Using Embodied AI for "Why" Questions in Systems Neuroscience

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

Neuroscience Institute Retreat

2024.12.04

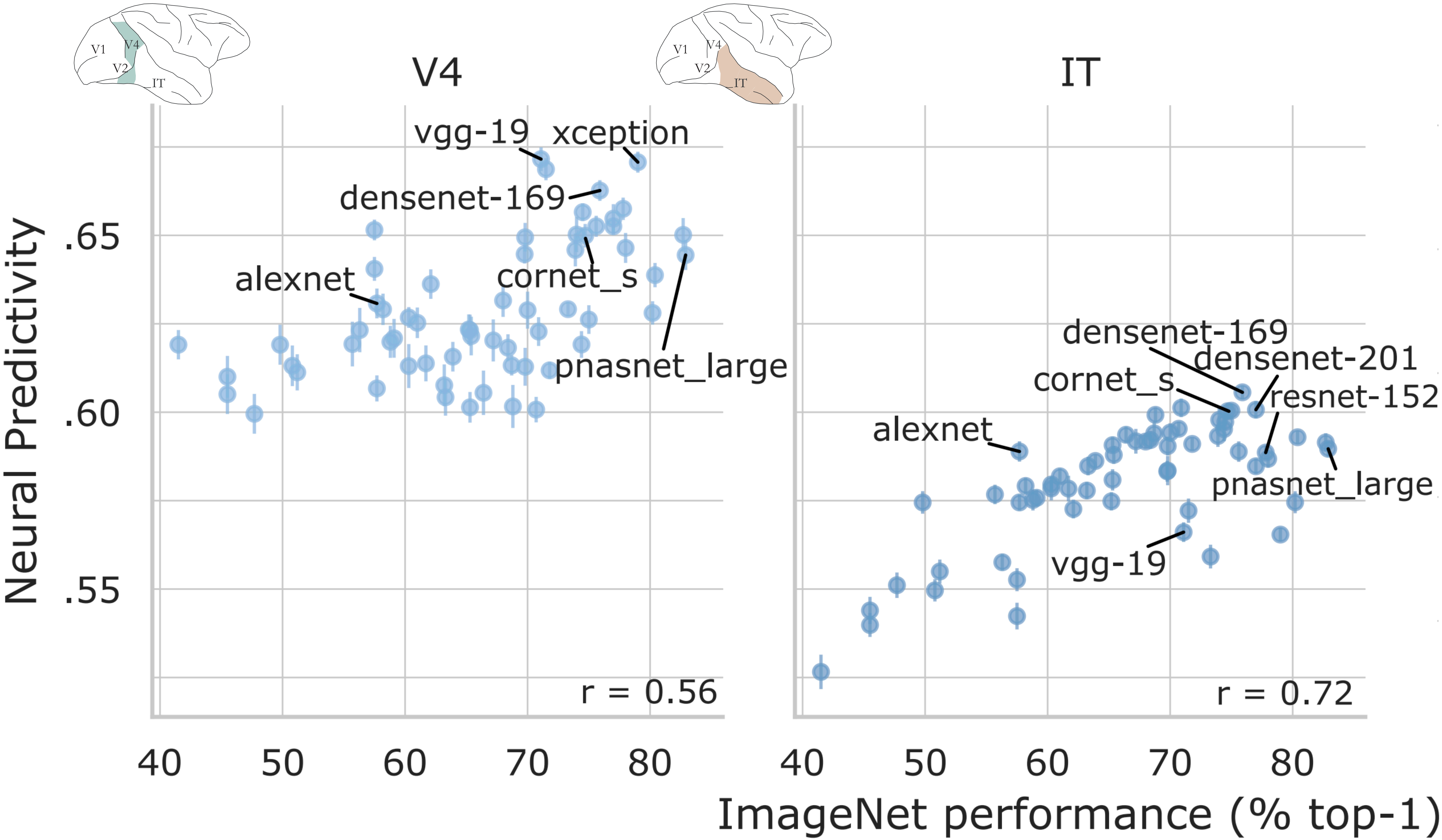
MACHINE LEARNING DEPARTMENT

Carnegie Mellon University
Neuroscience Institute

THE ROBOTICS INSTITUTE

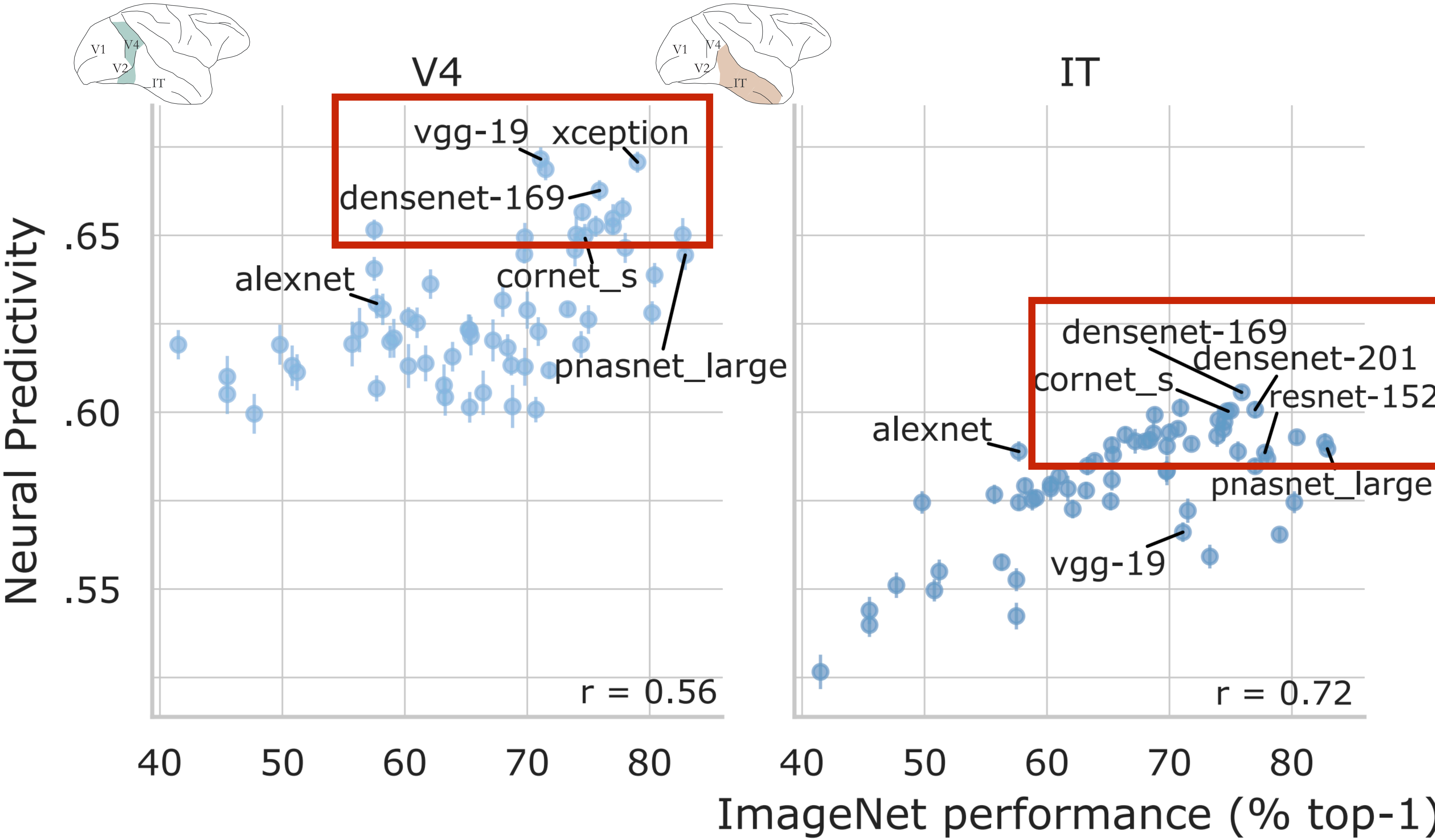
Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Similar predictivities among very different CNN architectures



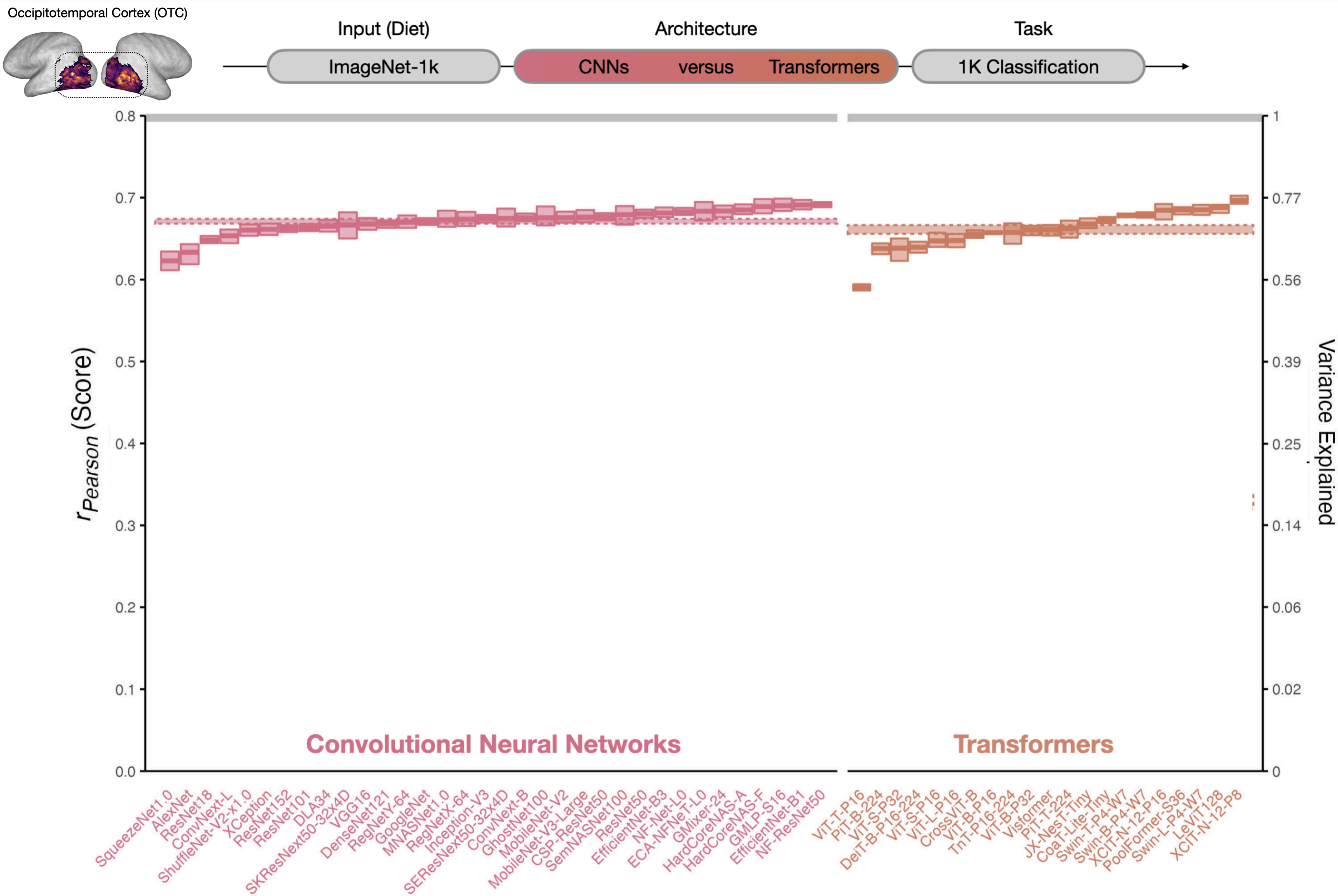
Schrimpf, Kubilius* et al. 2018*

Similar predictivities among very different CNN architectures



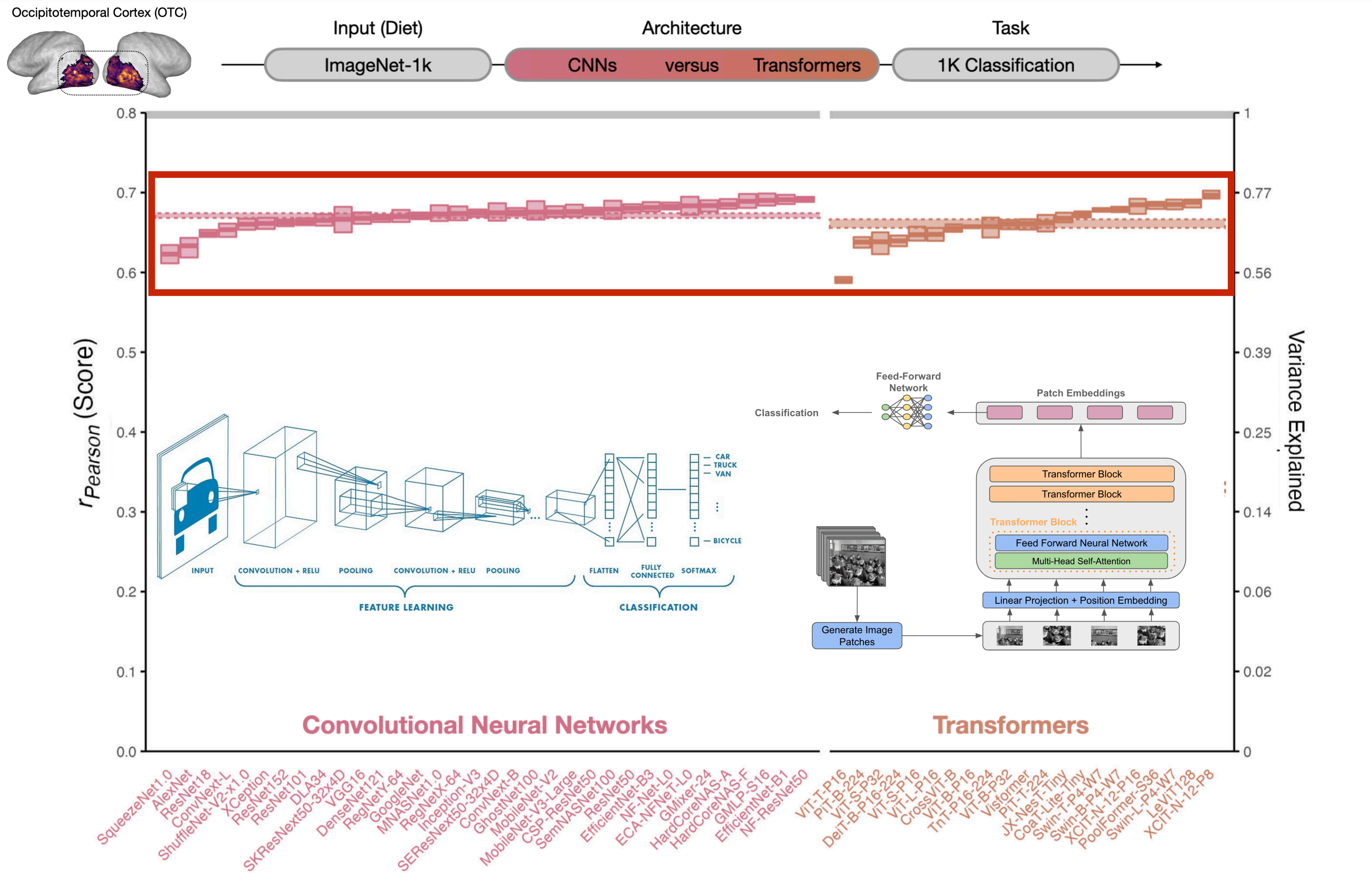
Schrimpf, Kubilius* et al. 2018*

Similar predictivities between CNNs vs. Transformers



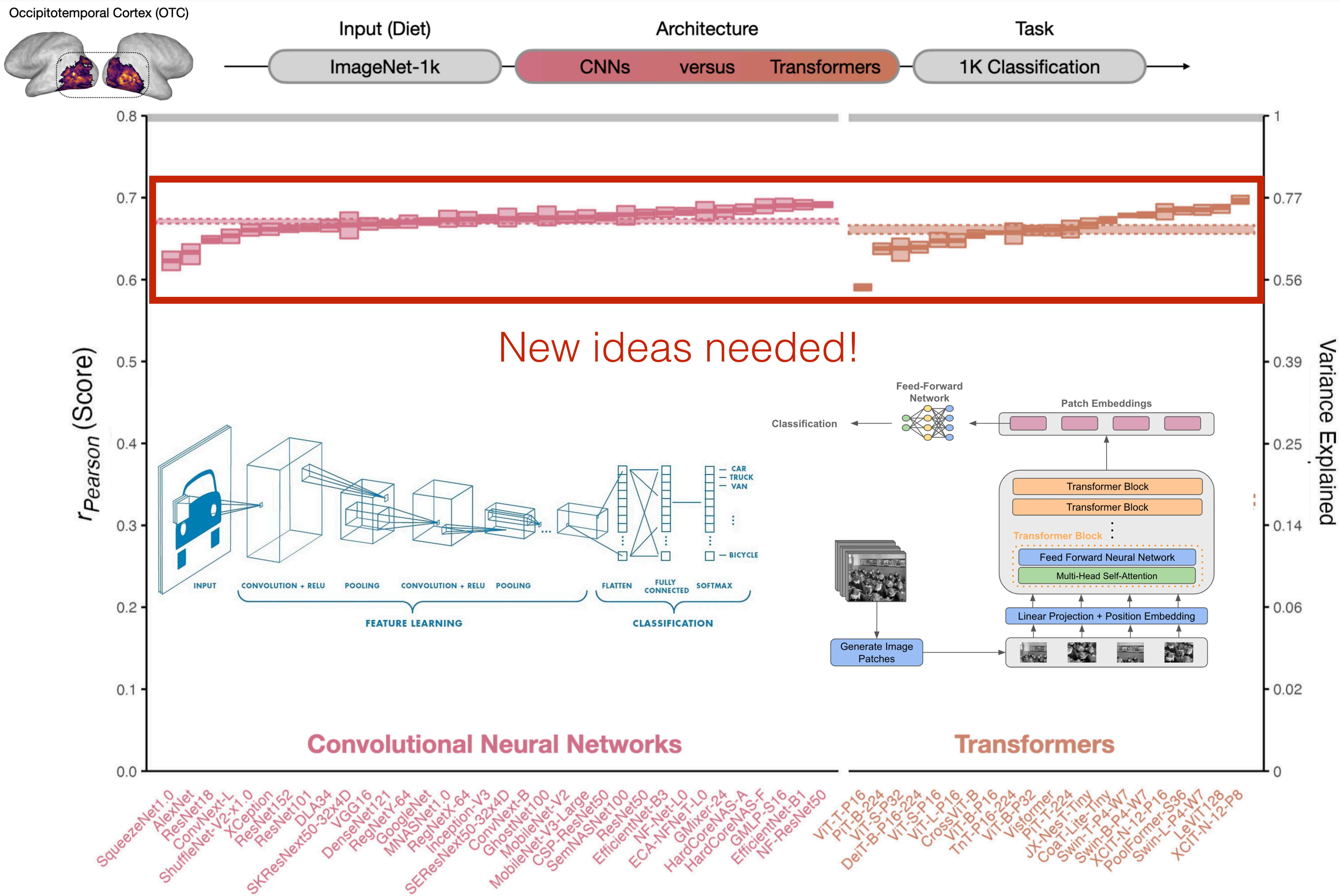
Conwell et al. 2023

Similar predictivities between CNNs vs. Transformers



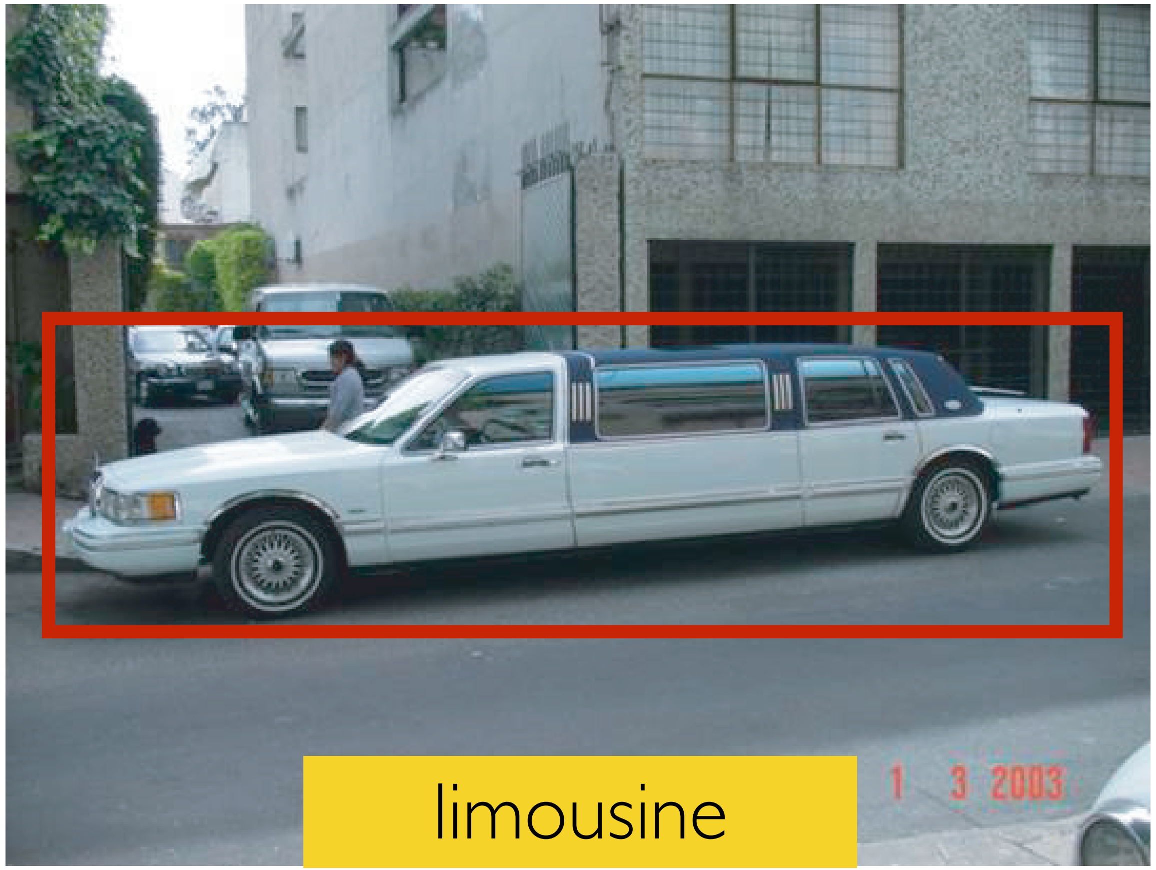
Conwell et al. 2023

Similar predictivities between CNNs vs. Transformers



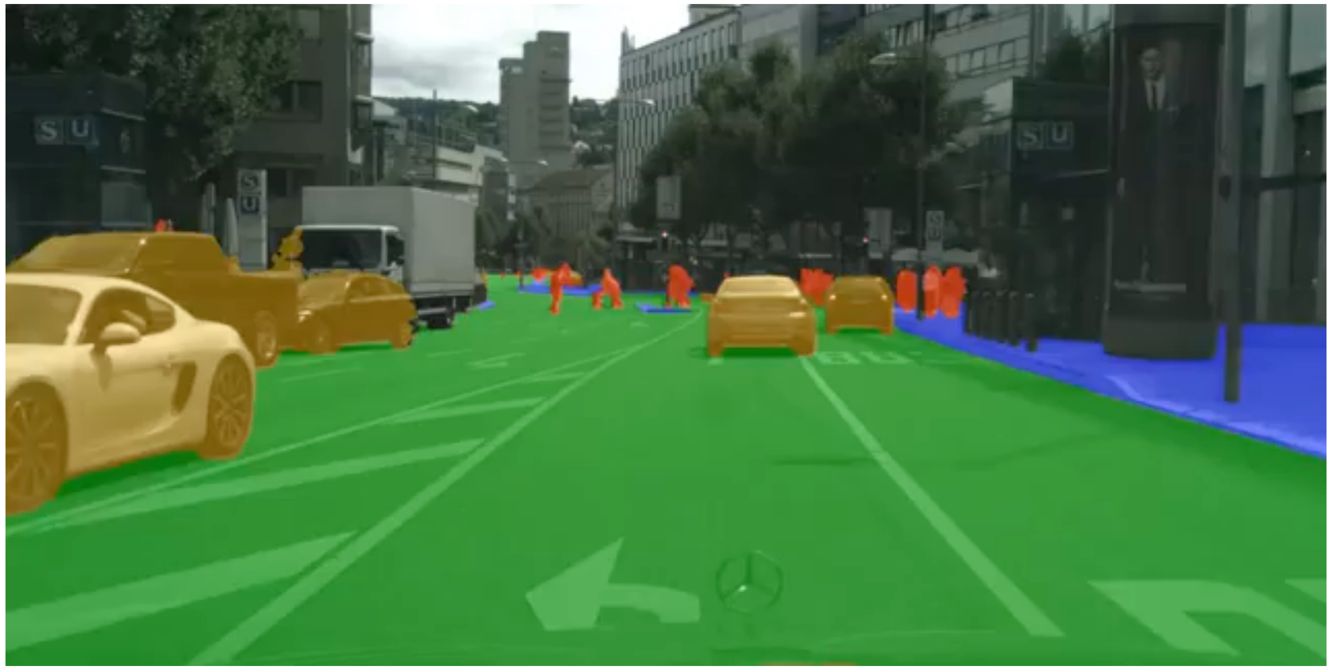
Conwell et al. 2023

We do a lot more than passive viewing…



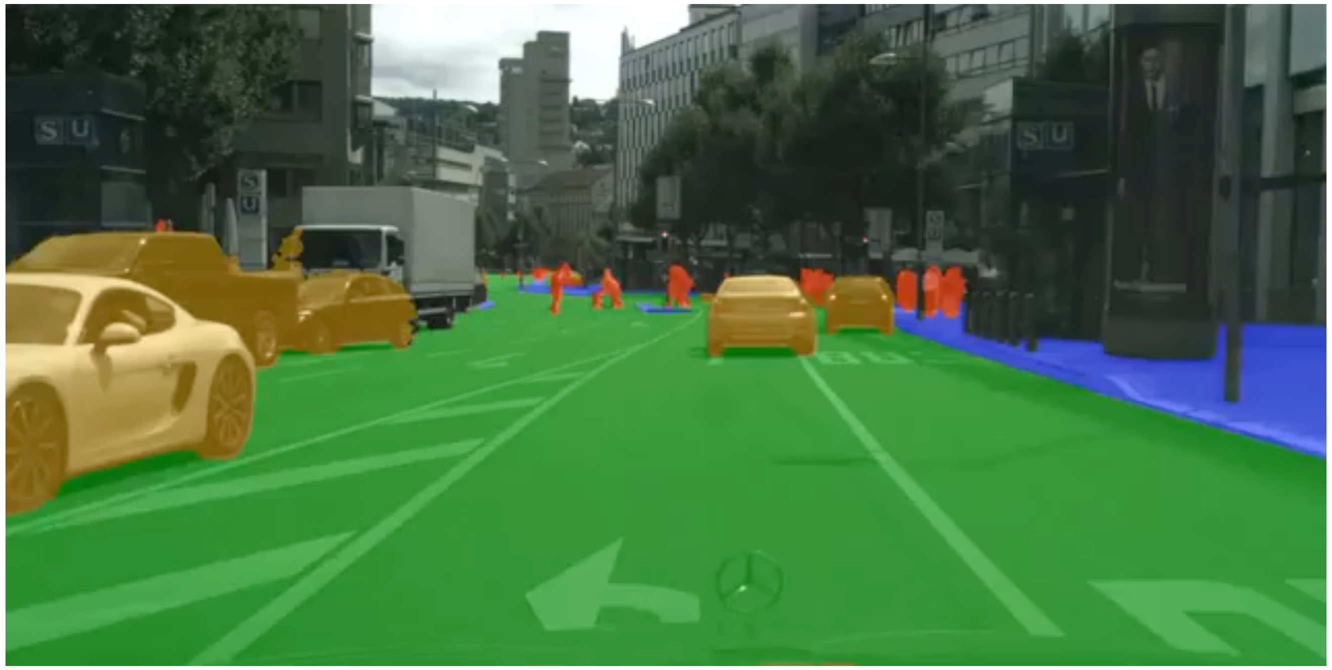
We do a lot more than passive viewing…

Scene Understanding



We do a lot more than passive viewing…

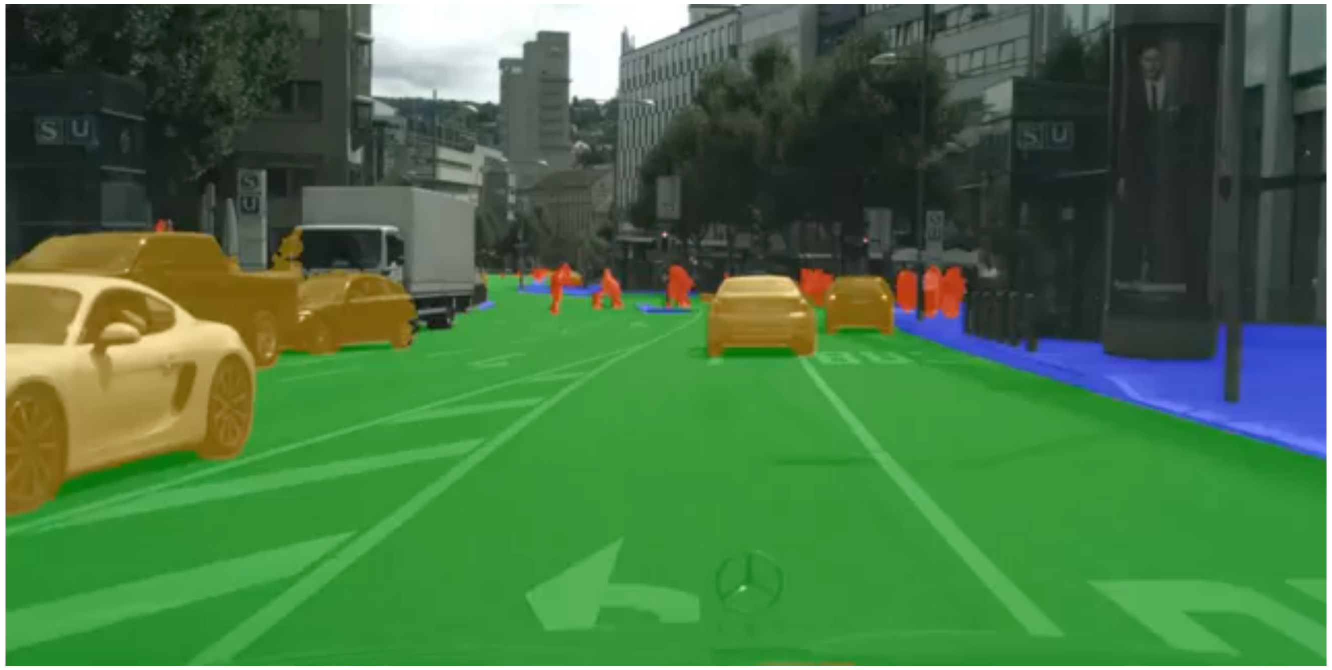
Scene Understanding



Multi-Step Planning

We do a lot more than passive viewing…

Scene Understanding

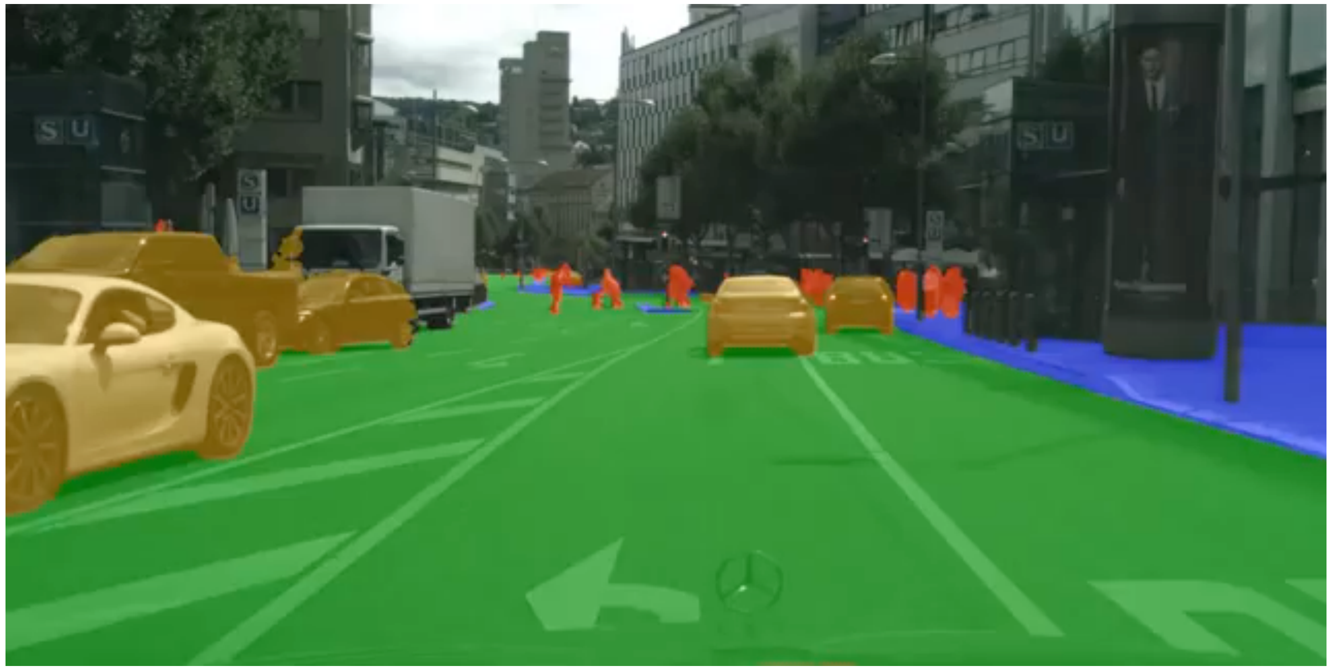


Multi-Step Planning

Navigation

We do a lot more than passive viewing…

Scene Understanding



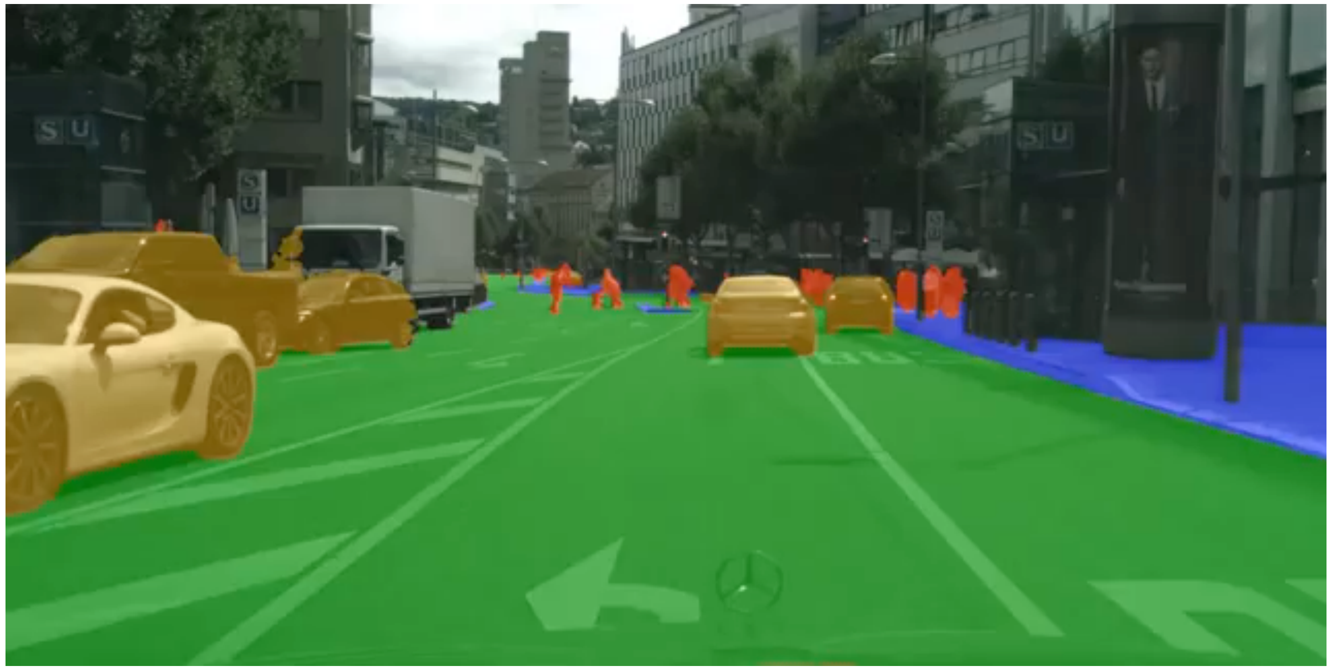
Multi-Step Planning

Navigation

Flexible Embodiment

We do a lot more than passive viewing…

Scene Understanding

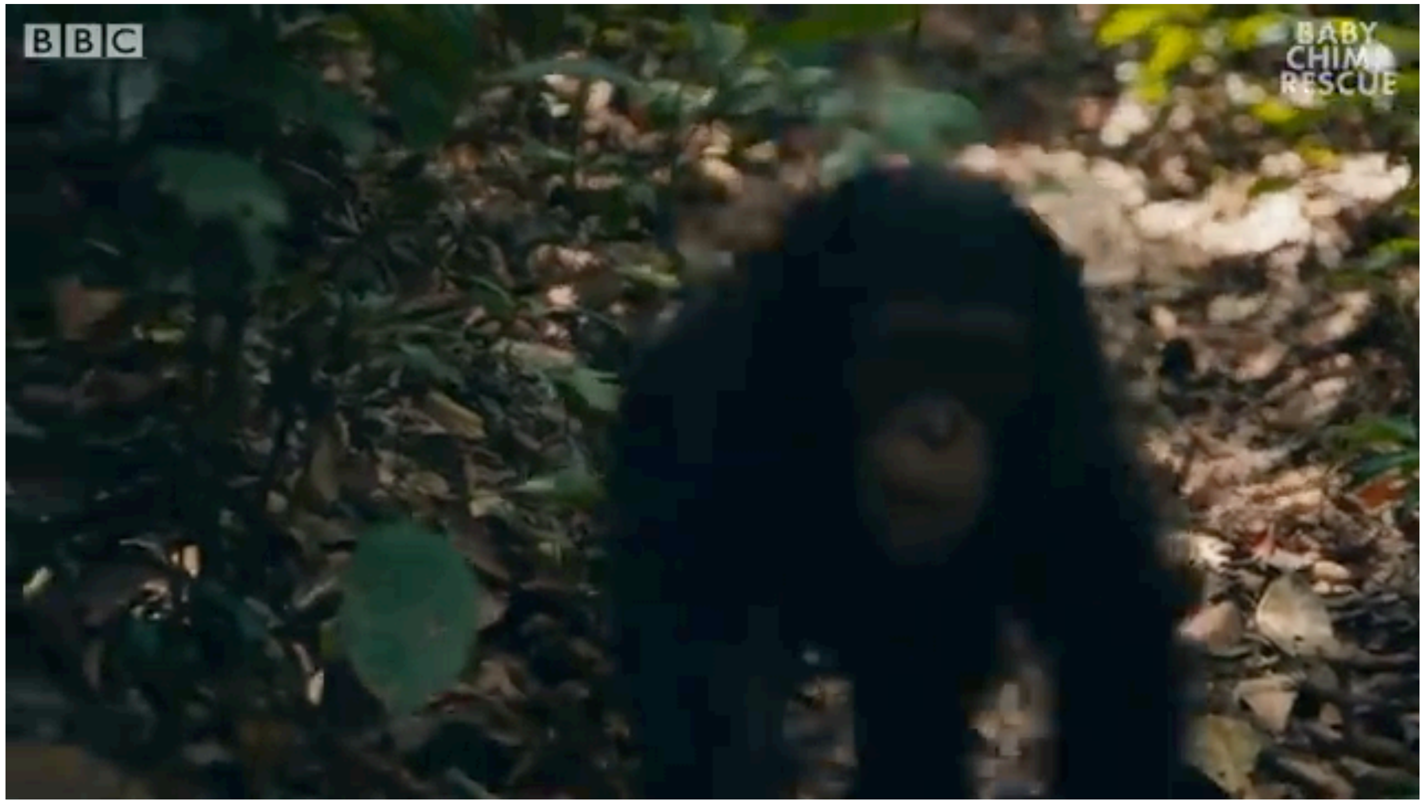


Multi-Step Planning

All of these behaviors are done in a body!

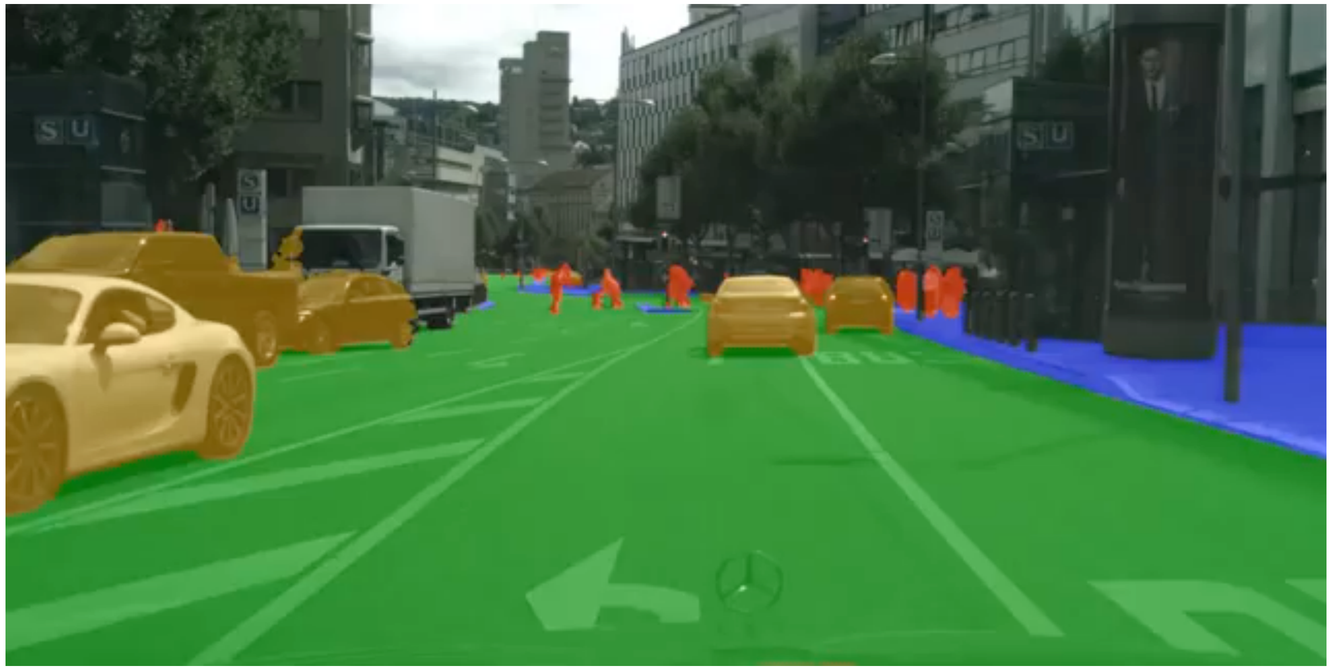
Navigation

Flexible Embodiment



We do a lot more than passive viewing…

Scene Understanding

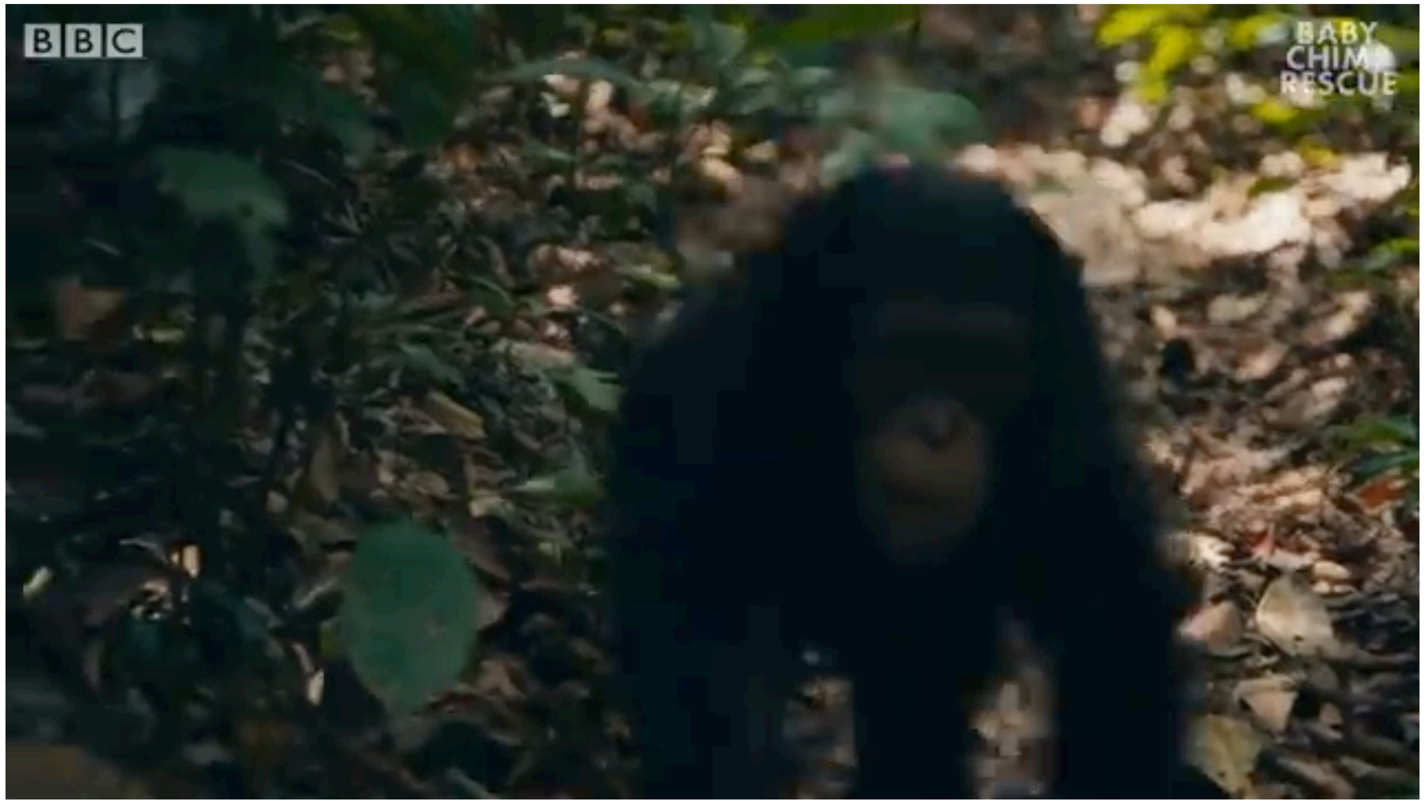


Multi-Step Planning

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

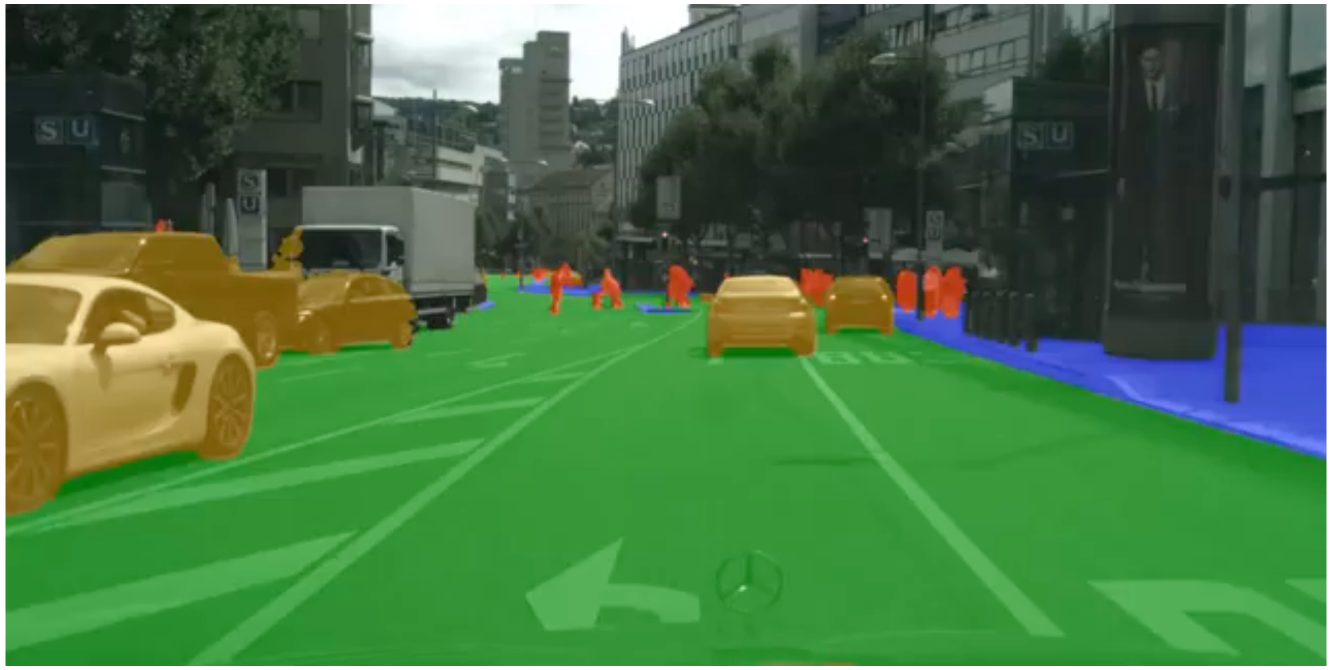
Navigation

Flexible Embodiment



We do a lot more than passive viewing…

Scene Understanding

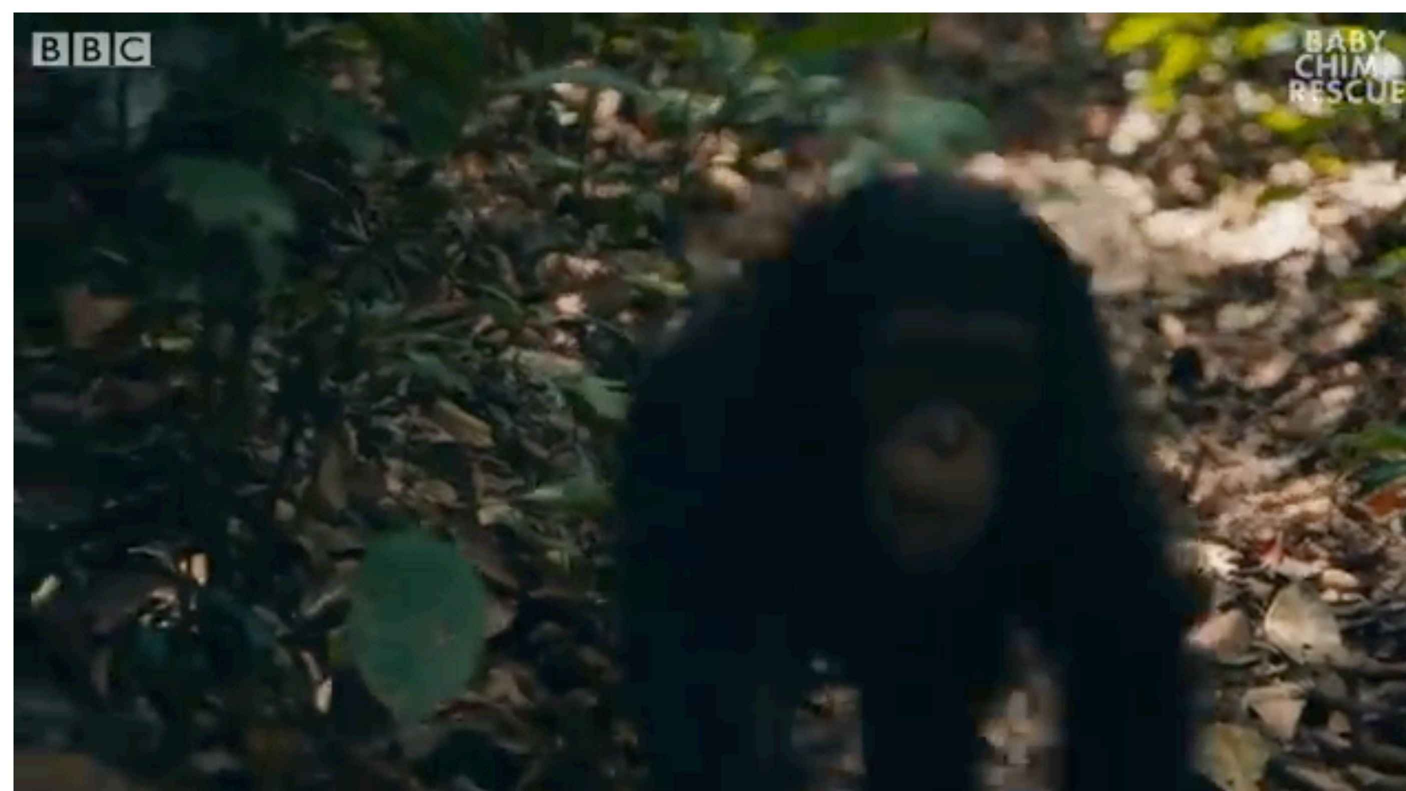


Multi-Step Planning

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

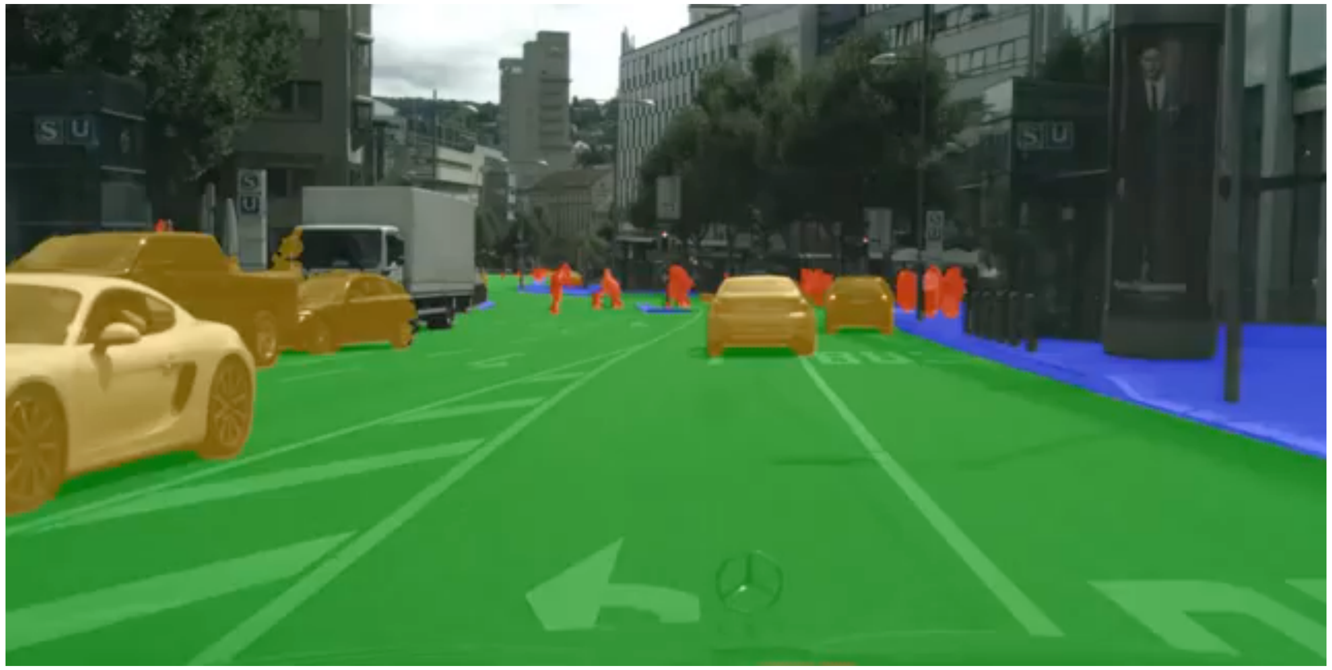
Navigation

Flexible Embodiment



We do a lot more than passive viewing…

Scene Understanding

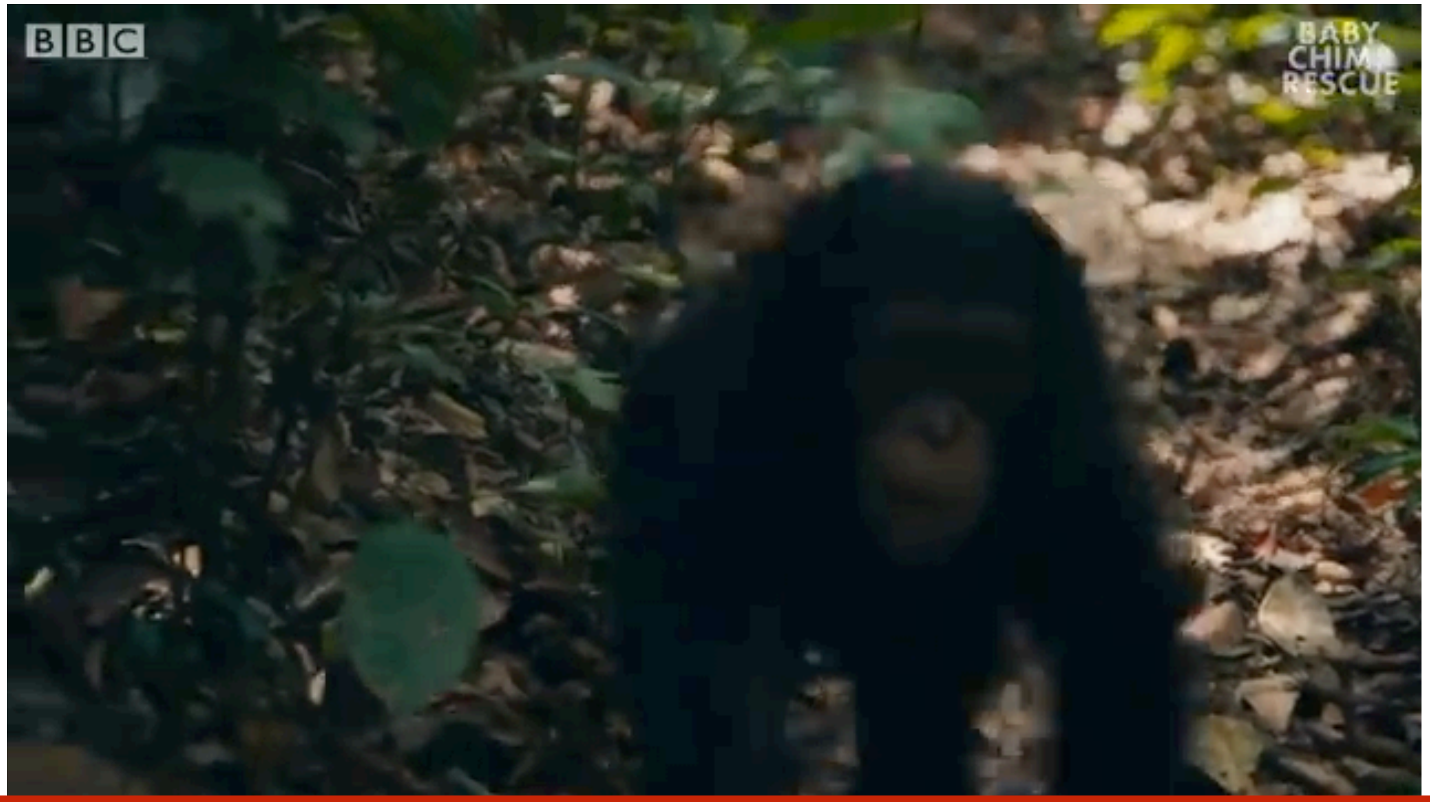


Multi-Step Planning

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

Navigation

Flexible Embodiment



Outline

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

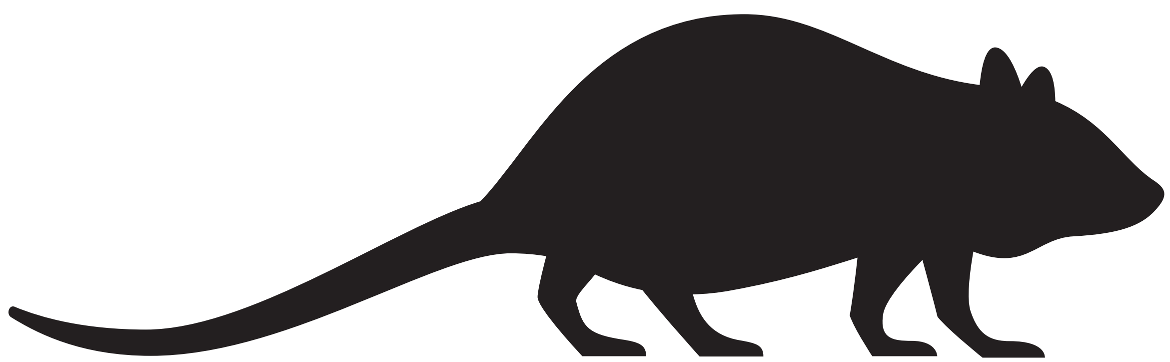
Outline

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

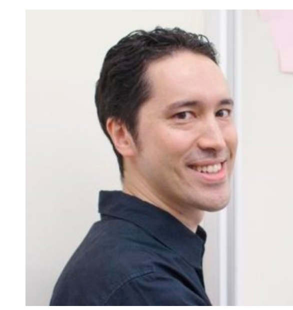
Mouse Visual Cortex as a Task-General, Limited Resource System

A. Nayebi*, N.C.L. Kong*, C. Zhuang, J.L. Gardner, A.M. Norcia, D.L.K. Yamins
Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation.
PLOS Computational Biology 2023


Nathan C.L. Kong*



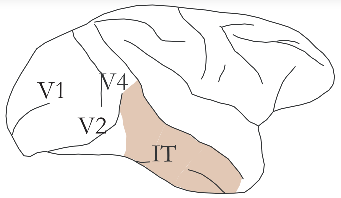

Chengxu Zhuang


Justin L. Gardner


Anthony M. Norcia

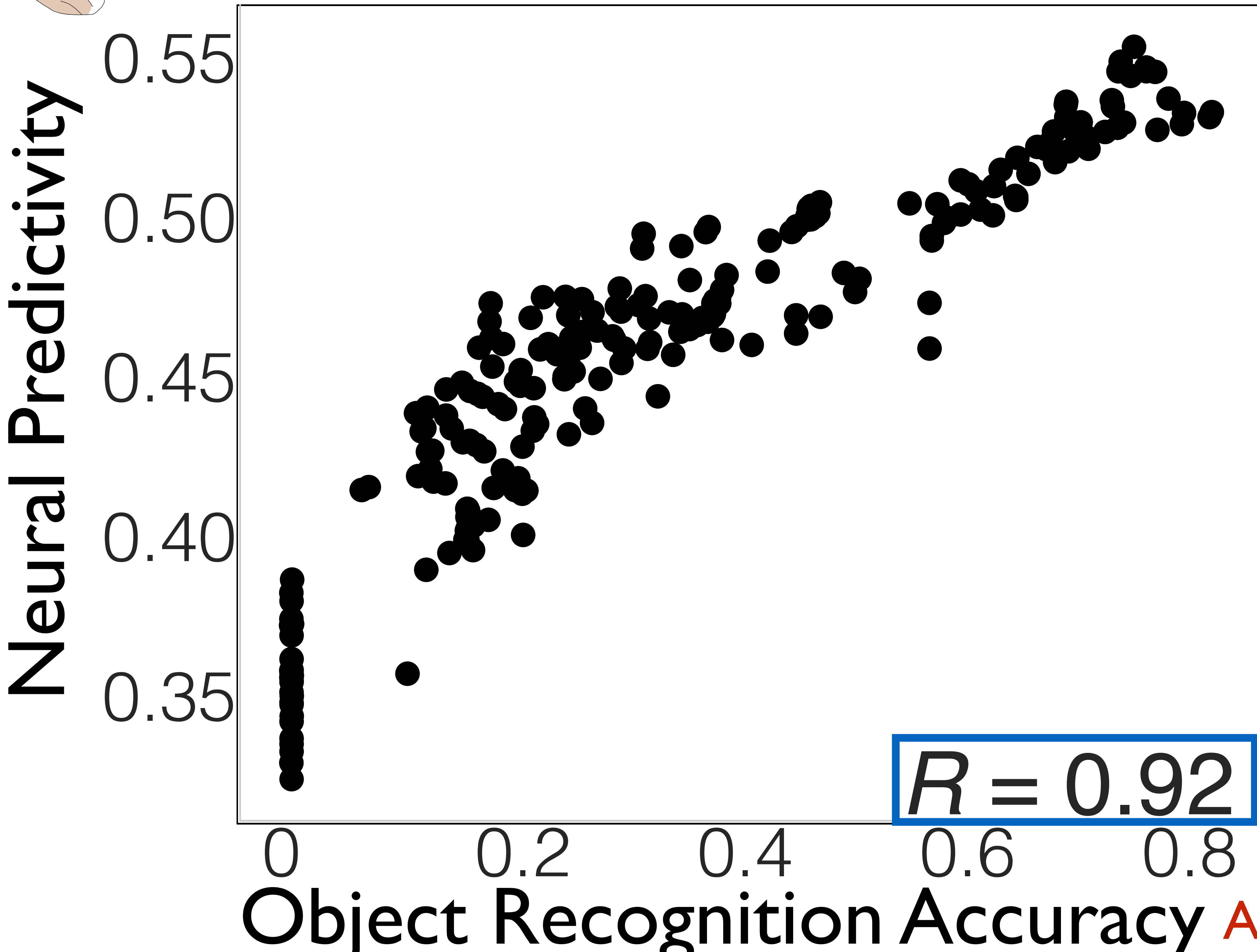

Daniel Yamins

Task Performance Correlated with Neural Predictivity



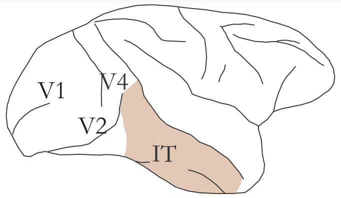
A Neuroscience Goal

Schrimpf, Kubilius* et al. 2018*

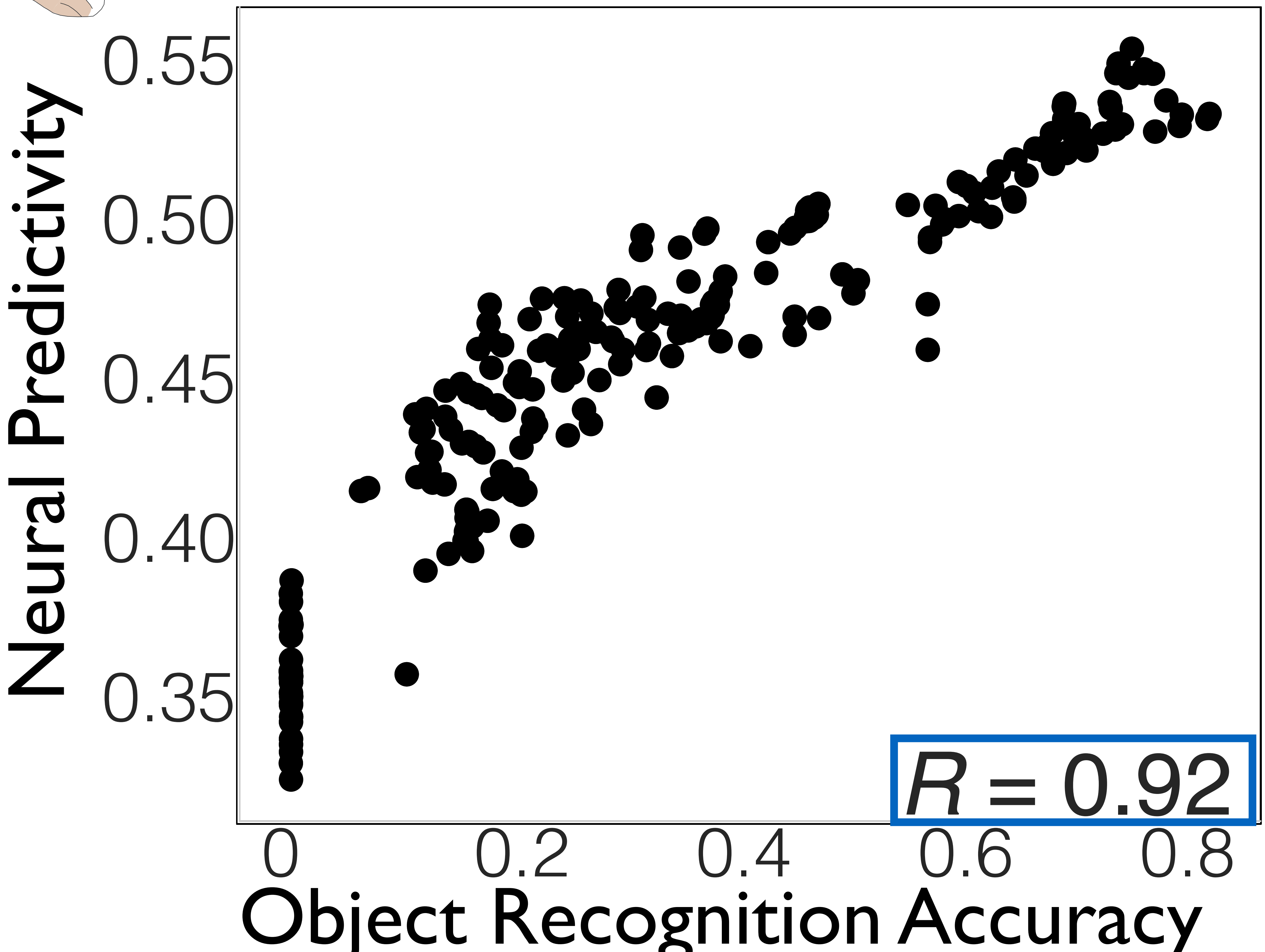


An AI Goal

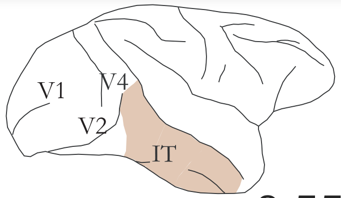
Task Performance Correlated with Neural Predictivity



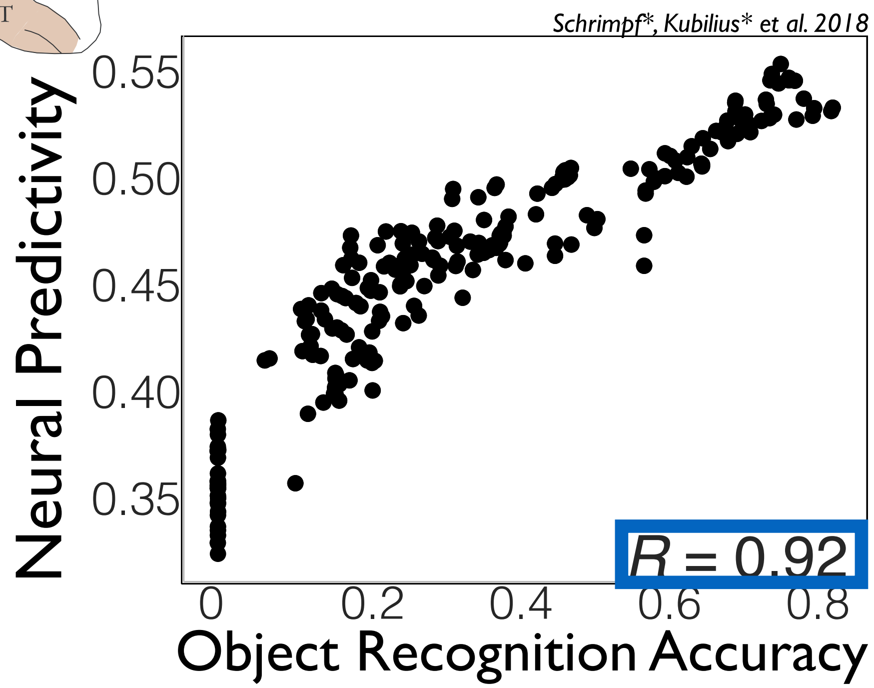
Schrimpf, Kubilius* et al. 2018*

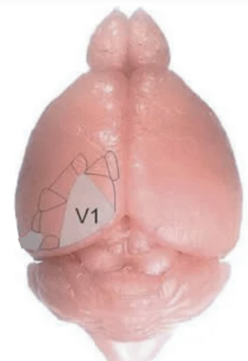


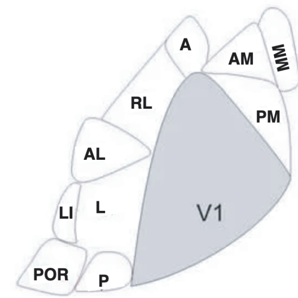
Task Performance Correlated with Neural Predictivity



Primates

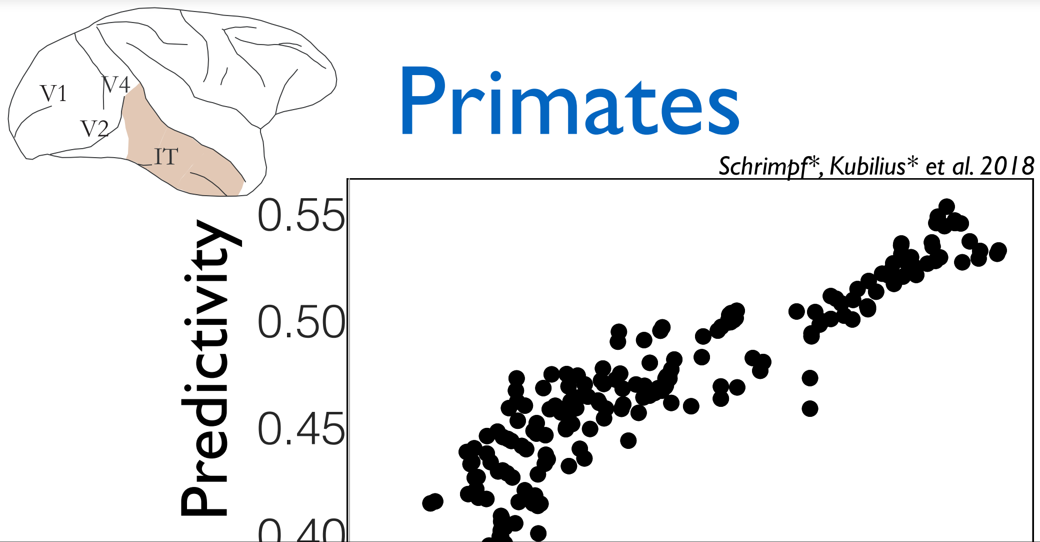


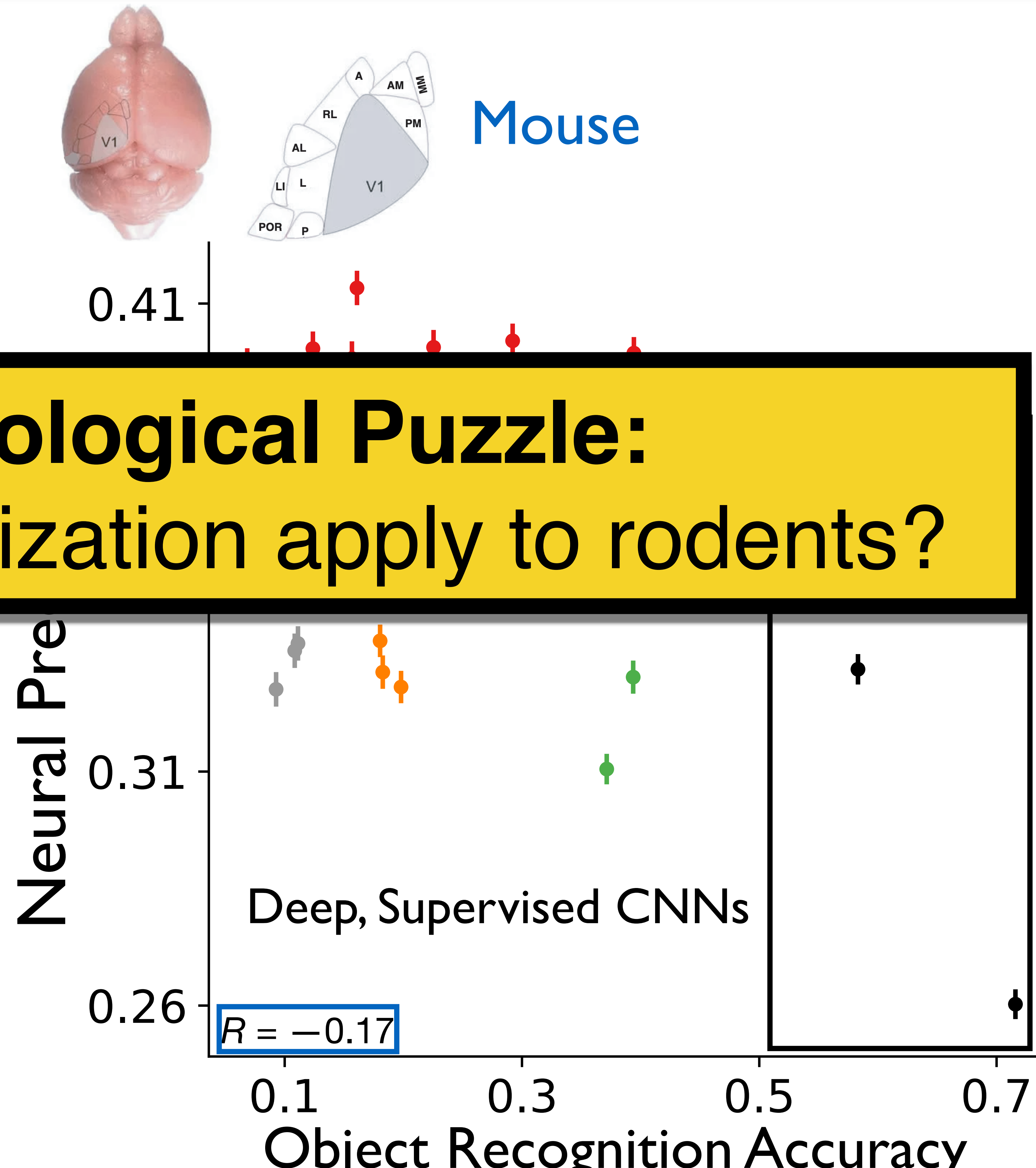




Mouse?

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

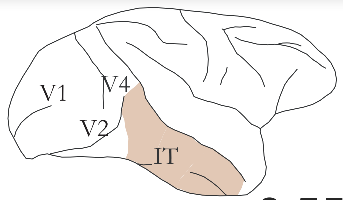




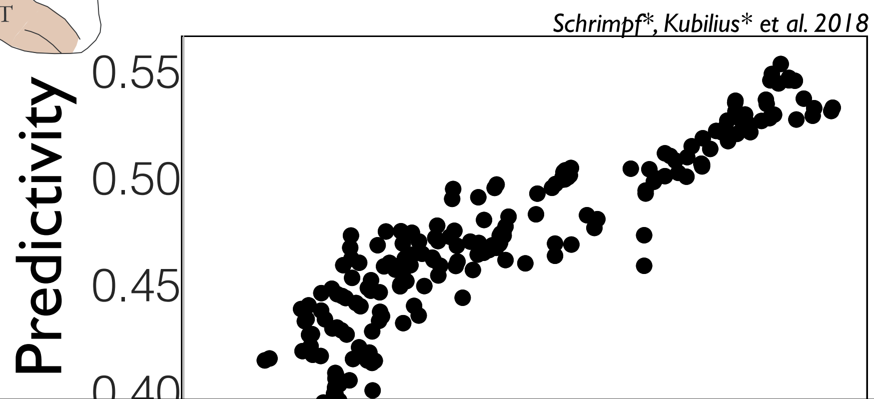
Neurobiological Puzzle:

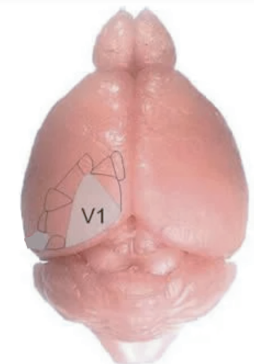
Does task-optimization apply to rodents?

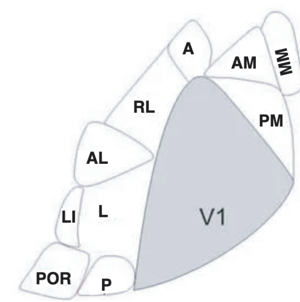
Object Categorization Ability **NOT** Correlated with Neural Predictivity



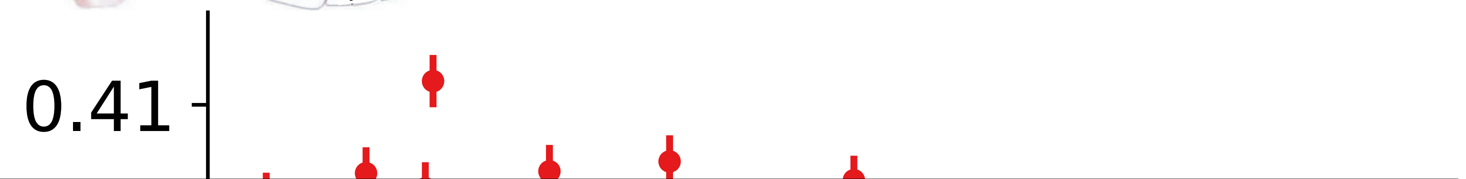
Primates







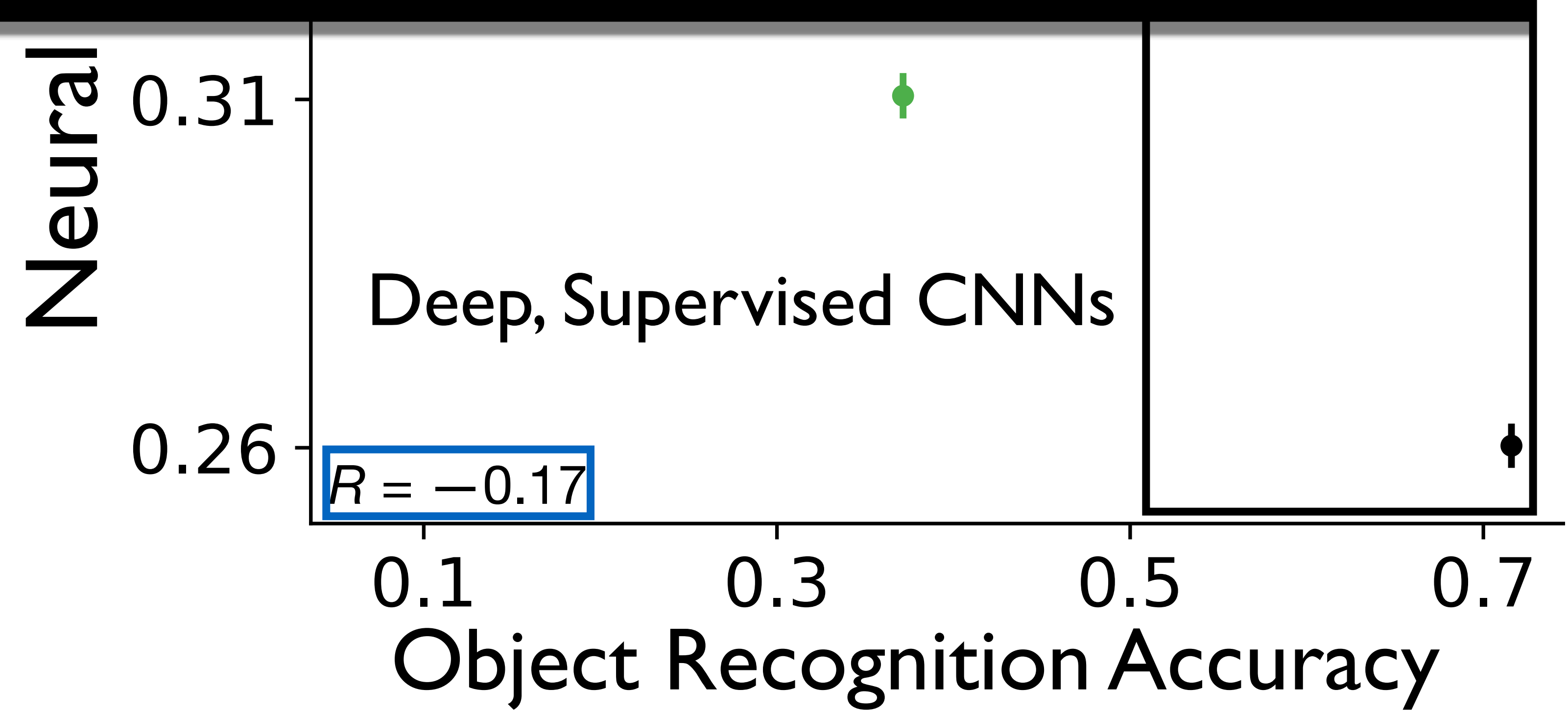
Mouse



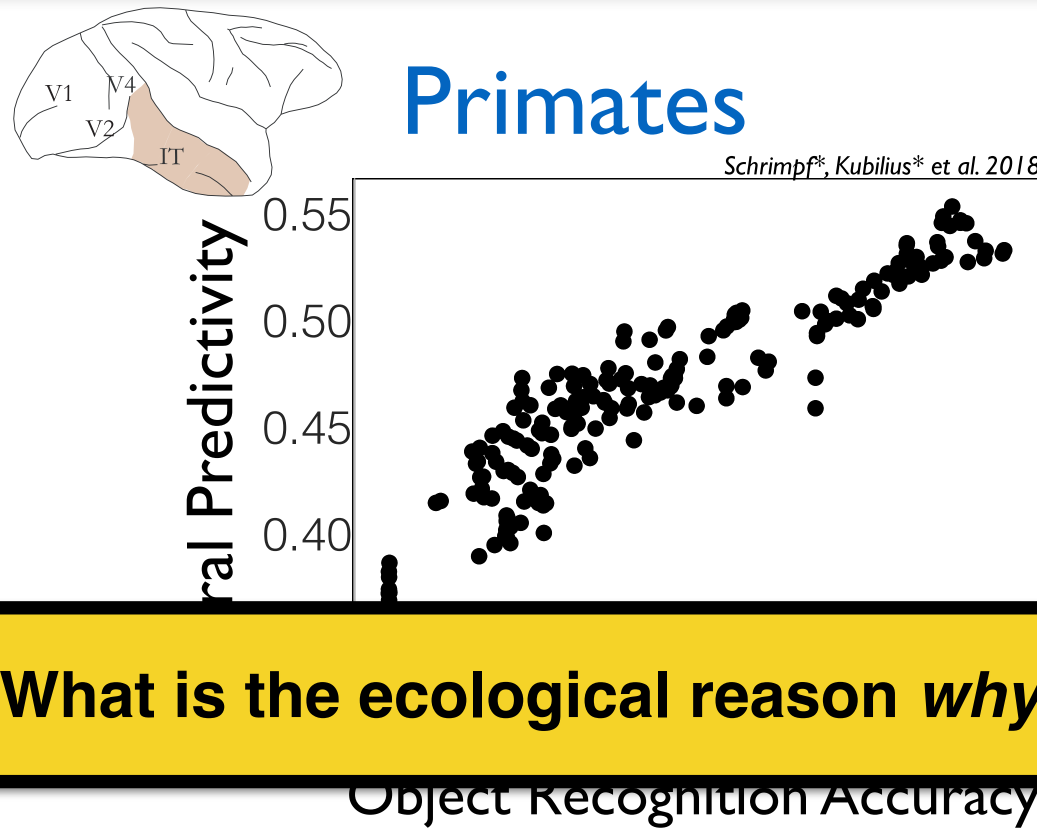
Neurobiological Puzzle:

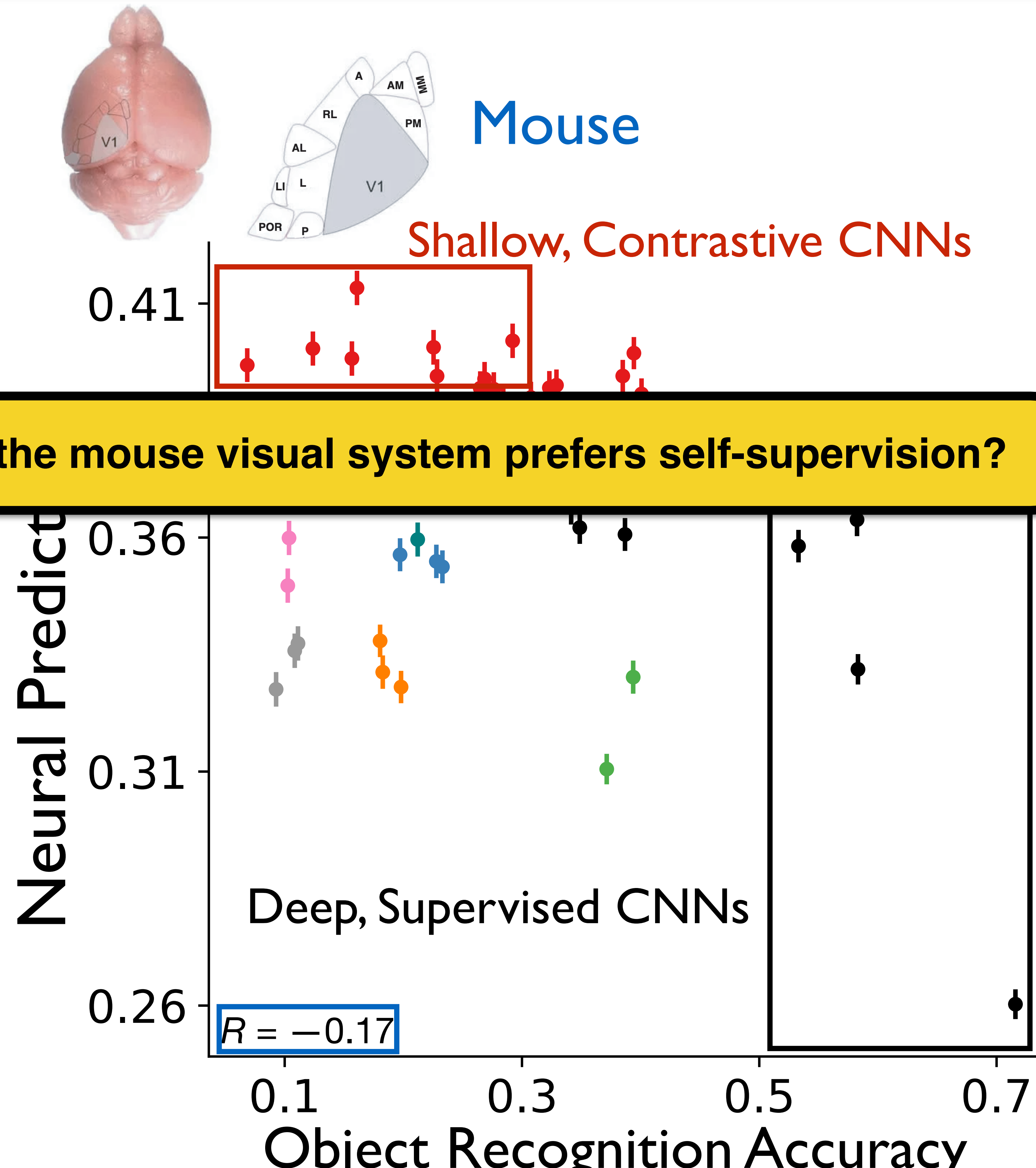
Does task-optimization apply to rodents?

Yes!



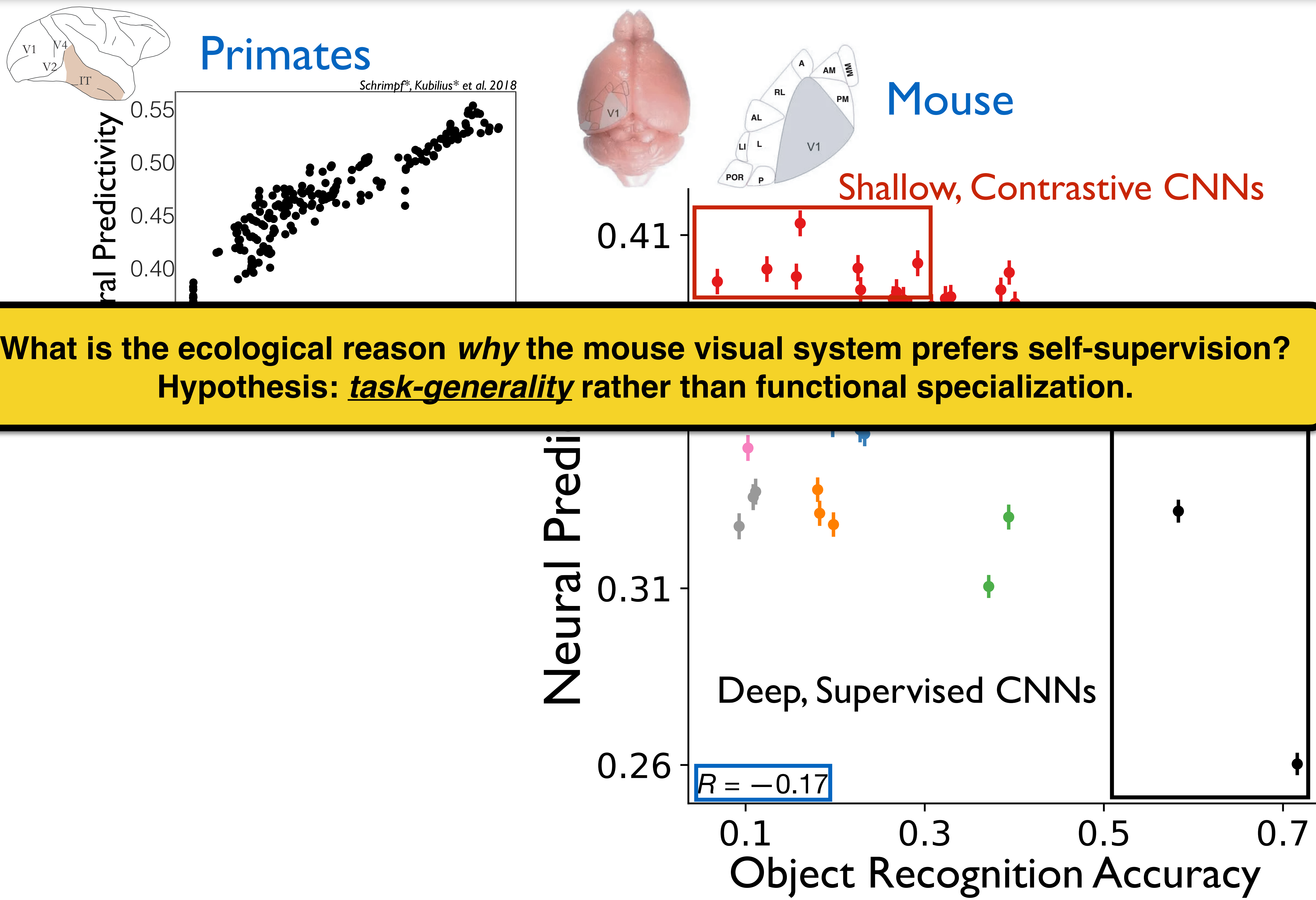
Object Categorization Ability **NOT** Correlated with Neural Predictivity





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

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



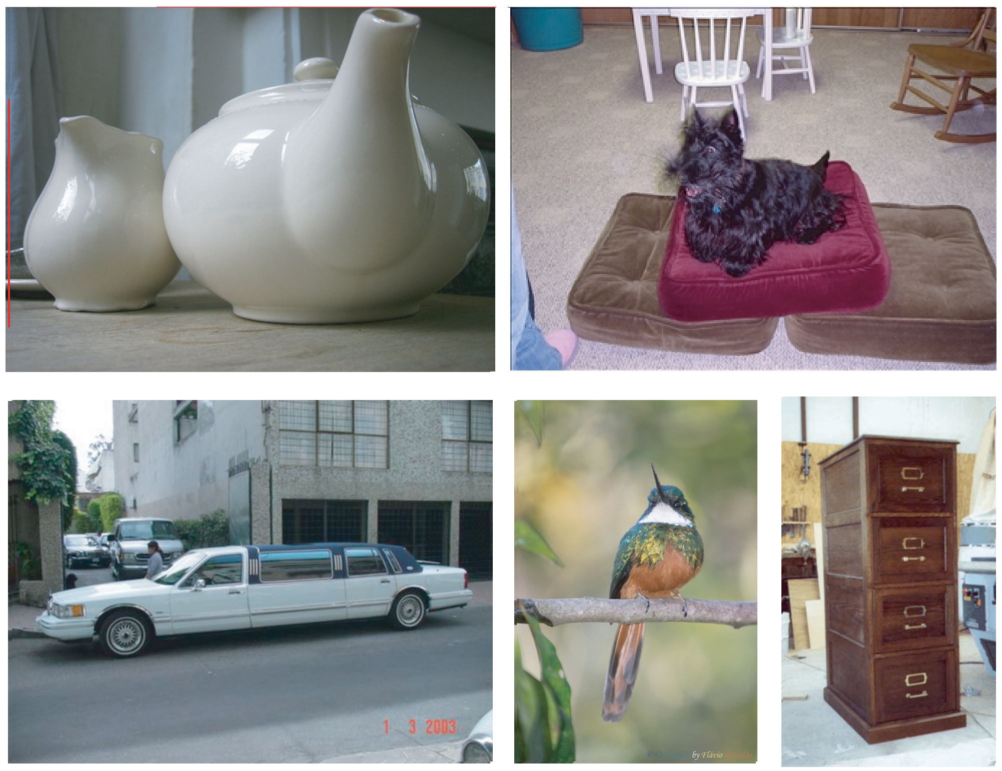
What is the ecological reason *why* the mouse visual system prefers self-supervision?
Hypothesis: *task-generality* rather than functional specialization.

Assessing Task-Generality

Assessing Task-Generality

Train

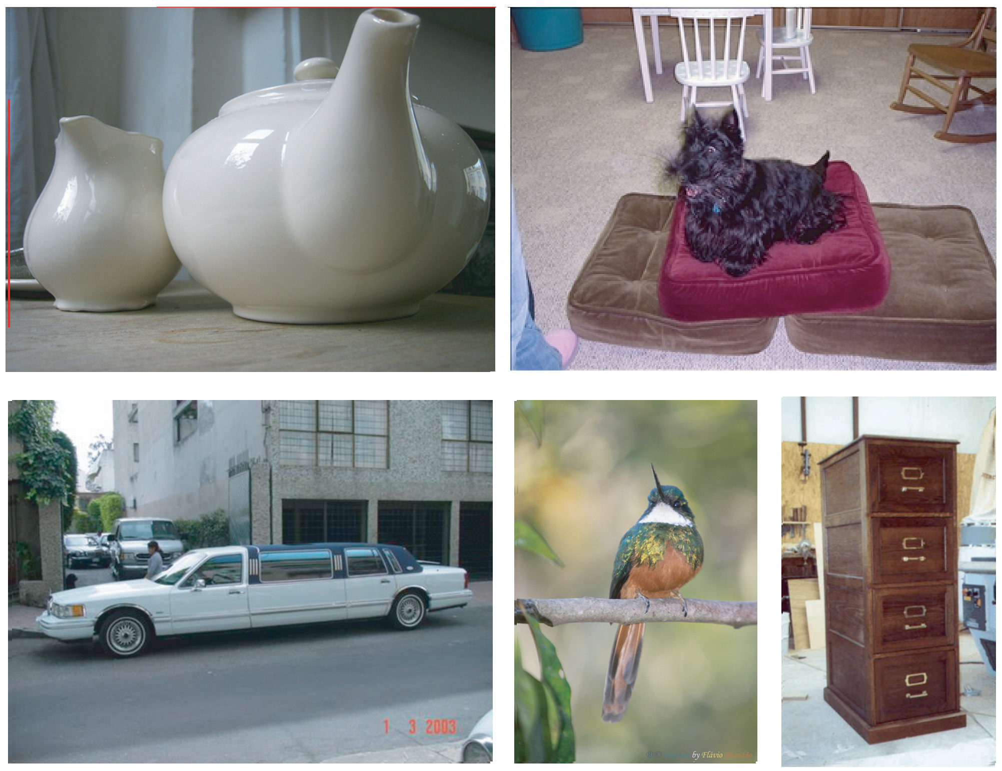
ImageNet



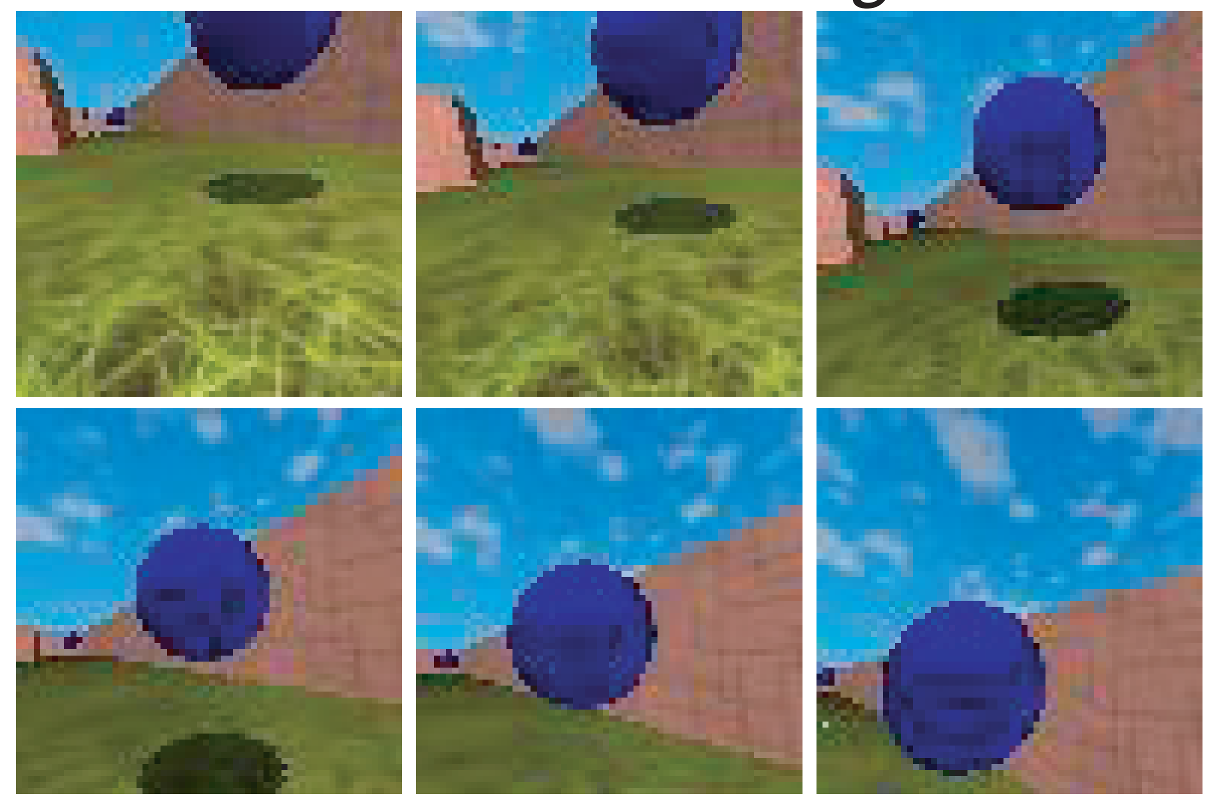
Assessing Task-Generality

Train | **Evaluate**

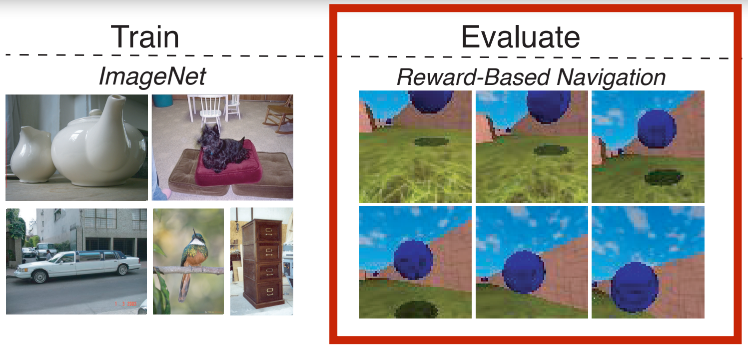
ImageNet



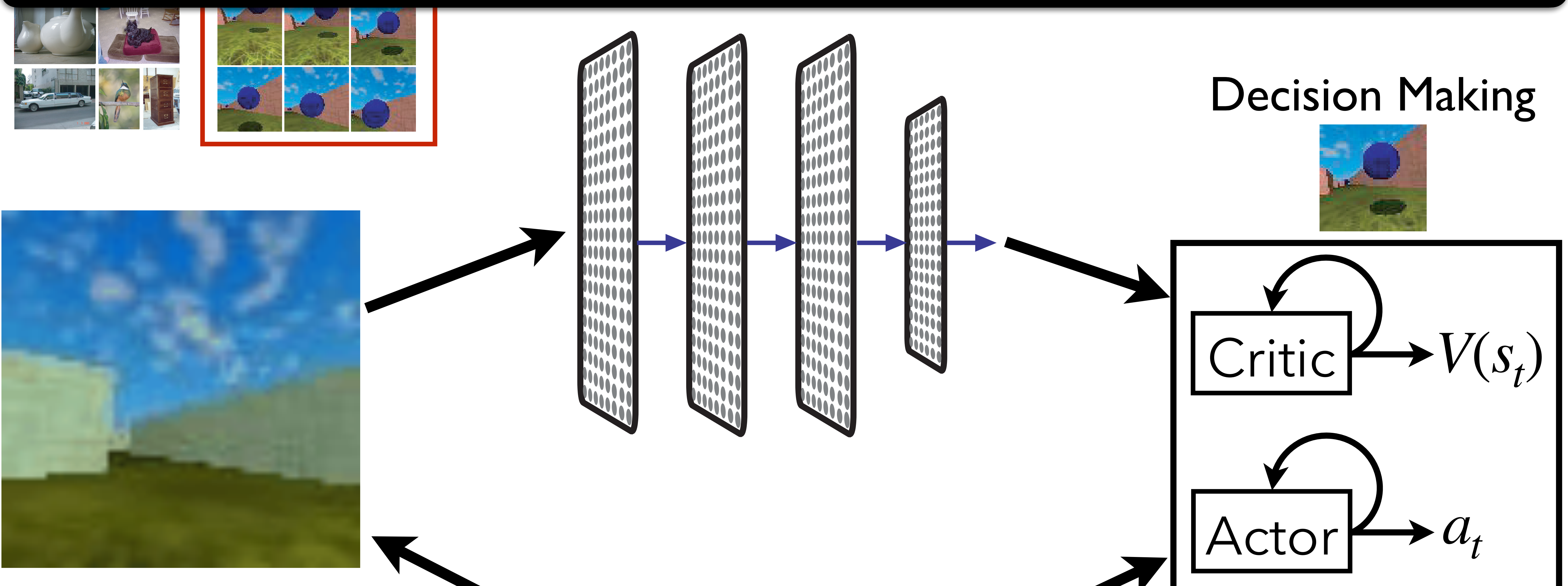
Reward-Based Navigation

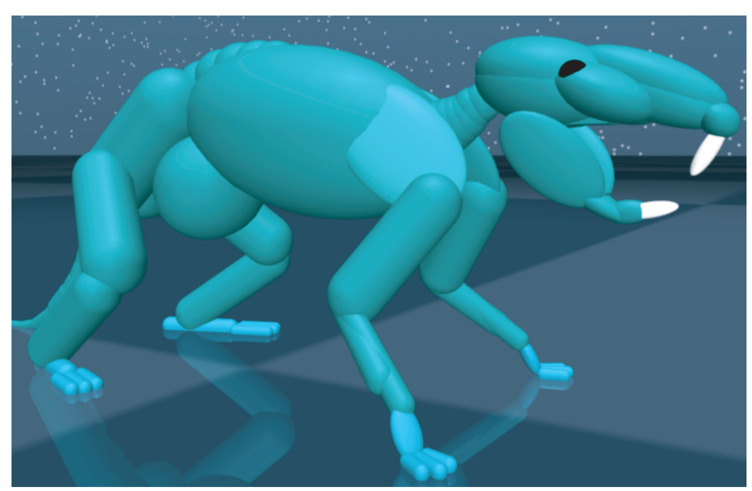


Assessing Task-Generality



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



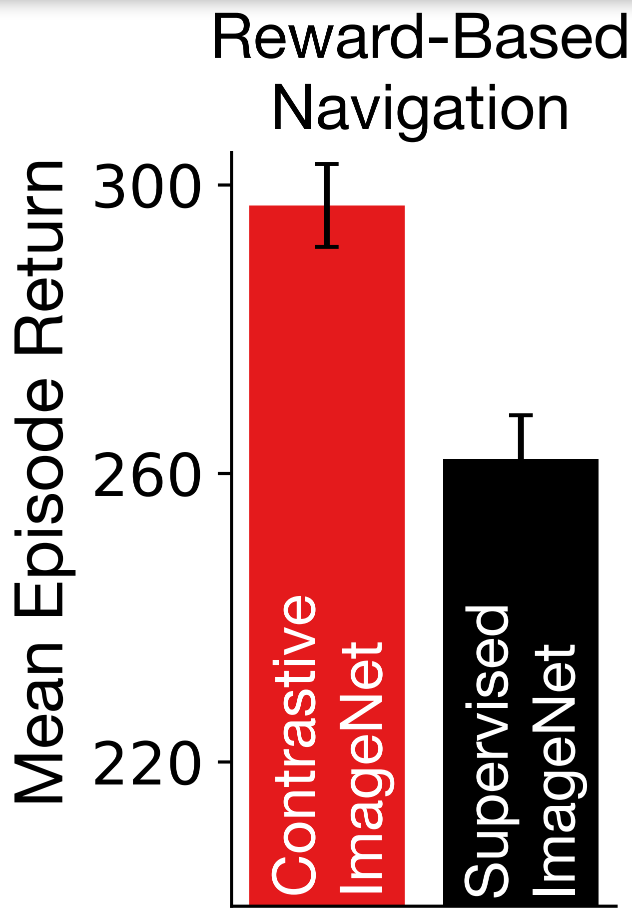


Biomechanical Model
(Joint angles, accelerometer, etc.)

Bence Ölveczky

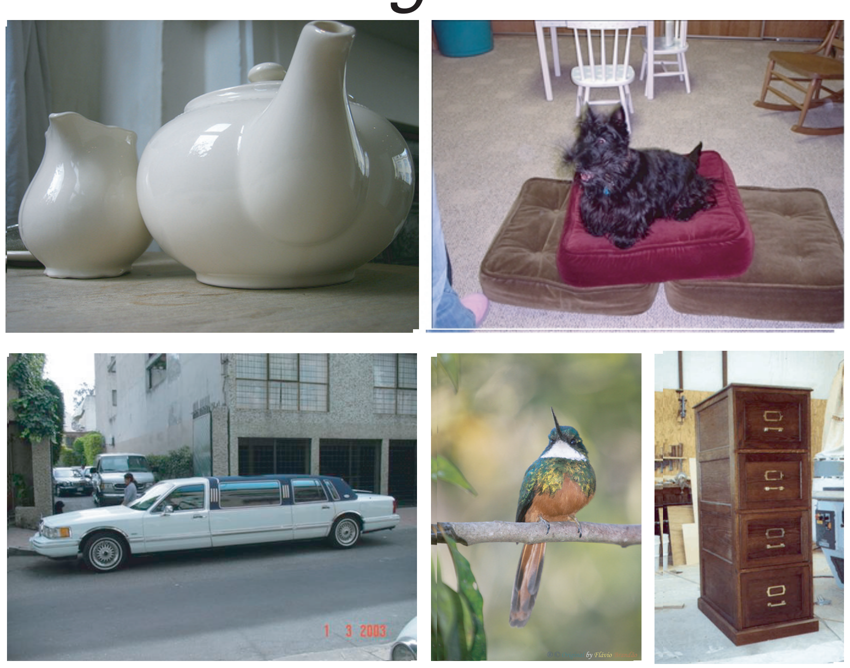
Contrastive Models Yield Better Transfer Performance

Contrastive Models Yield Better Transfer Performance

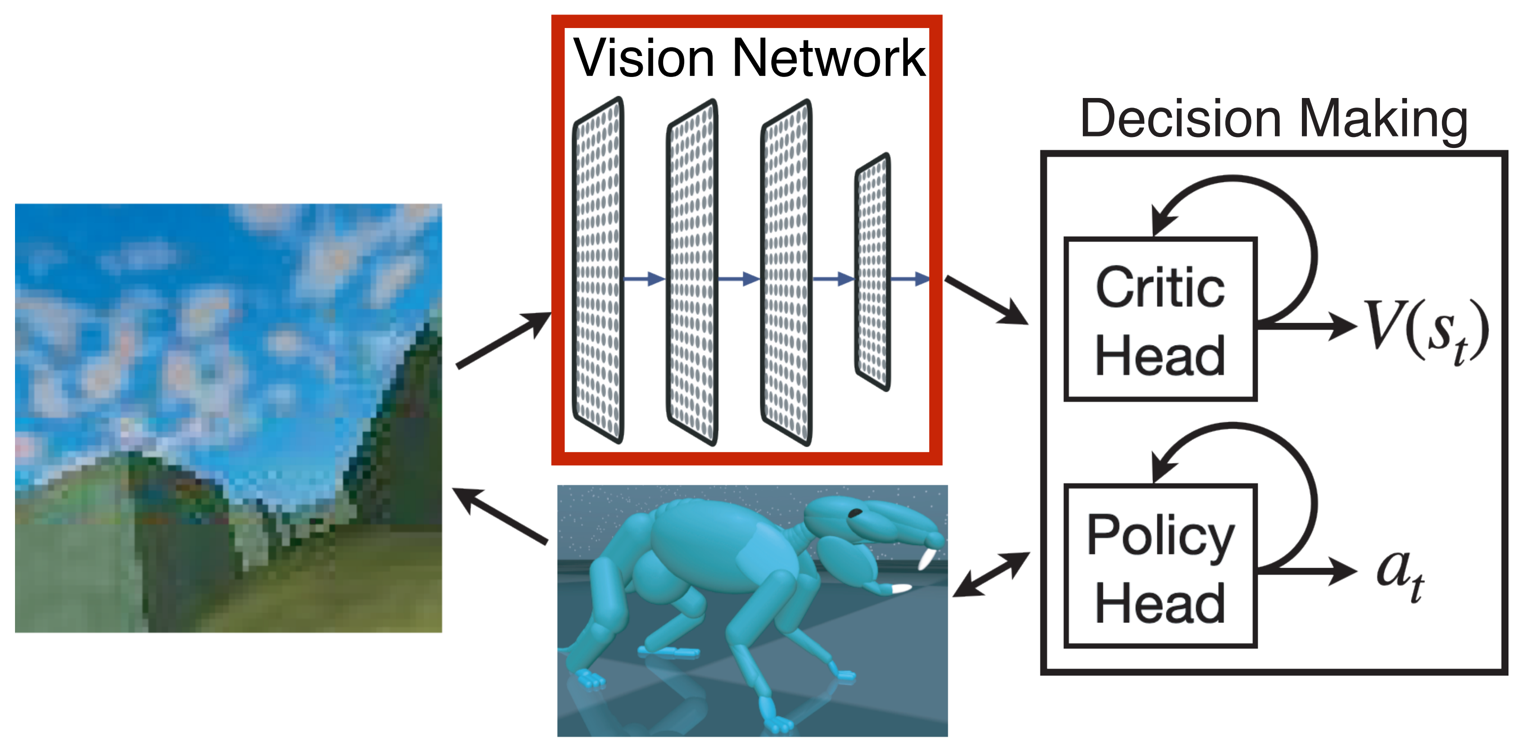


Train

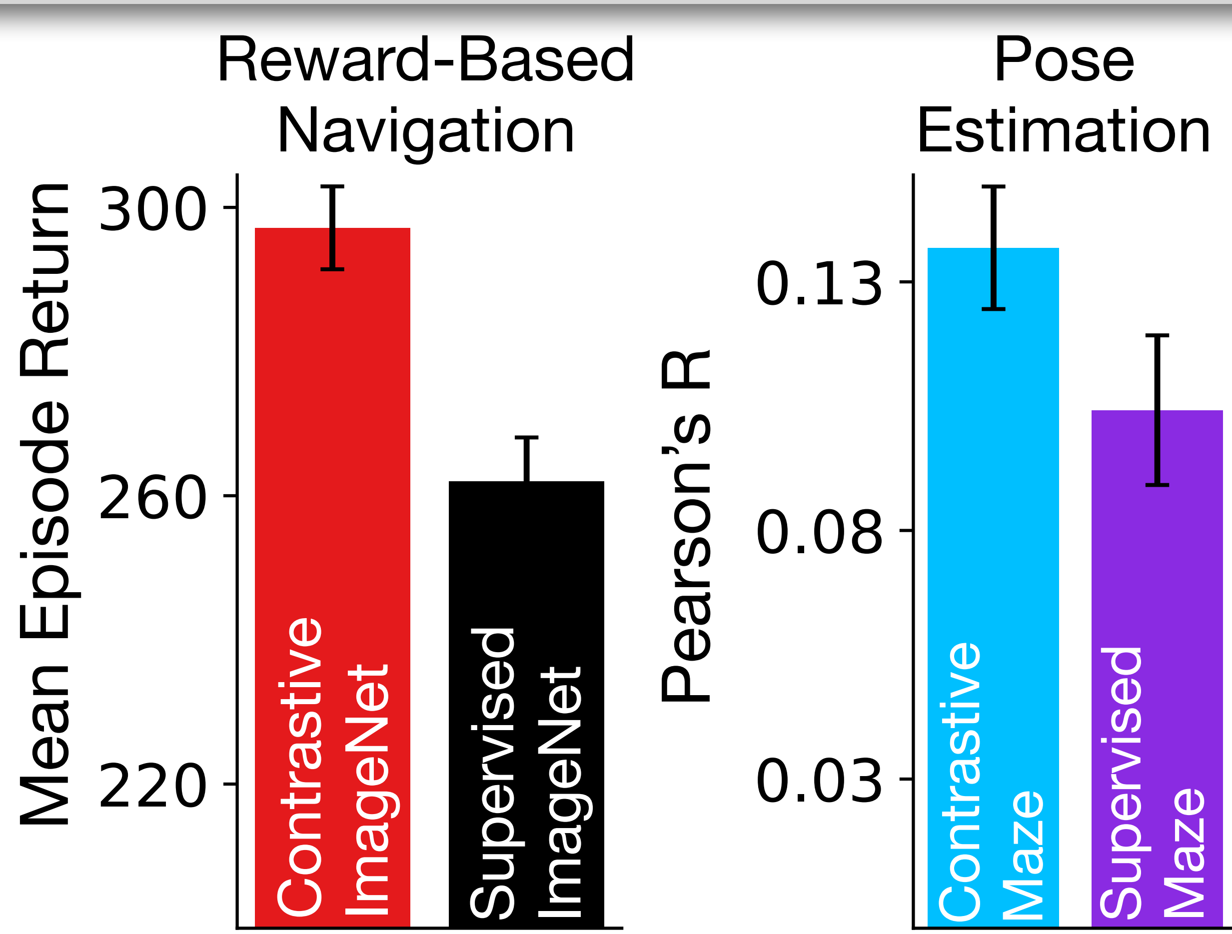
ImageNet



Evaluate

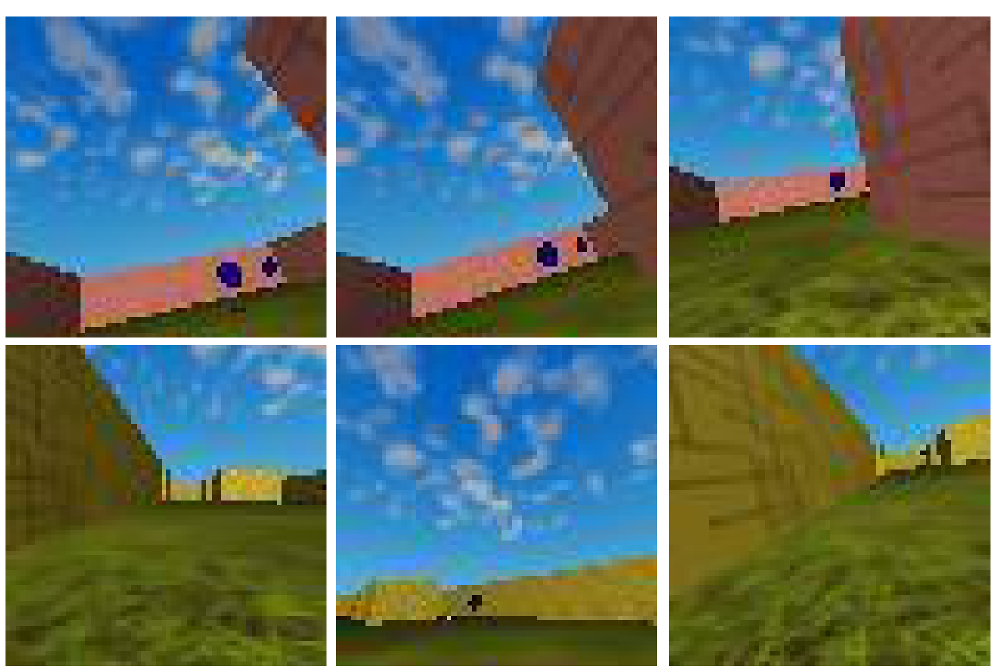


Contrastive Models Yield Better Transfer Performance



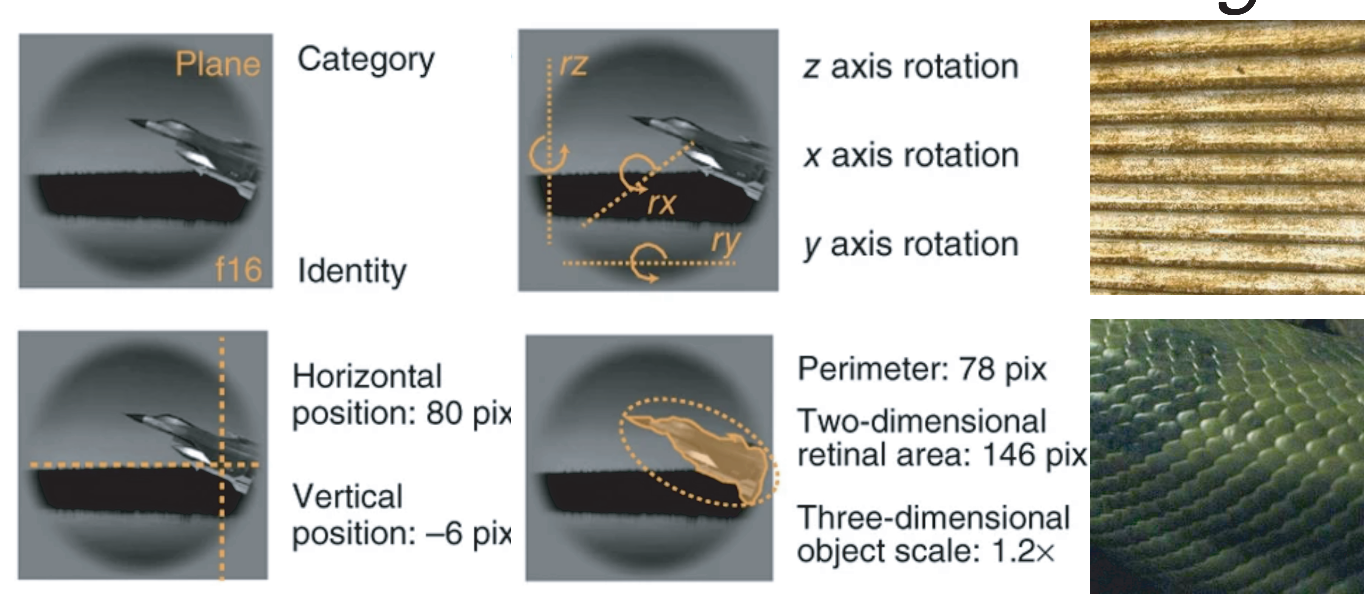
Train

Maze Environment



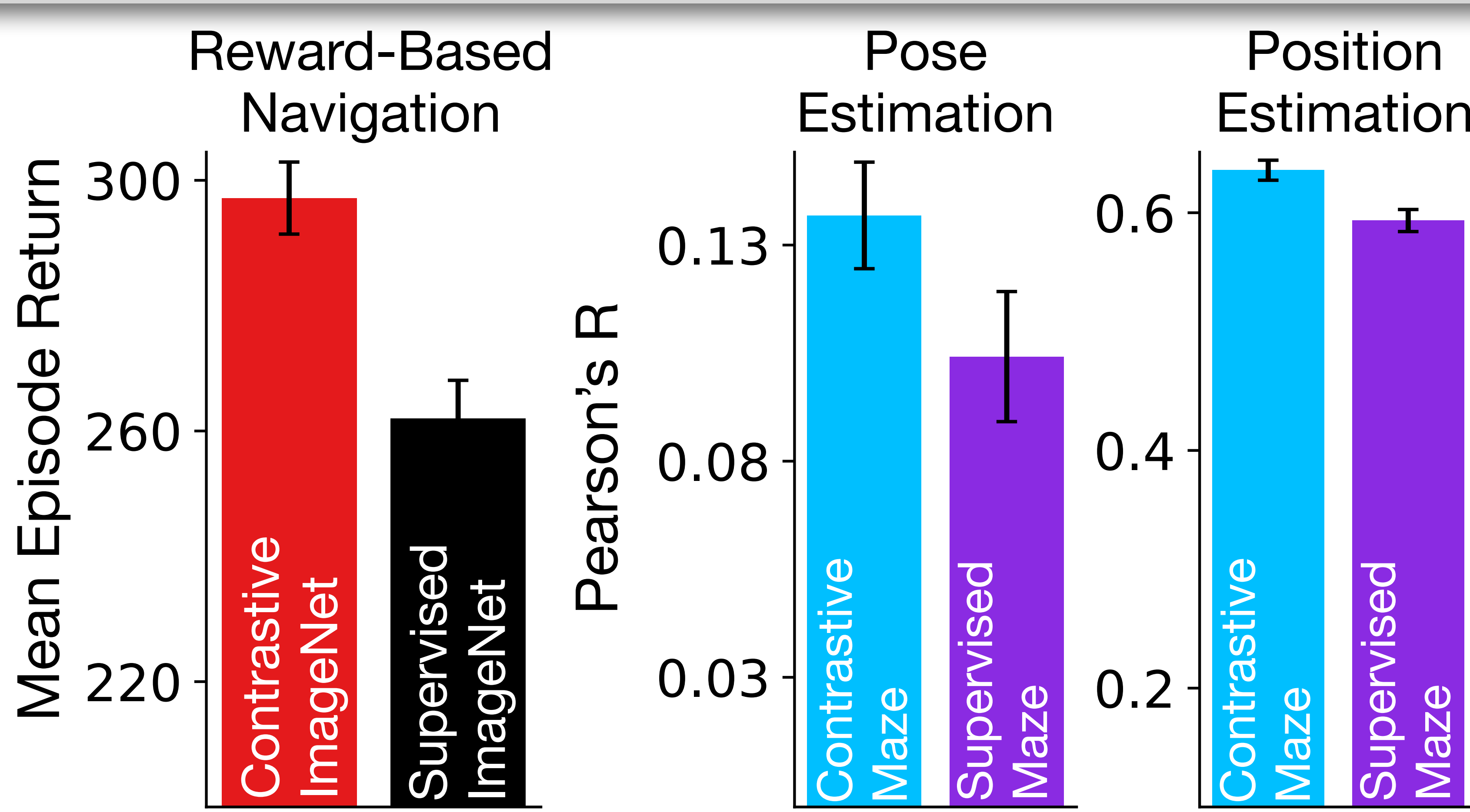
Evaluate

Visual Scene Understanding



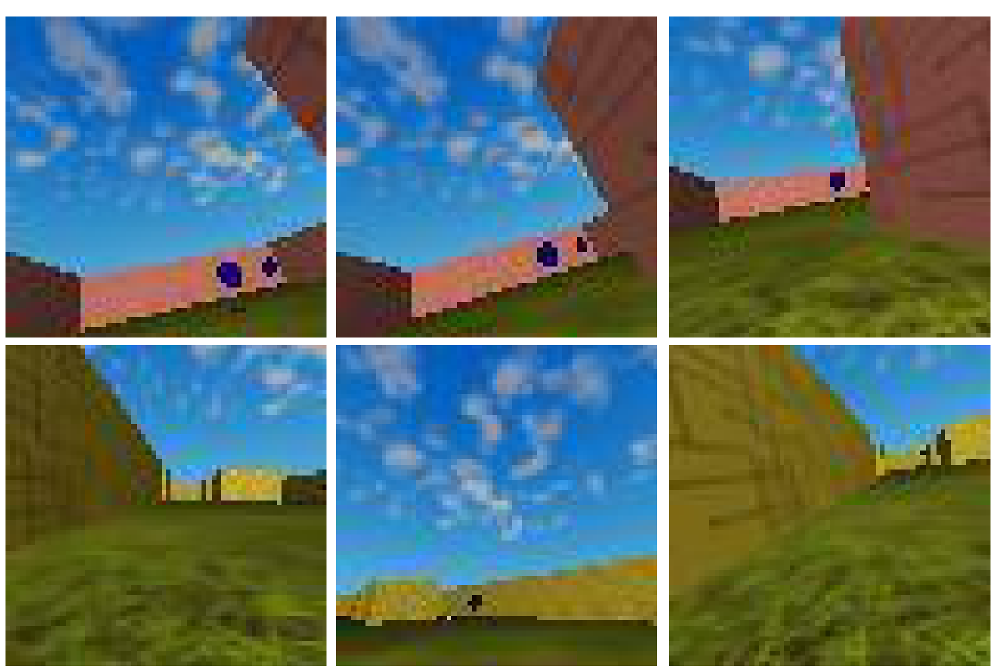
Object properties *Texture*

Contrastive Models Yield Better Transfer Performance



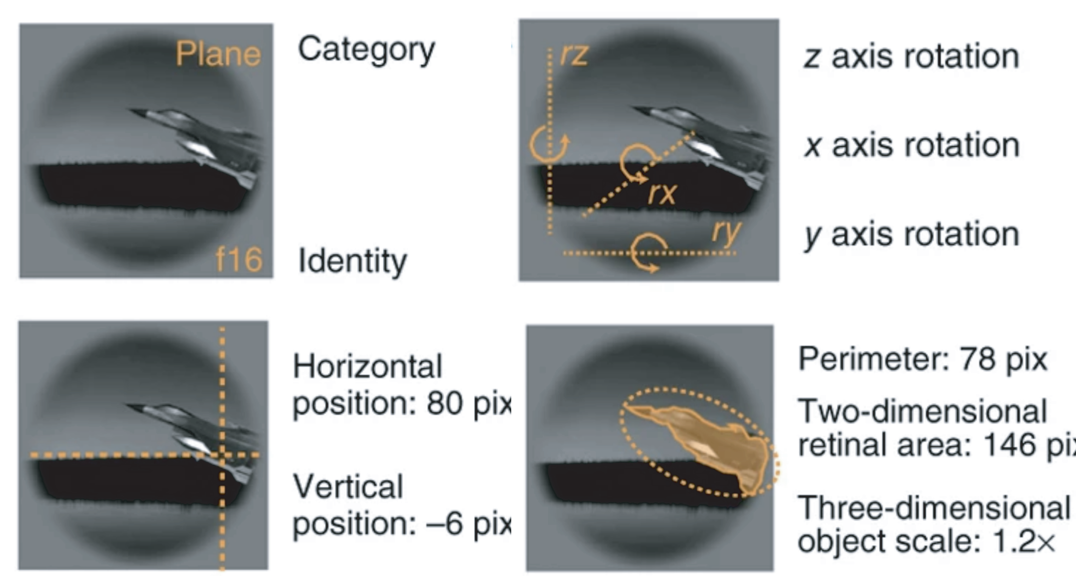
Train

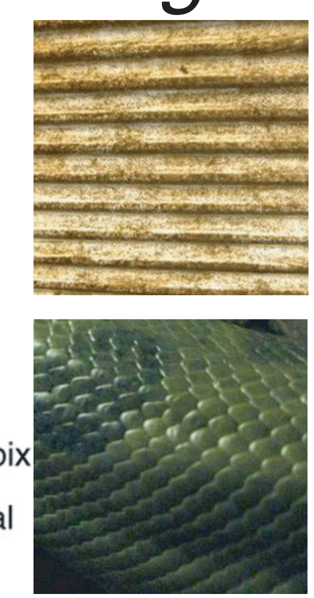
Maze Environment



Evaluate

Visual Scene Understanding

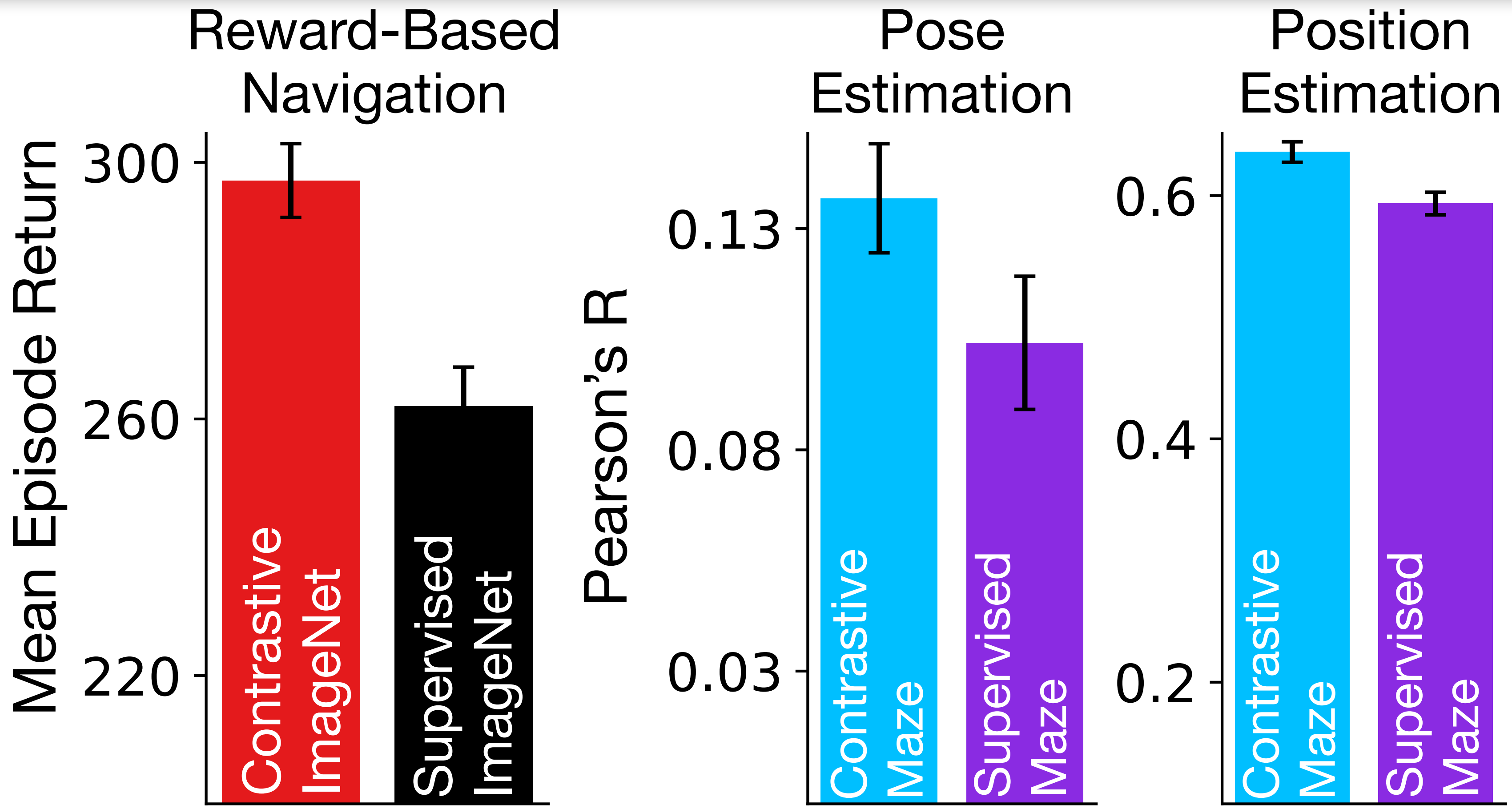


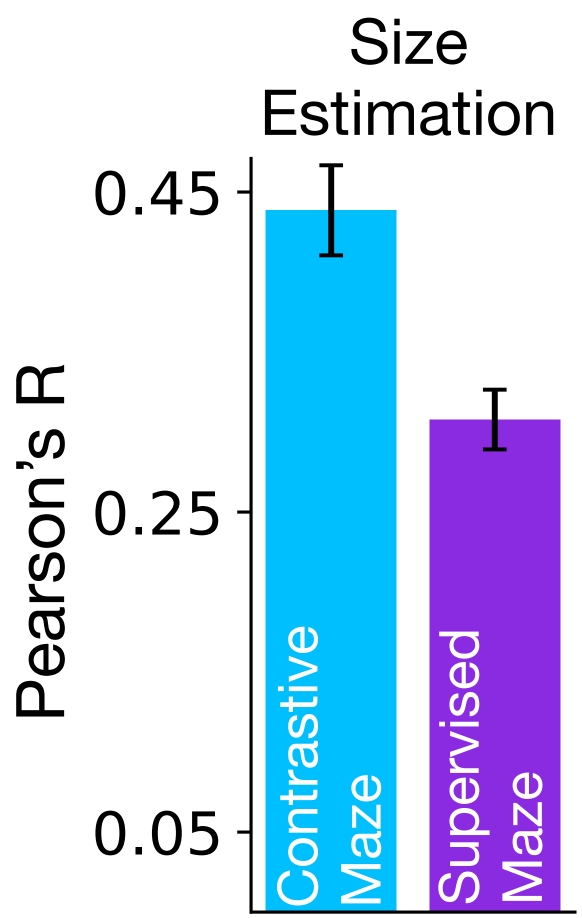


Object properties

Texture

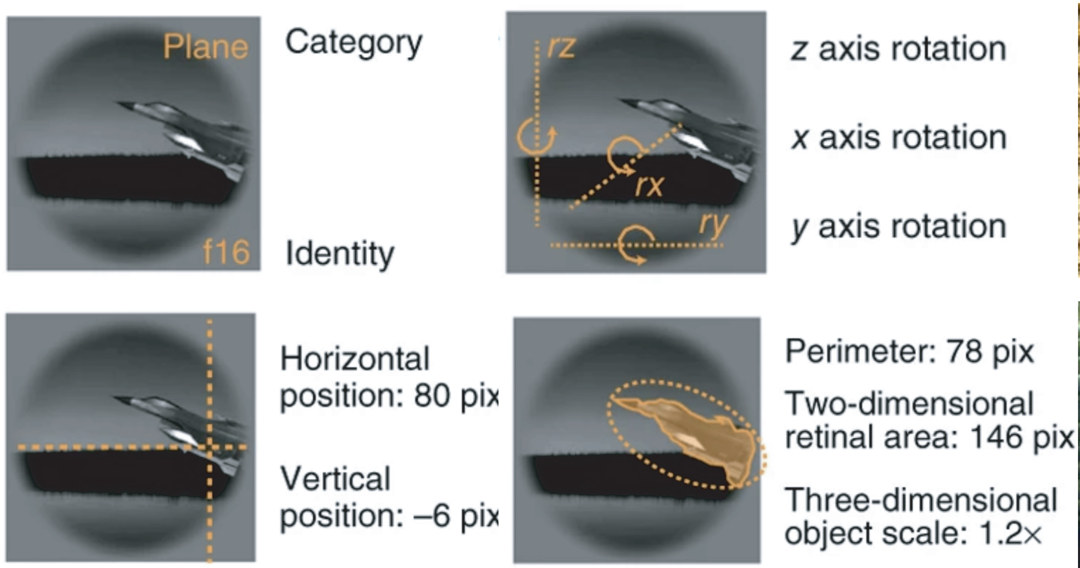
Contrastive Models Yield Better Transfer Performance



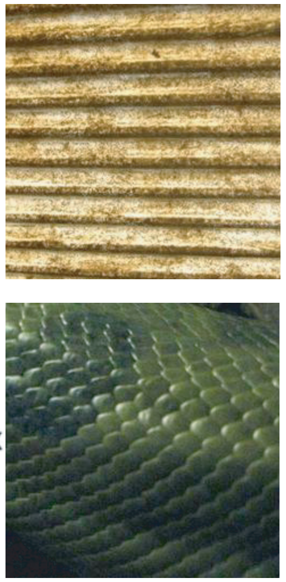


Evaluate

Visual Scene Understanding

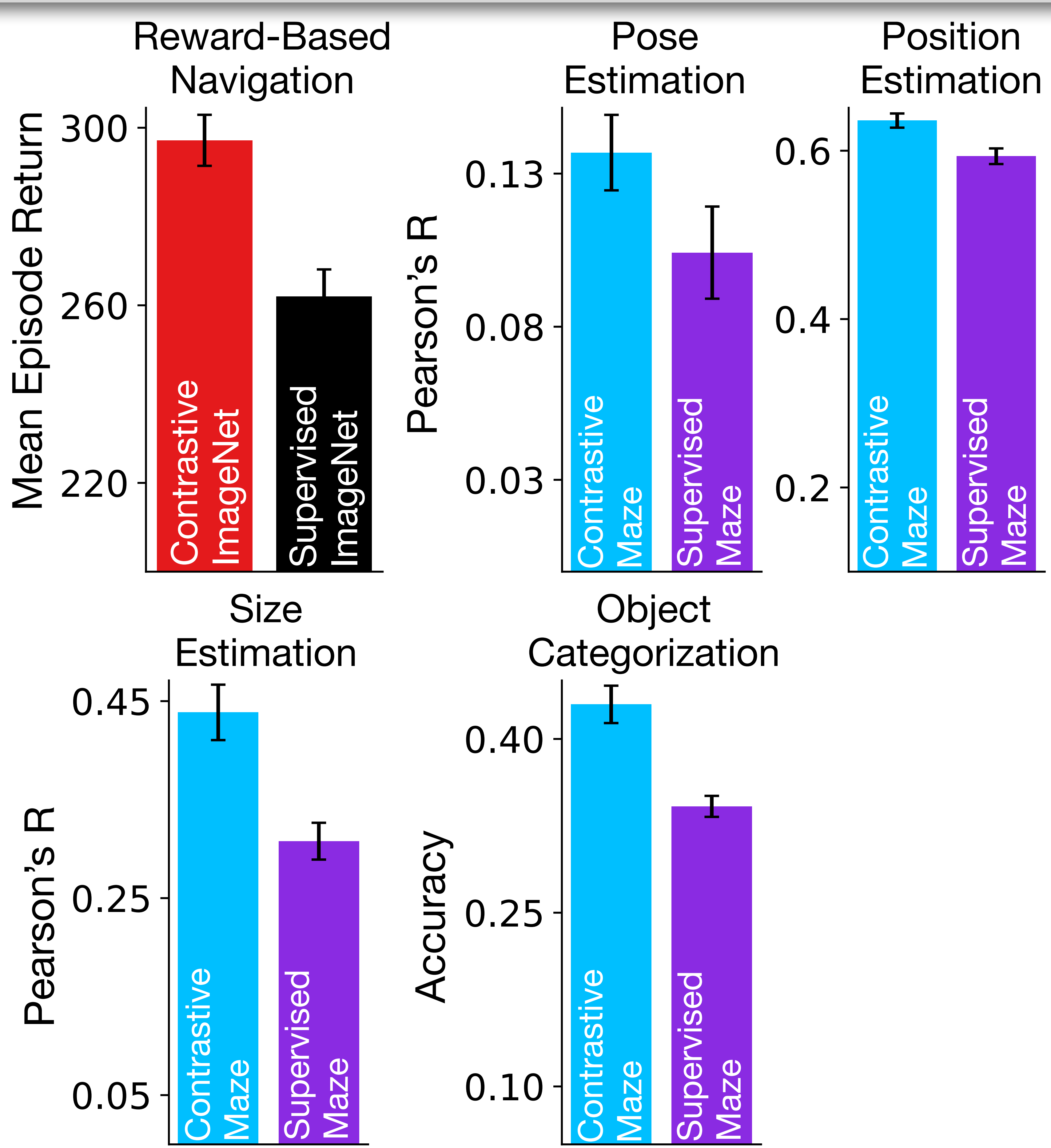


Object properties

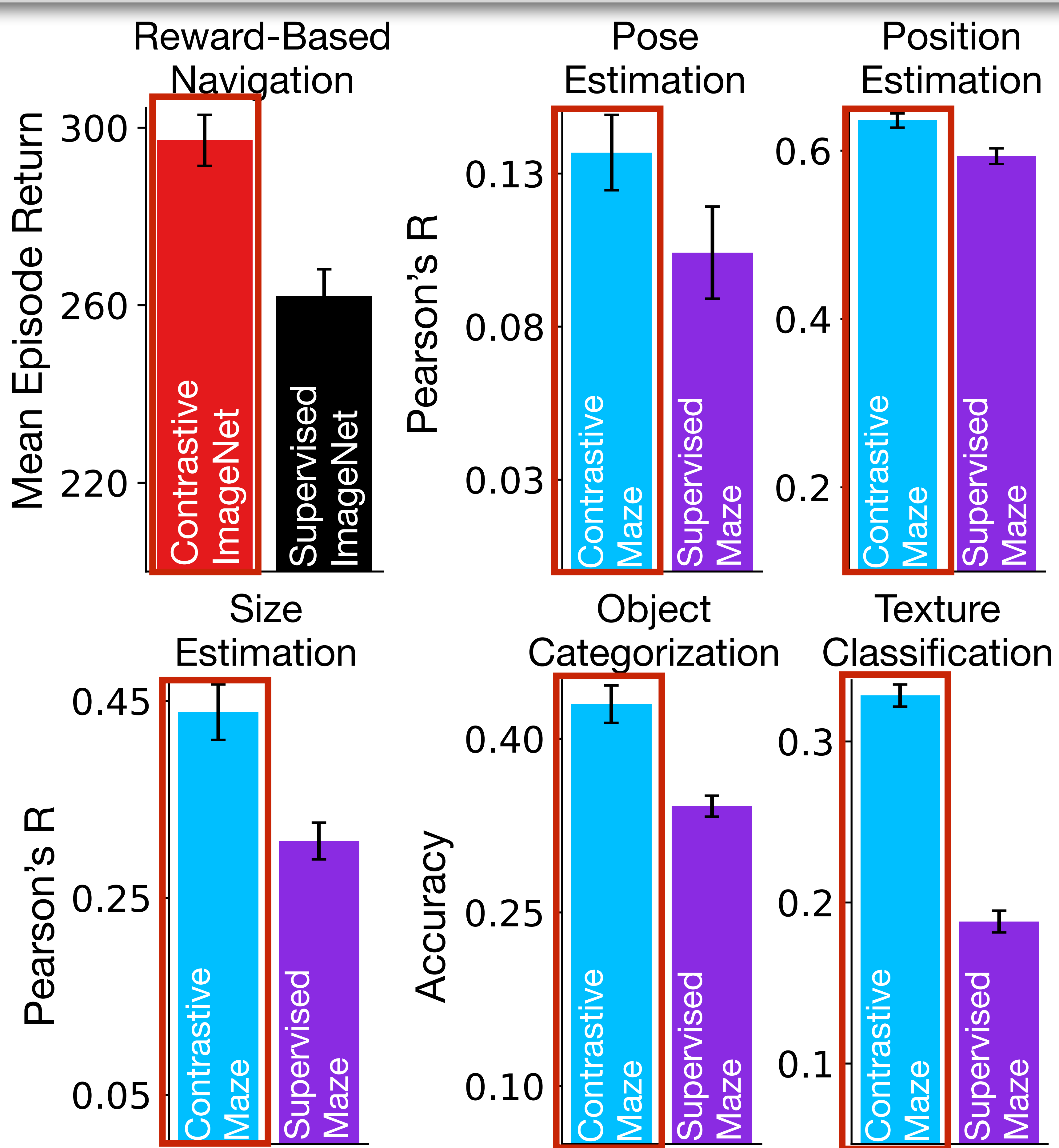


Texture

Contrastive Models Yield Better Transfer Performance



Contrastive Models Yield Better Transfer Performance



Outline

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

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

- Reusable Latent Representations for Primate Mental Simulation

- Heuristics for Interrogating Natural Intelligence

Outline

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

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

- Reusable Latent Representations for Primate Mental Simulation

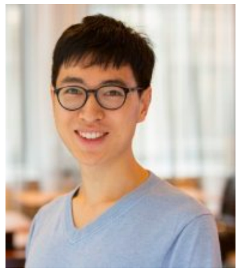
- Heuristics for Interrogating Natural Intelligence

Reusable Latent Representations for Primate Mental Simulation

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


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

Visually-Grounded Mental Simulation

Visually-Grounded Mental Simulation

<u>Infer:</u>
Has this ice block been out longer?

Visually-Grounded Mental Simulation

<u>Infer:</u>
Has this ice block been out longer?

Visually-Grounded Mental Simulation

<u>Infer:</u>
Has this ice block been out longer?

<u>Predict:</u>
Will this box support me?

Visually-Grounded Mental Simulation

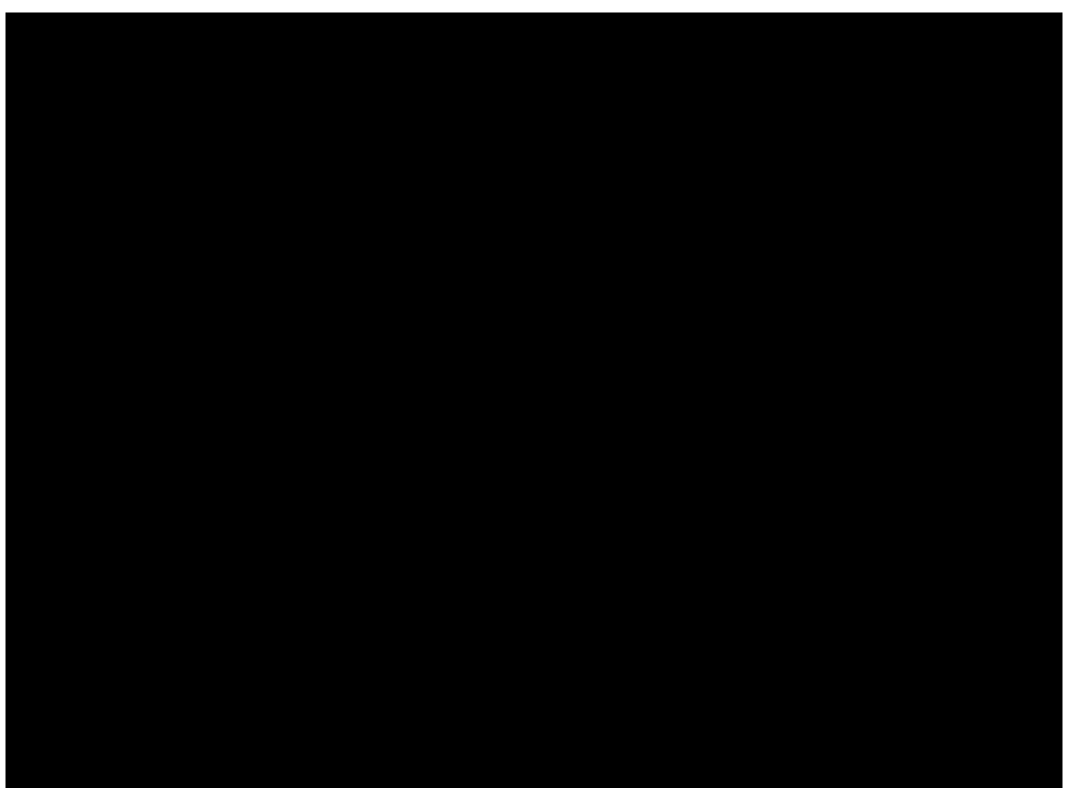
Infer:
Has this ice block been out longer?

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

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

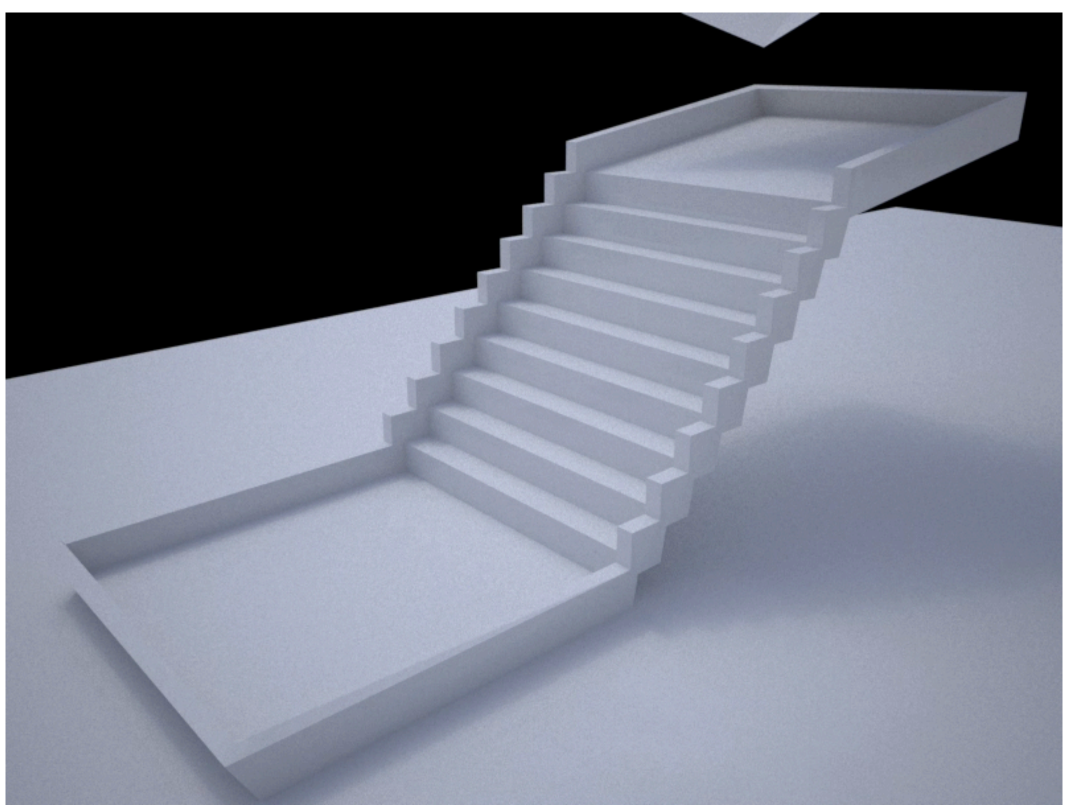


Plan:
How would I take these hats off the rack?

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

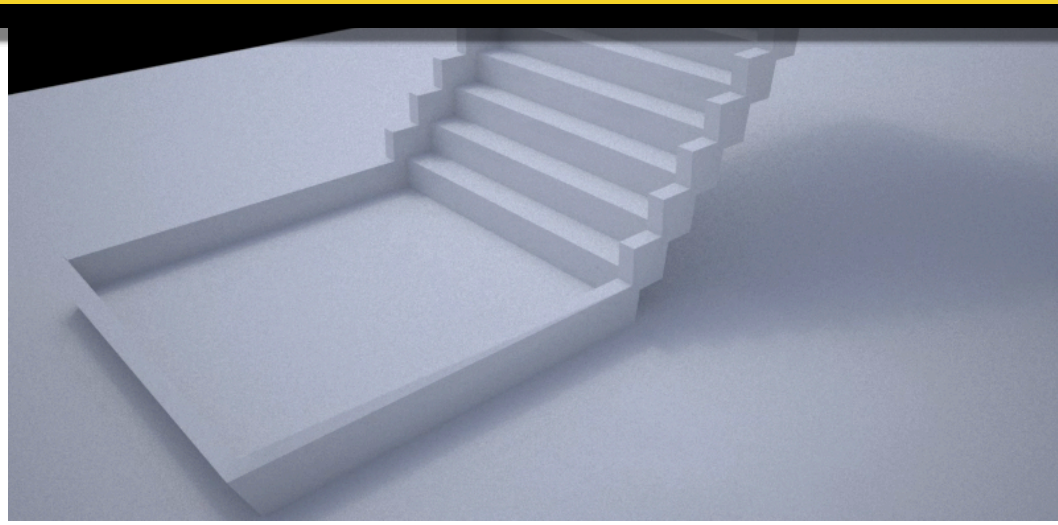


Plan:
How would I take these hats off the rack?

Predict:
Will this box support me?

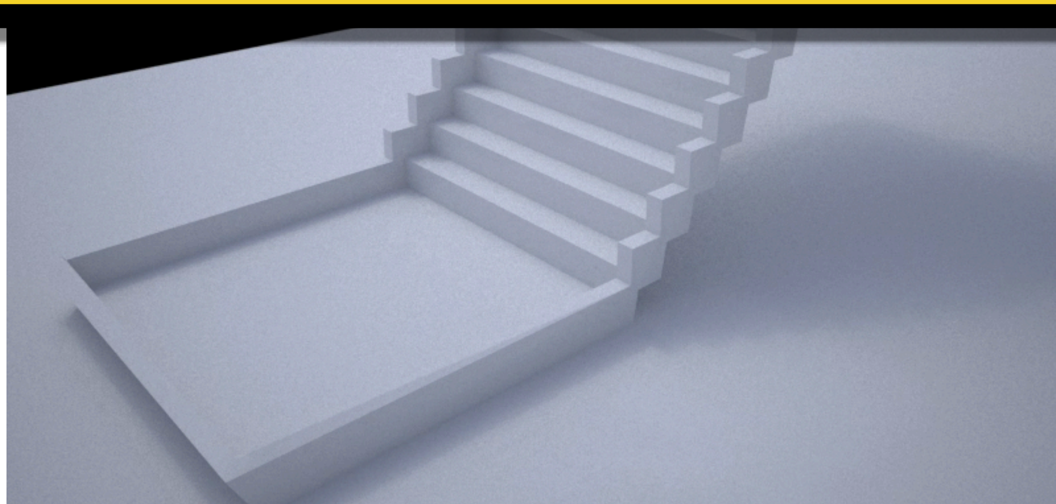
Neurobiological Puzzle:

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



Neurobiological Puzzle:

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



Defining Hypotheses

Defining Hypotheses

"Sensory-Cognitive Networks"

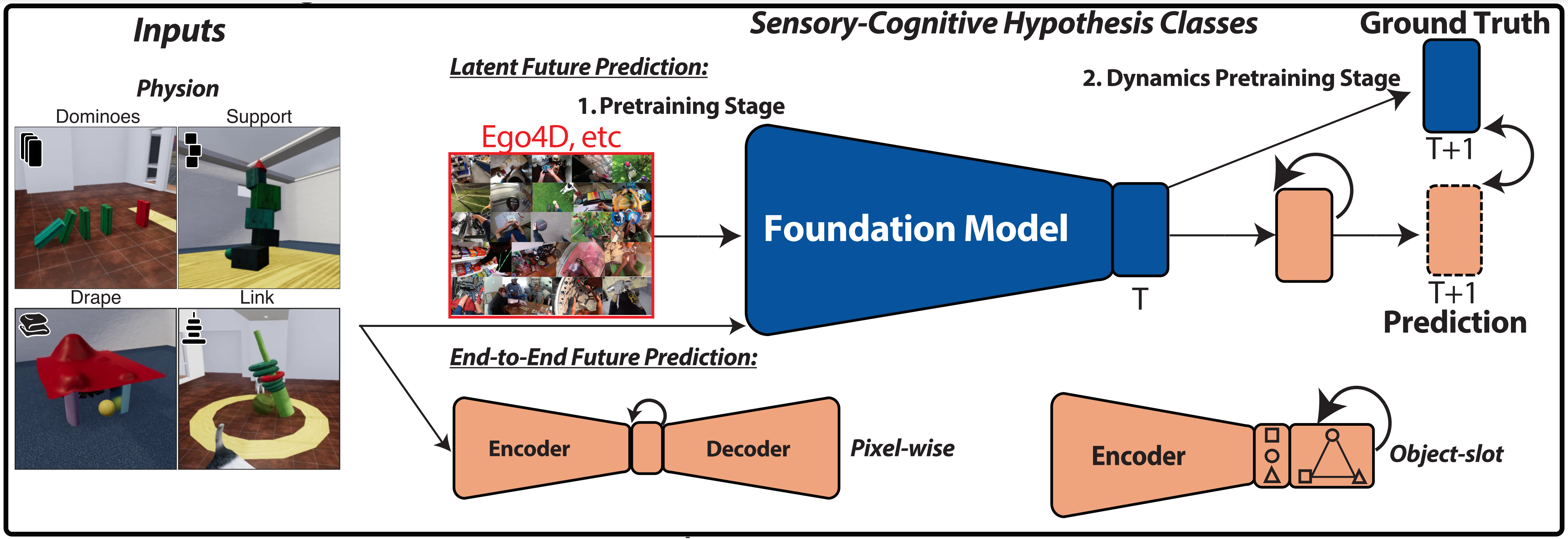
R1 (Input-Driven): Take in unstructured visual inputs across a range of physical phenomena.

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

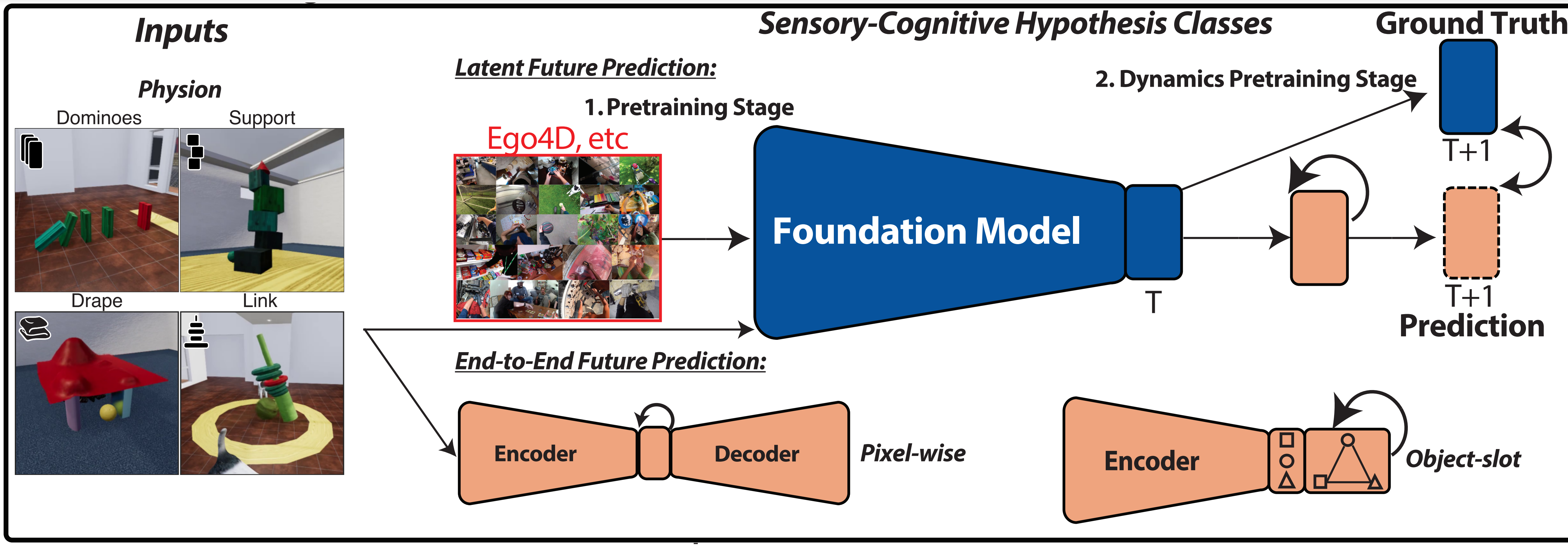
R3 (Neural Representations): Consist of internal units that can be compared to biological units (e.g. containing "artificial neurons").

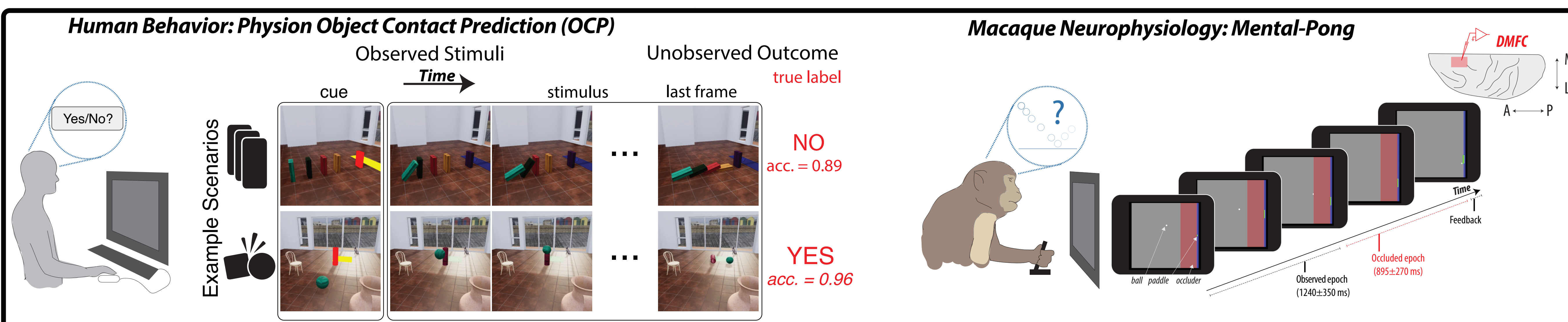
Overall Approach: Sensory-Cognitive Hypotheses

Overall Approach: Sensory-Cognitive Hypotheses

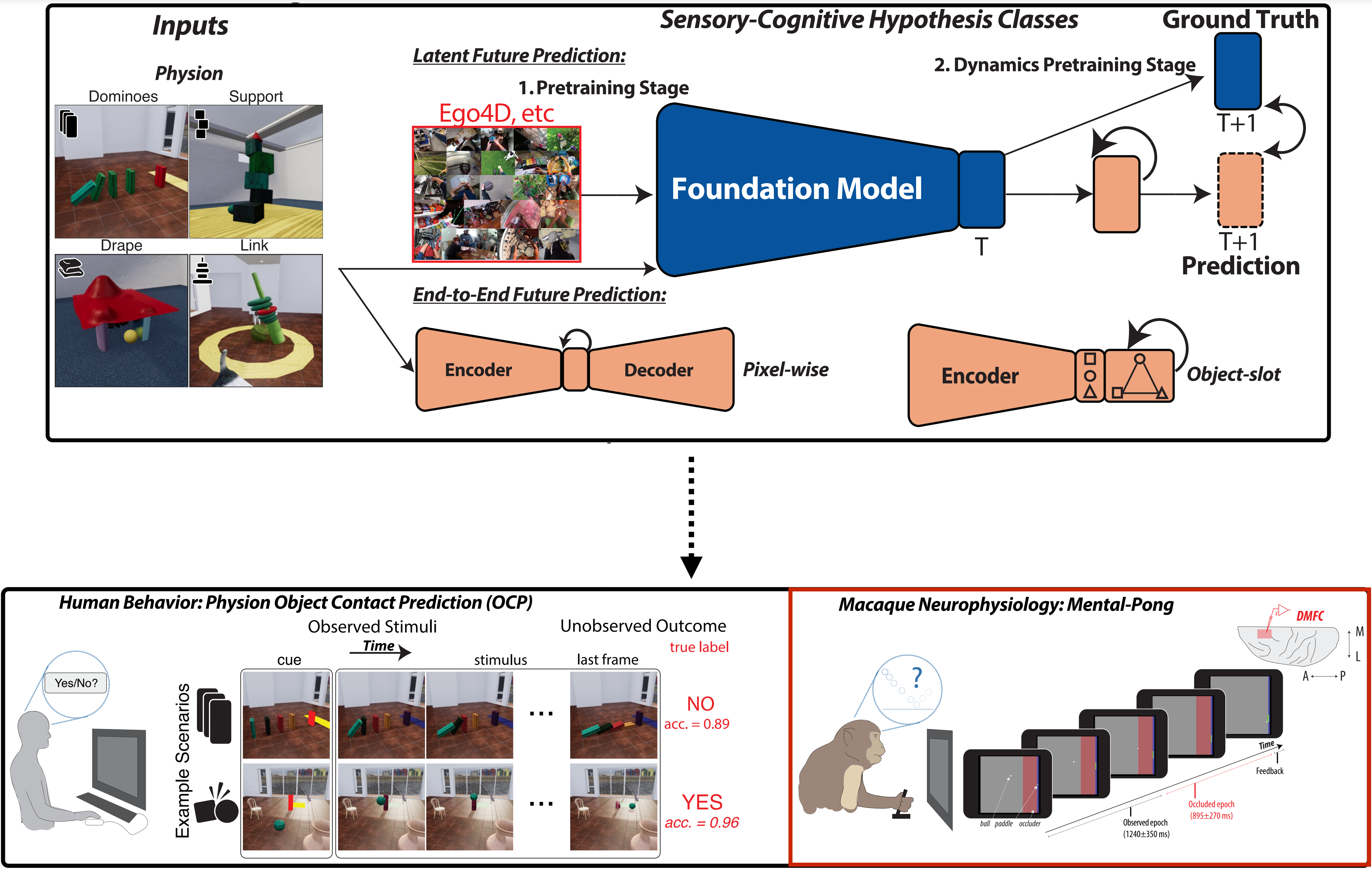


Overall Approach: Sensory-Cognitive Hypotheses

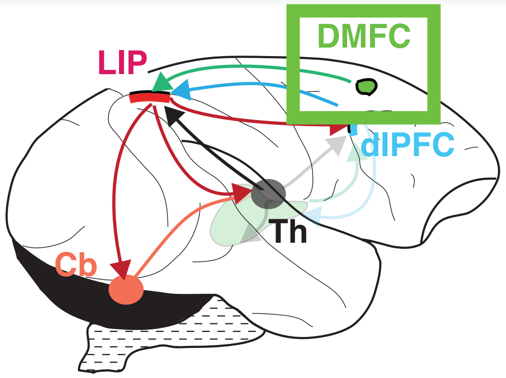




Macaque Neurophysiology: Mental Pong



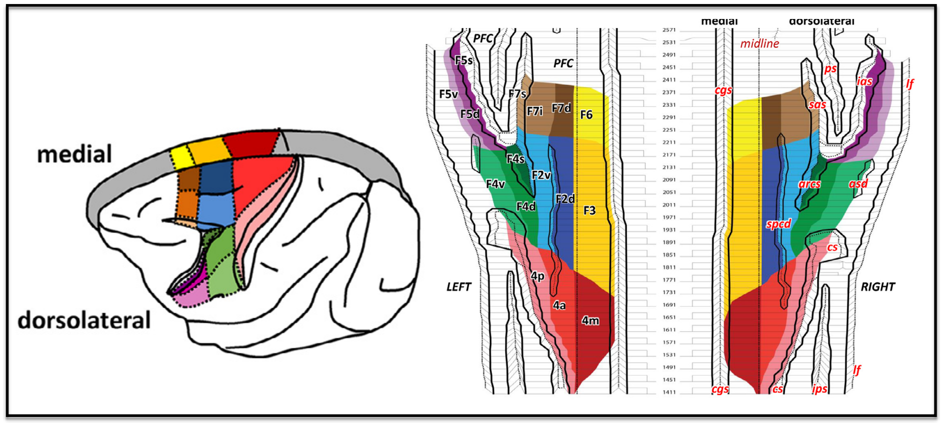
Macaque Neurophysiology: Mental Pong



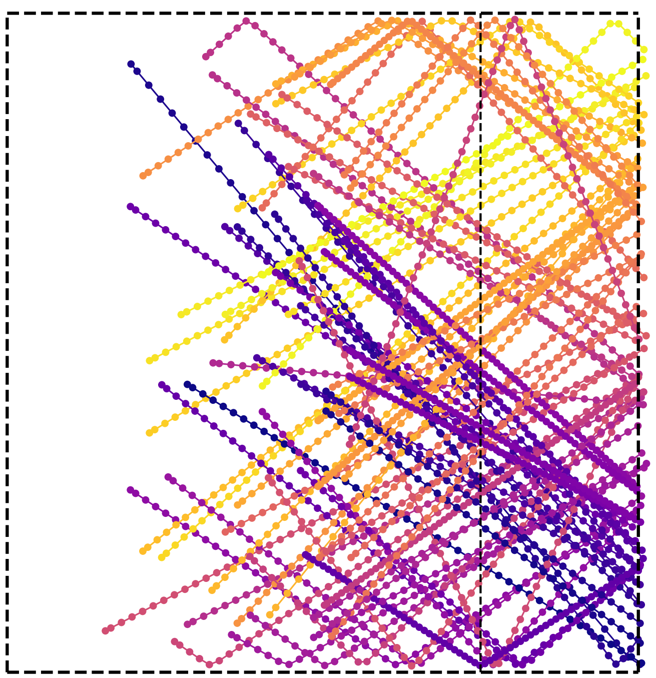
Fronto-Parietal Network

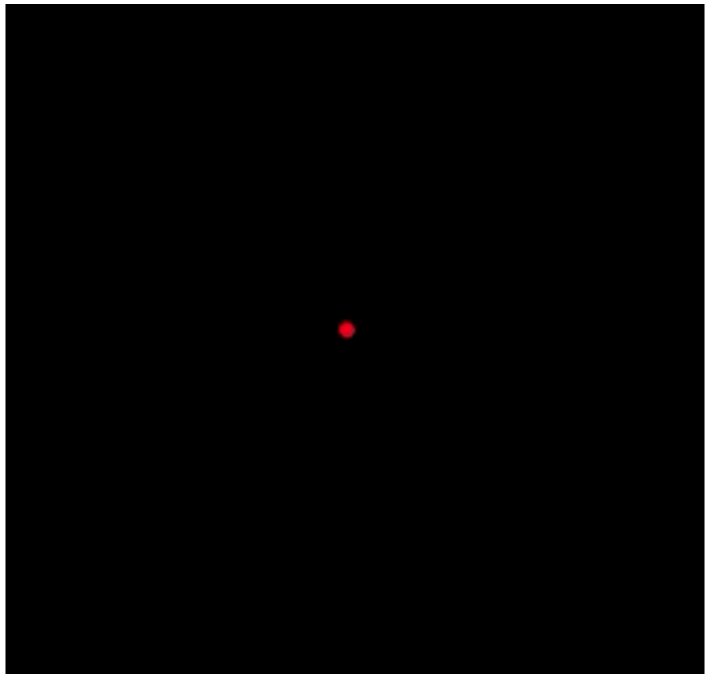
Dorsomedial frontal cortex (DMFC)

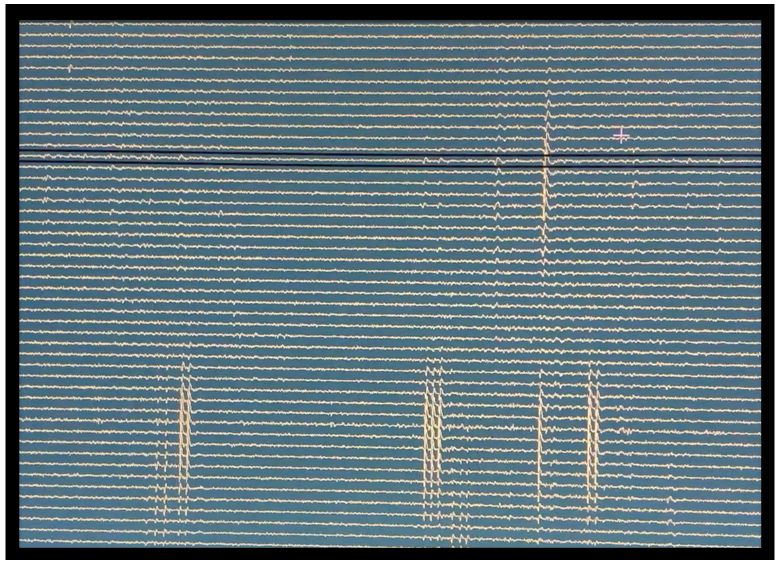
Monkey P *Monkey M*



79 conditions



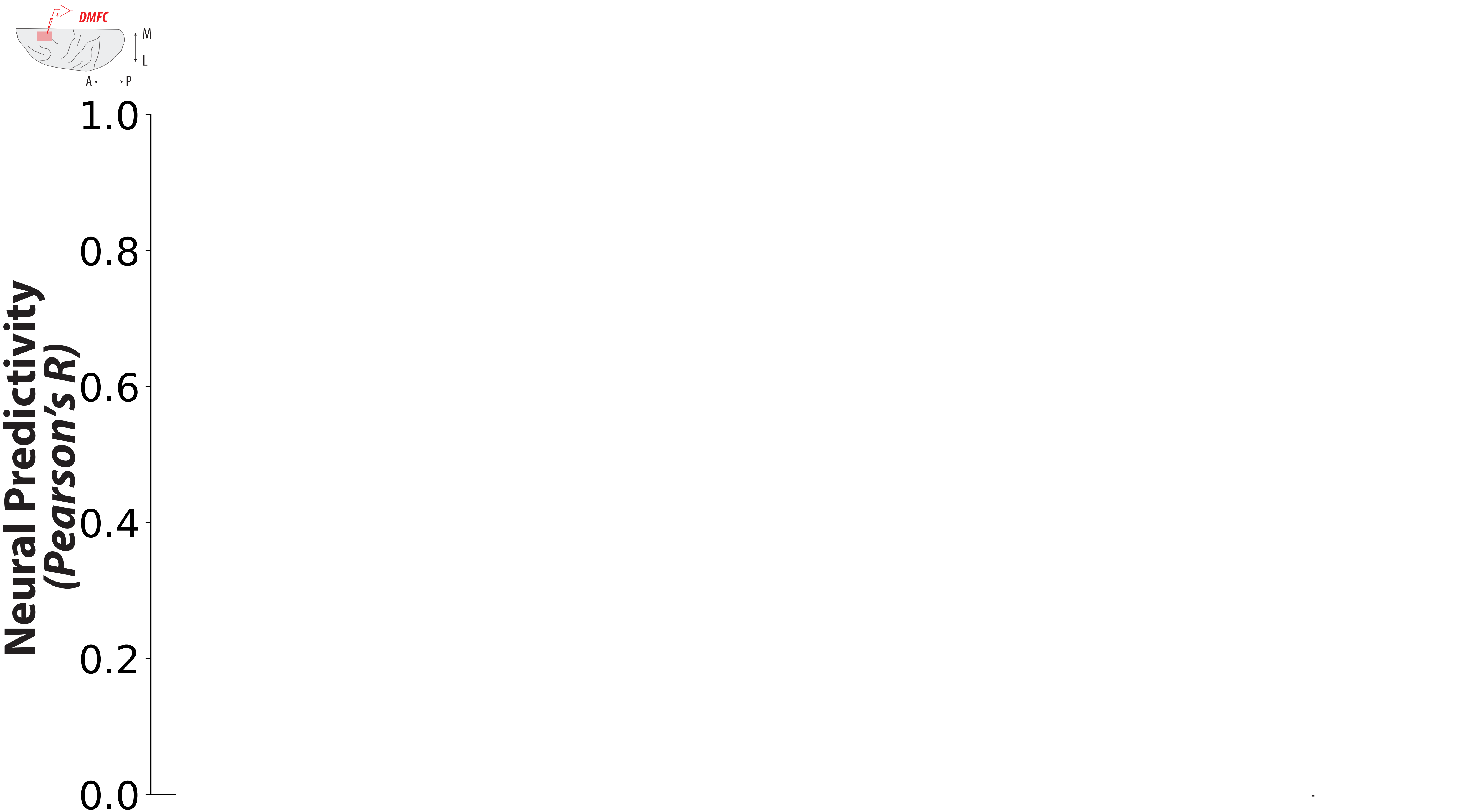




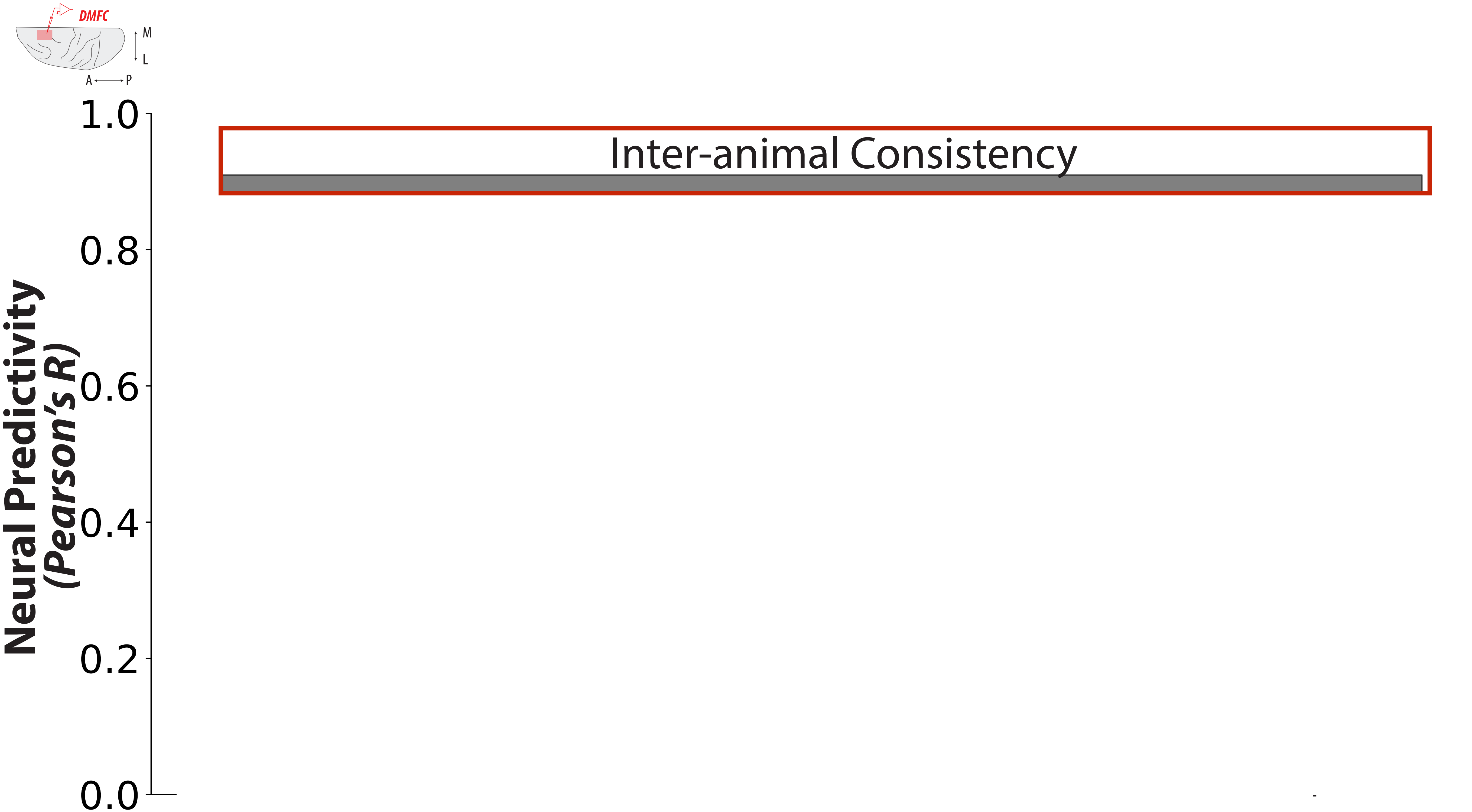
- *Data from two male adult monkeys*
- *79 subsampled M-Pong conditions*
- *64 channel v-probe (monkey P) and 384-channel Neuropixel probe (monkey M)*
- *Total of 1889 stable & reliable neurons recorded from DMFC*

Rishi Rajalingham

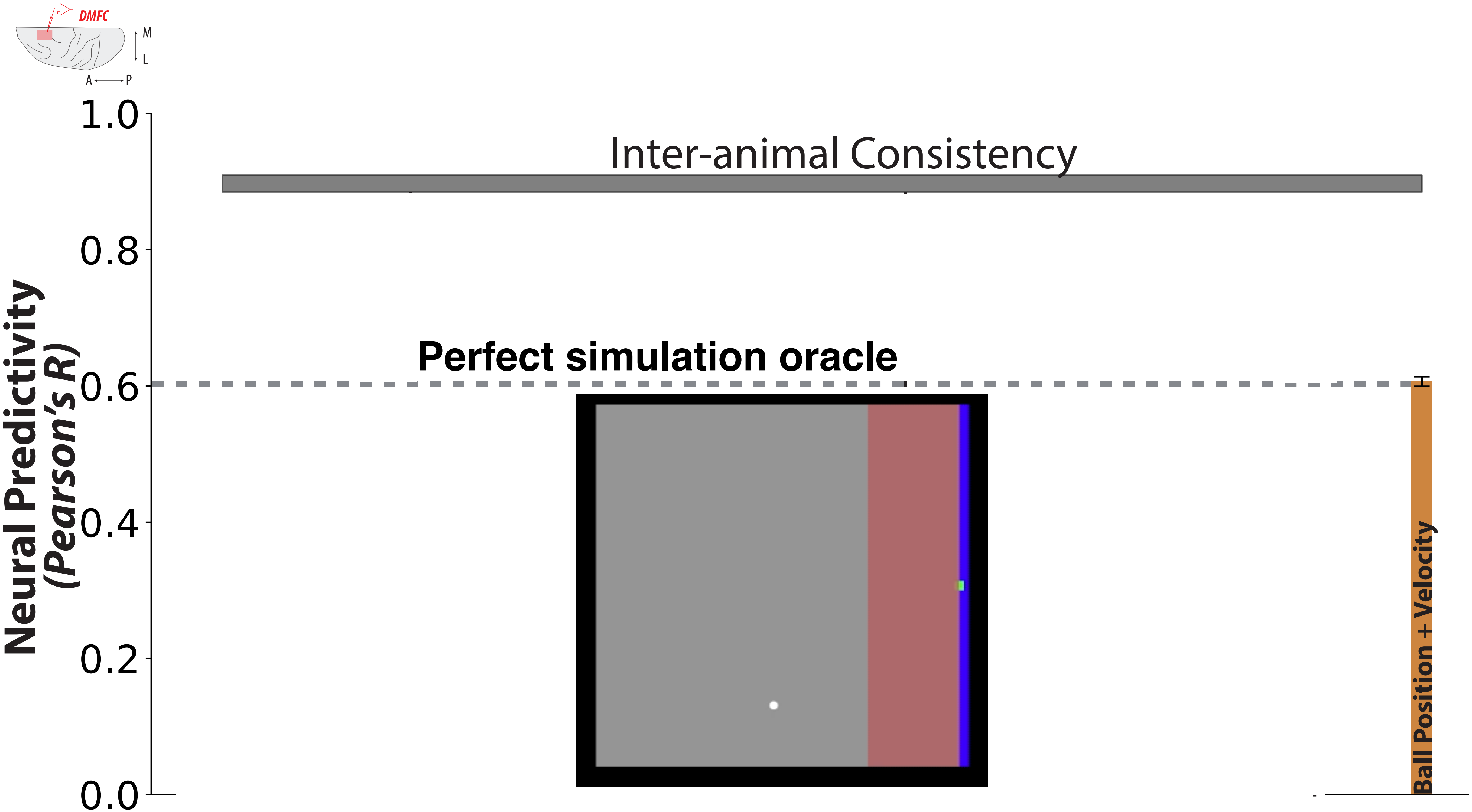
Macaque Neurophysiology: Mental Pong



Macaque Neurophysiology: Mental Pong



Perfect Simulation Oracle Predicts Neural Data Well

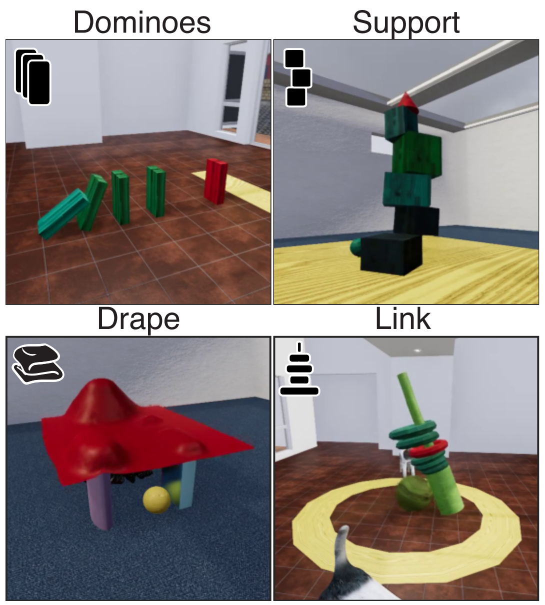


Functional Constraint Hypotheses

Inputs

Sensory-Cognitive Hypothesis Classes

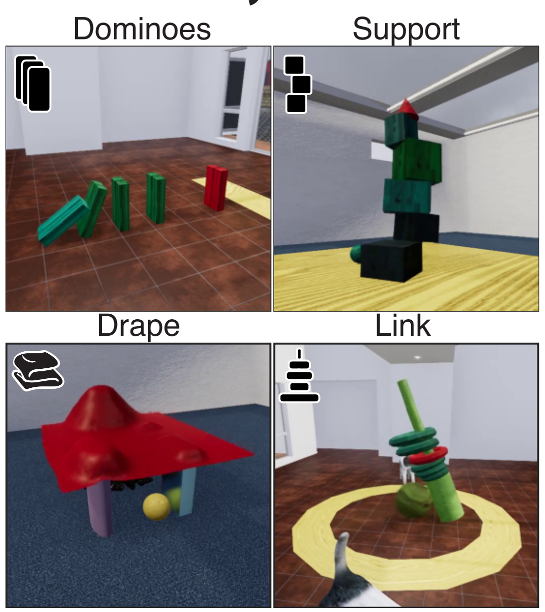
Physion



Hypothesis Class 1: Pixel-wise Future Prediction

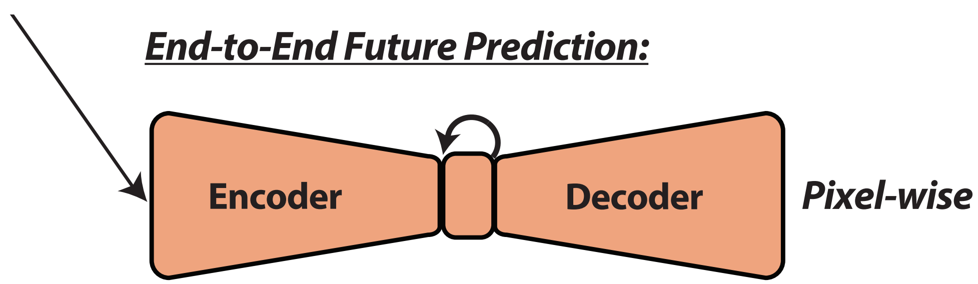
Inputs

Physion



Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



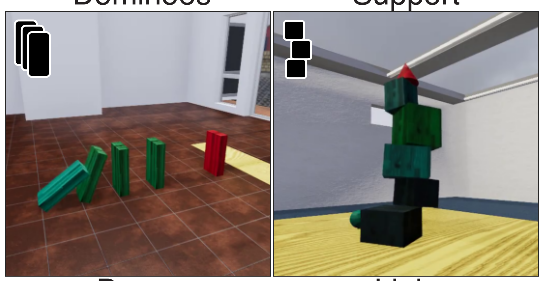
Hypothesis Class I: Pixel-wise Future Prediction

Inputs

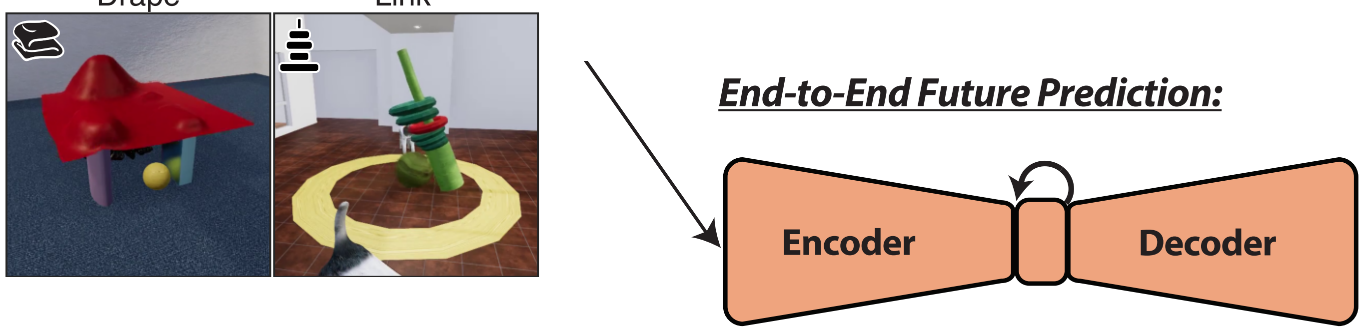
Sensory-Cognitive Hypothesis Classes

Physion

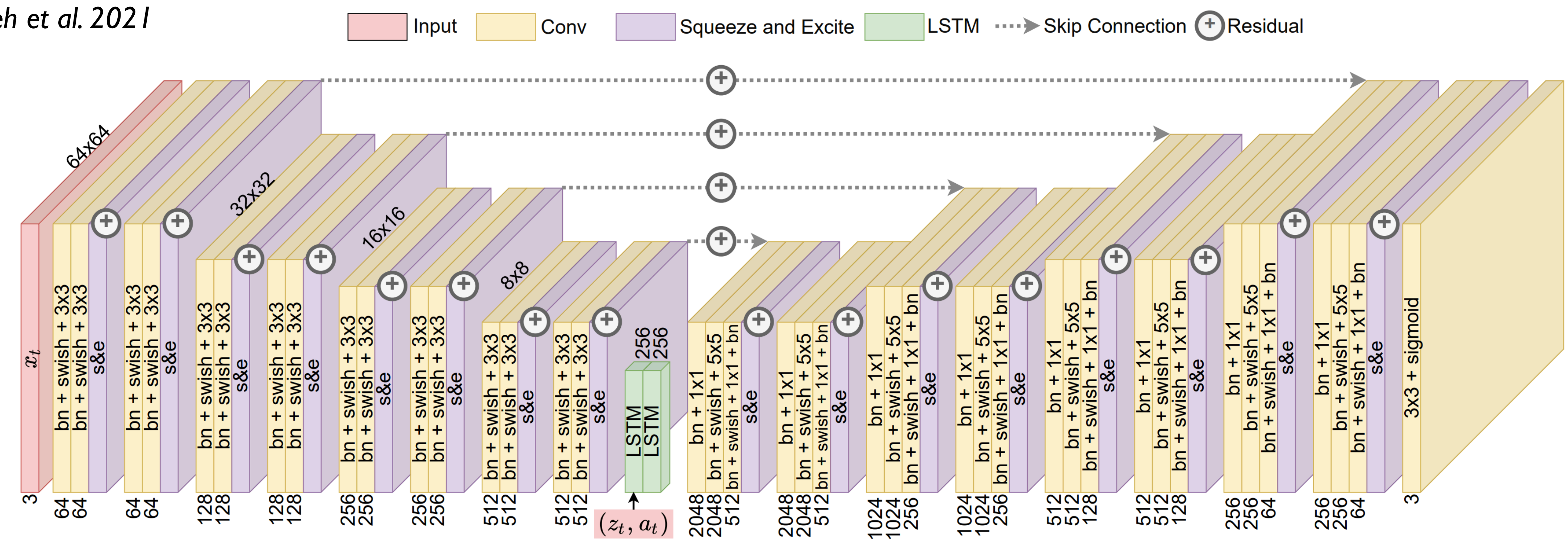
Dominoes Support



Drape Link



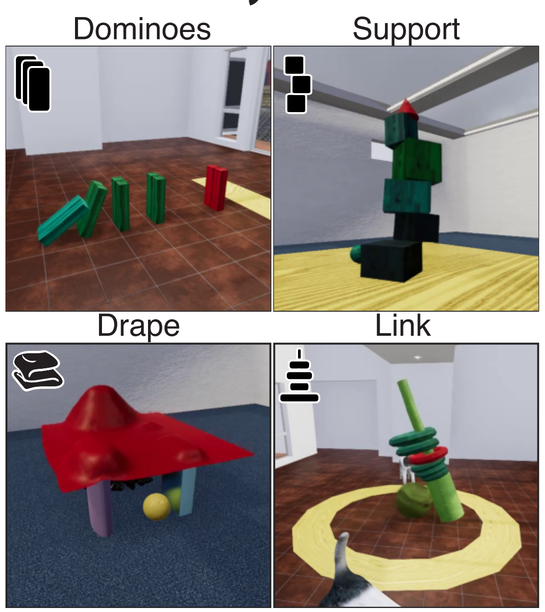
Babaeizadeh et al. 2021



Hypothesis Class I: Pixel-wise Future Prediction

Inputs

Physion

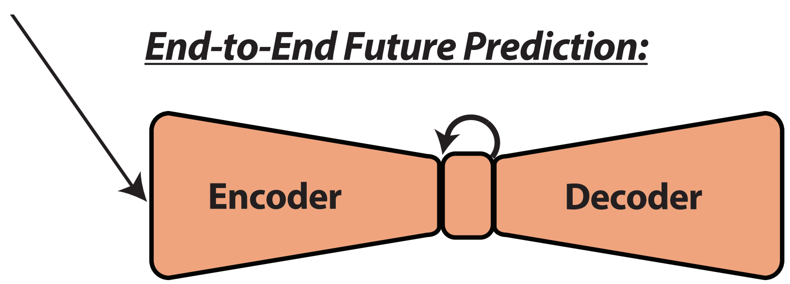


Sensory-Cognitive Hypothesis Classes

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

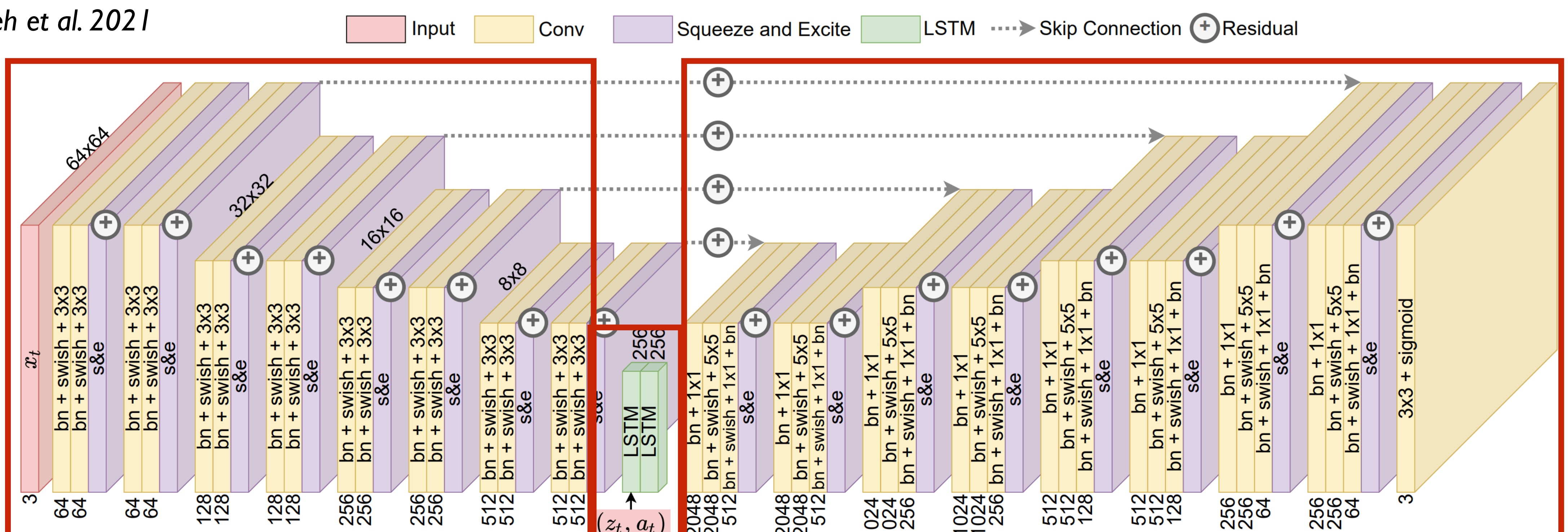
$$\mathcal{L} = \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\|_2^2 + \beta D_{KL}(\mathcal{N}(\mu, \sigma)\|\mathcal{N}(0, \mathbf{I}))$$

End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021

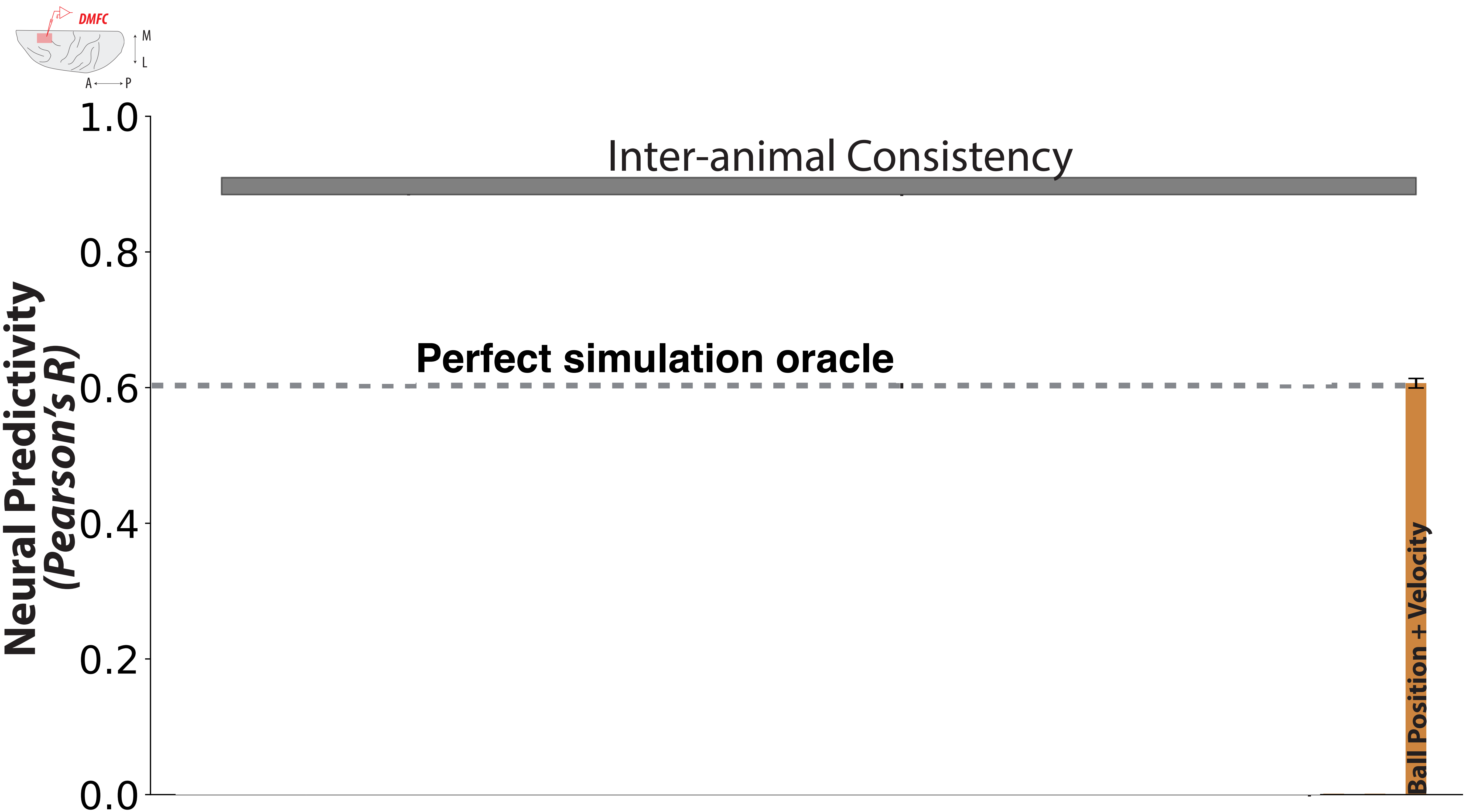


Visual Encoder ("Sensory")

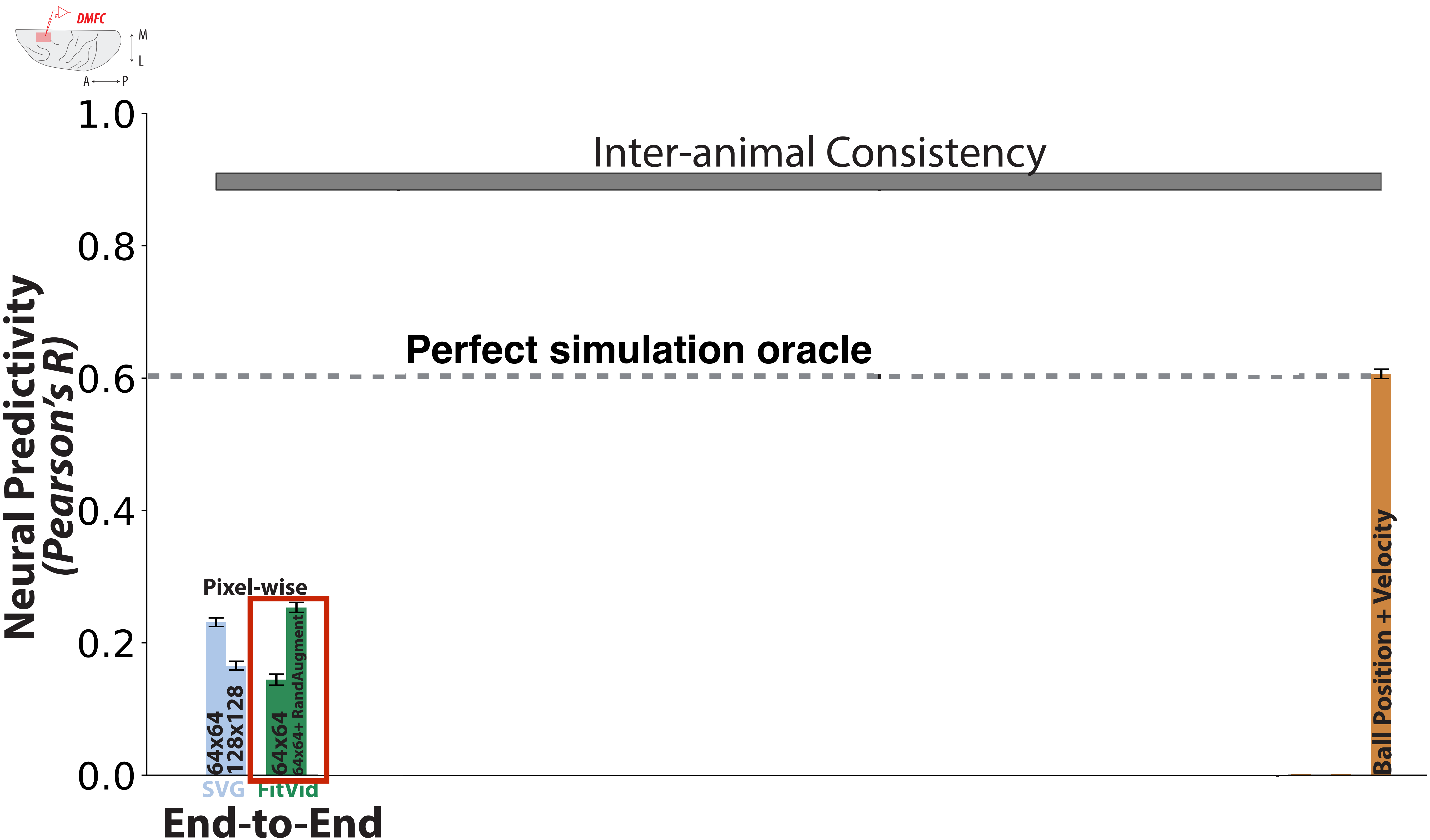
Dynamics Predictor ("Cognitive")

Frame Decoder ("Objective/Behavior")

Physical Simulation Oracles Predict Neural Data Well

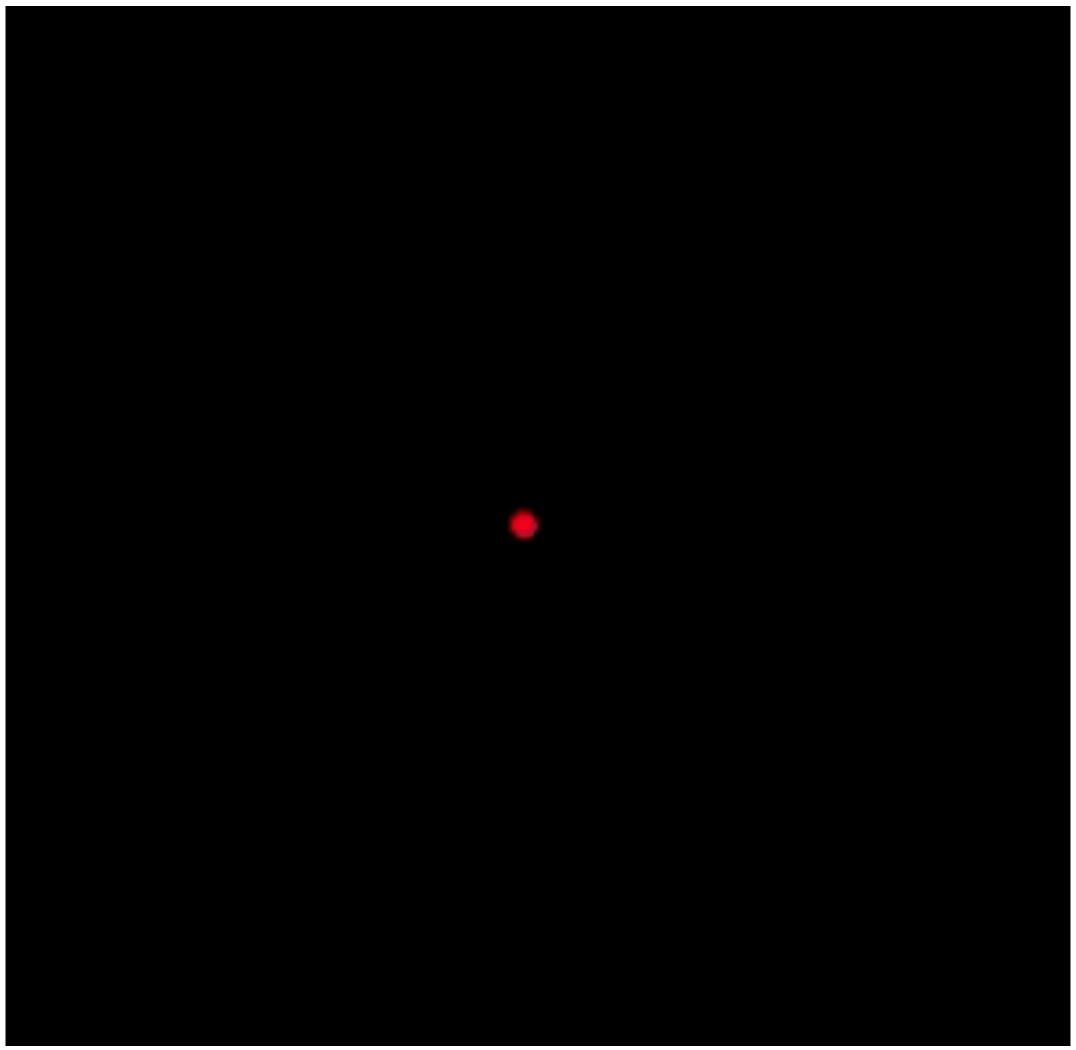


Pixel-wise Future Prediction Poorly Predicts Neurons

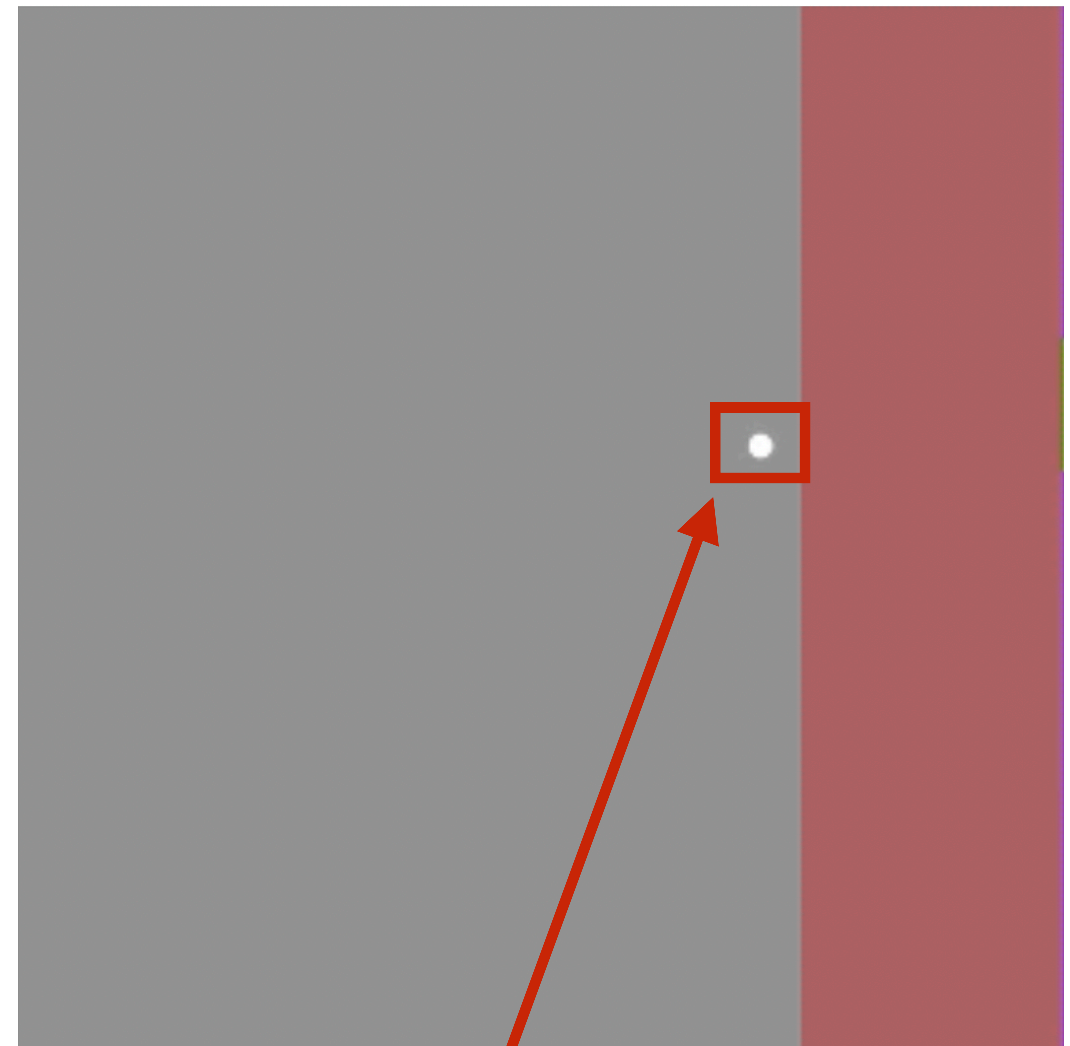


…and they struggle to generalize to Pong

Input Frames

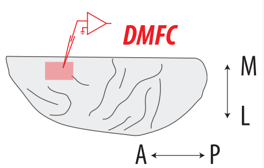


Predicted Frames

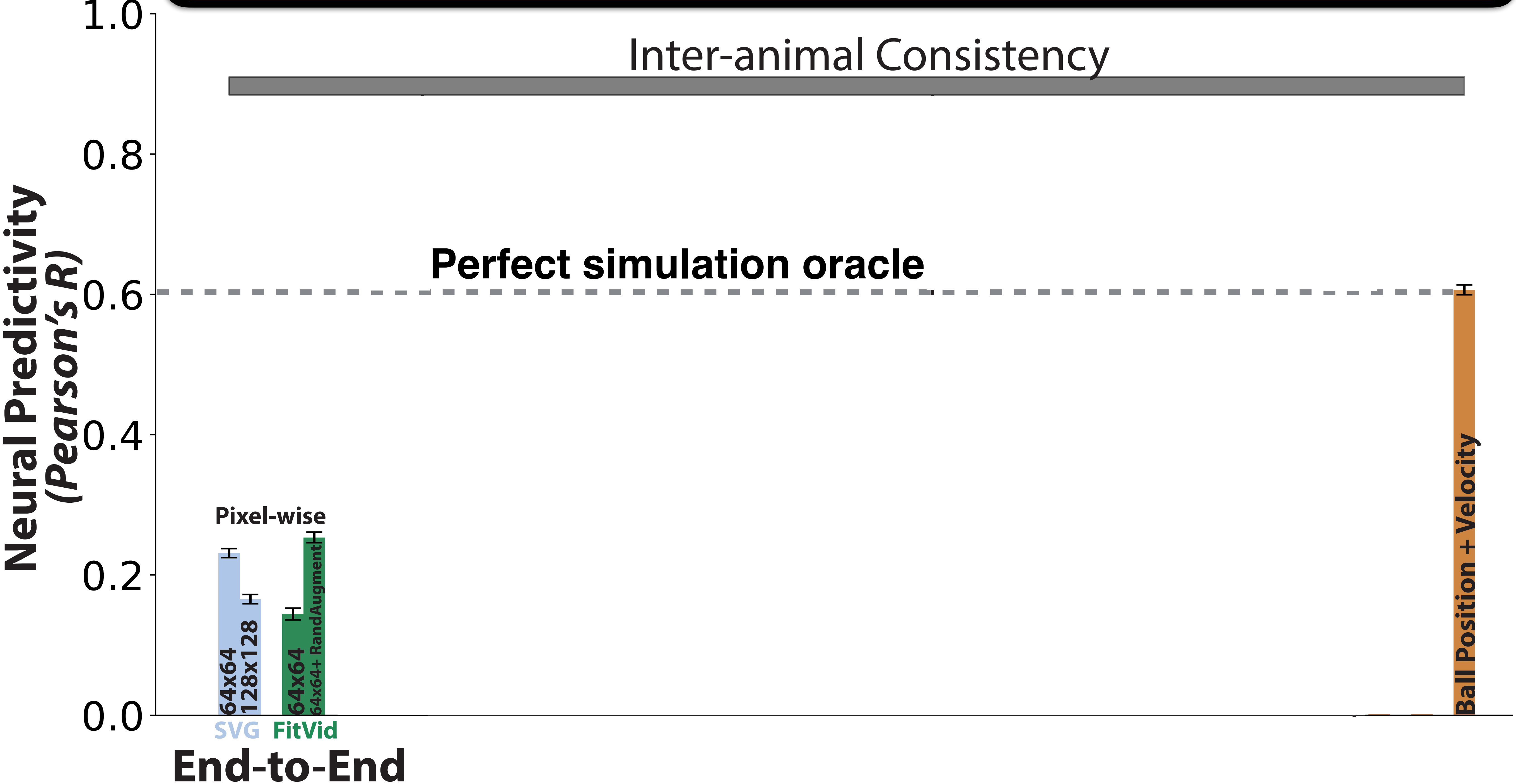


Ball stops at final input frame, in the model's "imagination"

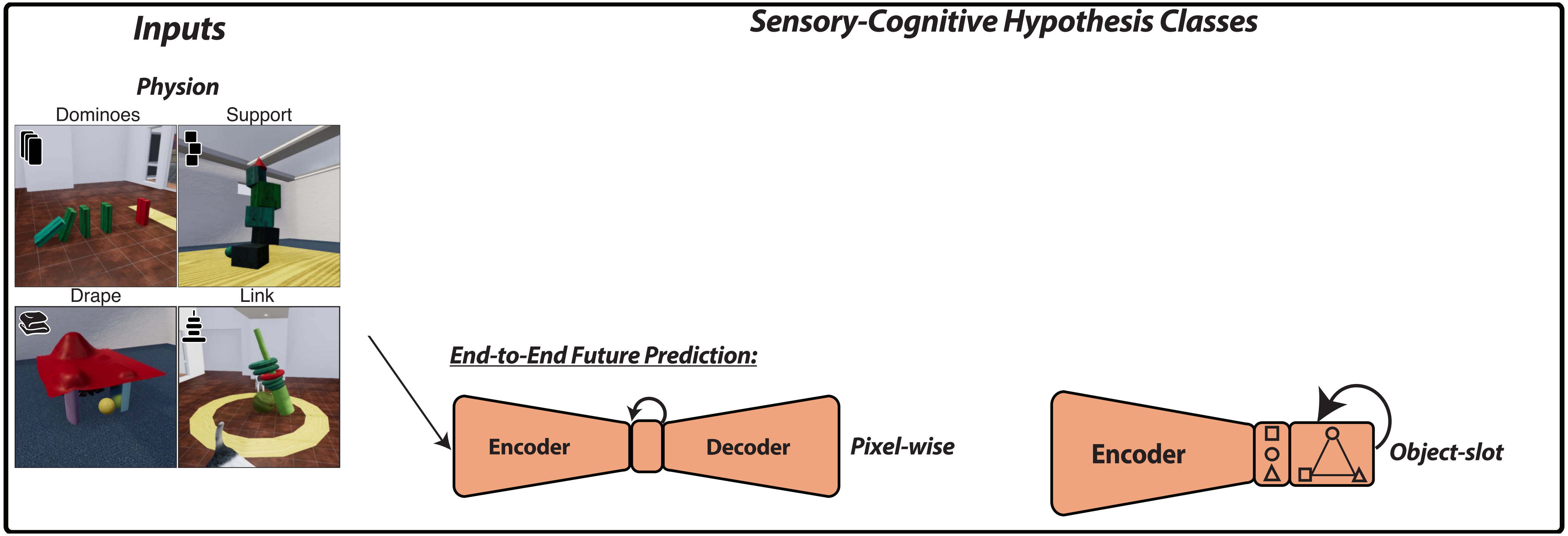
Pixel-wise Future Prediction Poorly Predicts Neurons



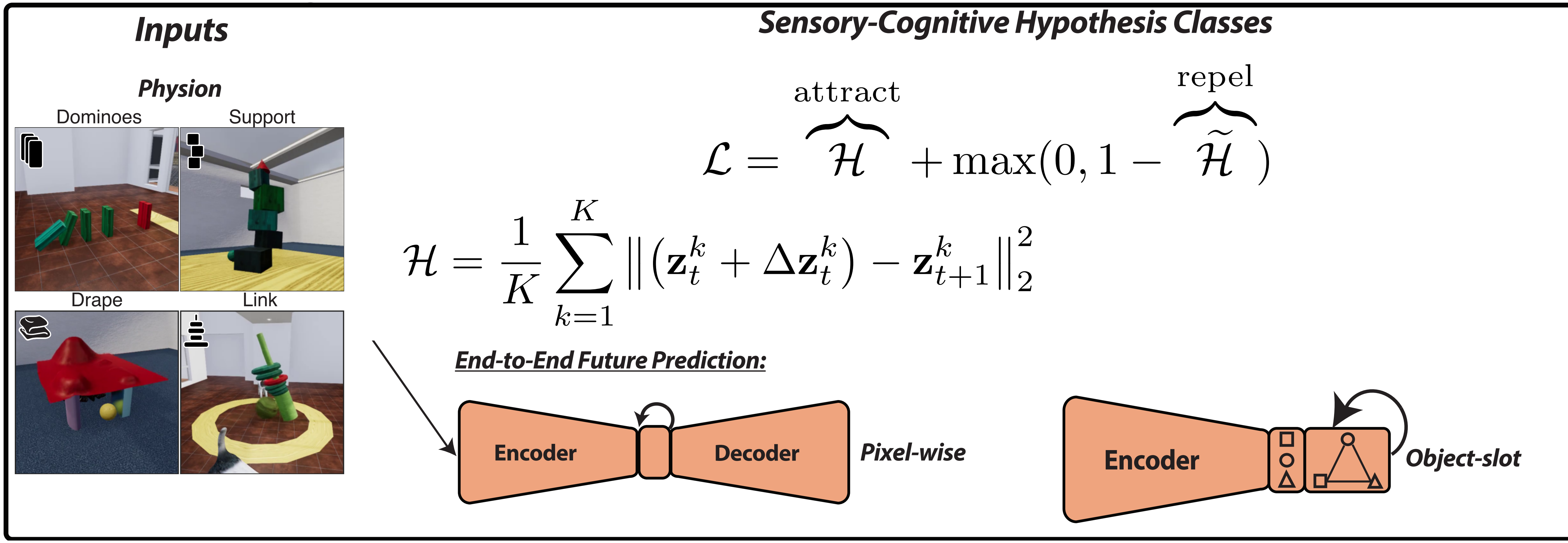
Perhaps DMFC predicts a "factorized" version of the scene?
How?



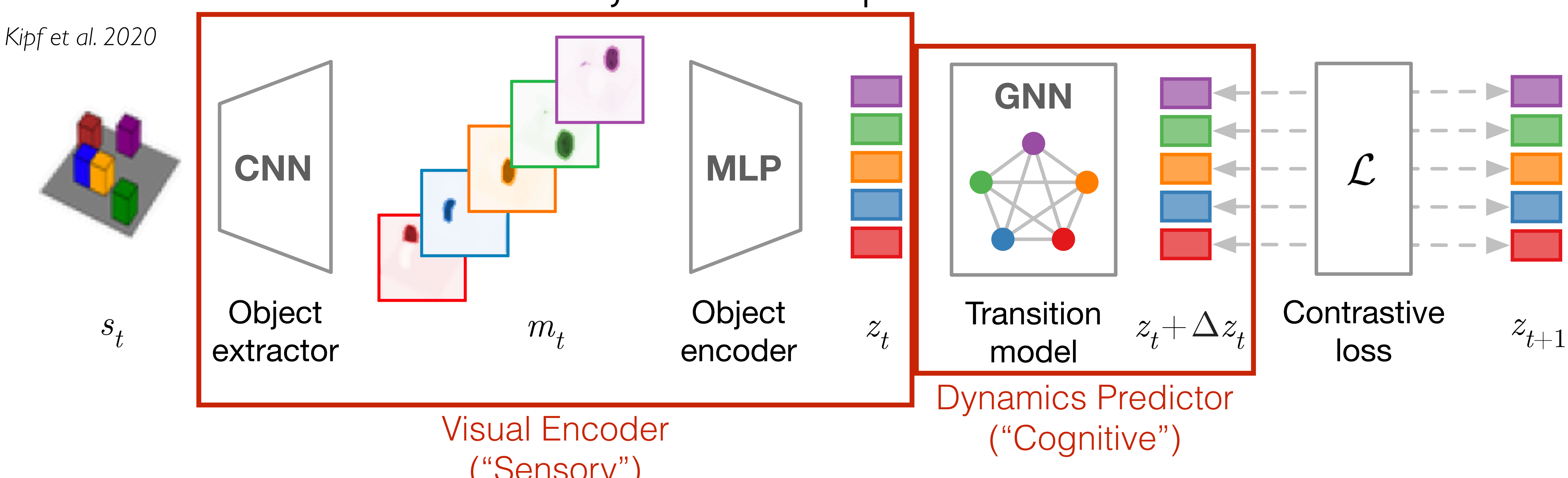
Hypothesis Class 2: Object Slots



Hypothesis Class 2: Object Slots



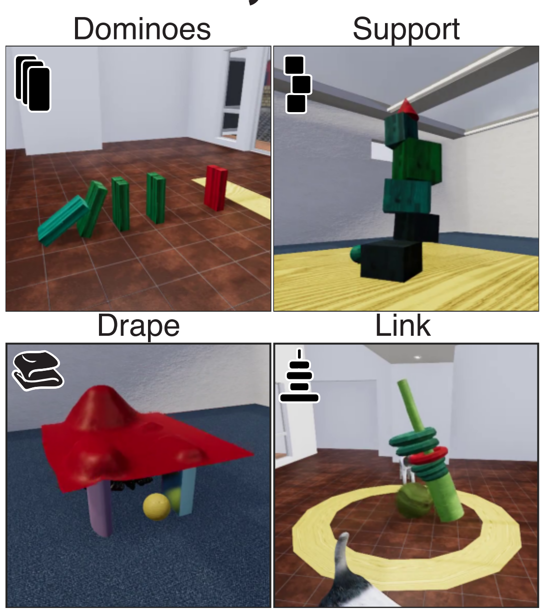
Predicts at the level of object slot representations and their relations



Hypothesis Class 2: Object Slots

Inputs

Physion



Sensory-Cognitive Hypothesis Classes

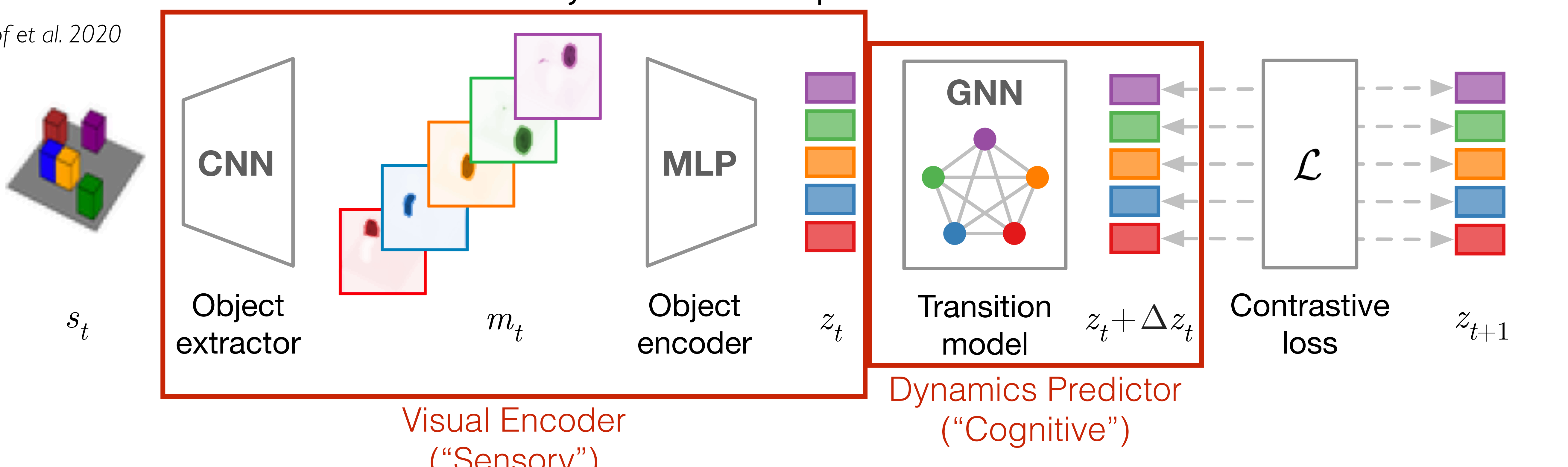
$$\mathcal{L} = \overbrace{\mathcal{H}}^{\text{attract}} + \max(0, 1 - \overbrace{\widetilde{\mathcal{H}}}^{\text{repel}})$$

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

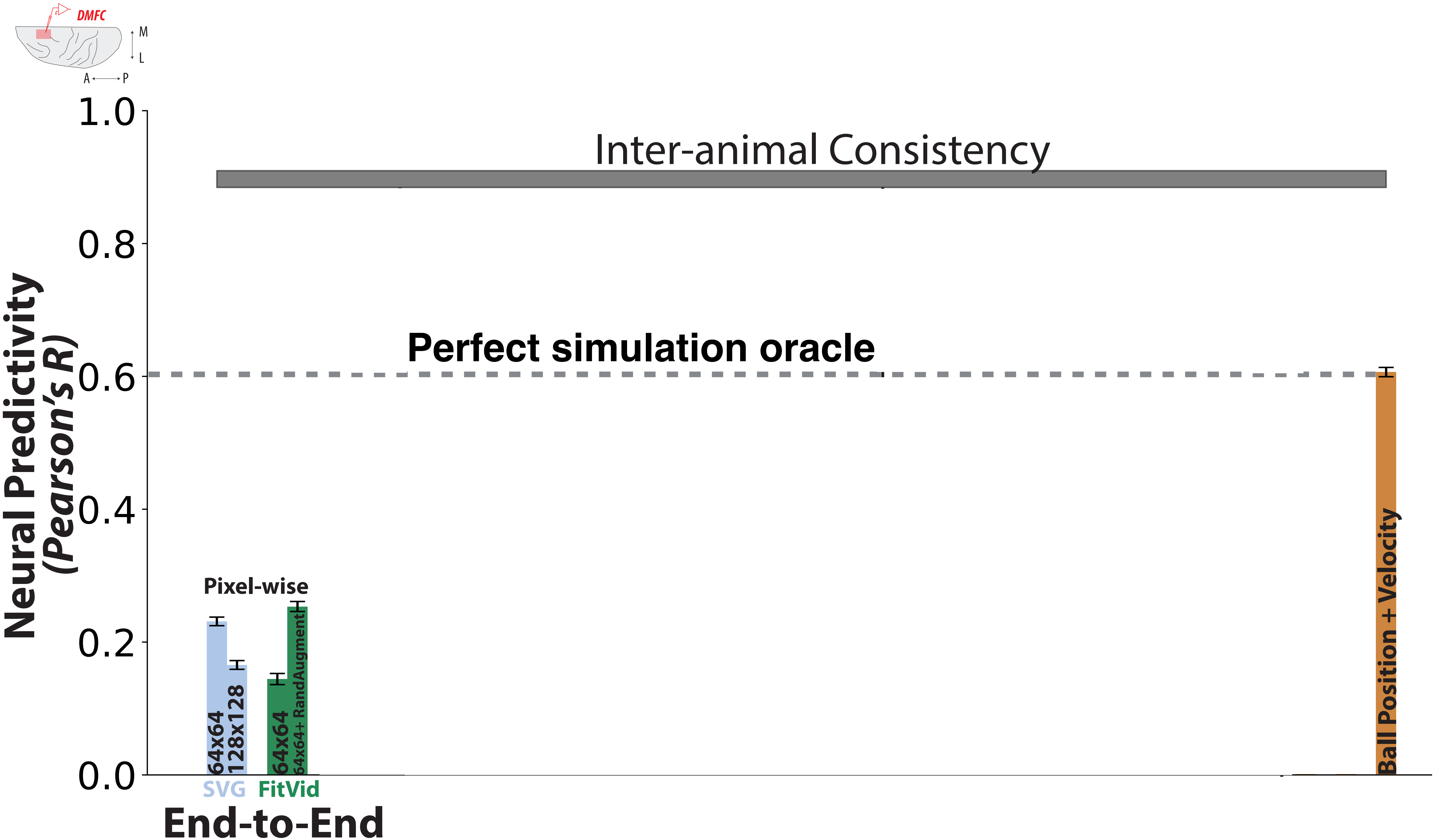
End-to-End Future Prediction:

Predicts at the level of object slot representations and their relations

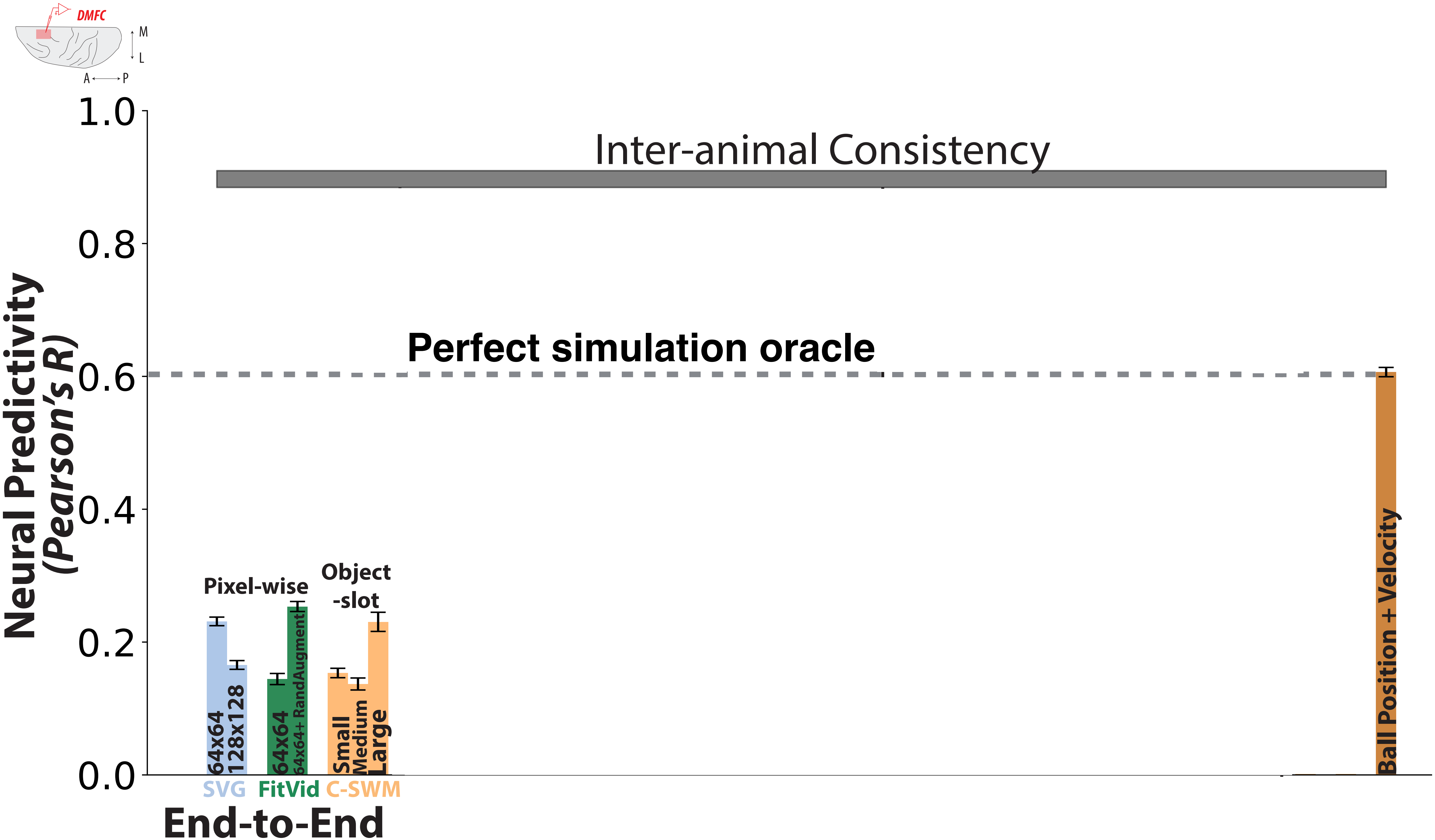
Kipf et al. 2020



Pixel-wise Future Prediction Poorly Predicts Neurons

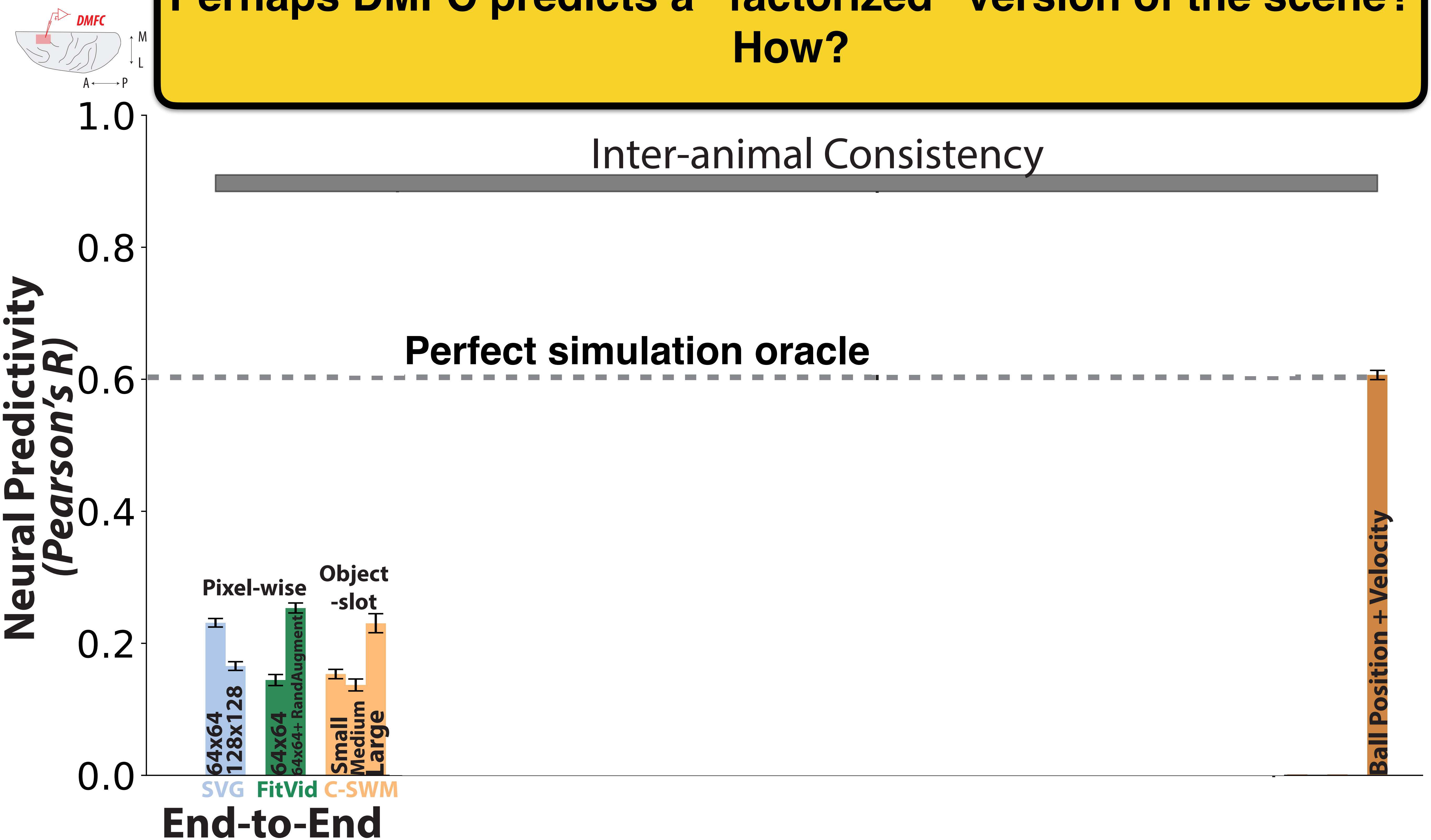


Object Slot Future Prediction Poorly Predicts Neurons



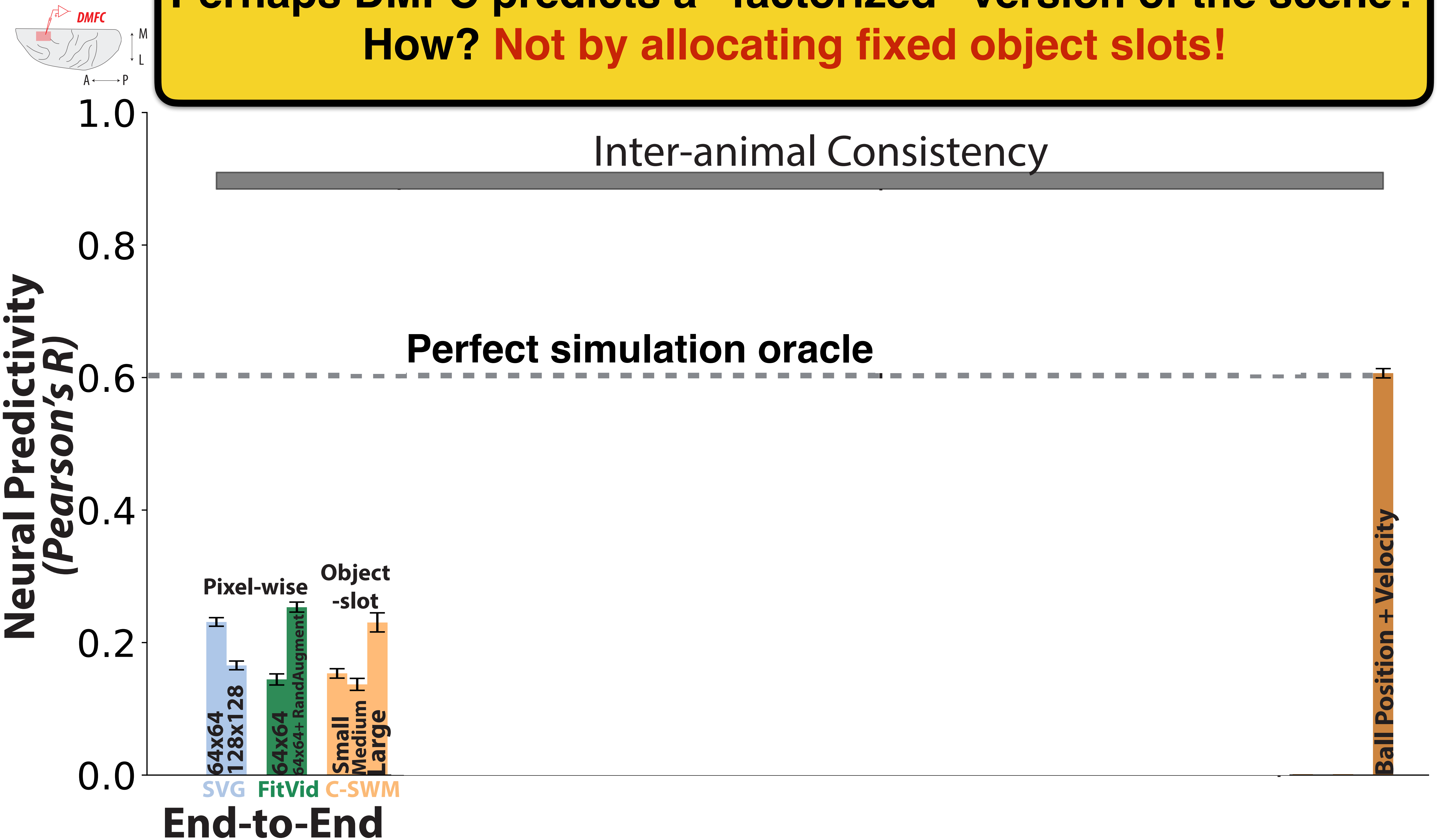
Object Slot Future Prediction Poorly Predicts Neurons

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

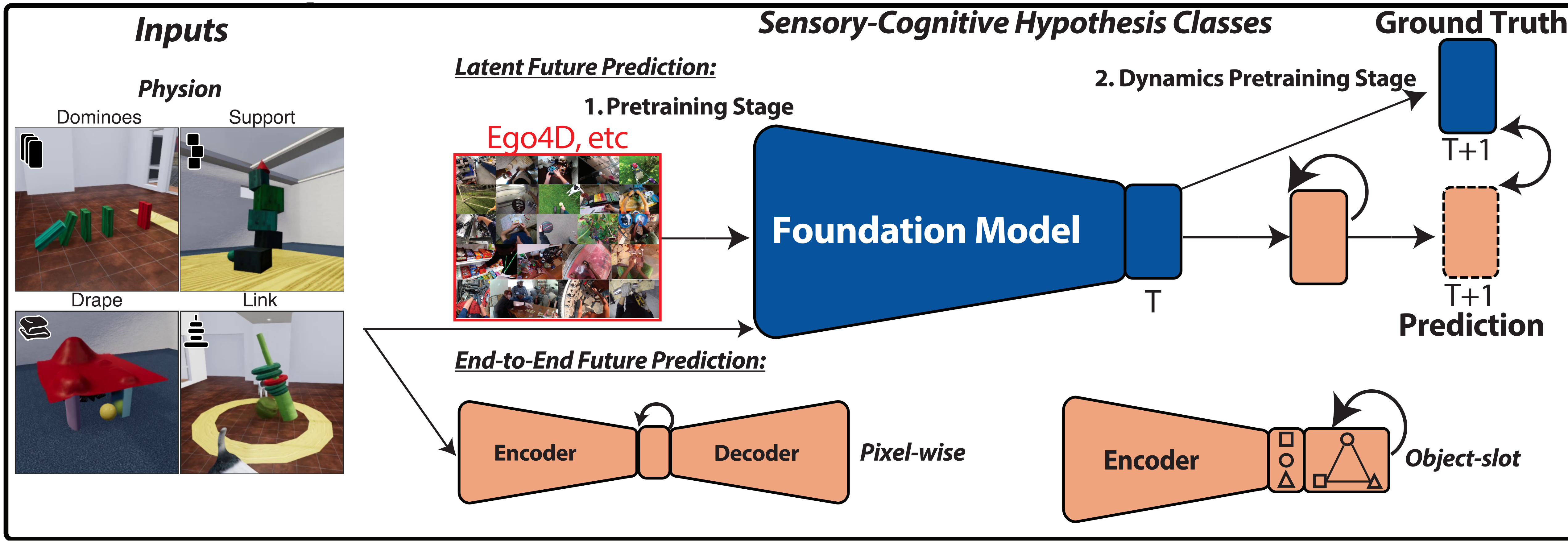


Object Slot Future Prediction Poorly Predicts Neurons

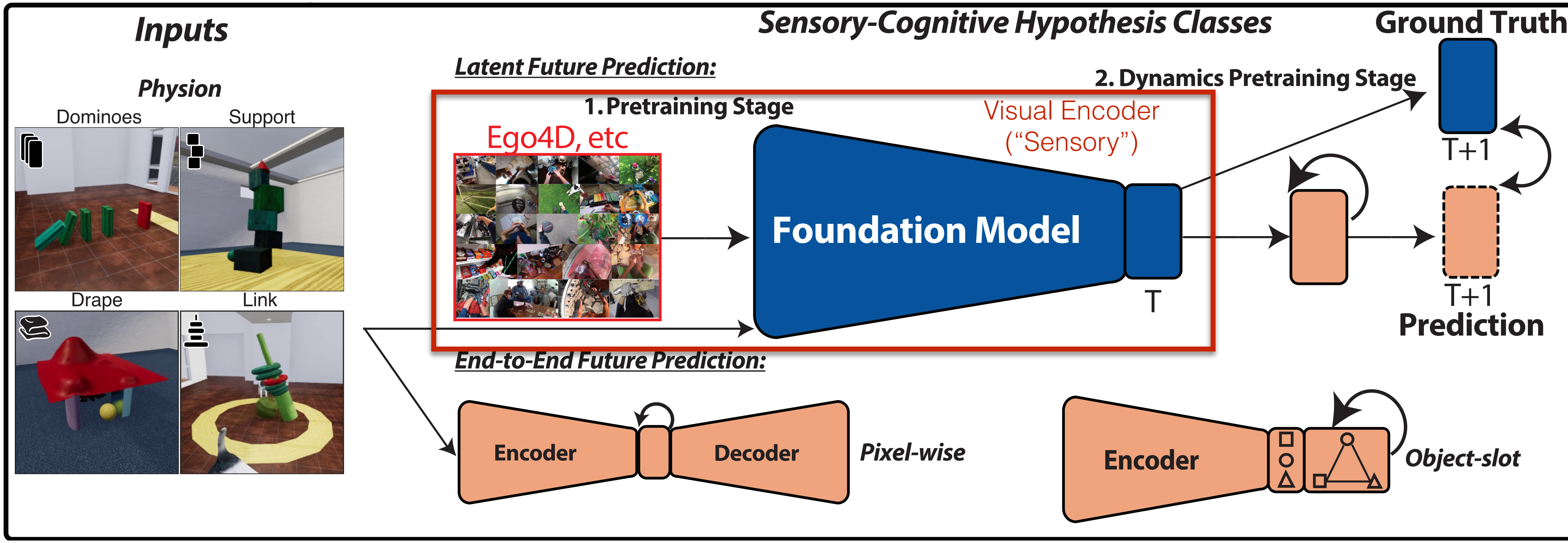
Perhaps DMFC predicts a "factorized" version of the scene?
How? Not by allocating fixed object slots!



Hypothesis Class 3: Latent Future Prediction

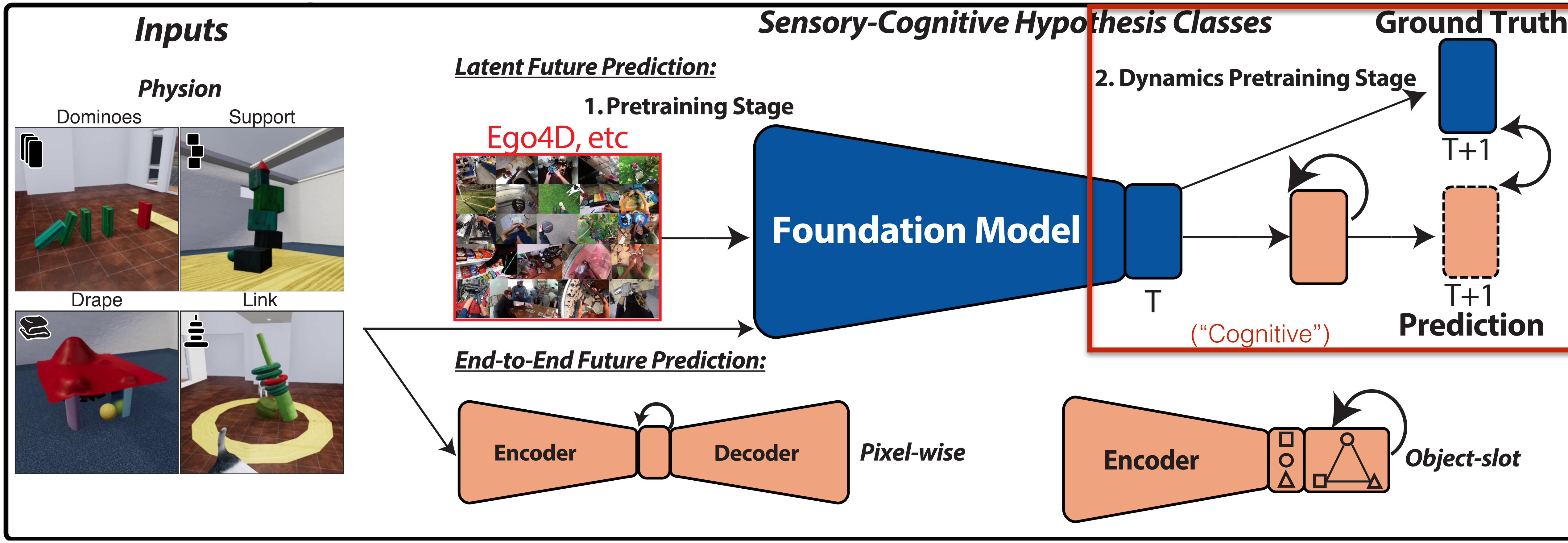


Hypothesis Class 3: Latent Future Prediction



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

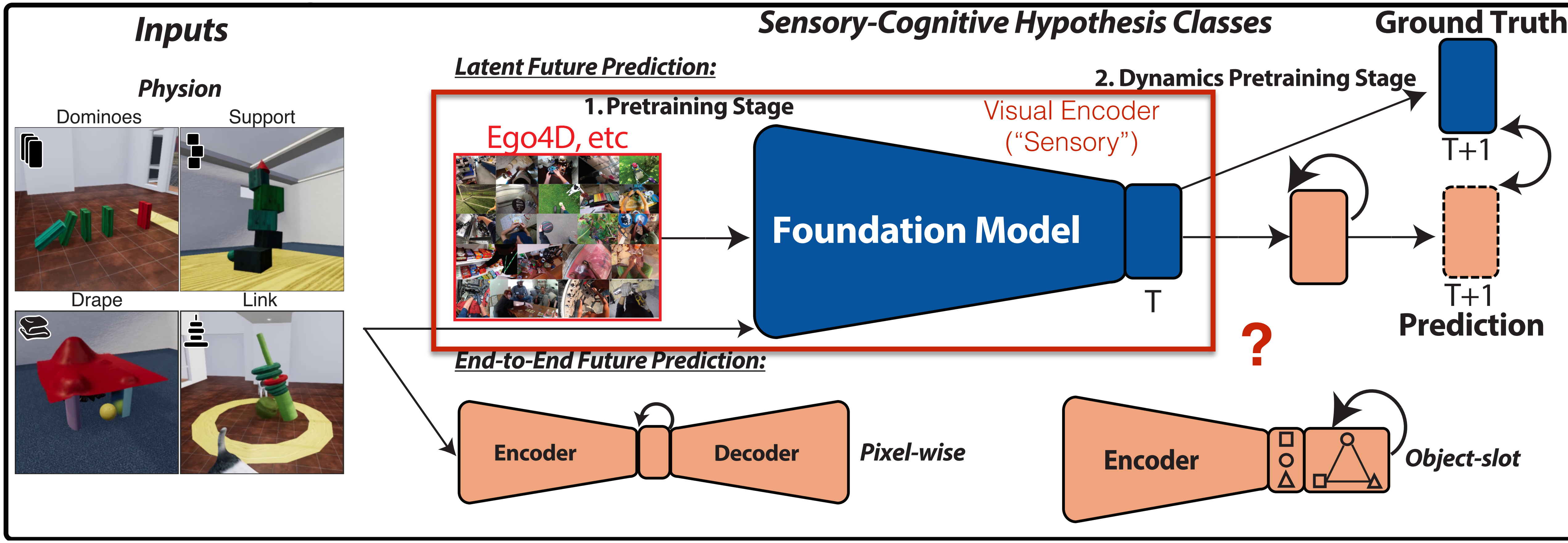
Hypothesis Class 3: Latent Future Prediction



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

Leverage these dynamics to do explicit future prediction

Hypothesis Class 3: Foundation Models

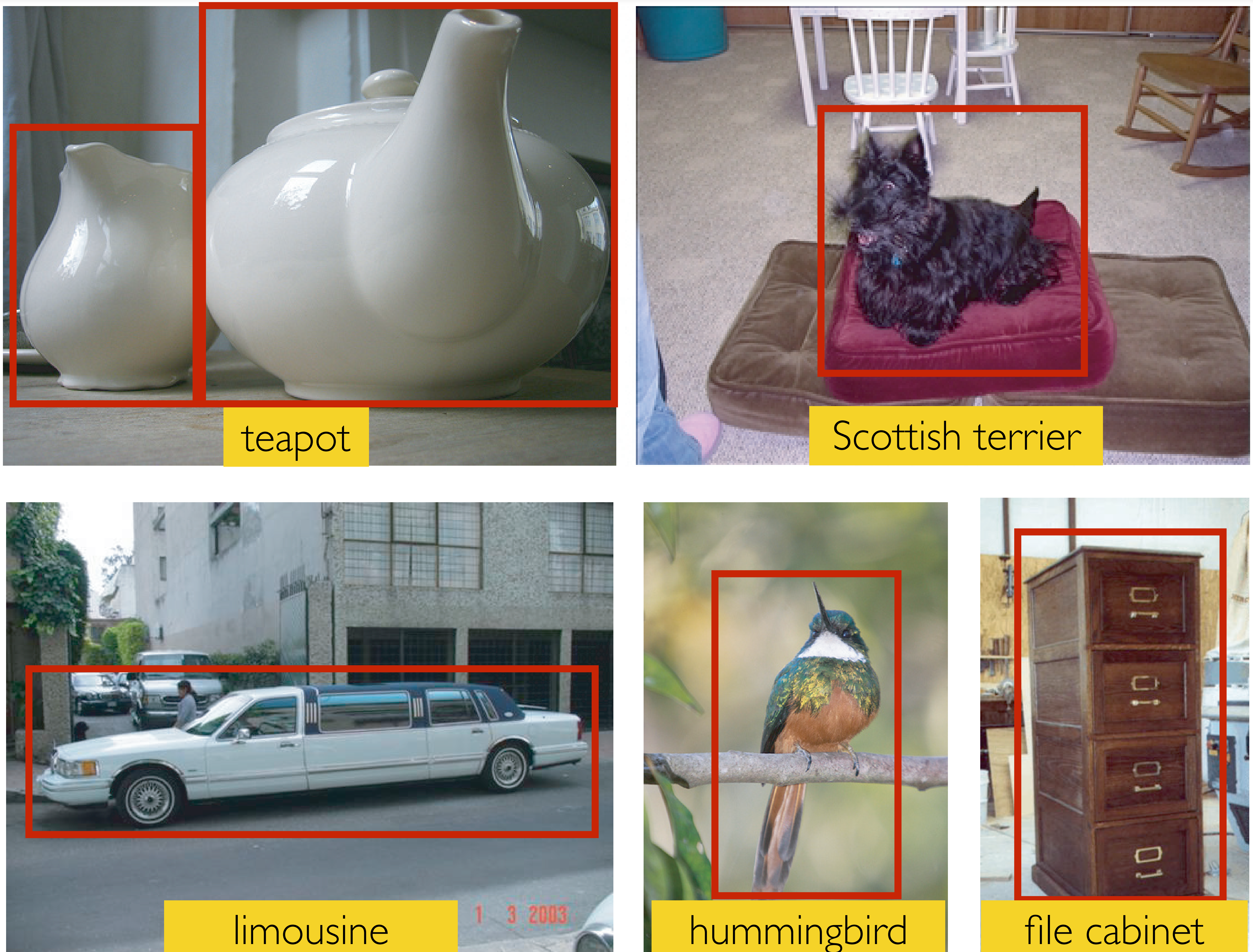


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

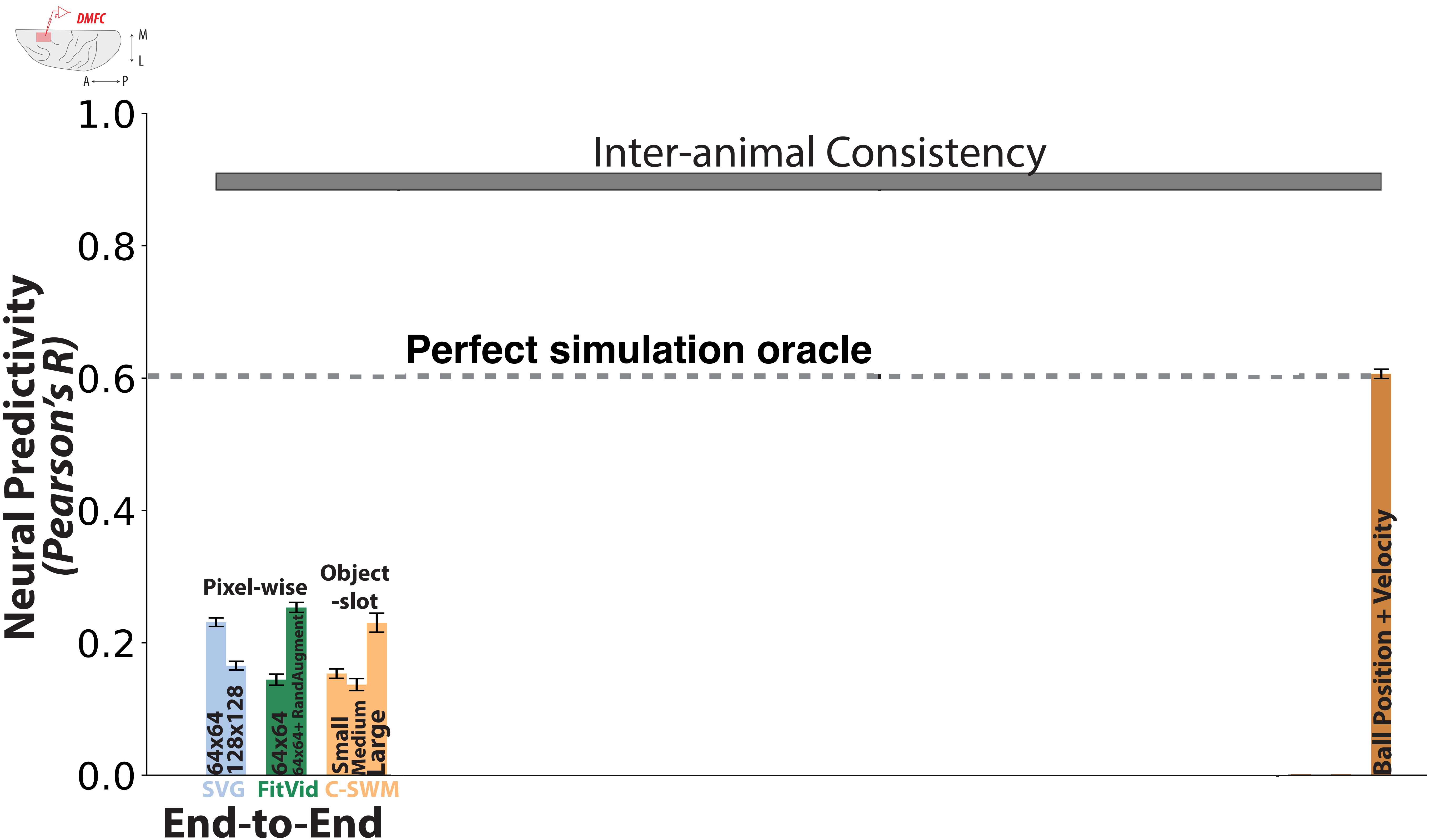
What vision task?

Leverage these dynamics to do explicit future prediction

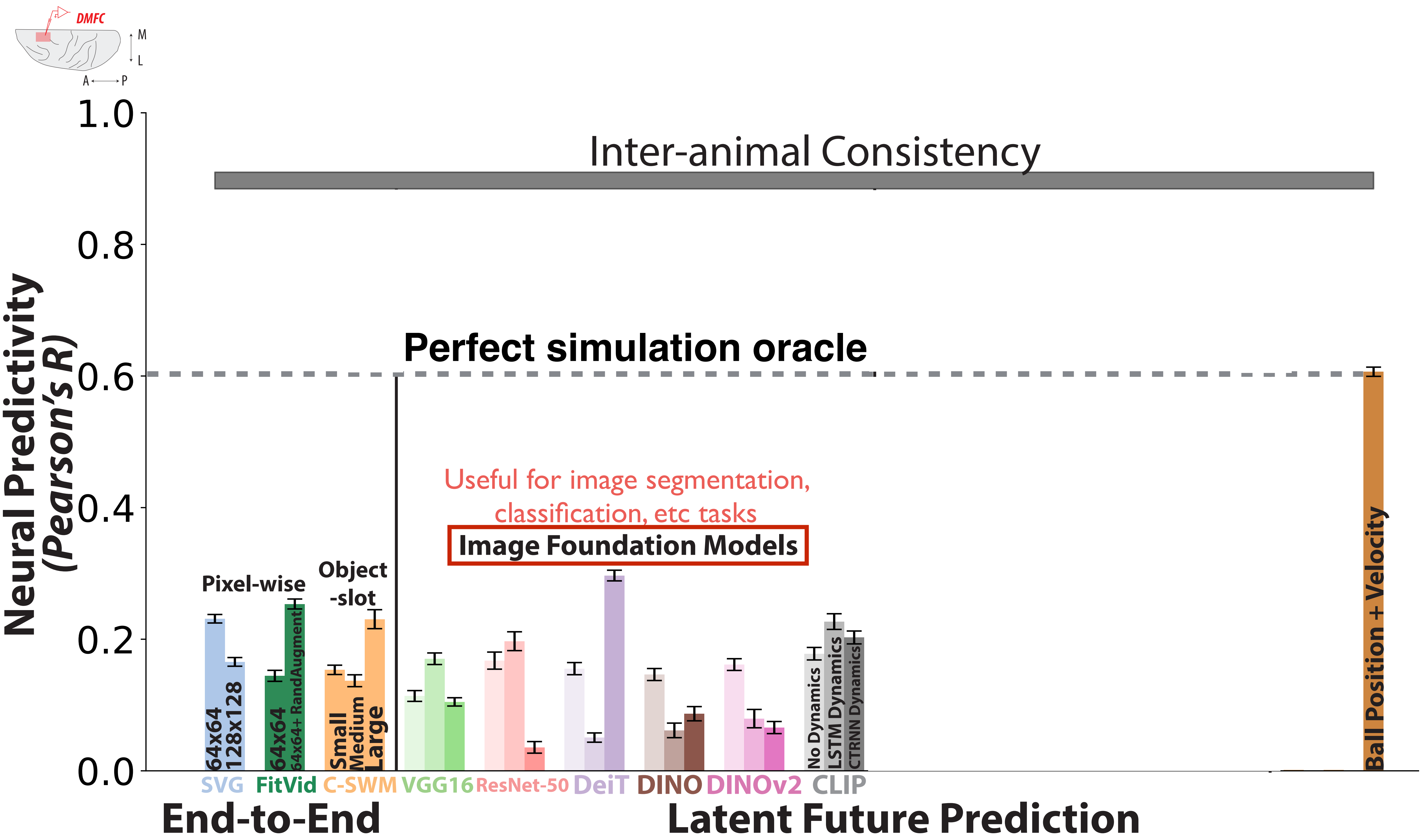
Hypothesis Class 3: Static Image Foundation Models



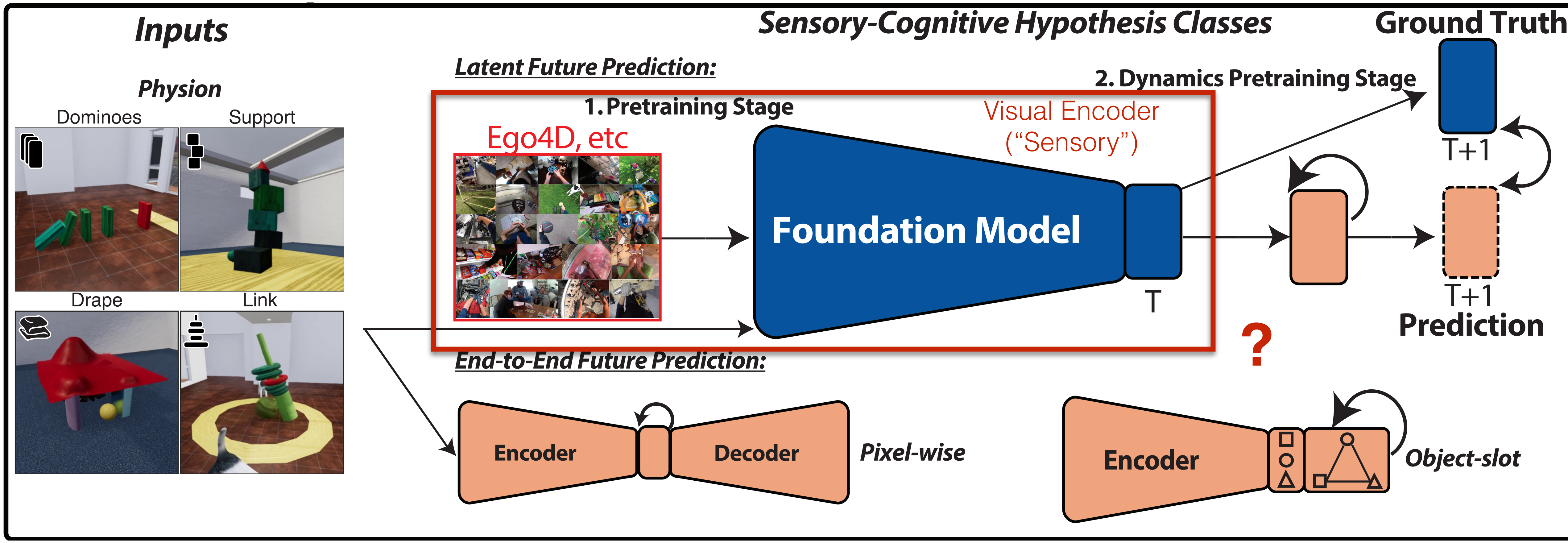
Object Slot Future Prediction Poorly Predicts Neurons



Static Image Foundation Future Prediction Poorly Predicts Neurons



Hypothesis Class 3: Foundation Models



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

What vision task?

We do far more than engage with static images!

Leverage these dynamics to do explicit future prediction

Hypothesis Class 3: Video Foundation Models

Ego4D: everyday activity around the world

$$\mathcal{L}_{contrastive} =$$

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

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

Ego4D: A massive-scale egocentric dataset

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

931 participants from 74 worldwide locations

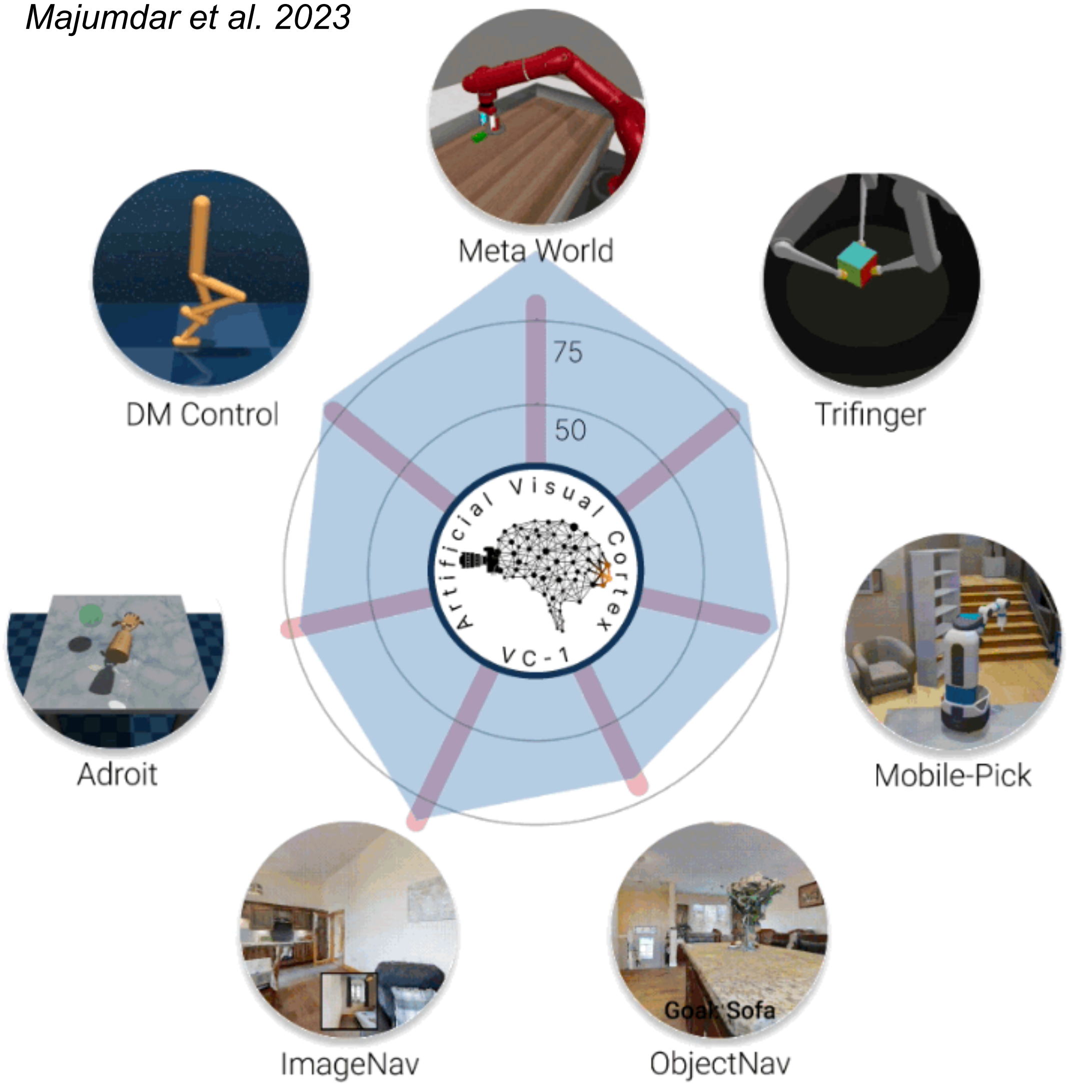
Multimodal: audio, 3D scans, IMU, stereo, multi-camera

Grauman et al. 2022

Hypothesis Class 3: Video Foundation Models

Ego4D: everyday activity around the world

Majumdar et al. 2023



Ego4D: A massive-scale egocentric dataset

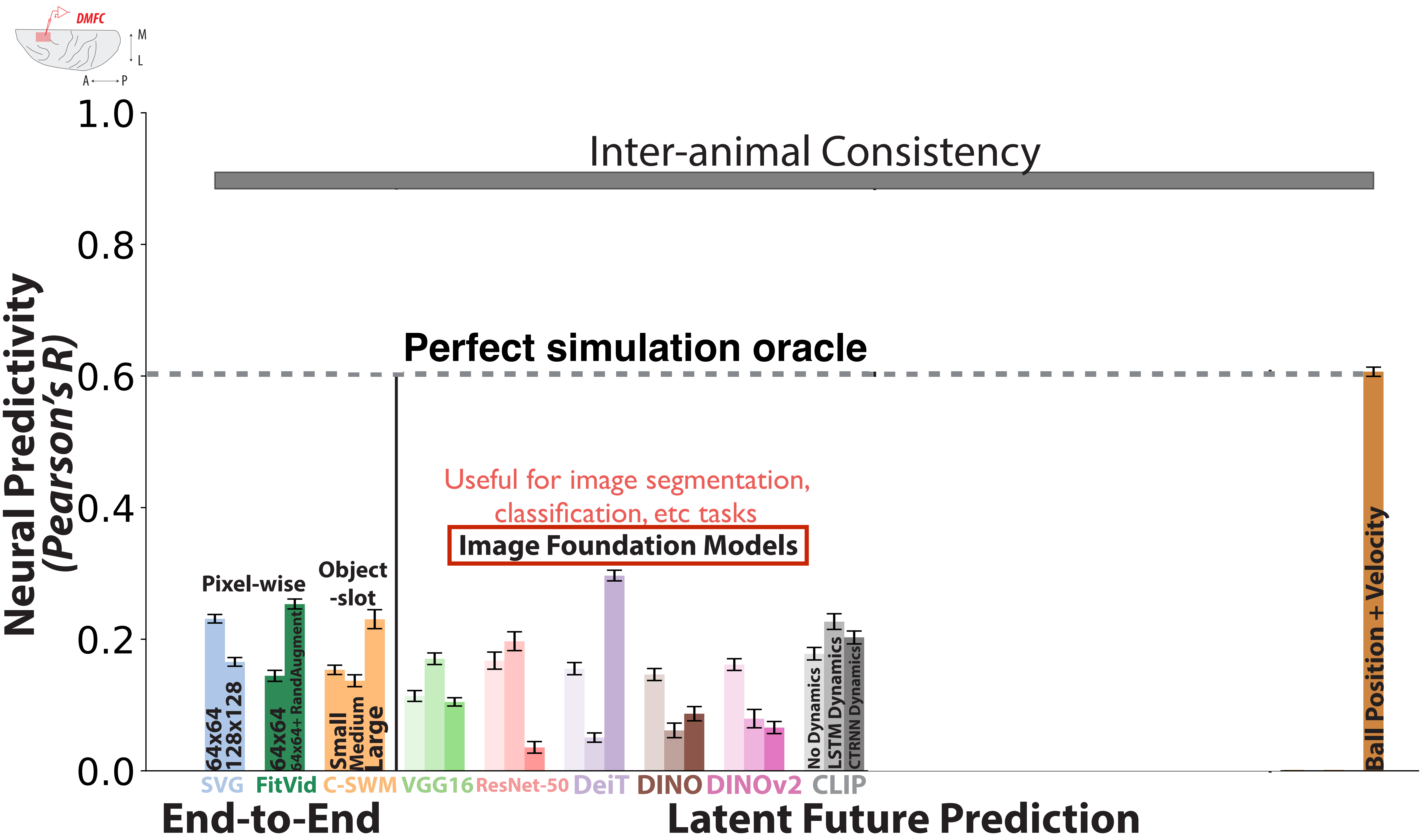
3,670 hours of in-the-wild daily life activity

931 participants from 74 worldwide locations

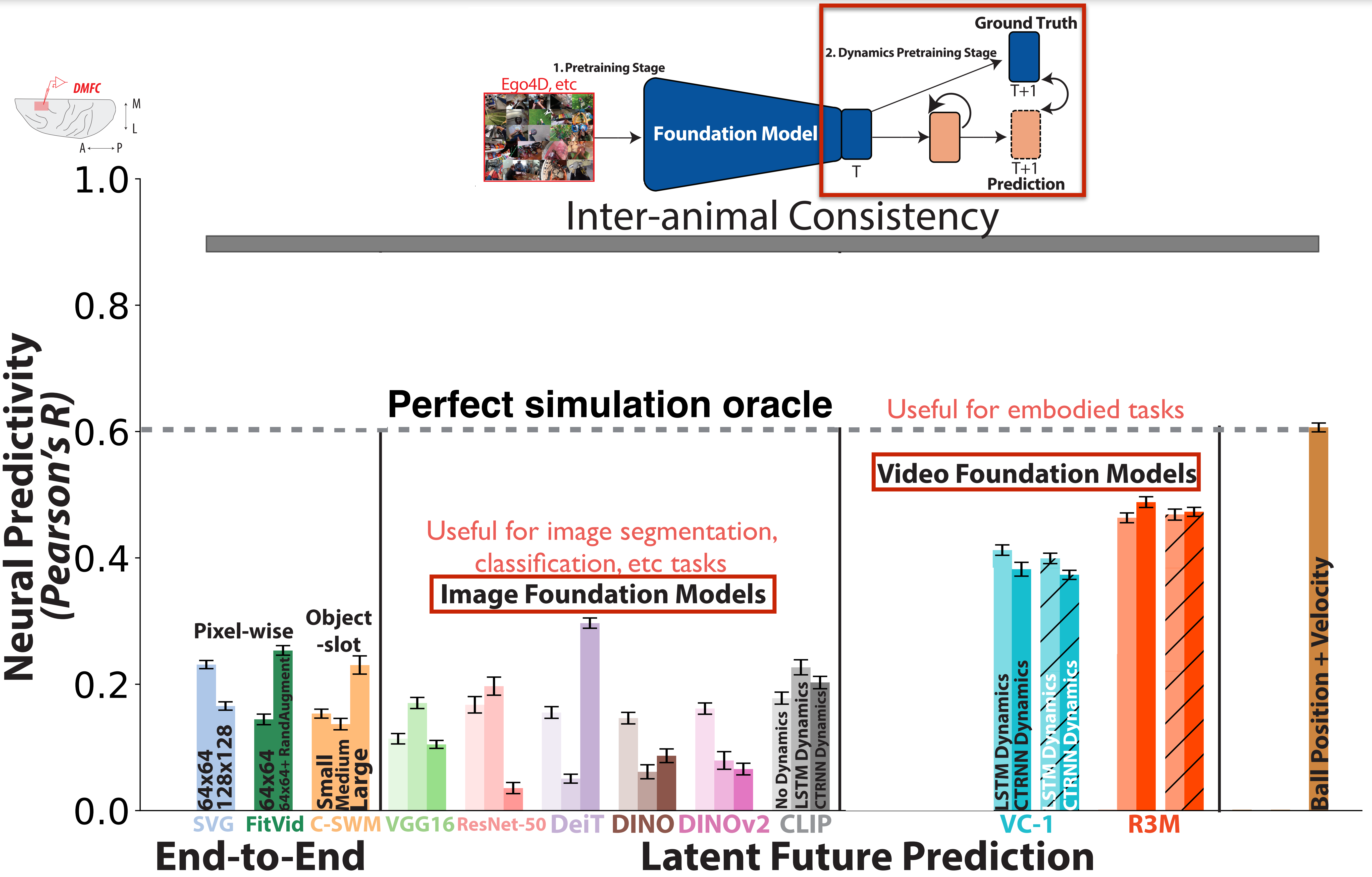
Multimodal: audio, 3D scans, IMU, stereo, multi-camera

Grauman et al. 2022

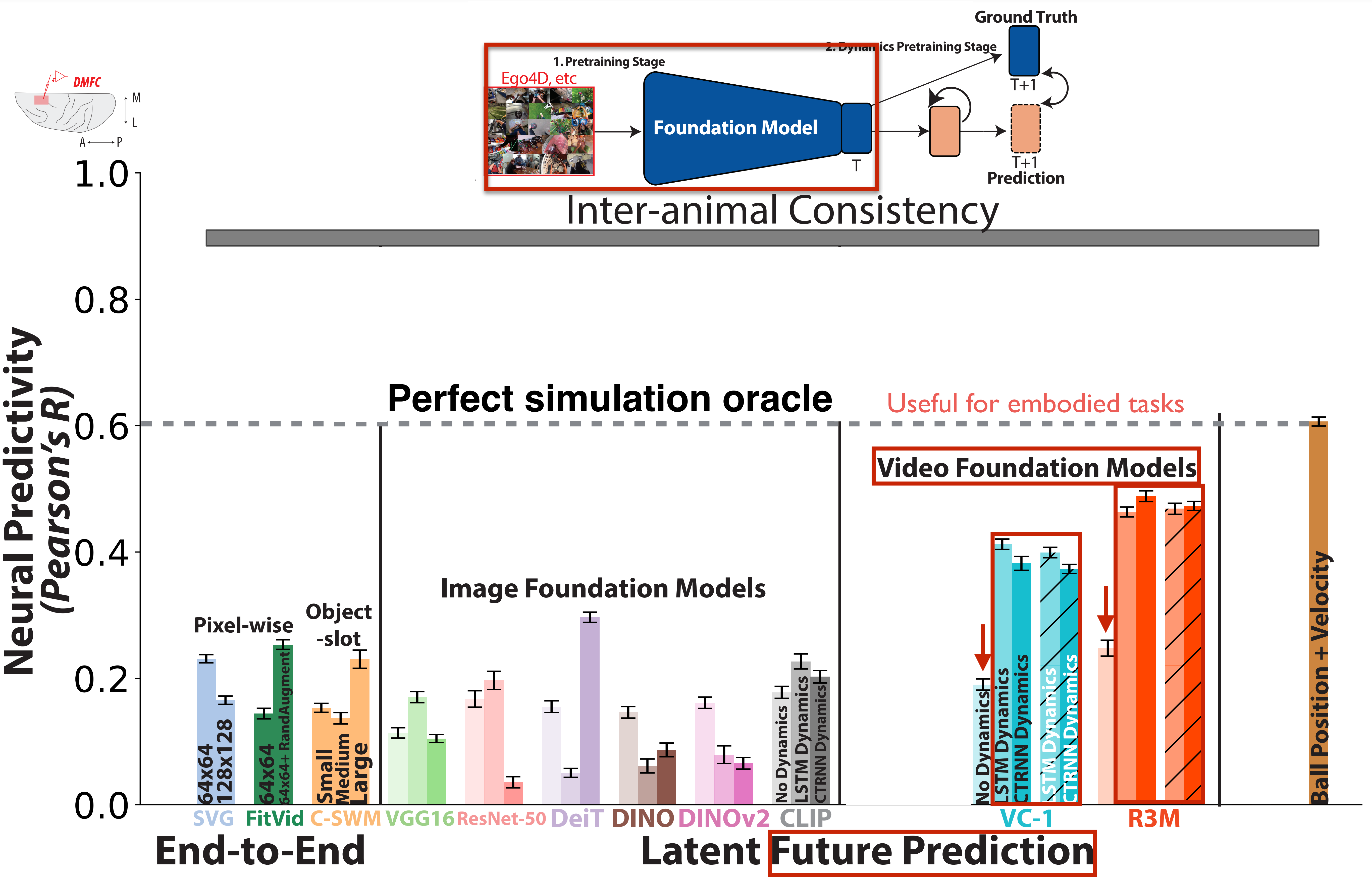
Static Image Foundation Future Prediction Poorly Predicts Neurons



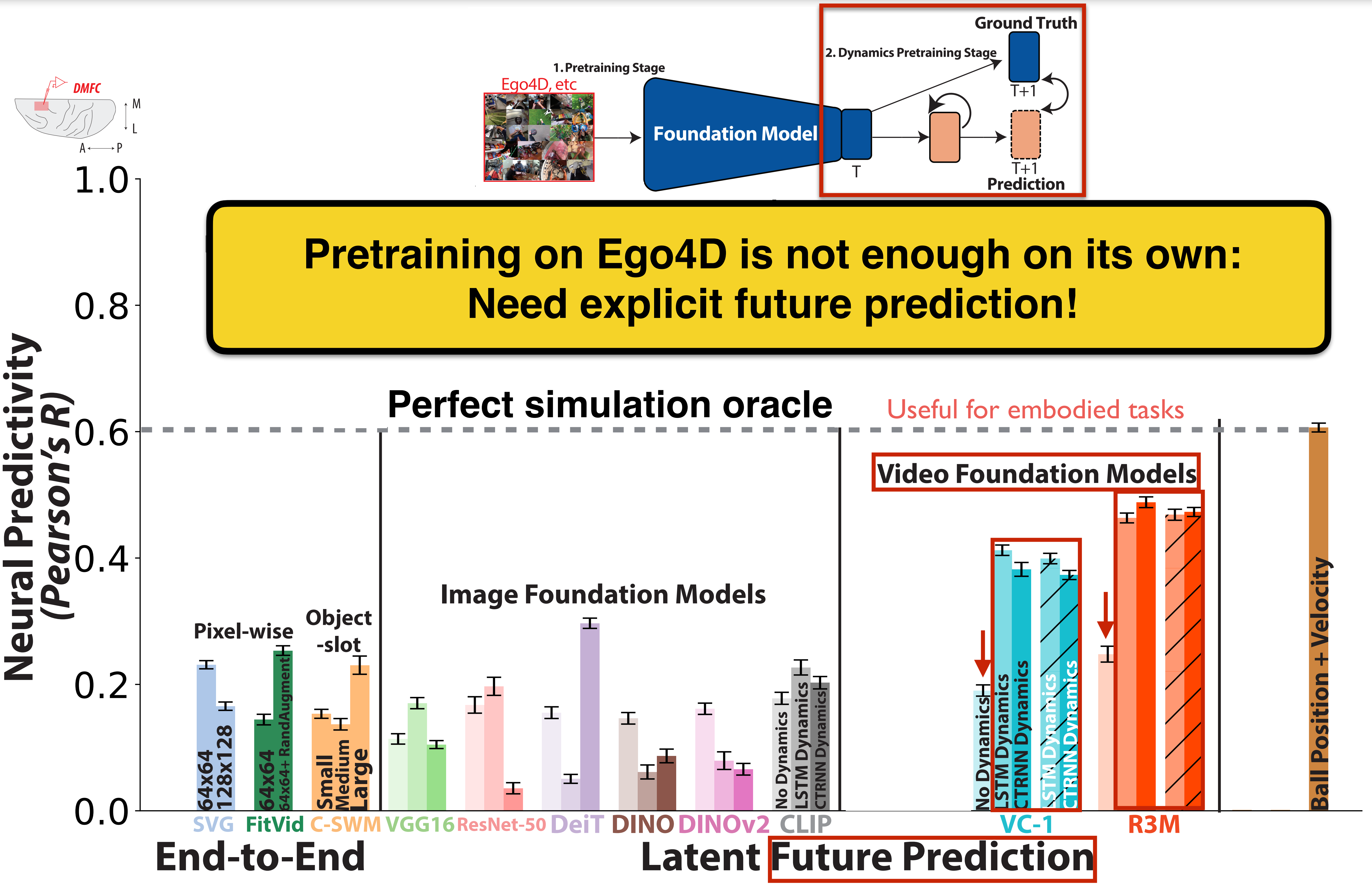
Video Foundation Future Prediction Best Predict Neurons



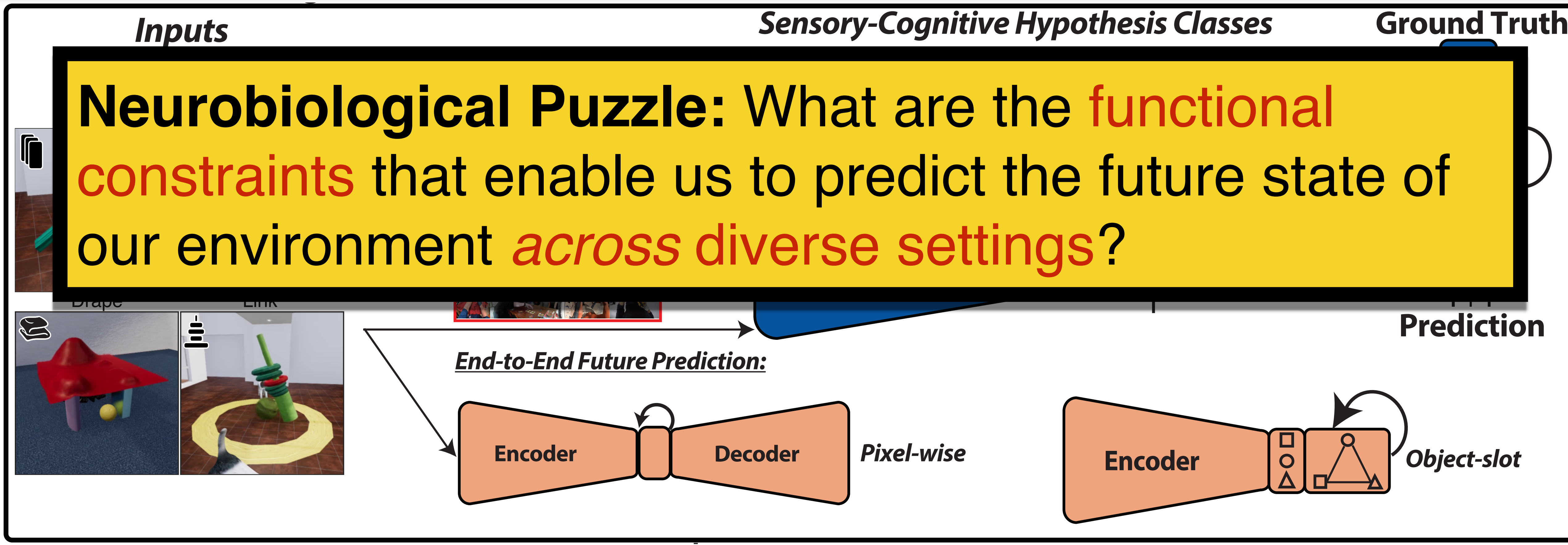
Video Foundation Future Prediction Best Predict Neurons



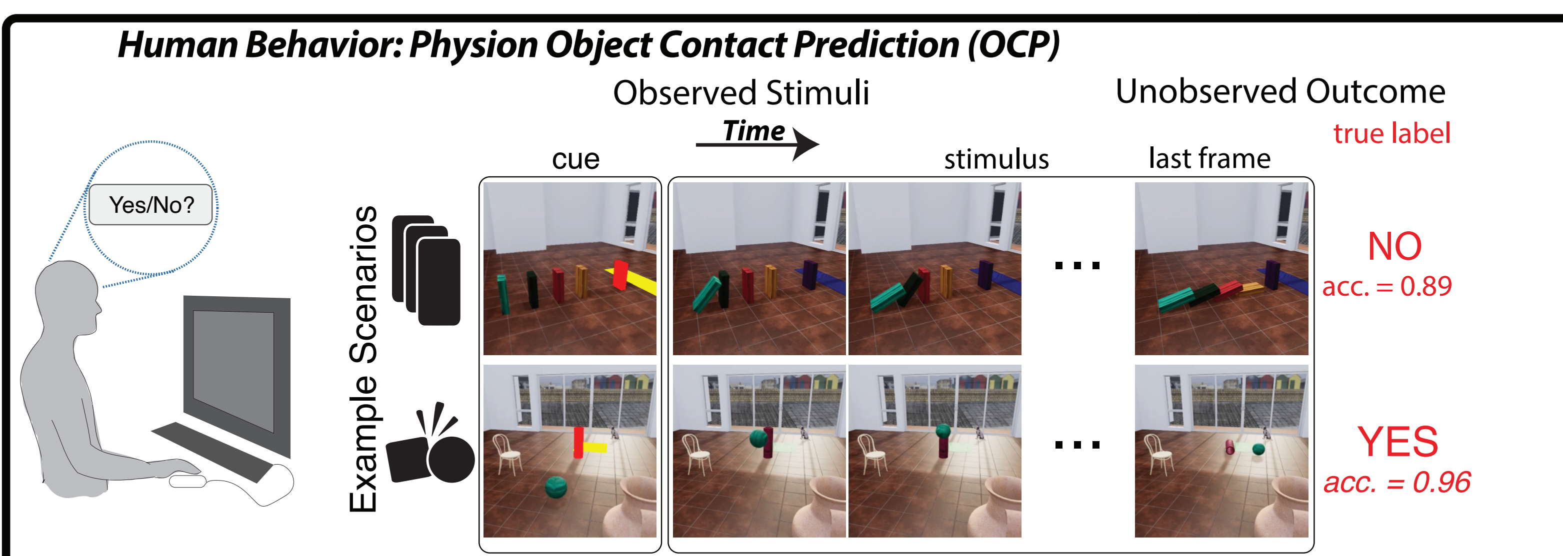
Video Foundation Future Prediction Best Predict Neurons

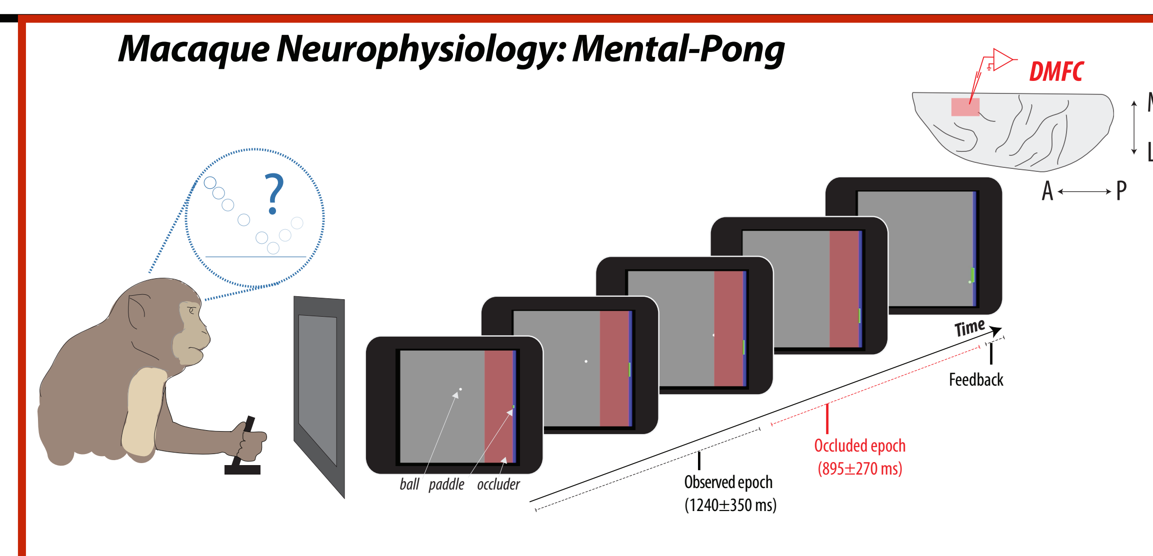


Macaque Neurophysiology: Mental Pong

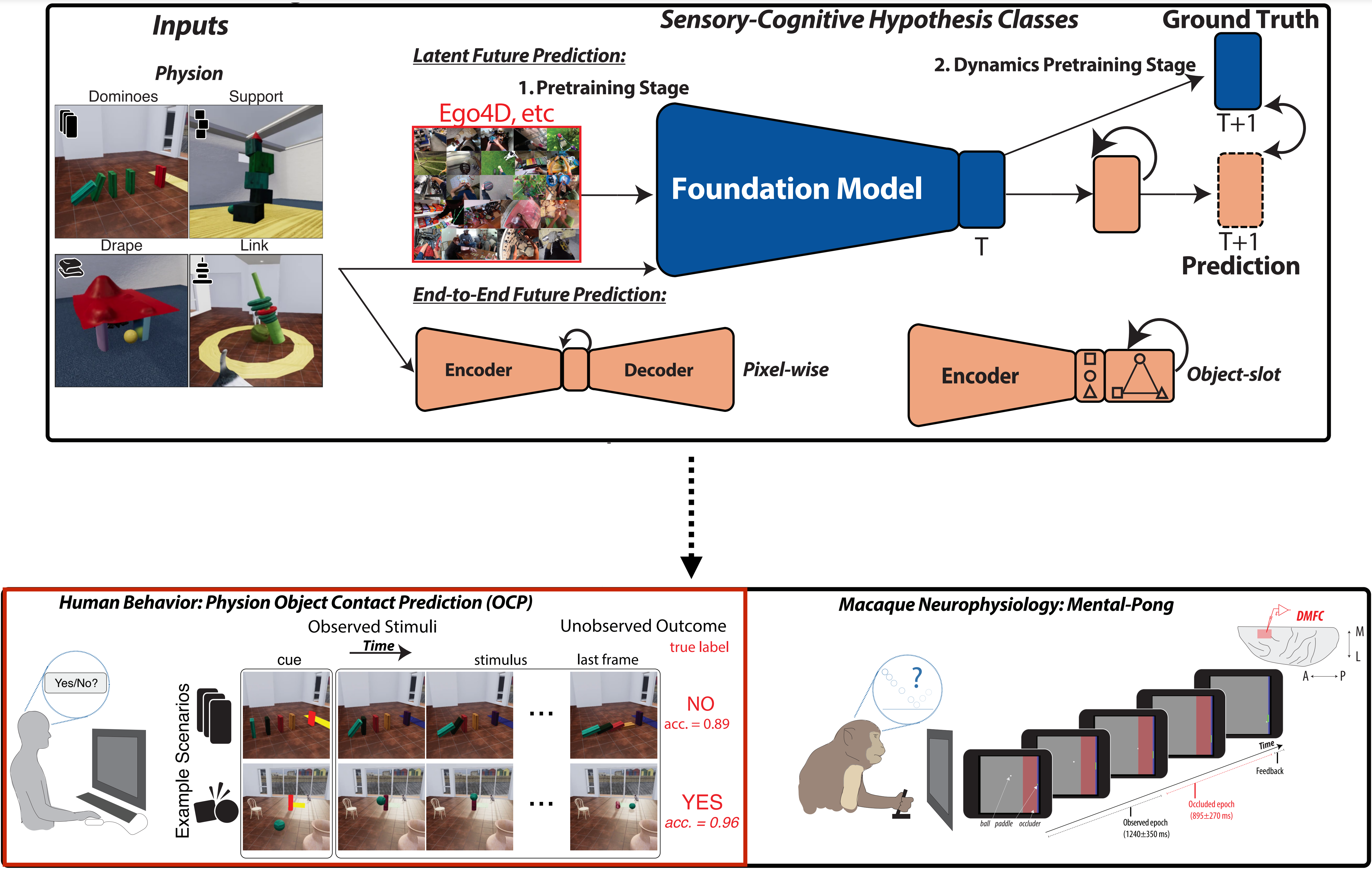


Neurobiological Puzzle: What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?





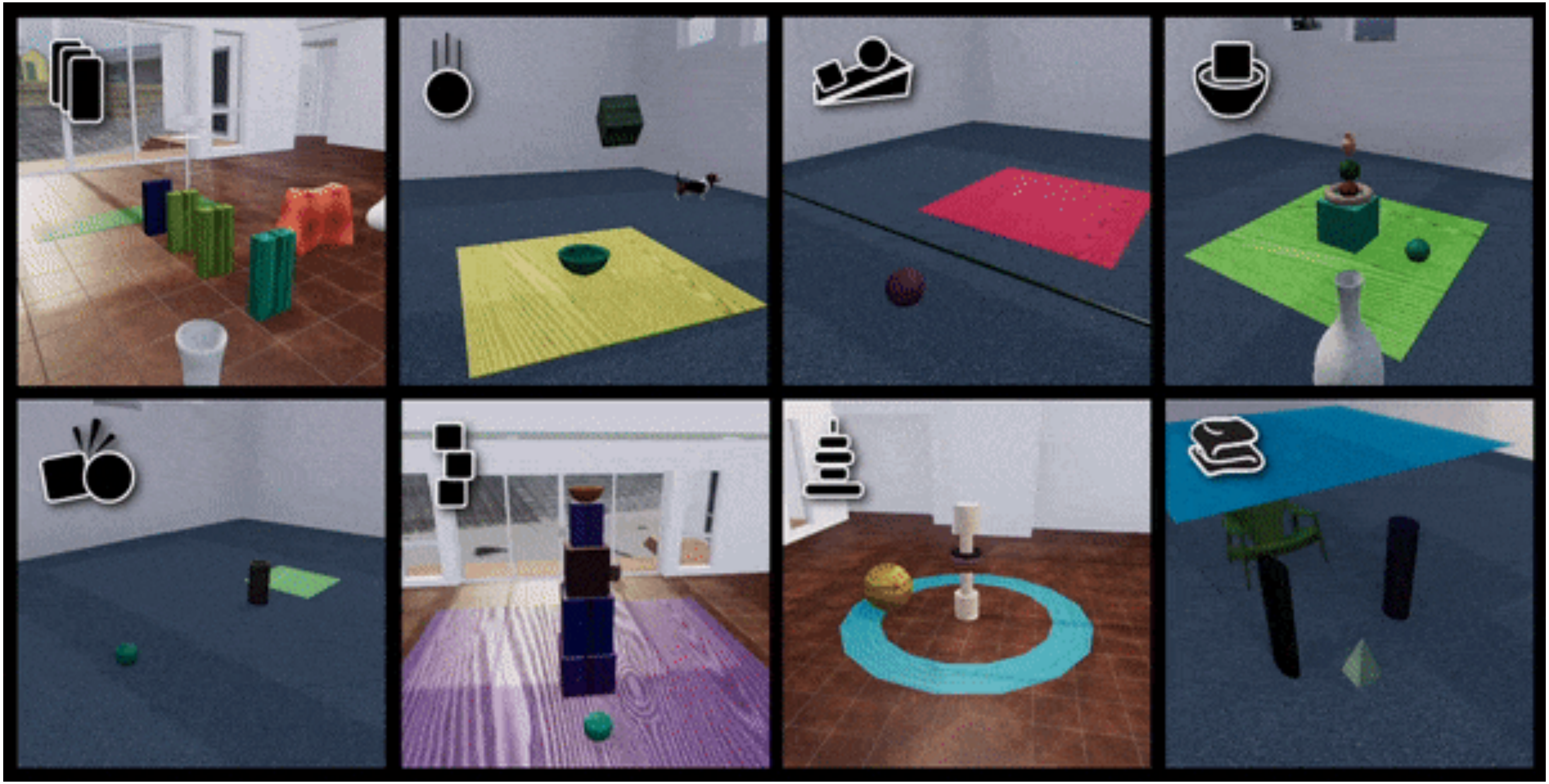
Human Behavior: Object Contact Prediction



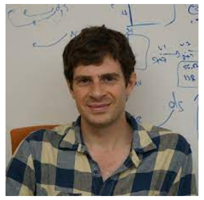
Object Contact Prediction Environment

Physion/ThreeD World (TDW)

Bear et al. 2021



Focus on everyday physical understanding



Daniel Bear

Joshua Tenenbaum

Daniel Yamins

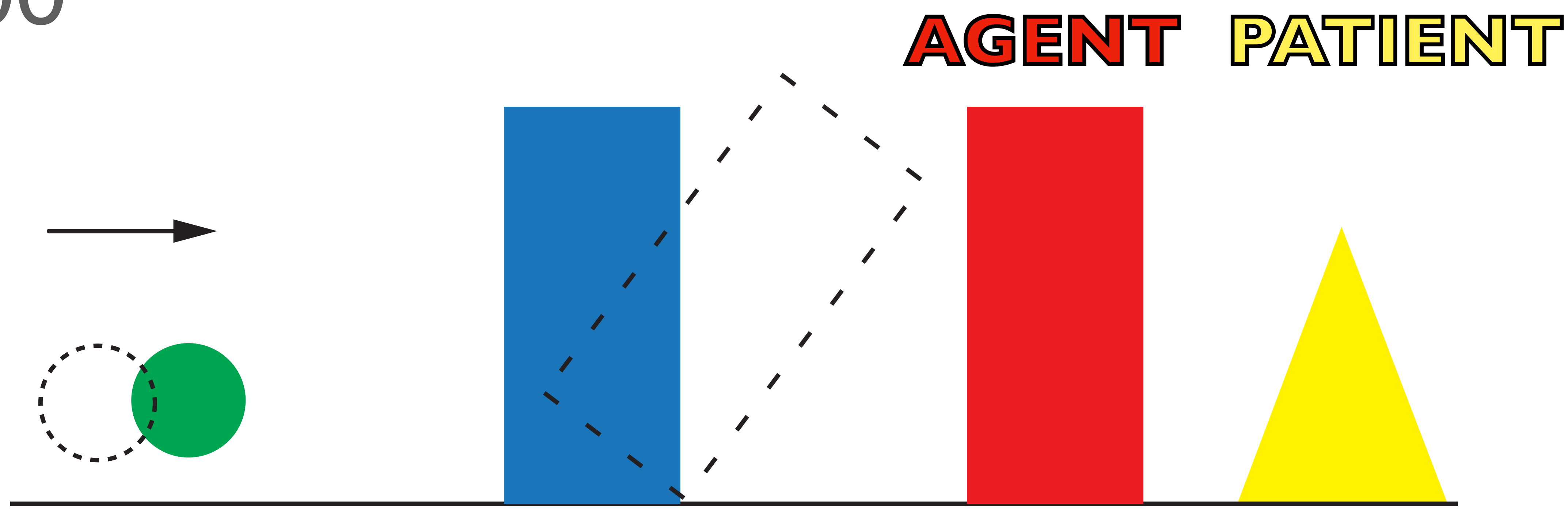
Judith Fan

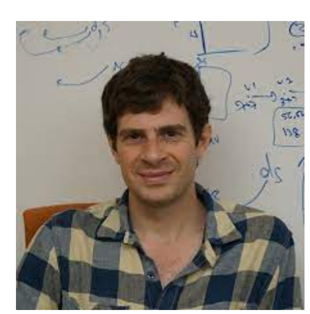
Human Behavior: Object Contact Prediction

Bear et al. 2021

n=100

"Will the agent object contact the patient object?"




Daniel Bear

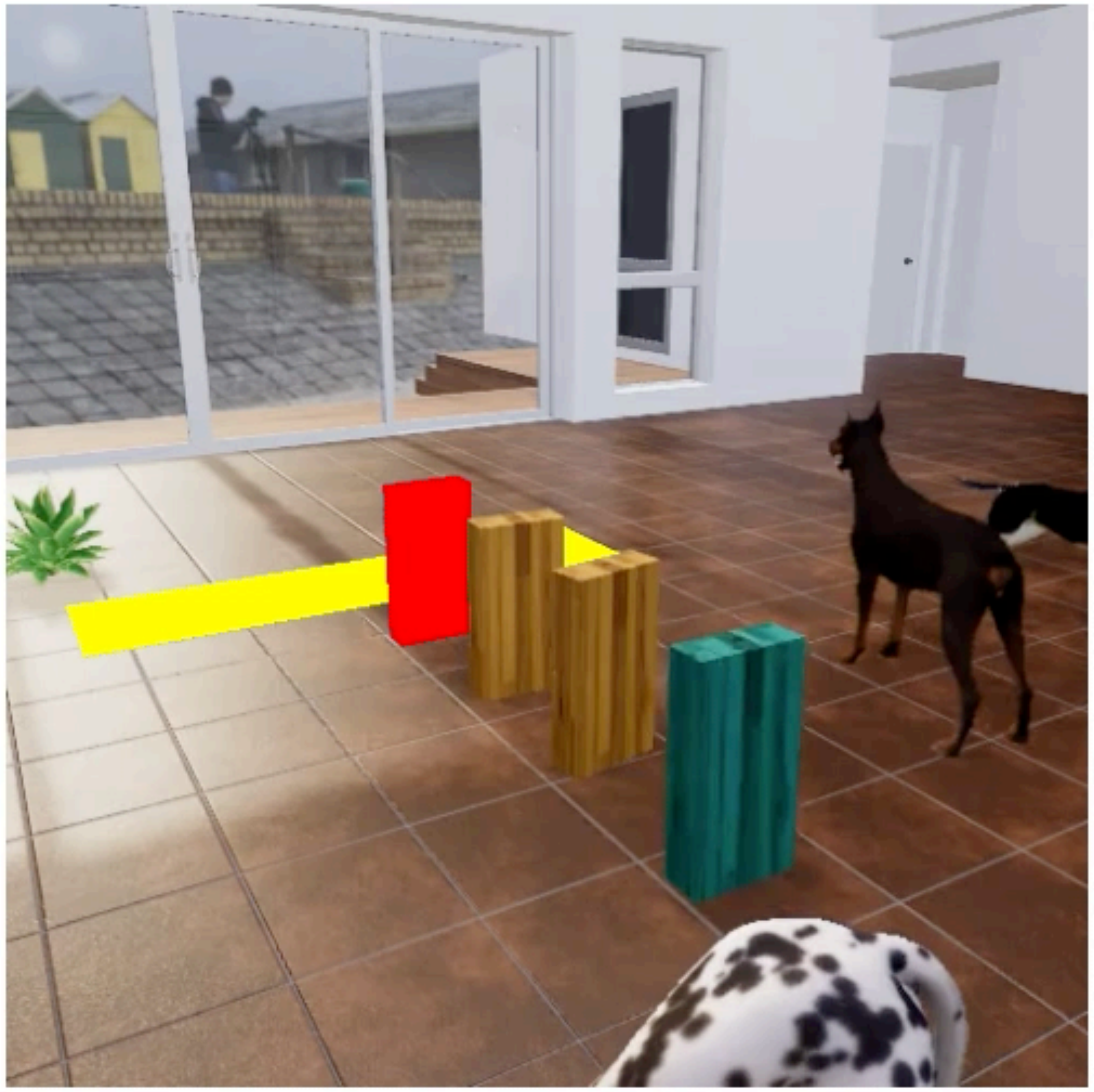

Joshua Tenenbaum


Daniel Yamins


Judith Fan

Completion Progress

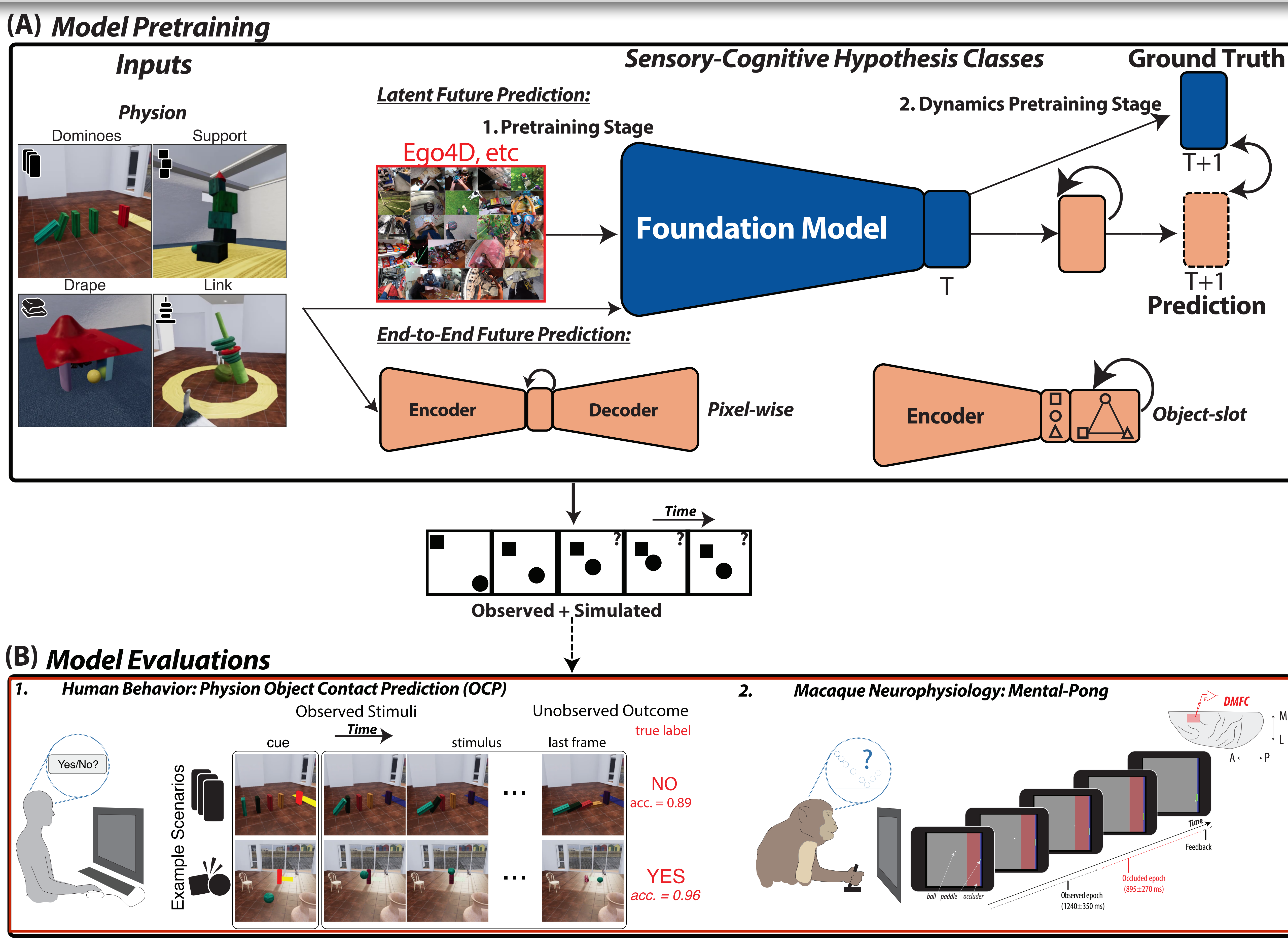
Bear et al. 2021



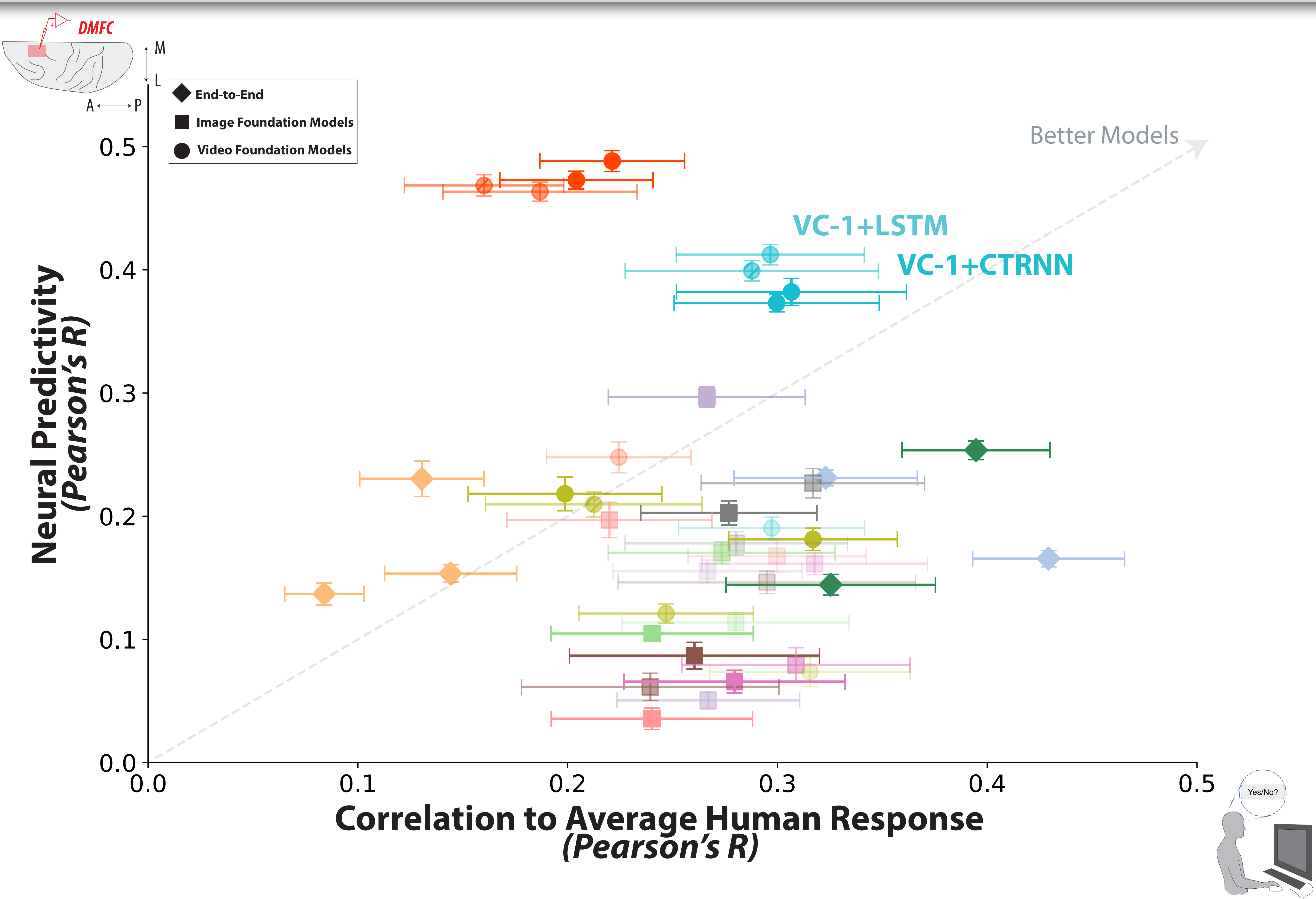
YES NO

Is the red object going to hit the yellow area?

Model Evaluations: Both Metrics

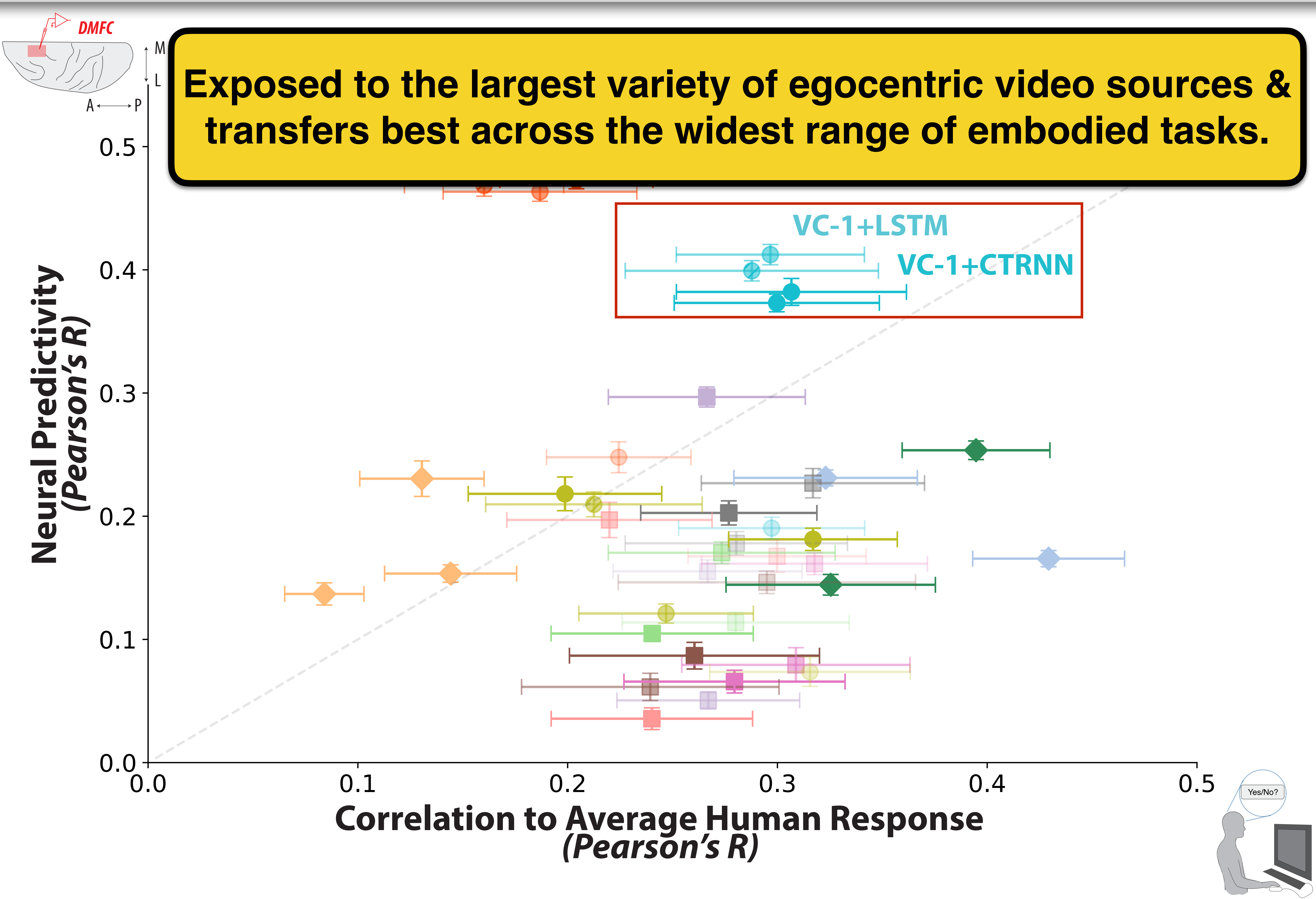


Dynamically-Equipped Video Foundation Models Can Match Both

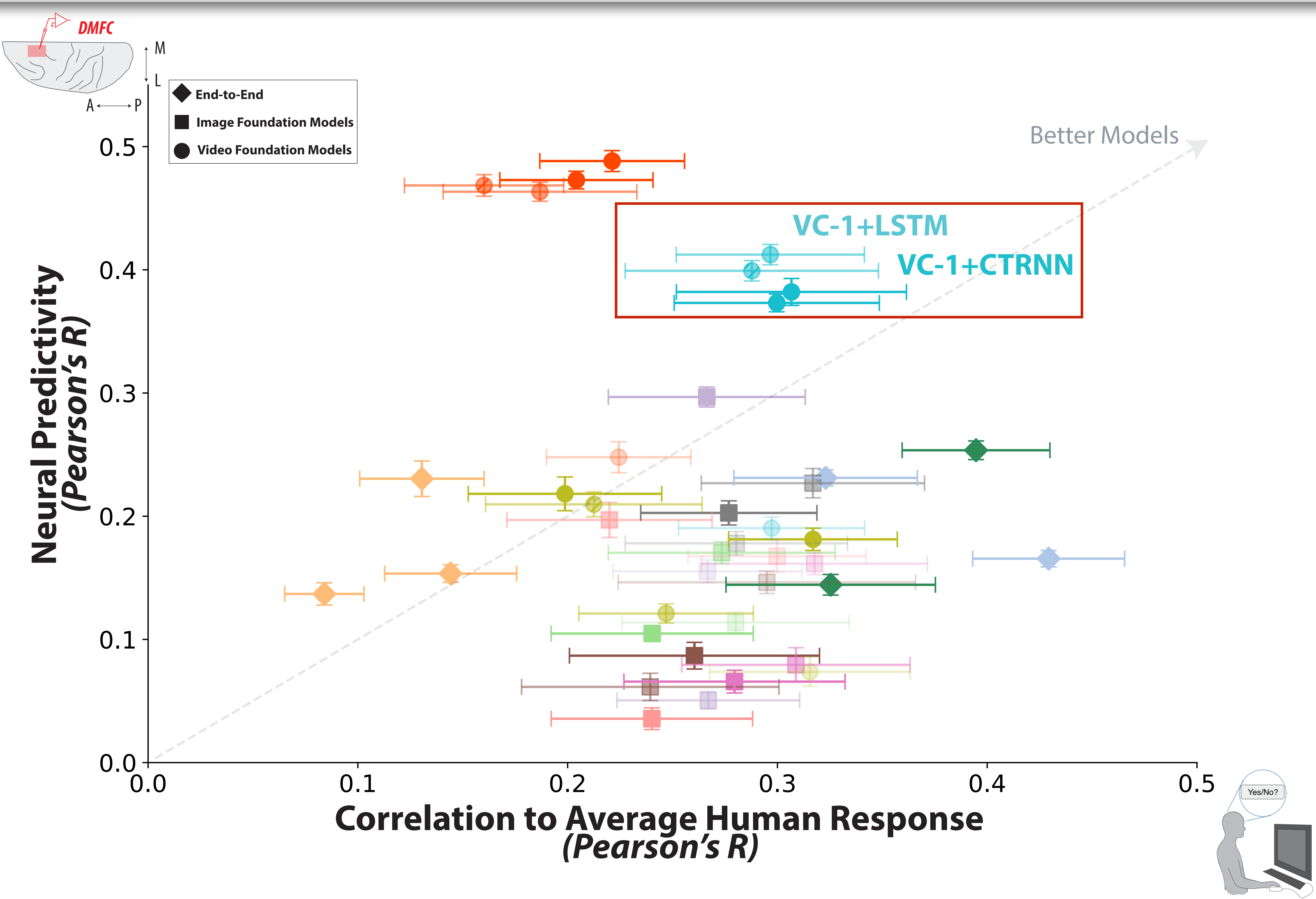


Dynamically-Equipped Video Foundation Models Can Match Both

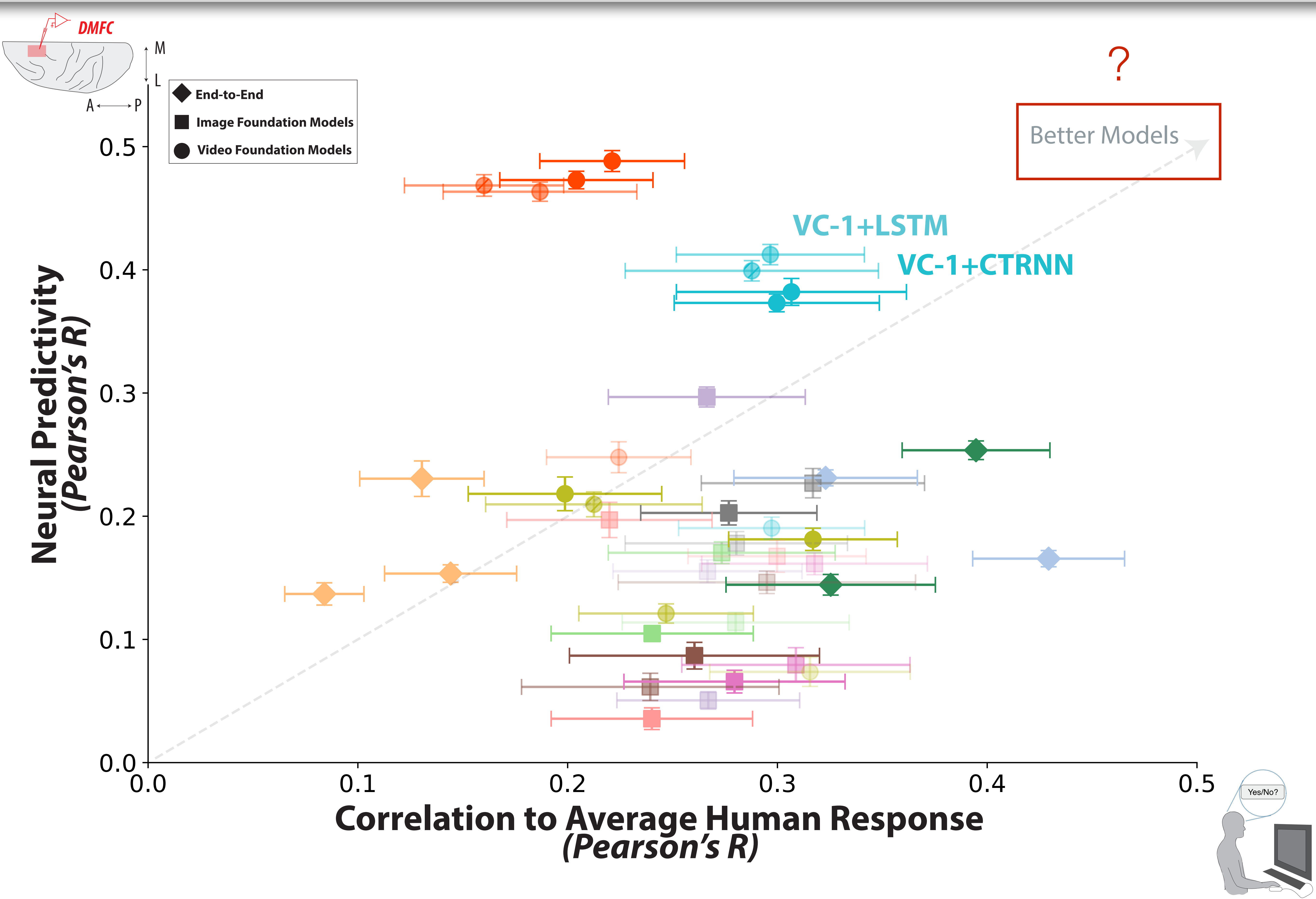
Exposed to the largest variety of egocentric video sources & transfers best across the widest range of embodied tasks.



Dynamically-Equipped Video Foundation Models Can Match Both



Future Directions



Outline

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

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

- Reusable Latent Representations for Primate Mental Simulation

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

- Heuristics for Interrogating Natural Intelligence

Outline

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

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

- Reusable Latent Representations for Primate Mental Simulation

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

- Heuristics for Interrogating Natural Intelligence

Heuristics for Interrogating Natural Intelligence

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

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

Heuristics for Interrogating Natural Intelligence

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

Heuristics for Interrogating Natural Intelligence

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

Heuristics for Interrogating Natural Intelligence

- **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

- **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.

Heuristics for Interrogating Natural Intelligence

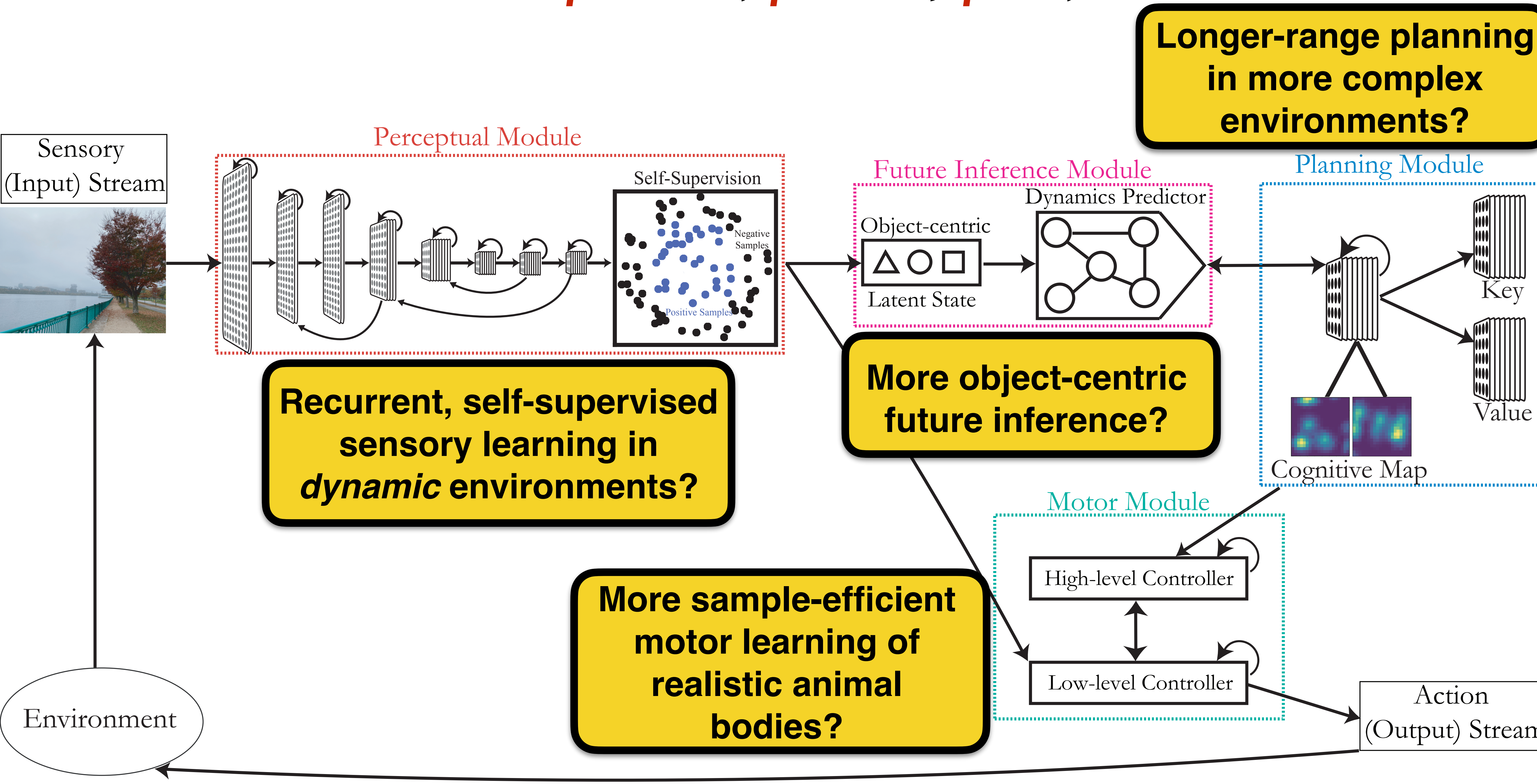
- **Incorporating Neuroscience Insights:**
  - **Connectomics:** Not usually a 1-1 mapping from a connectome to a functional model, and easy to get wrong. Rather, the best model often requires an *iterative balance* of functional optimization with macroscale structural constraints (e.g. shallow vs. deep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

Heuristics for Interrogating Natural Intelligence

- **<u>Incorporating Neuroscience Insights:</u>**
  - **Connectomics:** Not usually a 1-1 mapping from a connectome to a functional model, and easy to get wrong. Rather, the best model often requires an *iterative balance* of functional optimization with macroscale structural constraints (e.g. shallow vs. deep cortex).
  - **Ethology:** Ethology alone does not always give us the "correct" functional optimization target (e.g. for intermediate brain areas). Can be helpful to consider that the brain area is embodied in a larger cognitive agent.
- **<u>Incorporating AI Insights:</u>**
  - End-to-end reinforcement learning (RL) does *not* seem to give us neurally-aligned visual systems in both rodents and primates.
  - Suggests a possible functional modularization of optimization targets, with reusable SSL representations best matching visual areas overall.

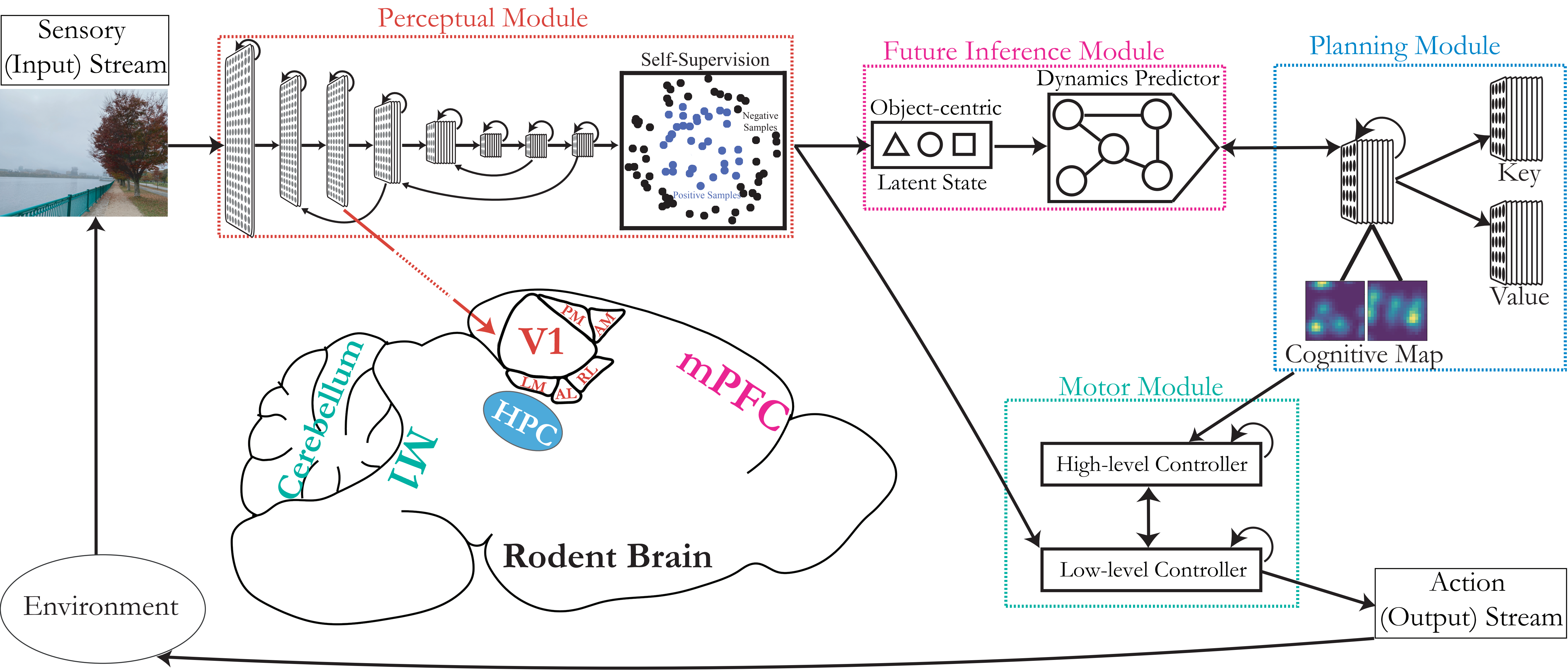
Next Steps: Modularized, Embodied Agents?

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

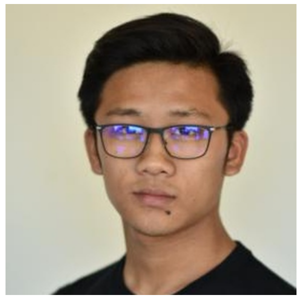


Next Steps: Modularized, Embodied Agents?

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



Acknowledgements


Nathan C.L. Kong


Chengxu Zhuang

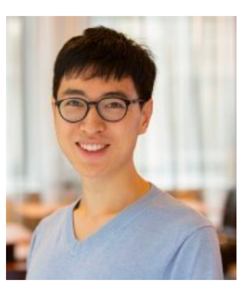

Justin L. Gardner


Anthony M. Norcia


Daniel Yamins


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

Contact:

anayebi@cs.cmu.edu

@aran_nayebi

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

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Funding:

Burroughs Wellcome Fund CASI Award

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

Stanford Neurosciences PhD Program

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