We probably need a new idea…or two?

Aran Nayebi

K. Lisa Yang ICoN Postdoctoral Fellow

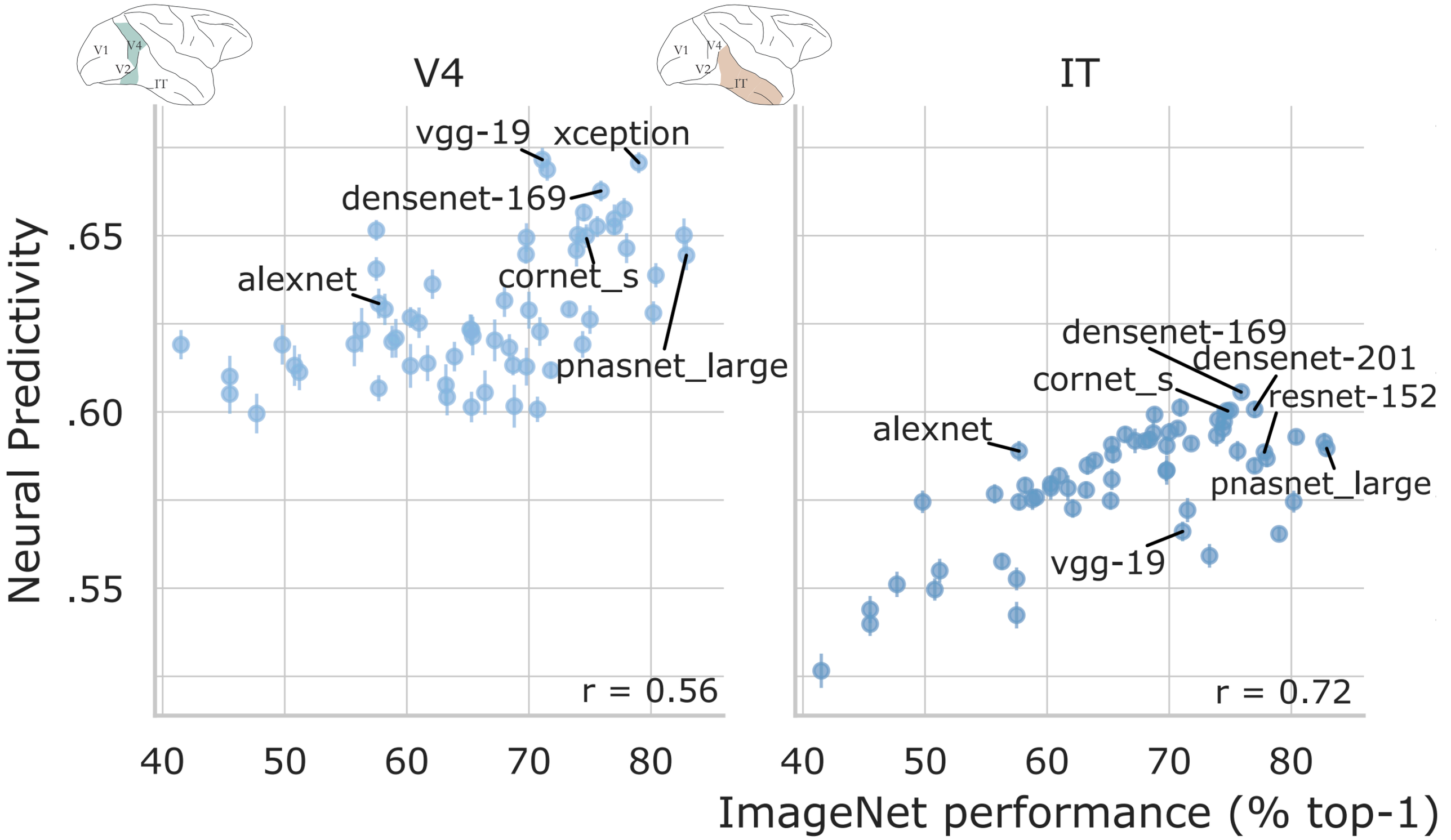
McGovern Institute, MIT

CCN 2023 GAC

2023.08.25

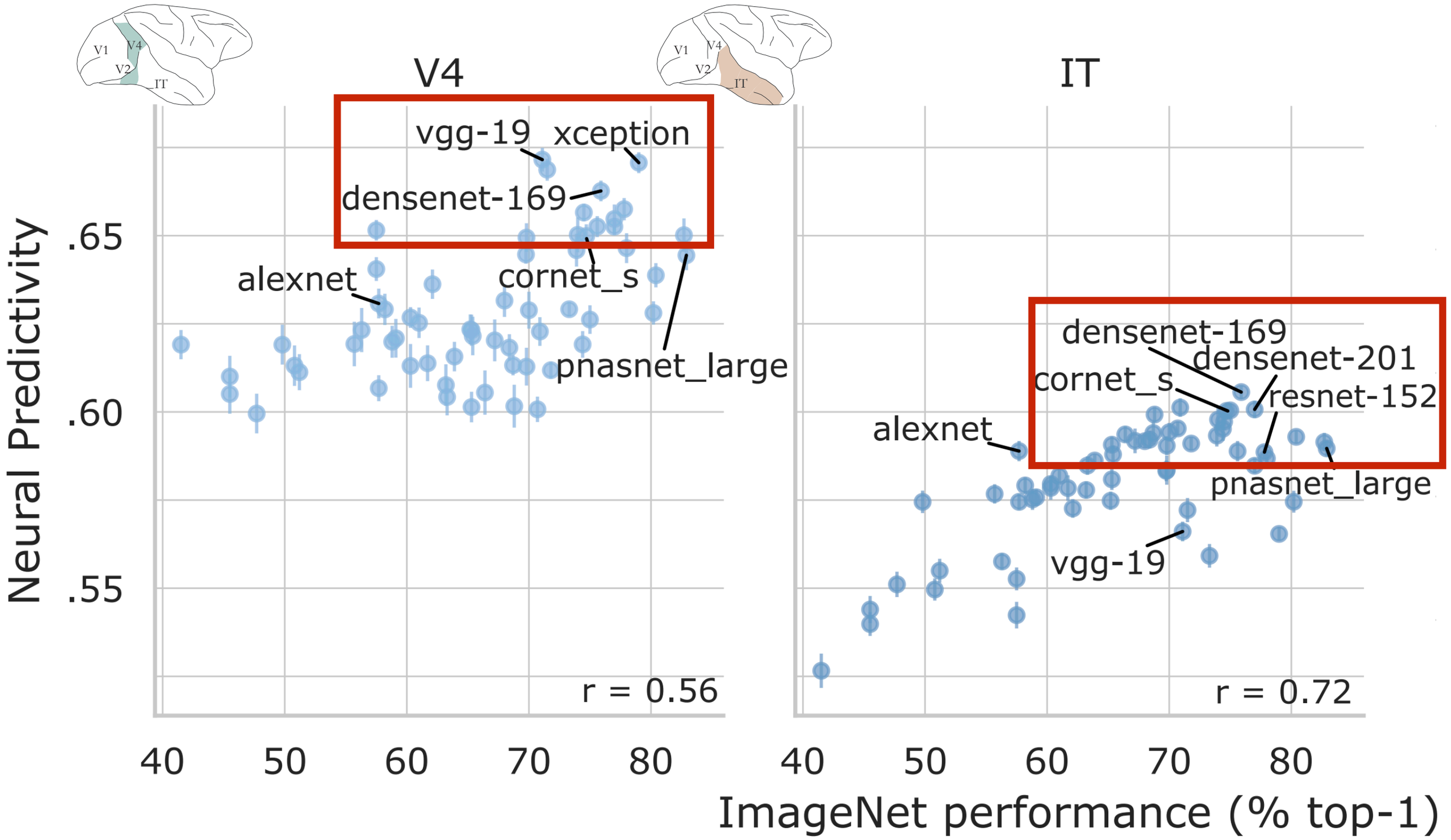
K. LISA YANG
ICoN CENTER
Integrative Computational Neuroscience

Similar predictivities among very different CNN architectures



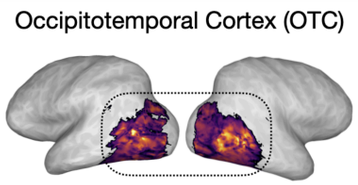
Schrimpf, Kubilius* et al. 2018*

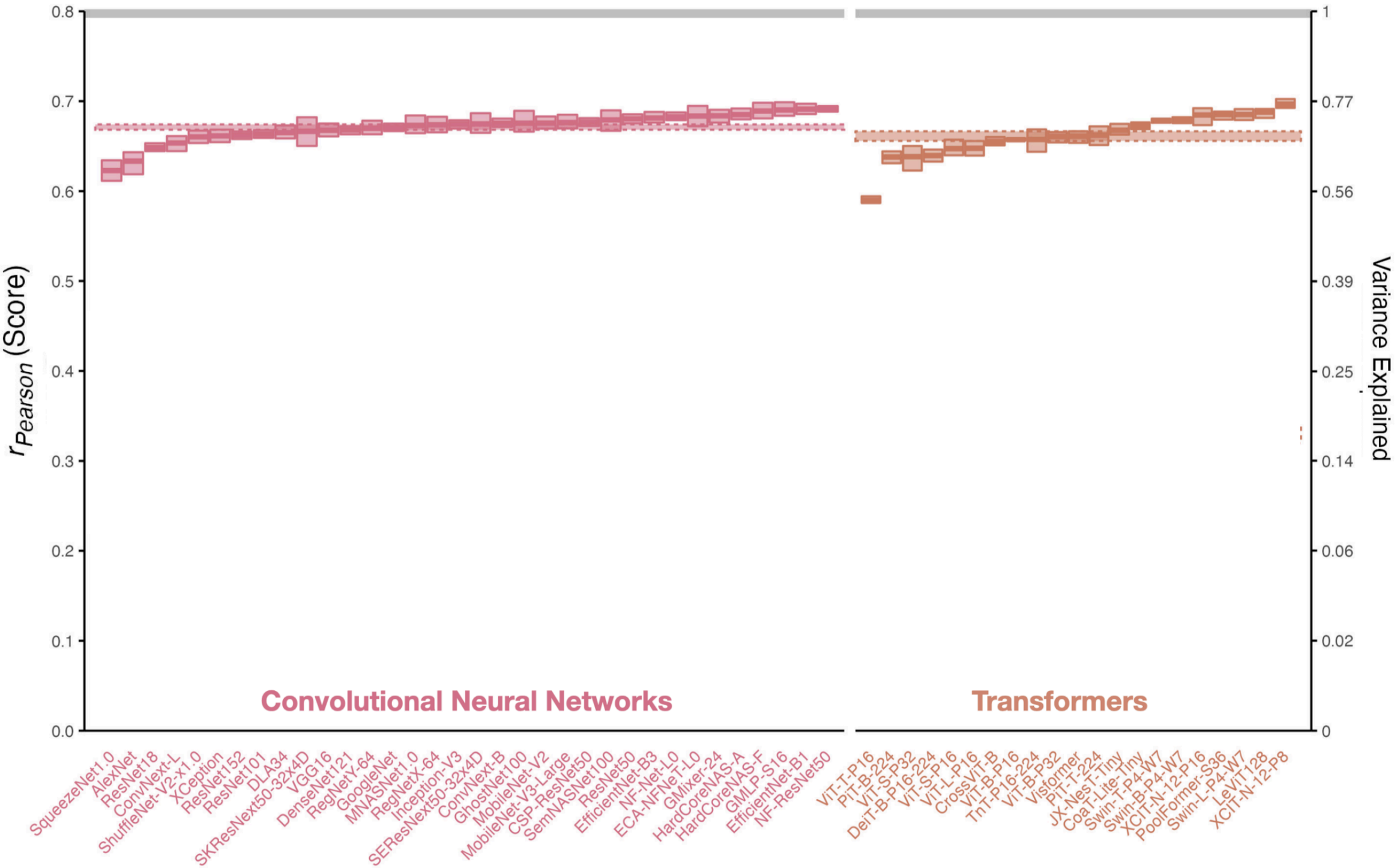
Similar predictivities among very different CNN architectures



Schrimpf, Kubilius* et al. 2018*

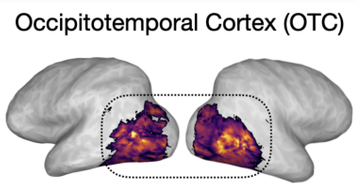
Similar predictivities between CNNs vs. Transformers

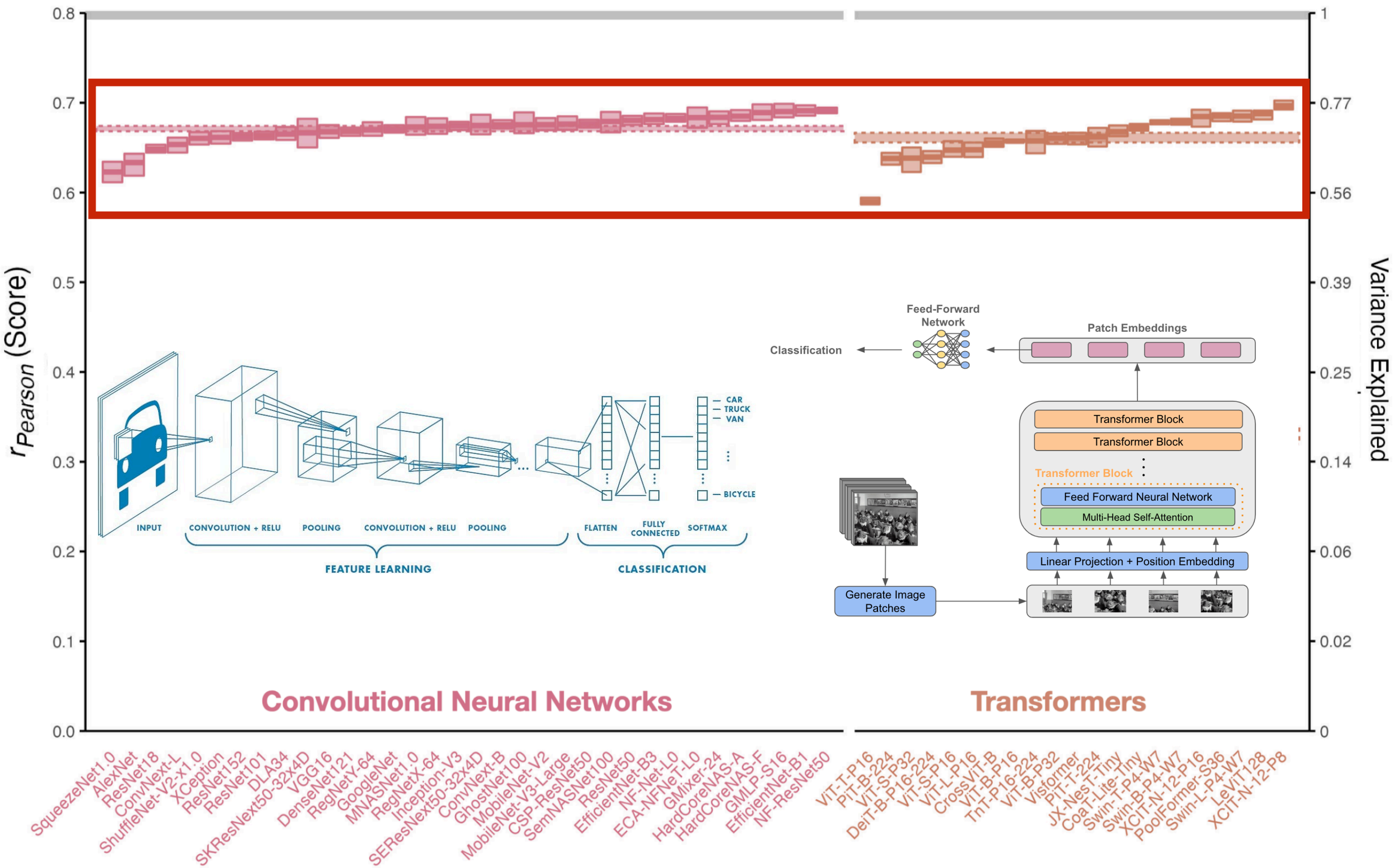




Conwell et al. 2023

Similar predictivities between CNNs vs. Transformers





Conwell et al. 2023

Why might this lack of distinction be?

Why might this lack of distinction be?

- **Metrics**
- **Brain Comparison Data**
- **Models**

Why might this lack of distinction be?

- **Metrics**
  - Correlation Based
  - Causal Manipulations
  - Mapping Transform Class

- **Brain Comparison Data**
  - Organism
  - Cognitive Tasks
  - Brain Area
  - Number of Units
  - Number of Conditions

- **Models**
  - Architecture
  - Loss Function
  - Dataset

Changing the Brain Comparison Data

- **Metrics**
  - *Correlation Based* (Regression)
  - Causal Manipulations
  - Mapping Transform Class

- **Brain Comparison Data**
  - *Organism* (Primates)
  - Cognitive Tasks
  - Brain Area
  - Number of Units
  - Number of Conditions

- **Models**
  - Architecture
  - Loss Function
  - Dataset

Changing the Brain Comparison Data

- **Metrics**
  - *Correlation Based* (Regression)
  - Causal Manipulations
  - Mapping Transform Class

Modify →

- **Brain Comparison Data**
  - *Organism* (Primates)
  - Cognitive Tasks
  - Brain Area
  - Number of Units
  - Number of Conditions

- **Models**
  - Architecture
  - Loss Function
  - Dataset

Changing the Brain Comparison Data

- **Metrics**
  - *Correlation Based* (Regression)
  - Causal Manipulations
  - Mapping Transform Class

Modify →

- **Brain Comparison Data**
  - *Organism* (Primates)
  - Cognitive Tasks (“Mental-Pong”)
  - Brain Area
  - Number of Units
  - Number of Conditions

- **Models**
  - Architecture
  - Loss Function
  - Dataset

Changing the Brain Comparison Data

- **Metrics**
  - *Correlation Based* (Regression)
  - Causal Manipulations
  - Mapping Transform Class

Modify →

- **Brain Comparison Data**
  - *Organism* (Primates)
  - Cognitive Tasks (“Mental-Pong”)
  - Brain Area (frontal cortex)
  - Number of Units
  - Number of Conditions

- **Models**
  - Architecture
  - Loss Function
  - Dataset

New Models Needed

- **Metrics**
  - *Correlation Based* (Regression)
  - Causal Manipulations
  - Mapping Transform Class

Modify →

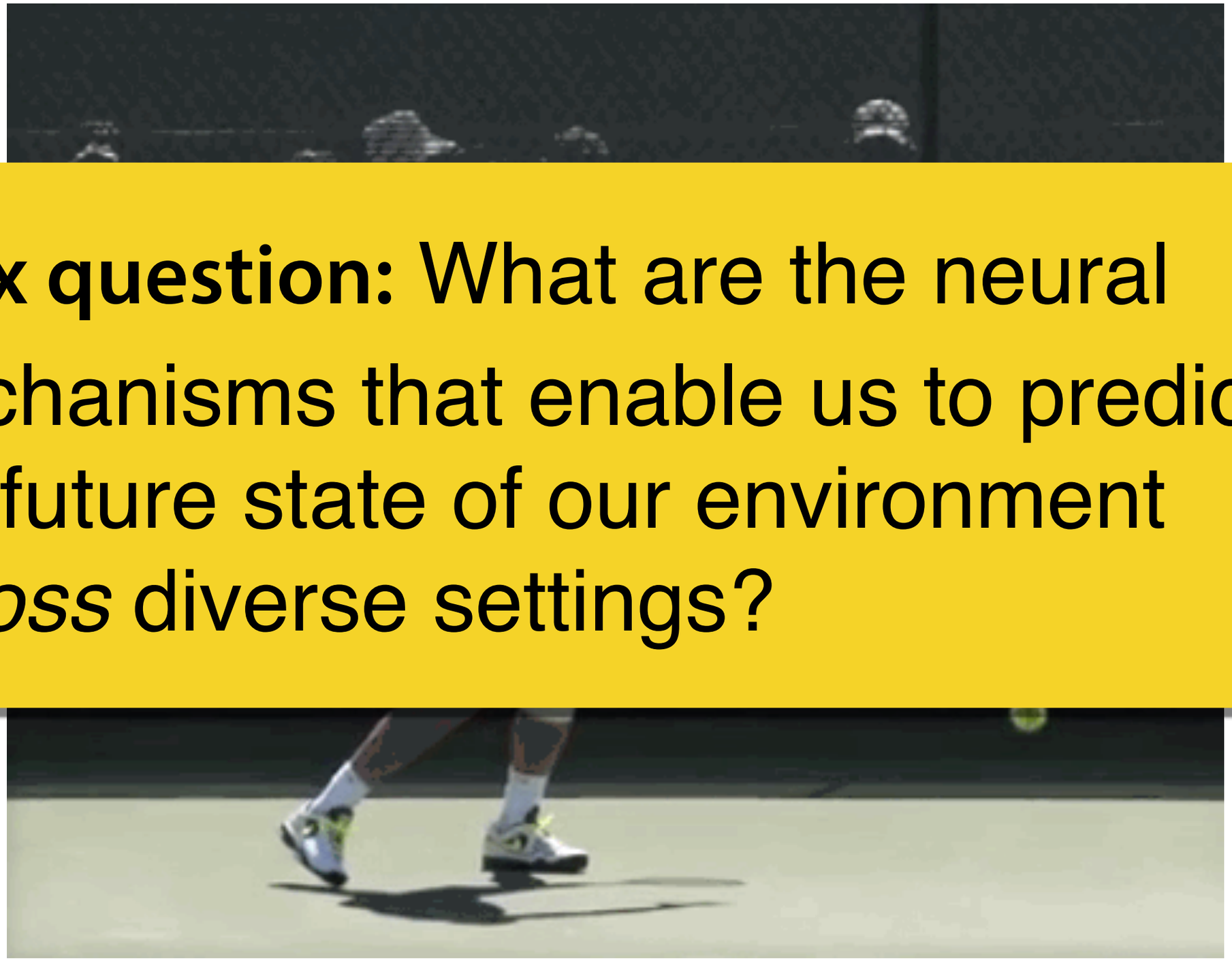
- **Brain Comparison Data**
  - *Organism* (Primates)
  - Cognitive Tasks ("Mental-Pong")
  - Brain Area (frontal cortex)
  - Number of Units
  - Number of Conditions

New ideas needed →

- **Models**
  - Architecture
  - Loss Function
  - Dataset

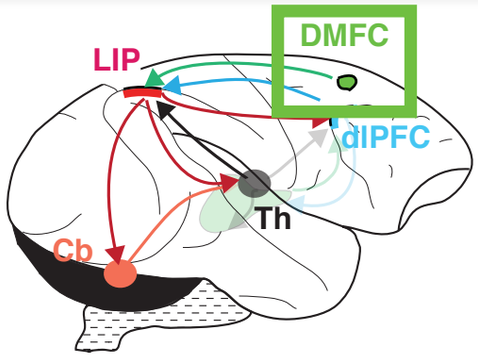
We do a lot more than passive viewing…

We do a lot more than passive viewing…



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

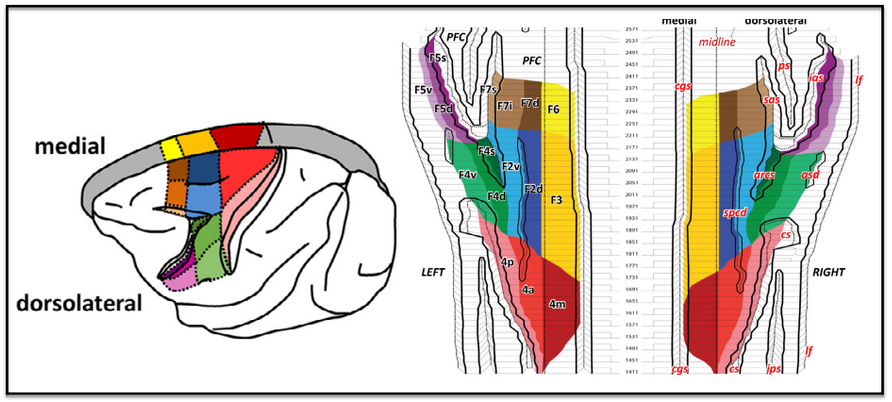
Mental-Pong Task & Macaque Neurophysiology



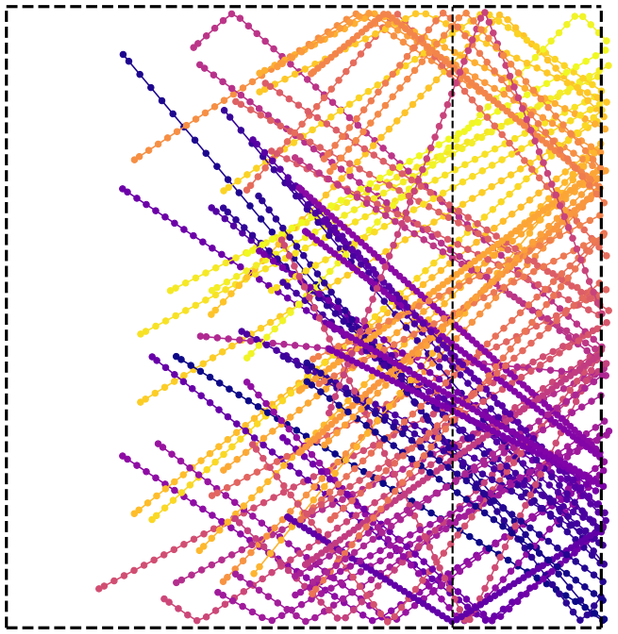
Fronto-Parietal Network

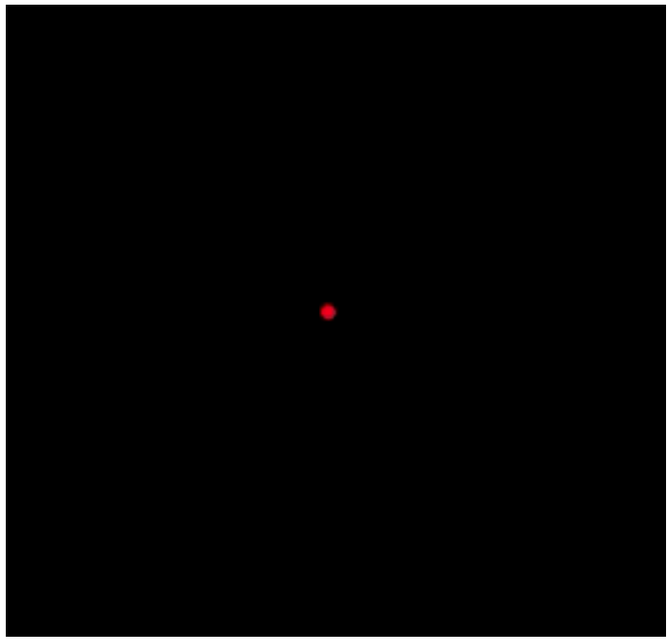
Dorsomedial frontal cortex (DMFC)

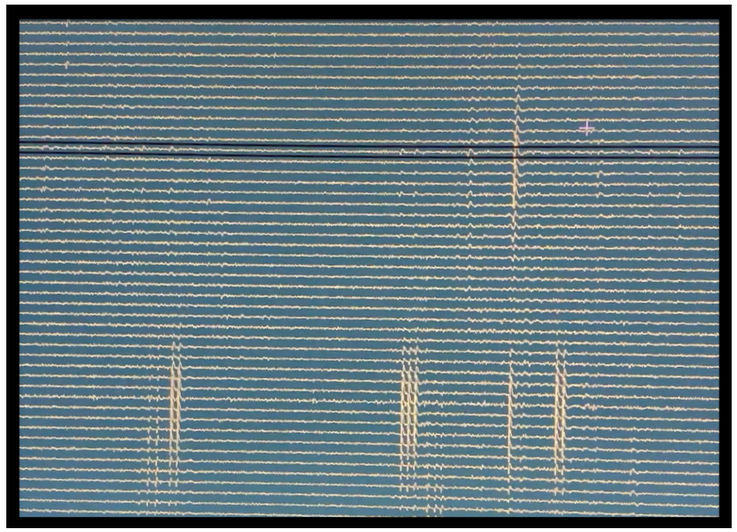
Monkey P *Monkey M*



79 conditions





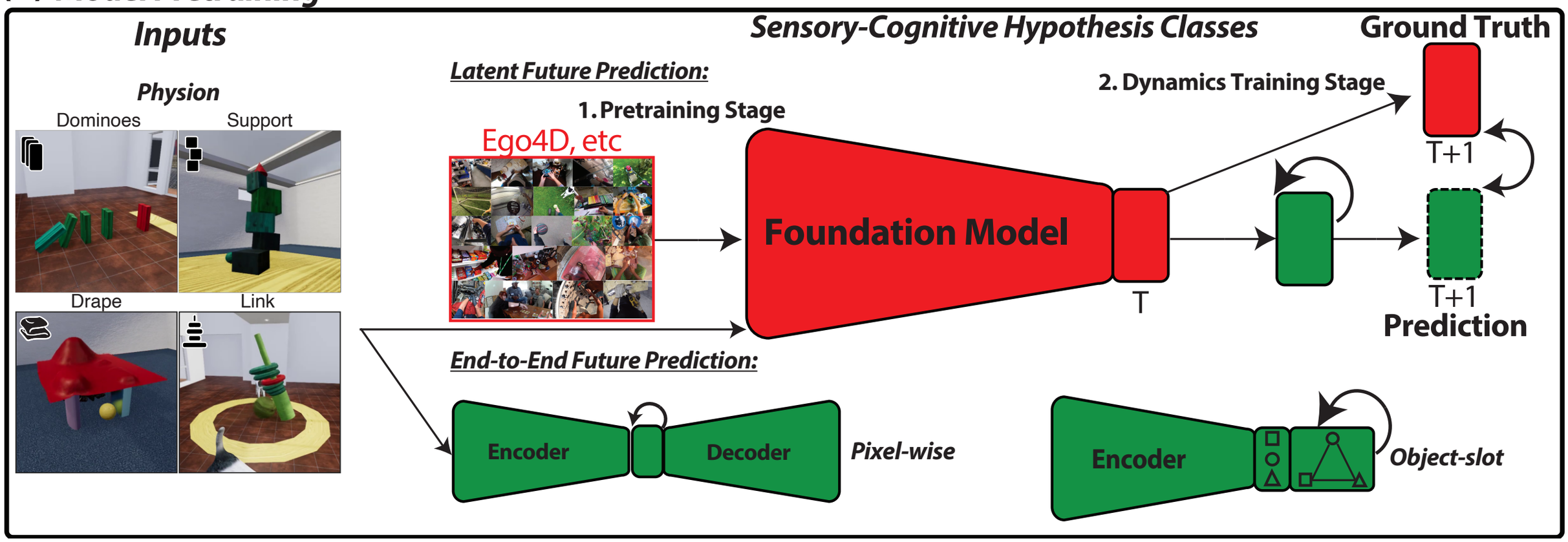


- *Data from two male adult monkeys*
- *79 subsampled M-Pong conditions*
- *64 channel v-probe (monkey P) and 384-channel Neuropixel probe (monkey M)*
- *Total of 1889 stable & reliable neurons recorded from DMFC*

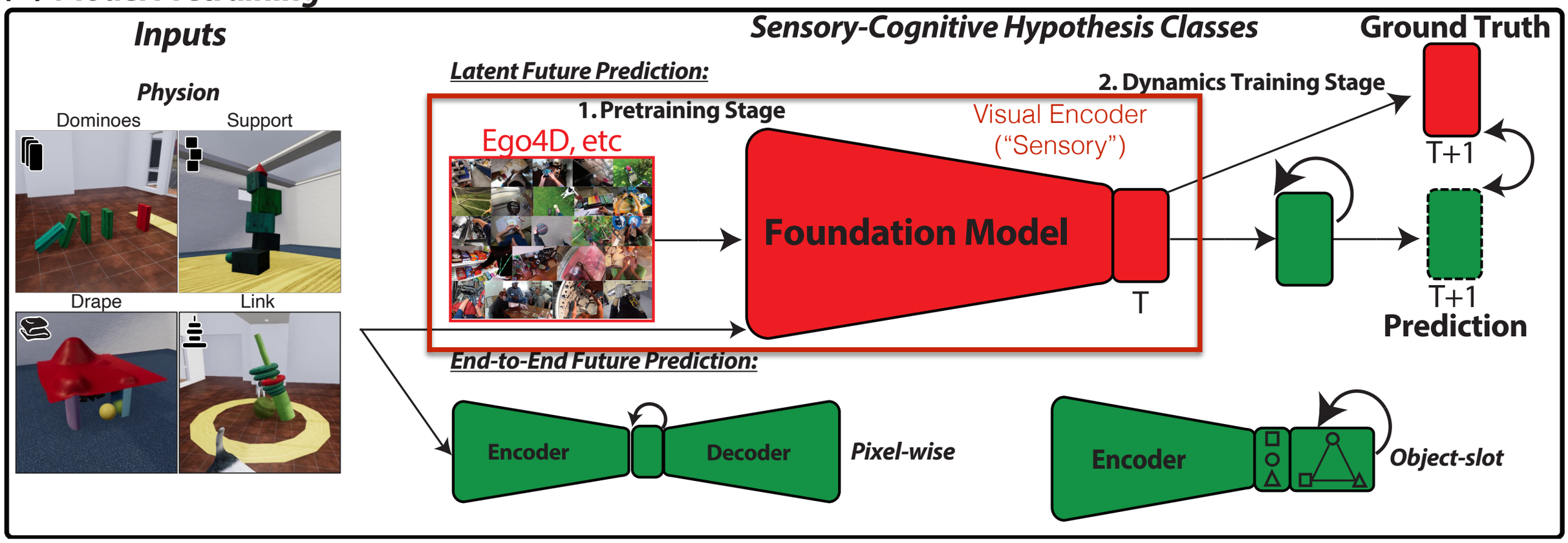
Rishi Rajalingham

Overall Approach

(A) *Model Pretraining*

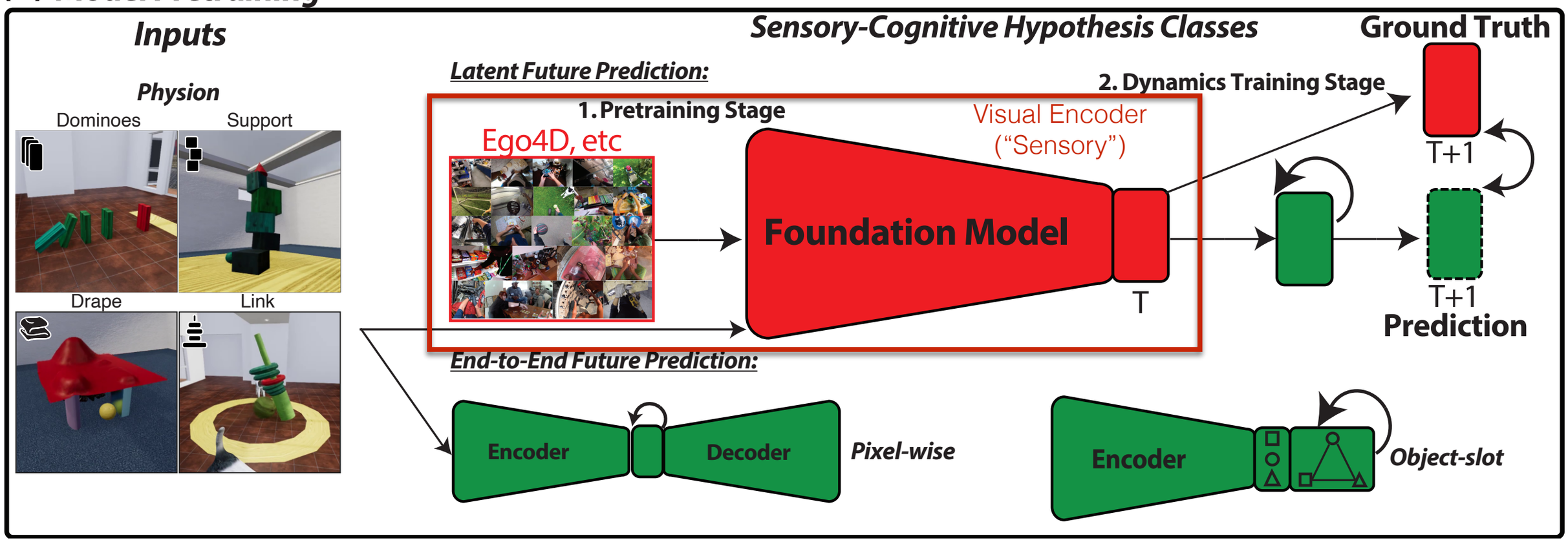


Overall Approach: Foundation Models



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

Overall Approach: Foundation Models

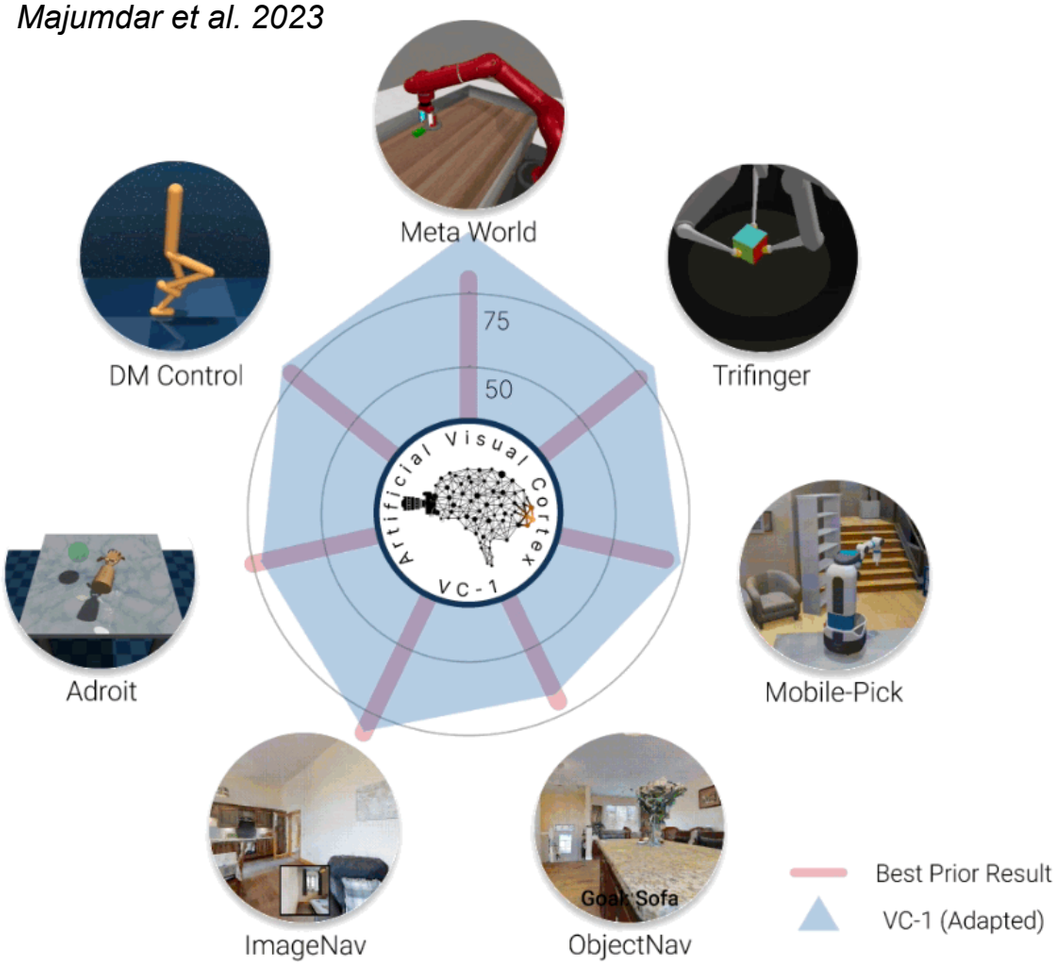


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

Emphasis on *reusability*!

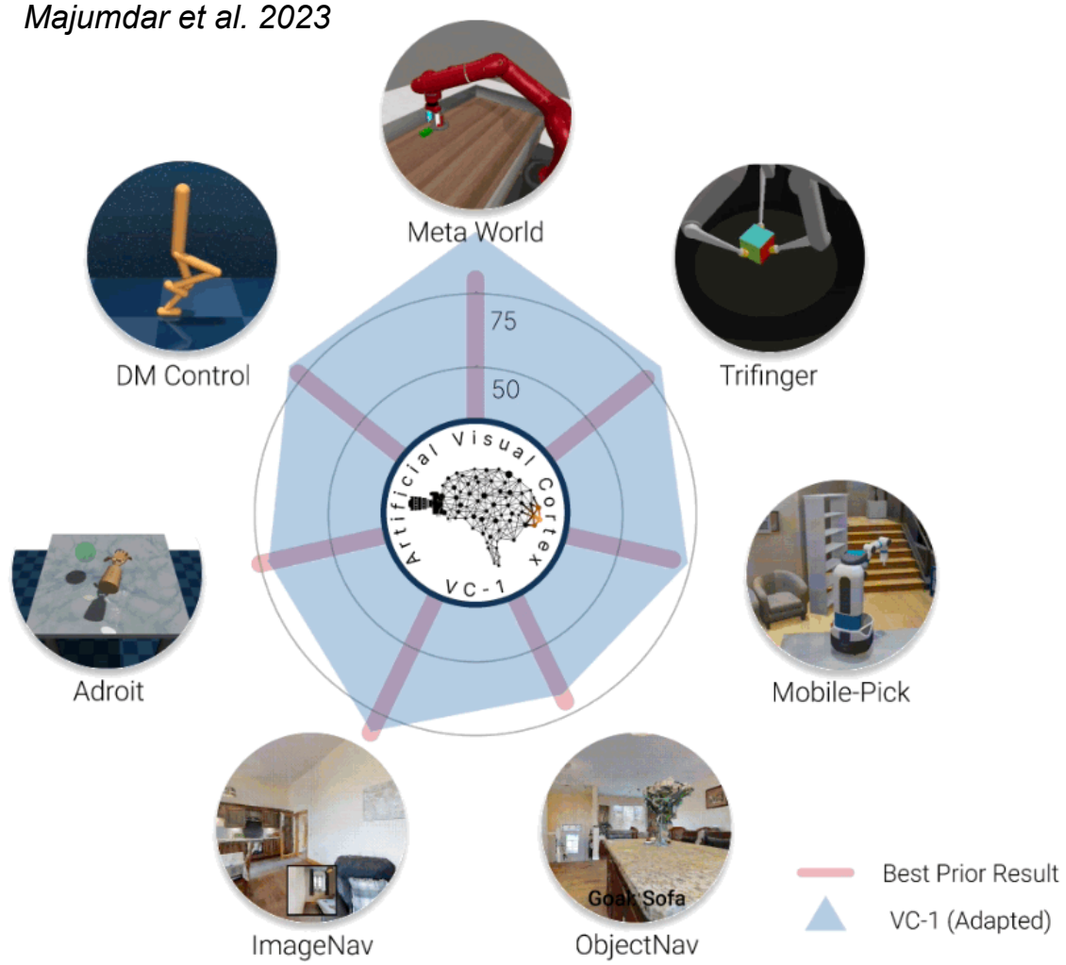
Overall Approach: Foundation Models

Majumdar et al. 2023



Overall Approach: Foundation Models

Majumdar et al. 2023



CortexBench

Ego4D: A massive-scale egocentric dataset

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

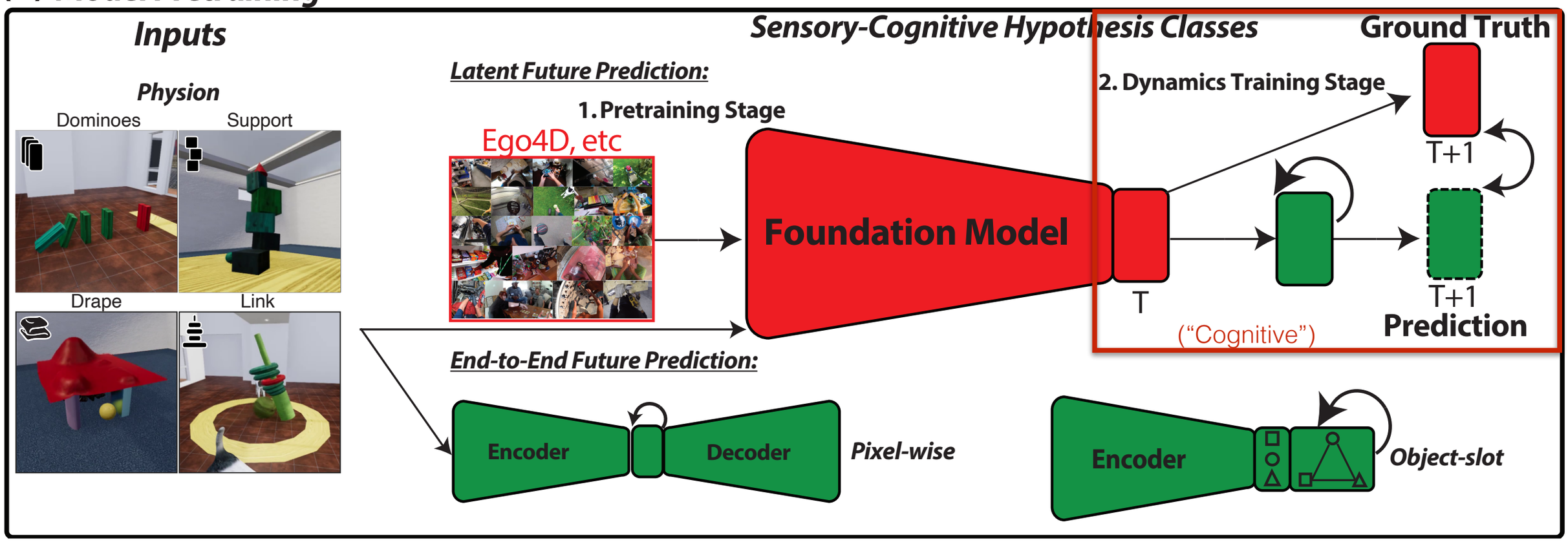
931 participants from 74 worldwide locations

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

Grauman et al. 2022

Ego4D: everyday activity around the world

Overall Approach: Foundation Models + Dynamics



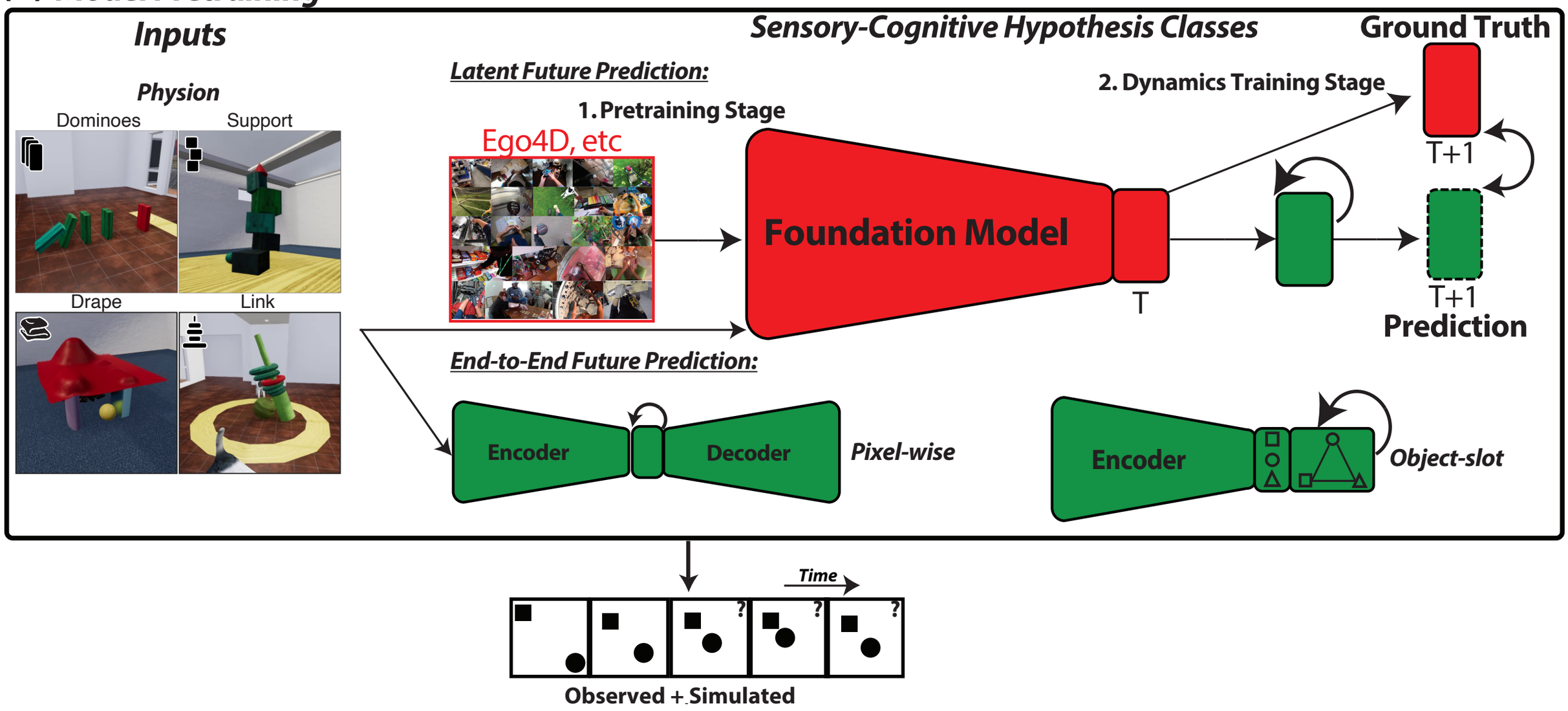
Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

Emphasis on *reusability*!

Leverage these dynamics to do explicit physical simulation

Overall Approach

(A) *Model Pretraining*



Overall Approach: Model Evaluations

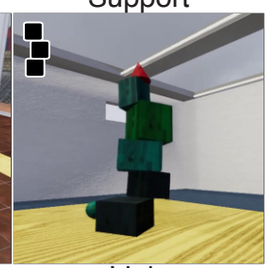
(A) *Model Pretraining*

Inputs

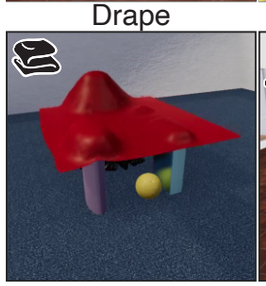
Physion

Dominoes

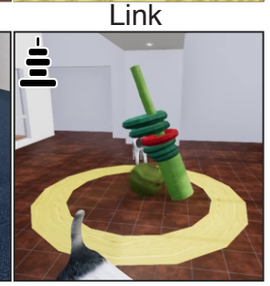
Support



Drape



Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T+1

T

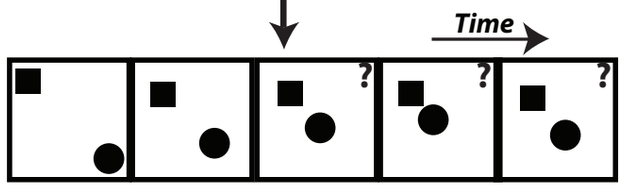
T+1

Prediction

End-to-End Future Prediction:

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

Encoder ***Object-slot***



Observed + Simulated

(B) *Model Evaluations*

Overall Approach: Model Evaluations (Human Behavior)

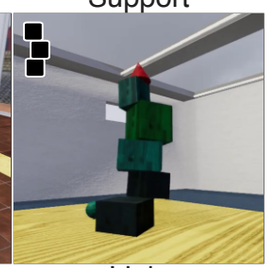
(A) *Model Pretraining*

Inputs

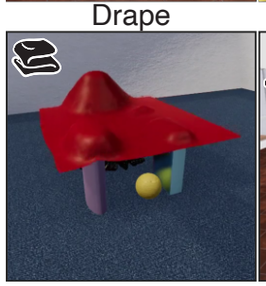
Physion

Dominoes

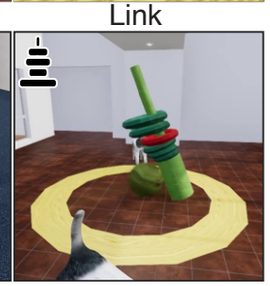
Support



Drape



Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T

T+1

T+1

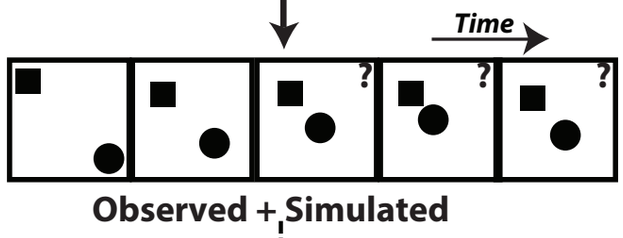
Prediction

End-to-End Future Prediction:

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

Encoder ***Object-slot***

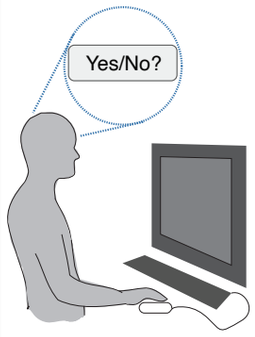
Time



Observed + Simulated

(B) *Model Evaluations*

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



Observed Stimuli

Unobserved Outcome

true label

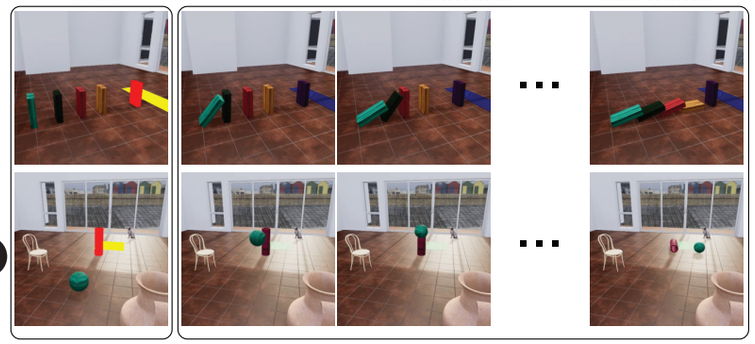
Time

cue

stimulus

last frame

Example Scenarios



NO

acc. = 0.89

YES

acc. = 0.96

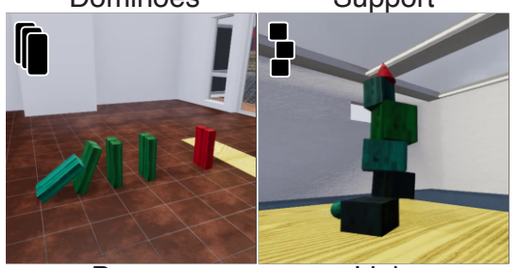
Overall Approach: Model Evaluations (Macaque Physiology)

(A) *Model Pretraining*

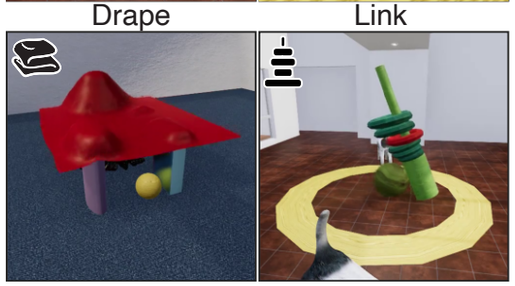
Inputs

Physion

Dominoes | Support



Drape | Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T

T+1

T+1

Prediction

End-to-End Future Prediction:

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

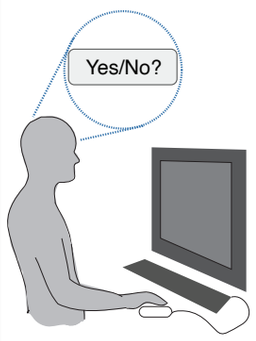
Encoder ***Object-slot***

Time

Observed + Simulated

(B) *Model Evaluations*

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



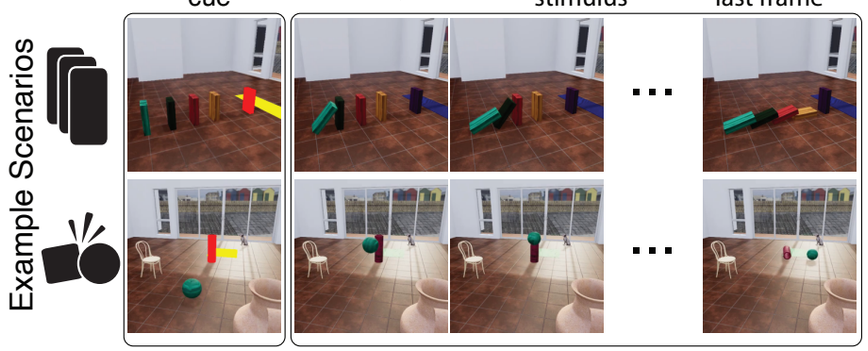
Observed Stimuli — Unobserved Outcome

Time

cue — stimulus — last frame

true label

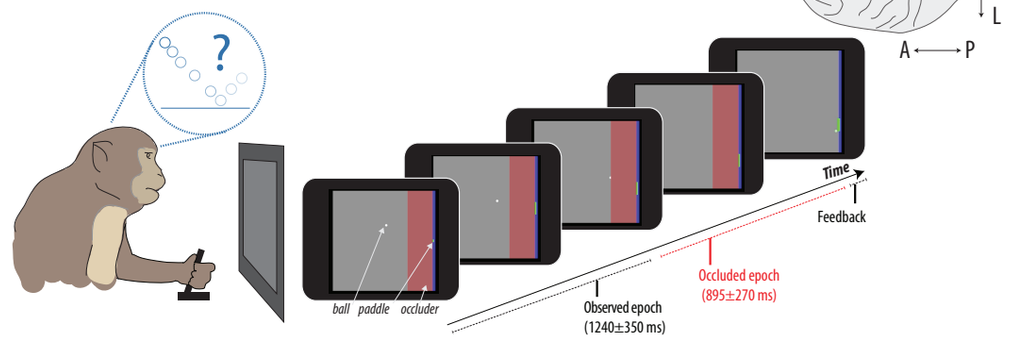
Example Scenarios



NO
acc. = 0.89

YES
acc. = 0.96

2. *Macaque Neurophysiology: Mental-Pong*



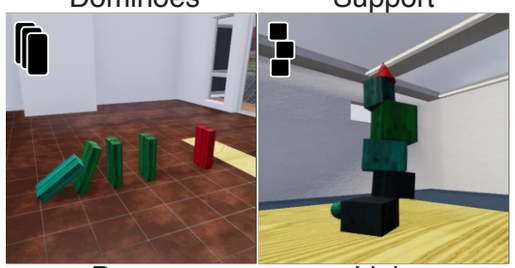
Model Evaluations: Macaque Neurophysiology

(A) *Model Pretraining*

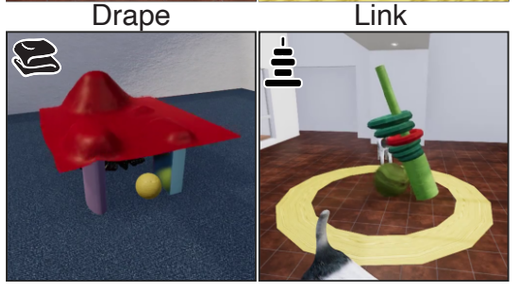
Inputs

Physion

Dominoes | Support



Drape | Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

T

2. Dynamics Training Stage

T+1

T+1

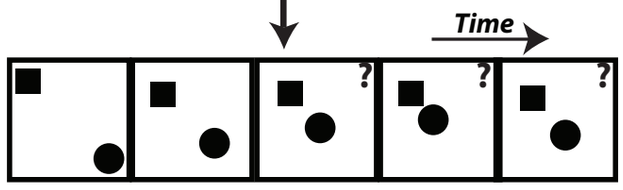
Prediction

End-to-End Future Prediction:

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

Encoder ***Object-slot***

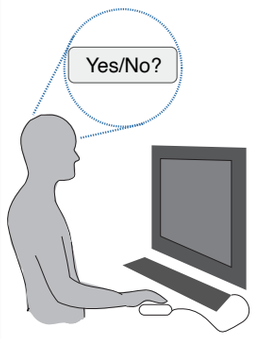
Time



Observed + Simulated

(B) *Model Evaluations*

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

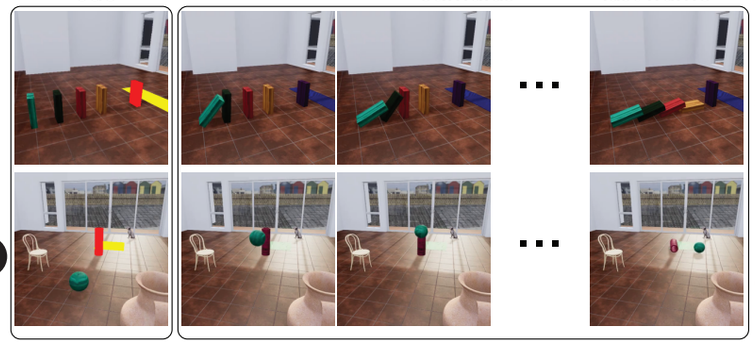


Observed Stimuli | Unobserved Outcome

Time

cue | stimulus | last frame | true label

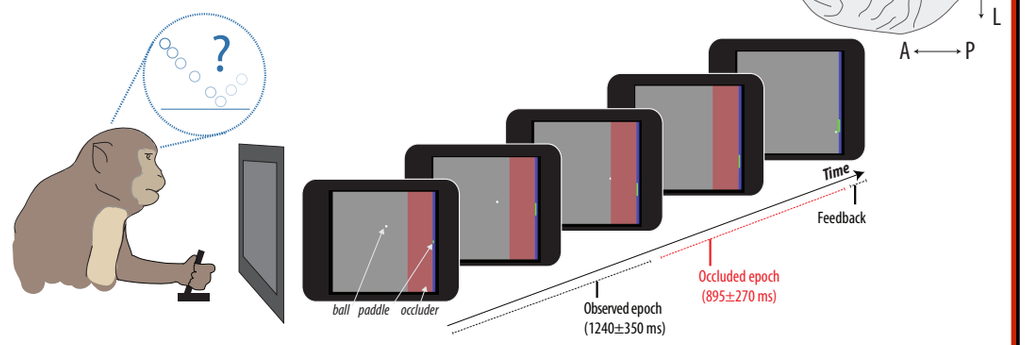
Example Scenarios



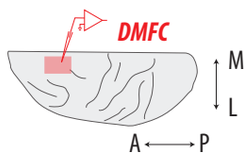
NO
acc. = 0.89

YES
acc. = 0.96

2. *Macaque Neurophysiology: Mental-Pong*



Neural response predictivity strongly separates models



Inter-animal Consistency

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise
64x64
128x128
64x64
64x64+ RandAugment
Object -slot
Small
Medium
Large

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

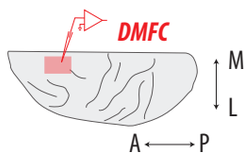
Ball Position
Ball Velocity
Ball Position + Velocity

SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP VIP VC-1 R3M

End-to-End

Latent Future Prediction

Neural response predictivity strongly separates models



Inter-animal Consistency

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Pixel-wise
64x64
128x128
64x64
64x64+ RandAugment
Object -slot
Small
Medium
Large

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

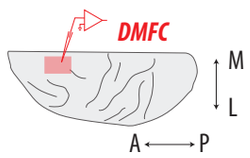
Ball Position
Ball Velocity
Ball Position + Velocity

SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP VIP VC-1 R3M

End-to-End

Latent Future Prediction

Neural response predictivity strongly separates models



Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Inter-animal Consistency

Pixel-wise
64x64
128x128
64x64
64x64+ RandAugment
Object -slot
Small
Medium
Large

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

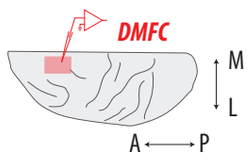
Ball Position
Ball Velocity
Ball Position + Velocity

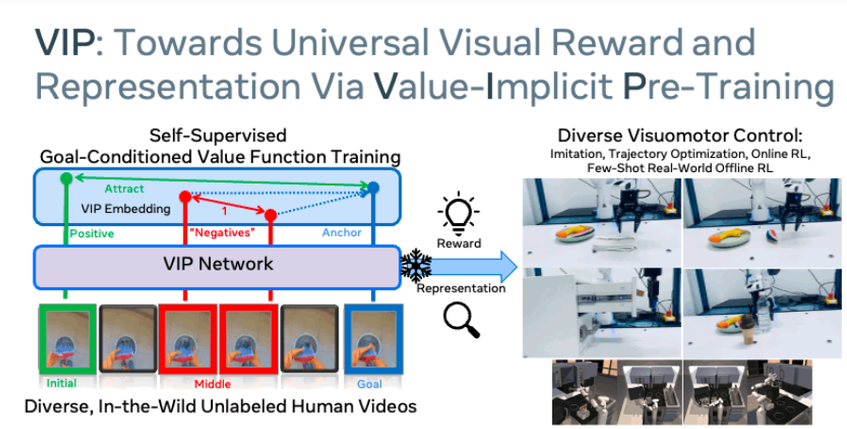
SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP VIP VC-1 R3M

End-to-End

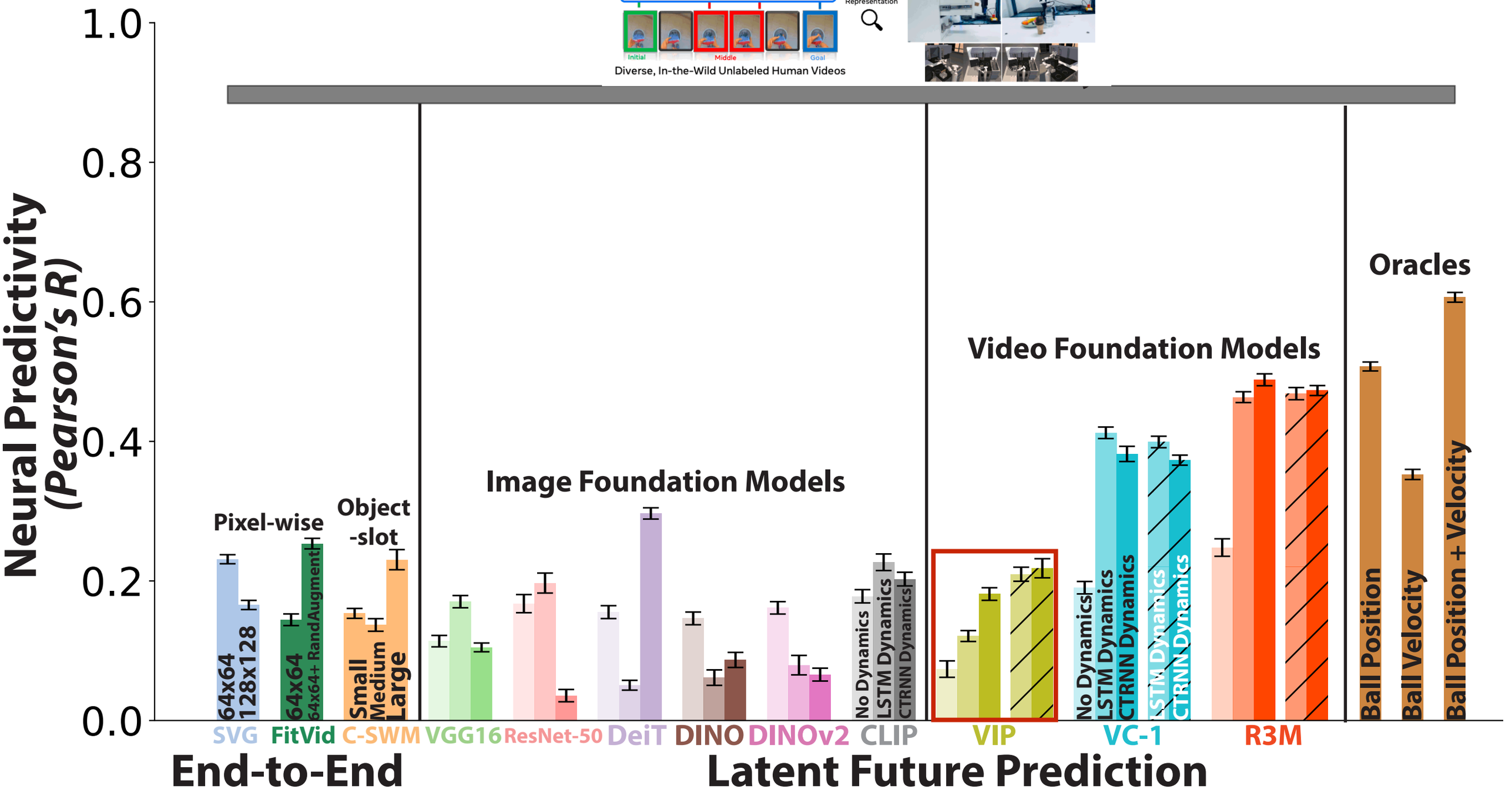
Latent Future Prediction

Neural response predictivity strongly separates models

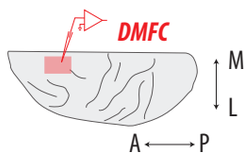




Ma et al. 2023



Neural response predictivity strongly separates models



Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Inter-animal Consistency

Pixel-wise
64x64
128x128
64x64
64x64+ RandAugment
Object-slot
Small
Medium
Large

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

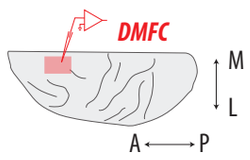
Ball Position
Ball Velocity
Ball Position + Velocity

SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP VIP VC-1 R3M

End-to-End

Latent Future Prediction

Neural response predictivity strongly separates models



Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Inter-animal Consistency

Most useful for image segmentation, classification, etc

Image Foundation Models

Most useful for robotic manipulation tasks

Video Foundation Models

Oracles

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

Object-slot
Small
Medium
Large

No Dynamics
LSTM Dynamics
CTRNN Dynamics

No Dynamics
LSTM Dynamics
CTRNN Dynamics

LSTM Dynamics
CTRNN Dynamics

Ball Position
Ball Velocity
Ball Position + Velocity

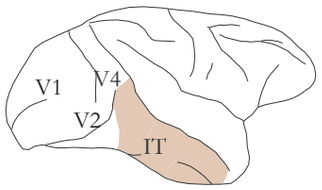
SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP VIP VC-1 R3M

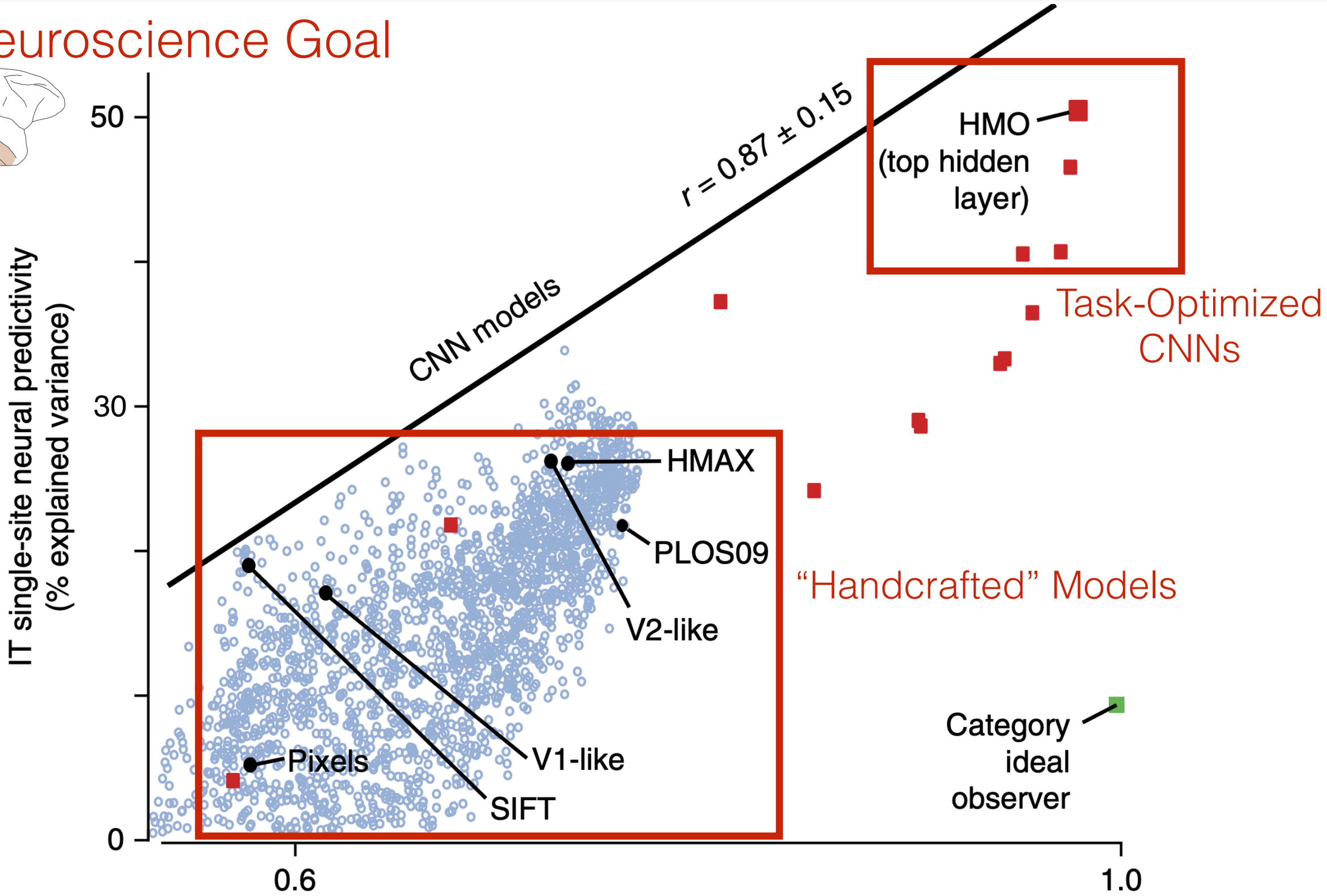
End-to-End

Latent Future Prediction

Prior Results in Inferior Temporal (IT) Cortex

A Neuroscience Goal





Yamins, Hong* et al. 2014*

An AI Goal

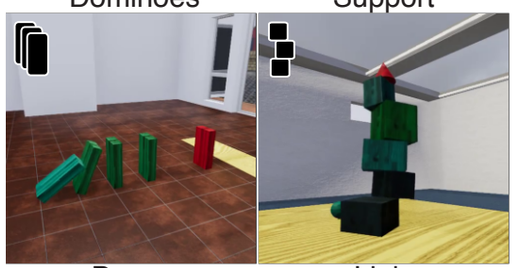
Model Evaluations: Object Contact Prediction (OCP)

(A) *Model Pretraining*

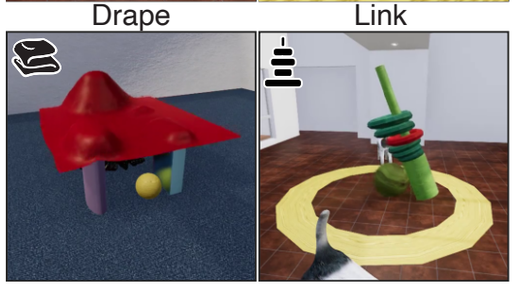
Inputs

Physion

Dominoes | Support



Drape | Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T

T+1

T+1

Prediction

End-to-End Future Prediction:

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

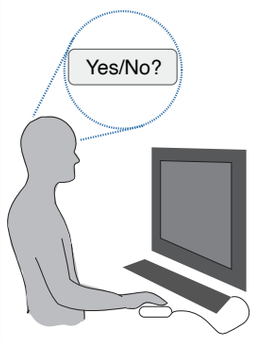
Encoder ***Object-slot***

Time

Observed + Simulated

(B) *Model Evaluations*

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



Observed Stimuli

Unobserved Outcome

true label

Time

cue

stimulus

last frame

Example Scenarios

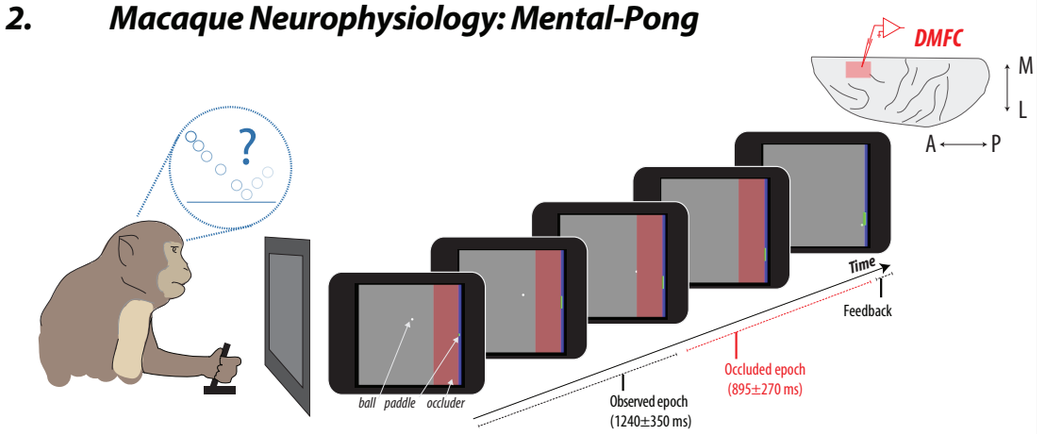
NO

acc. = 0.89

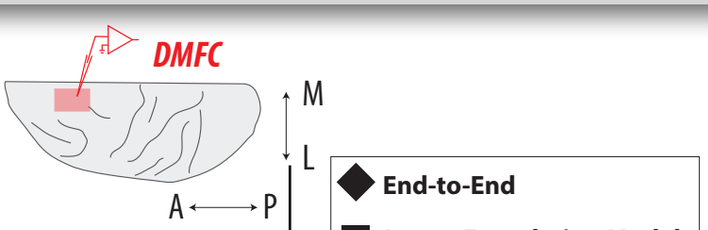
YES

acc. = 0.96

2. *Macaque Neurophysiology: Mental-Pong*



Video Foundation Models Can Better Match Both



End-to-End
Image Foundation Models
Video Foundation Models

Better Models

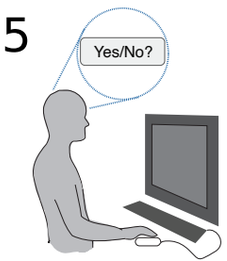
VC-1+LSTM

VC-1+CTRNN

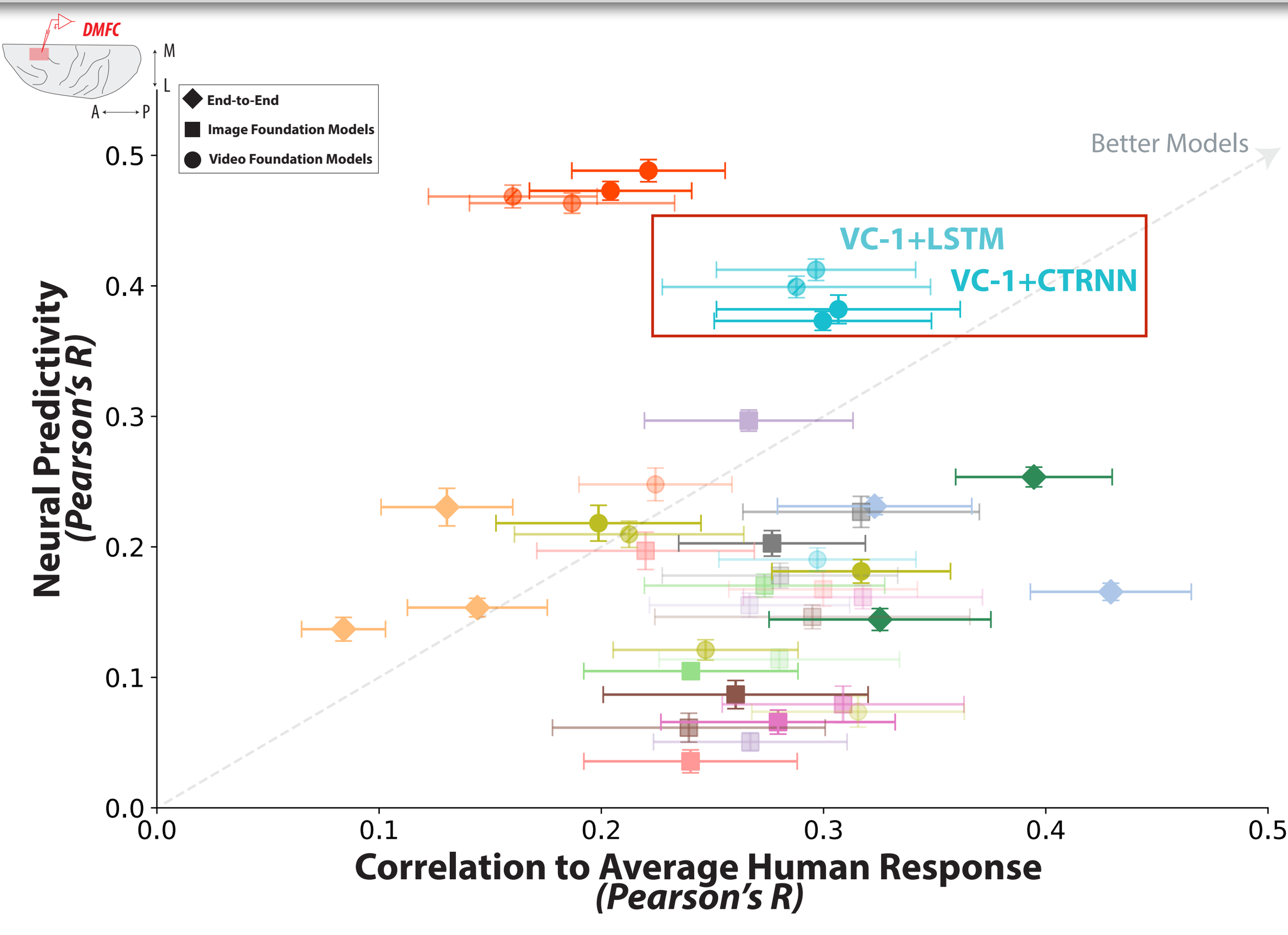
Neural Predictivity (Pearson's R)

0.0
0.1
0.2
0.3
0.4
0.5

Correlation to Average Human Response (Pearson's R)



Video Foundation Models Can Better Match Both



Takeaways so far…

1. Mental simulation appears to be primarily relevant to predicting the **future** state of the environment in a suitable **latent** space.

Takeaways so far…

1. Mental simulation appears to be primarily relevant to predicting the **future** state of the environment in a suitable **latent** space.

2. In particular, this latent space is highly constrained -- it doesn't appear to 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.

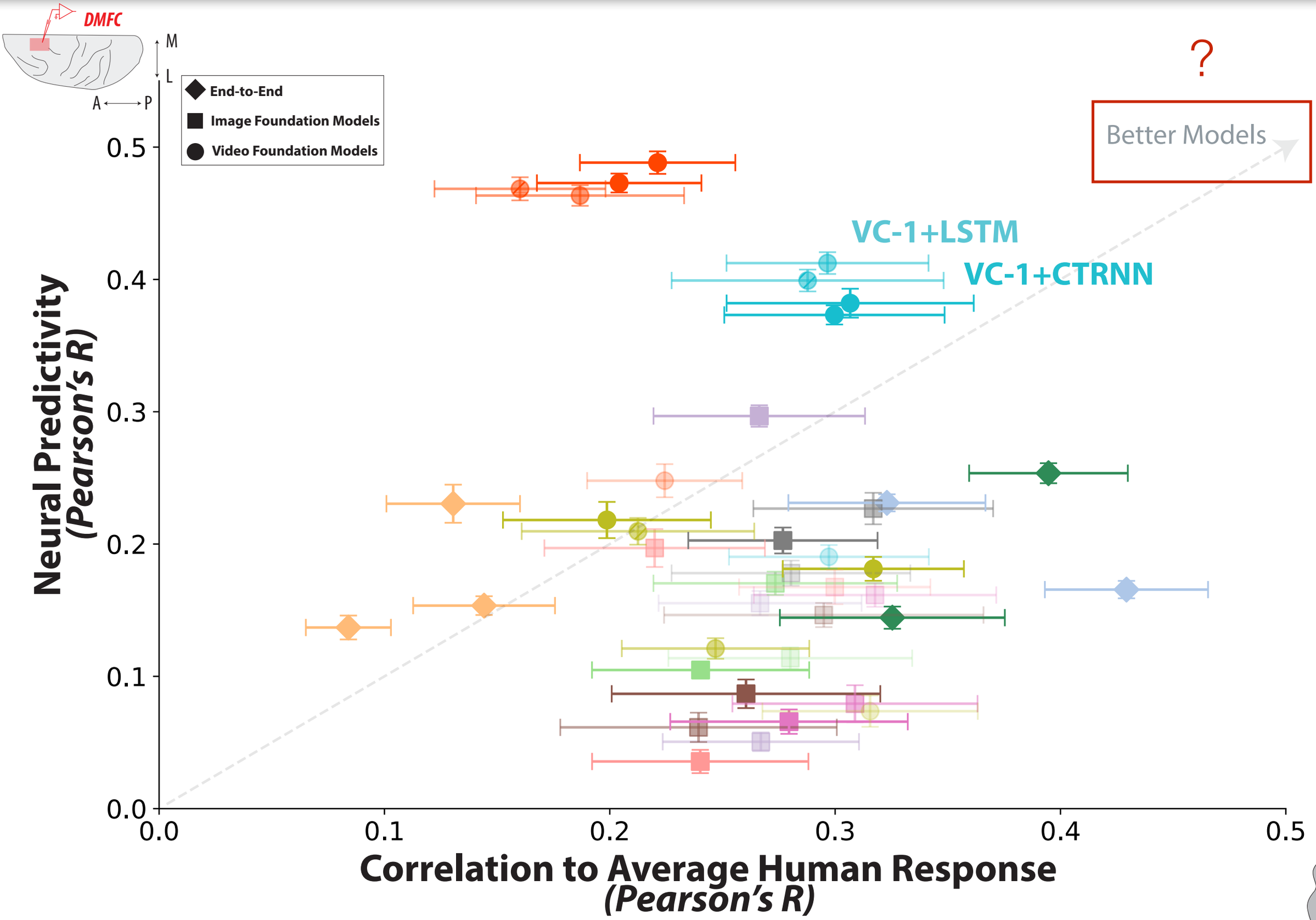
Takeaways so far…

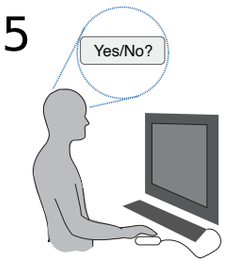
1. Mental simulation appears to be primarily relevant to predicting the **future** state of the environment in a suitable **latent** space.

2. In particular, this latent space is highly constrained -- it doesn't appear to 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.

3. So far a correspondence between the ability to predict neural & behavioral responses, and developing useful representations for Embodied AI more generally (rather than classic computer vision tasks e.g. classification, segmentation, etc).

Future Directions



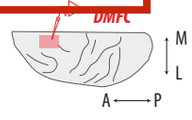


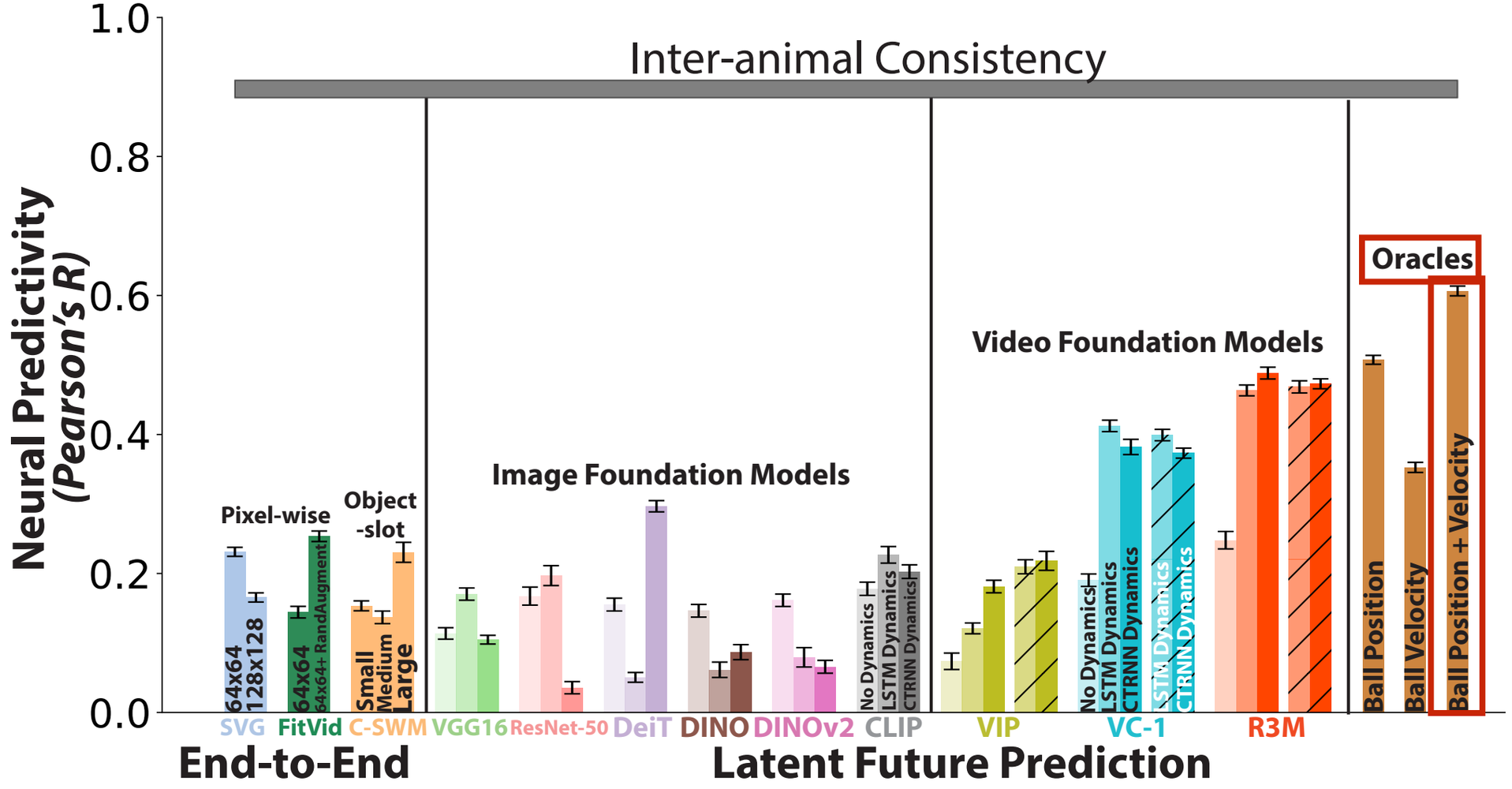
Future Directions

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

Future Directions

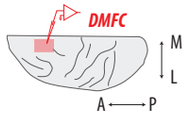
1. **Sensory:** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation:

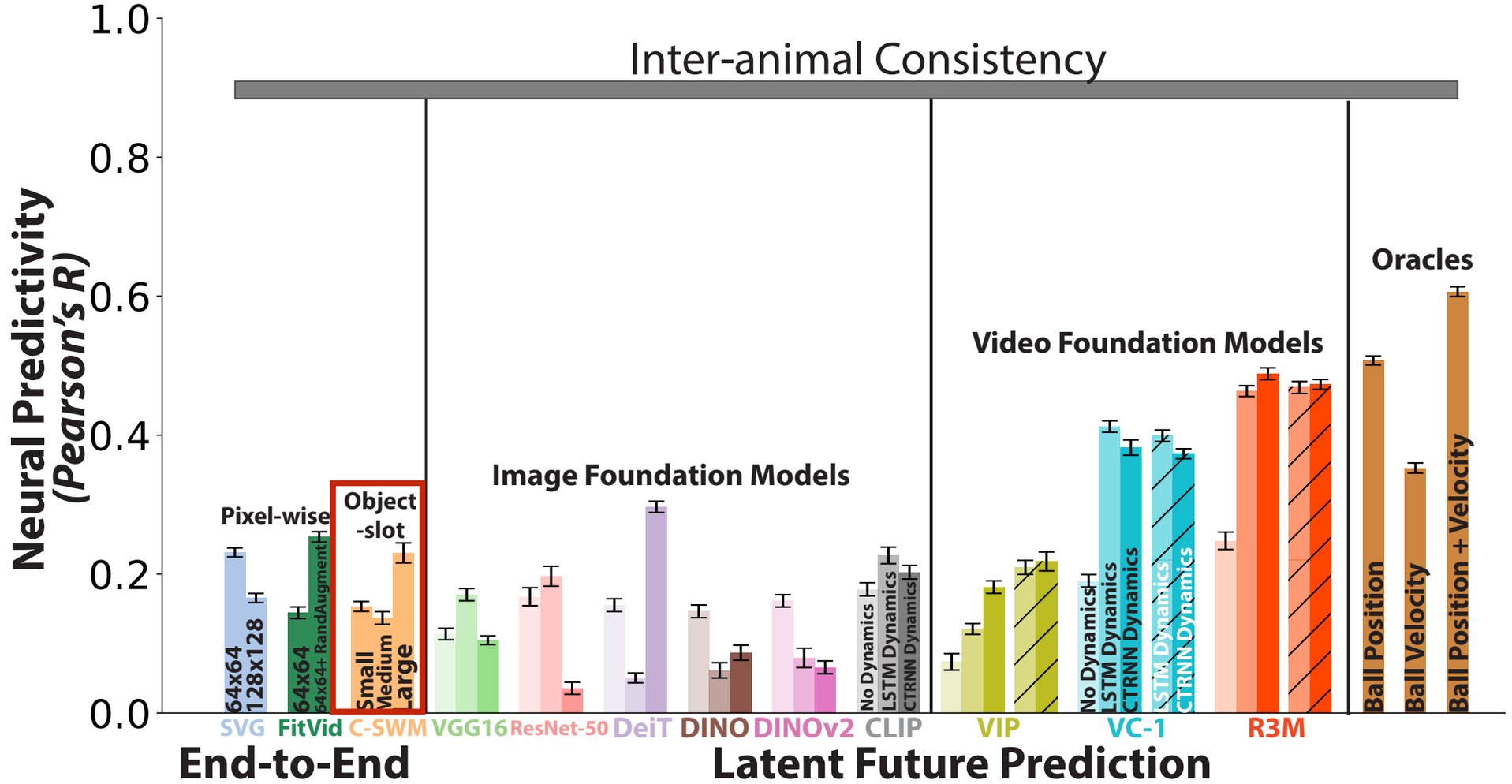




Future Directions

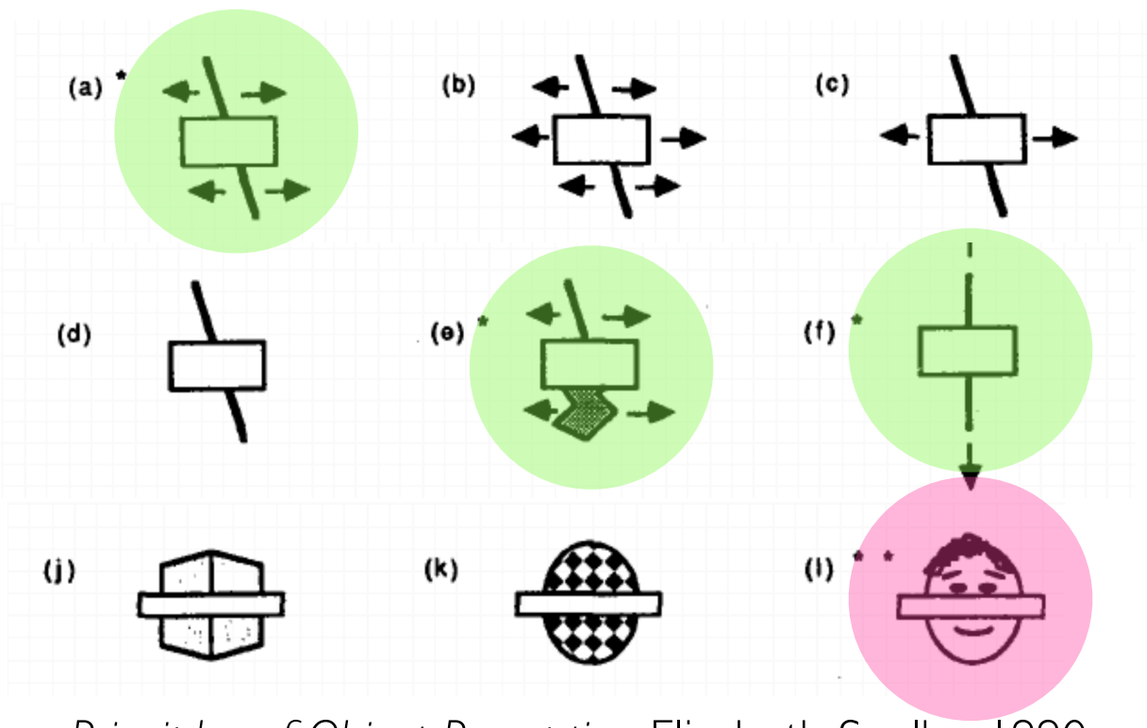
1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation:





Future Directions

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

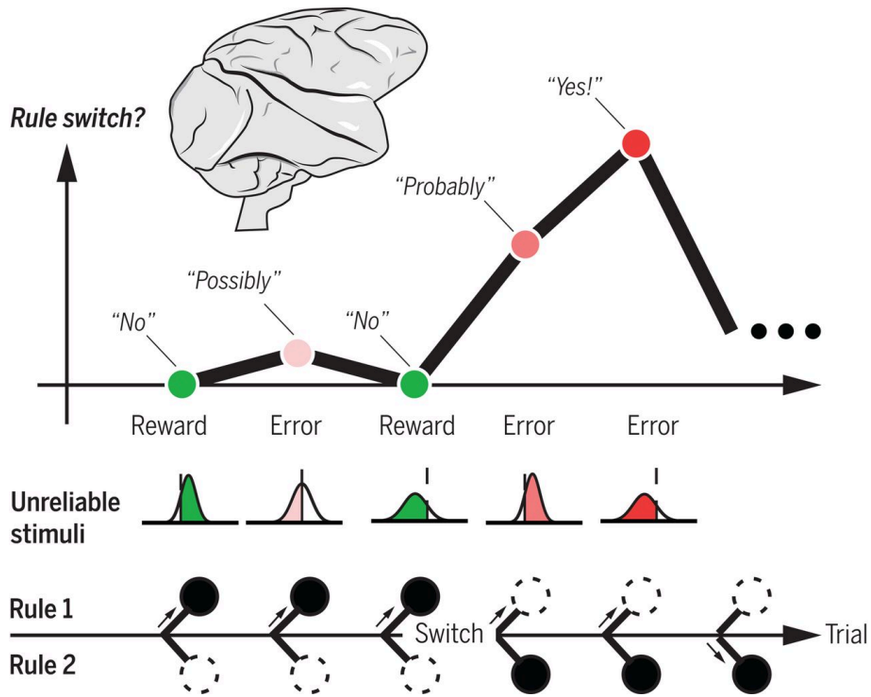
1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation: **object-centric, video foundation model**?

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

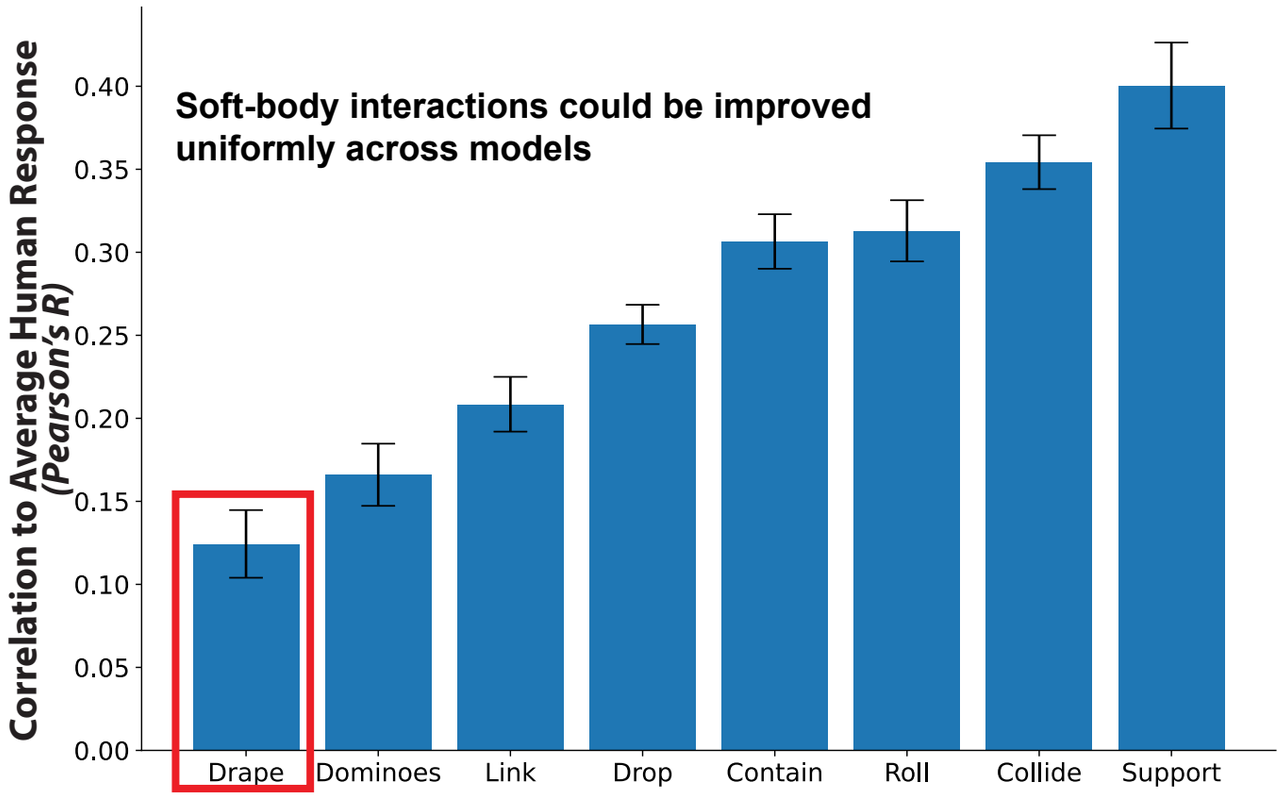


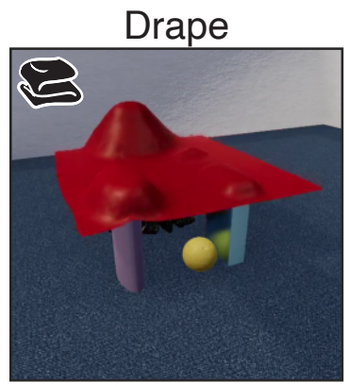
Future Directions: Learning Diverse Material Properties

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

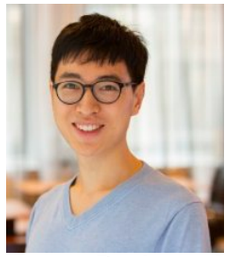




Acknowledgements


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

Contact:
anayebi@mit.edu
@aran_nayebi

Preprint: https://arxiv.org/abs/2305.11772

K. LISA YANG ICoN CENTER
Integrative Computational Neuroscience

MCGOVERN INSTITUTE

brain+cognitive sciences

SIMONS FOUNDATION

THE HELEN HAY WHITNEY FOUNDATION

JazLab


YangLab