Goal-Driven Models of Physical Understanding

Aran Nayebi
McGovern Institute, MIT

CS 375/Psych 249
Stanford University
2024.02.13

MCGOVERN INSTITUTE

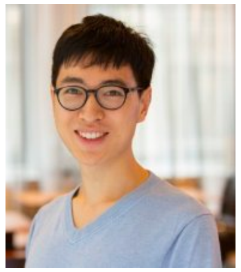
K. LISA YANG ICoN CENTER
Integrative Computational Neuroscience

Visually-Grounded Mental Simulation

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


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

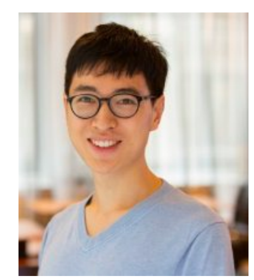
https://arxiv.org/abs/2305.11772

Visually-Grounded Mental Simulation

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


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

https://arxiv.org/abs/2305.11772

Motivation


氷みくじ
氷みくじ


WHOLE WHEAT
Müeller
PRO-SERIES
MULTI
LODGE
CAST IRON
levoit

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

Motivation

Infer:
Has this ice block been out longer?

Predict:
Will this box support me?

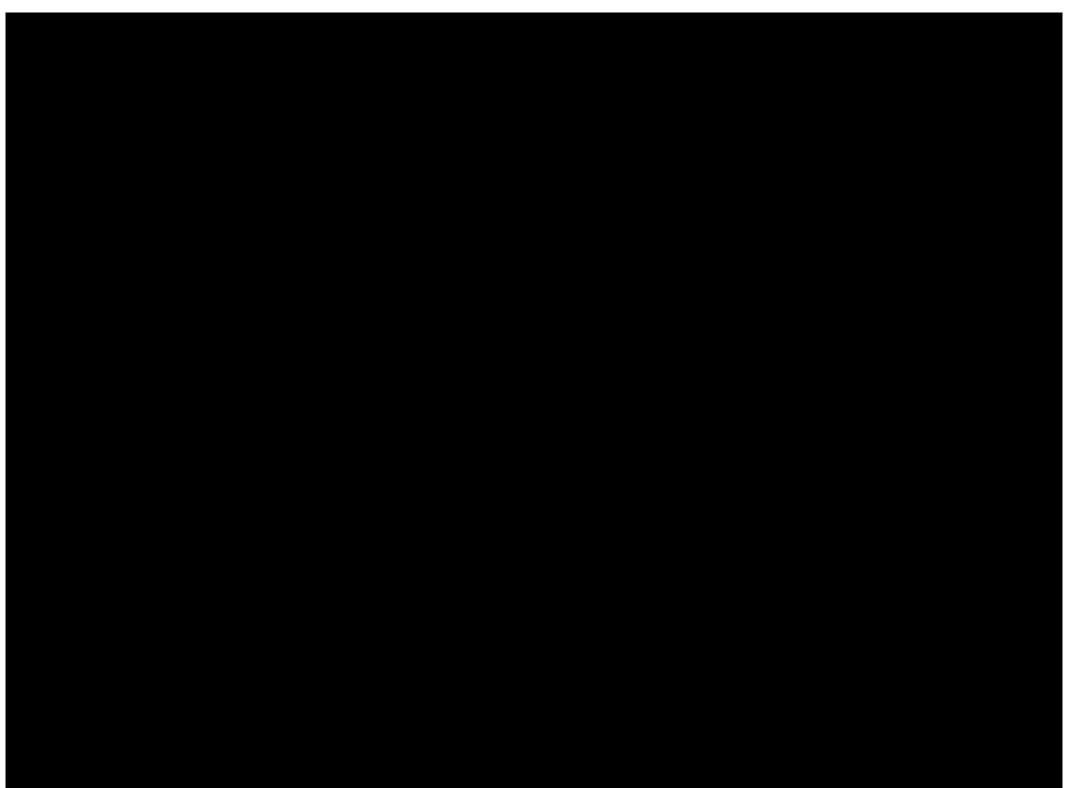
Motivation

Infer:
Has this ice block been out longer?

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

Predict:
Will this box support me?

Motivation



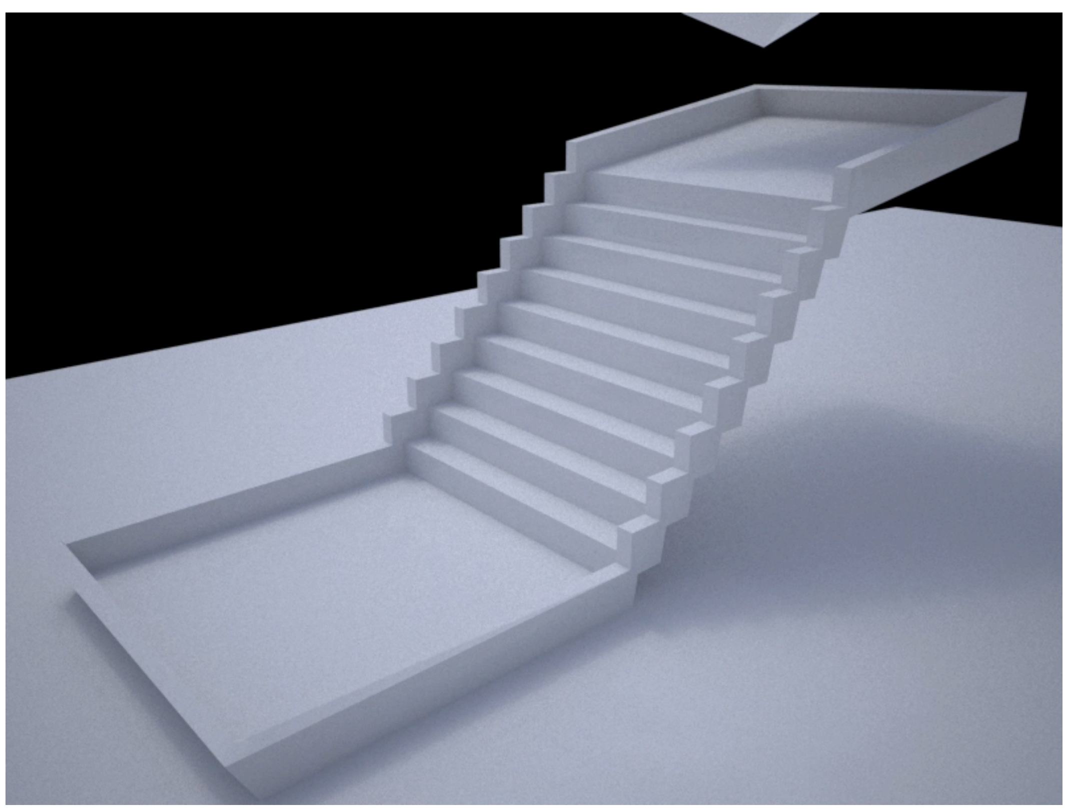
Infer:
Has this ice block been out longer?

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

Predict:
Will this box support me?

Motivation

Infer:
Has this ice block been out longer?



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

Predict:
Will this box support me?

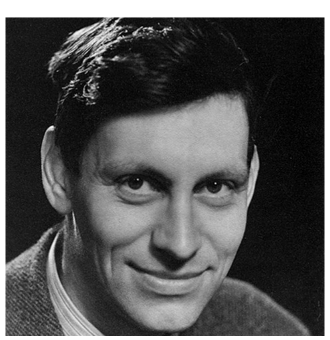
The Mental Simulation Hypothesis

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



Kenneth Craik

The Mental Simulation Hypothesis

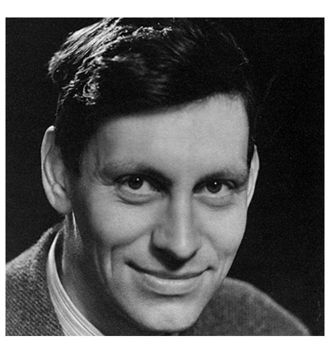
The Nature of Explanation

Pre-dates the modern computer!

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.**



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.**

Focus on *physical* simulation

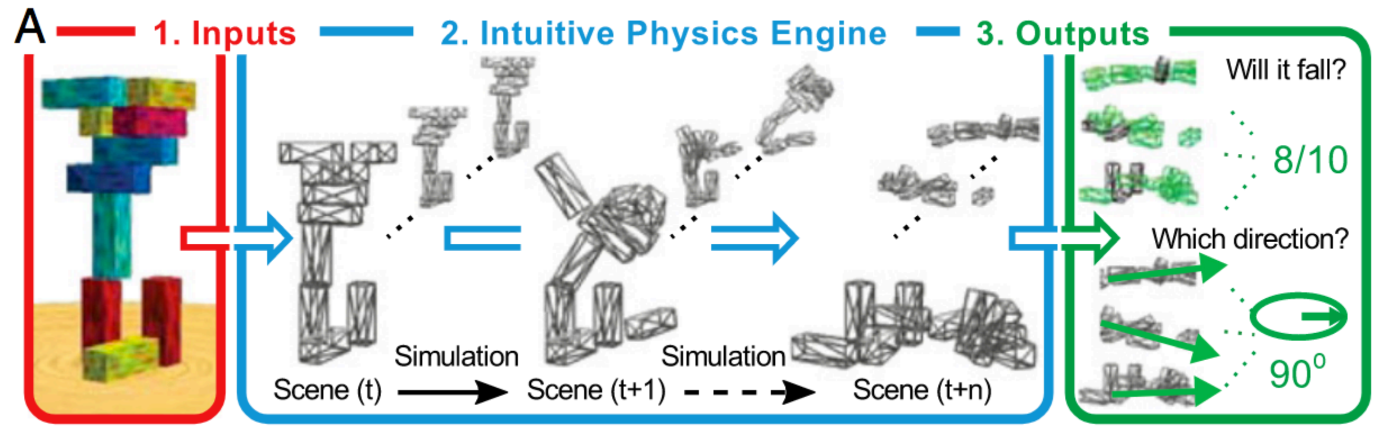
The Mental Simulation Hypothesis: Behavioral Evidence

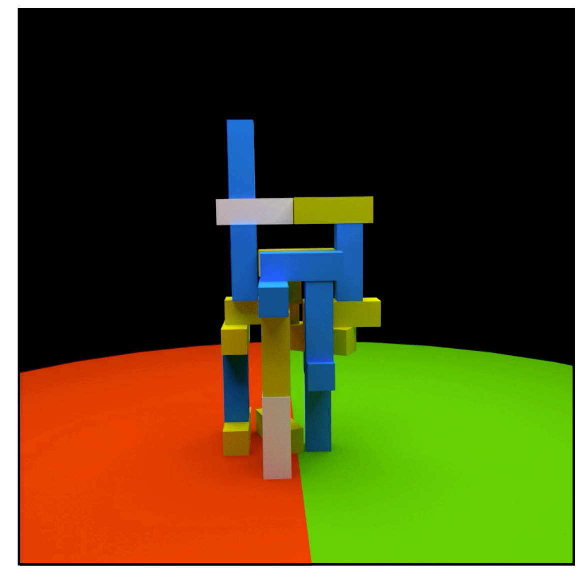
The Nature of Explanation

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.**

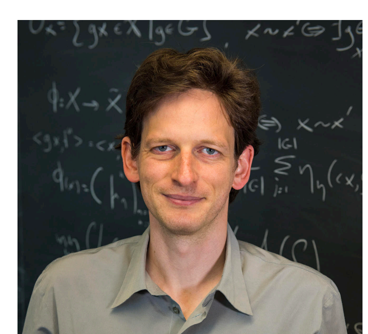




Ullman et al. 2017 *Battaglia, Hamrick, Tenenbaum 2013*

Intuitive Physics Engine (IPE) can match human physical judgements


Peter Battaglia

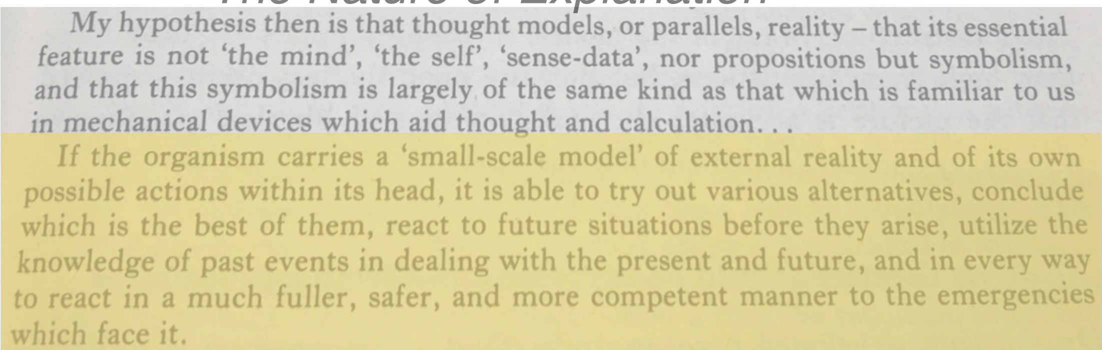

Tomer Ullman


Jessica Hamrick


Joshua Tenenbaum

The Mental Simulation Hypothesis: Human Neuroimaging Evidence

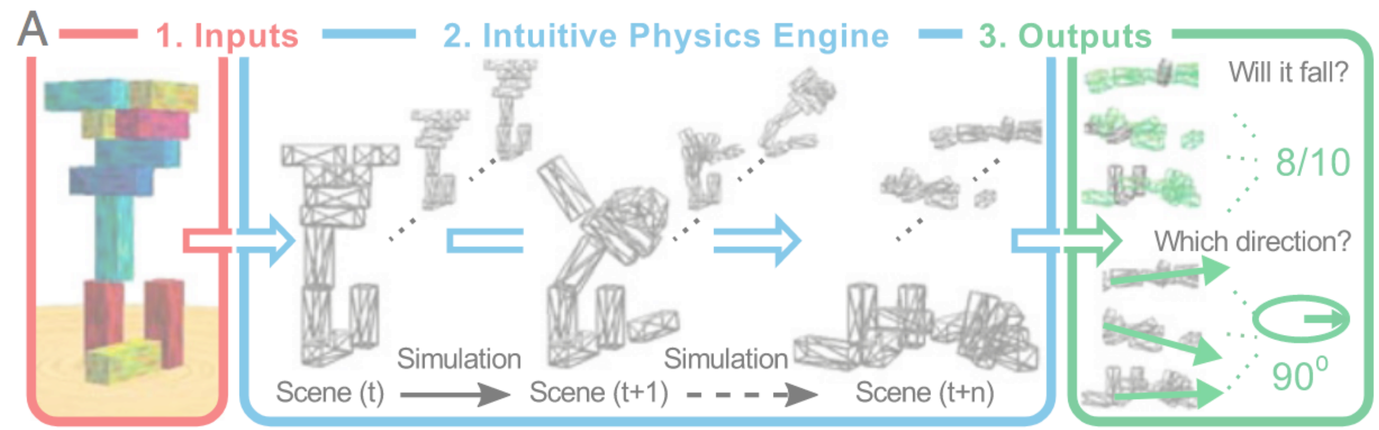
The Nature of Explanation

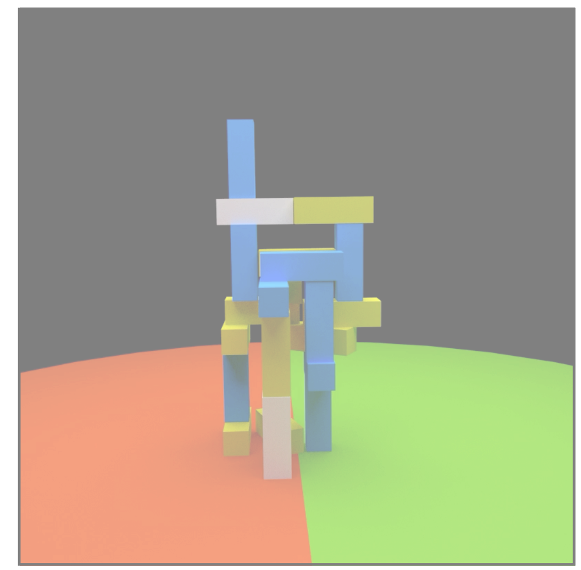

My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation.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

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.**

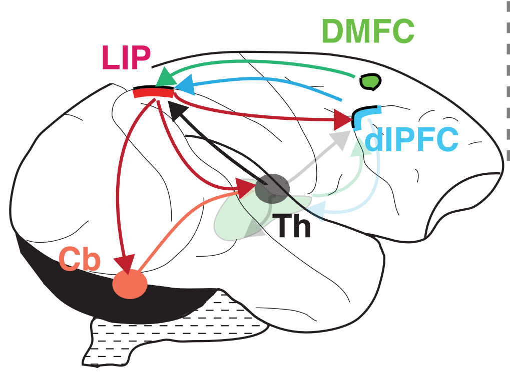
The Brain's "Physics Engine"



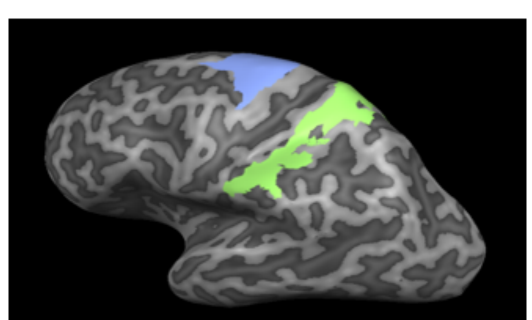


Ullman et al. 2017

Battaglia, Hamrick, Tenenbaum 2013



Fronto-Parietal Network



Nancy Kanwisher

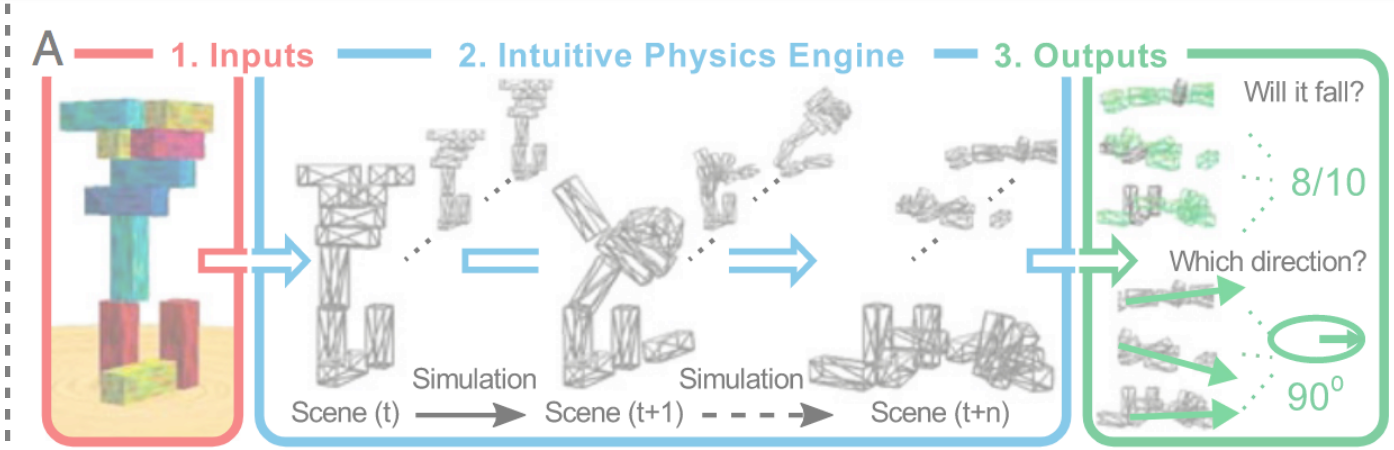
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

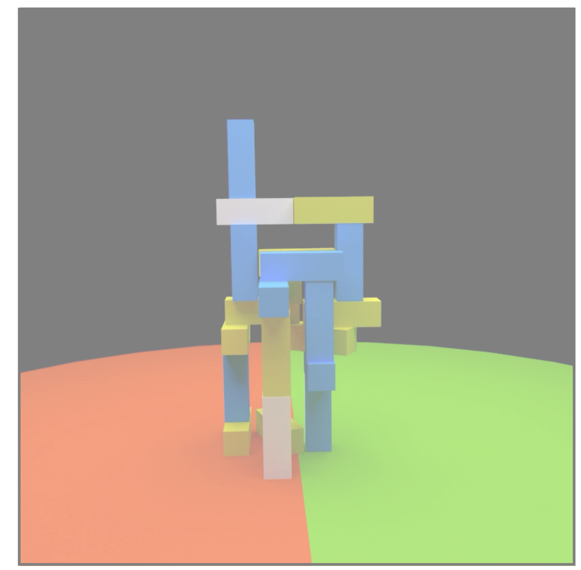
The Nature of Explanation

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.**



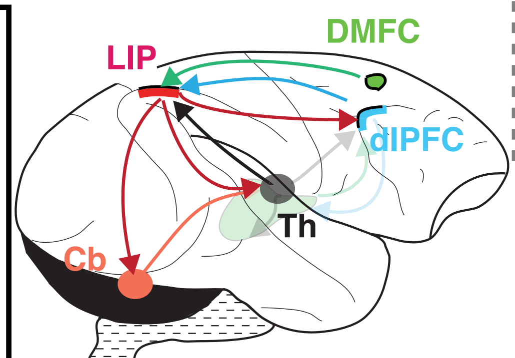


Ullman et al. 2017

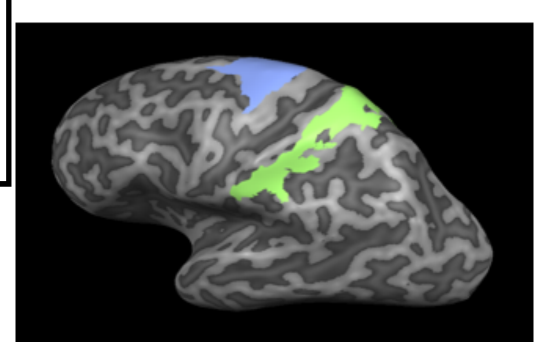
Battaglia, Hamrick, Tenenbaum 2013

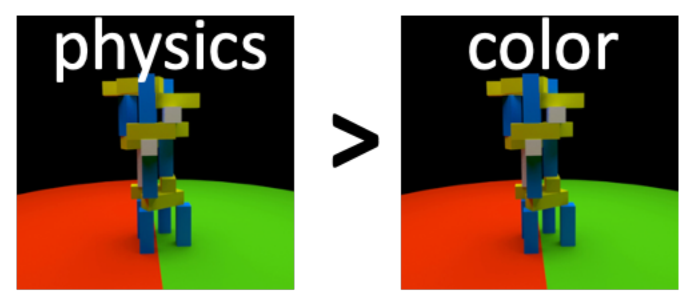
The Brain's "Physics Engine"

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



Fronto-Parietal Network





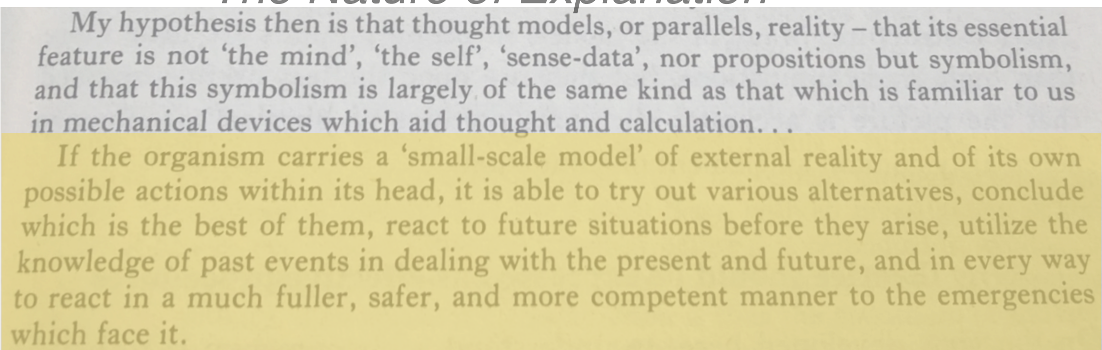
Fischer et al. 2016

Jason Fischer

Nancy Kanwisher

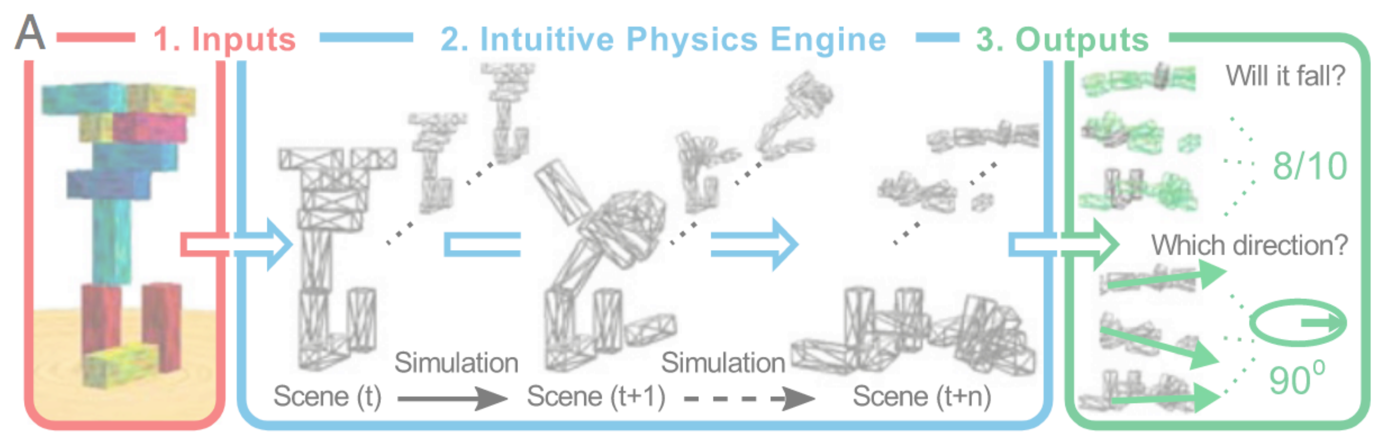
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

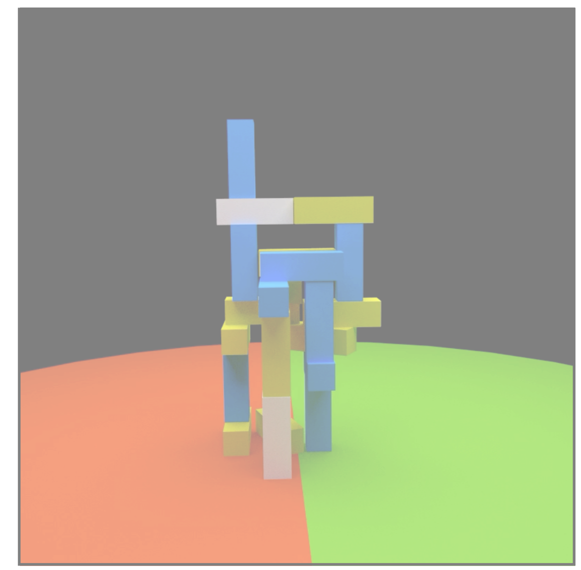
The Nature of Explanation


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.**

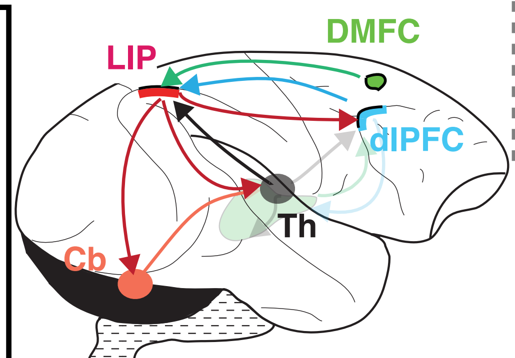




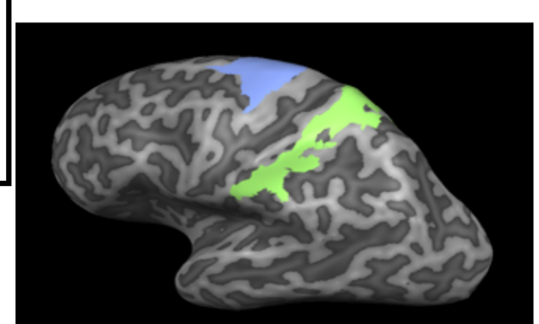
Ullman et al. 2017 *Battaglia, Hamrick, Tenenbaum 2013*

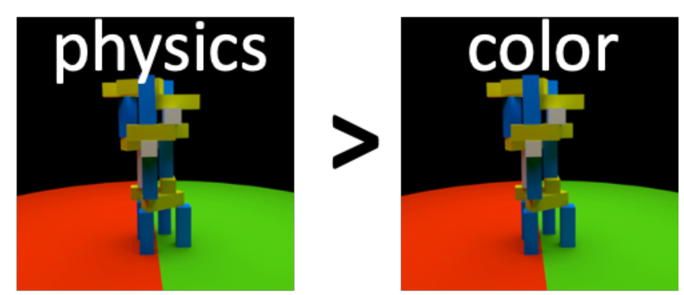
The Brain's "Physics Engine"

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

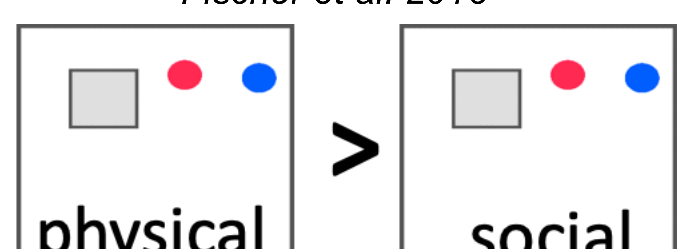


Fronto-Parietal Network





Fischer et al. 2016



Schwettmann et al. 2019

Sarah Schwettmann

Nancy Kanwisher

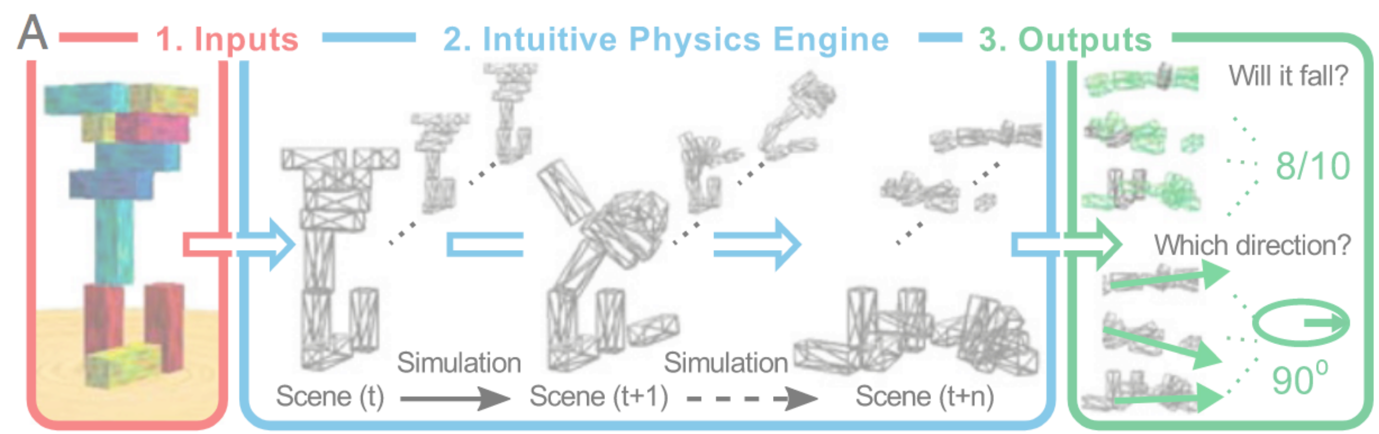
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

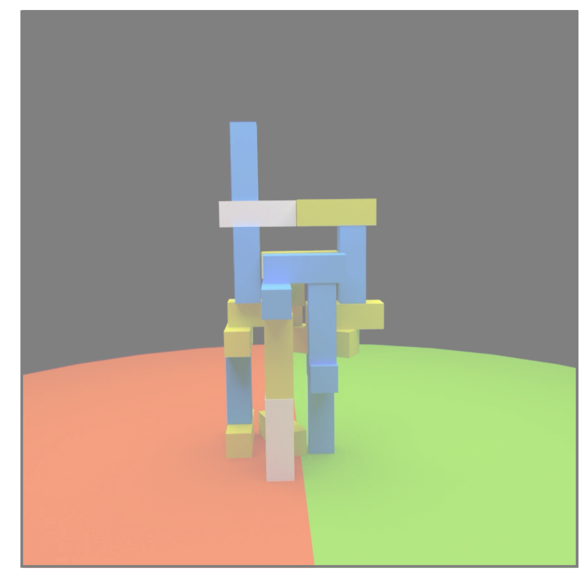
The Nature of Explanation

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.**

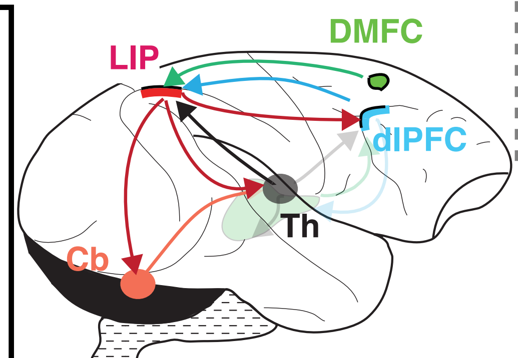




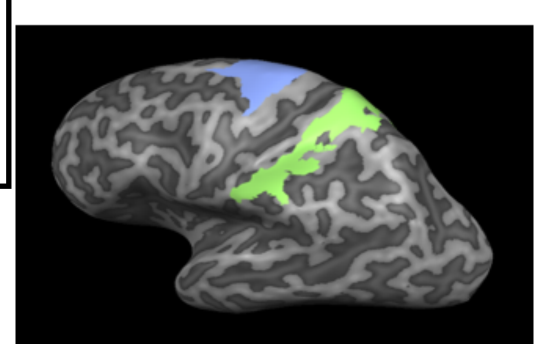
Ullman et al. 2017 Battaglia, Hamrick, Tenenbaum 2013

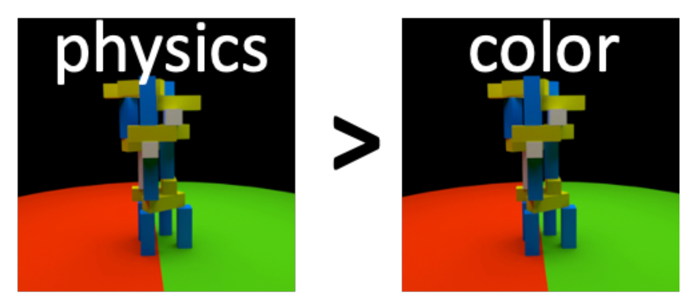
The Brain's "Physics Engine"

- 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*)

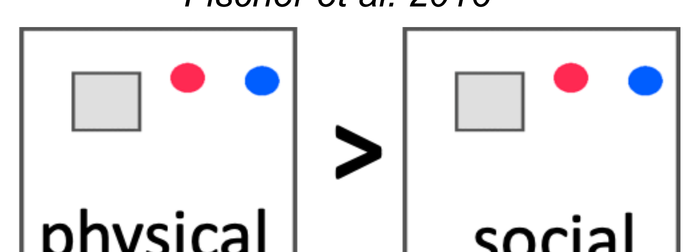


Fronto-Parietal Network





Fischer et al. 2016



Schwettmann et al. 2019

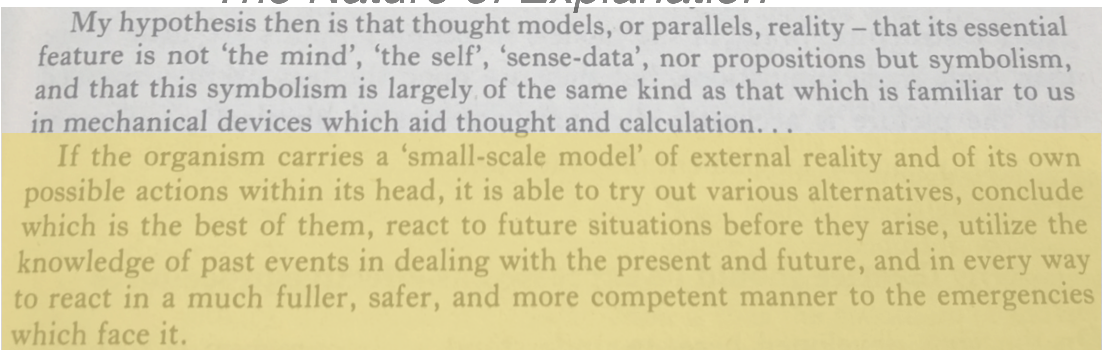
Pramod et al. 2022

RT Pramod

Nancy Kanwisher

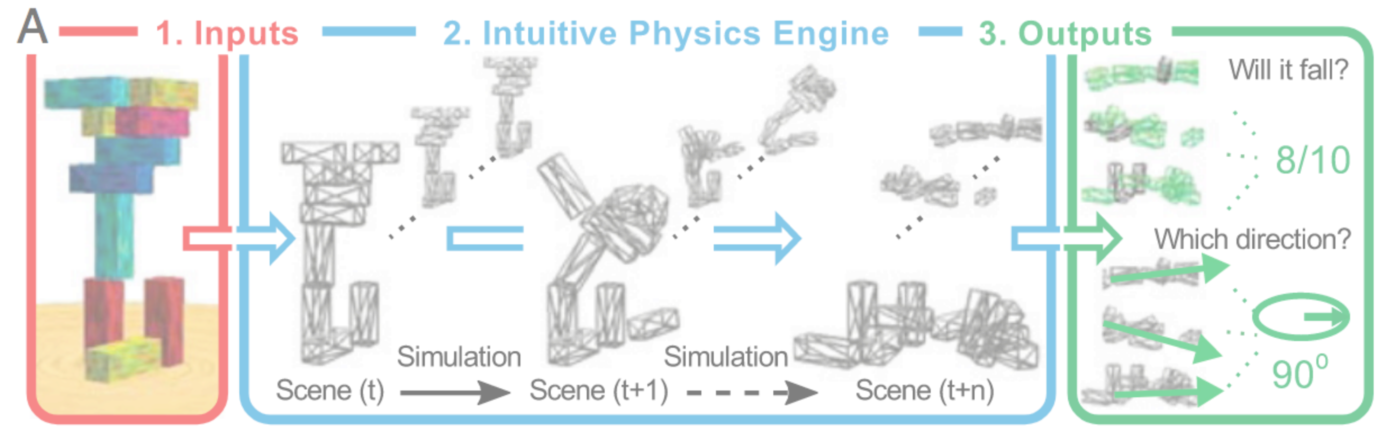
The Mental Simulation Hypothesis: Primate Electrophysiological Evidence

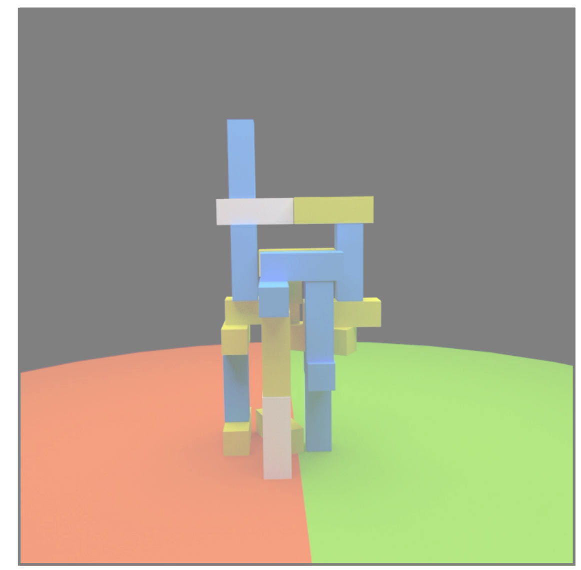
The Nature of Explanation


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. . .

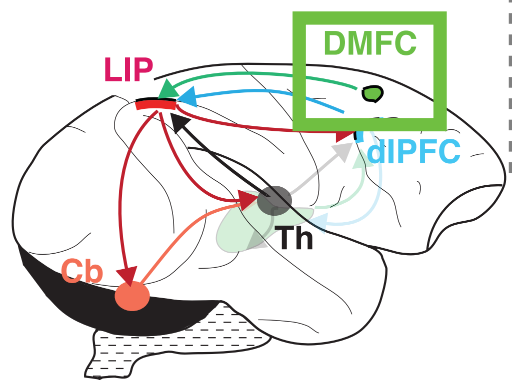
If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present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

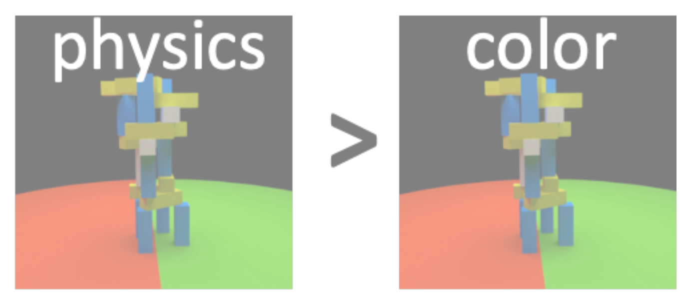




Ullman et al. 2017 *Battaglia, Hamrick, Tenenbaum 2013*

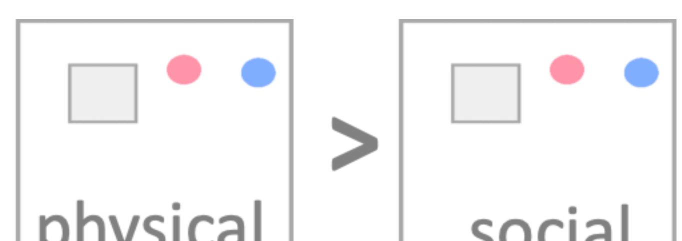


Fronto-Parietal Network

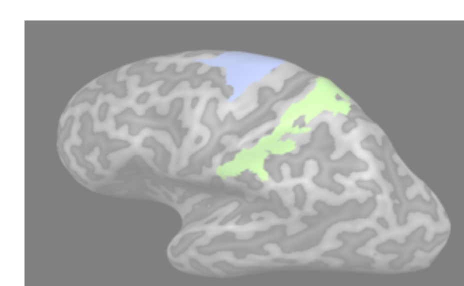


Fischer et al. 2016

Pramod et al. 2022



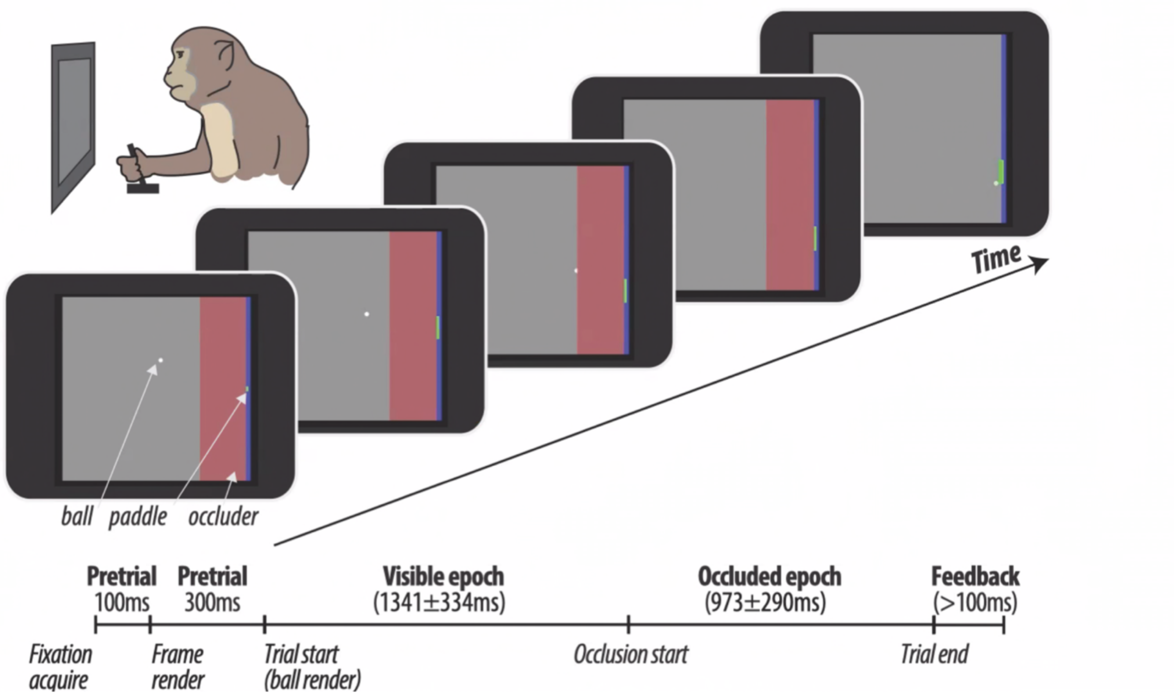
Schwettmann et al. 2019



Rishi Rajalingham

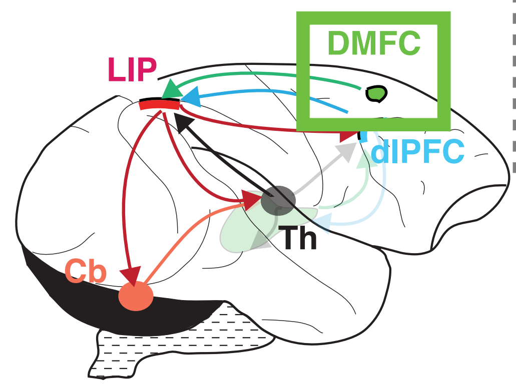
Mehrdad Jazayeri

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



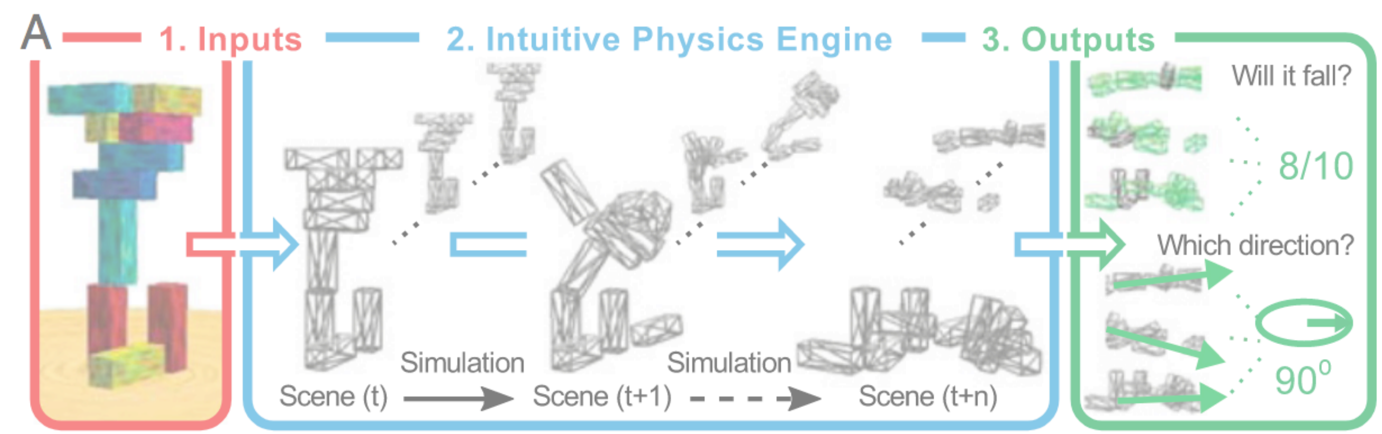
Fronto-Parietal Network

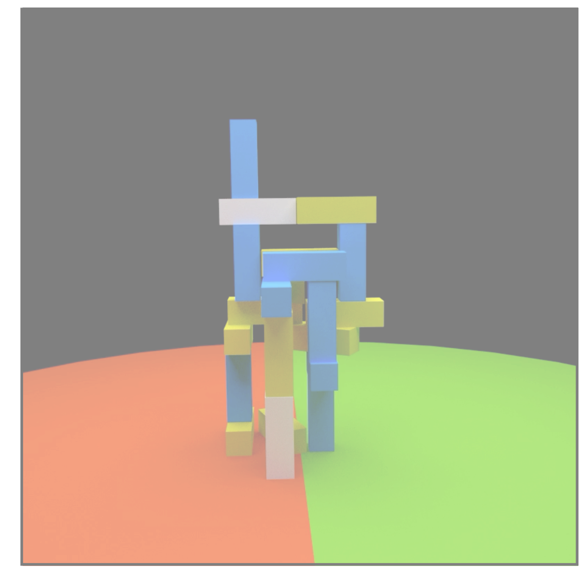
Dynamic tracking of objects in the macaque dorsomedial frontal cortex

Rishi Rajalingham, Hansem Sohn, Mehrdad Jazayeri
doi: https://doi.org/10.1101/2022.06.24.497529

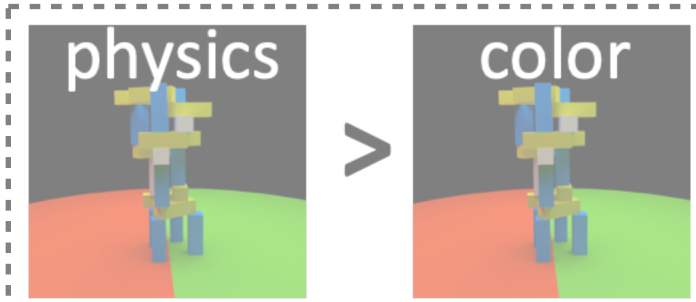

Rishi Rajalingham


Mehrdad Jazayeri



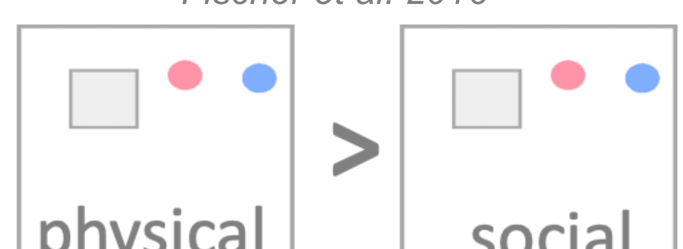


Ullman et al. 2017 *Battaglia, Hamrick, Tenenbaum 2013*

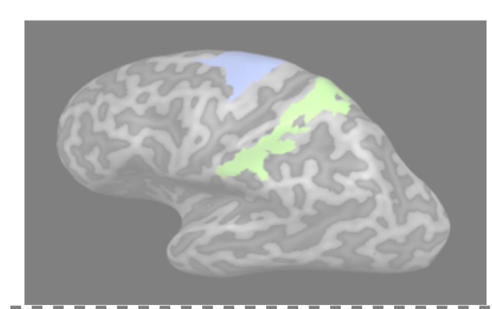


Fischer et al. 2016

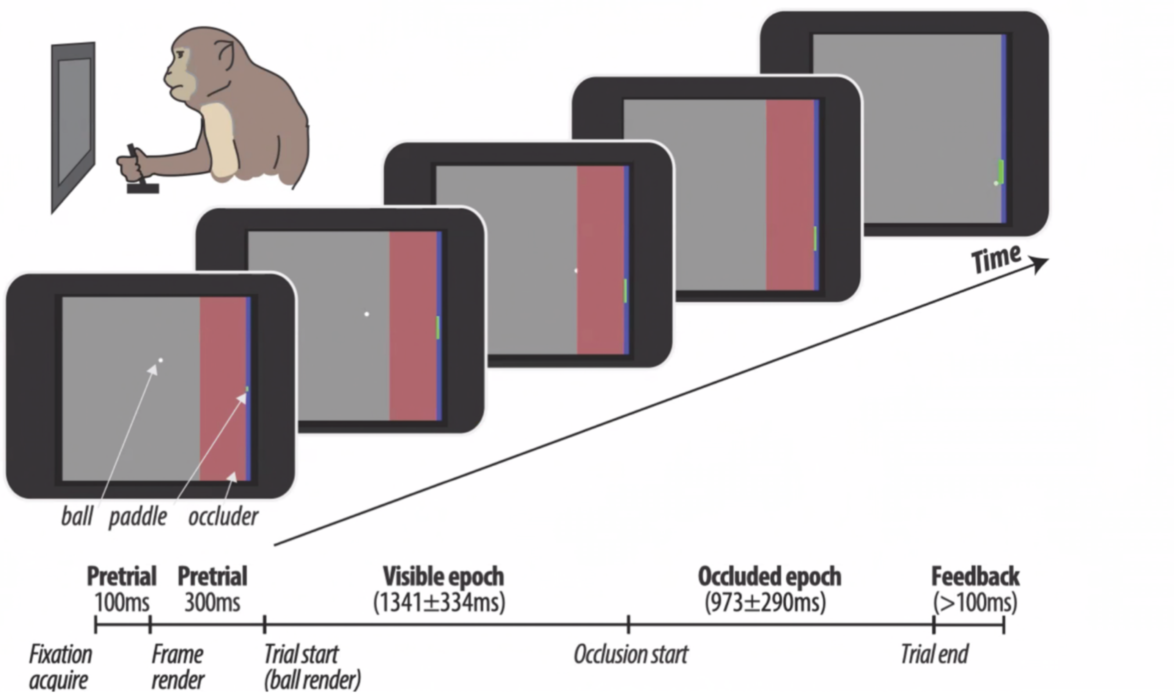
Pramod et al. 2022

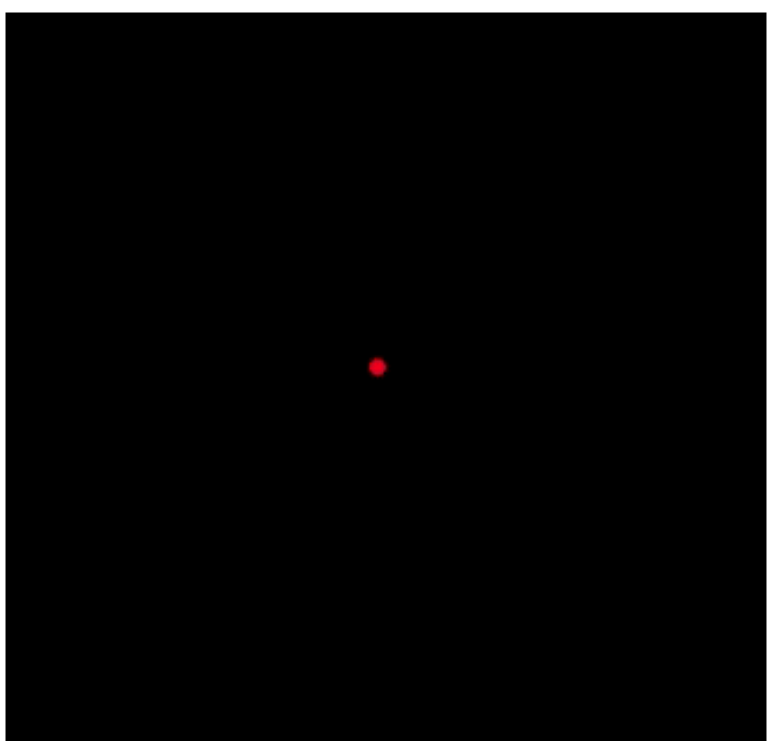


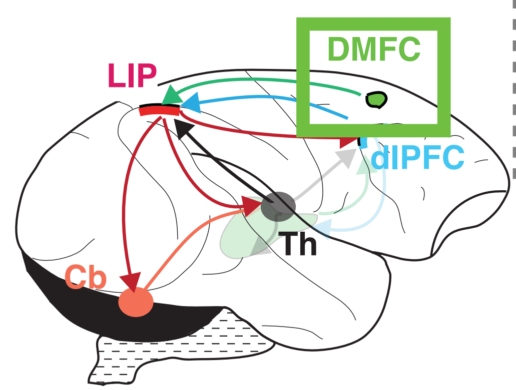
Schwettmann et al. 2019



The Mental Simulation Hypothesis: Primate Electrophysiological Evidence

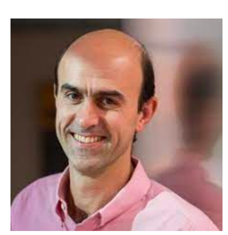




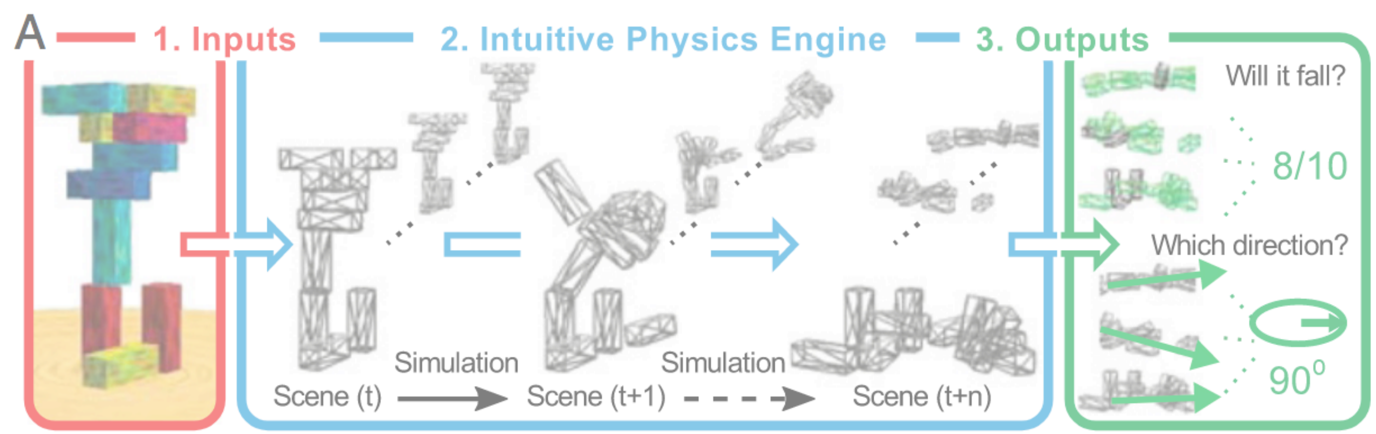


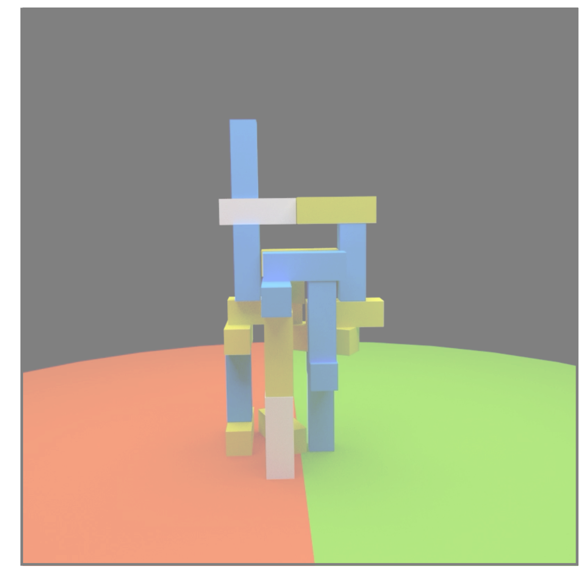
Fronto-Parietal Network

Rishi Rajalingham



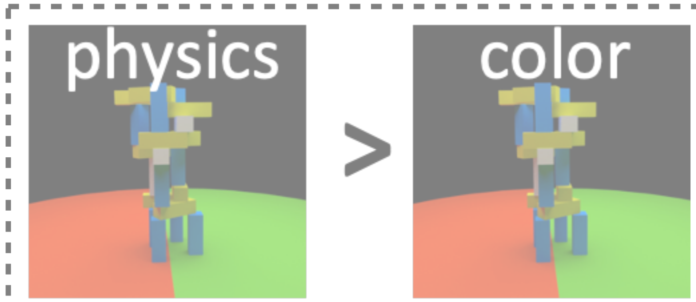
Mehrdad Jazayeri





Ullman et al. 2017

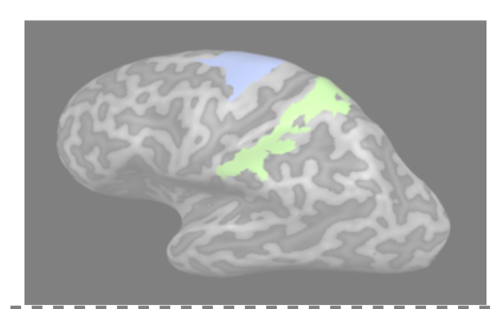
Battaglia, Hamrick, Tenenbaum 2013



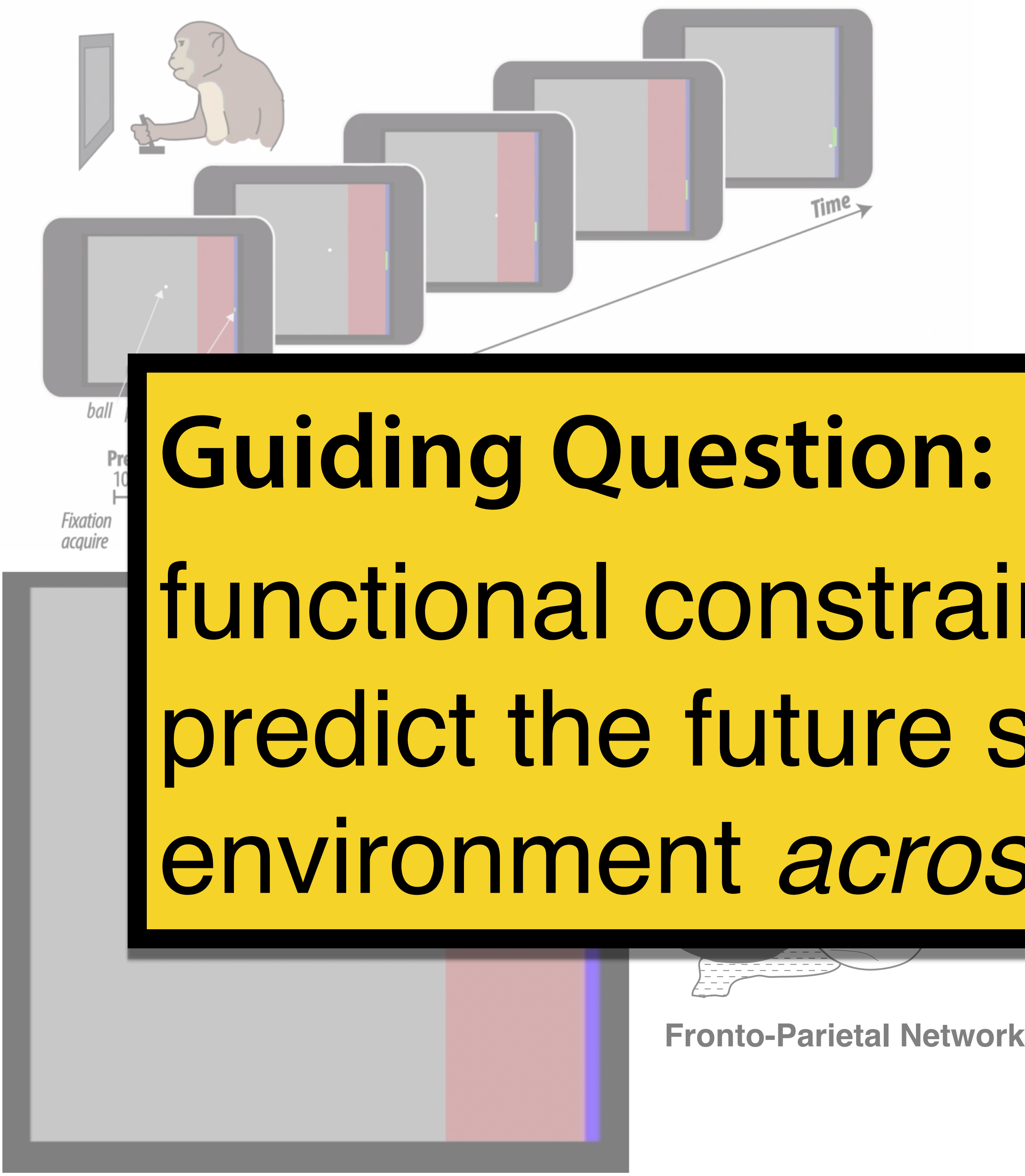
Fischer et al. 2016

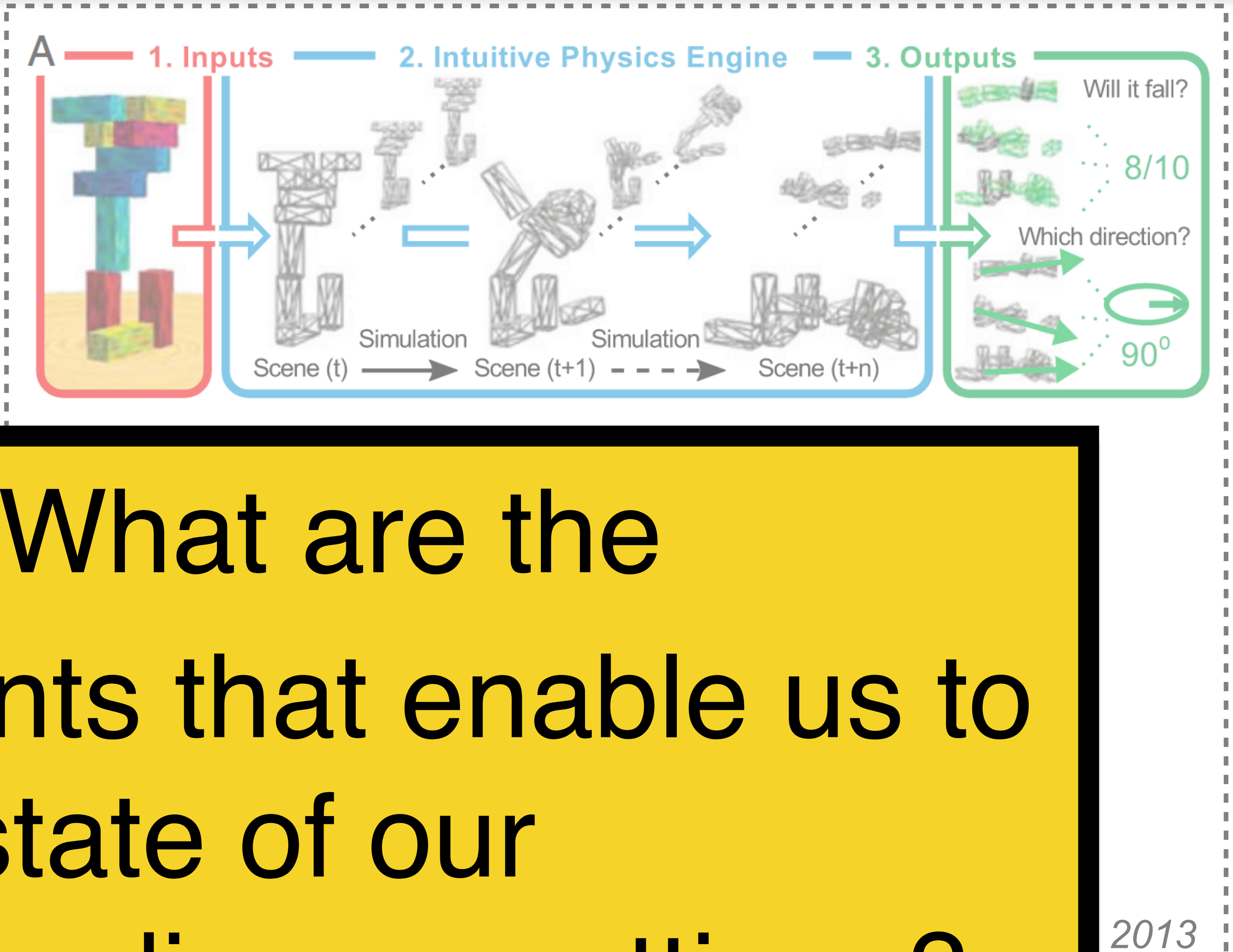
Pramod et al. 2022

Schwettmann et al. 2019

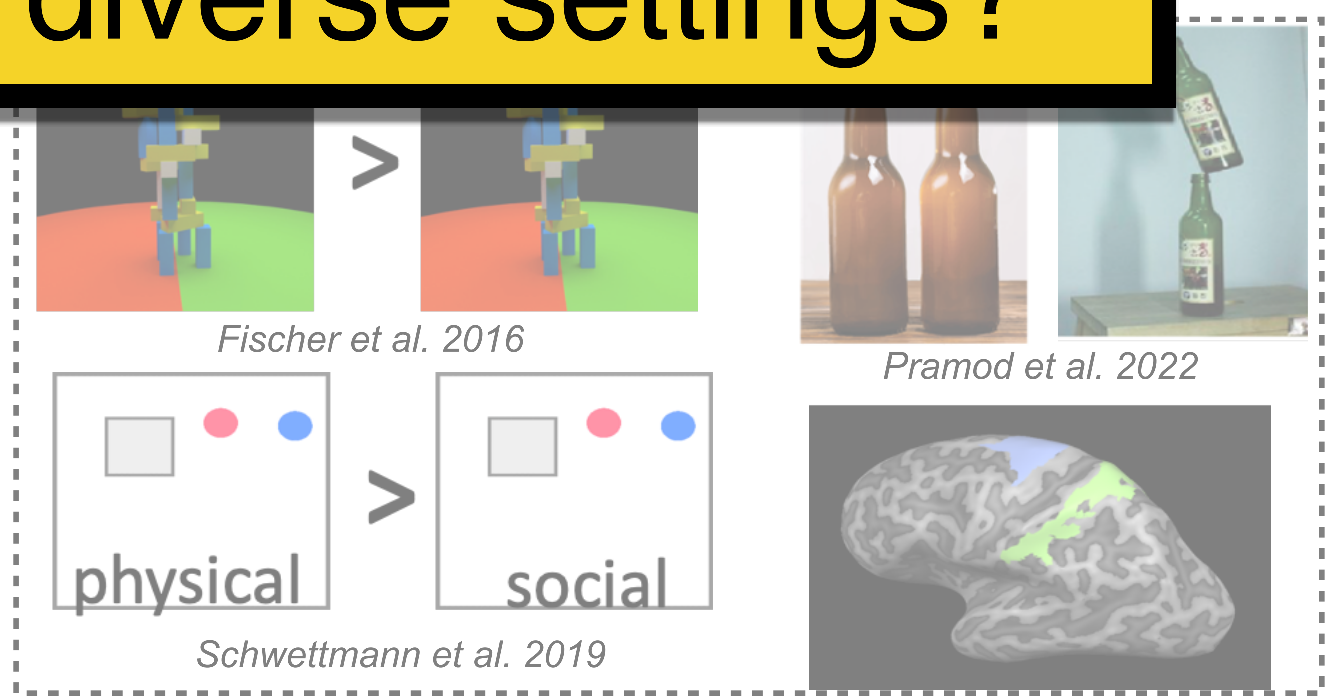


Functional Constraints of Mental Simulation *Across* Environments?

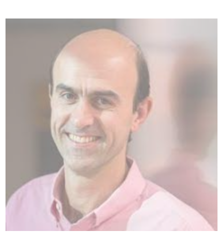




Guiding Question: What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?




Rishi Rajalingham


Mehrdad Jazayeri

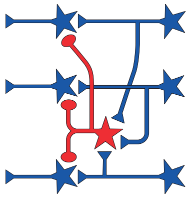
Defining Hypotheses: Goal-Driven Approach

L = *learning rule*

"Natural selection + plasticity"

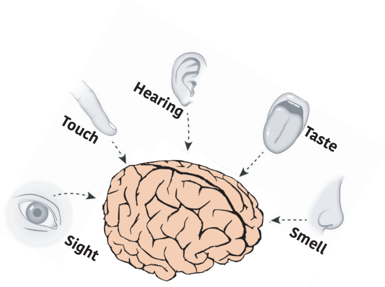
T = *task loss*

"Ecological niche/ behavior"



"Circuit"

A = *architecture class*



"Environment"

D = *data stream*

Defining Hypotheses: Goal-Driven Approach

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

"Circuit"

A = *architecture class*

"Environment"

D = *data stream*

Defining Hypotheses: Goal-Driven Approach

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

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

"Circuit"

A = *architecture class*

"Environment"

D = *data stream*

Defining Hypotheses: Goal-Driven Approach

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

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

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

"Circuit"

A = *architecture class*

"Environment"

D = *data stream*

Defining Hypotheses: Goal-Driven Approach

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

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").

"Circuit"

A = *architecture class*

"Environment"

D = *data stream*

Defining Hypotheses: Goal-Driven Approach

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

"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").

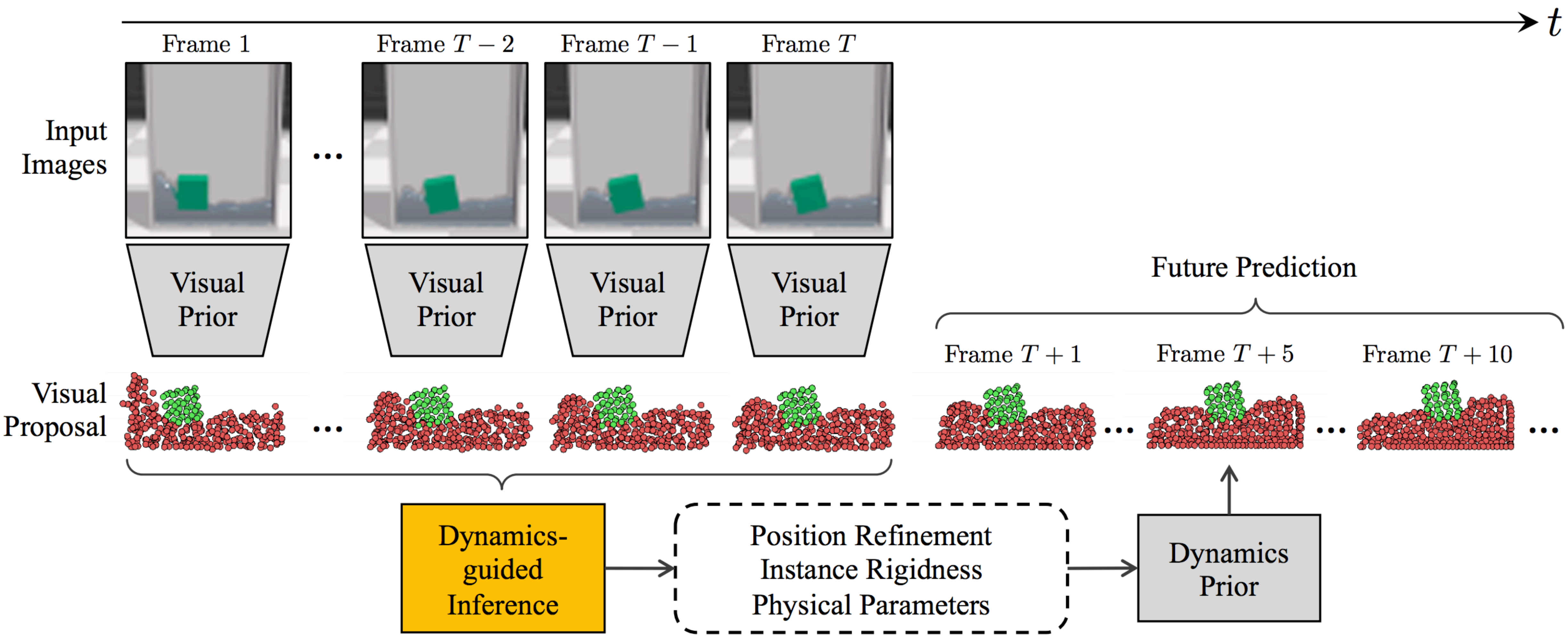
"Circuit"

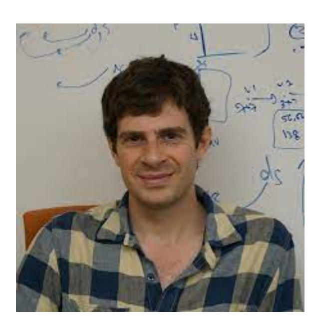
A = *architecture class*

"Environment"

D = *data stream*

Comparing Oracle Models to Human Physical Prediction





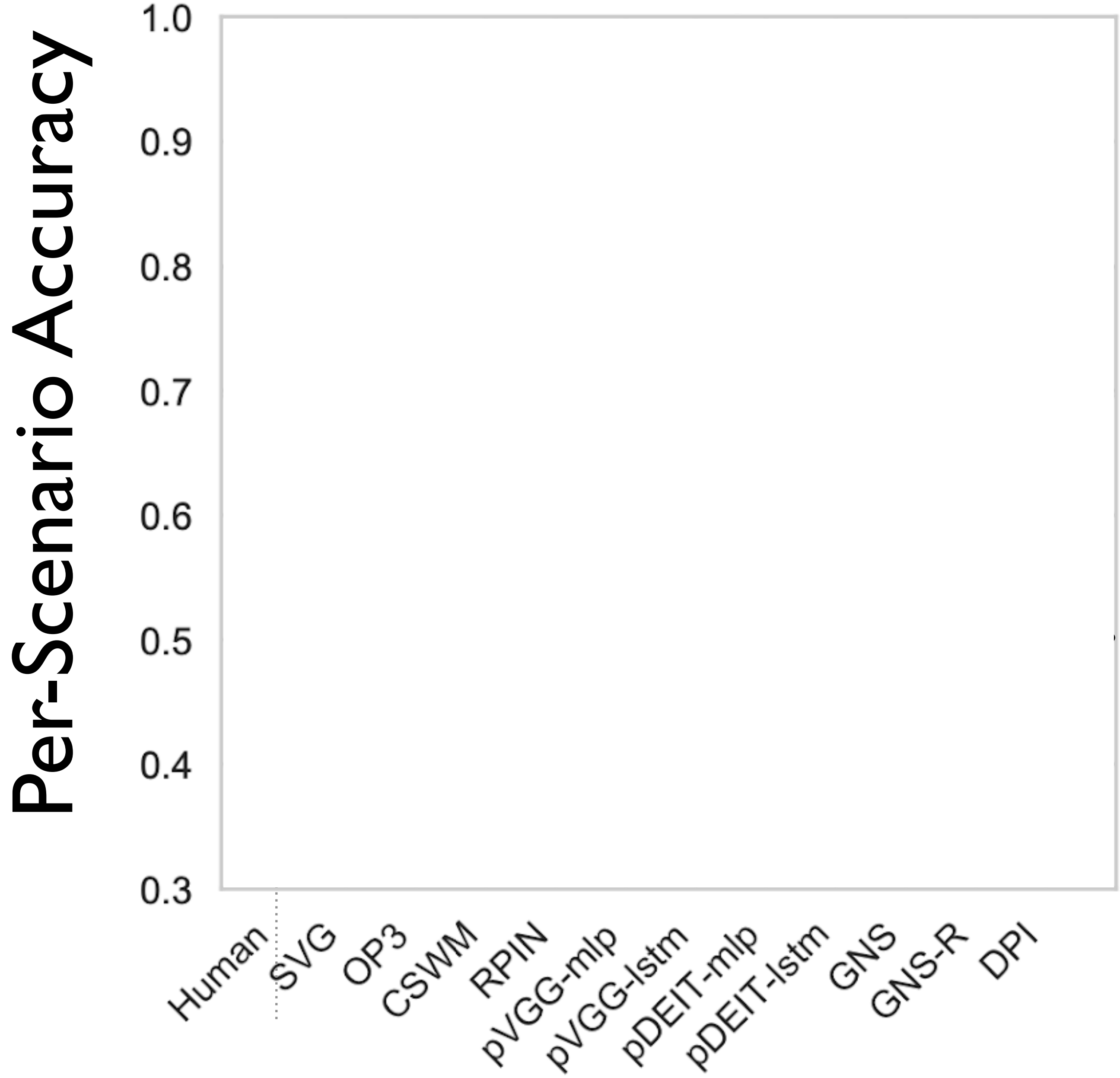
Daniel Bear

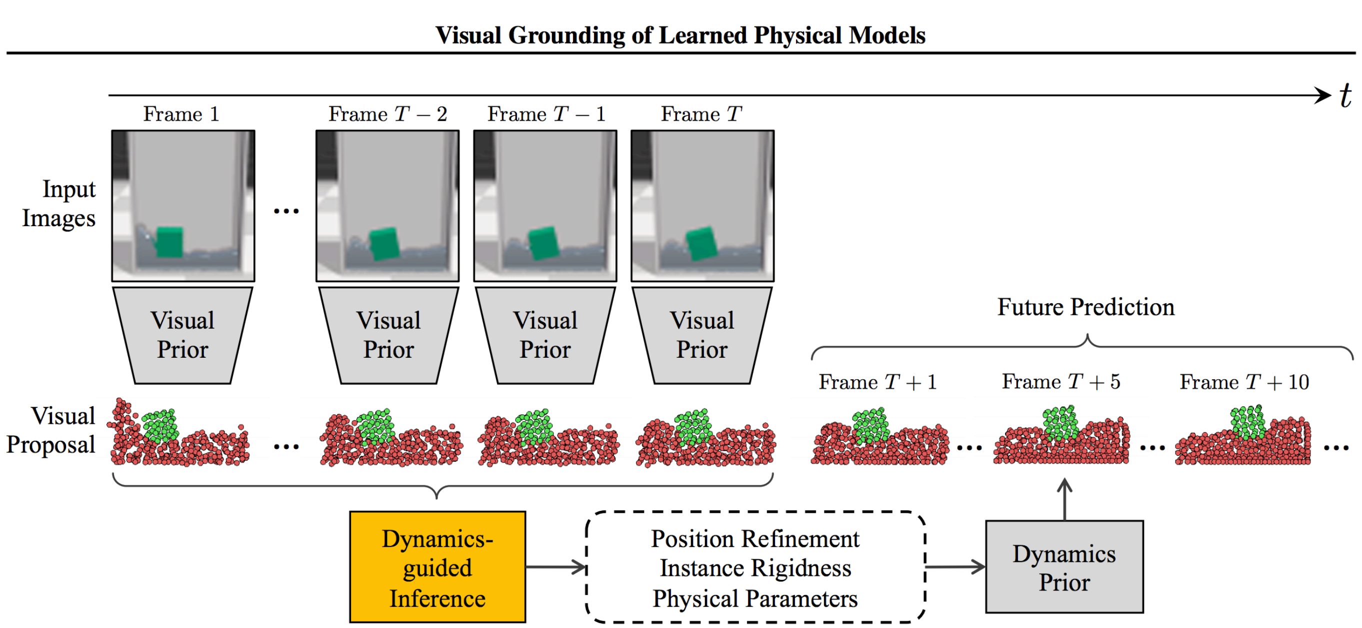
Joshua Tenenbaum

Daniel Yamins

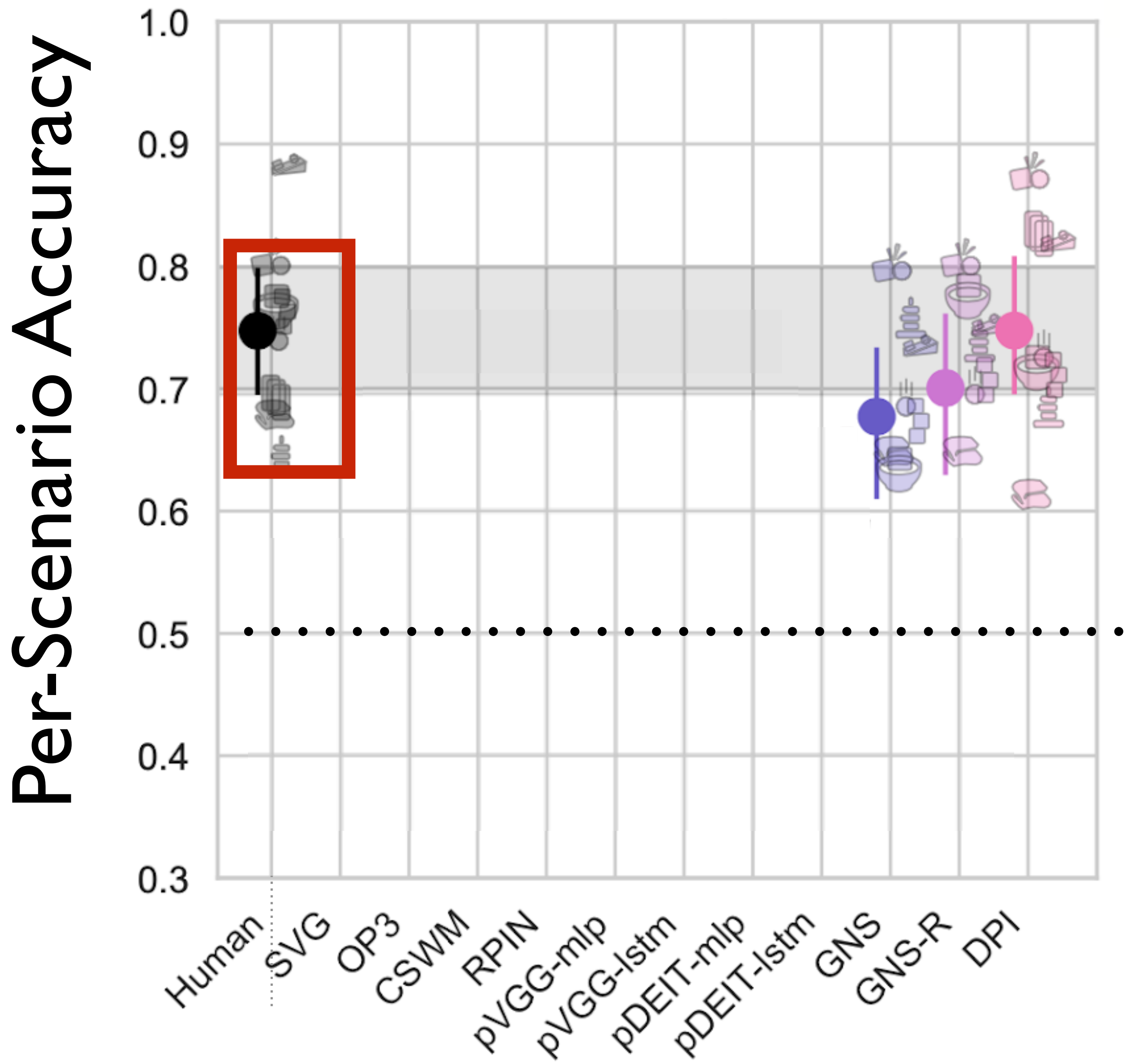
Judith Fan

Comparing Oracle Models to Human Physical Prediction

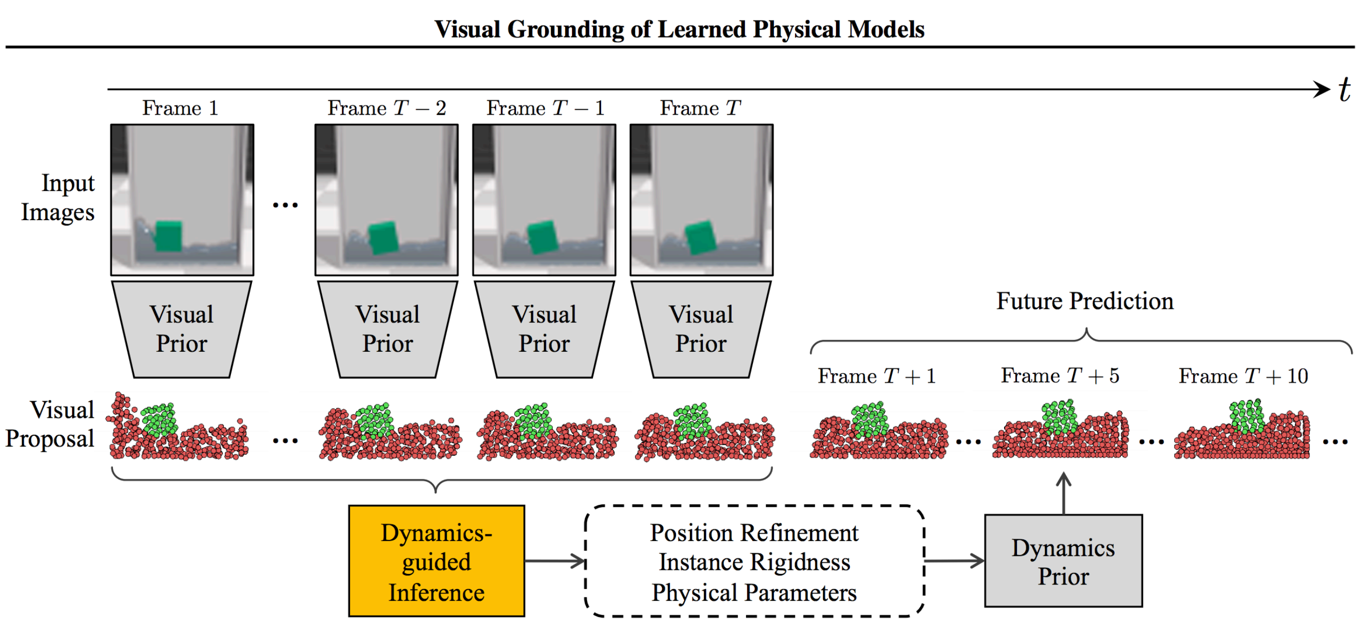




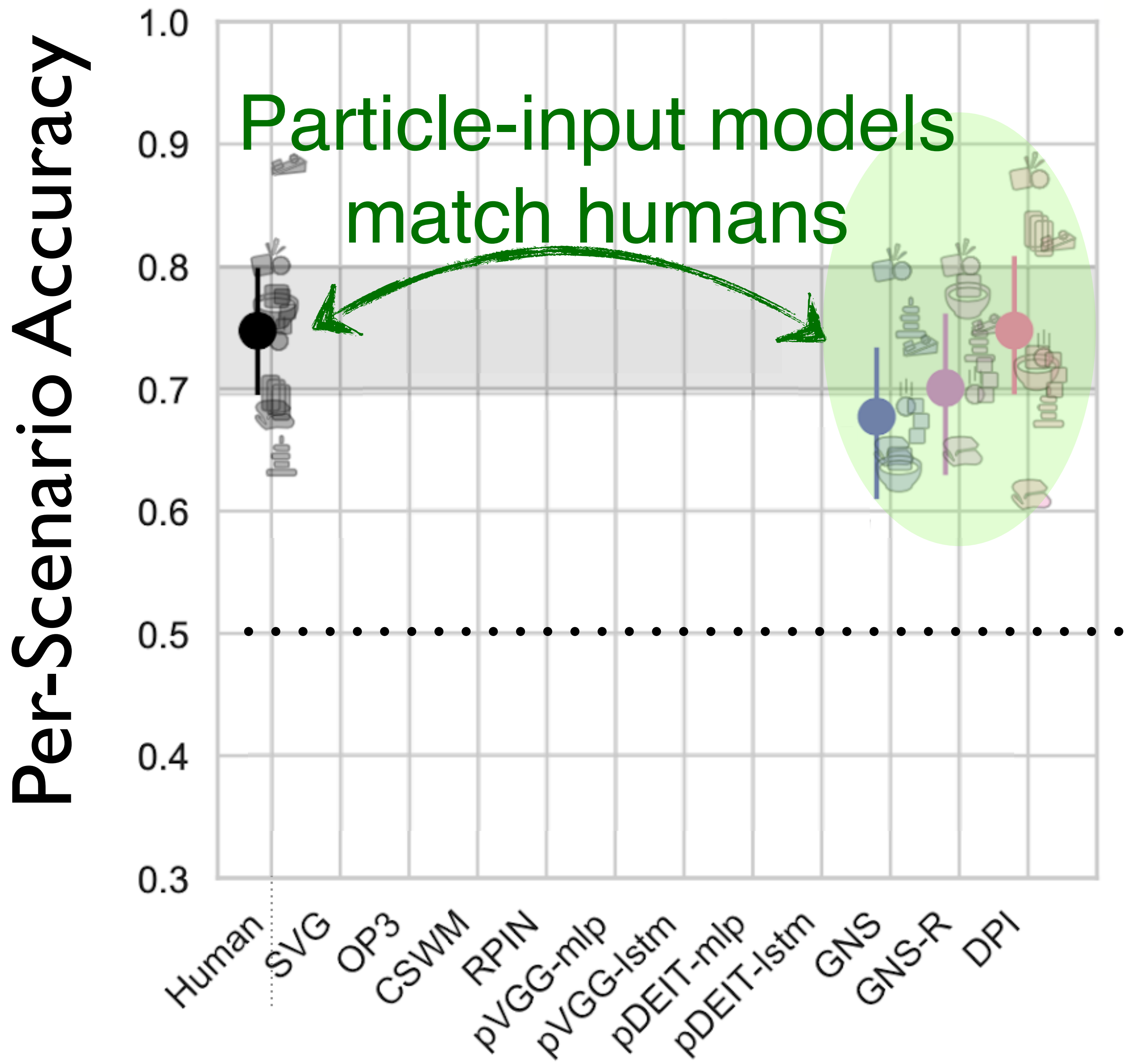
Comparing Oracle Models to Human Physical Prediction



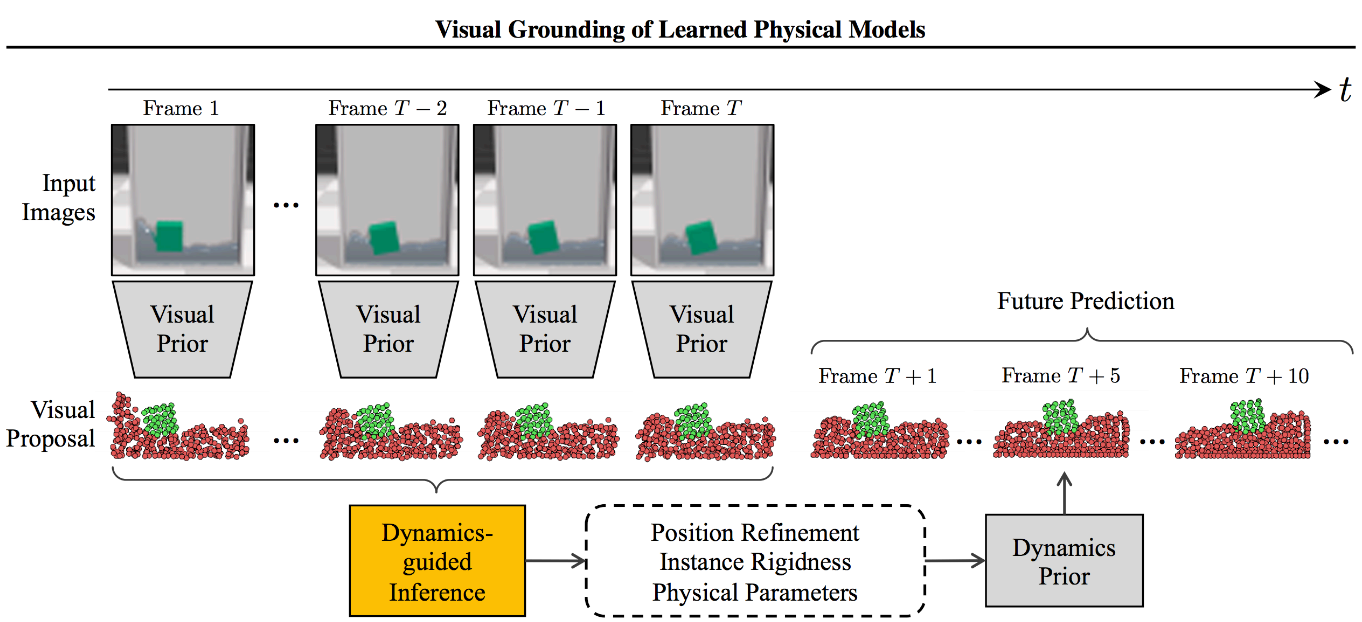
- Humans are good but not perfect



Comparing Oracle Models to Human Physical Prediction



- Humans are good but not perfect
- *Particle*-input models match or exceed human performance — having an explicit physical scene description helps tremendously!

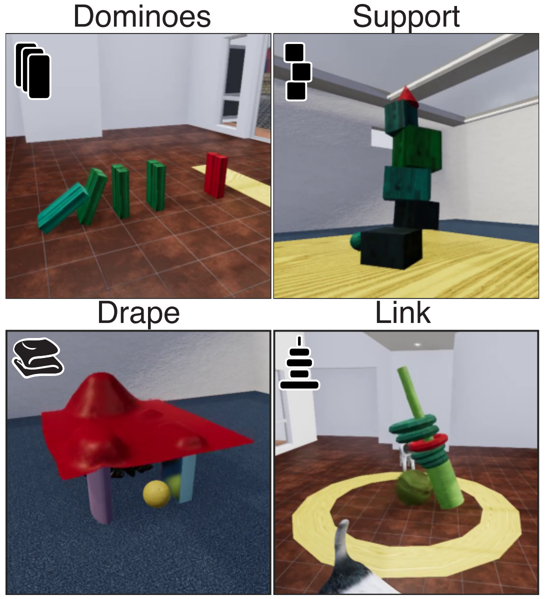


Overall Approach

Inputs

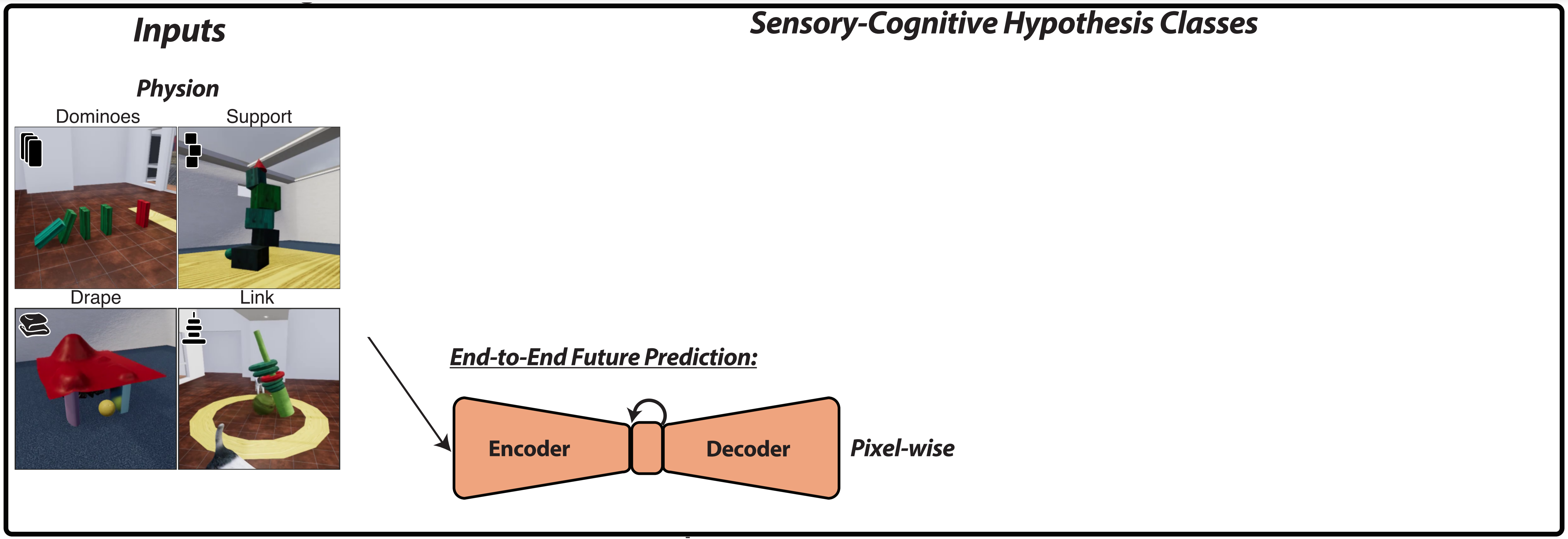
Sensory-Cognitive Hypothesis Classes

Physion



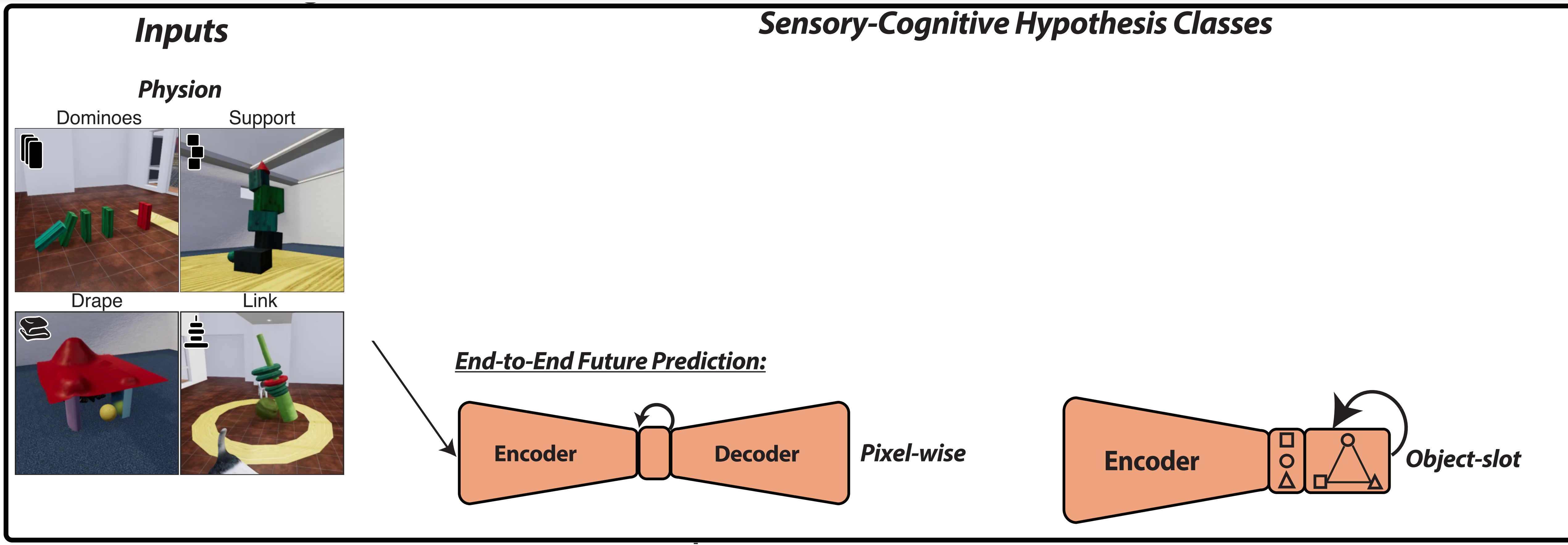
Guiding Question: What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

Overall Approach



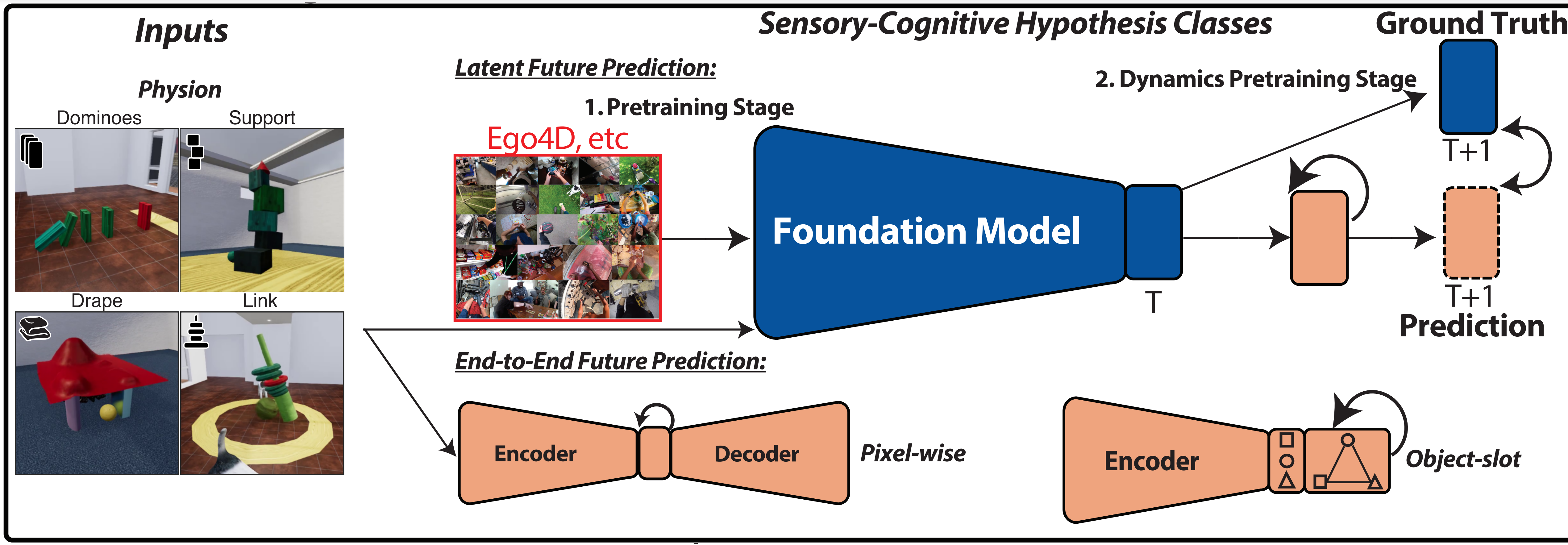
Guiding Question: What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

Overall Approach



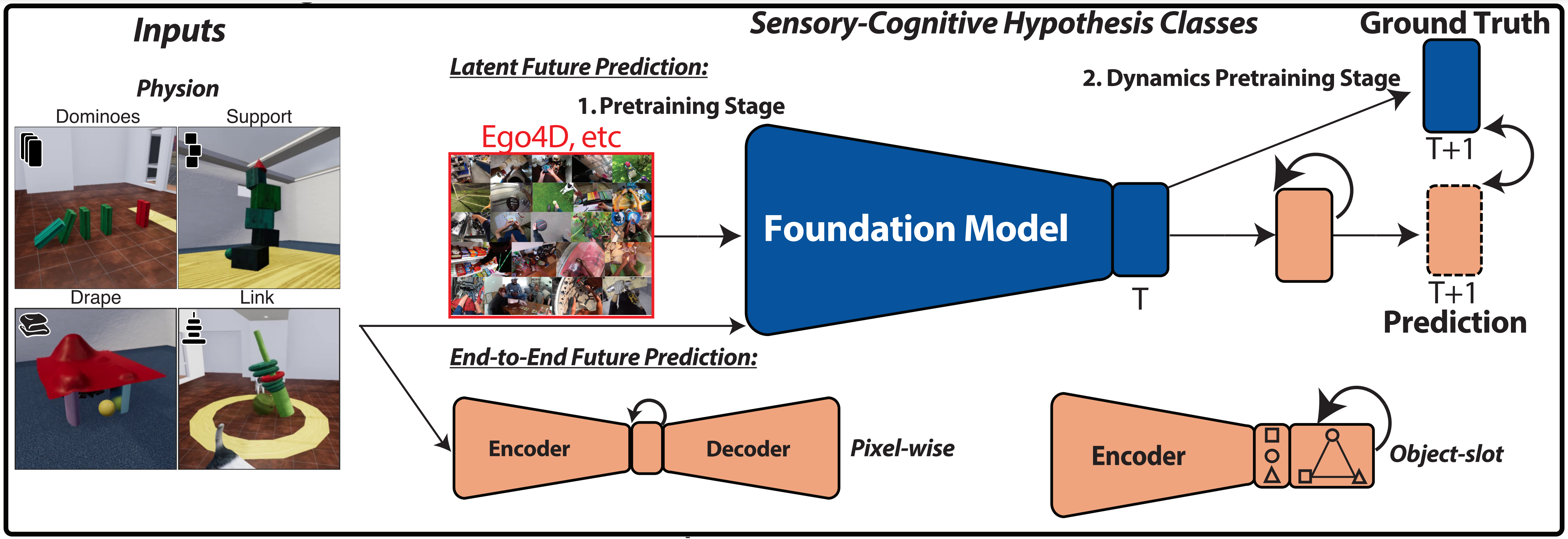
Guiding Question: What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

Overall Approach



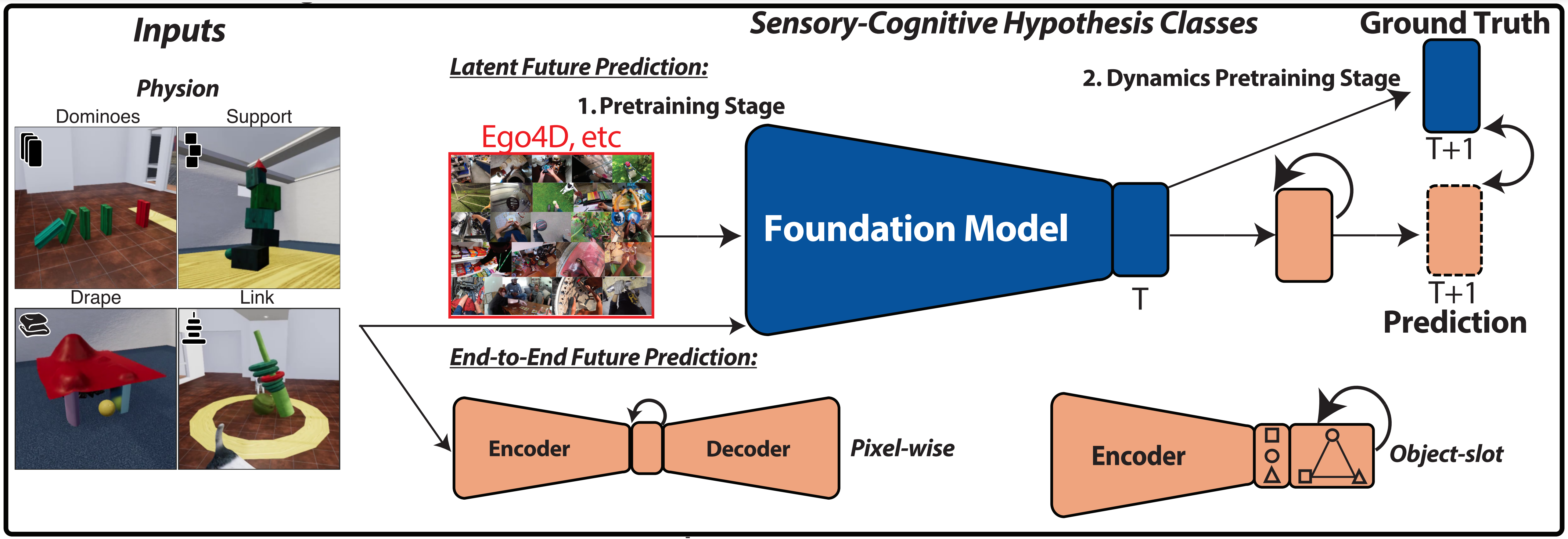
Guiding Question: What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

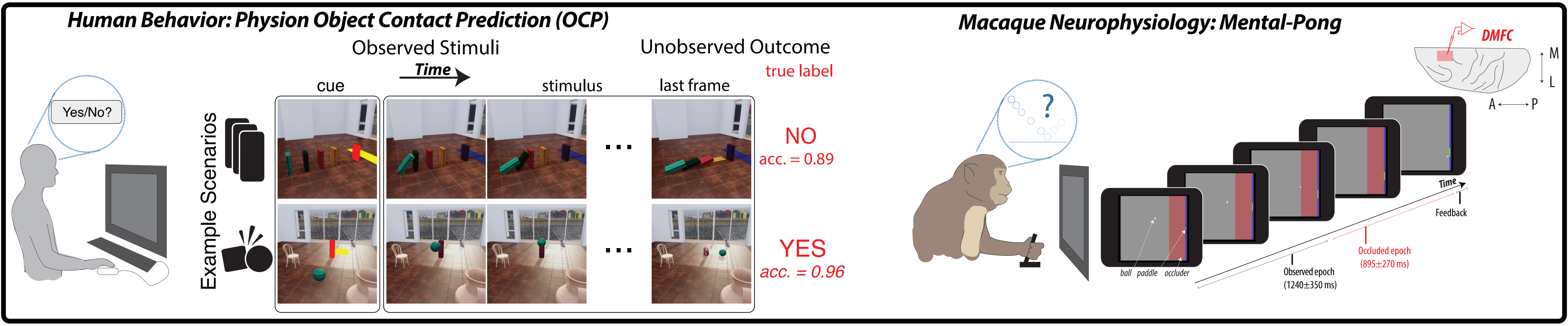
Overall Approach



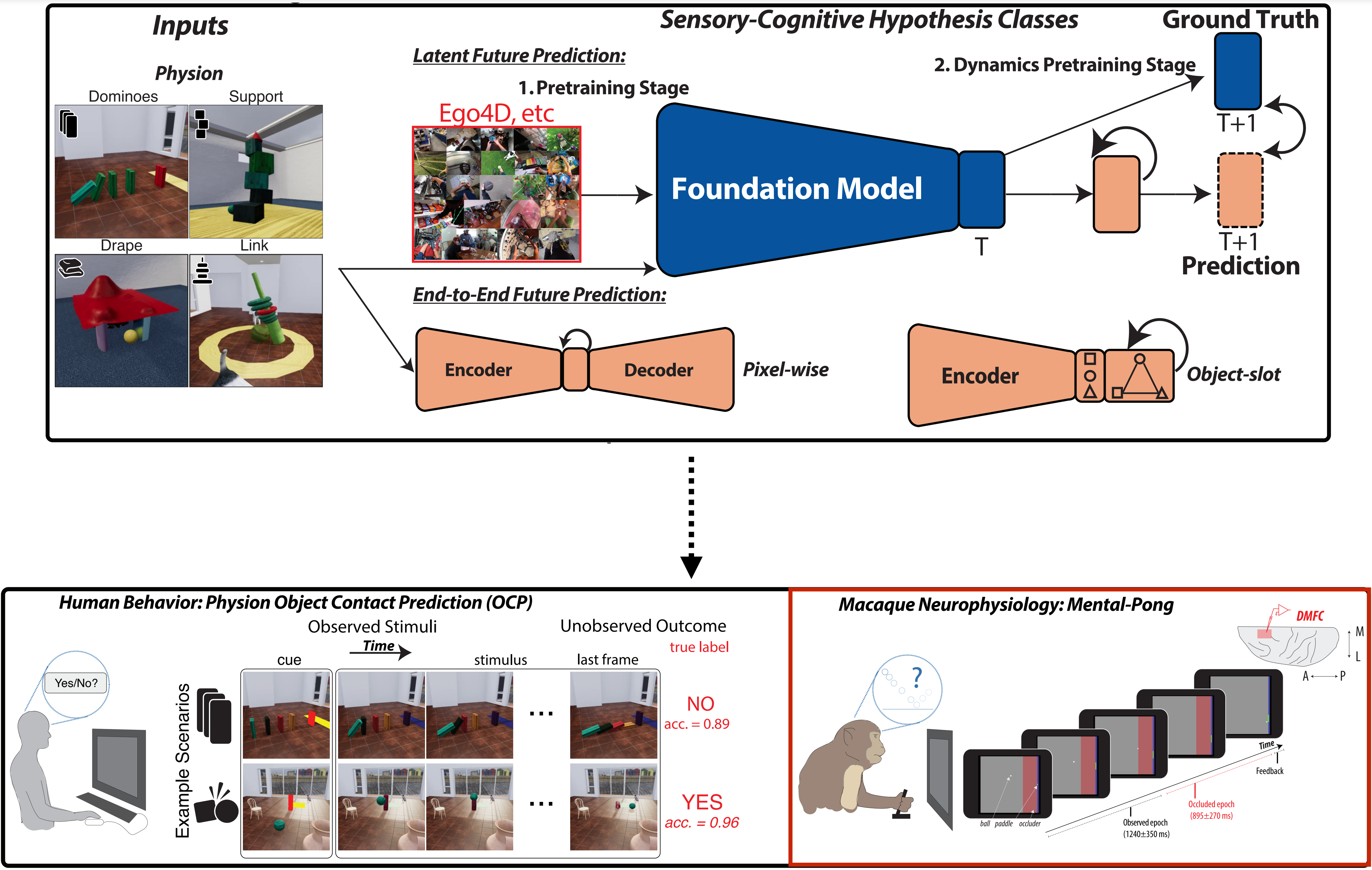
Guiding Question: What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?

Overall Approach

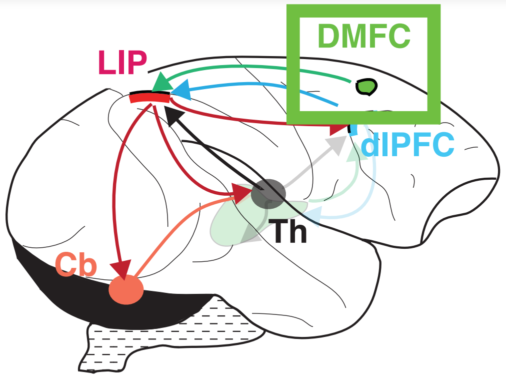




Macaque Neurophysiology: Mental Pong

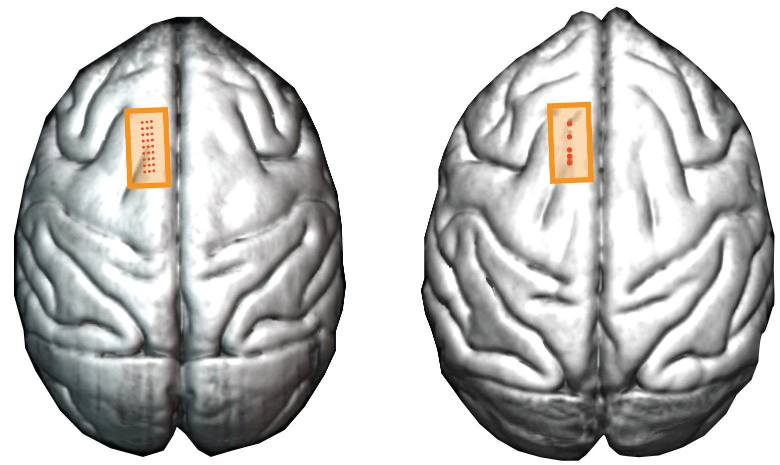


Model Evaluations: Macaque Neurophysiology

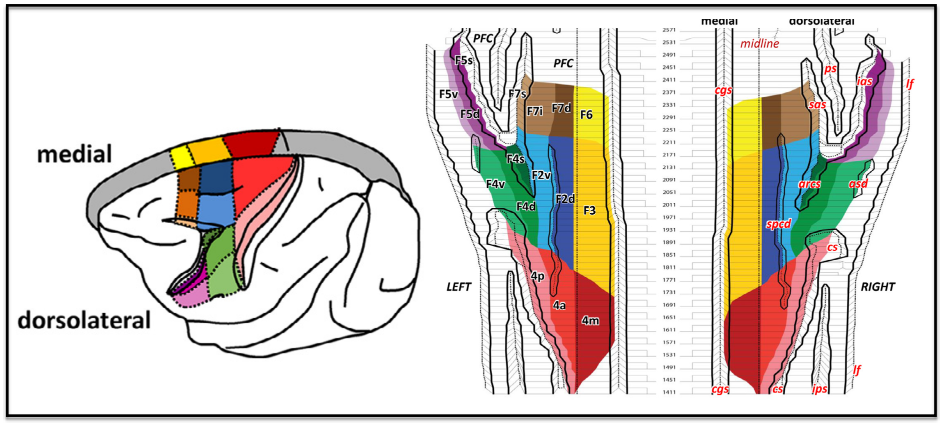


Fronto-Parietal Network

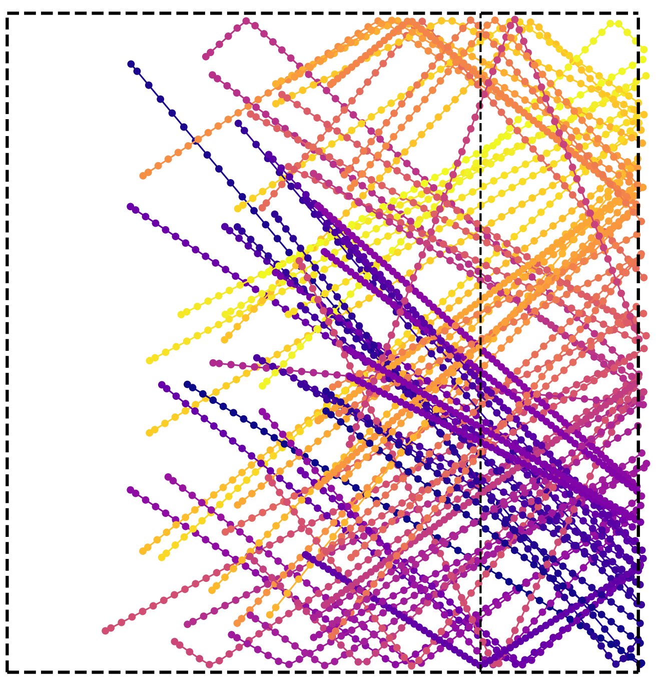
Dorsomedial frontal cortex (DMFC)

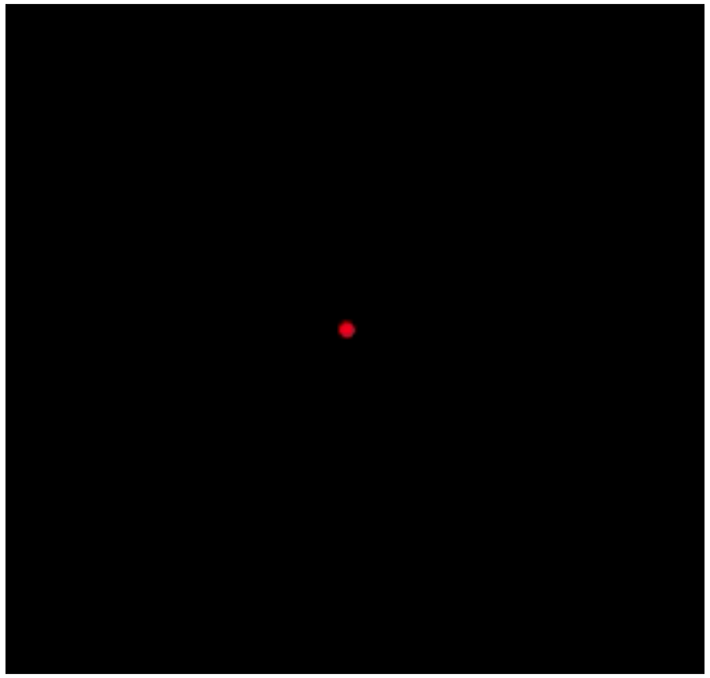


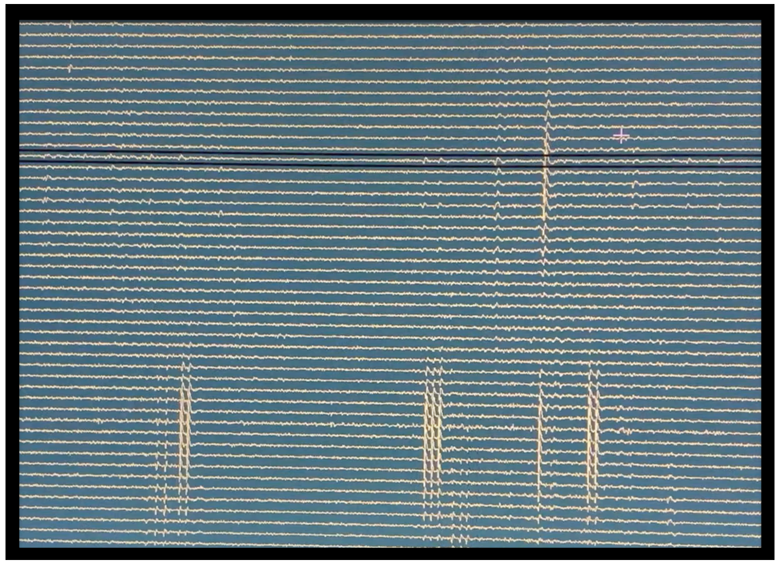
Monkey P *Monkey M*



79 conditions



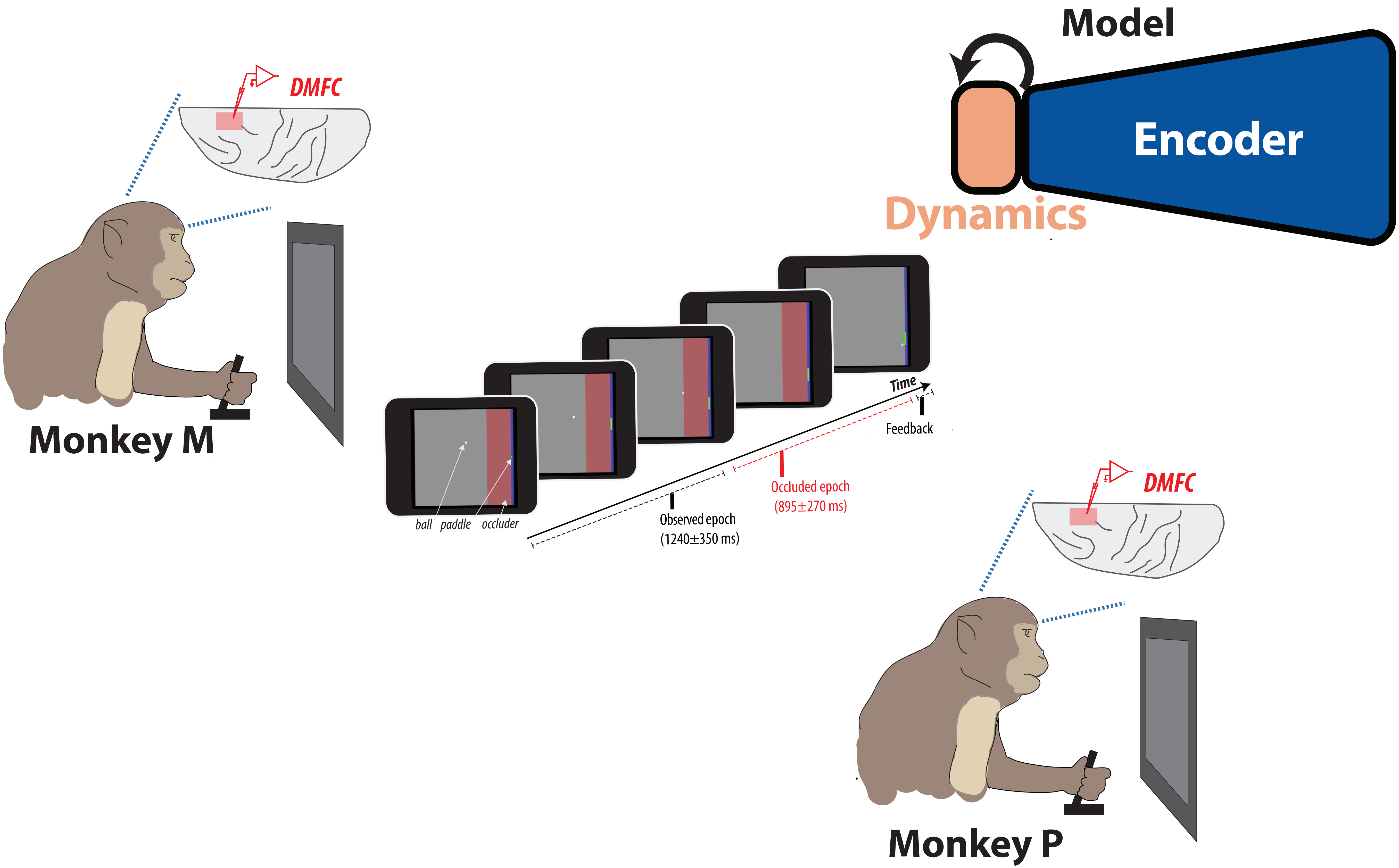




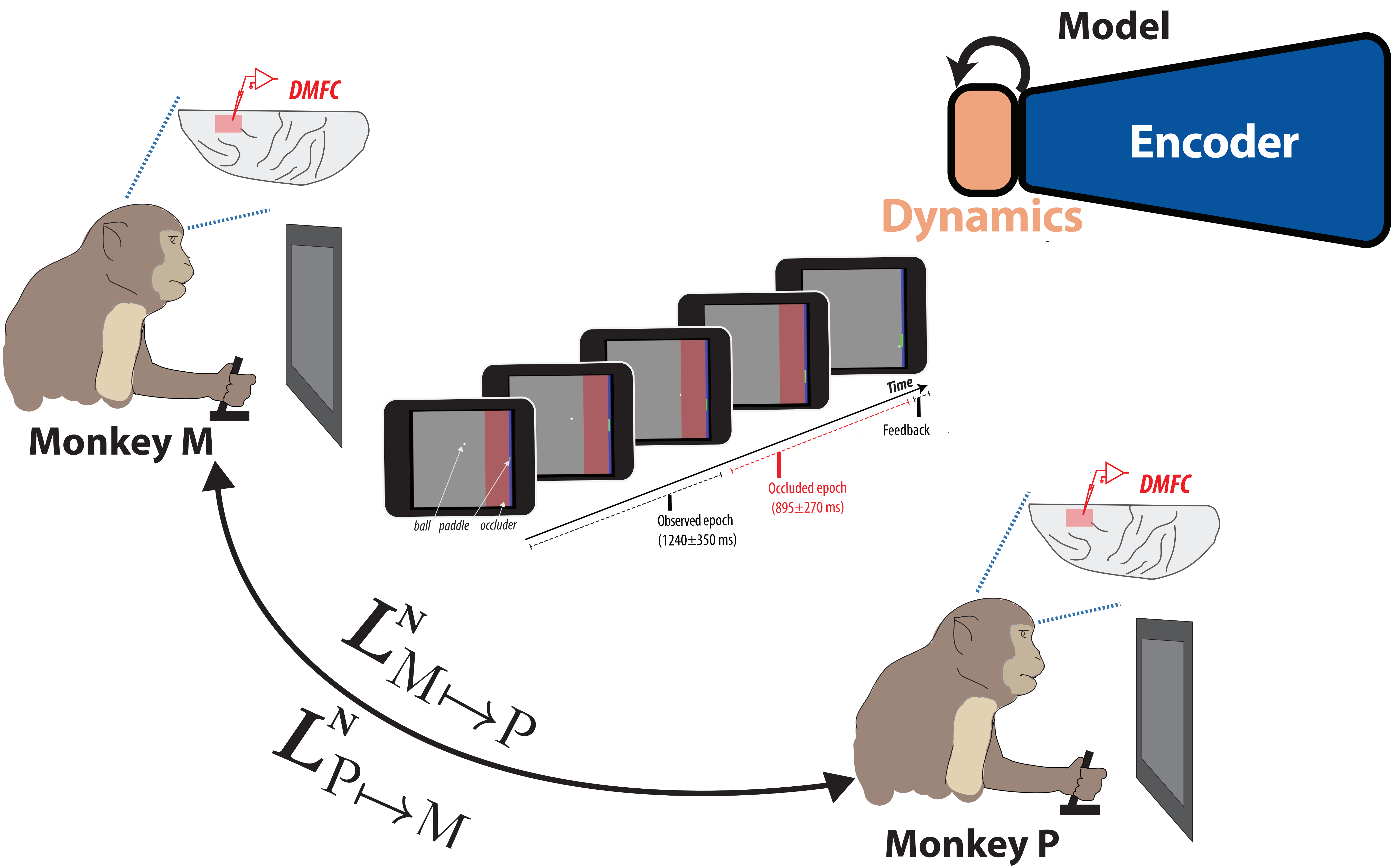
- *Data from two male adult monkeys*
- *79 subsampled M-Pong conditions*
- *64 channel v-probe (monkey P) and 384-channel Neuropixel probe (monkey M)*
- *Total of 1889 stable & reliable neurons recorded from DMFC*

Rishi Rajalingham

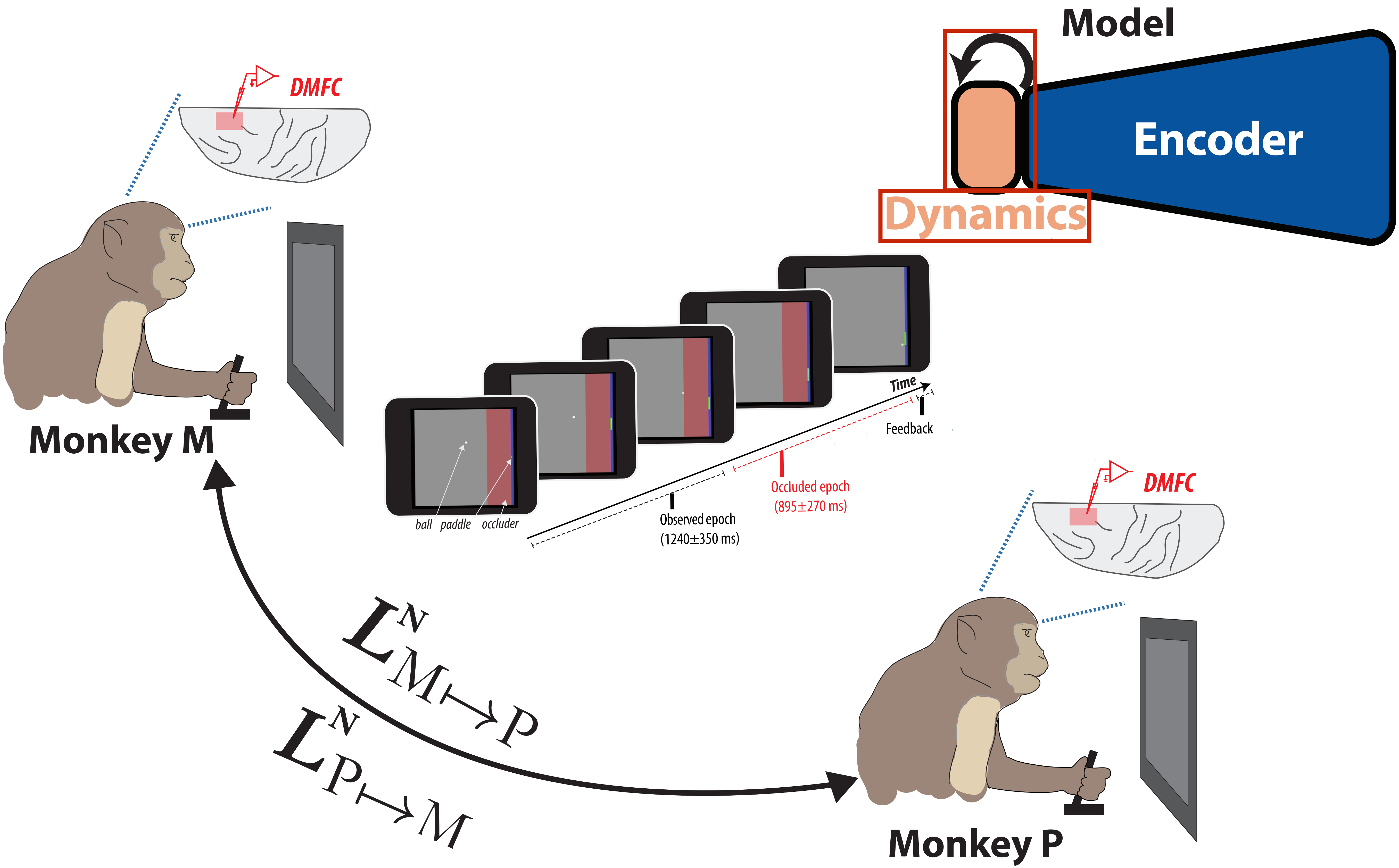
Assessing Model Similarity: Neural Response Predictivity



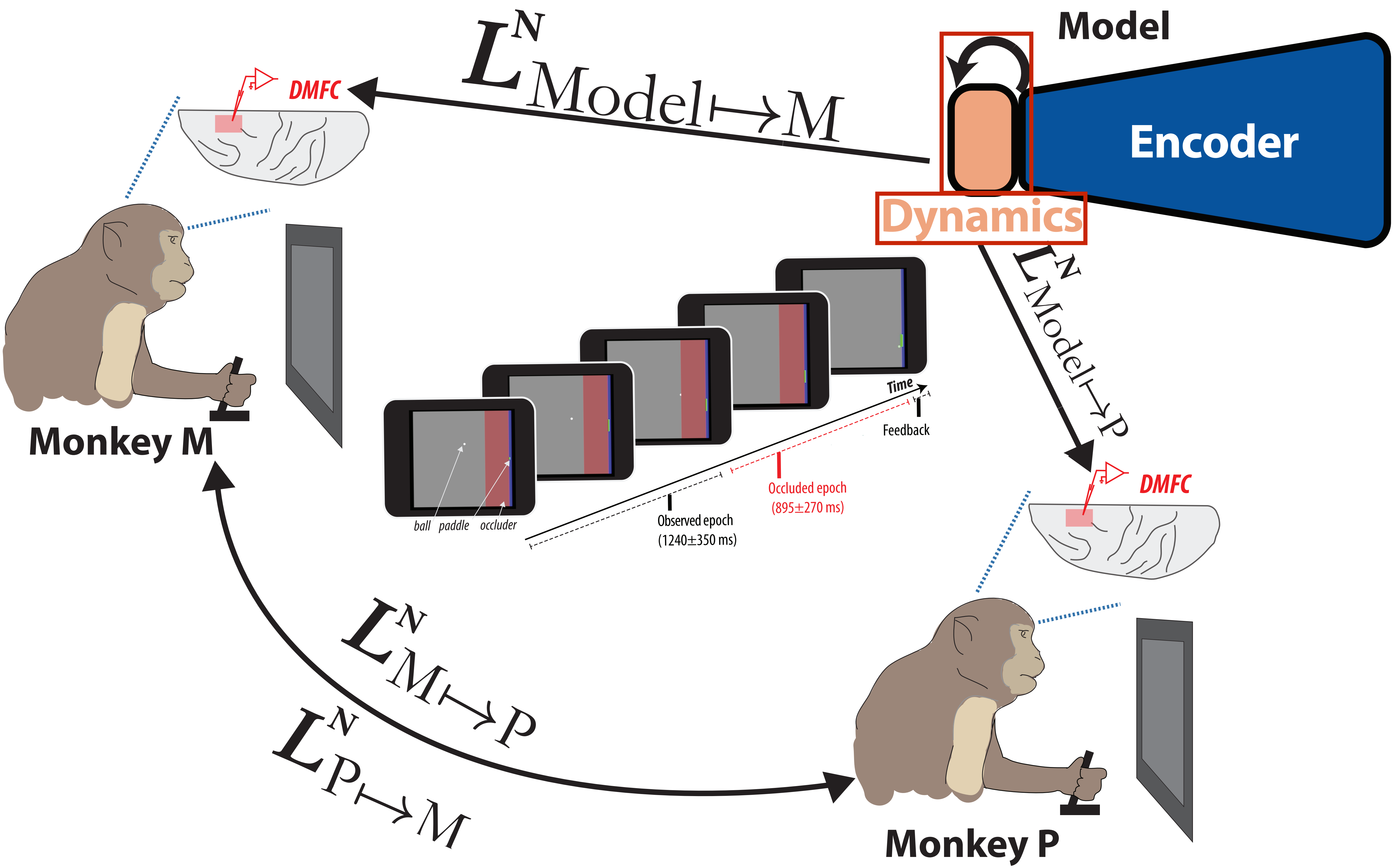
Assessing Model Similarity: Neural Response Predictivity



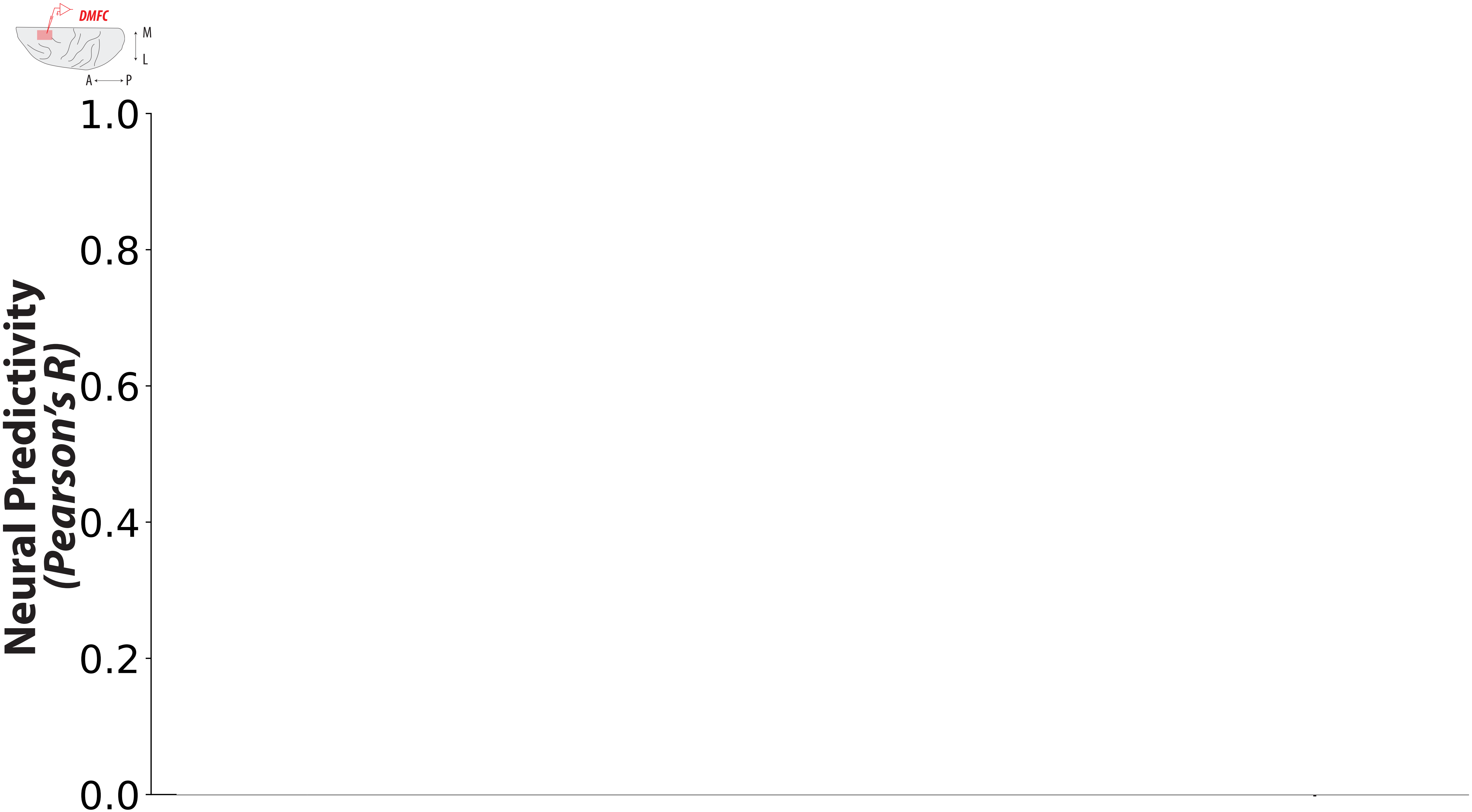
Assessing Model Similarity: Neural Response Predictivity



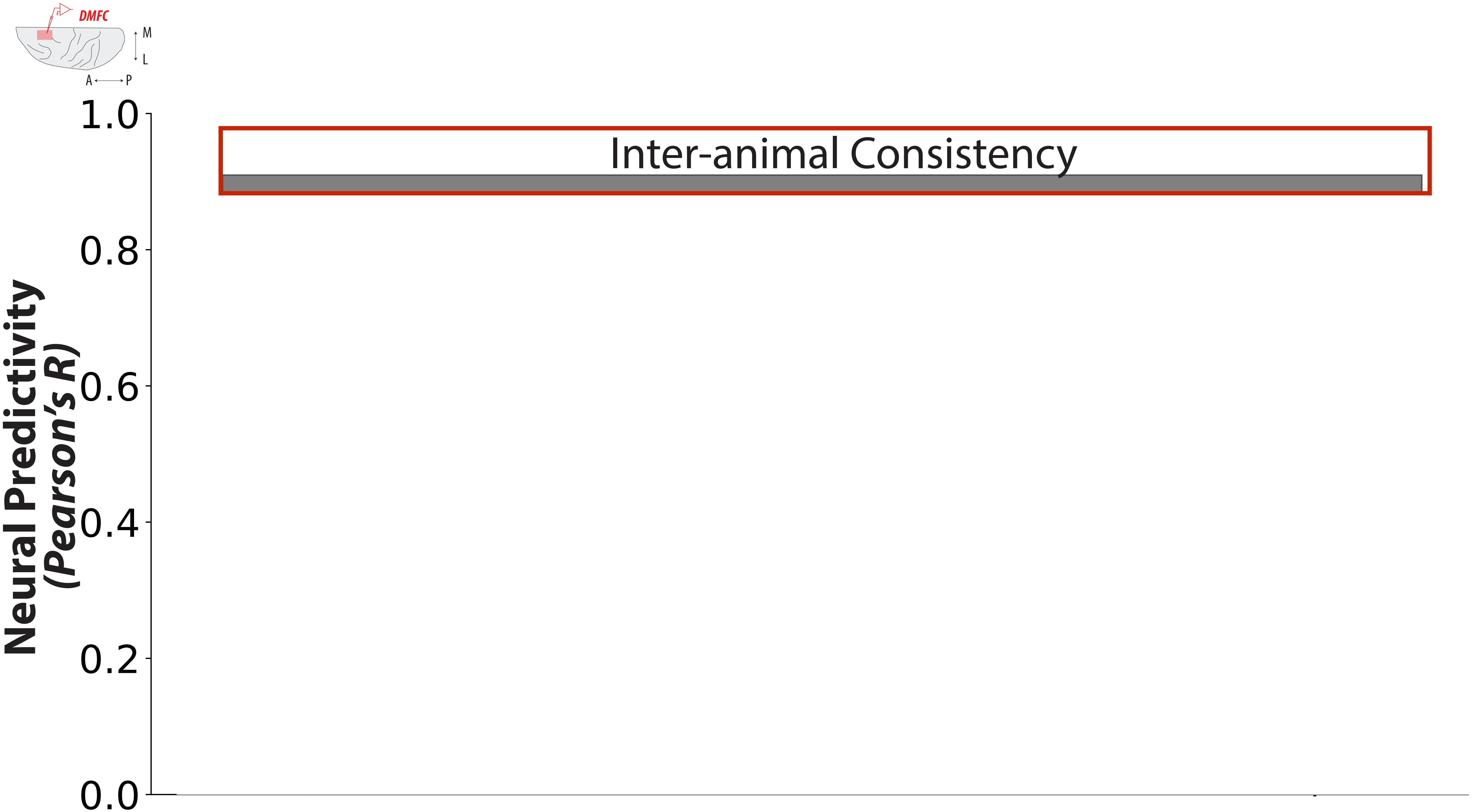
Assessing Model Similarity: Neural Response Predictivity



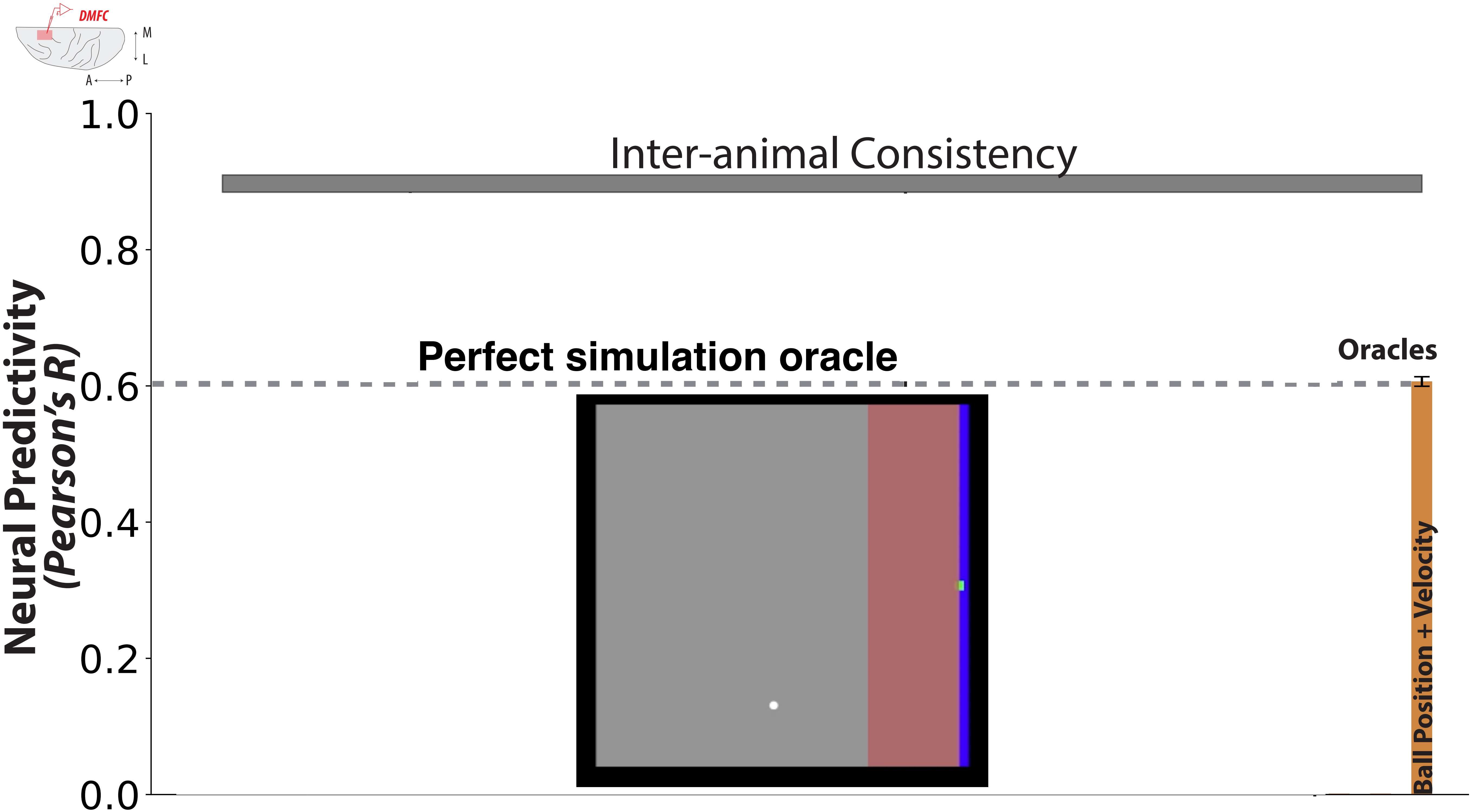
Macaque Neurophysiology: Mental Pong



Macaque Neurophysiology: Mental Pong



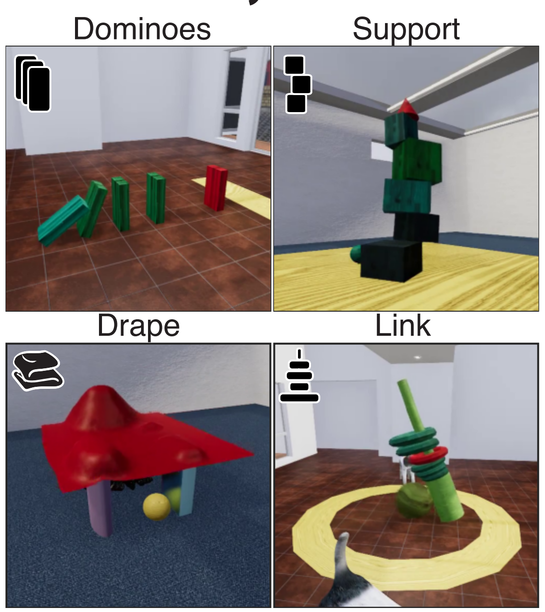
Physical Simulation Oracles Predict Neural Data Well



Functional Constraint Hypotheses

Inputs

Physion



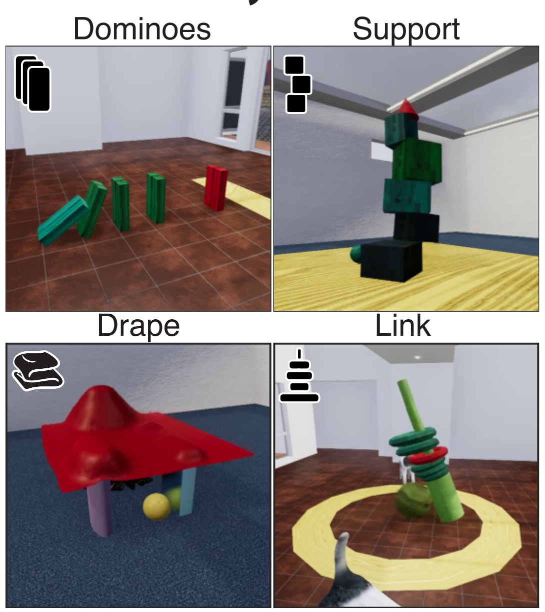
Sensory-Cognitive Hypothesis Classes

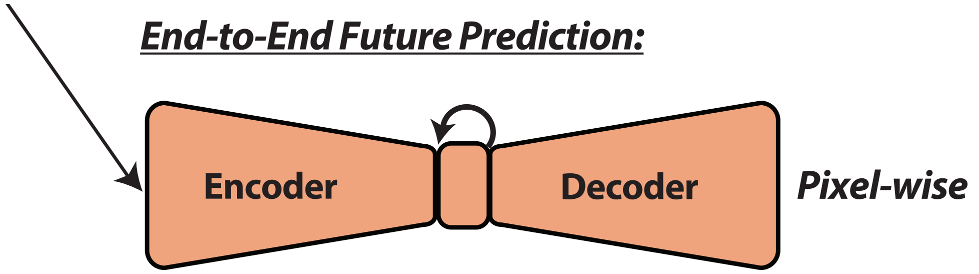
Hypothesis Class 1: Pixel-wise Future Prediction

Inputs

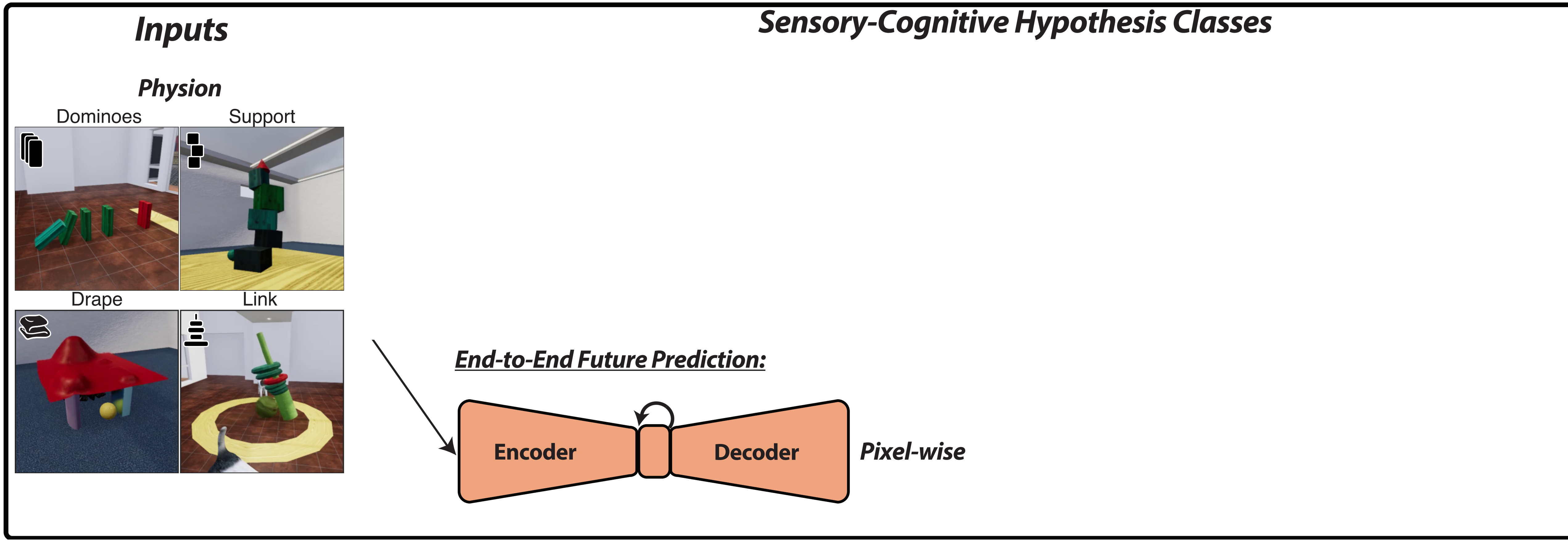
Sensory-Cognitive Hypothesis Classes

Physion

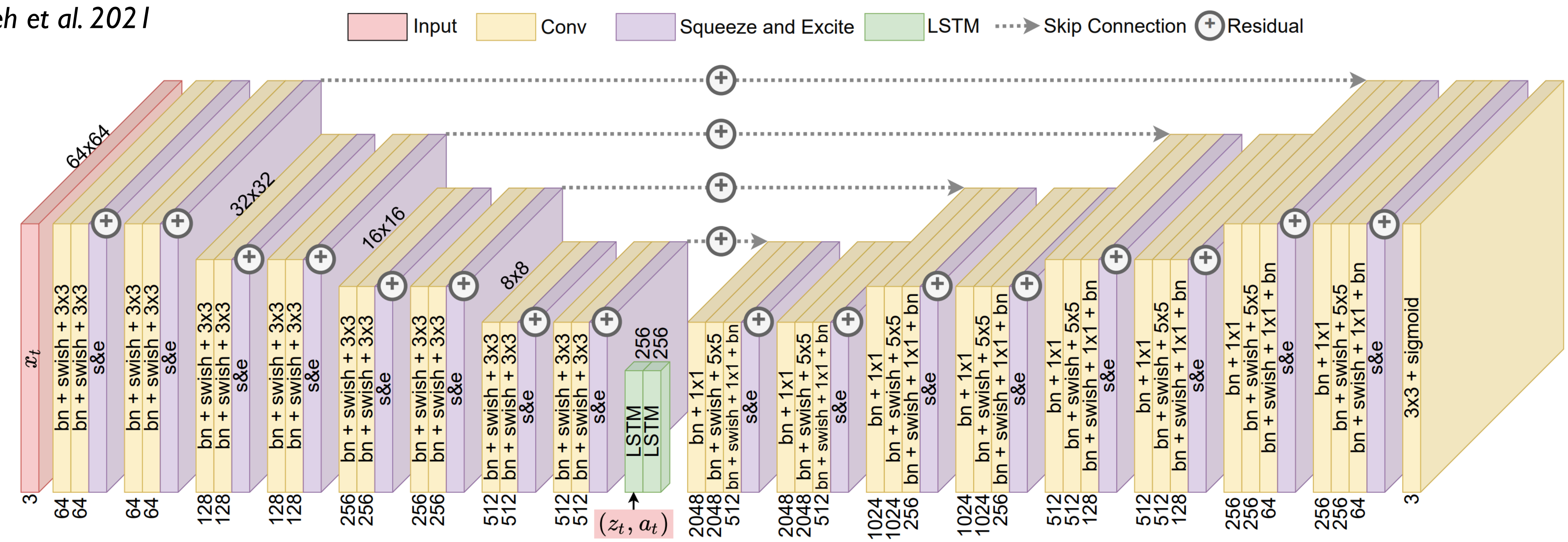




Hypothesis Class I: Pixel-wise Future Prediction



Babaeizadeh et al. 2021

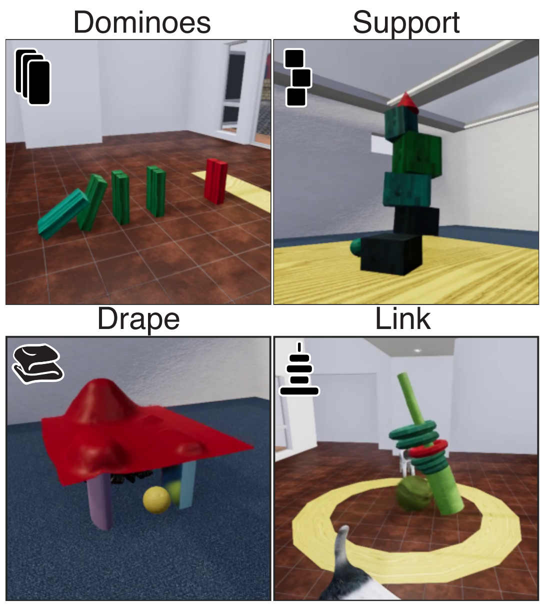


Hypothesis Class I: Pixel-wise Future Prediction

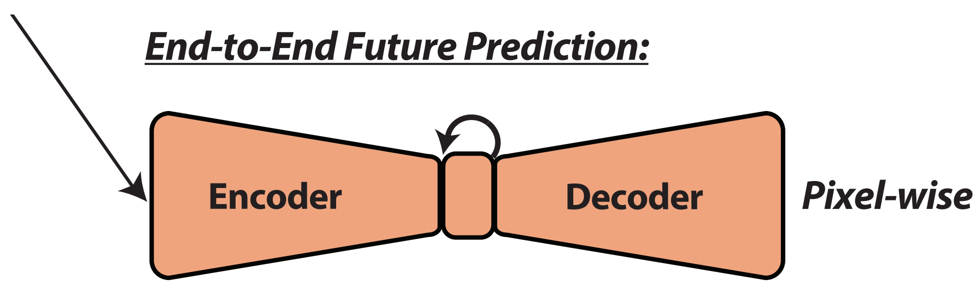
Inputs

Sensory-Cognitive Hypothesis Classes

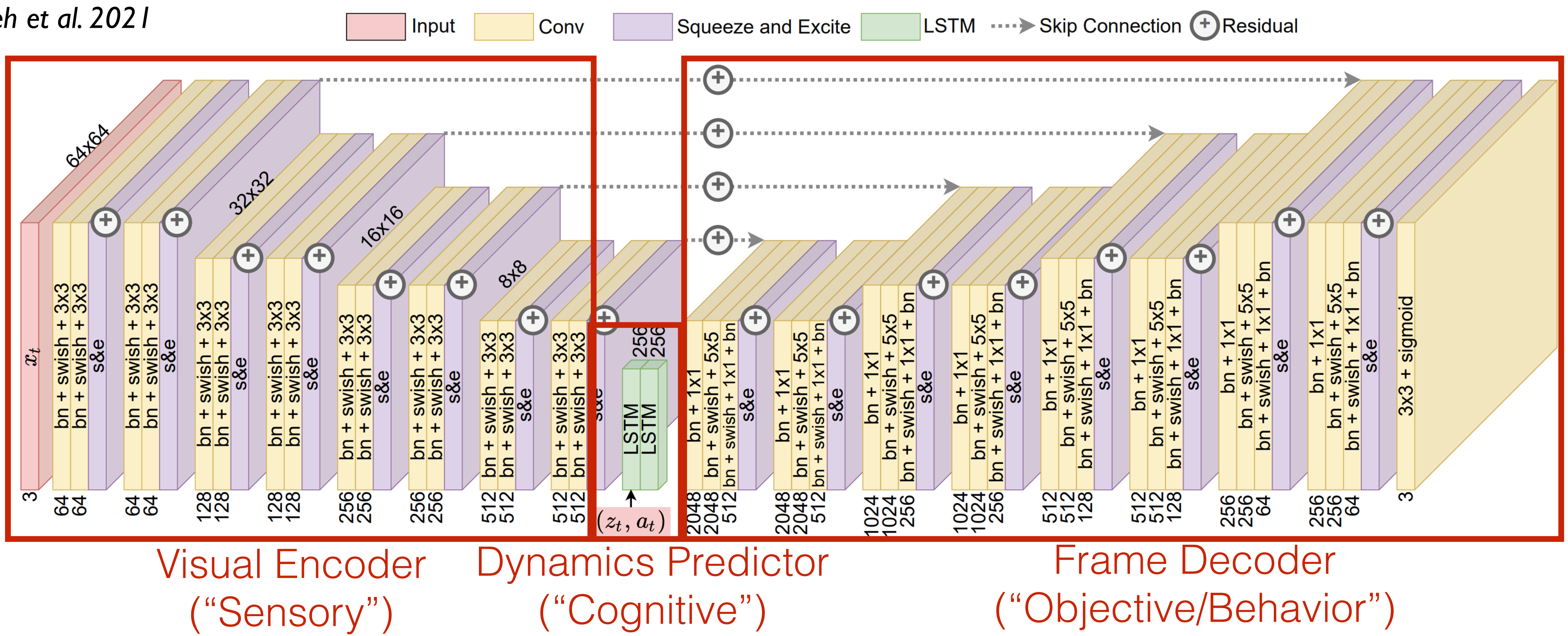
Physion



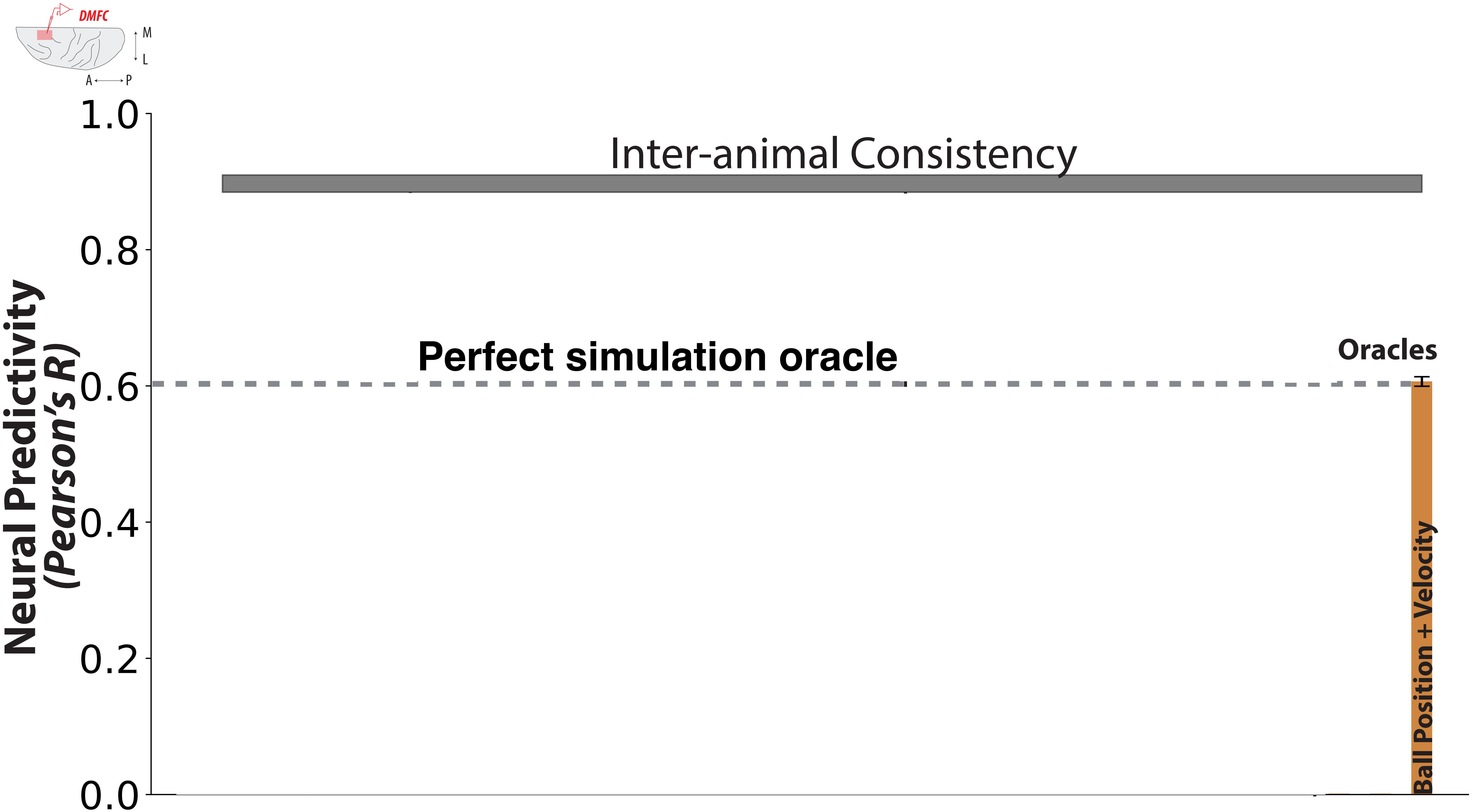
Predicts the future at the *pixel* resolution of the sensory input **(very detailed)**



Babaeizadeh et al. 2021

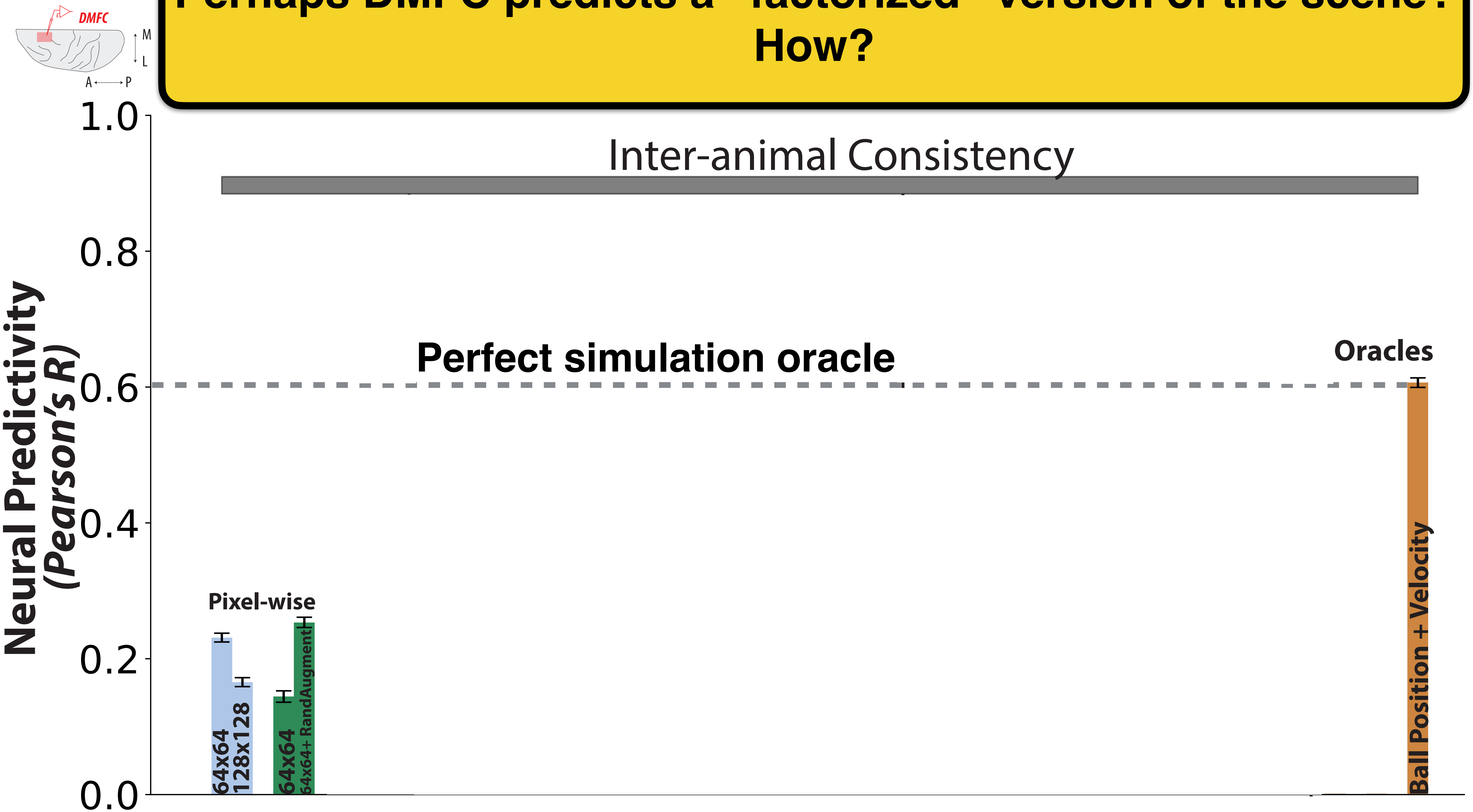


Physical Simulation Oracles Predict Neural Data Well

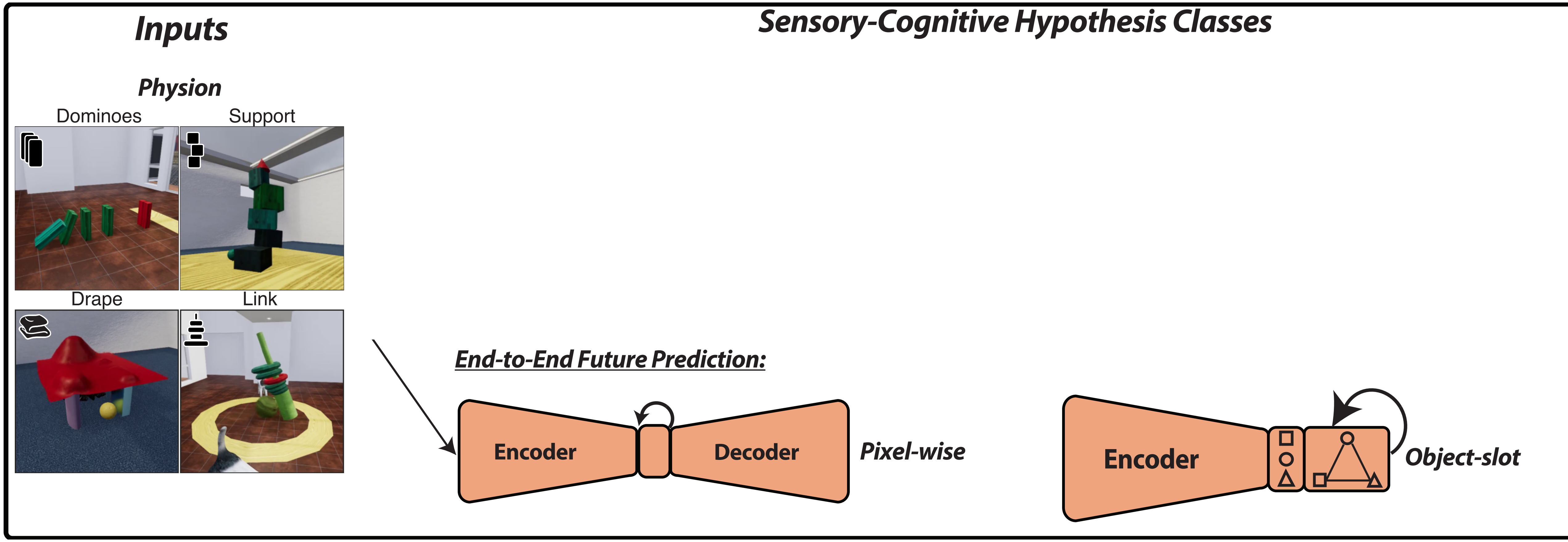


Pixel-wise Future Prediction Poorly Predicts Neurons

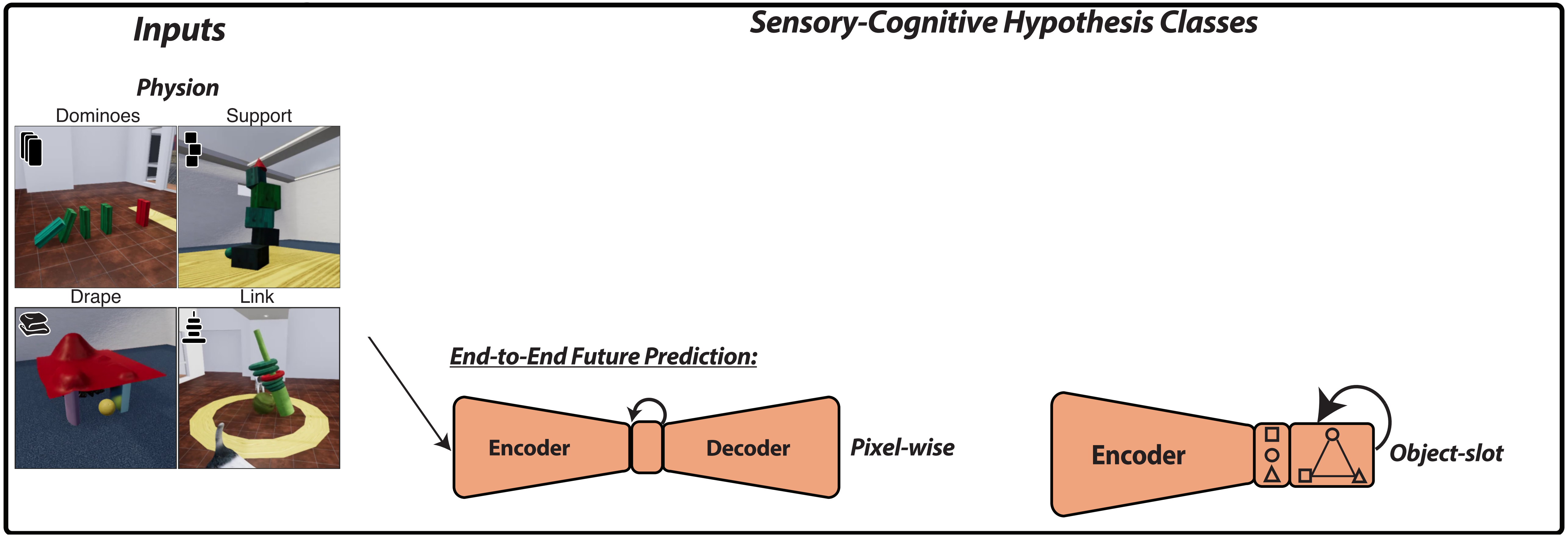
Perhaps DMFC predicts a "factorized" version of the scene? How?



Hypothesis Class 2: Object Slots

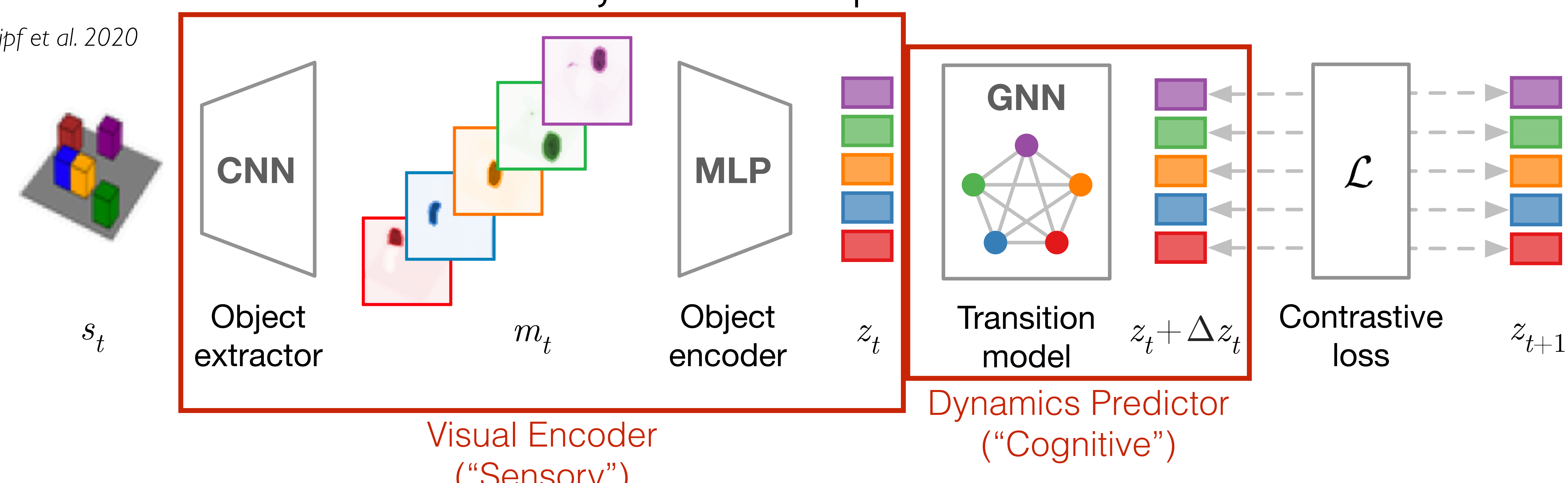


Hypothesis Class 2: Object Slots

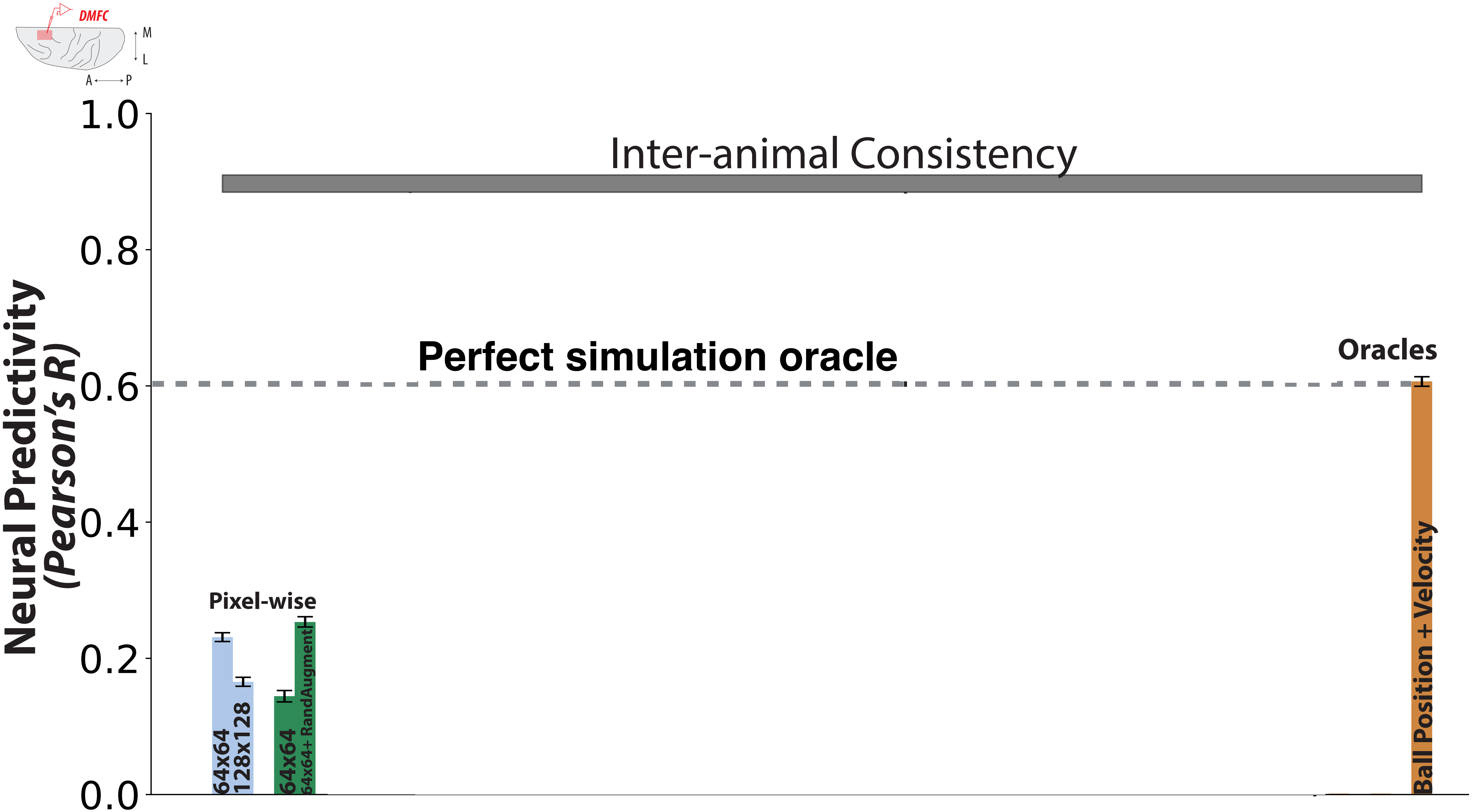


Predicts at the level of object slot representations and their relations

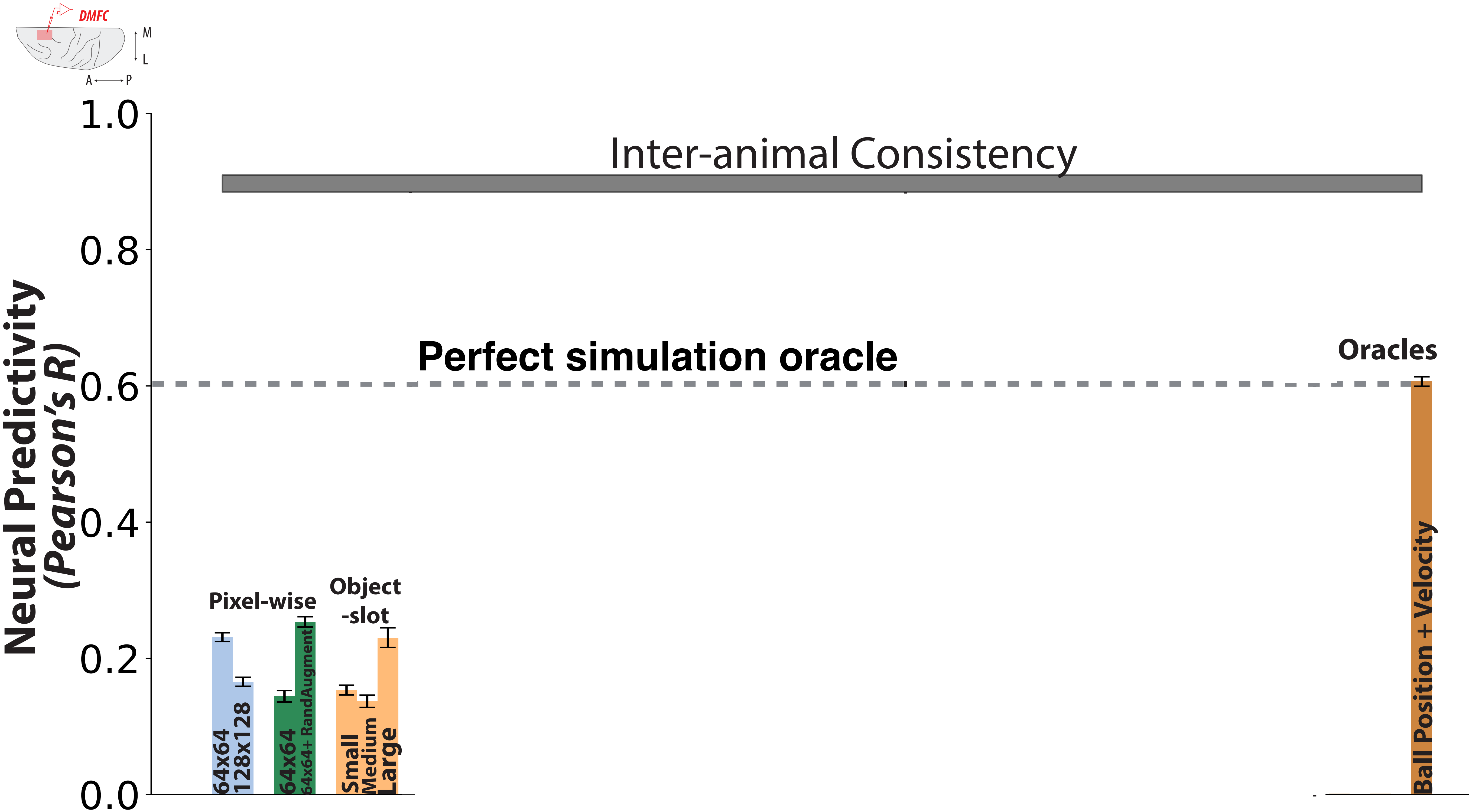
Kipf et al. 2020



Pixel-wise Future Prediction Poorly Predicts Neurons

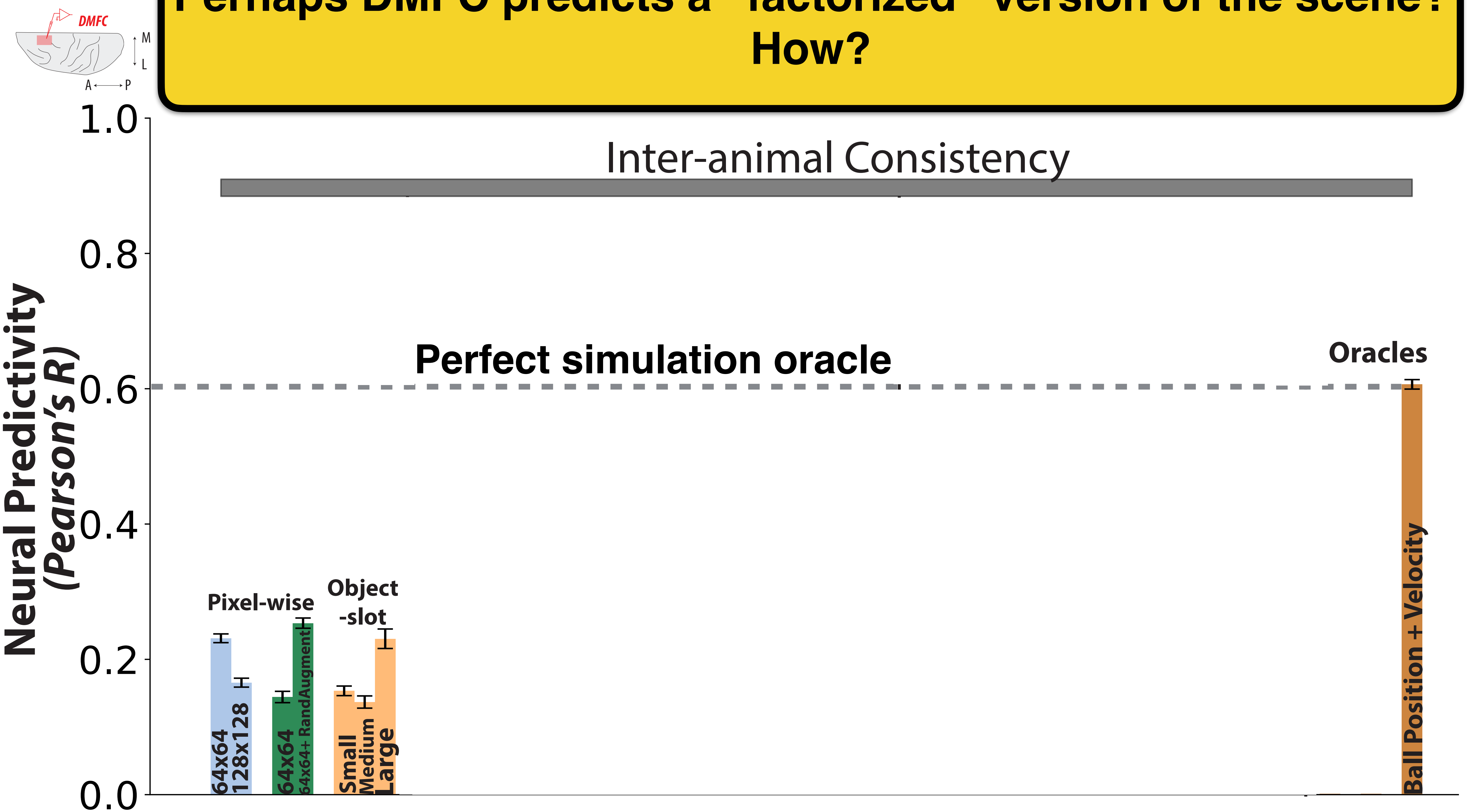


Object-Slot Future Prediction Poorly Predicts Neurons



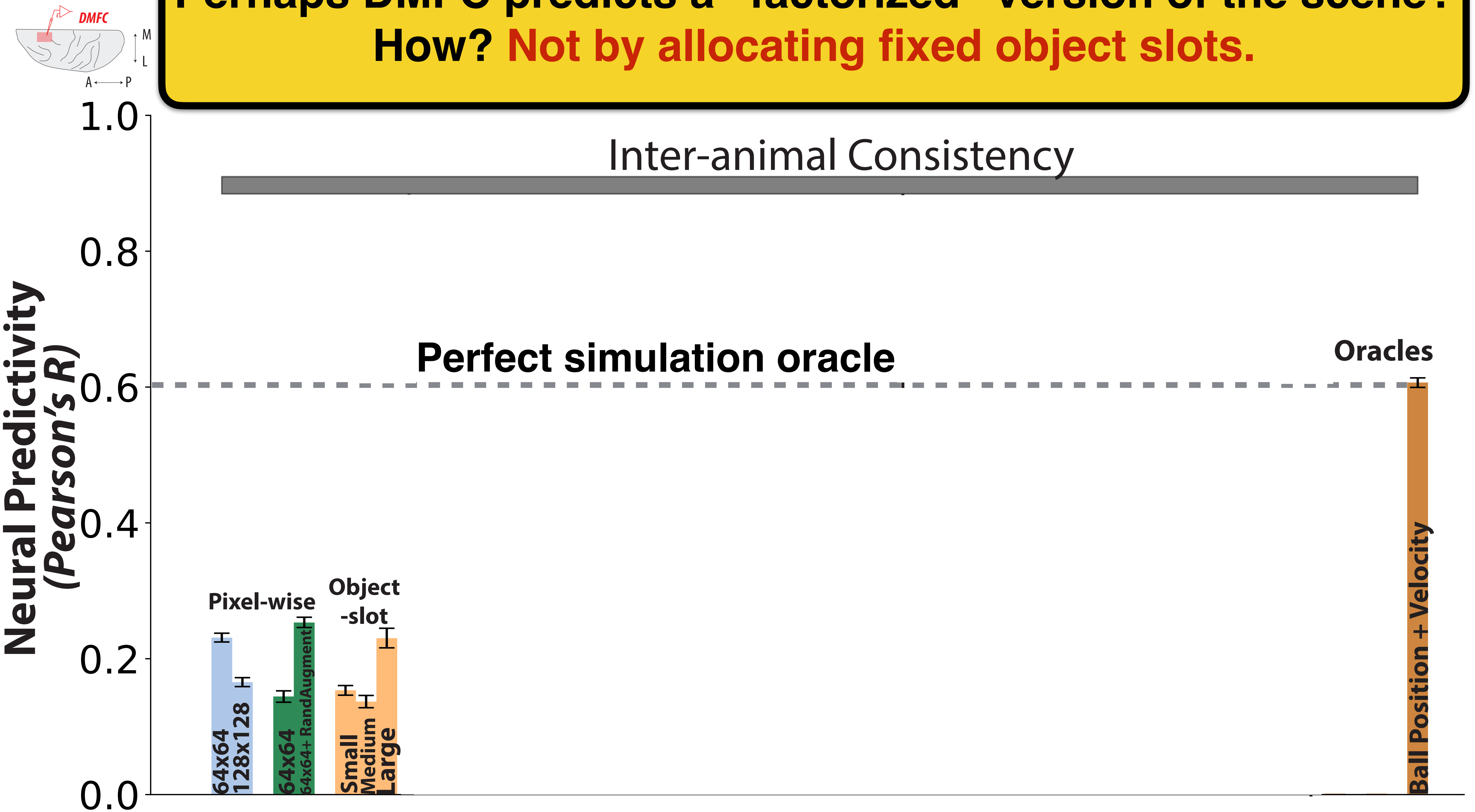
Object-Slot Future Prediction Poorly Predicts Neurons

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

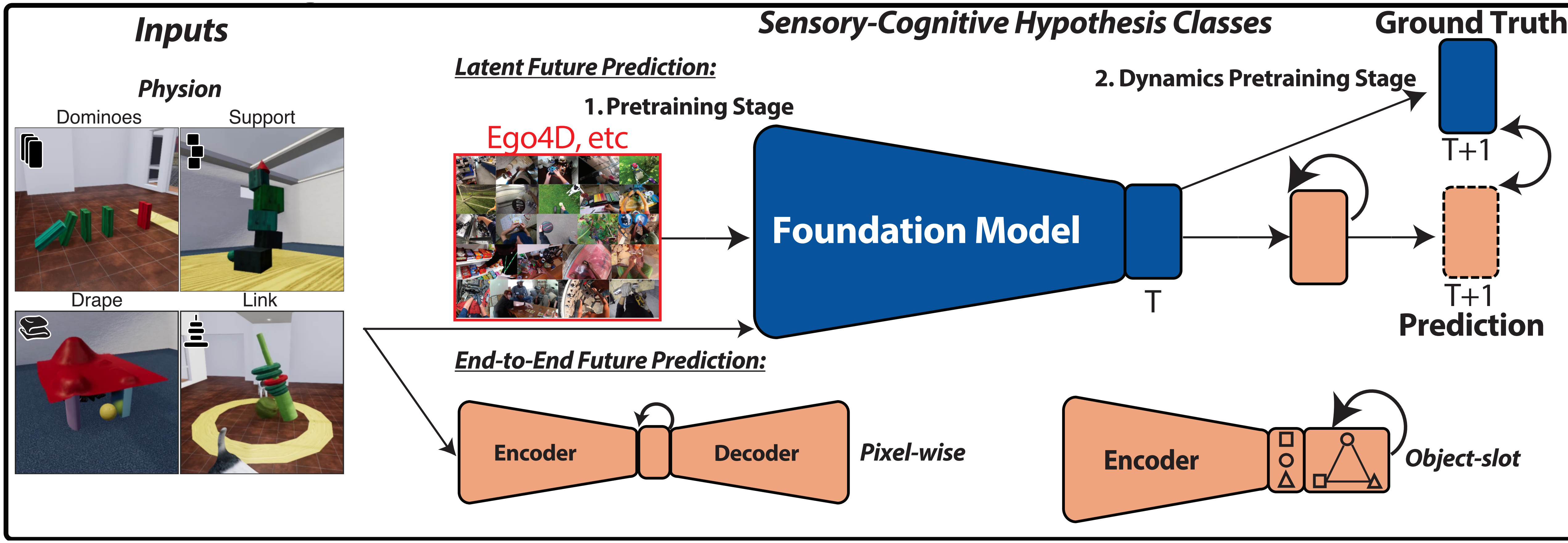


Object-Slot Future Prediction Poorly Predicts Neurons

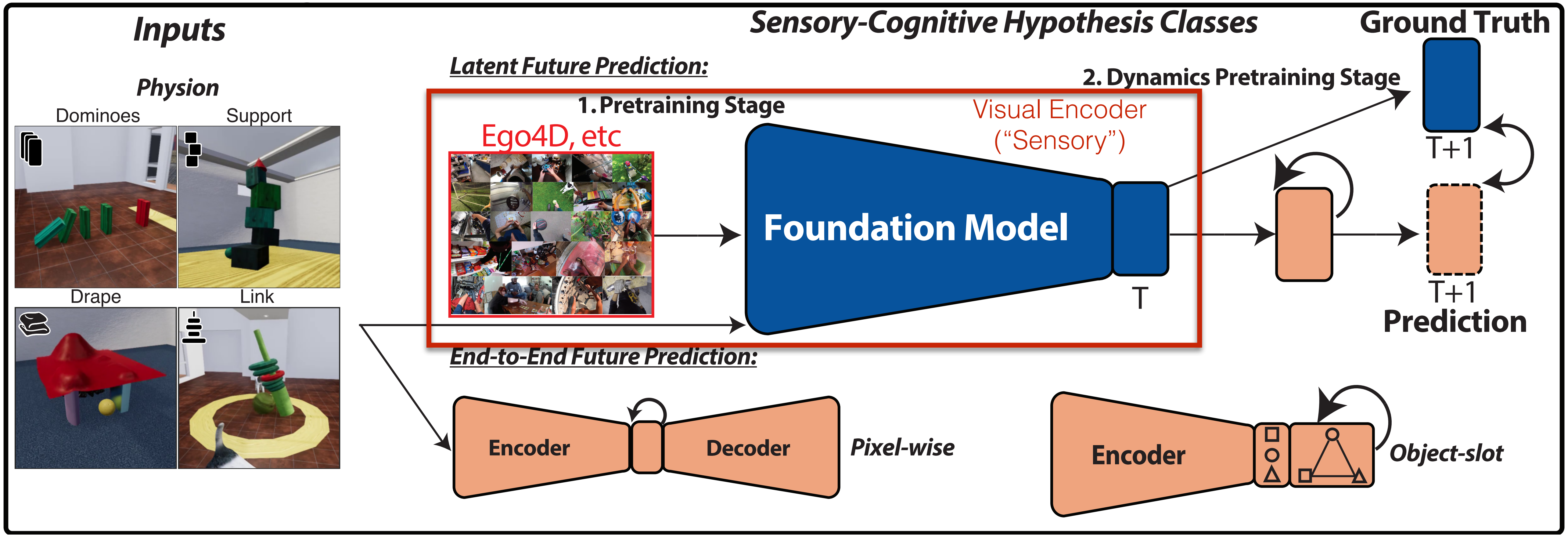
Perhaps DMFC predicts a "factorized" version of the scene?
How? Not by allocating fixed object slots.



Hypothesis Class 3: Latent Future Prediction

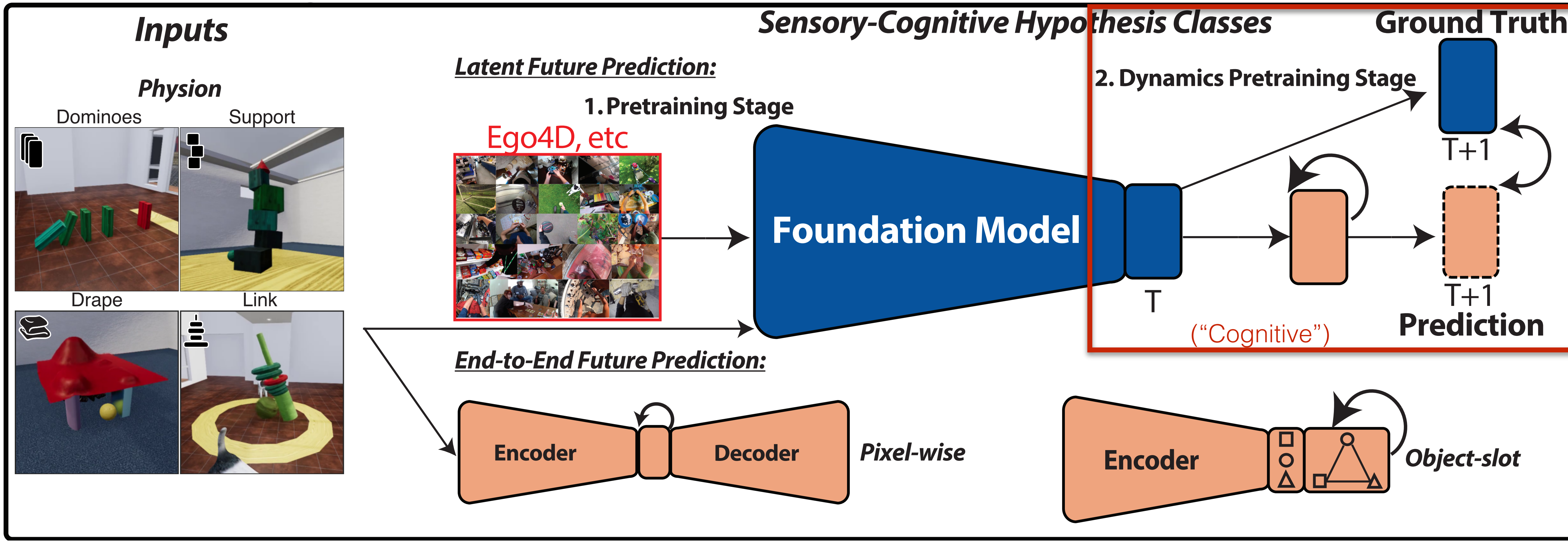


Hypothesis Class 3: Latent Future Prediction



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

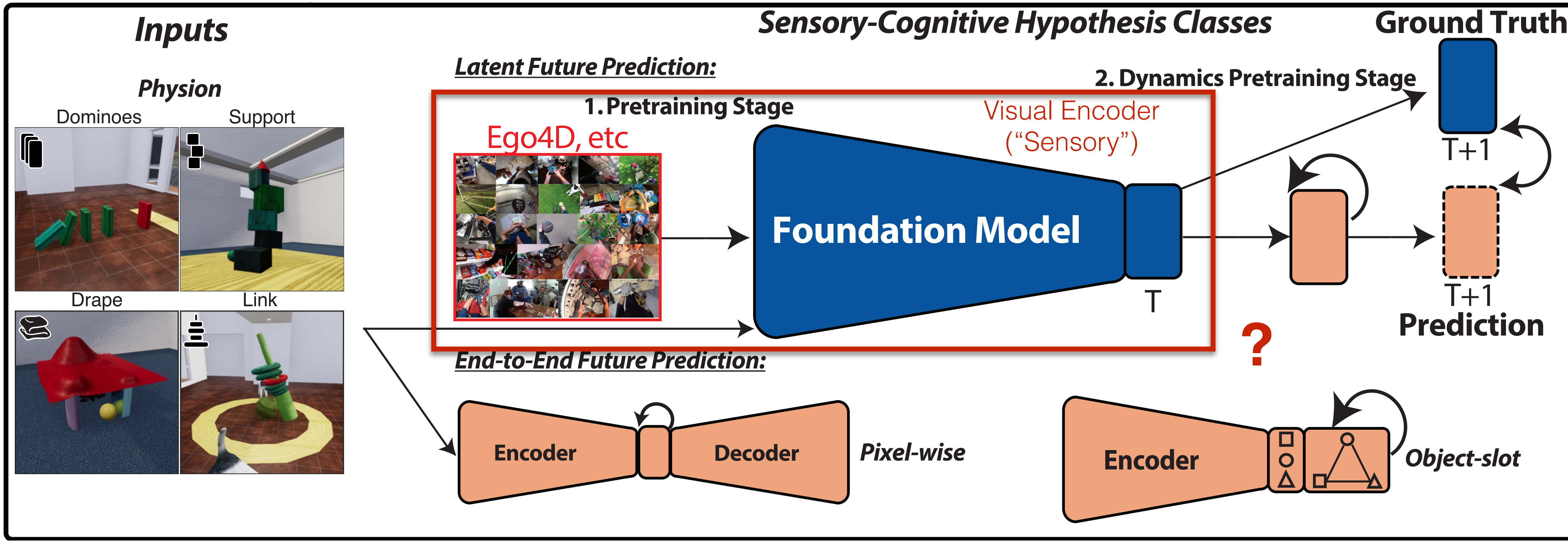
Hypothesis Class 3: Latent Future Prediction



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

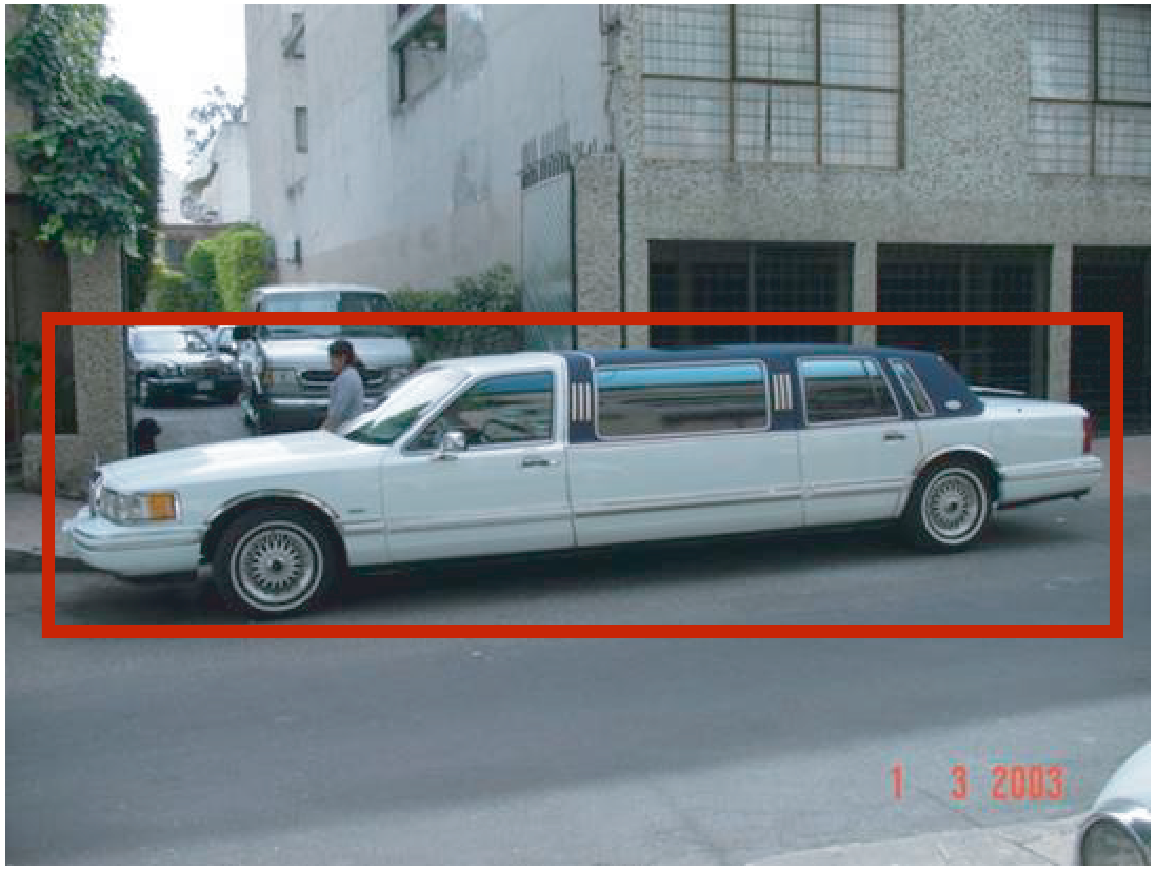


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

What vision task?

Leverage these dynamics to do explicit future prediction

Hypothesis Class 3: Image Foundation Models



Object-Slot Future Prediction Poorly Predicts Neurons

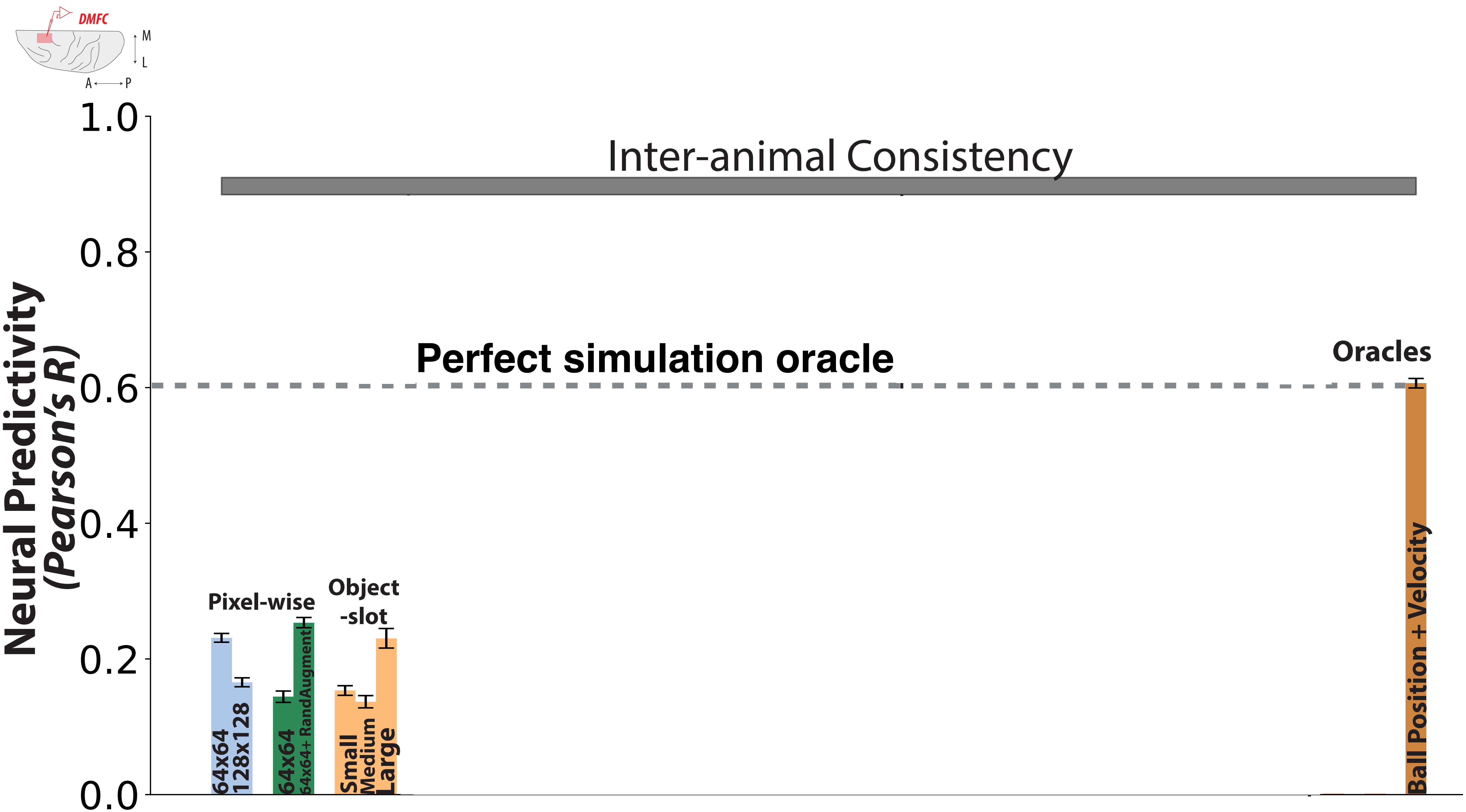
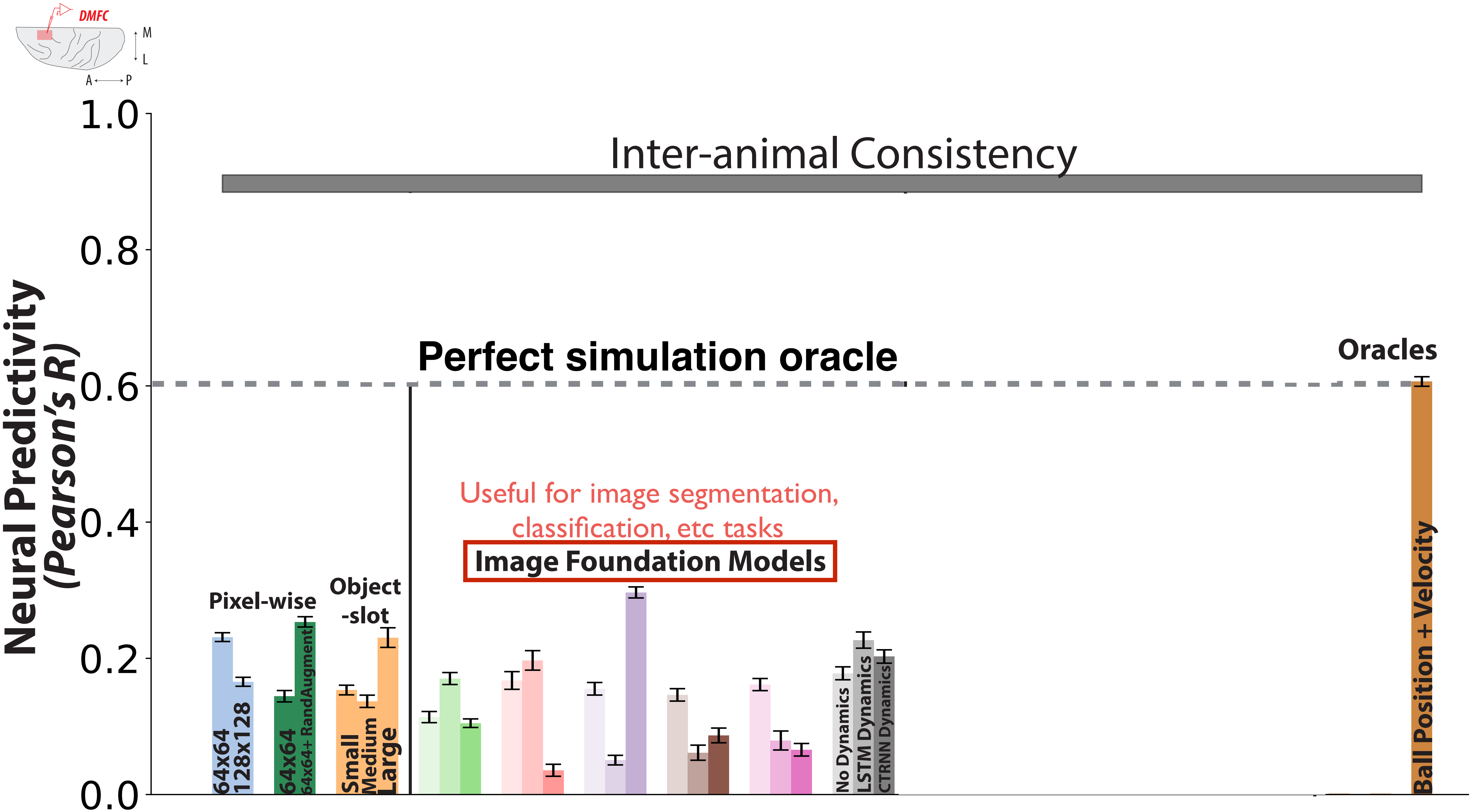
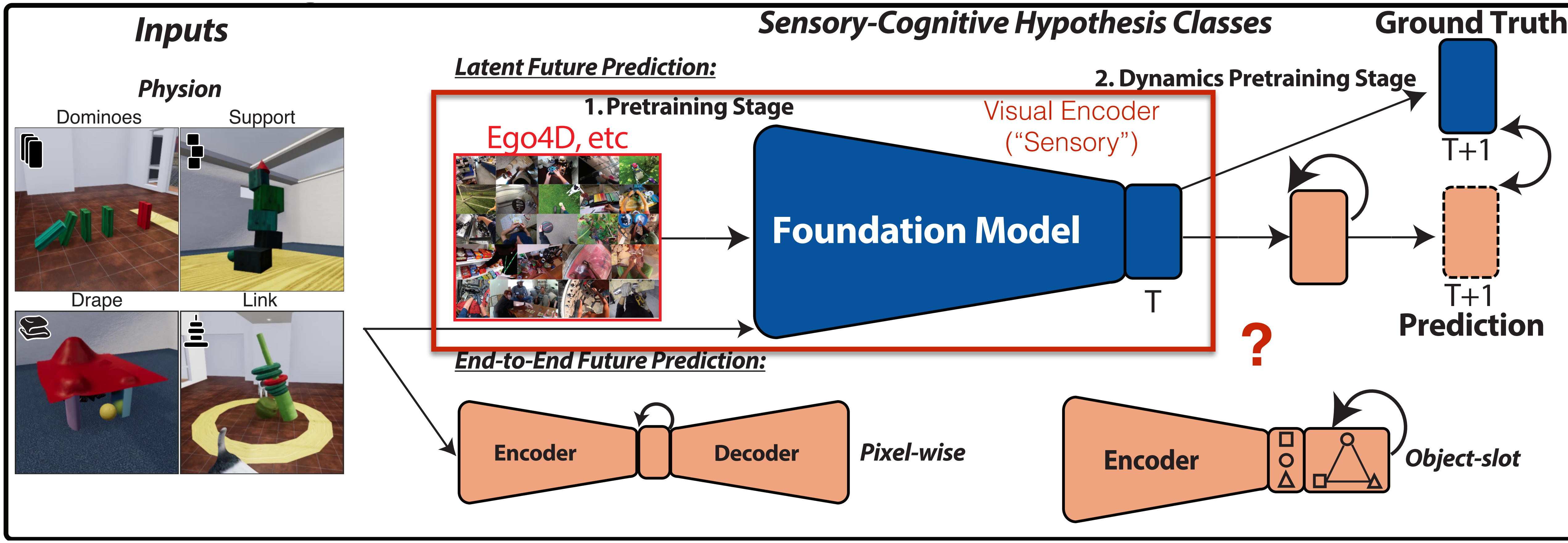


Image Foundation Future Prediction Poorly Predicts Neurons



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

Ego4D: A massive-scale egocentric dataset

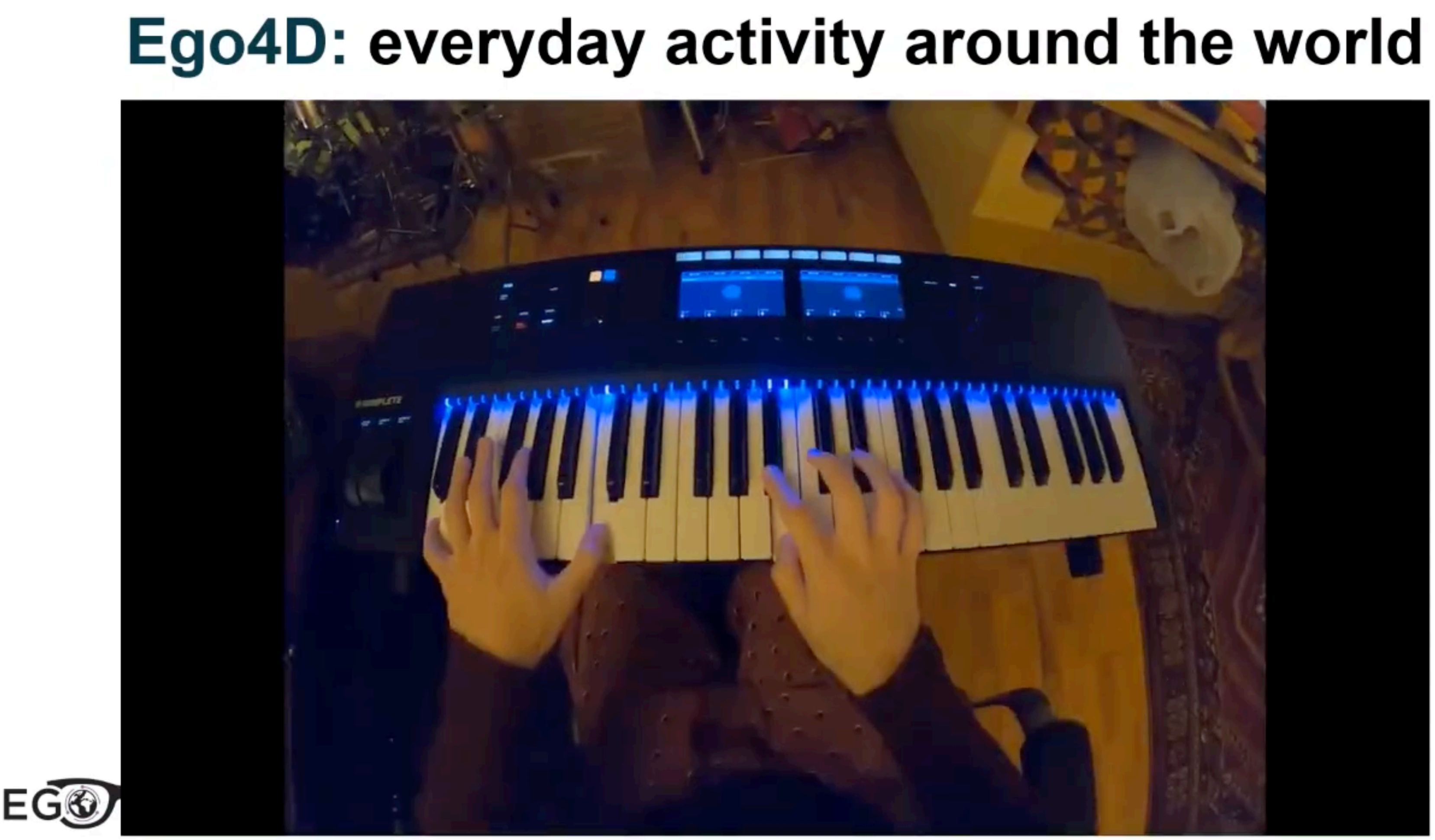
3,670 hours of in-the-wild daily life activity

931 participants from 74 worldwide locations

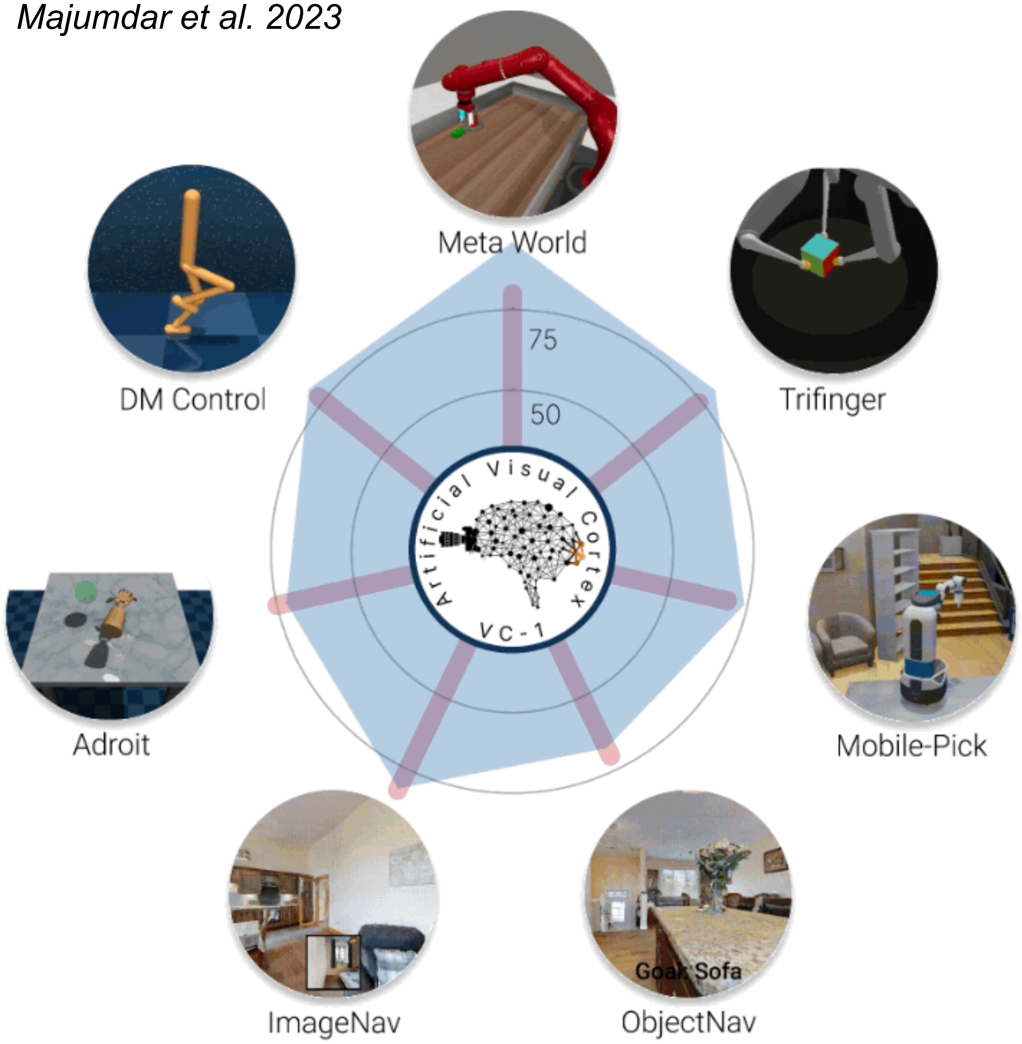
Multimodal: audio, 3D scans, IMU, stereo, multi-camera

Grauman et al. 2022

Hypothesis Class 3: Video Foundation Models



Majumdar et al. 2023

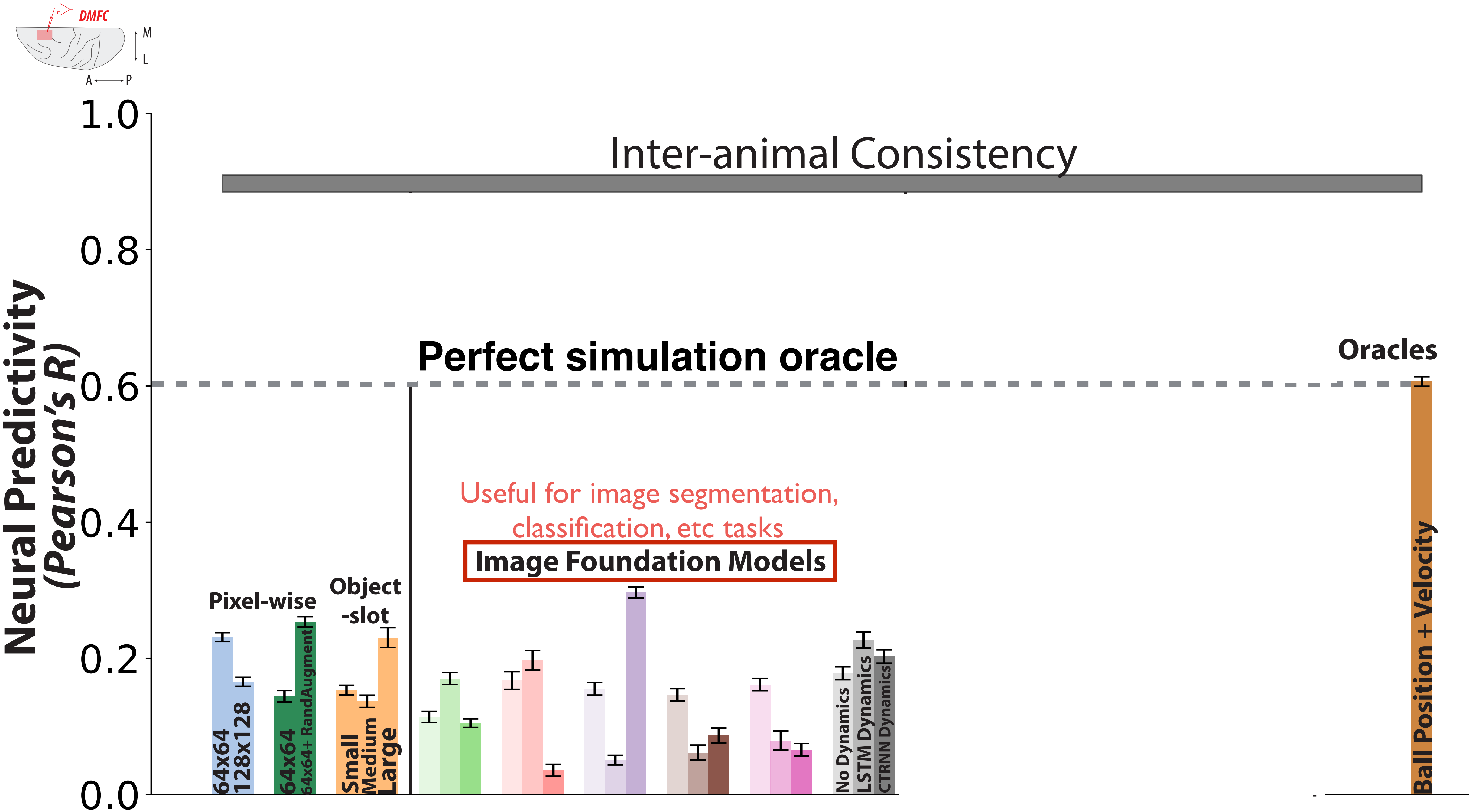


Ego4D: A massive-scale egocentric dataset

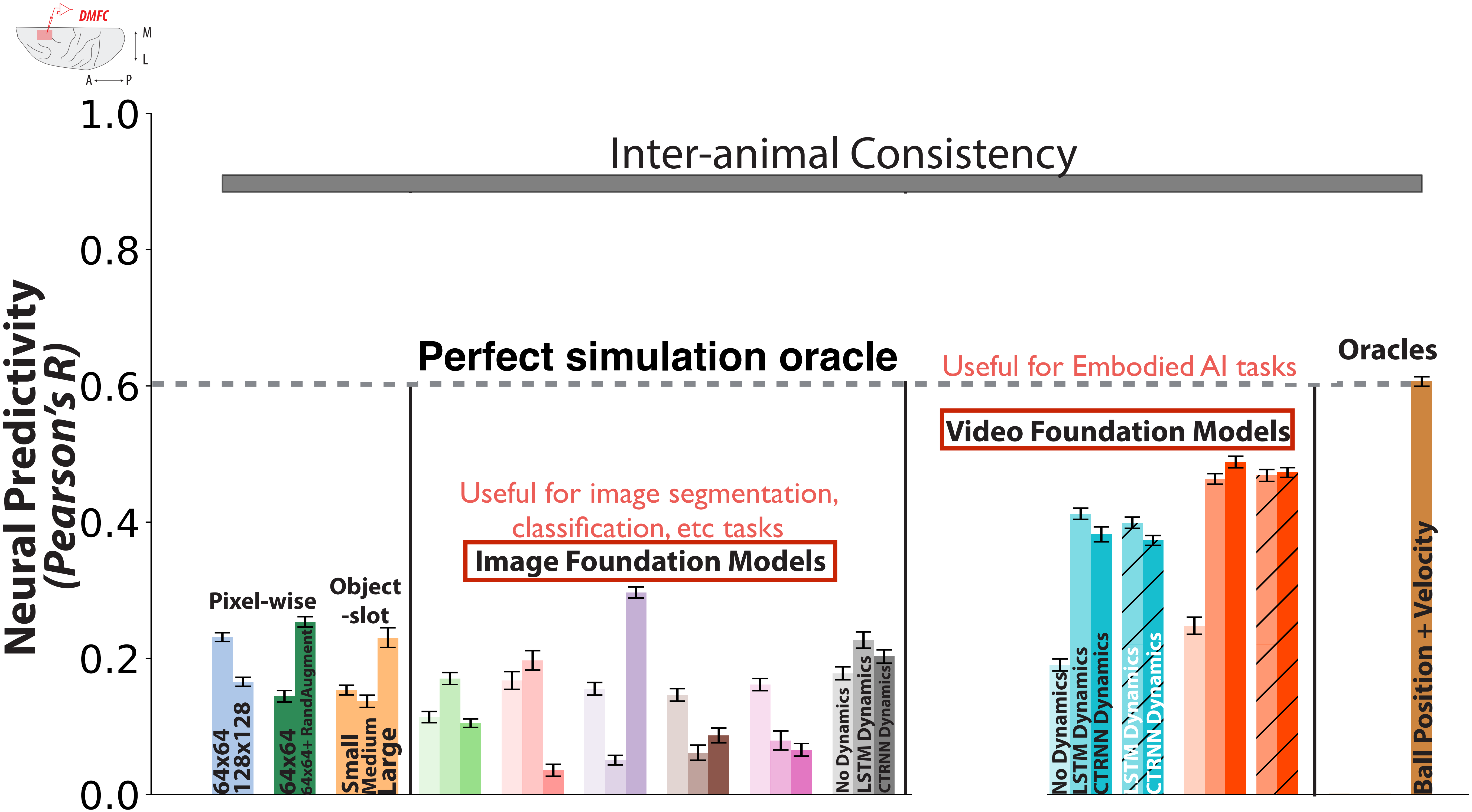
Reusability but with egocentric *videos* rather than static images!

Grauman et al. 2022

Image Foundation Future Prediction Poorly Predicts Neurons

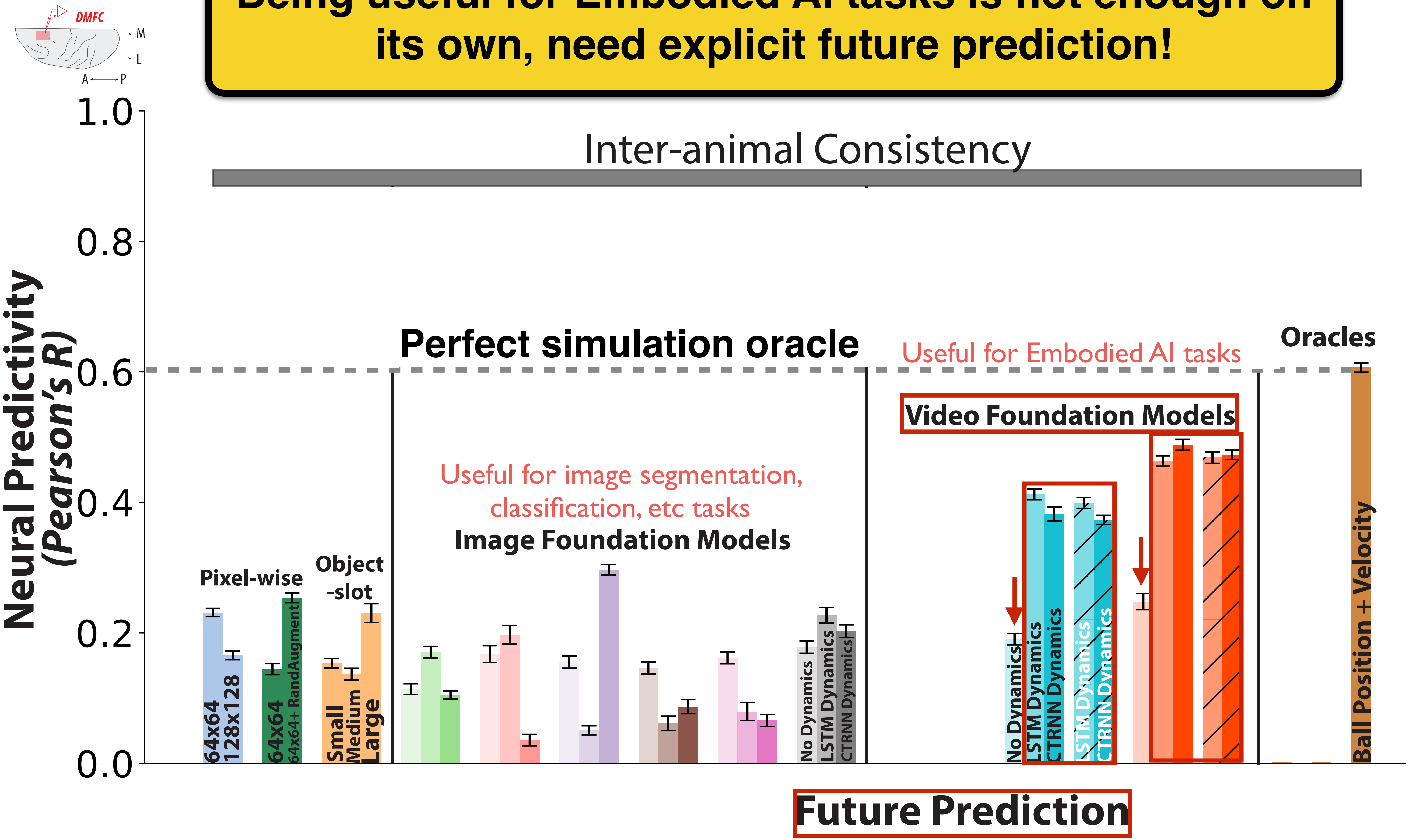


Video Foundation Future Prediction Best Predict Neurons



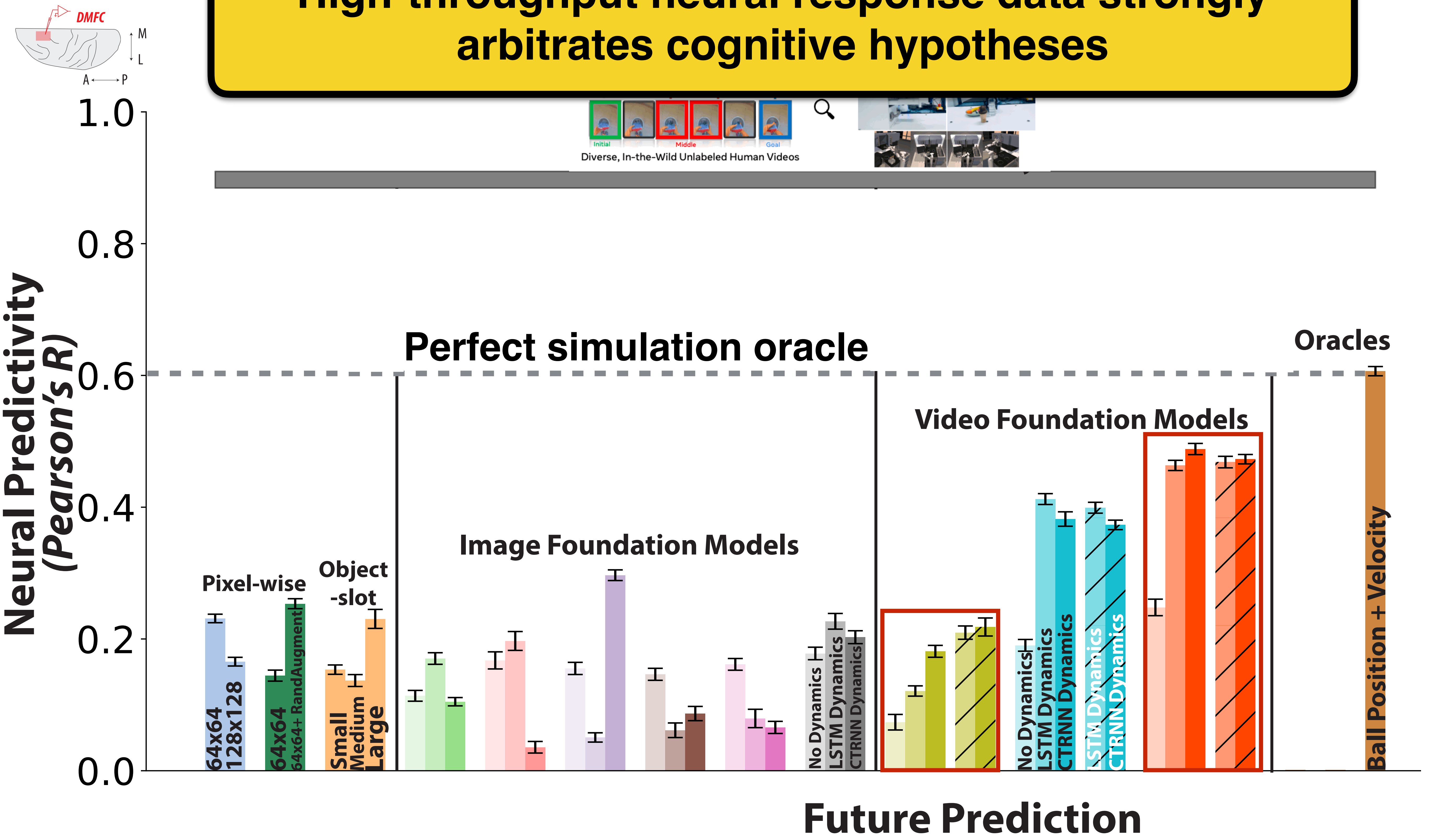
Video Foundation Future Prediction Best Predict Neurons

Being useful for Embodied AI tasks is not enough on its own, need explicit future prediction!

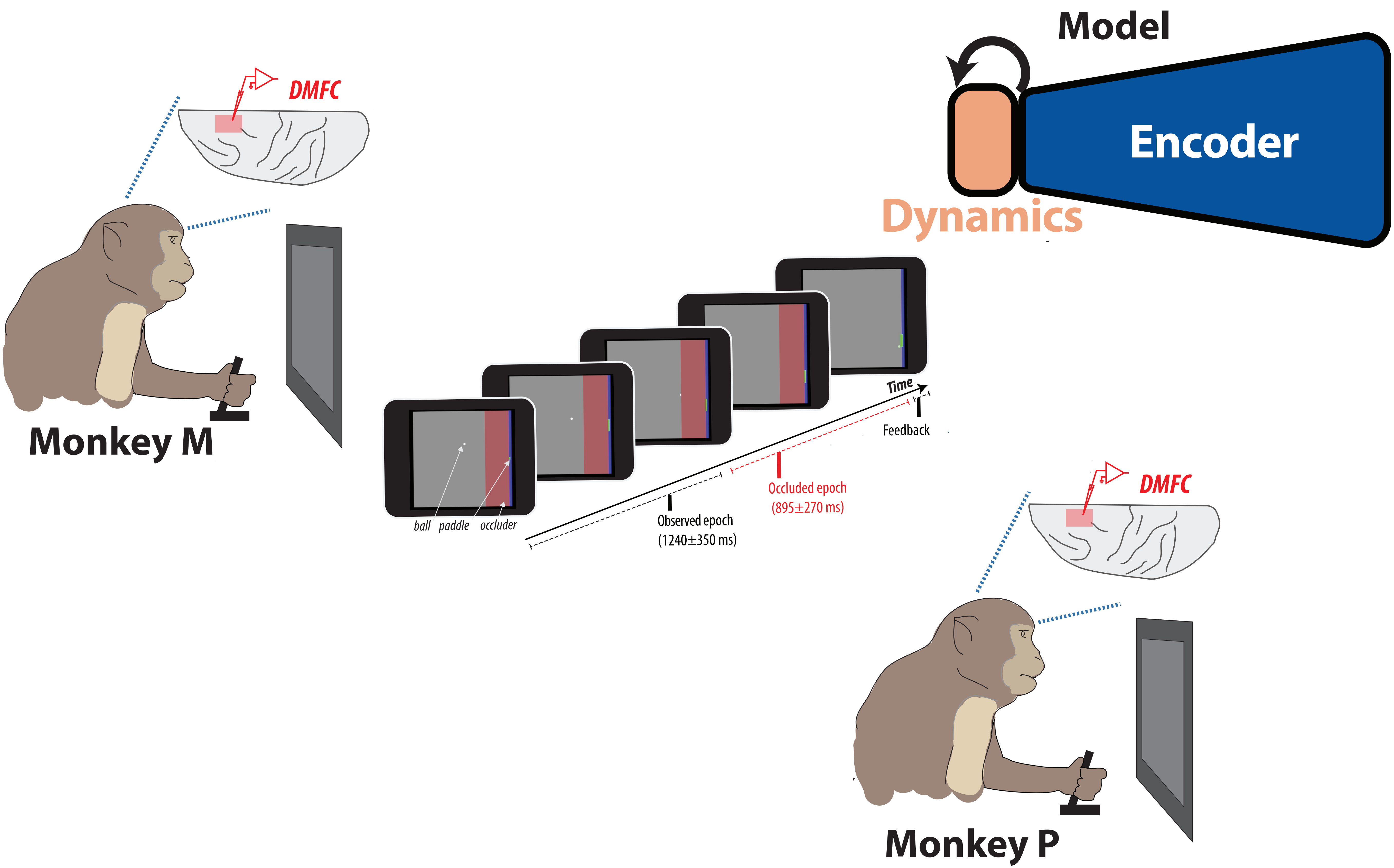


Video Foundation Future Prediction Best Predict Neurons

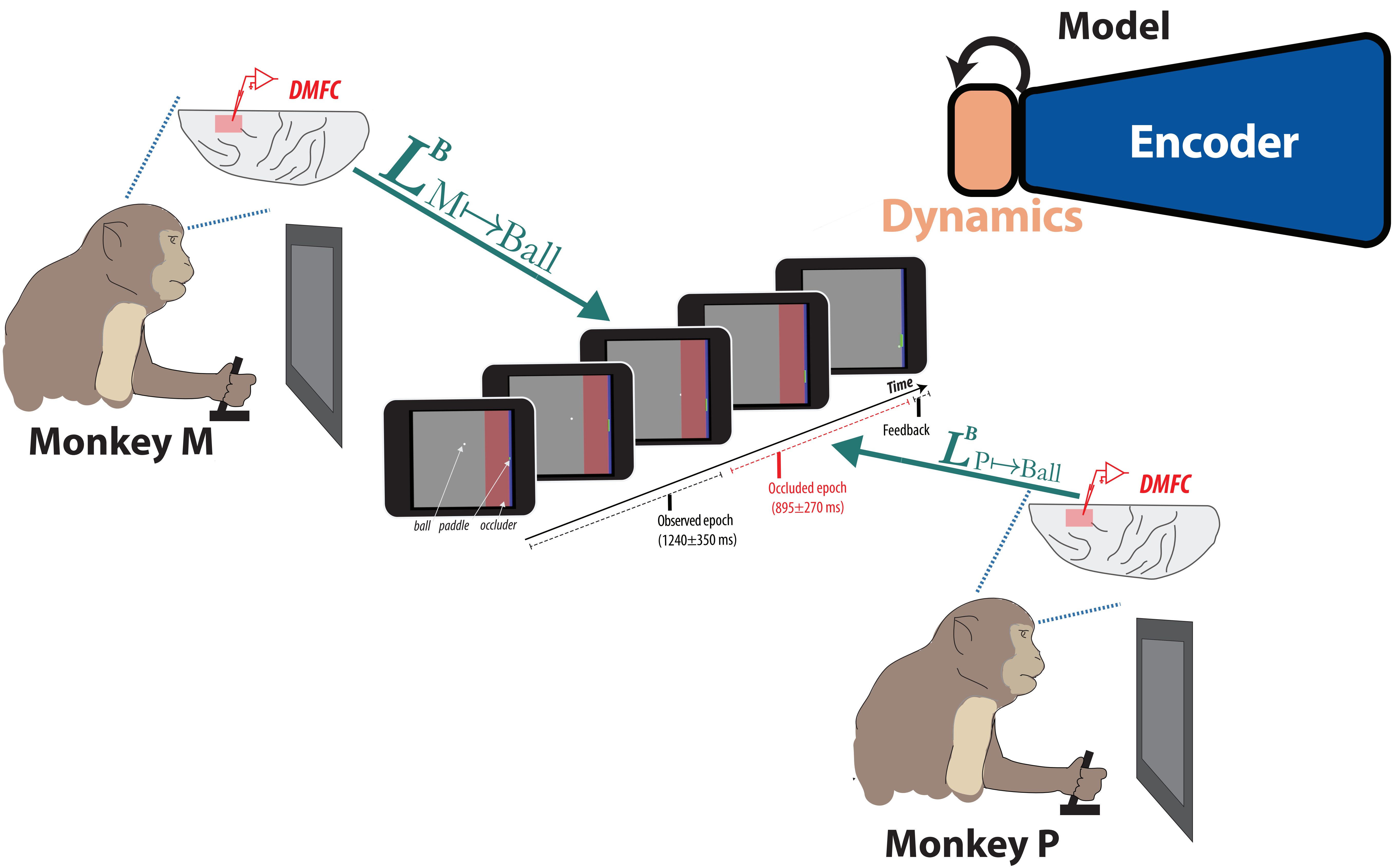
High-throughput neural response data strongly arbitrates cognitive hypotheses



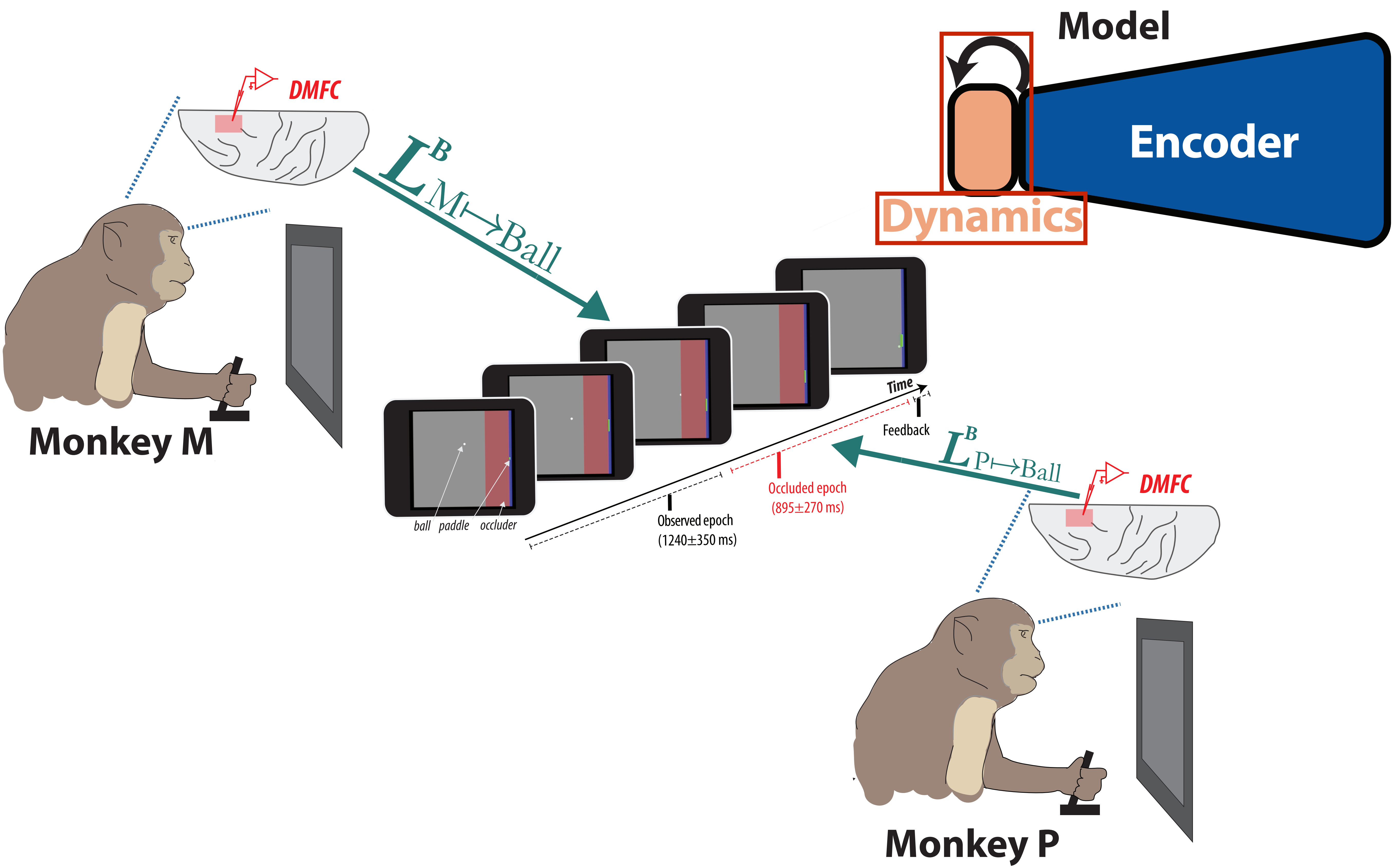
Assessing Model Similarity: Ground Truth State Decoding



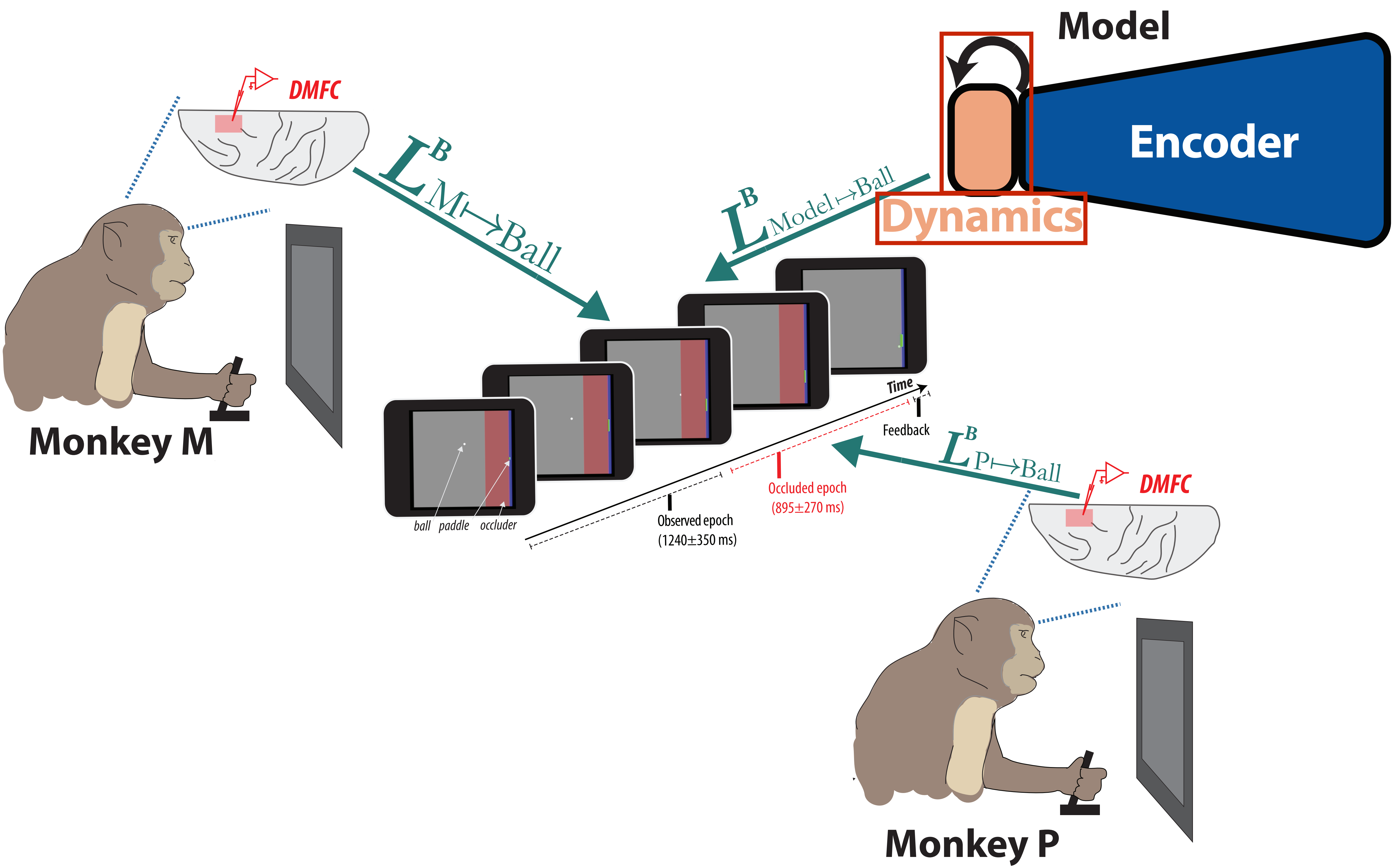
Assessing Model Similarity: Ground Truth State Decoding



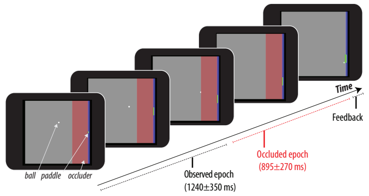
Assessing Model Similarity: Ground Truth State Decoding

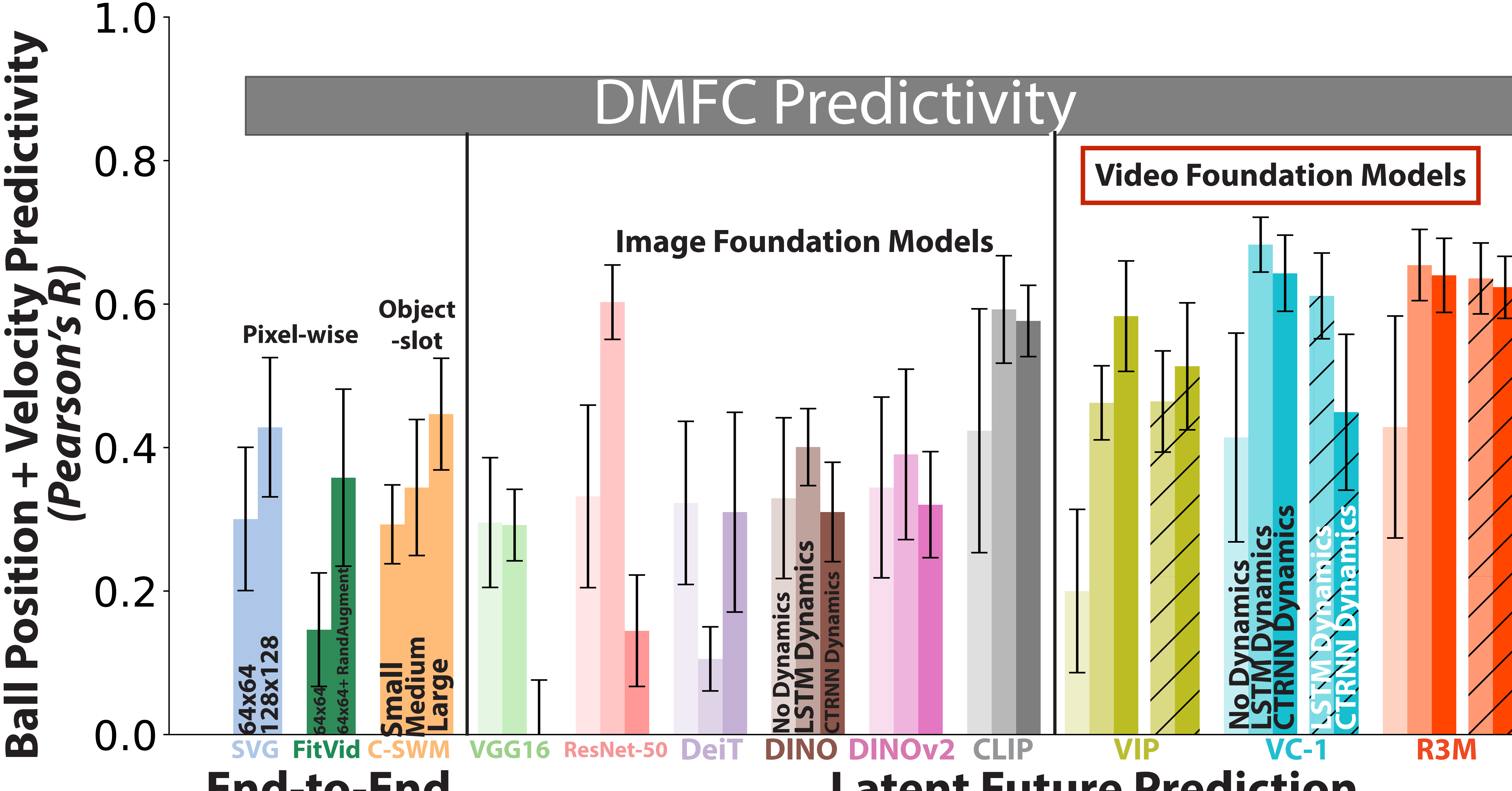


Assessing Model Similarity: Ground Truth State Decoding

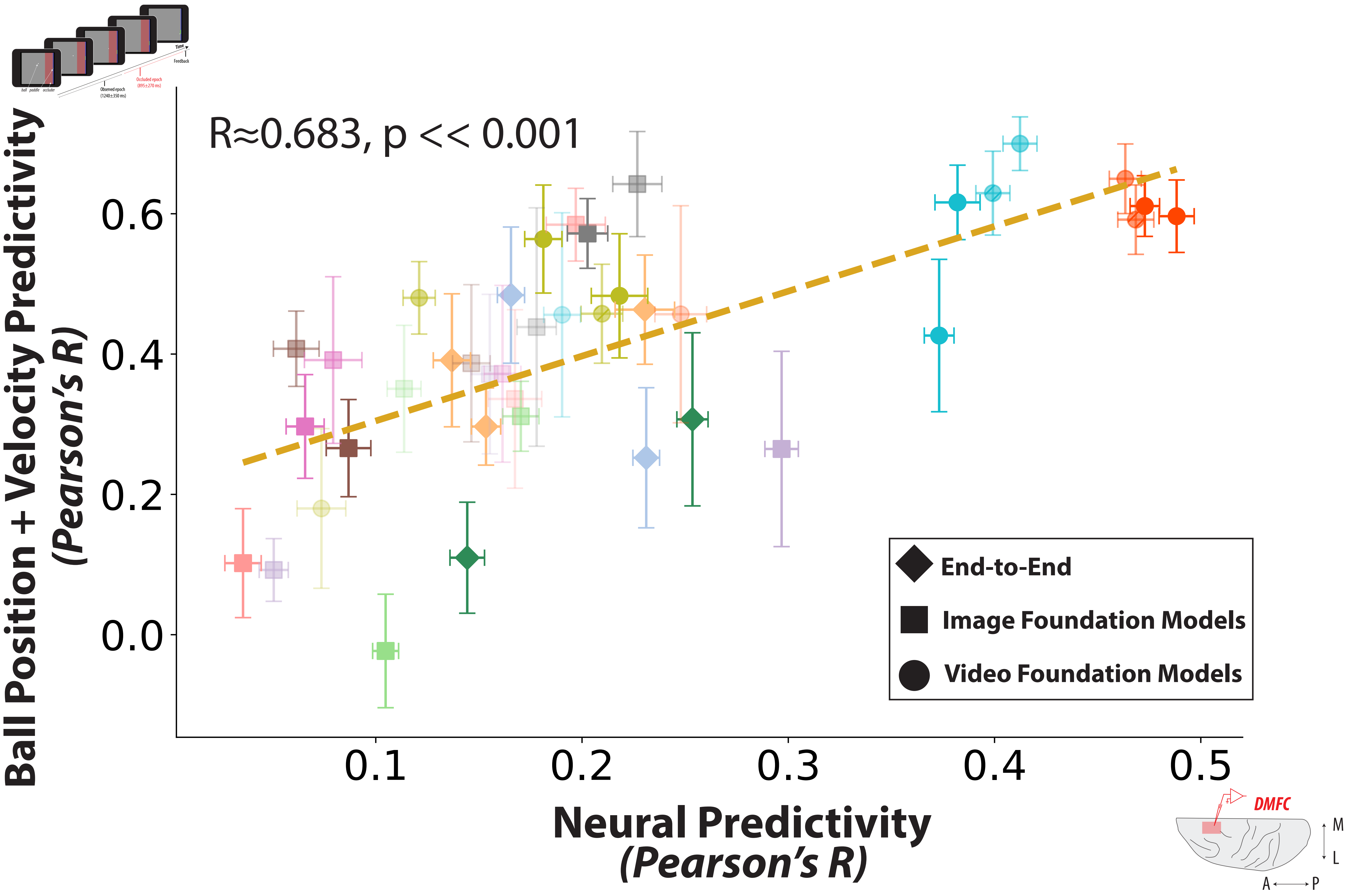


Best models approach ground truth state predictivity ceiling

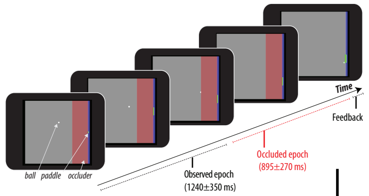


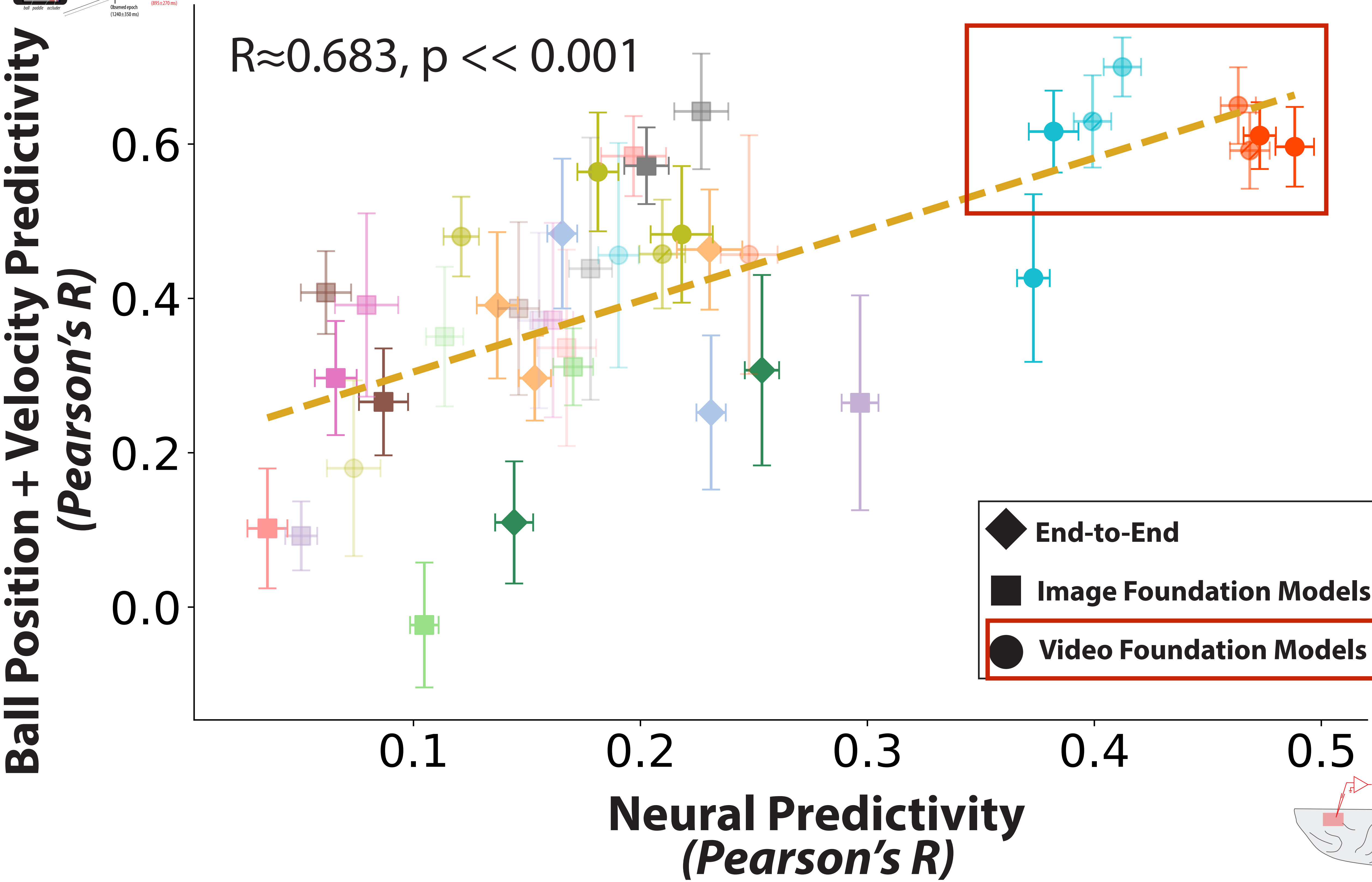


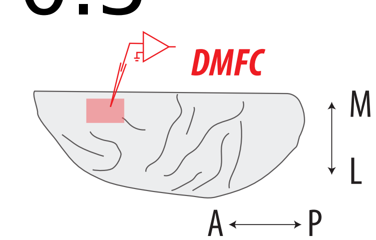
Predicting neurons is relevant to simulating the ball



Predicting neurons is relevant to simulating the ball

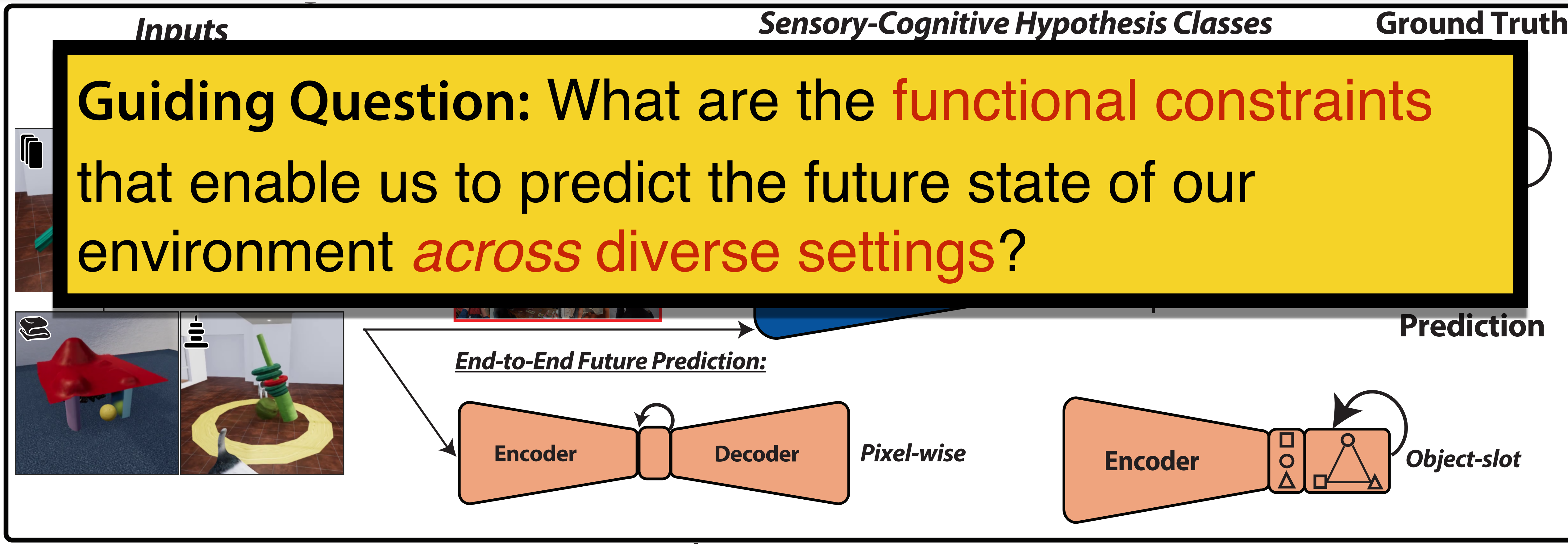


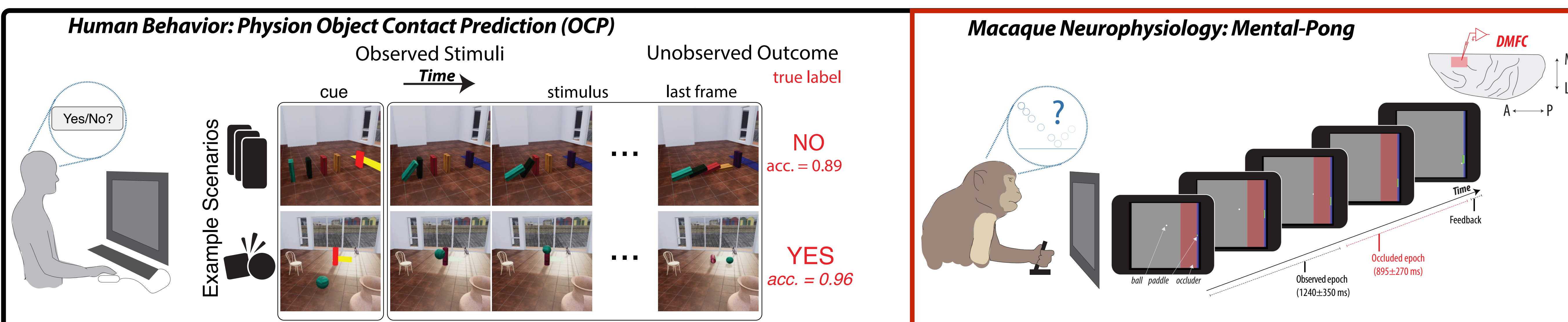




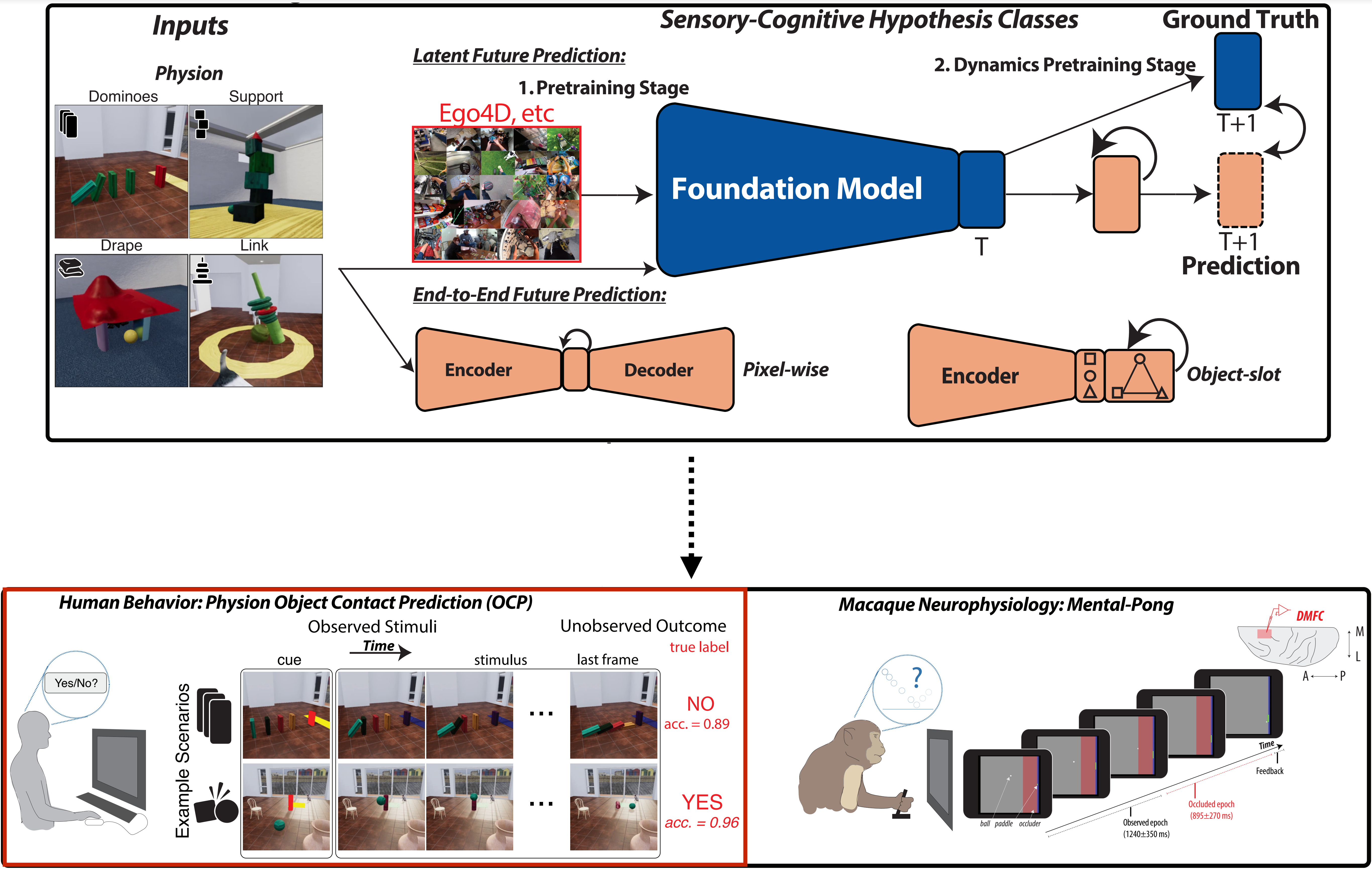
Macaque Neurophysiology: Mental Pong

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





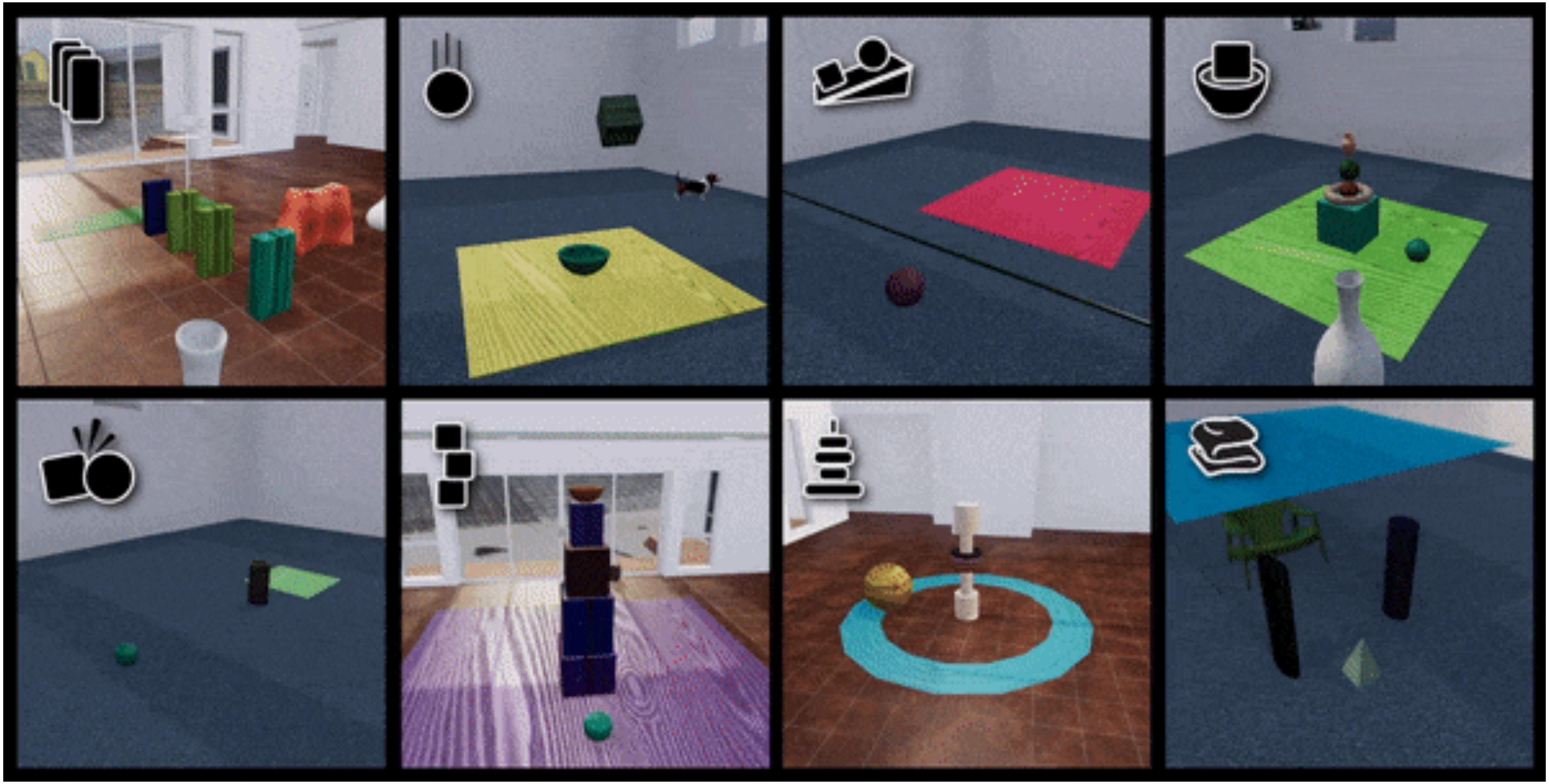
Human Behavior: Object Contact Prediction



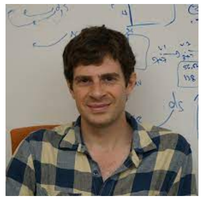
Object Contact Prediction Environment

Physion/ThreeD World (TDW)

Bear et al. 2021



Focus on everyday physical understanding



Daniel Bear

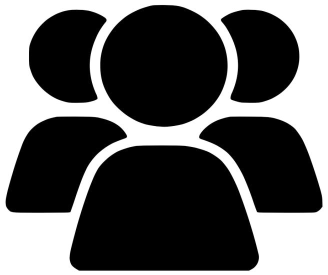
Joshua Tenenbaum

Daniel Yamins

Judith Fan

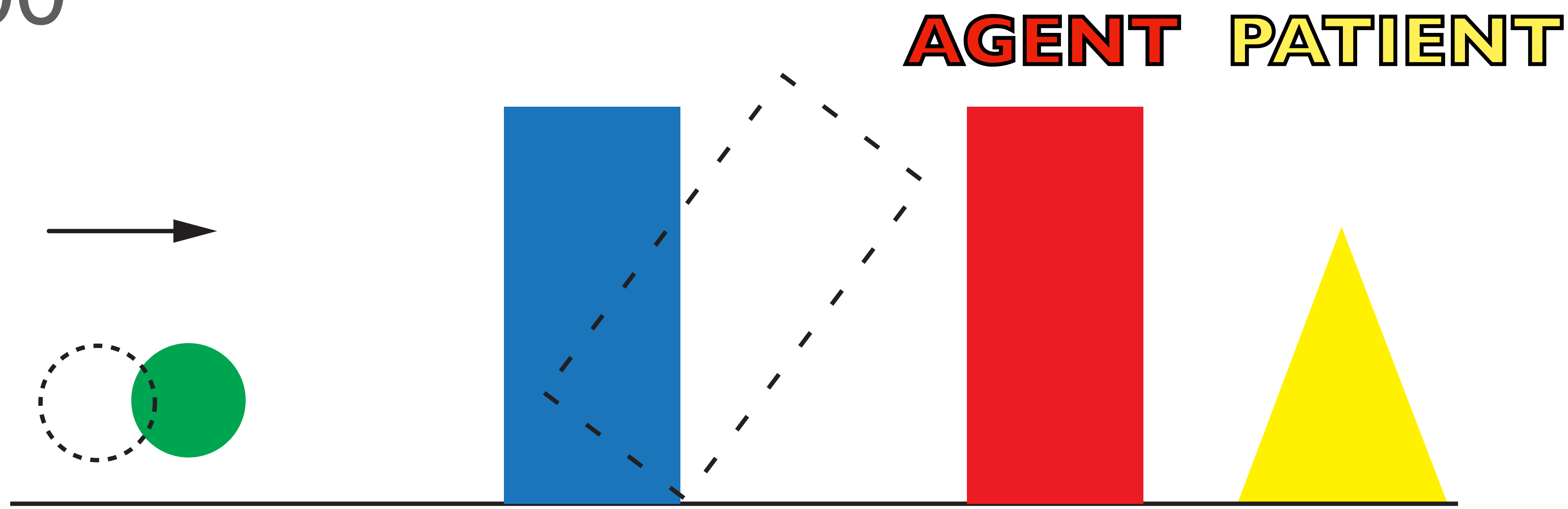
Human Behavior: Object Contact Prediction

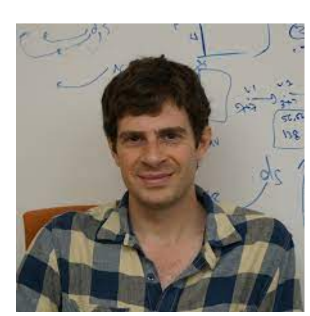
Bear et al. 2021



n=100

"Will the agent object contact the patient object?"




Daniel Bear

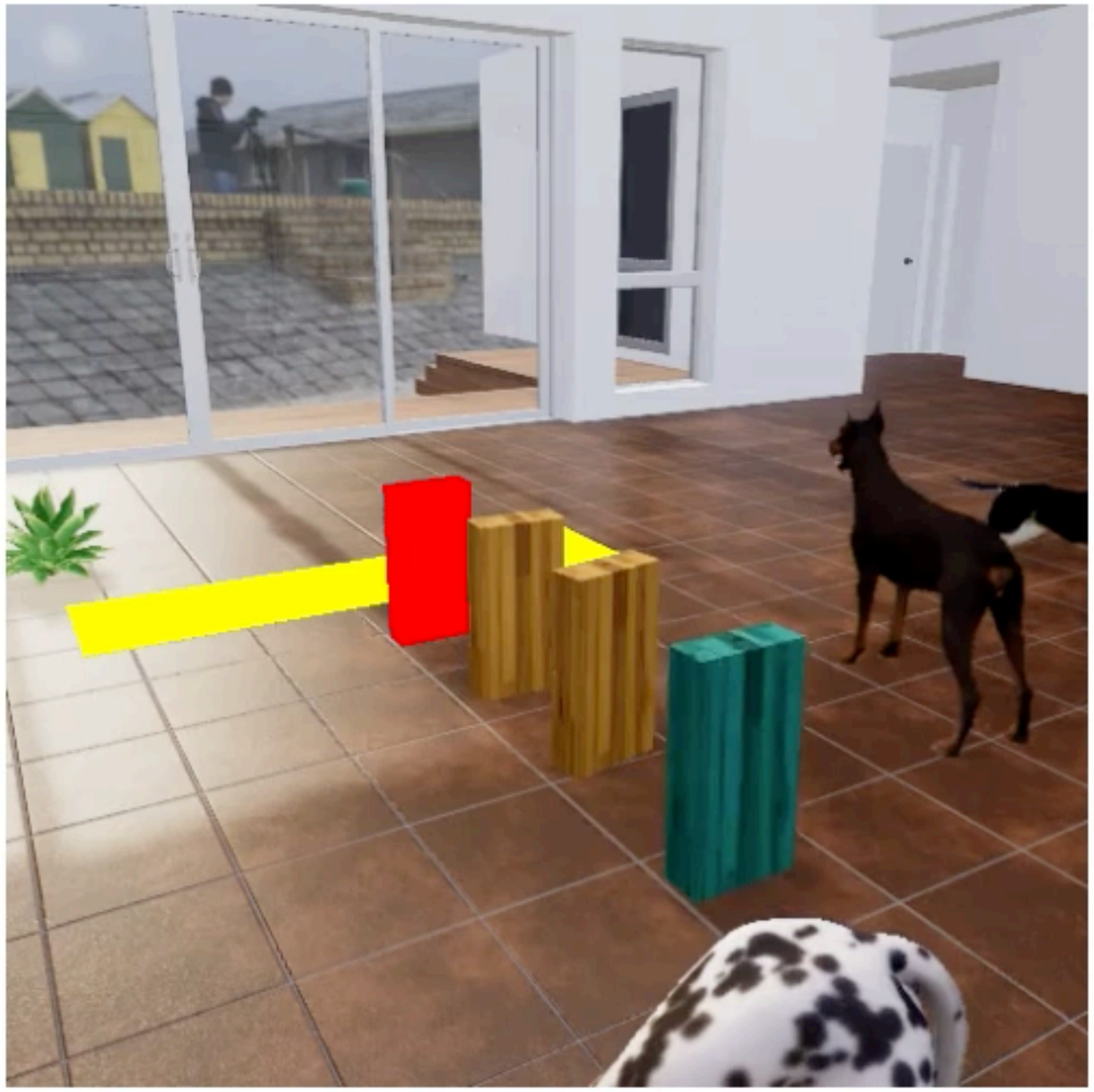

Joshua Tenenbaum


Daniel Yamins


Judith Fan

Completion Progress

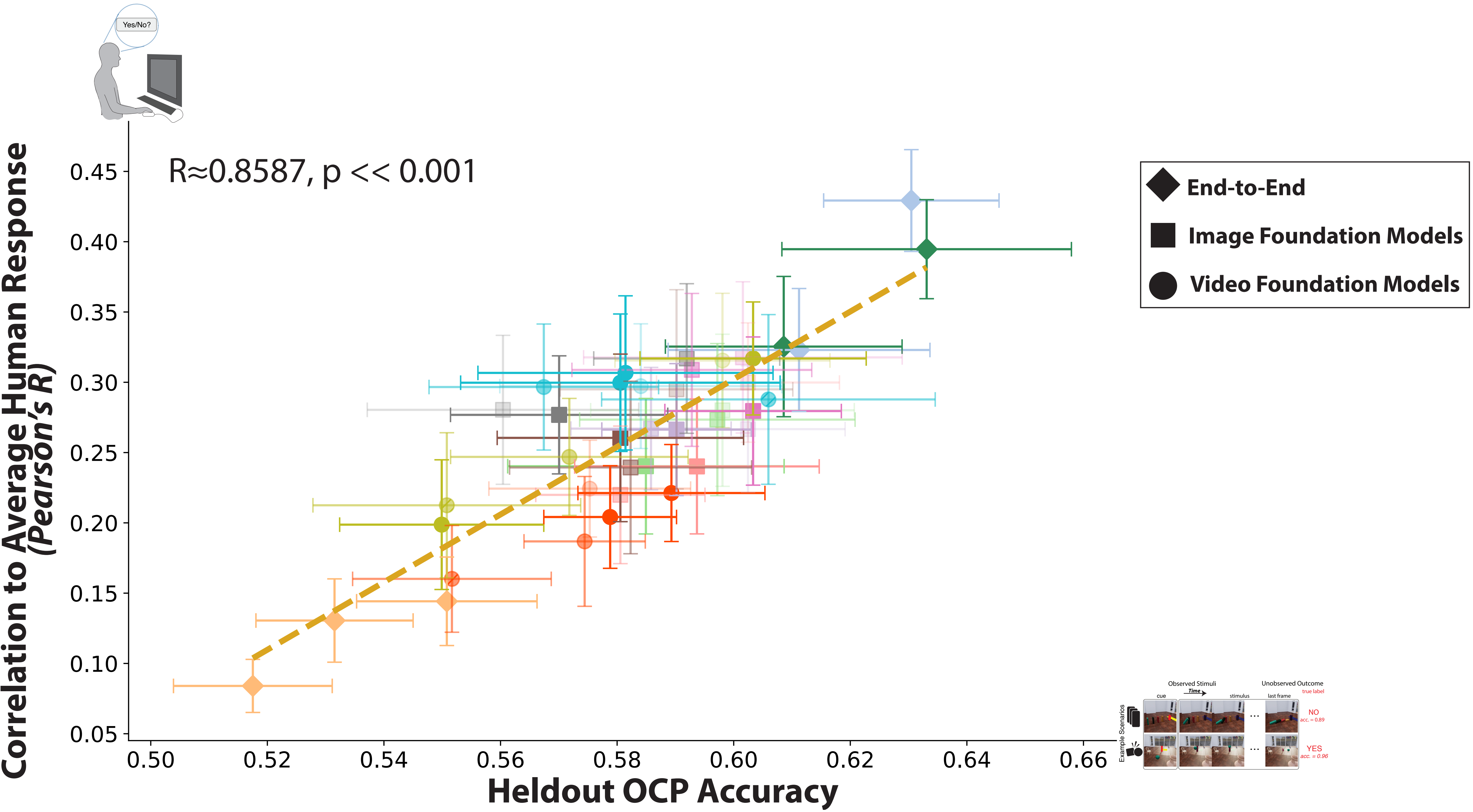
Bear et al. 2021



YES NO

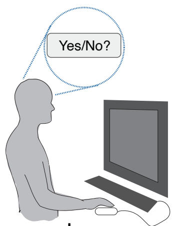
Is the red object going to hit the yellow area?

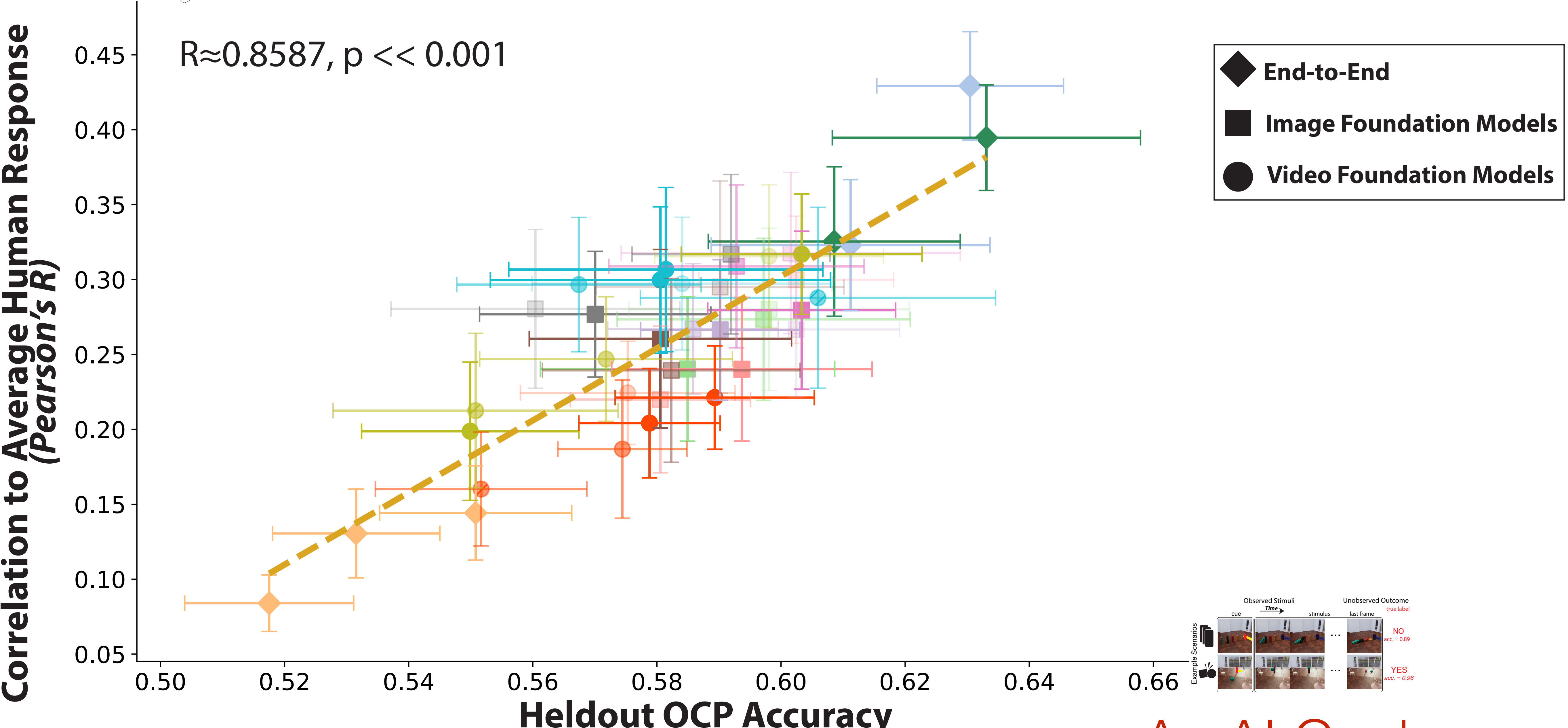
OCP Accuracy & Matching Human Error Patterns Are Related



OCP Accuracy & Matching Human Error Patterns Are Related

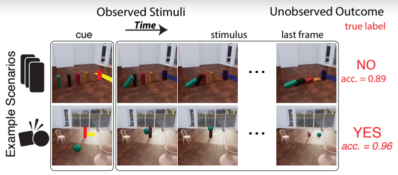
A Cognitive Goal

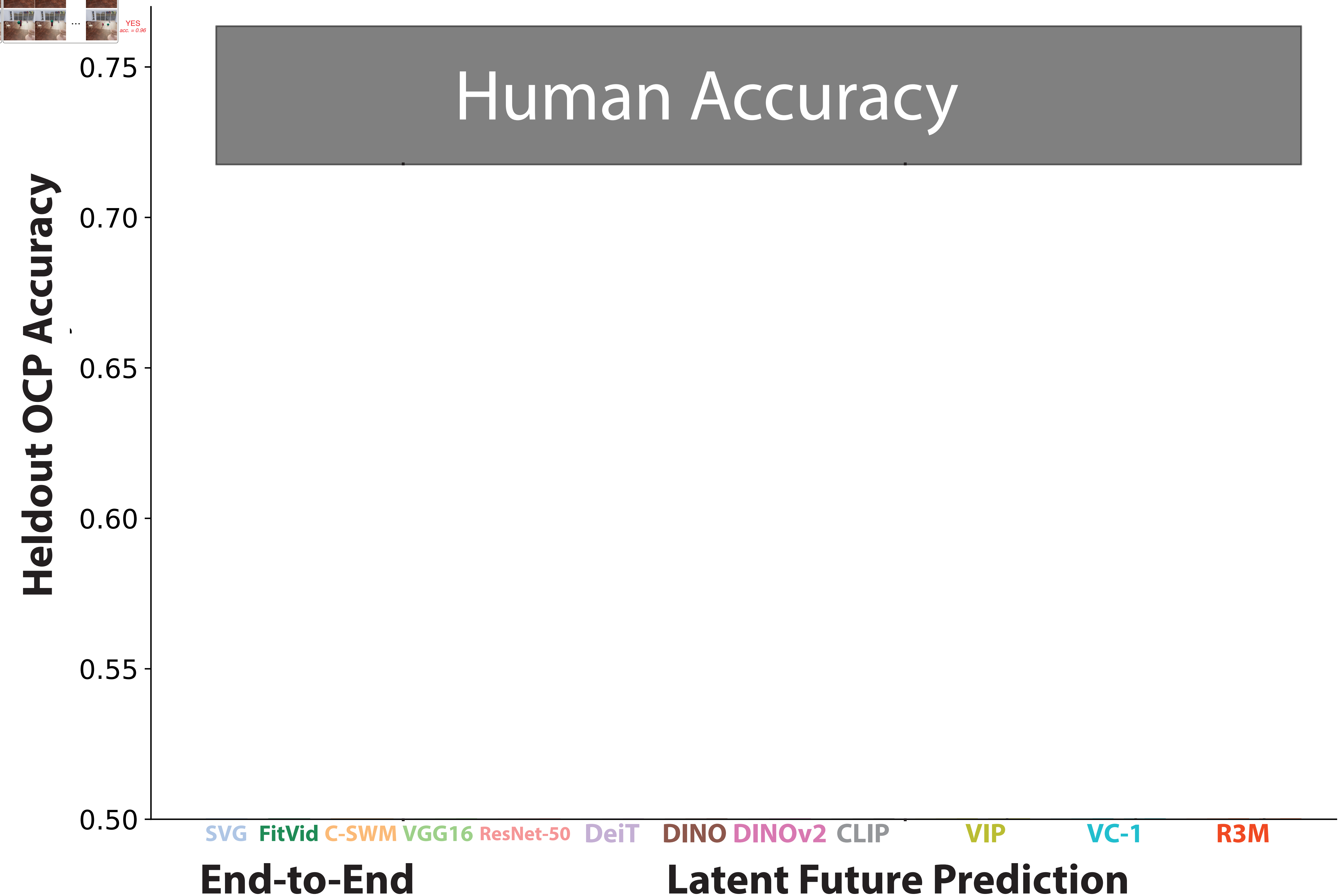




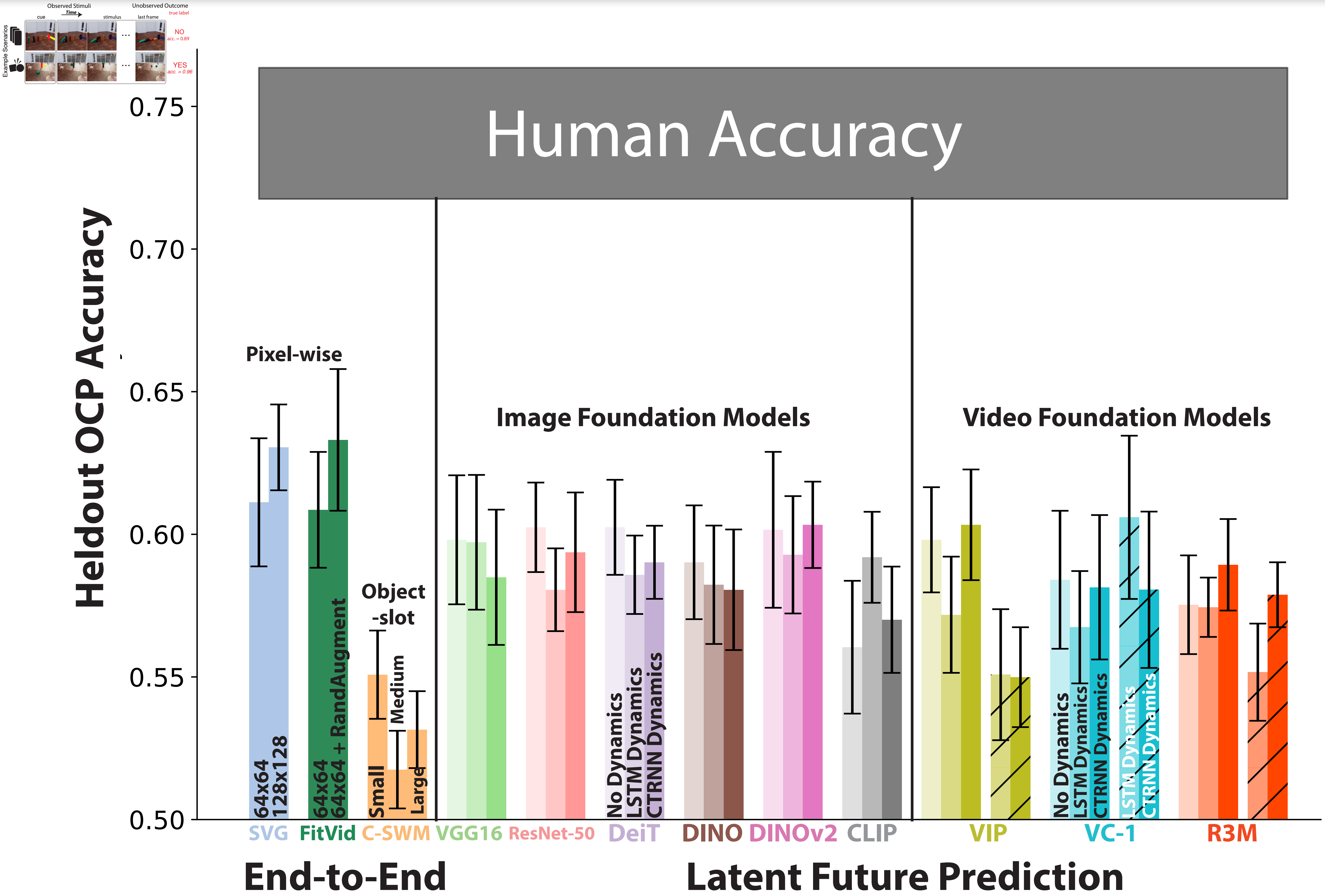
An AI Goal

Comparing Visually-Grounded Models to Human Judgements

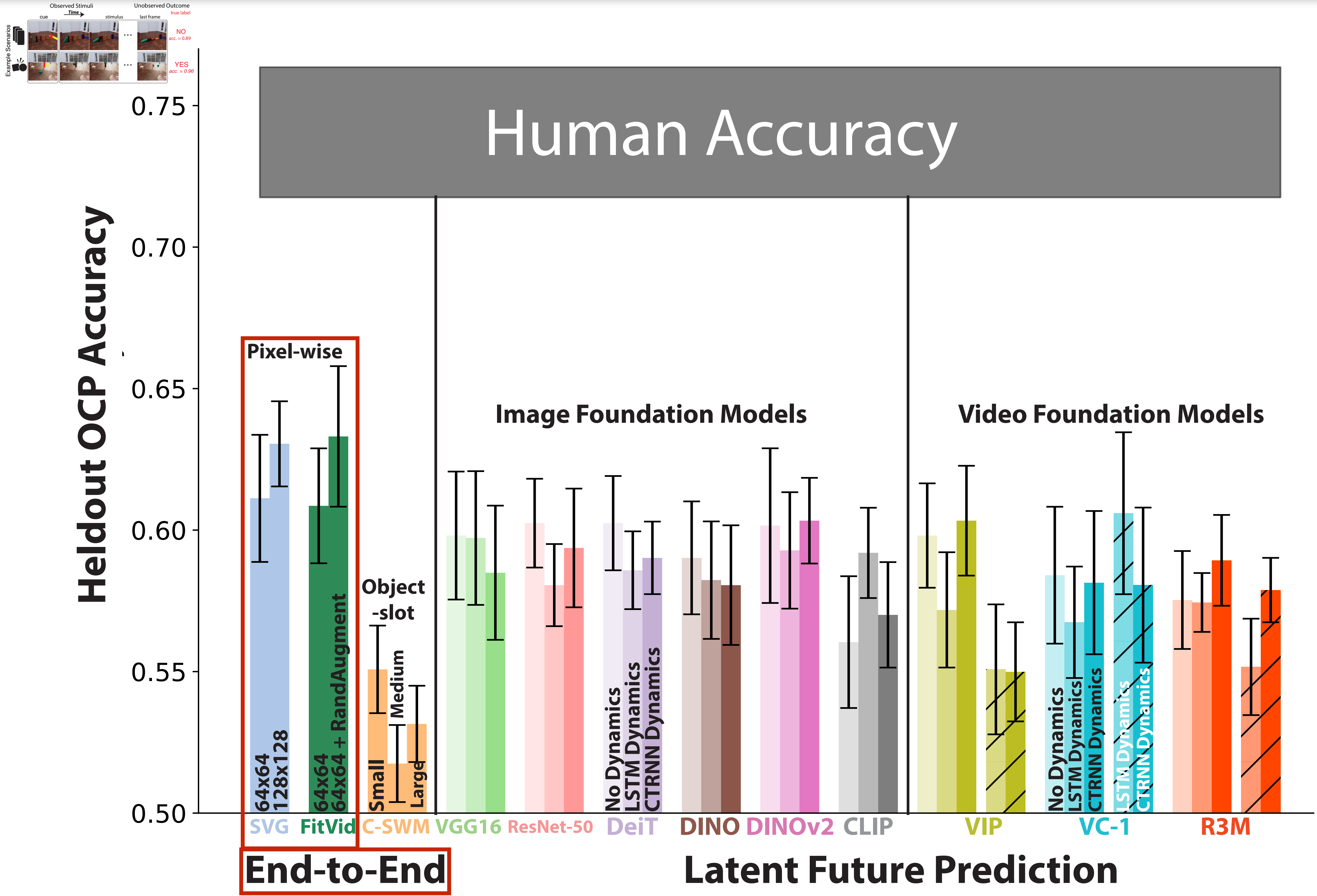




Comparing Visually-Grounded Models to Human Judgements

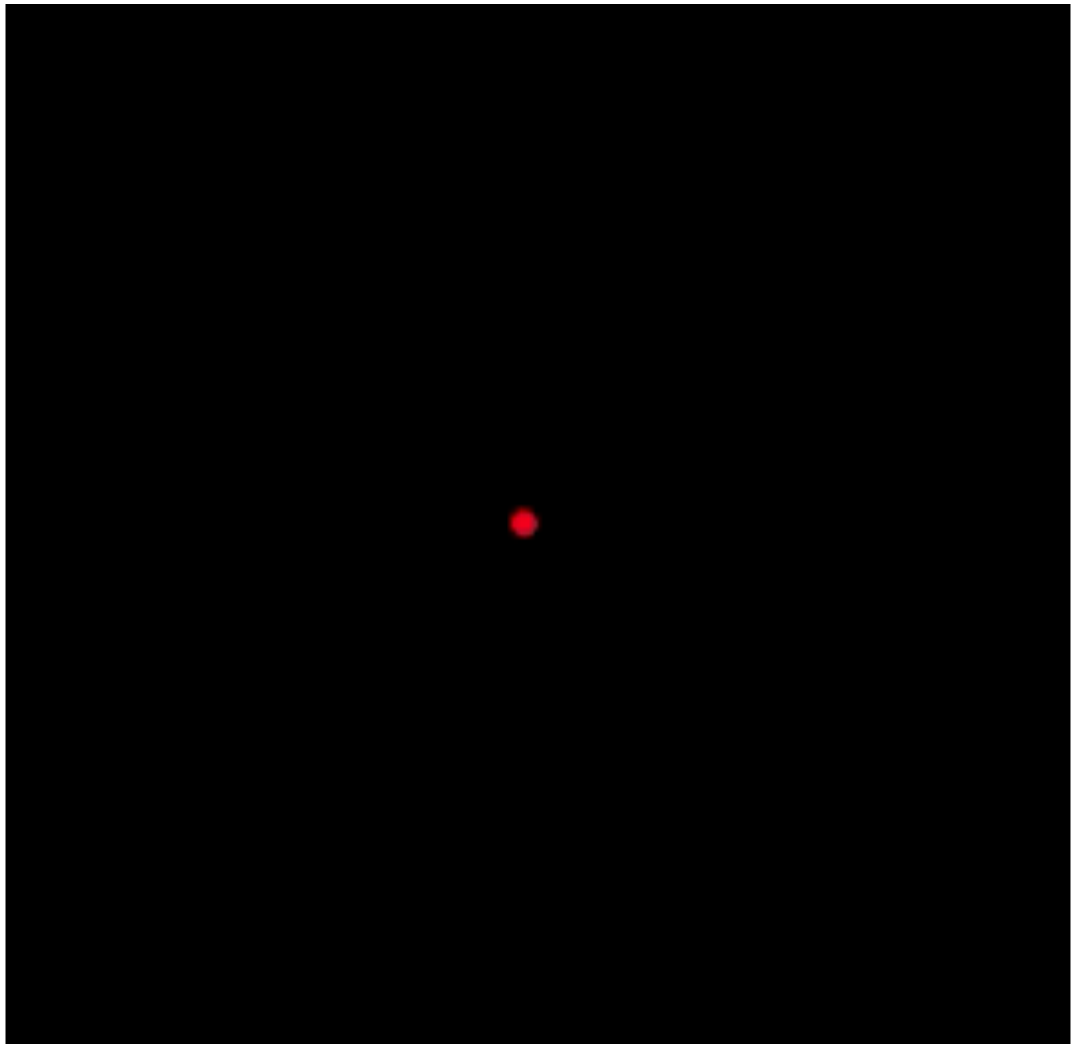


Pixel-wise future predictors are best in the *same* environment

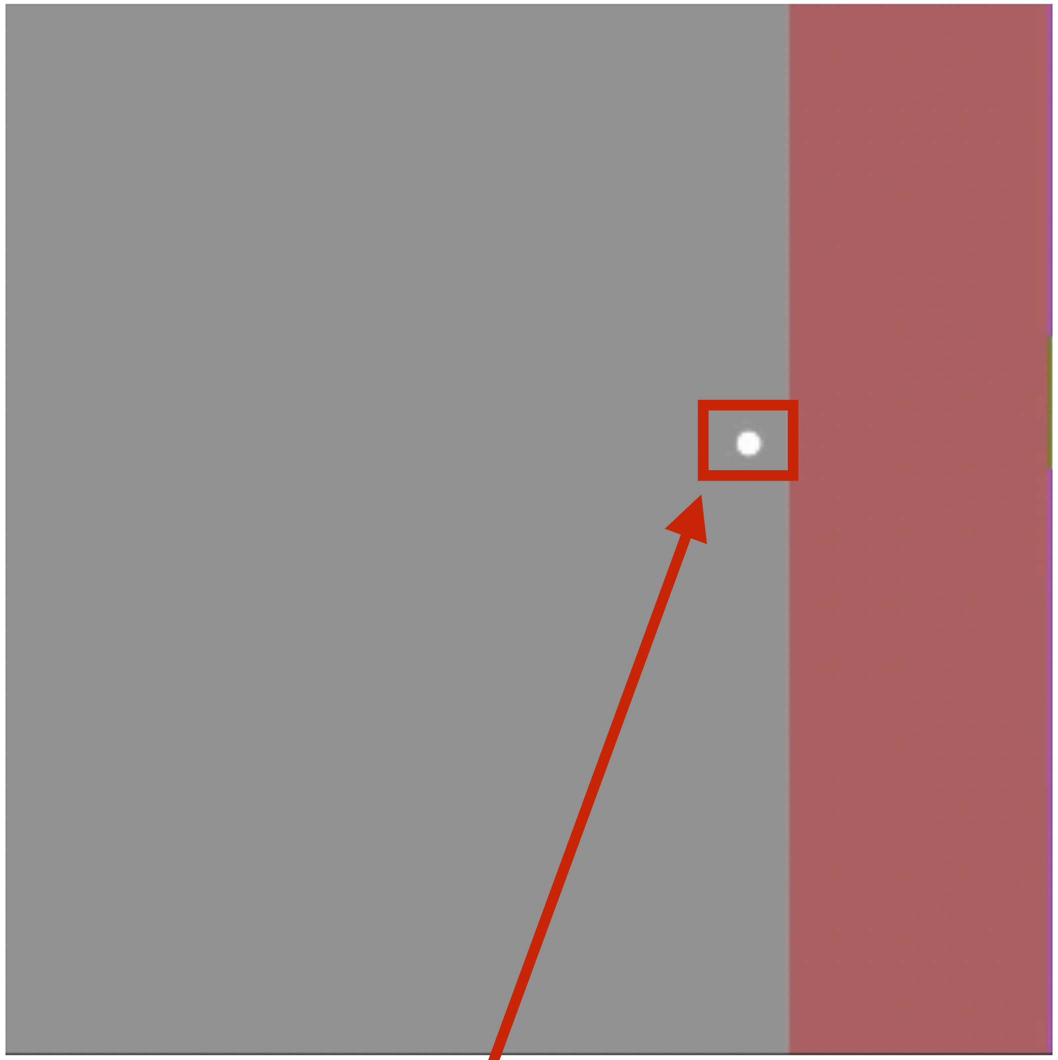


…but they struggle to generalize to Pong

Input Frames

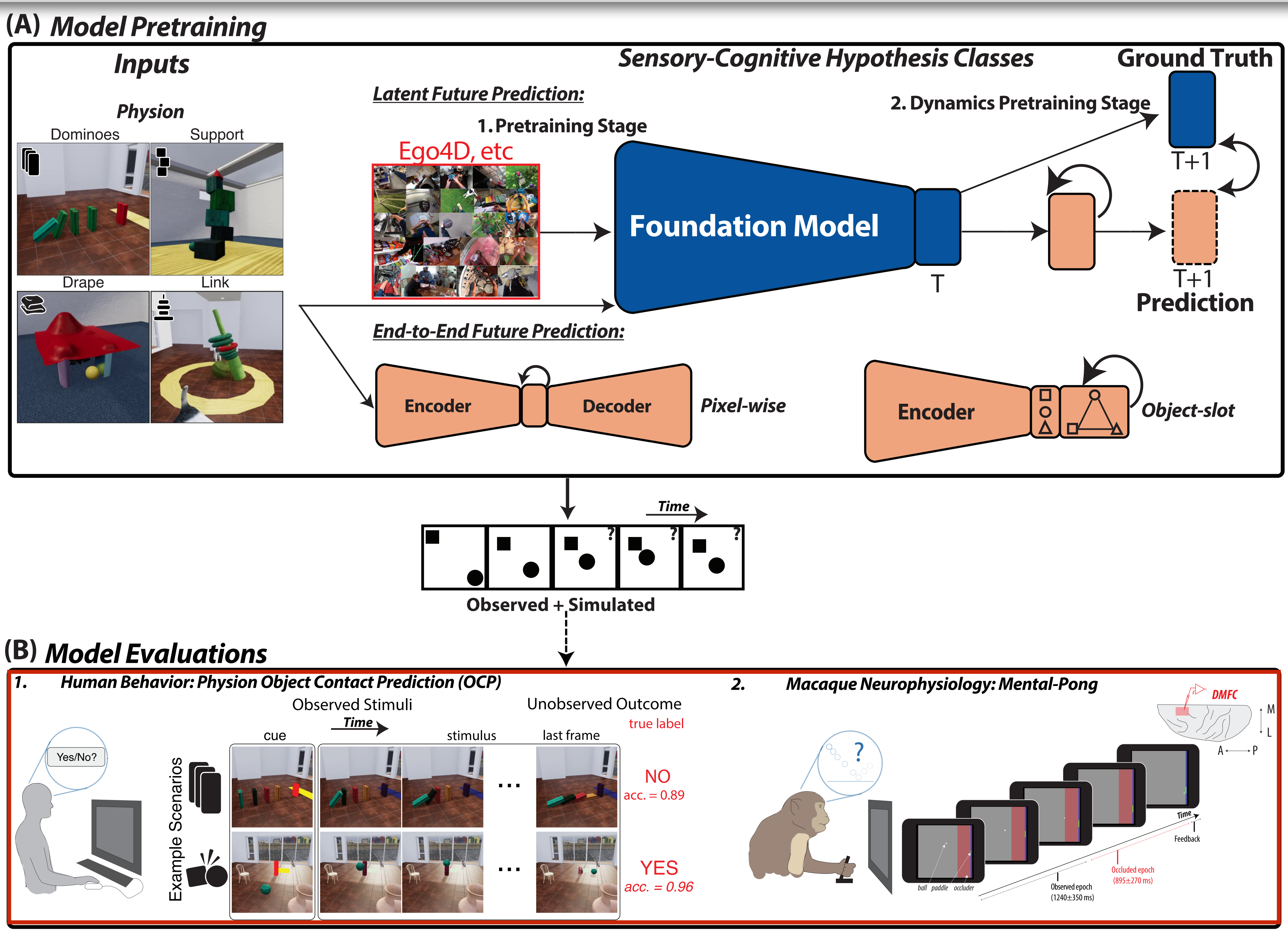


Predicted Frames

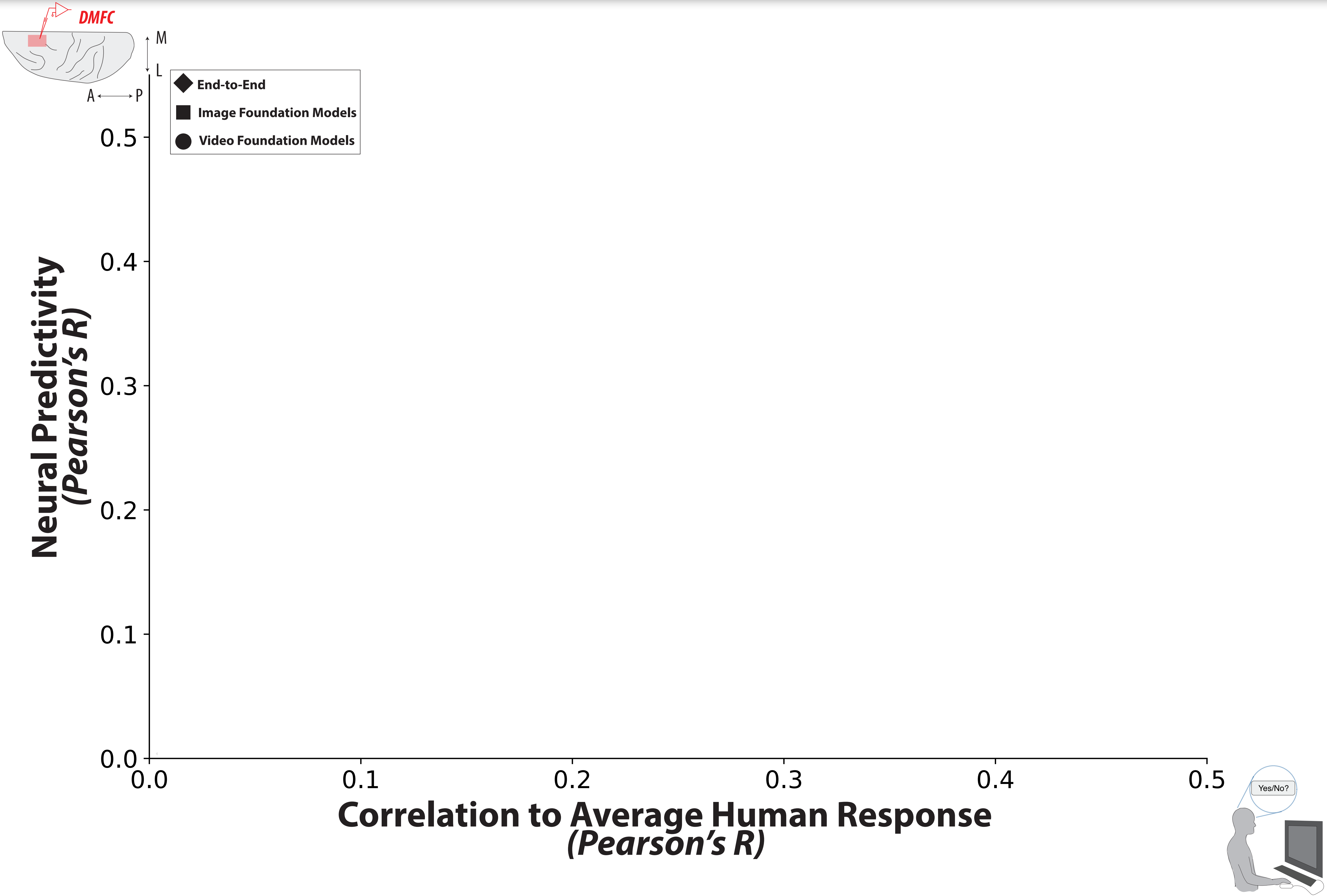


Ball stops at final input frame, in the model's "imagination"

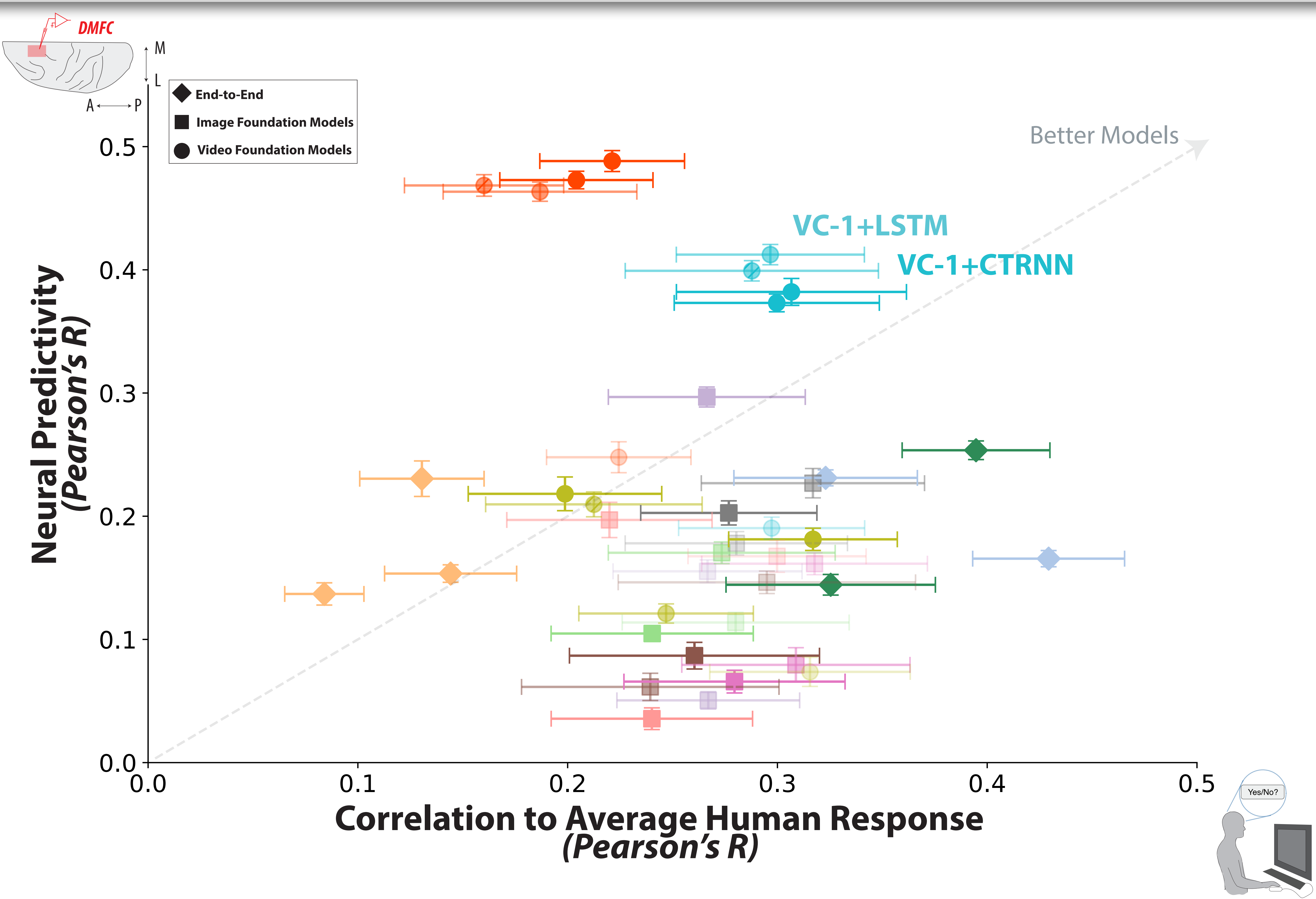
Model Evaluations: What About Both Metrics?



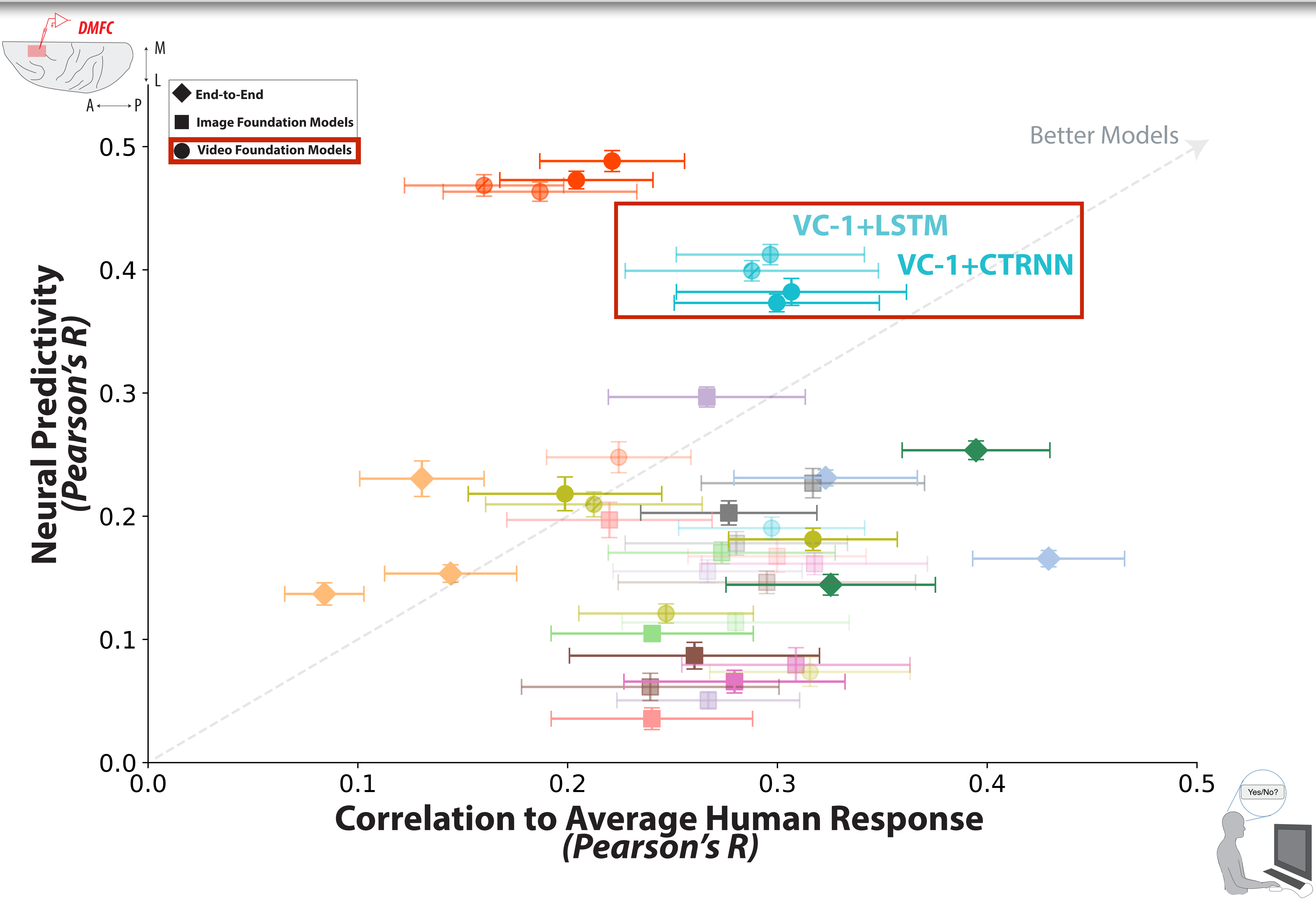
Comparing to Both Human Behavioral and Neural Response Patterns



Dynamically-Equipped Video Foundation Models Can Match Both

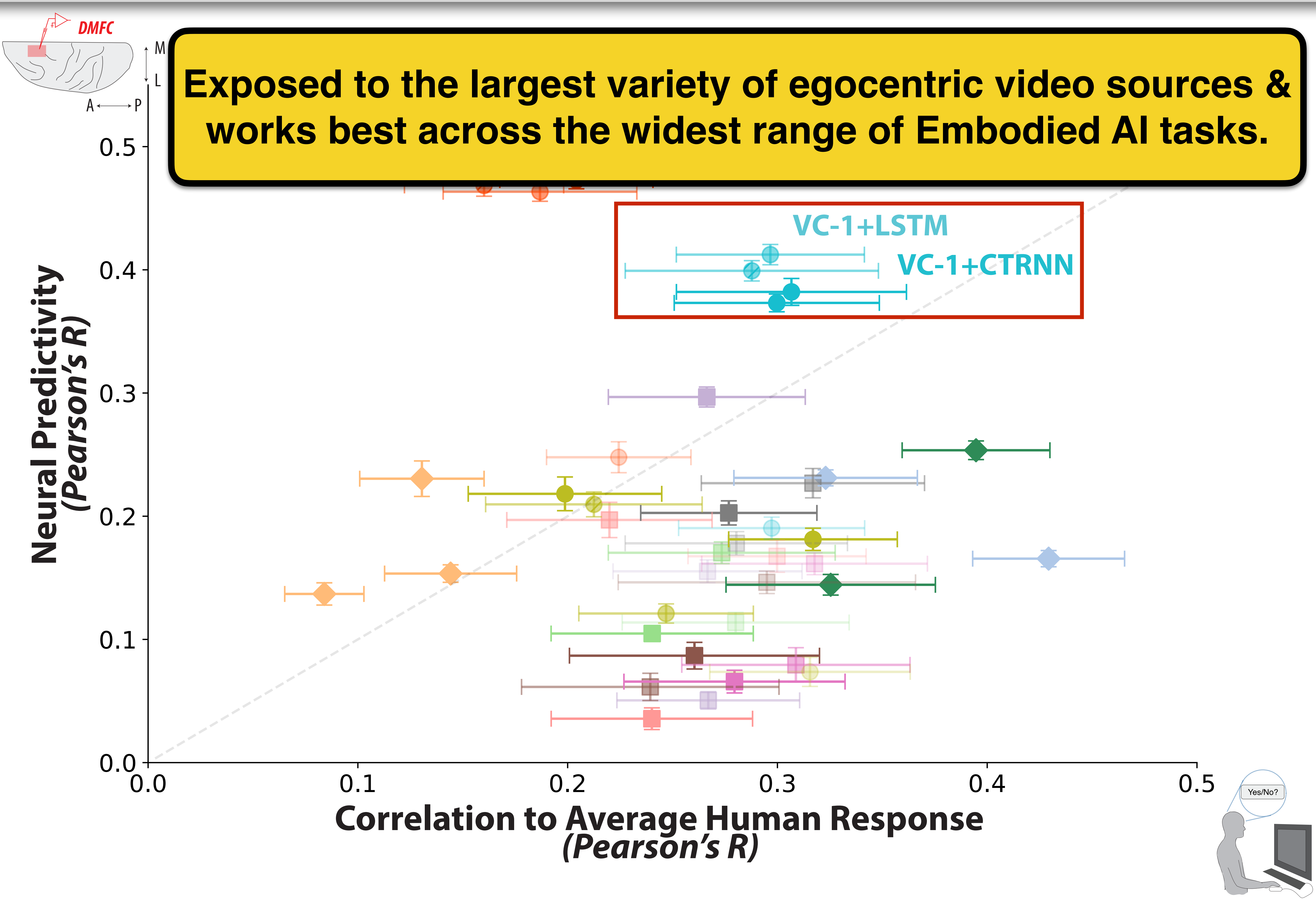


Dynamically-Equipped Video Foundation Models Can Match Both

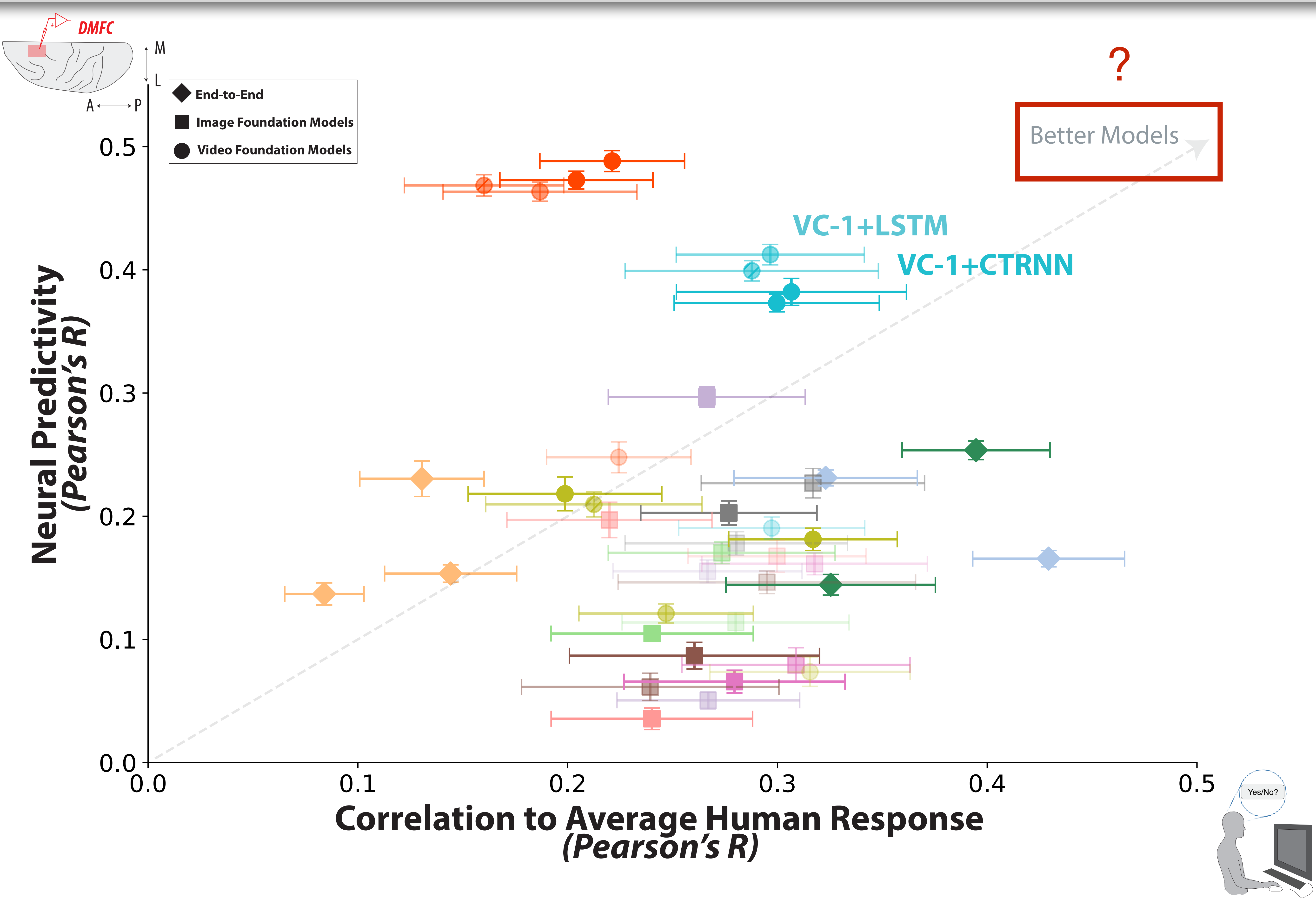


Dynamically-Equipped Video Foundation Models Can Match Both

Exposed to the largest variety of egocentric video sources & works best across the widest range of Embodied AI tasks.



Towards More Robust Future Inference

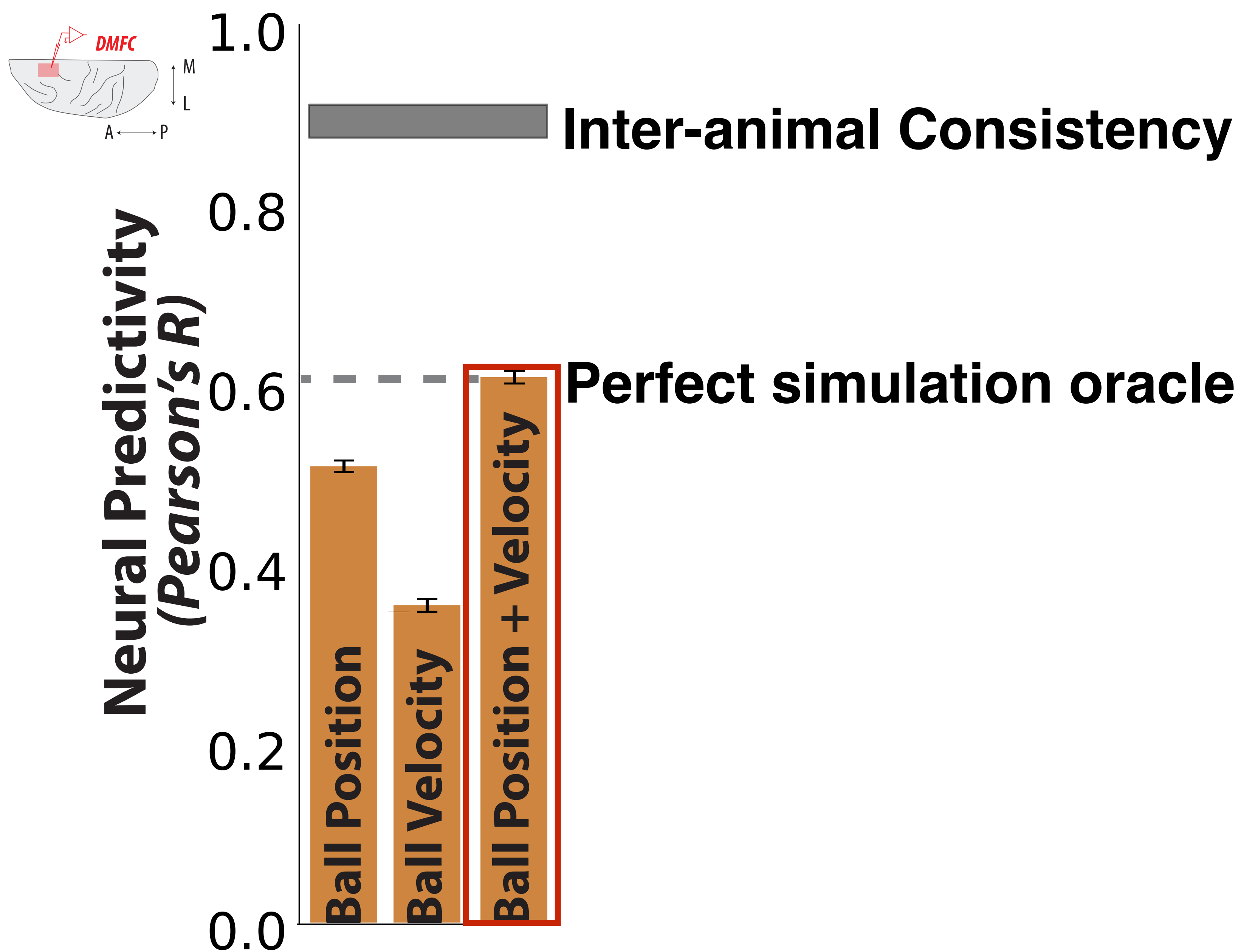


Future Directions

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

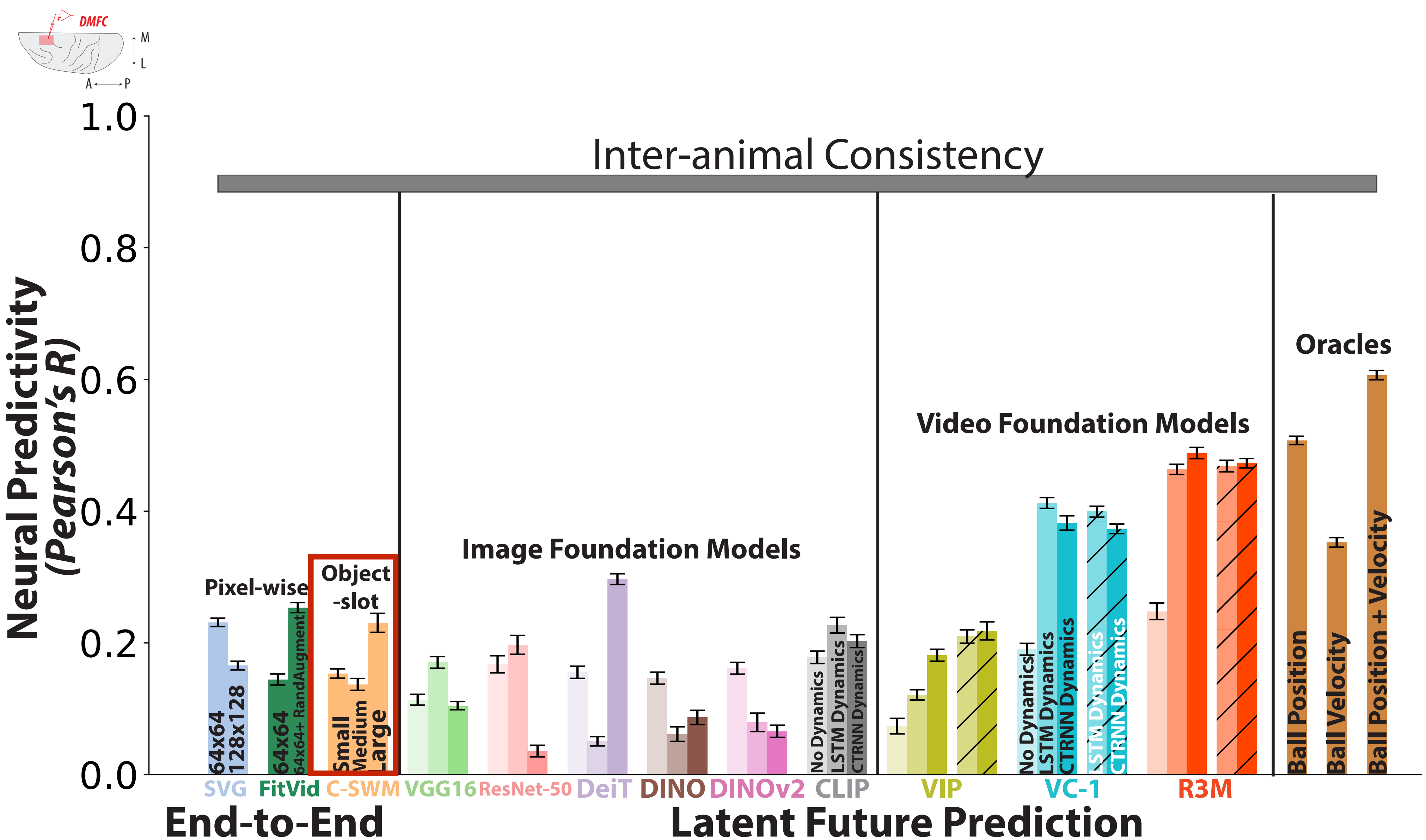
Future Directions

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



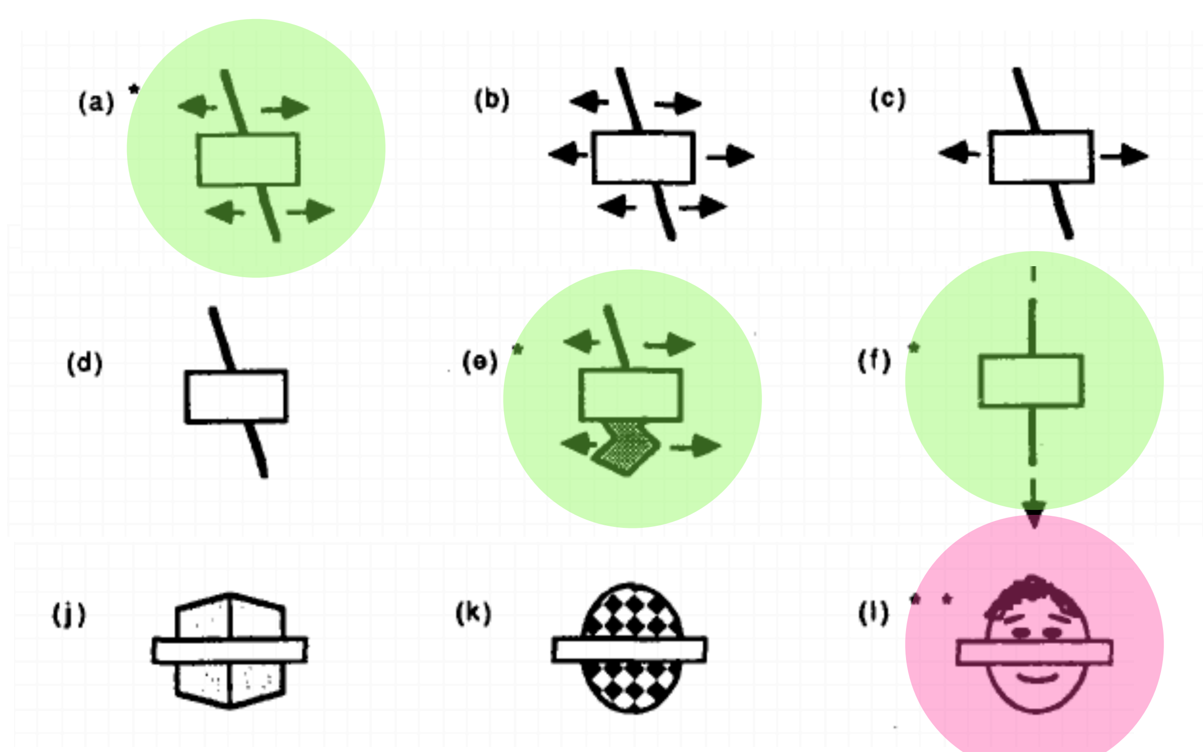
Future Directions

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



Future Directions

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



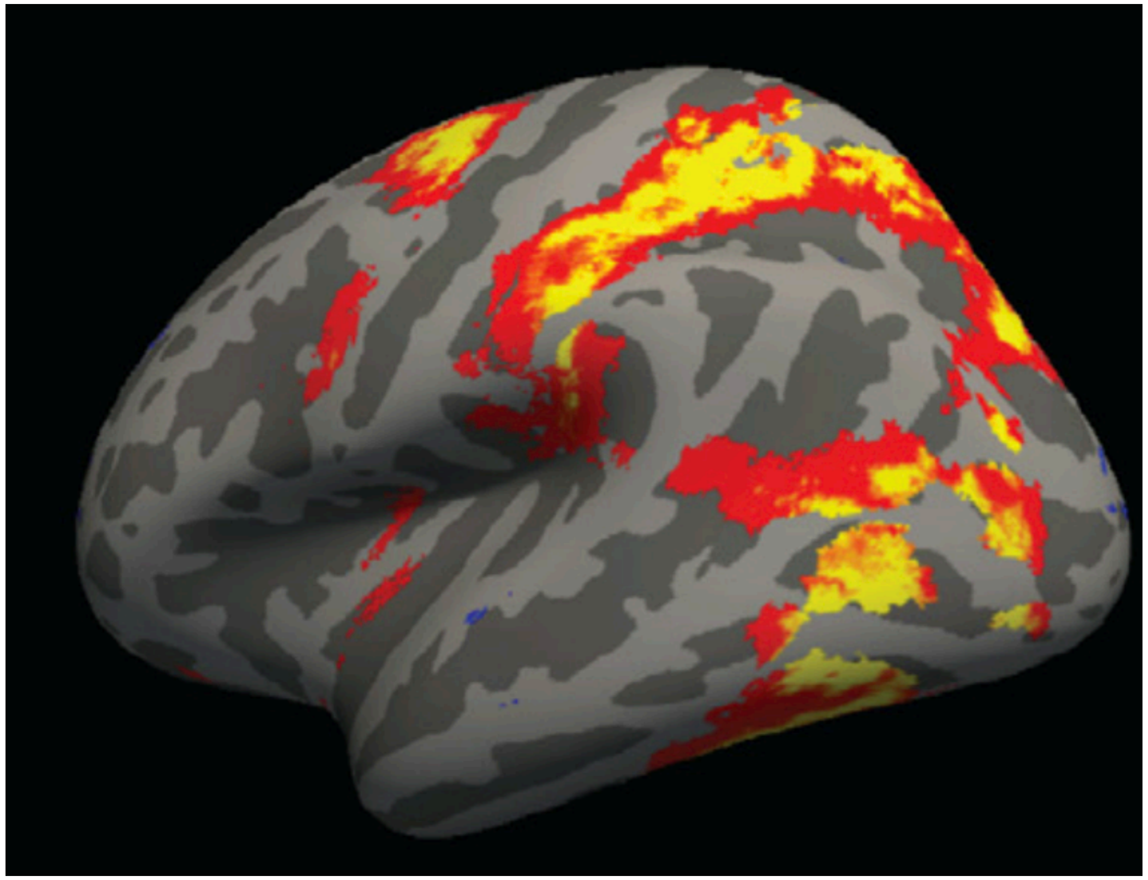
Principles of Object Perception Elizabeth Spelke, 1990

Elizabeth Spelke

Future Directions

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

2. **Cognitive:** Does the "physics engine" use a hierarchy of timescales to represent multiple possibilities?



Future Directions: Learning Diverse Material Properties

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

2. **Cognitive:** Does the "physics engine" use a hierarchy of timescales to represent multiple possibilities?

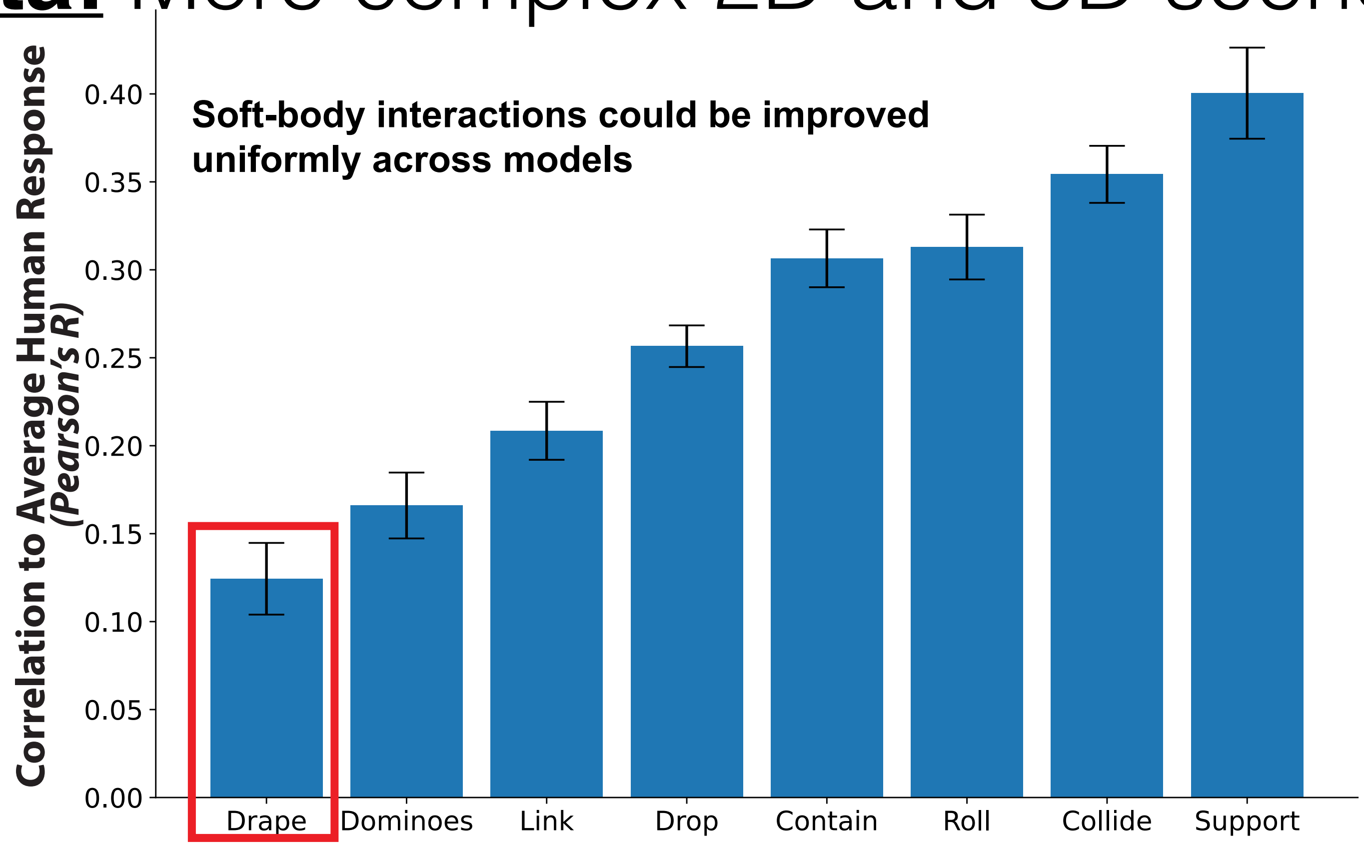
3. **Data:** More complex 2D and 3D scenes/real world objects

Future Directions: Learning Diverse Material Properties

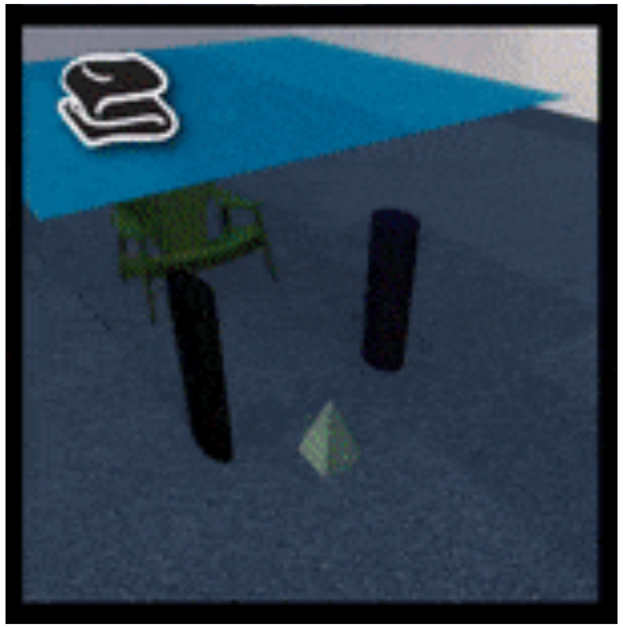
1. **<u>Sensory:</u>** Better leverage temporal relationships to build a more "factorized" *and* reusable representation: **object-centric, video foundation model**?

2. **<u>Cognitive:</u>** Does the "physics engine" use a hierarchy of timescales to represent multiple possibilities?

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



Drape



Takeaways

L = *learning rule*

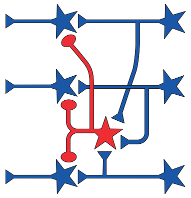
"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

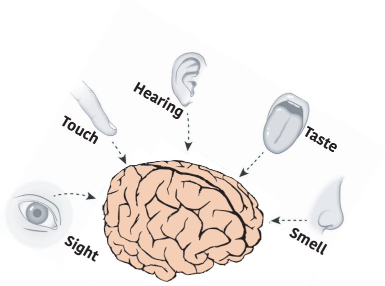
<u>Guiding Question:</u>

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



"Circuit"

A = *architecture class*



"Environment"

D = *data stream*

Takeaways

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

latent future prediction

<u>Guiding Question:</u>

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

SSL video foundation encoder + recurrent neural network

"Circuit"

A = *architecture class*

egocentric videos

"Environment"

D = *data stream*

Takeaways

Guiding Question:

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

Findings:

Mental simulation crucially involves explicit future prediction of a visual scene description.

Takeaways

Guiding Question:

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

Findings:

Mental simulation crucially involves explicit future prediction of a visual scene description.

This scene description is *not* fine-grained at the level of pixels, but must be "factorized" somehow.

Takeaways

<u>Guiding Question:</u>

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

<u>Findings:</u>

Mental simulation crucially involves explicit future prediction of a visual scene description.

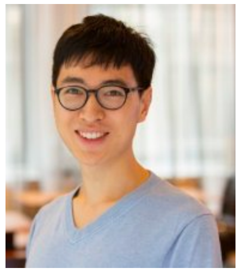
This scene description is *not* fine-grained at the level of pixels, but must be "factorized" somehow.

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

Acknowledgements


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

Contact:
anayebi@mit.edu
@aran_nayebi

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

NeurIPS 2023

K. LISA YANG ICoN CENTER Integrative Computational Neuroscience

McGOVERN INSTITUTE

brain+cognitive sciences

SIMONS FOUNDATION

THE HELEN HAY WHITNEY FOUNDATION

JazLab


YangLab