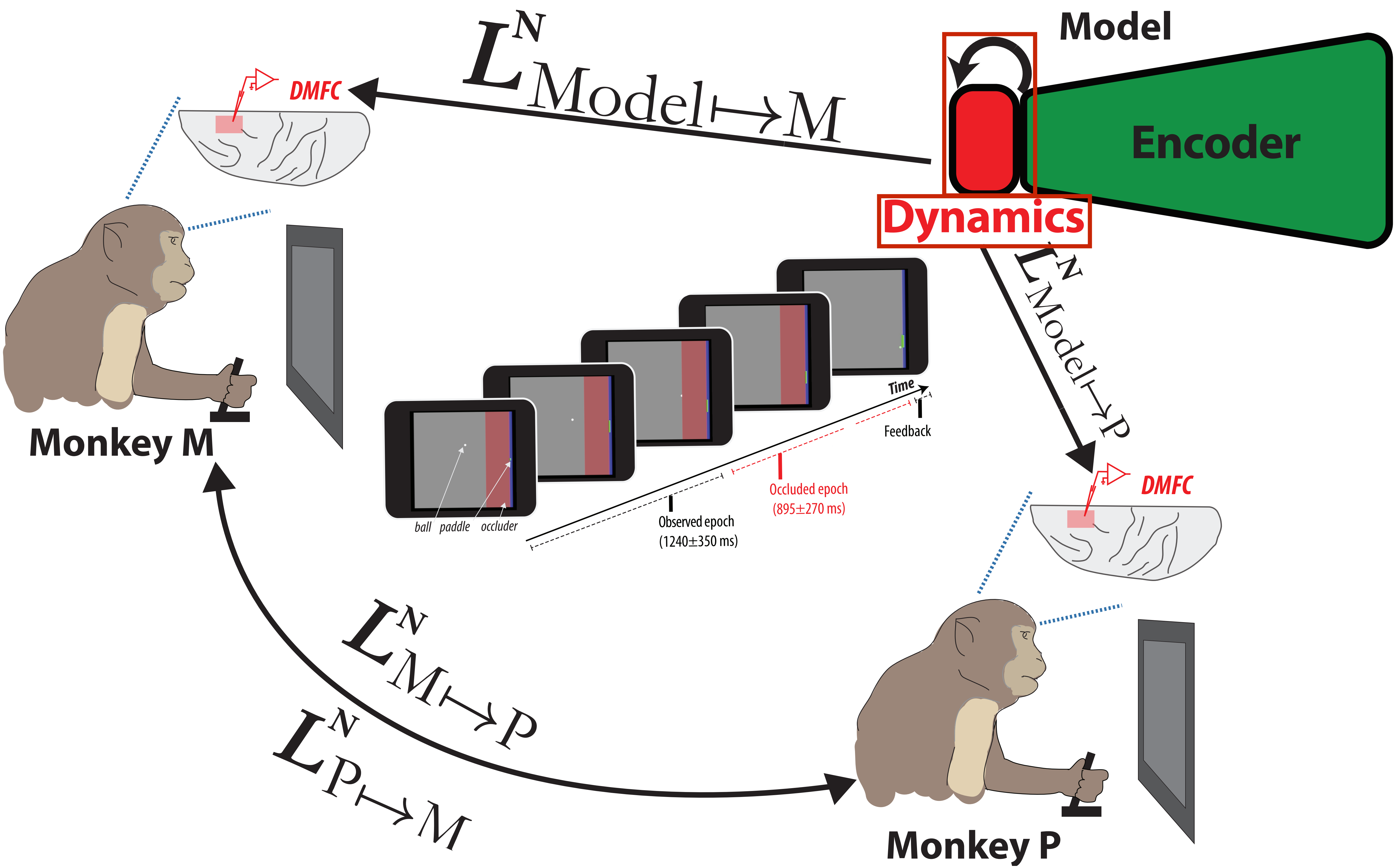
Assessing Model Similarity: Neural Response Predictivity



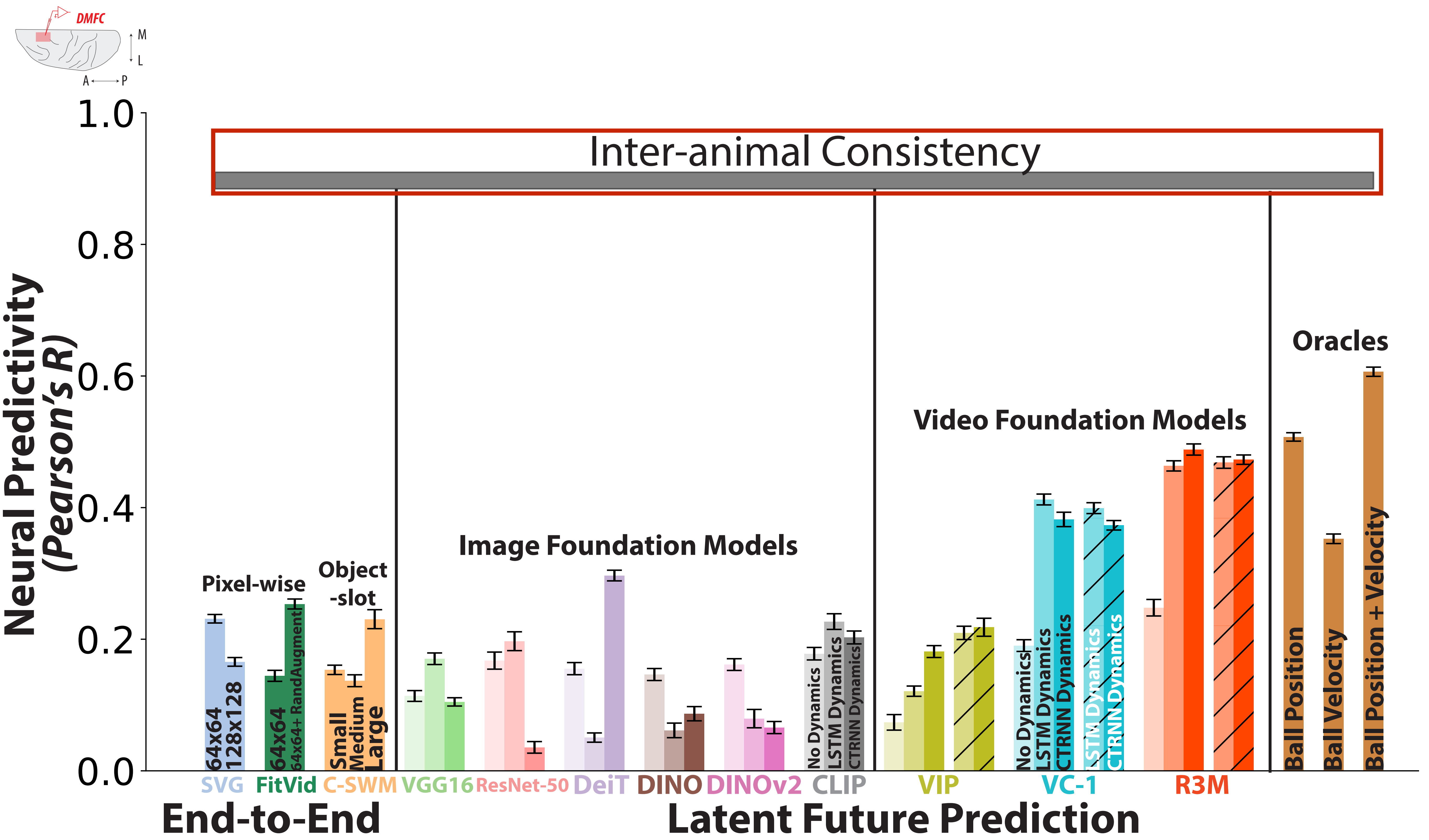
Assessing Model Similarity: Neural Response Predictivity



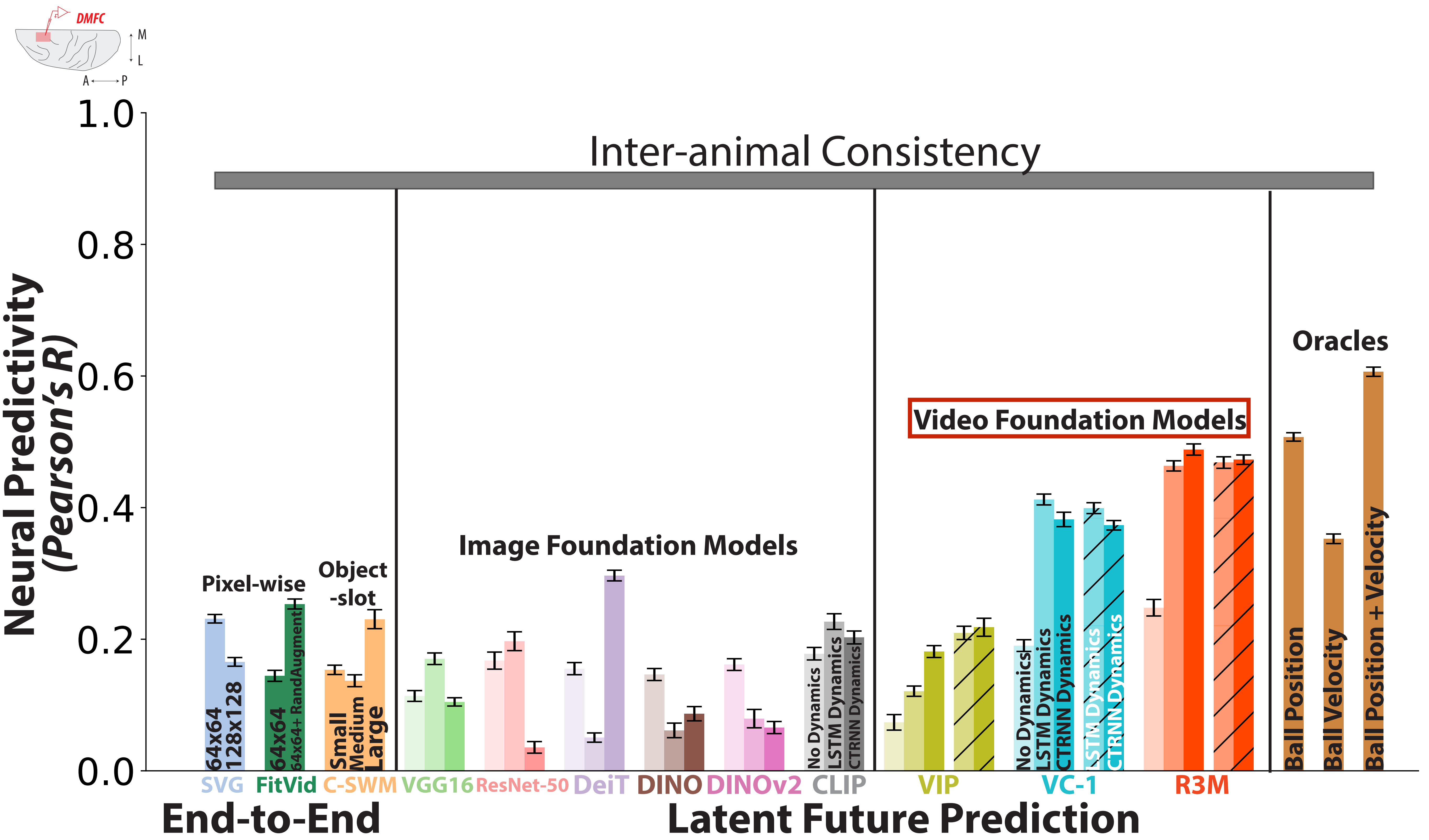
Assessing Model Similarity: Neural Response Predictivity



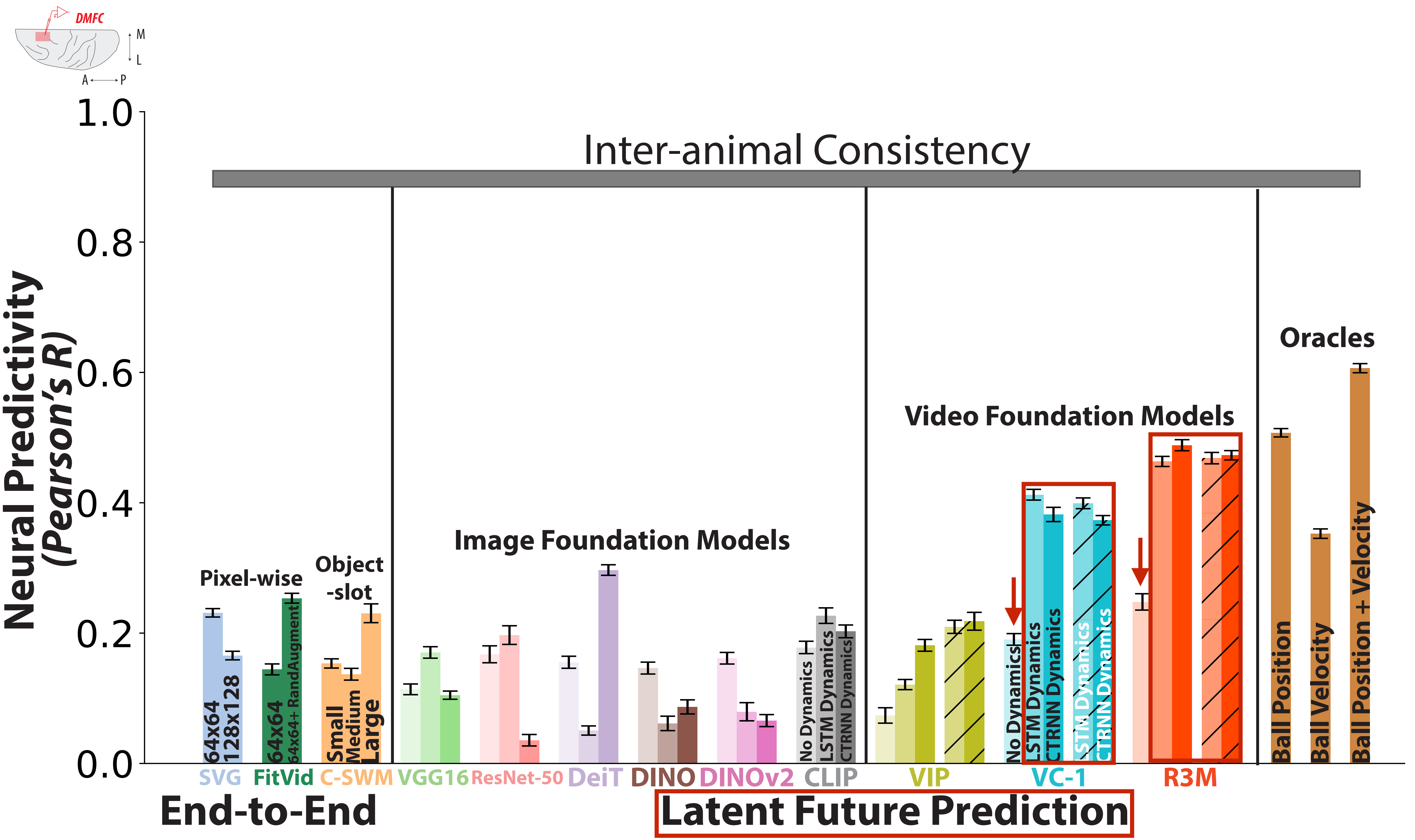
Neural response predictivity strongly separates models



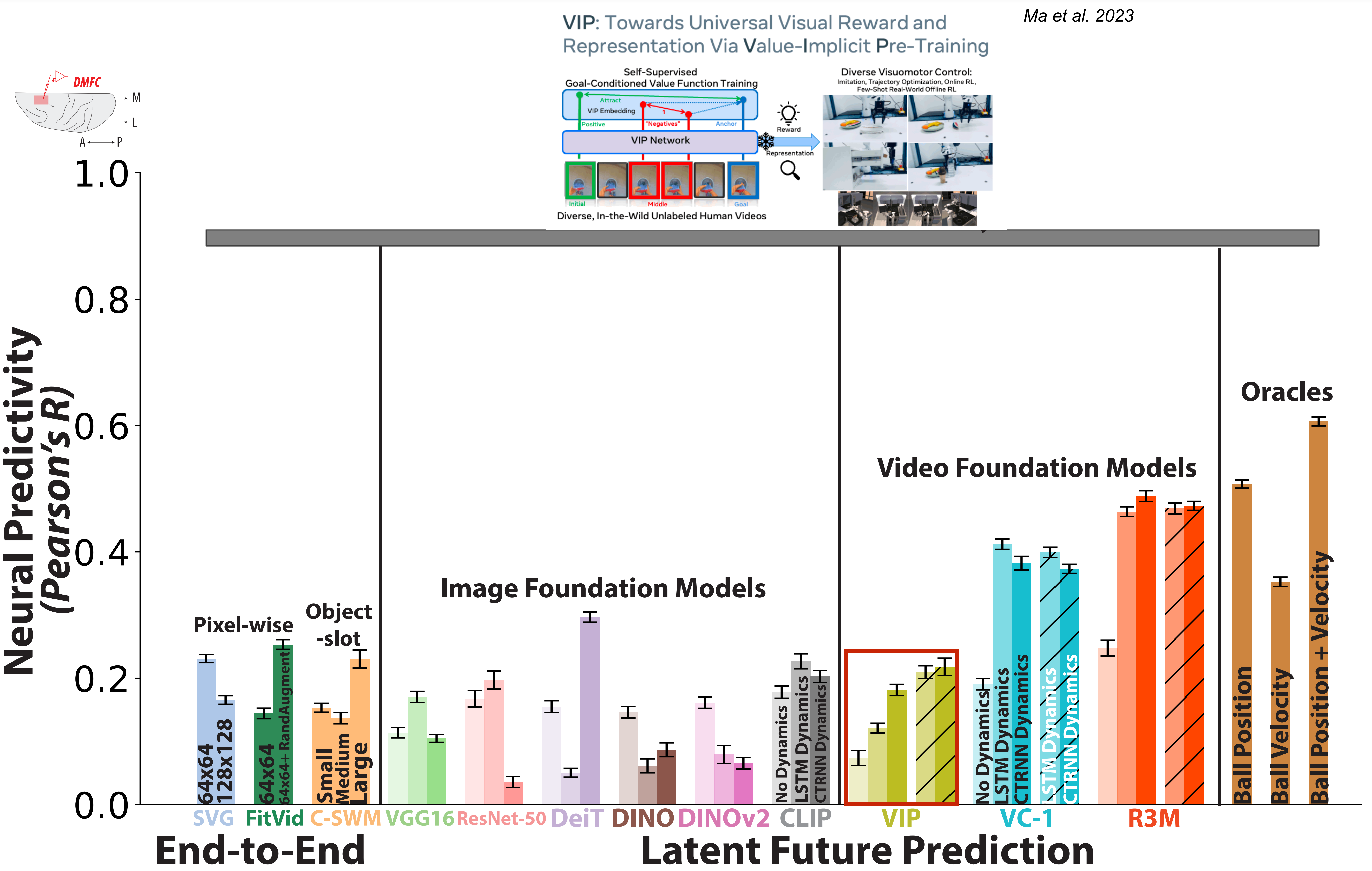
Neural response predictivity strongly separates models



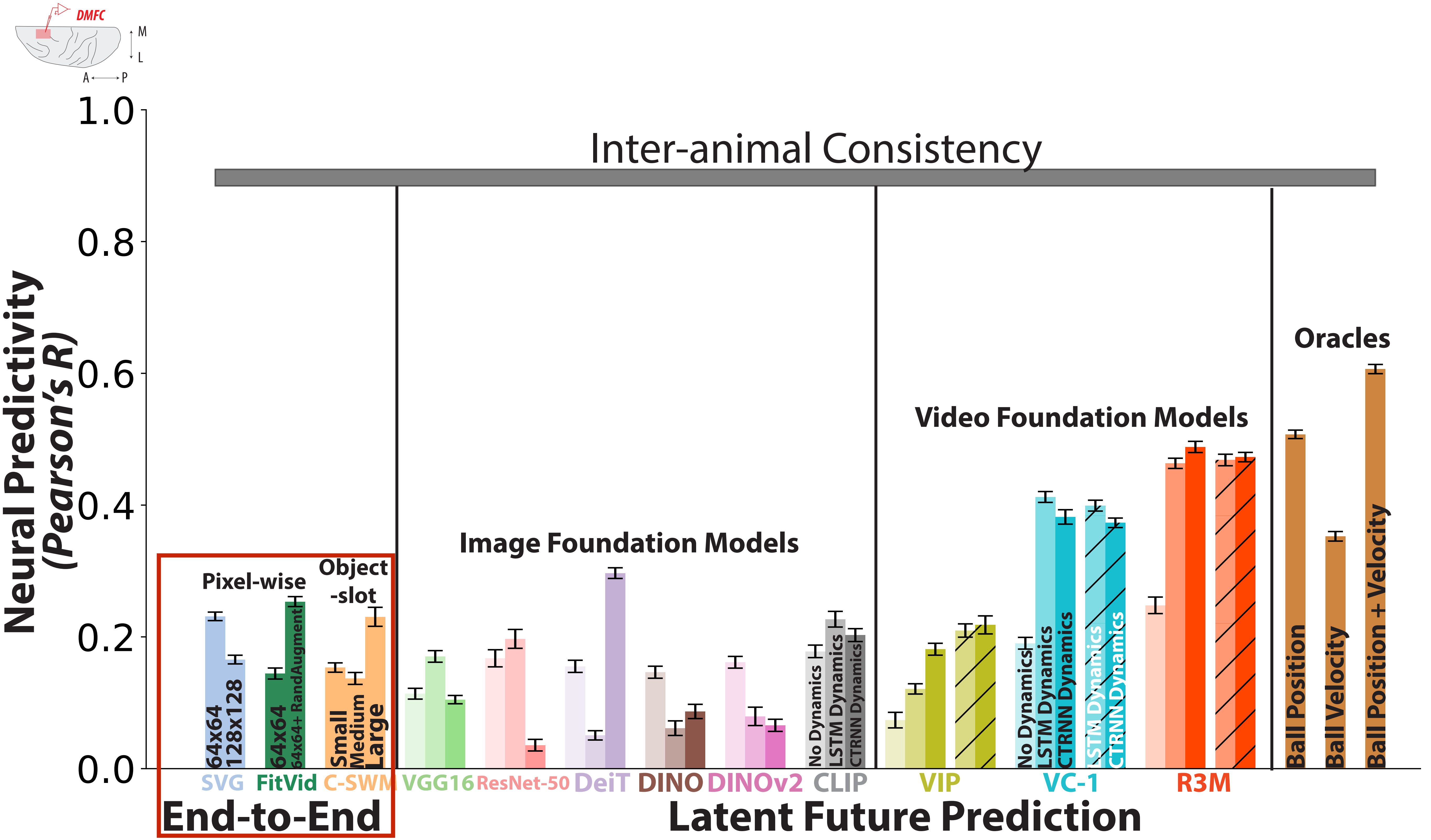
Neural response predictivity strongly separates models



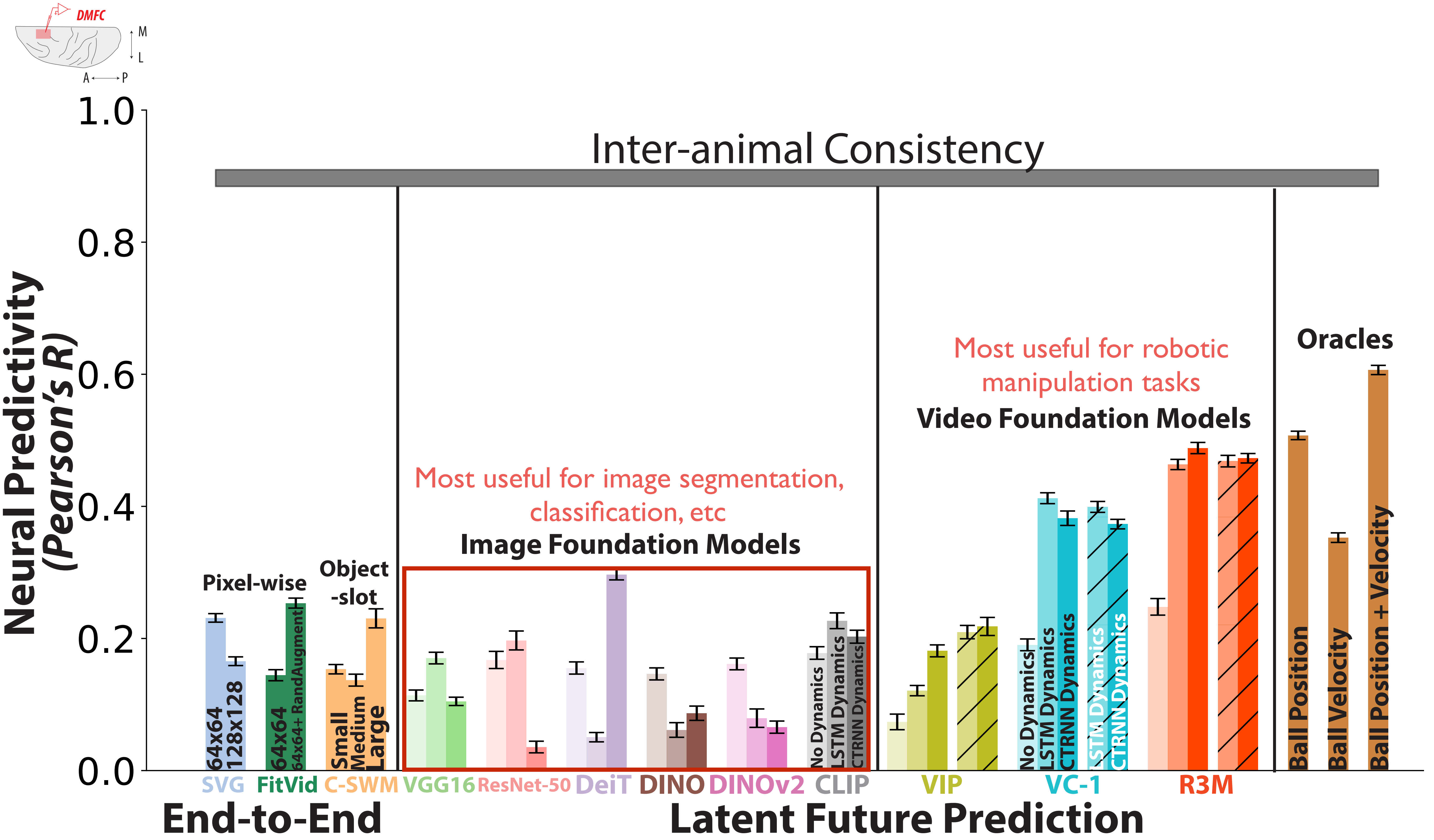
Neural response predictivity strongly separates models



Neural response predictivity strongly separates models

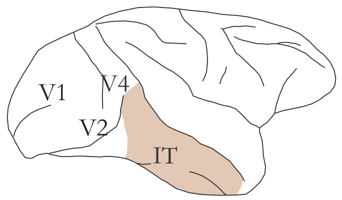


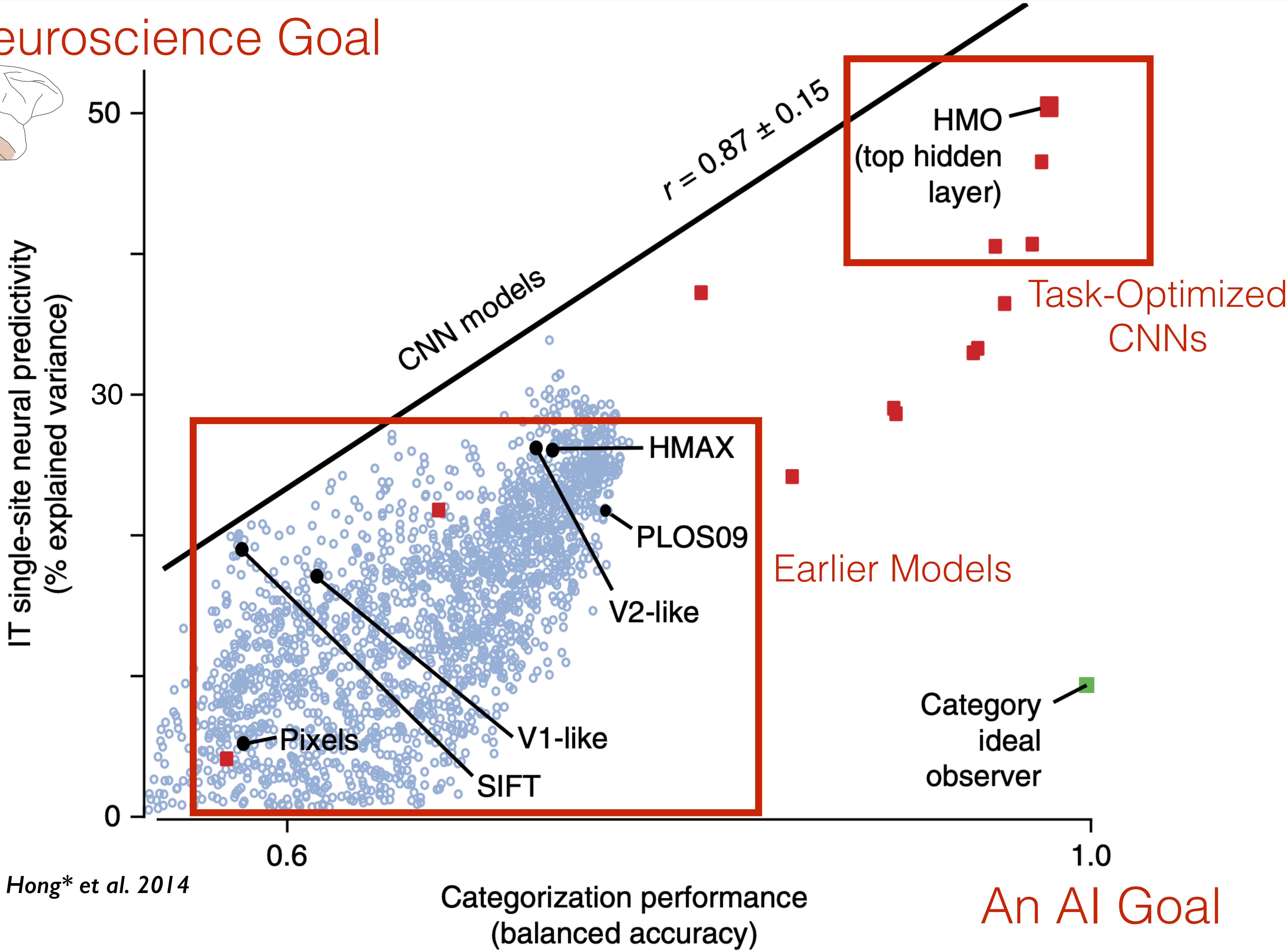
Neural response predictivity strongly separates models



Prior Results in Inferior Temporal (IT) Cortex

A Neuroscience Goal

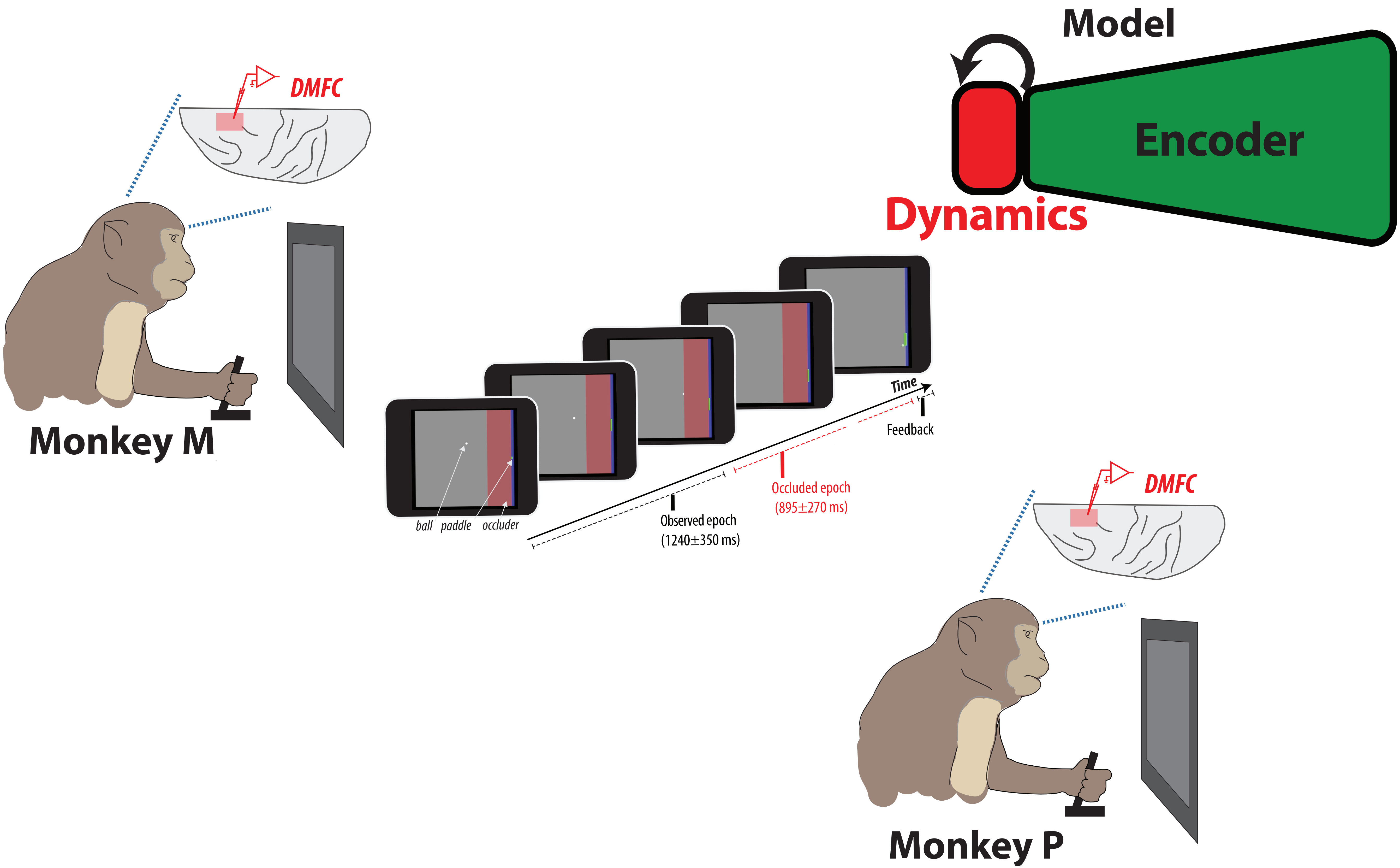




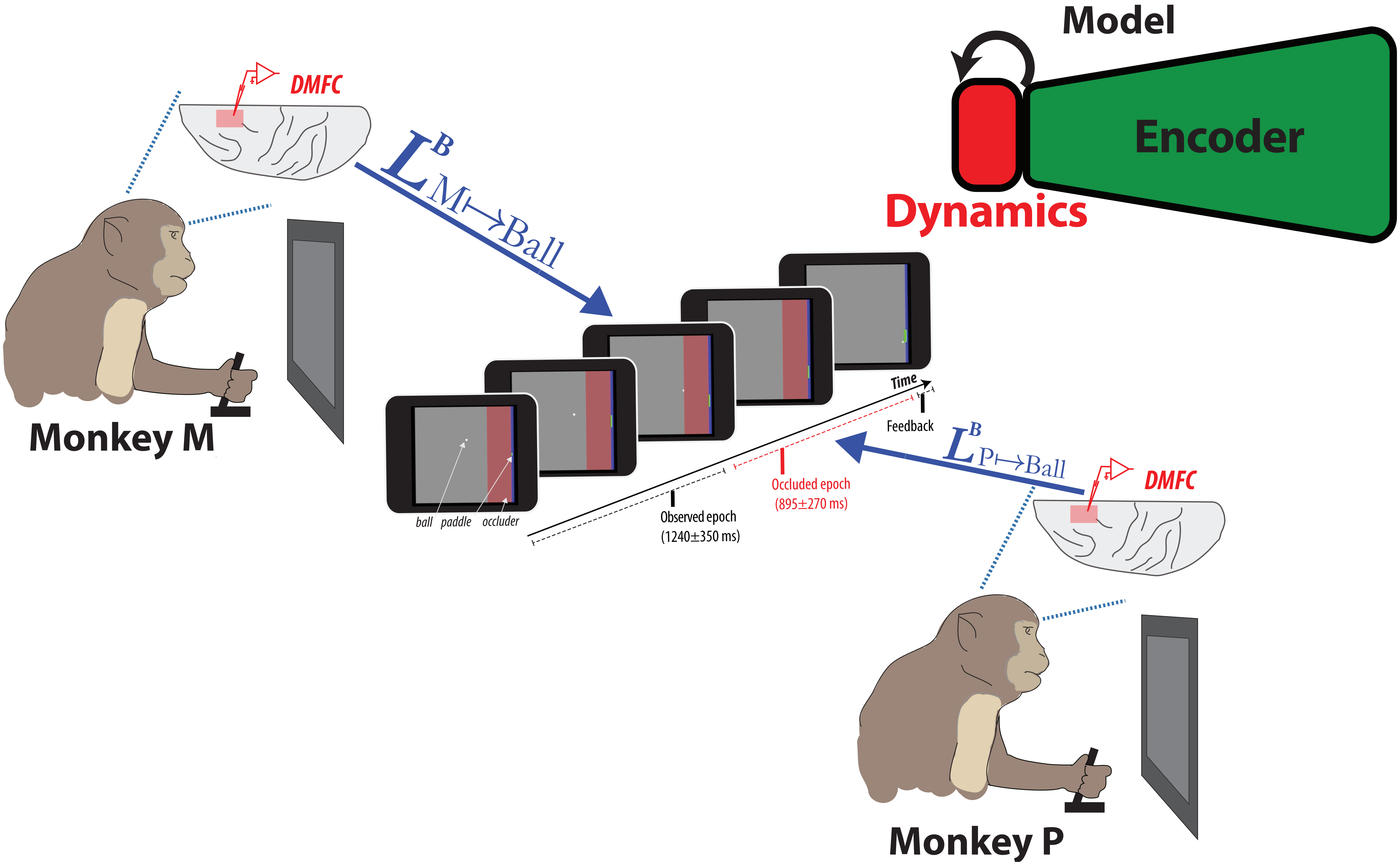
Yamins, Hong* et al. 2014*

An AI Goal

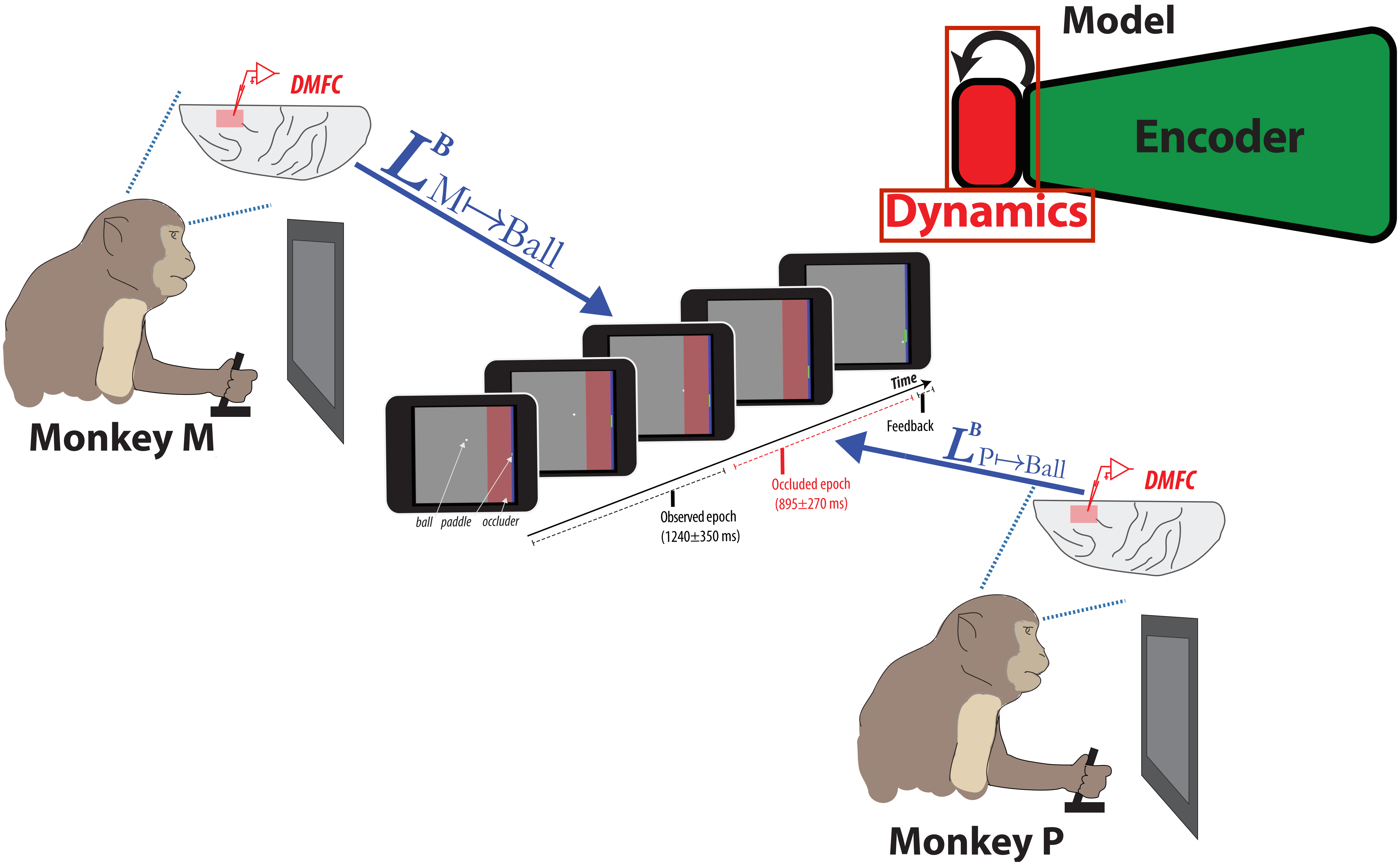
Assessing Model Similarity: Ground Truth State Decoding



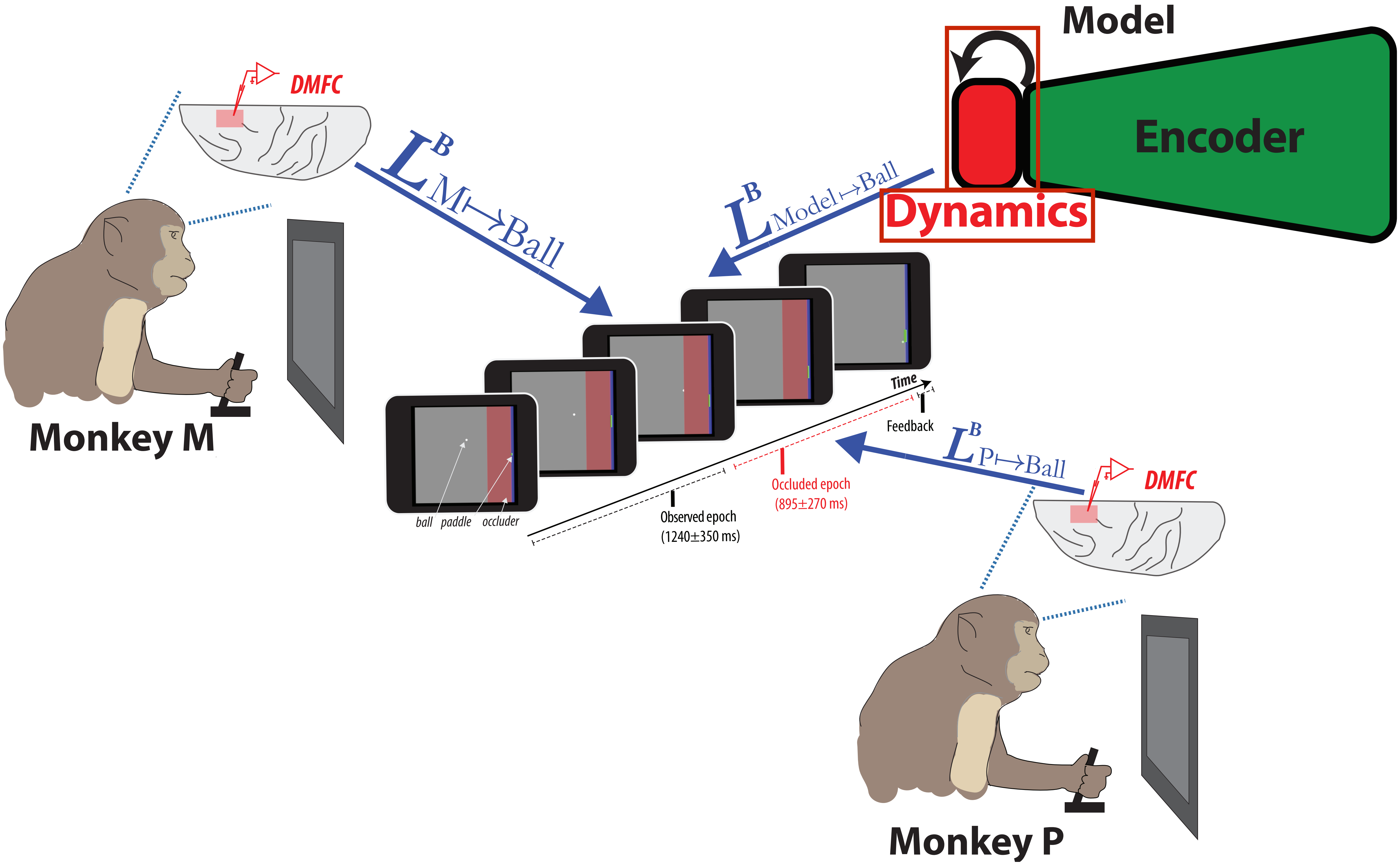
Assessing Model Similarity: Ground Truth State Decoding



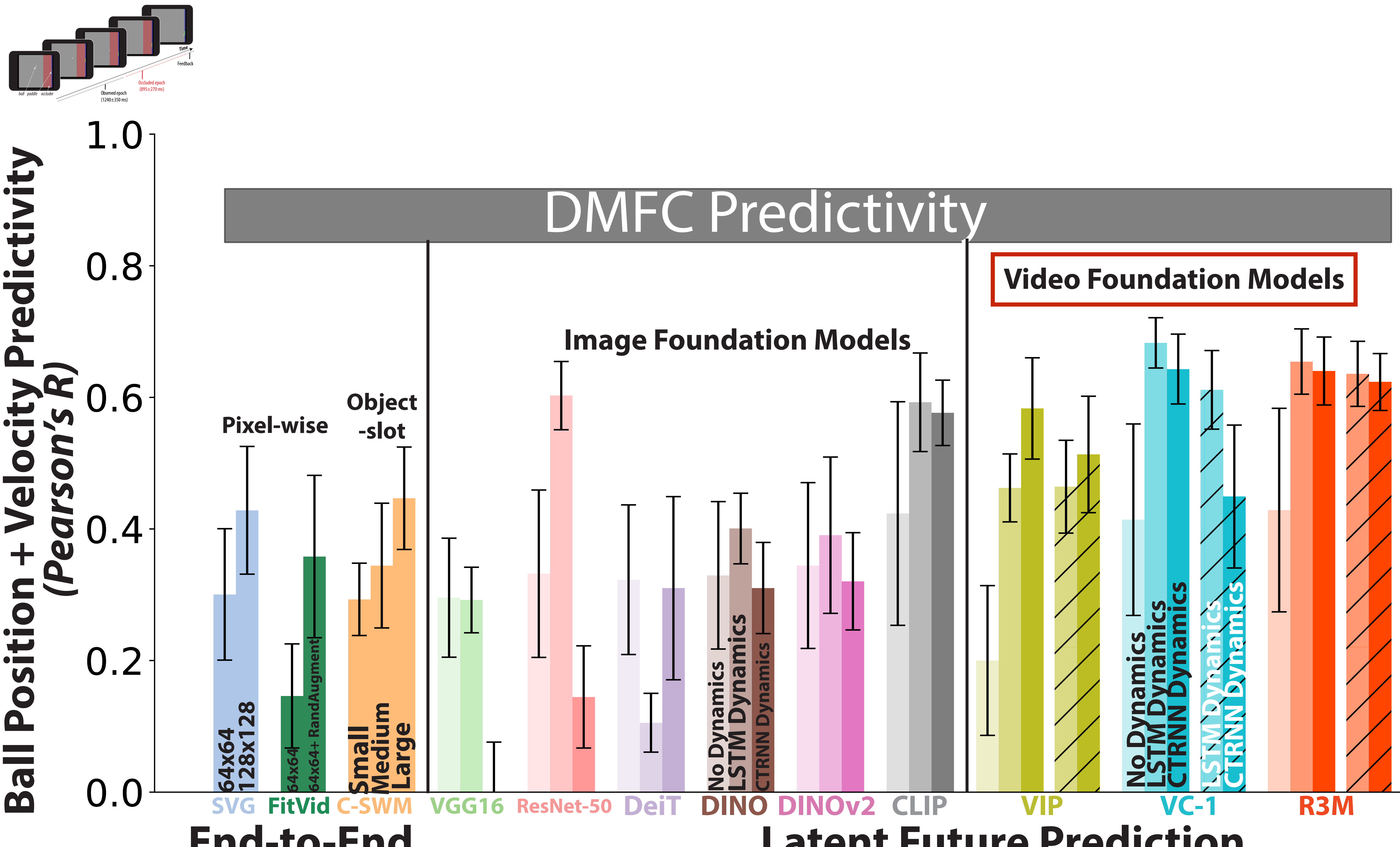
Assessing Model Similarity: Ground Truth State Decoding



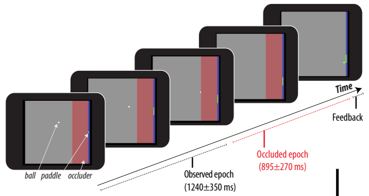
Assessing Model Similarity: Ground Truth State Decoding

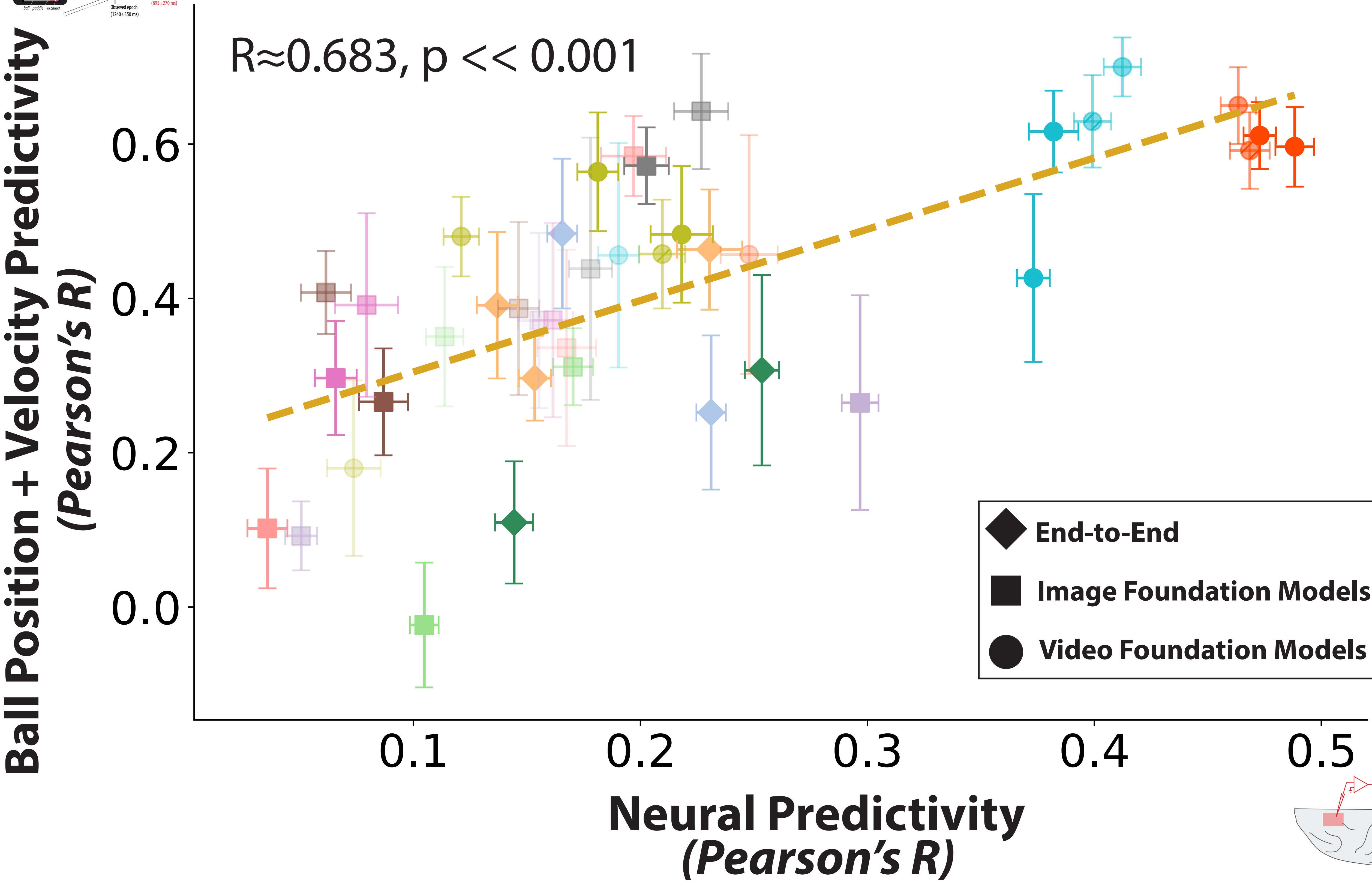


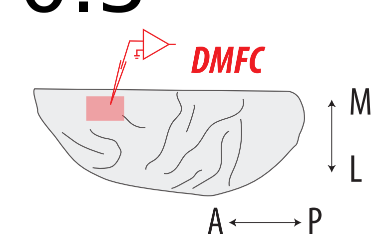
Best models approach ground truth state predictivity ceiling



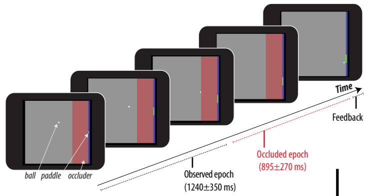
Predicting neurons is relevant to simulating the ball

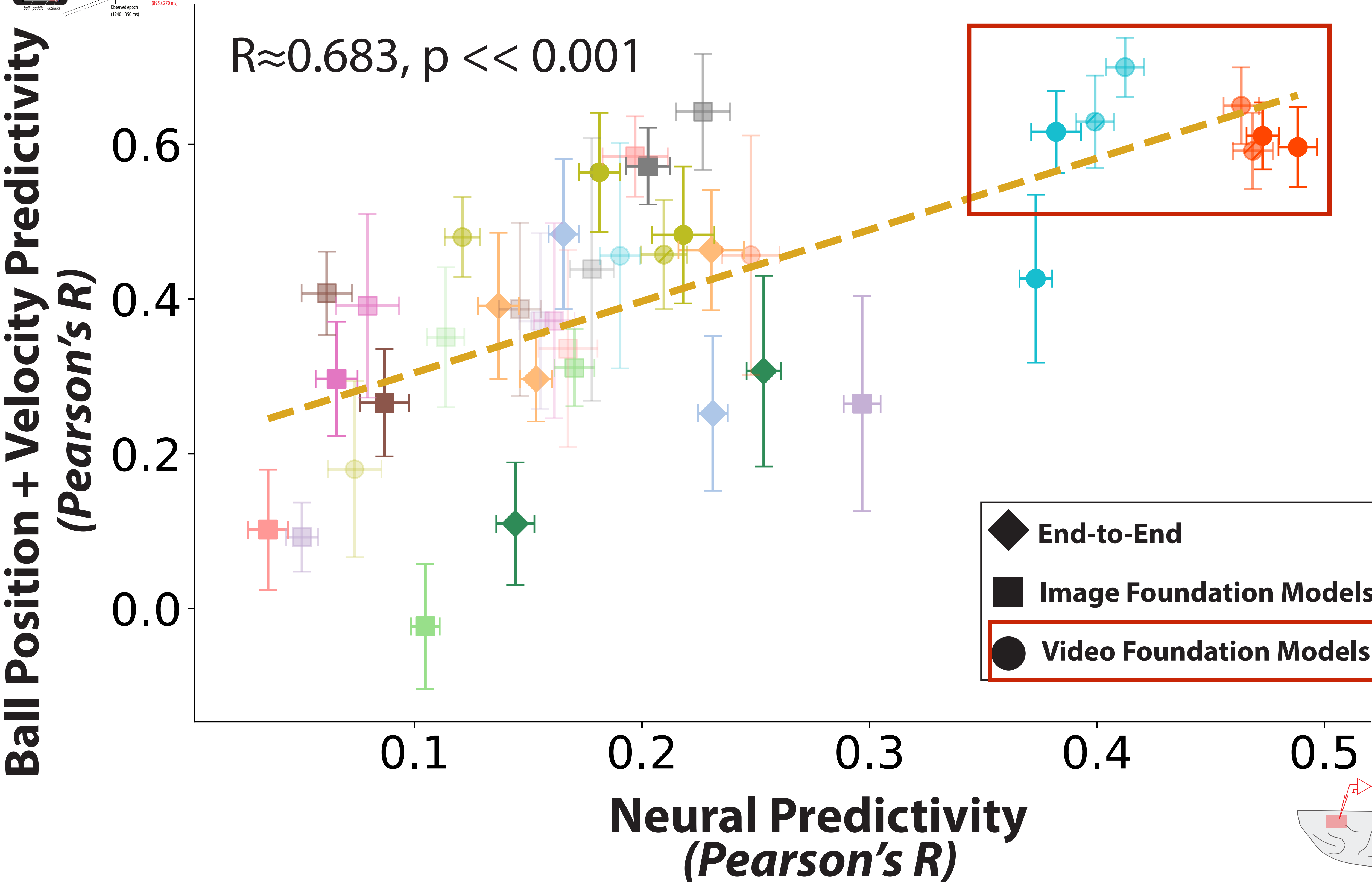


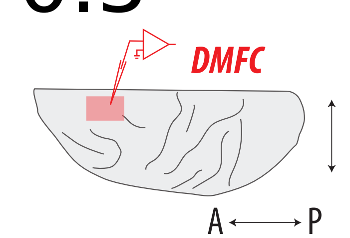




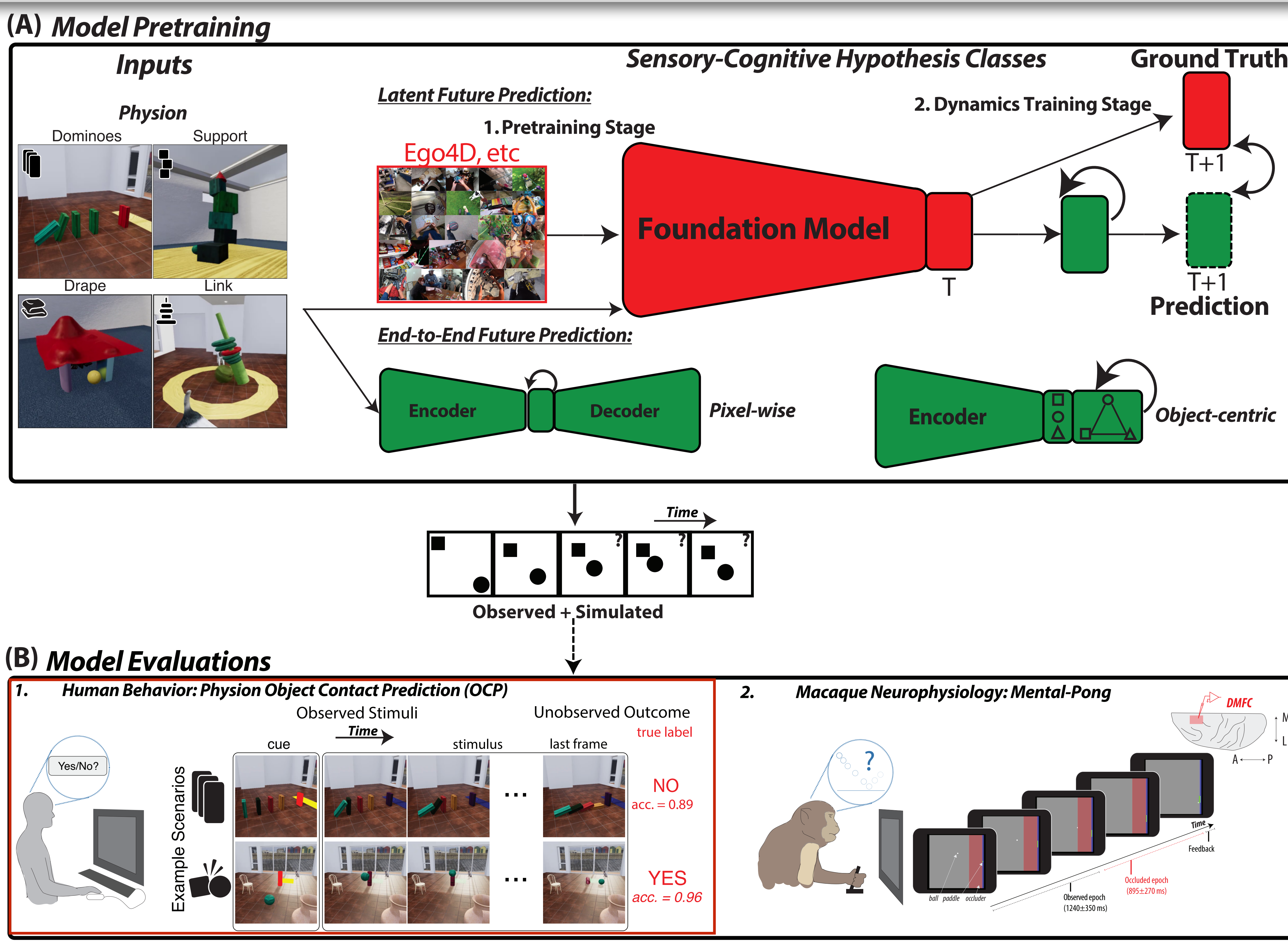
Predicting neurons is relevant to simulating the ball







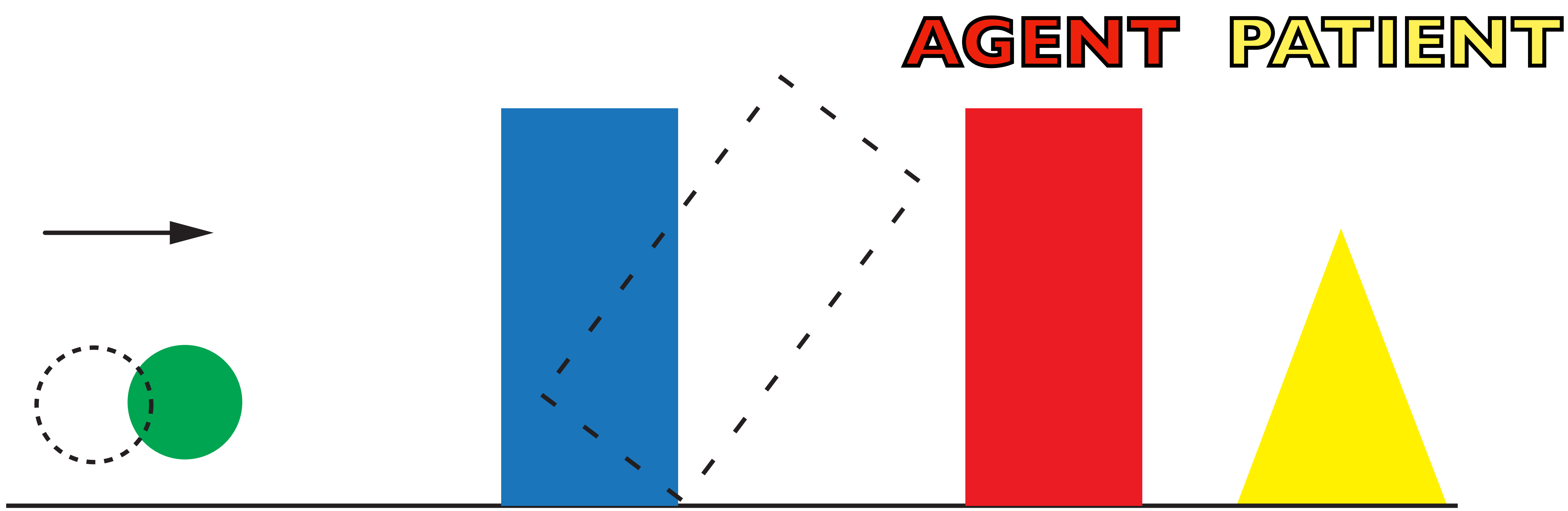
Model Evaluations: Object Contact Prediction (OCP)



Model Evaluations: Object Contact Prediction (OCP)

Bear et al. 2021

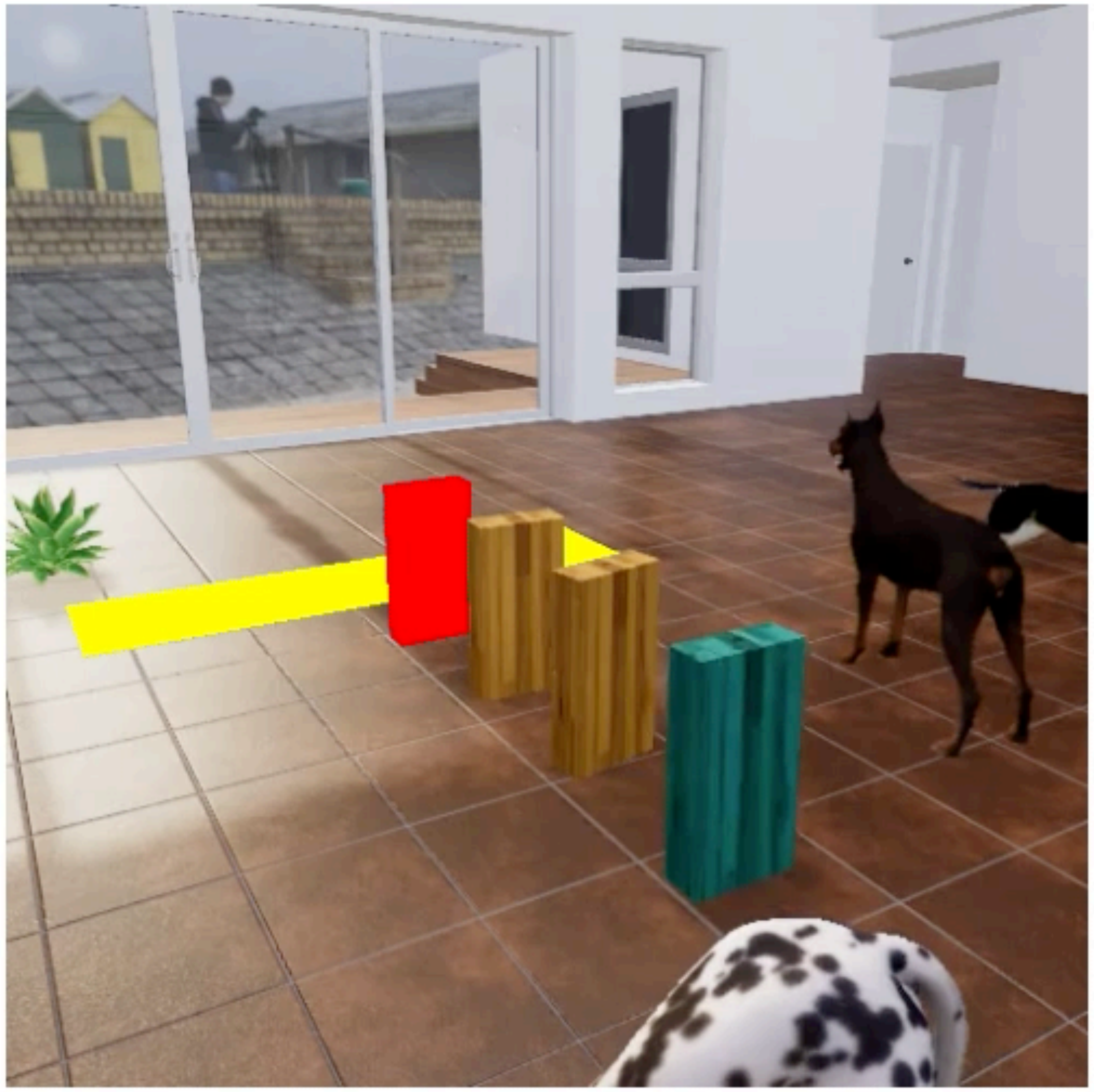
"Will the agent object contact the patient object?"



Daniel Bear | Joshua Tenenbaum | Daniel Yamins | Judith Fan

Completion Progress

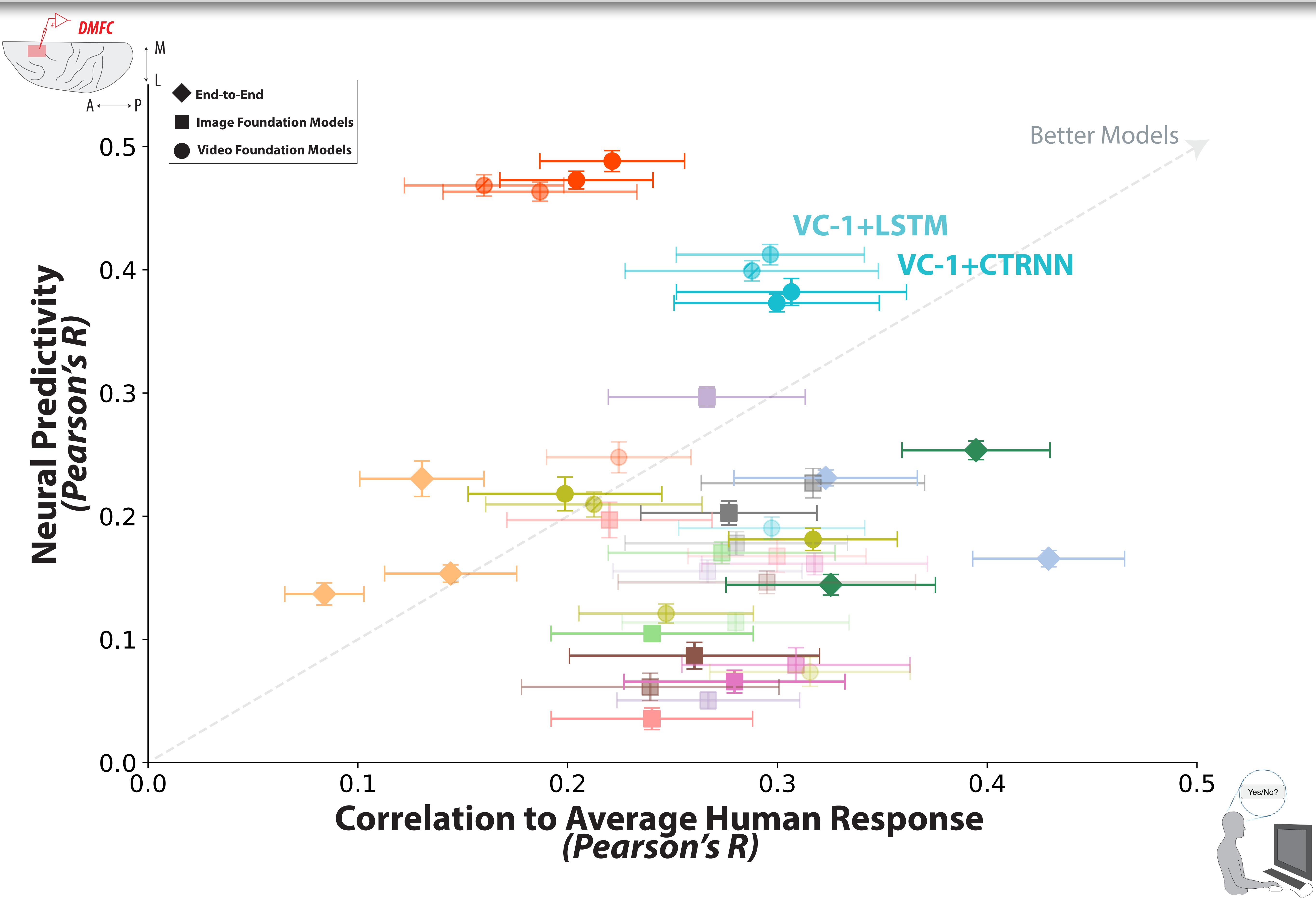
Bear et al. 2021



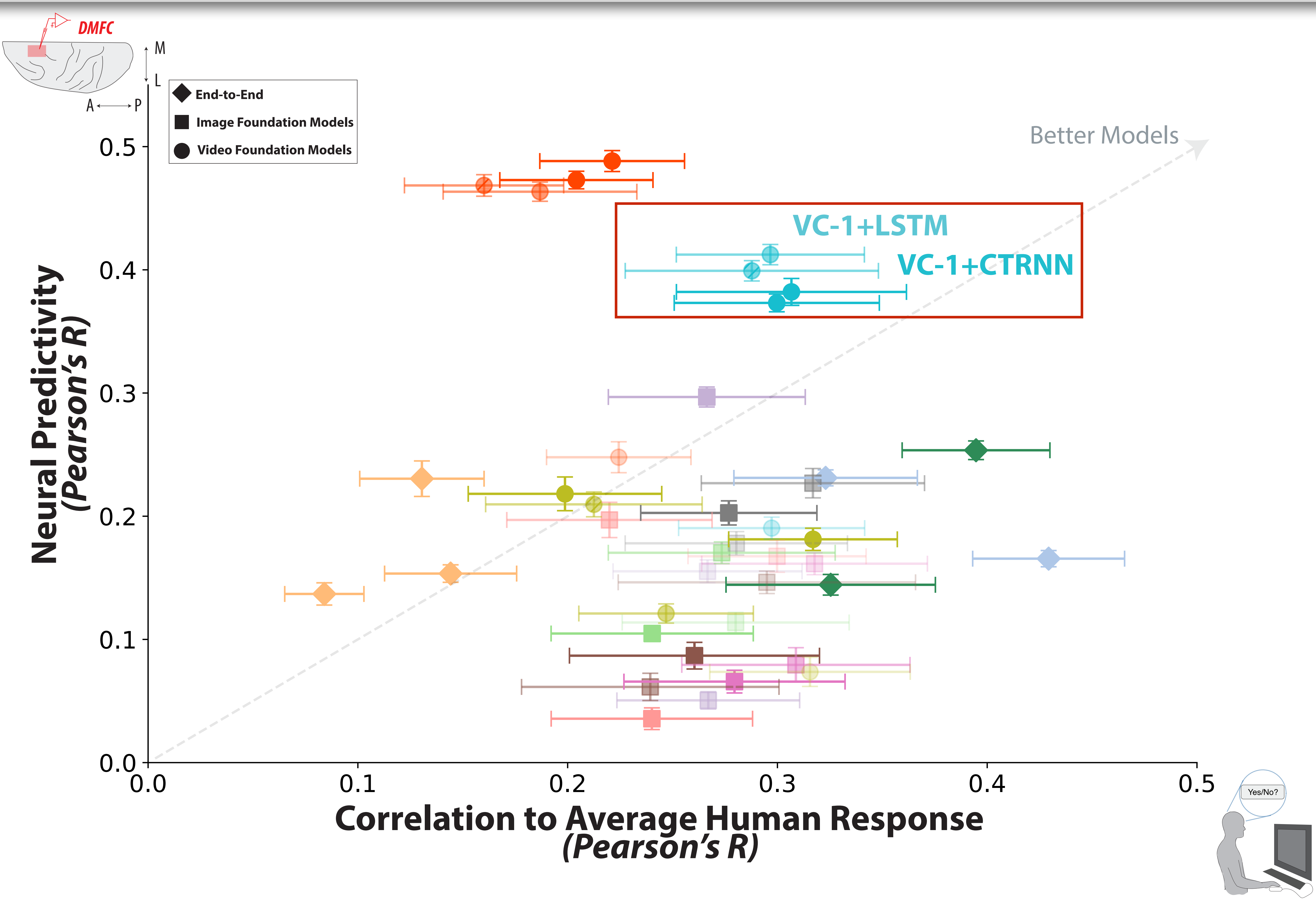
YES NO

Is the red object going to hit the yellow area?

Dynamically Equipped Sensorimotor Foundation Models Can Match Both



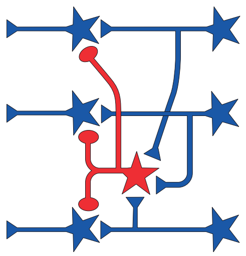
Dynamically Equipped Sensorimotor Foundation Models Can Match Both



Takeaways

A = *architecture class*

1. *"Circuit"*

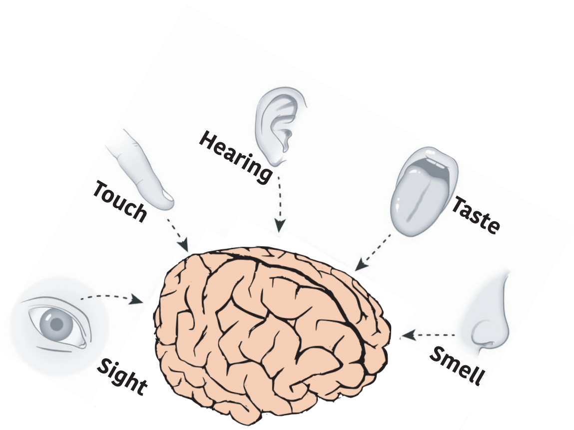


3. *"Ecological niche/behavior"*

T = *task loss*

<u>Neurobiological Puzzle:</u>

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



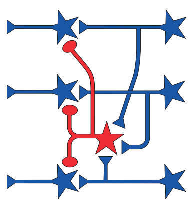
2. *"Environment"*

D = *data stream*

Takeaways

A = *architecture class*

1. *"Circuit"*



T = *task loss*

3. *"Ecological niche/behavior"*

latent future prediction

<u>Neurobiological Puzzle:</u>

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

<u>Findings:</u>

Mental simulation appears to be primarily relevant to predicting the environment future state in a suitable, highly-constrained *latent* space.

Latent space does not appear to just consist of bespoke object slots or prioritize fine-grained details (e.g. at the level of pixels), but rather mainly has to be *reusable* across *dynamic* scenes.

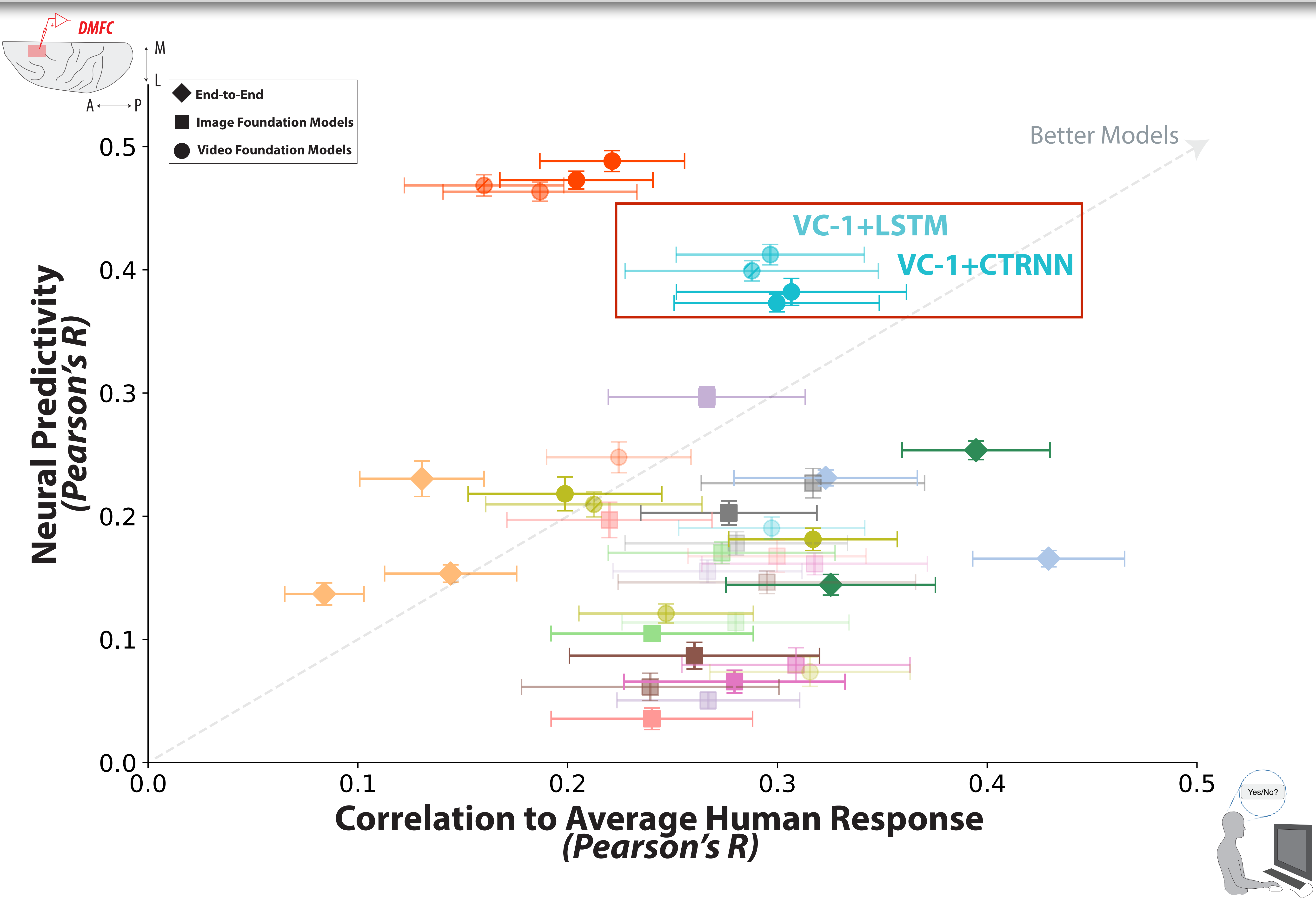
egocentric videos

2. *"Environment"*

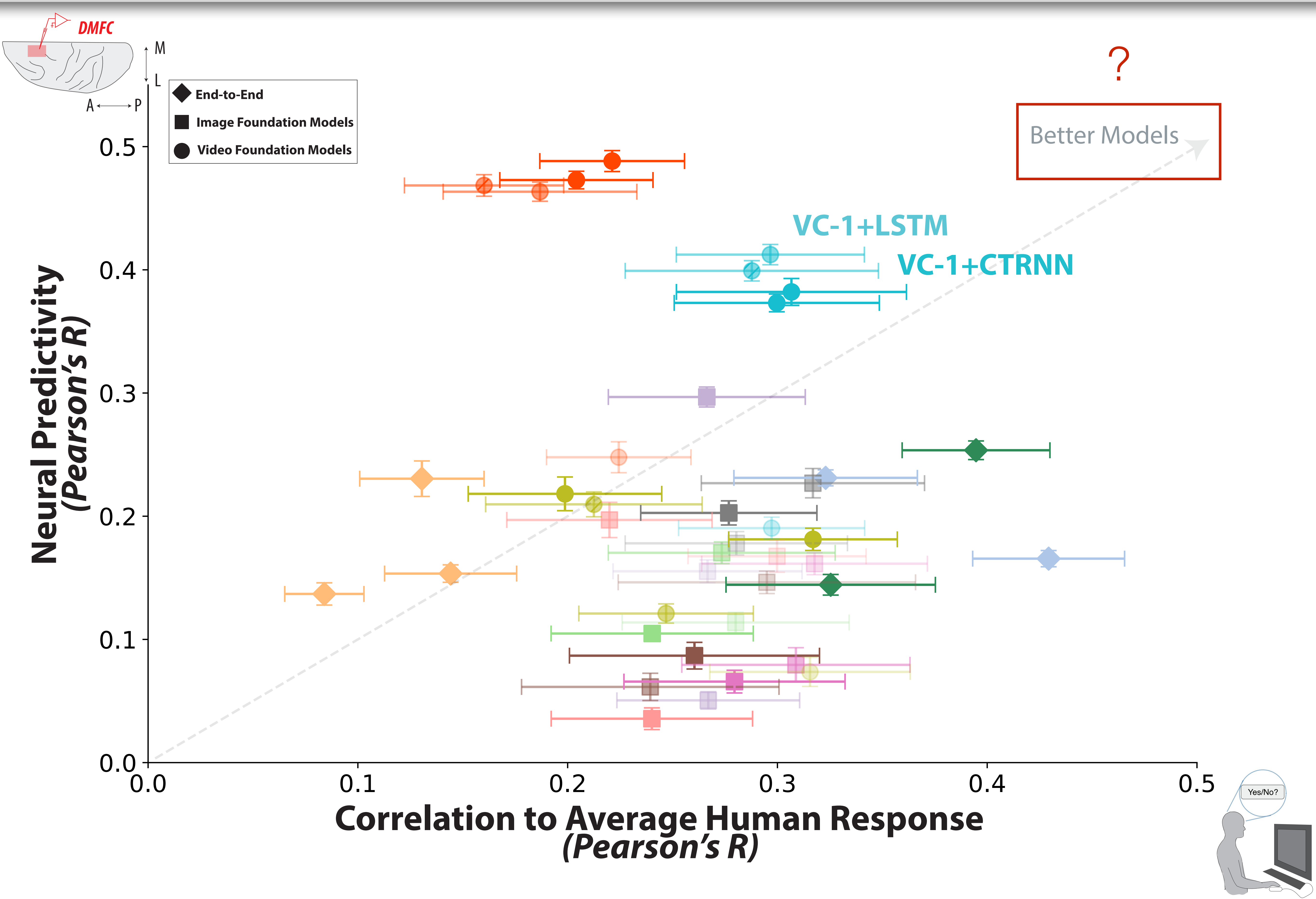
D = *data stream*

Future Directions

Future Directions



Future Directions

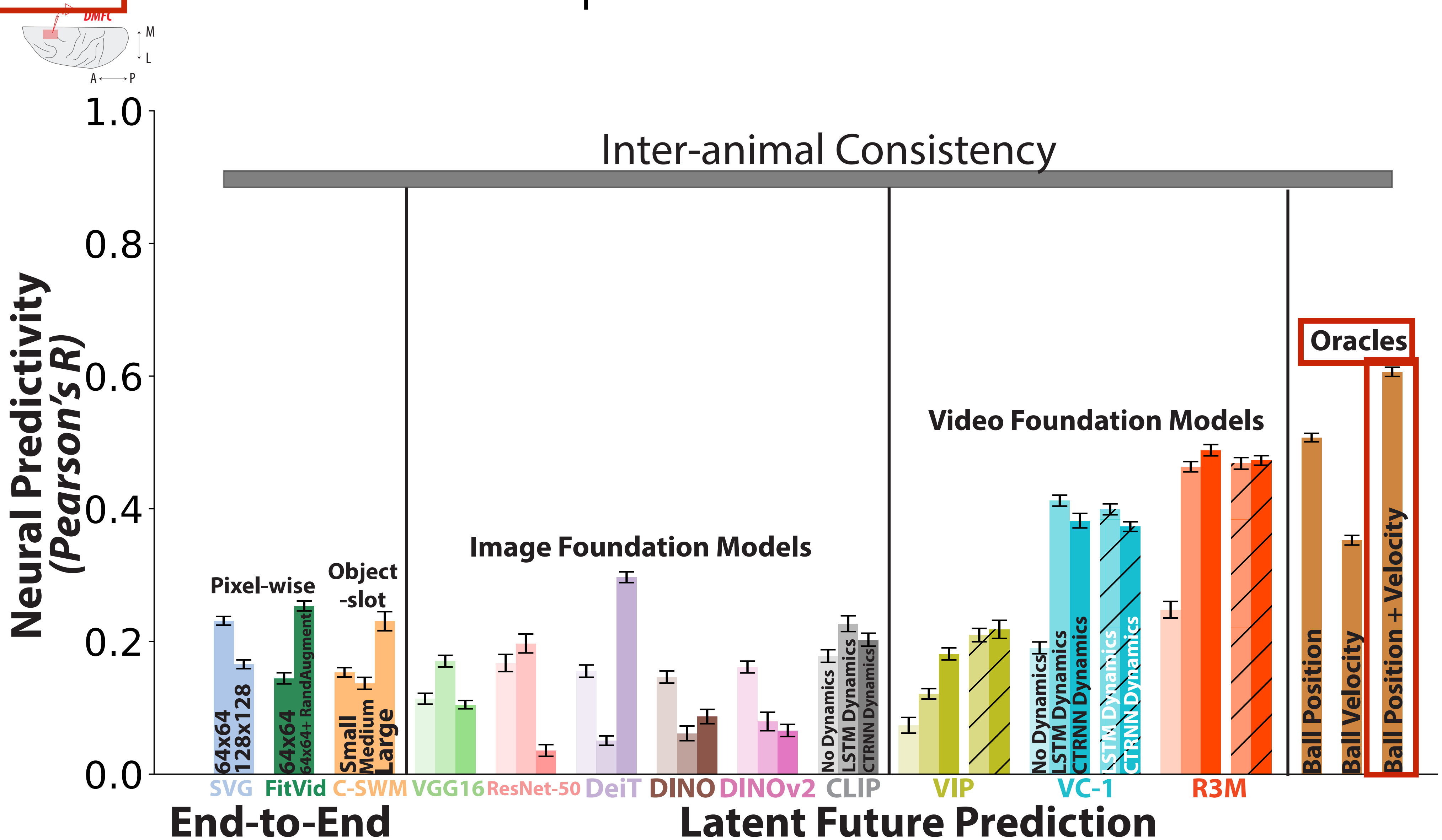


Future Directions

1. **Sensory:** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation:

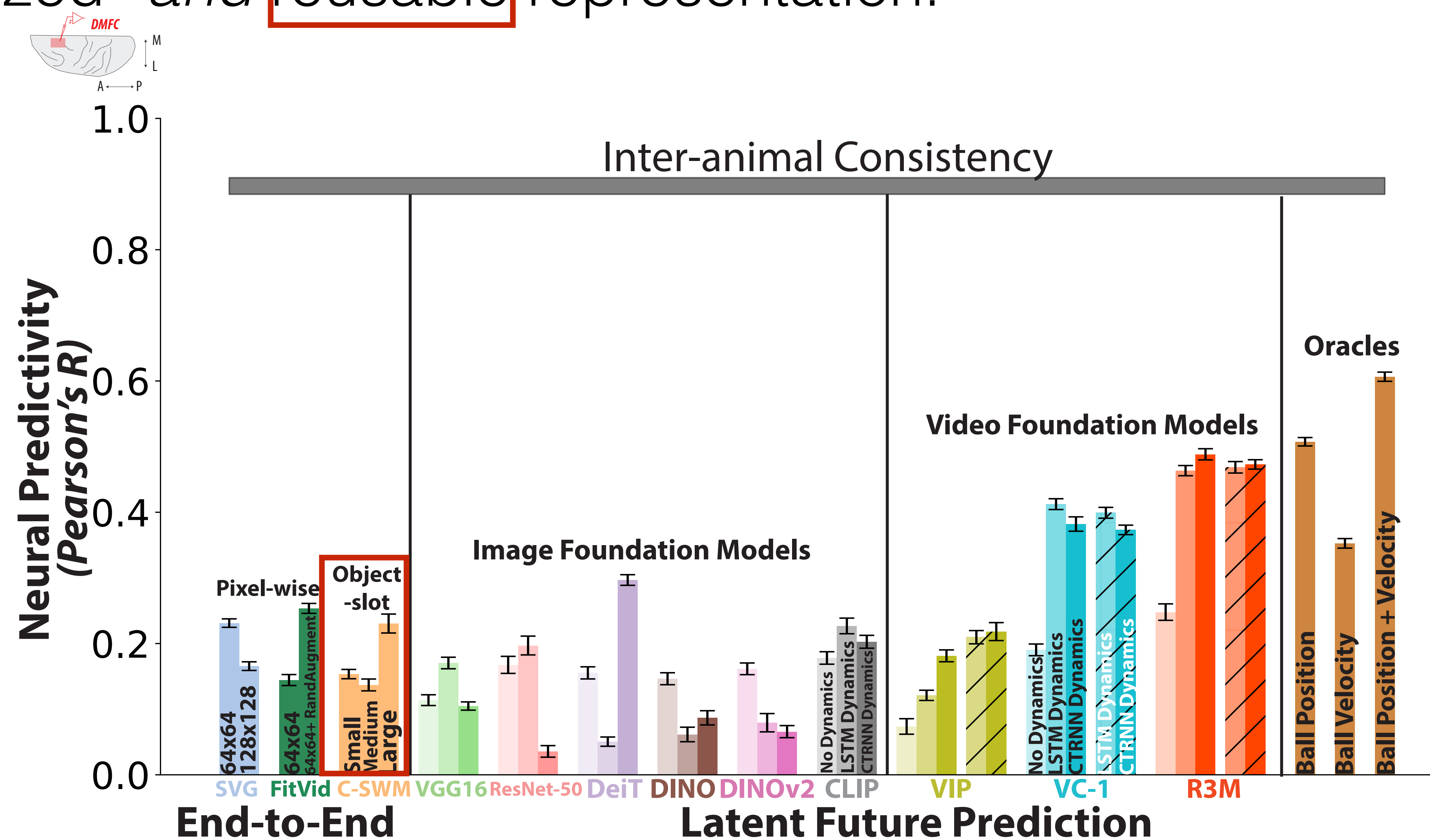
Future Directions

1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation:



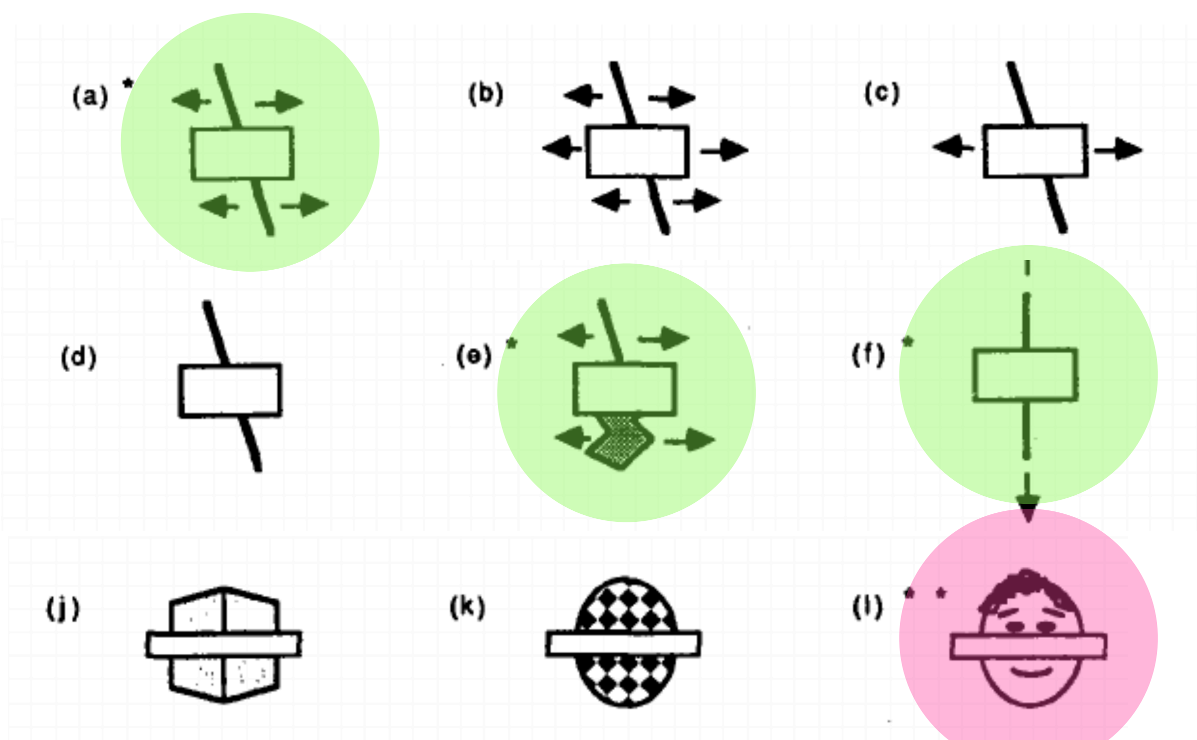
Future Directions

1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more “factorized” *and* reusable representation:



Future Directions

1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation: **object-centric, video foundation model**?



Principles of Object Perception Elizabeth Spelke, 1990

Elizabeth Spelke

Future Directions

1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation: **object-centric, video foundation model**?

2. **<u>Cognitive:</u>** Hierarchy/modularization of timescales in dynamics?

Future Directions

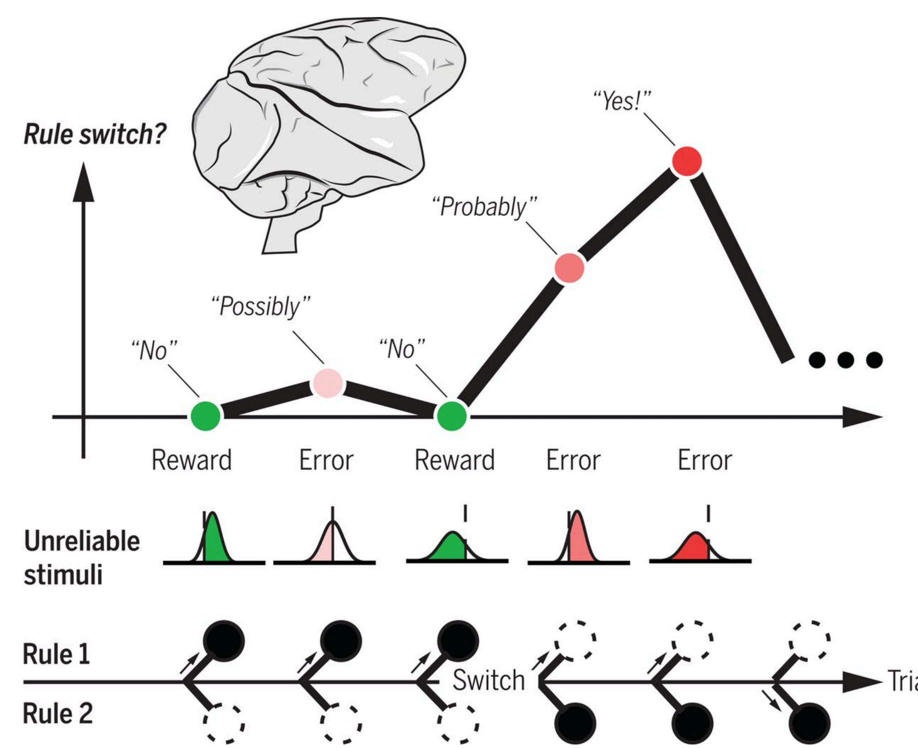
1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation: **object-centric, video foundation model**?

2. **<u>Cognitive:</u>** Hierarchy/modularization of timescales in dynamics?

Hierarchical reasoning by neural circuits in the frontal cortex

MORTEZA SARAFYAZD AND MEHRDAD JAZAYERI Authors Info & Affiliations

SCIENCE • 17 May 2019 • Vol 364, Issue 6441 • DOI: 10.1126/science.aav8911

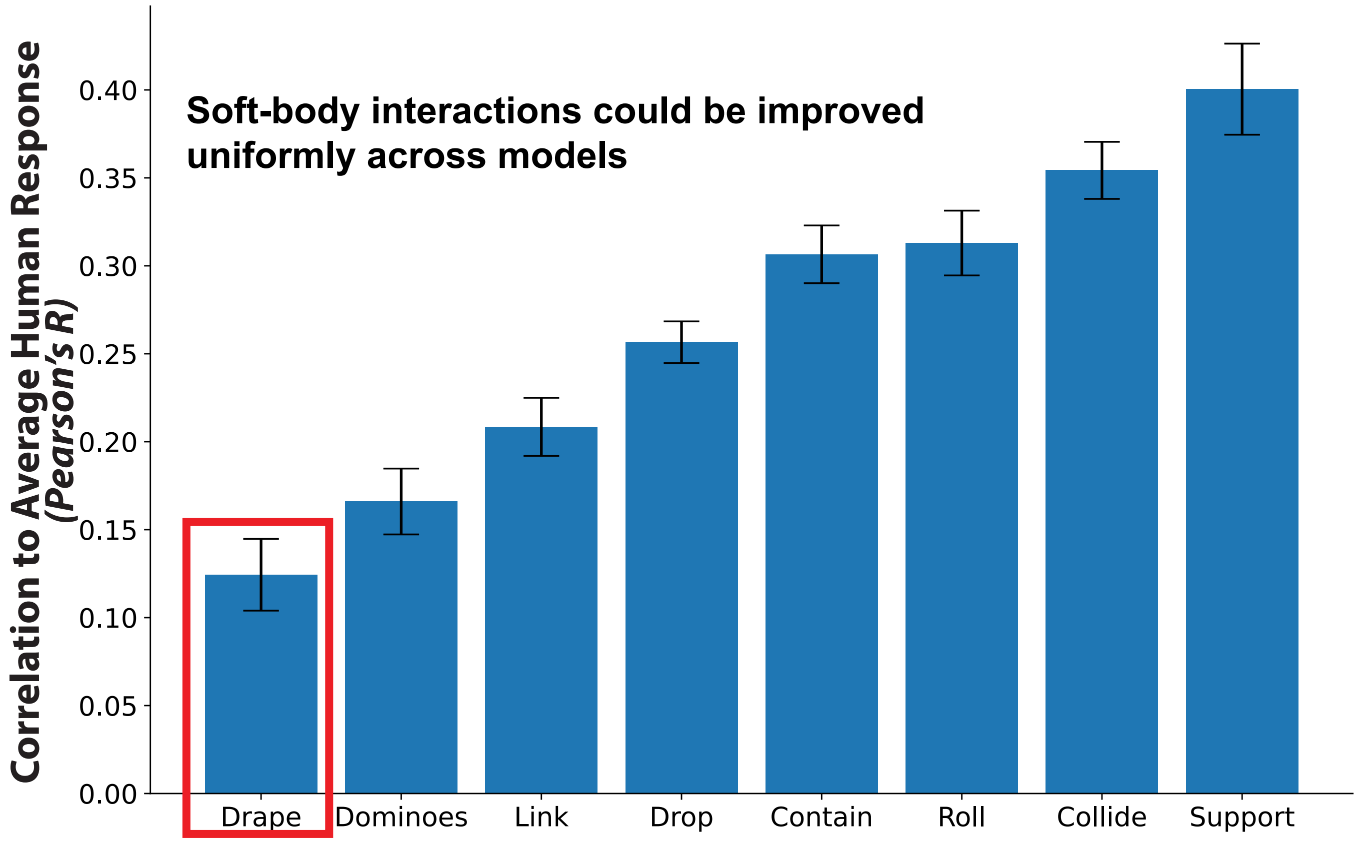


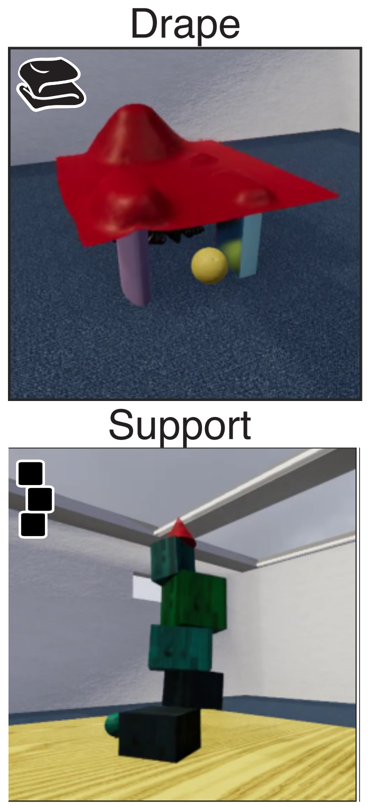
Future Directions: Learning Diverse Material Properties

1. **Sensory:** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation: **object-centric, video foundation model**?

2. **Cognitive:** Hierarchy/modularization of timescales in dynamics?

3. **Data:** More complex 2D and 3D scenes/real world objects





Outline

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

Heuristics for Interrogating Natural Intelligence

- **Incorporating Neuroscience Insights:**

- **Incorporating AI Insights:**

Heuristics for Interrogating Natural Intelligence

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

Heuristics for Interrogating Natural Intelligence

- **Incorporating Neuroscience Insights:**
  - **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:**
- **Incorporating AI Insights:**

Heuristics for Interrogating Natural Intelligence

- **Incorporating Neuroscience Insights:**
  - **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.
- **Incorporating AI Insights:**

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

- **Incorporating Neuroscience Insights:**
    - **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.
- **Incorporating AI Insights:**
    - 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.

Acknowledgements

Contact:
anayebi@mit.edu
@aran_nayebi


Nathan C.L. Kong


Chengxu Zhuang

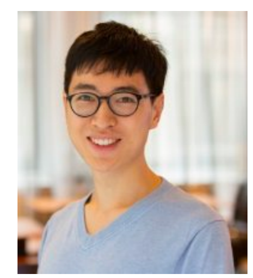

Justin L. Gardner


Anthony M. Norcia


Daniel Yamins


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

K. LISA YANG ICoN CENTER
Integrative Computational Neuroscience

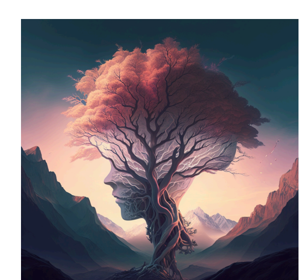
McGOVERN INSTITUTE

brain+cognitive sciences

SIMONS FOUNDATION

THE HELEN HAY WHITNEY FOUNDATION

JazLab


YangLab