Neural Foundations of Mental Simulation: Future Prediction of Latent Representations on Dynamic Scenes

Aran Nayebi

In collaboration with:

Rishi Rajalingham, Mehrdad Jazayeri, and Guangyu Robert Yang

BCS/MIBR/PILM Retreat

2023.06.04

Motivation


氷みくじ
氷に張りつきます
氷みくじ
氷に張りつきます


WHOLE WHEAT
Mueller
PRO-SERIES
MULTI
LODGE
CAST IRON
LOVE


NEW
Specialty Drinks
CHOKING
NO SMOKING
5094M
F 32803

Motivation

Infer:
Has this ice block been out longer?

Motivation

Infer:
Has this ice block been out longer?

Predict:
Are these stacks stable?

Motivation

Infer:
Has this ice block been out longer?

Predict:
Are these stacks stable?

Predict:
Will this box support me?

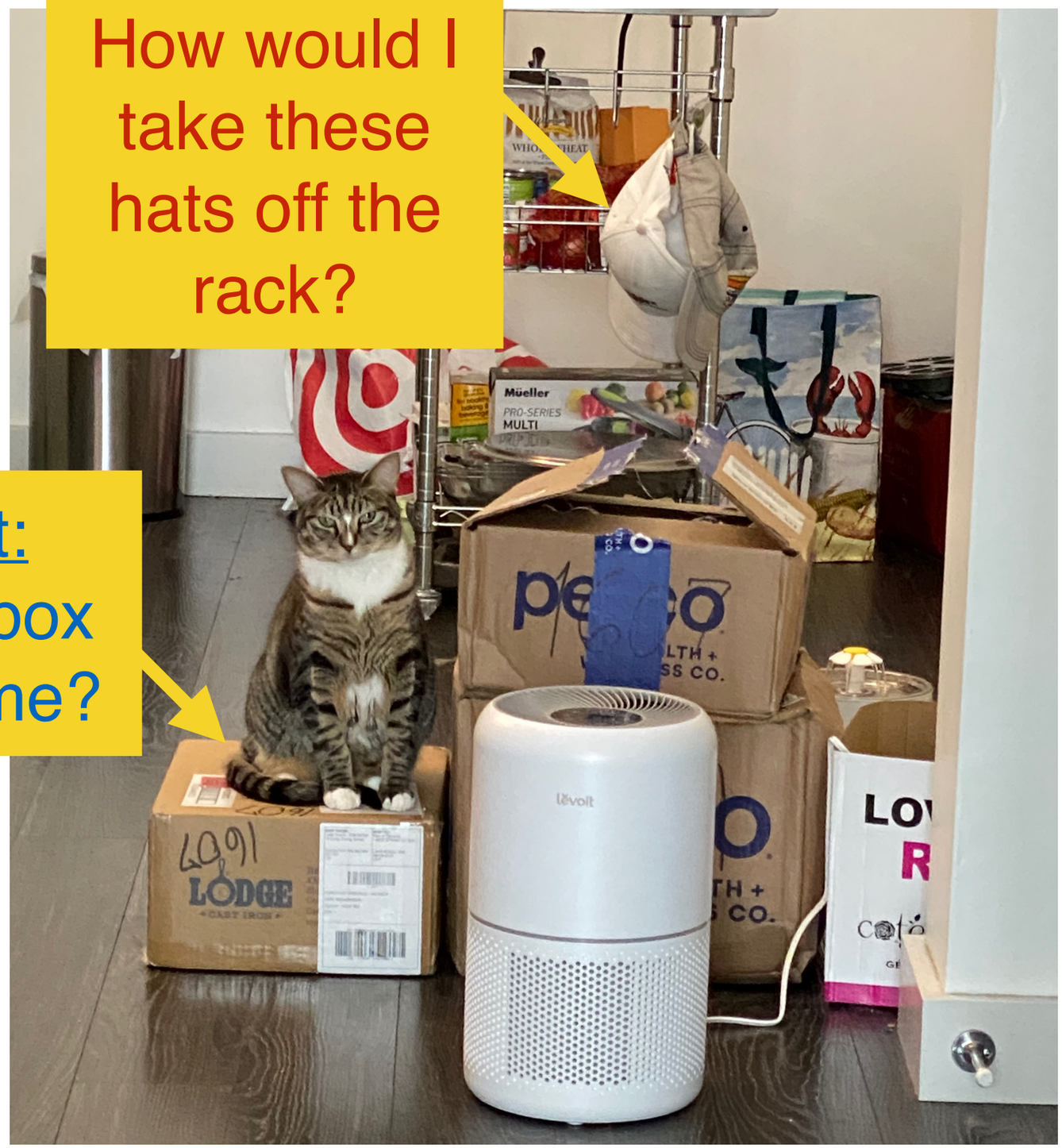
Motivation

Infer:
Has this ice block been out longer?

Predict:
Are these stacks stable?

Predict:
Will this box support me?

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



The Mental Simulation Hypothesis

The Nature of Explanation

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

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

Craik (1943)

Kenneth Craik

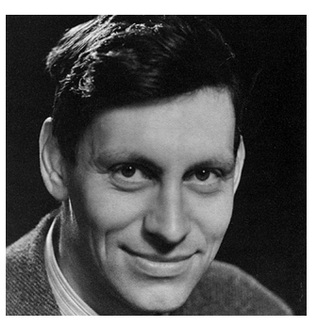
The Mental Simulation Hypothesis

The Nature of Explanation

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

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

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

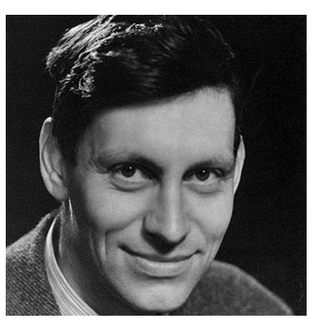
The Nature of Explanation

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

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

⟵ Pre-dates the modern computer!

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



Kenneth Craik

The Mental Simulation Hypothesis: Behavioral Evidence

The Nature of Explanation

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

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

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

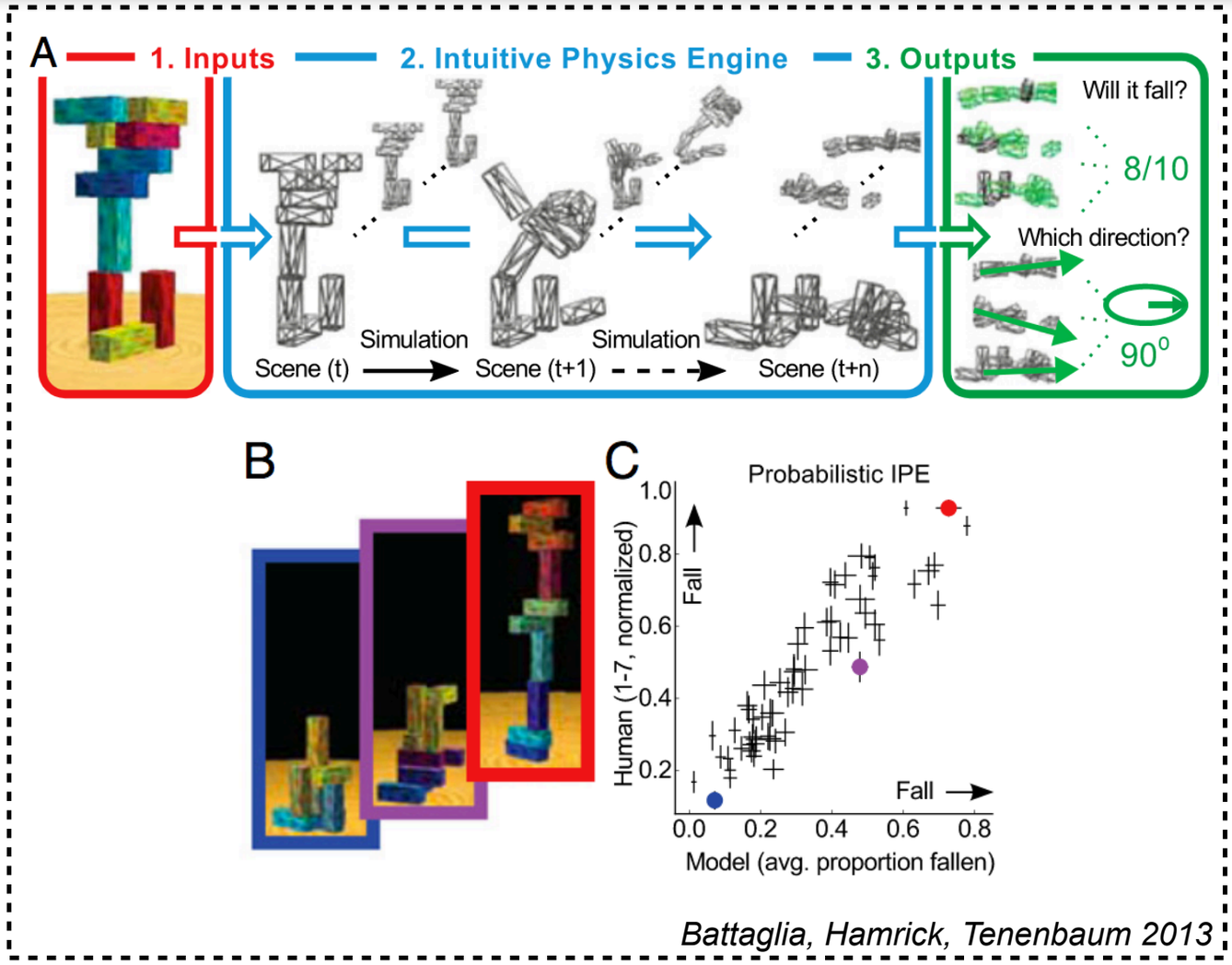
The Mental Simulation Hypothesis: Behavioral Evidence

The Nature of Explanation

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

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

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



Battaglia, Hamrick, Tenenbaum 2013

Intuitive Physics Engine (IPE) can match human physical judgements

Peter Battaglia

Jessica Hamrick

Joshua Tenenbaum

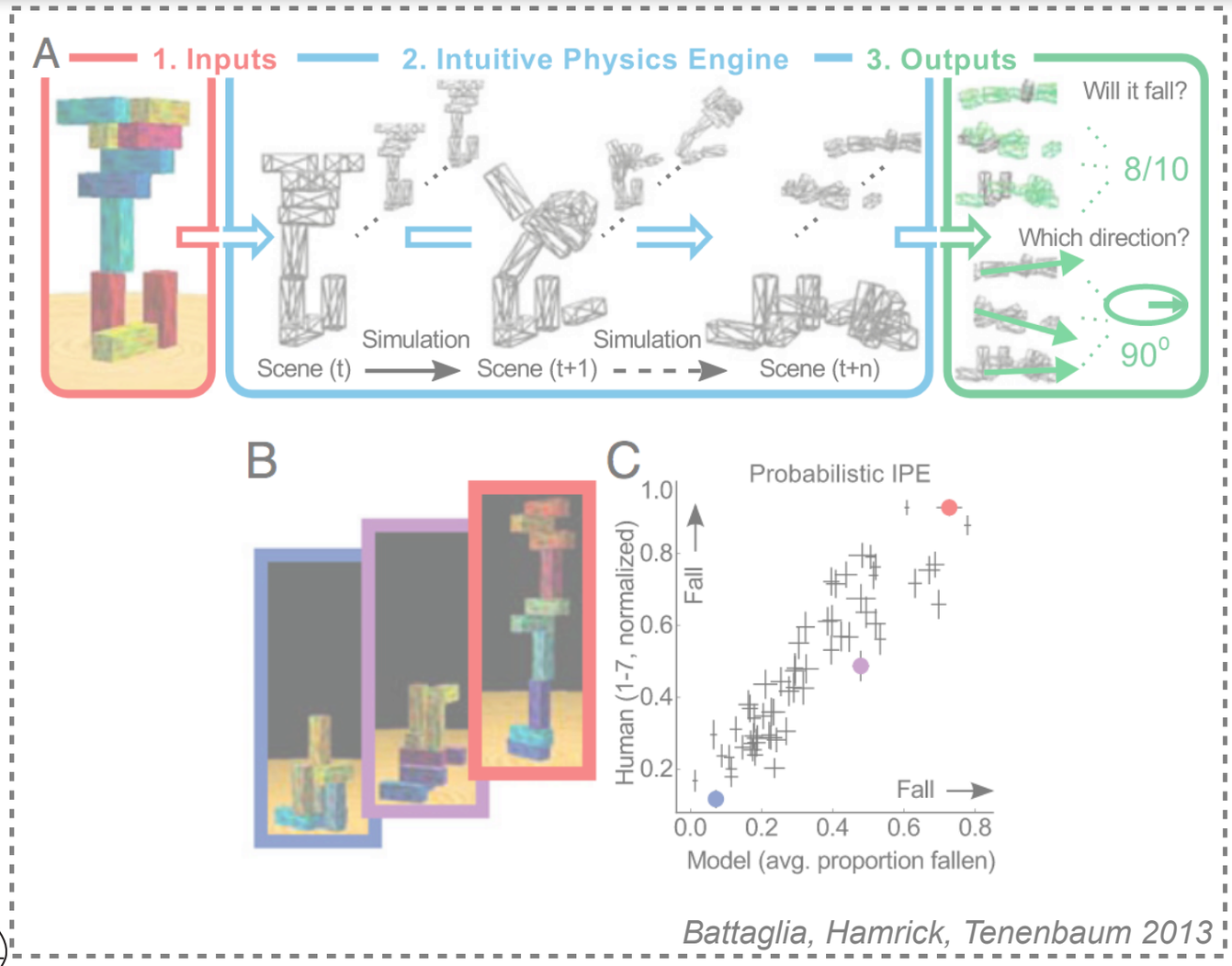
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

The Nature of Explanation

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

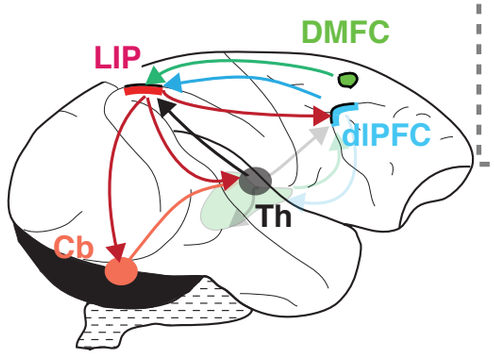
If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present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

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

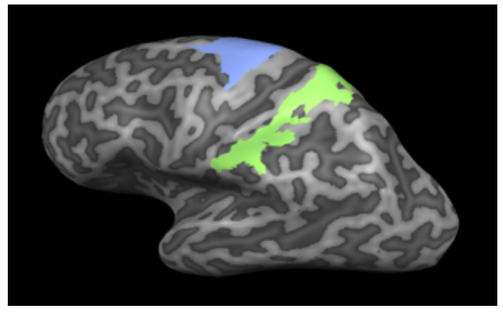


Battaglia, Hamrick, Tenenbaum 2013

The Brain's "Physics Engine"



Fronto-Parietal Network




Joshua Tenenbaum


Nancy Kanwisher

The Mental Simulation Hypothesis: Human Neuroimaging Evidence

The Nature of Explanation

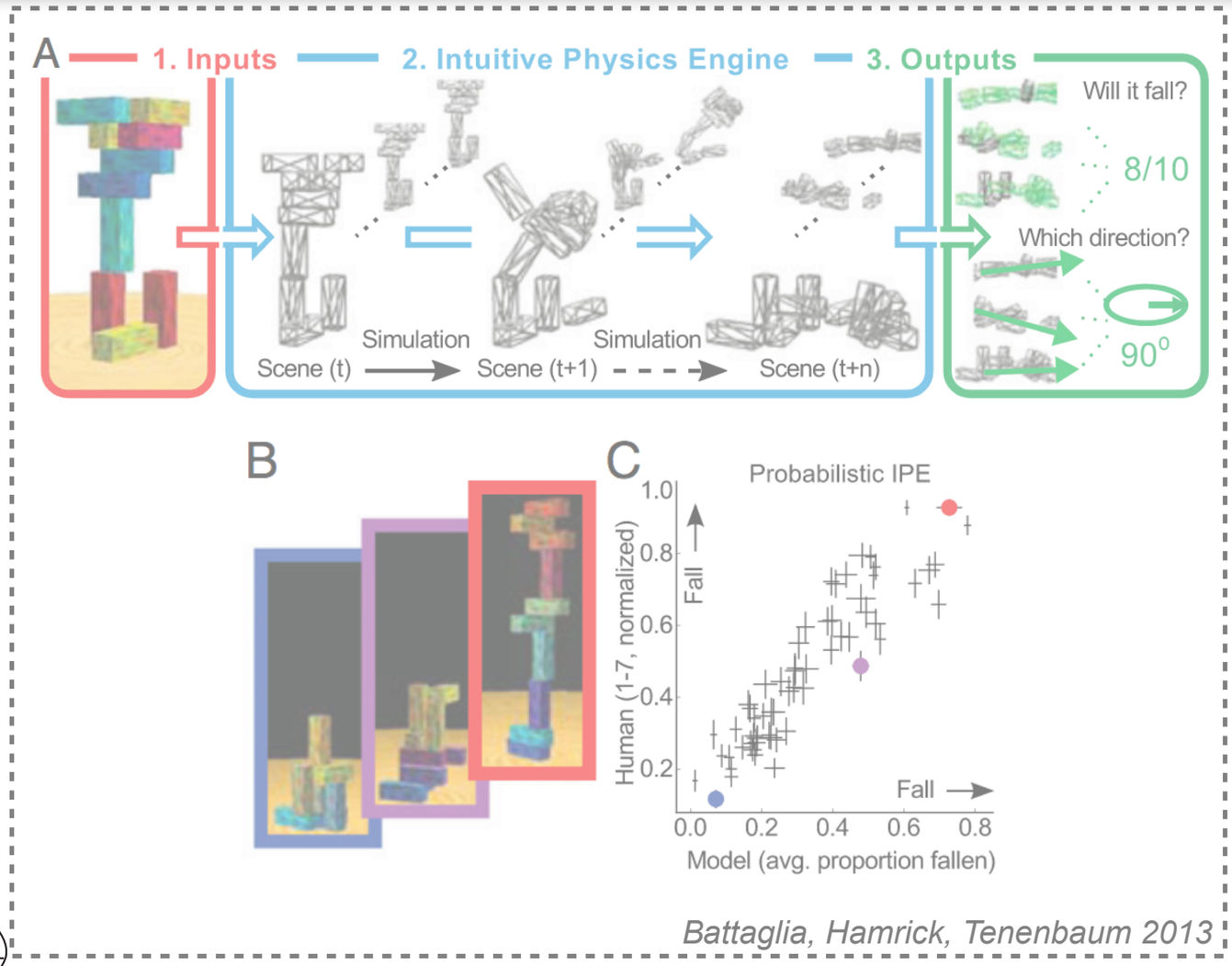
My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation.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

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

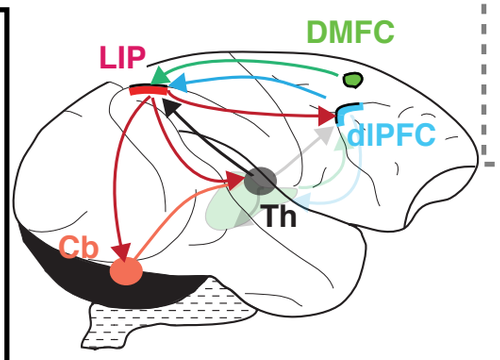
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**

The Brain's "Physics Engine"

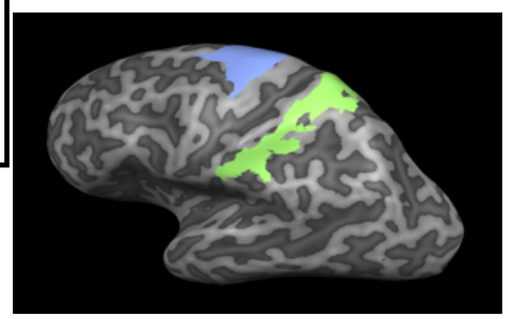
- A network of brain regions recruited by physical inferences (*Fischer et al. 2016*)

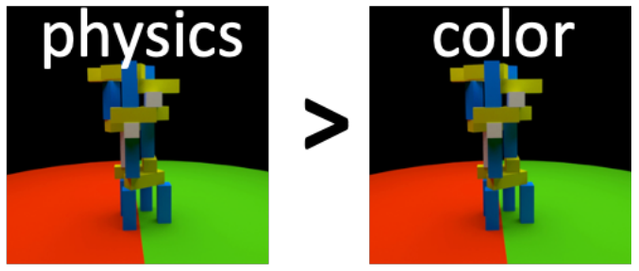


Battaglia, Hamrick, Tenenbaum 2013



Fronto-Parietal Network





Fischer et al. 2016

Joshua Tenenbaum

Nancy Kanwisher

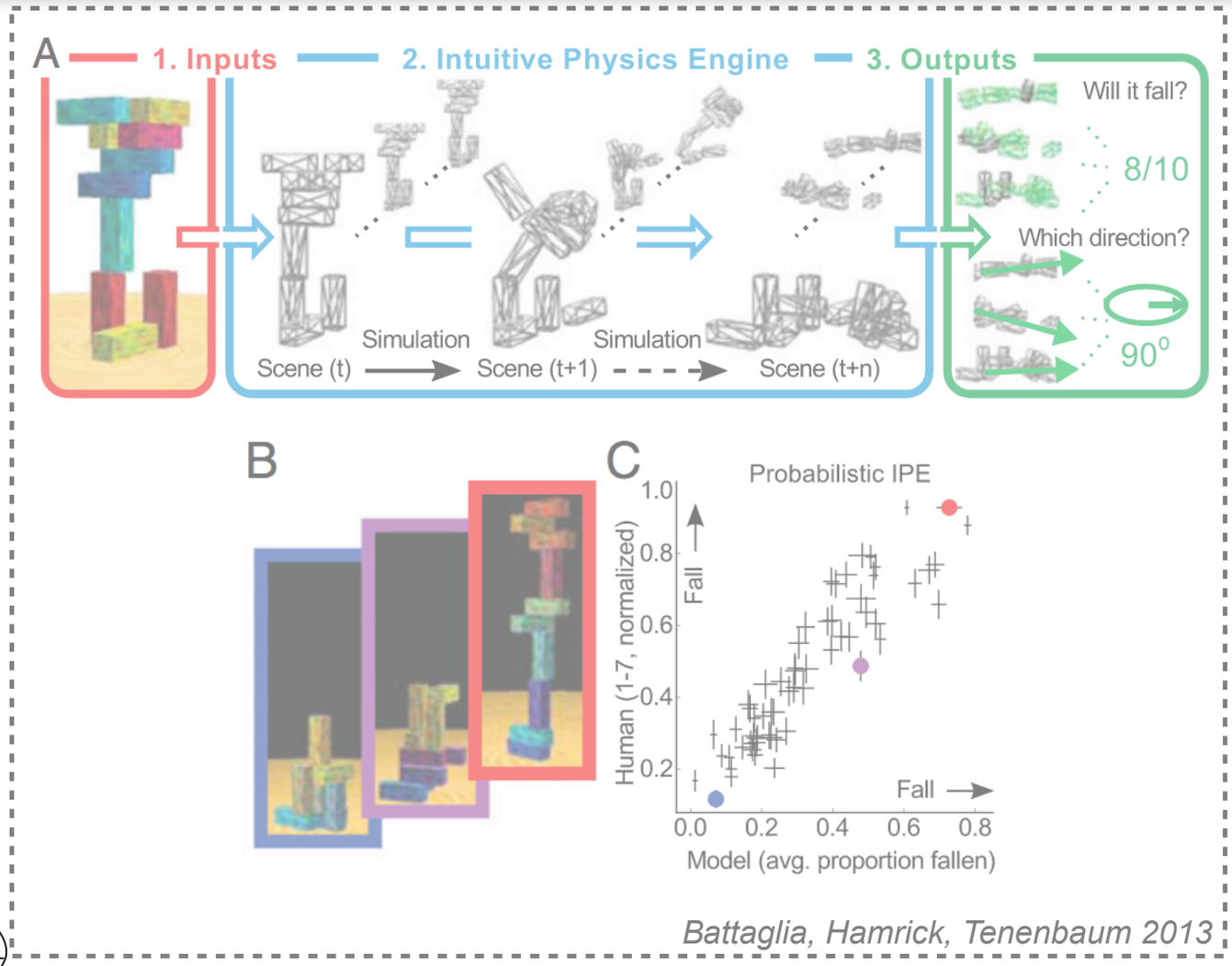
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

The Nature of Explanation

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

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

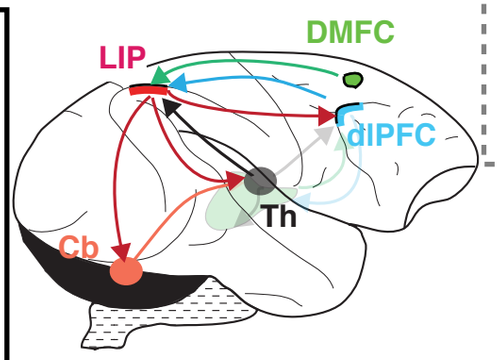
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



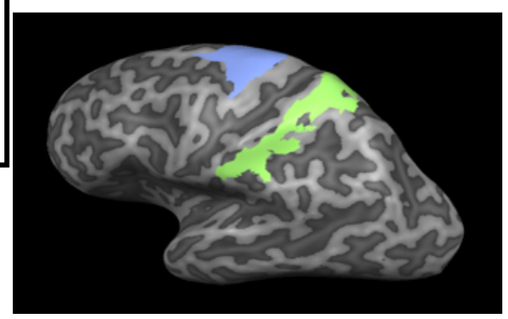
Battaglia, Hamrick, Tenenbaum 2013

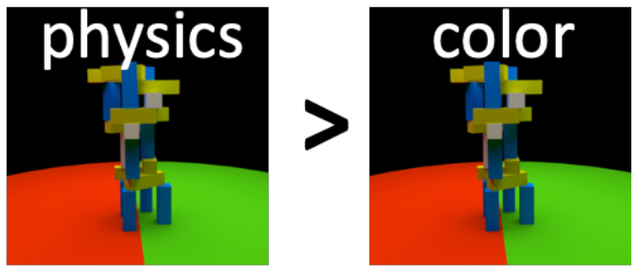
The Brain's "Physics Engine"

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

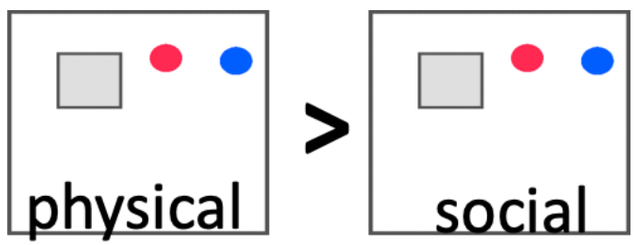


Fronto-Parietal Network





Fischer et al. 2016



Schwettmann et al. 2019

Joshua Tenenbaum

Nancy Kanwisher

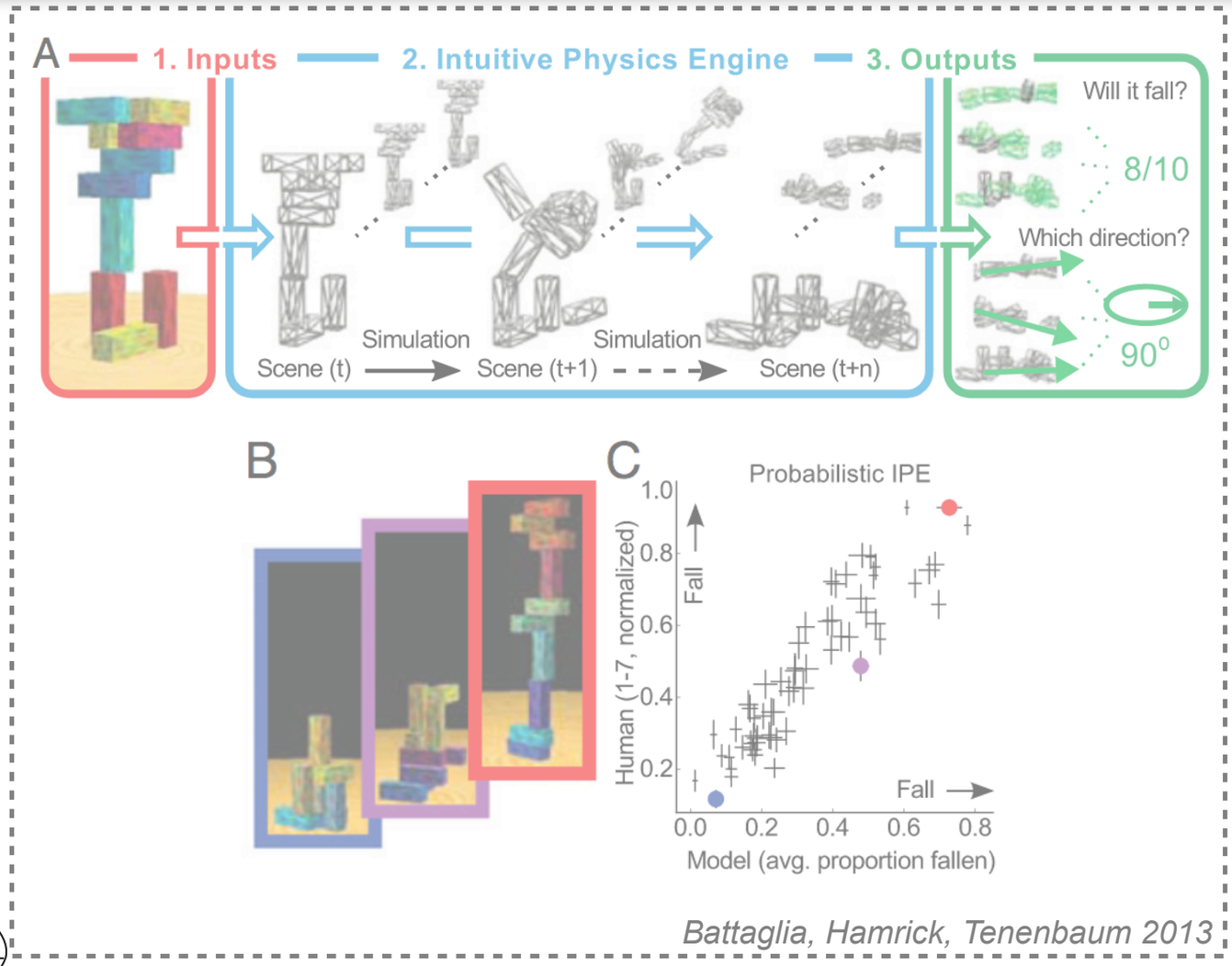
The Mental Simulation Hypothesis: Human Neuroimaging Evidence

The Nature of Explanation

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

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

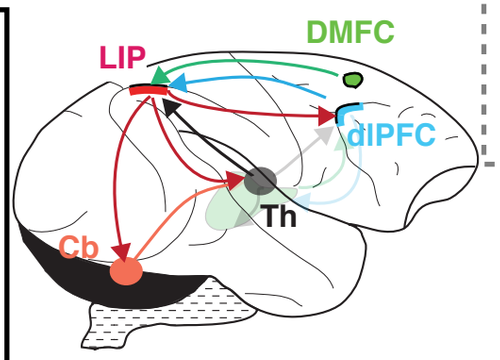
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



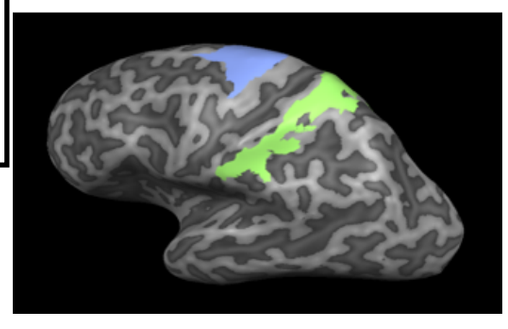
Battaglia, Hamrick, Tenenbaum 2013

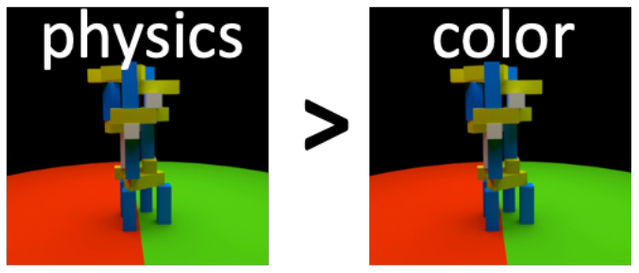
The Brain's "Physics Engine"

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

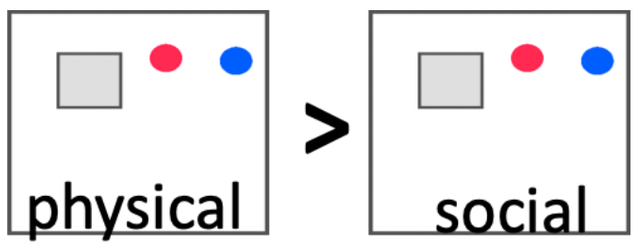


Fronto-Parietal Network





Fischer et al. 2016



Schwettmann et al. 2019

Pramod et al. 2022

Joshua Tenenbaum

Nancy Kanwisher

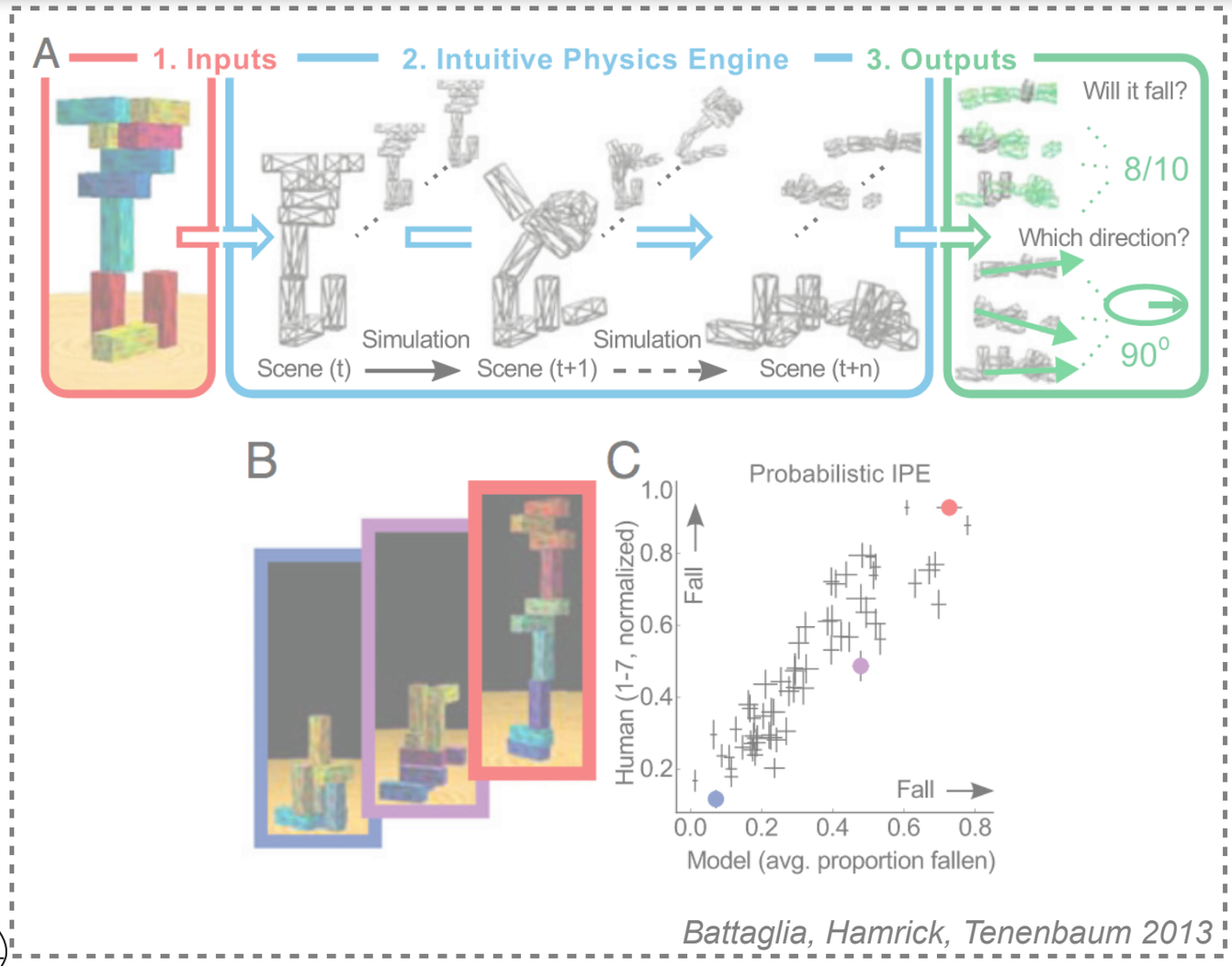
The Mental Simulation Hypothesis: Primate Electrophysiological Evidence

The Nature of Explanation

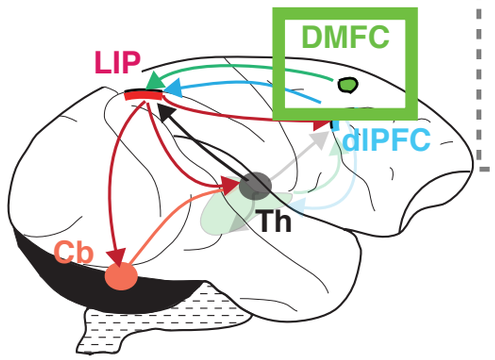
My hypothesis then is that thought models, or parallels, reality – that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation.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

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

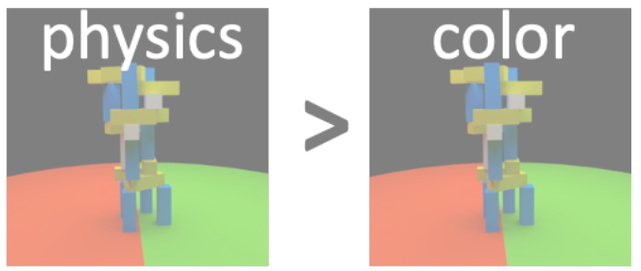
Craik (1943): The brain builds **mental models** of the external physical world, that support physical inferences via **mental simulations.**



Battaglia, Hamrick, Tenenbaum 2013

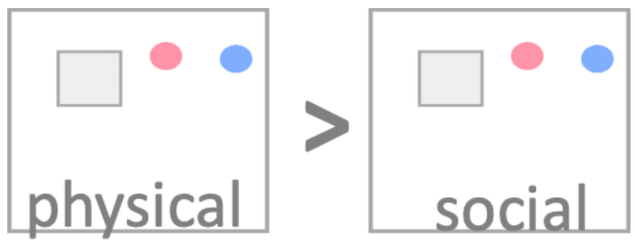


Fronto-Parietal Network

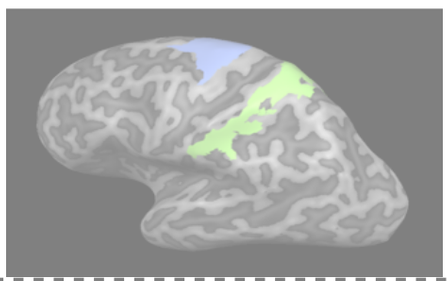


Fischer et al. 2016

Pramod et al. 2022



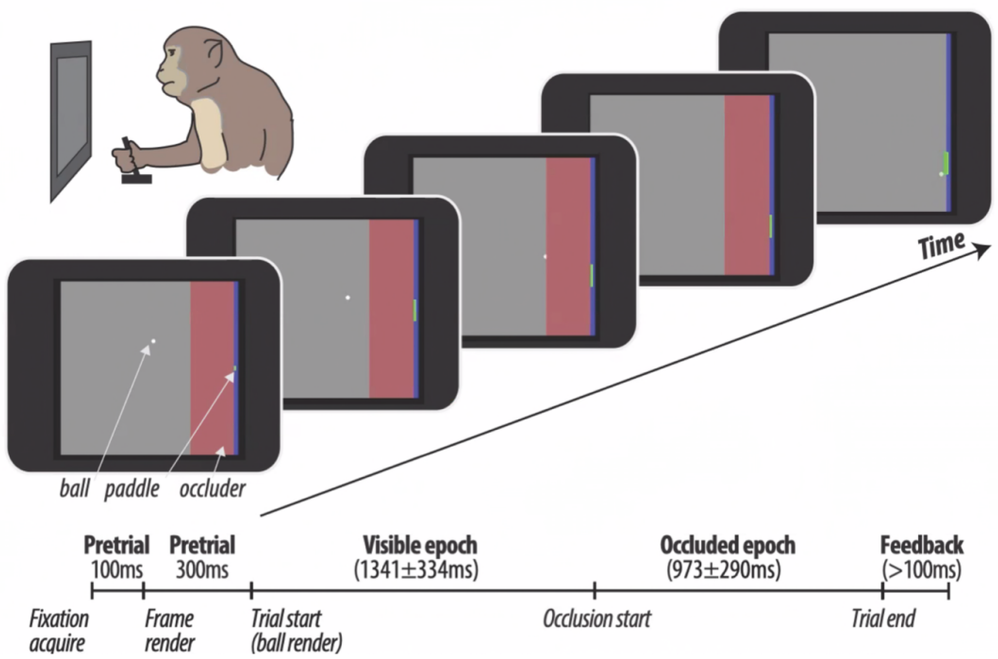
Schwettmann et al. 2019



Rishi Rajalingham

Mehrdad Jazayeri

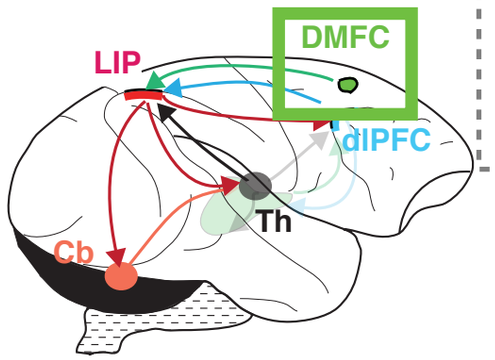
The Mental Simulation Hypothesis: Primate Electrophysiological Evidence



The role of mental simulation in primate physical inference abilities

Rishi Rajalingham, Aida Piccato, Mehrdad Jazayeri

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



Fronto-Parietal Network

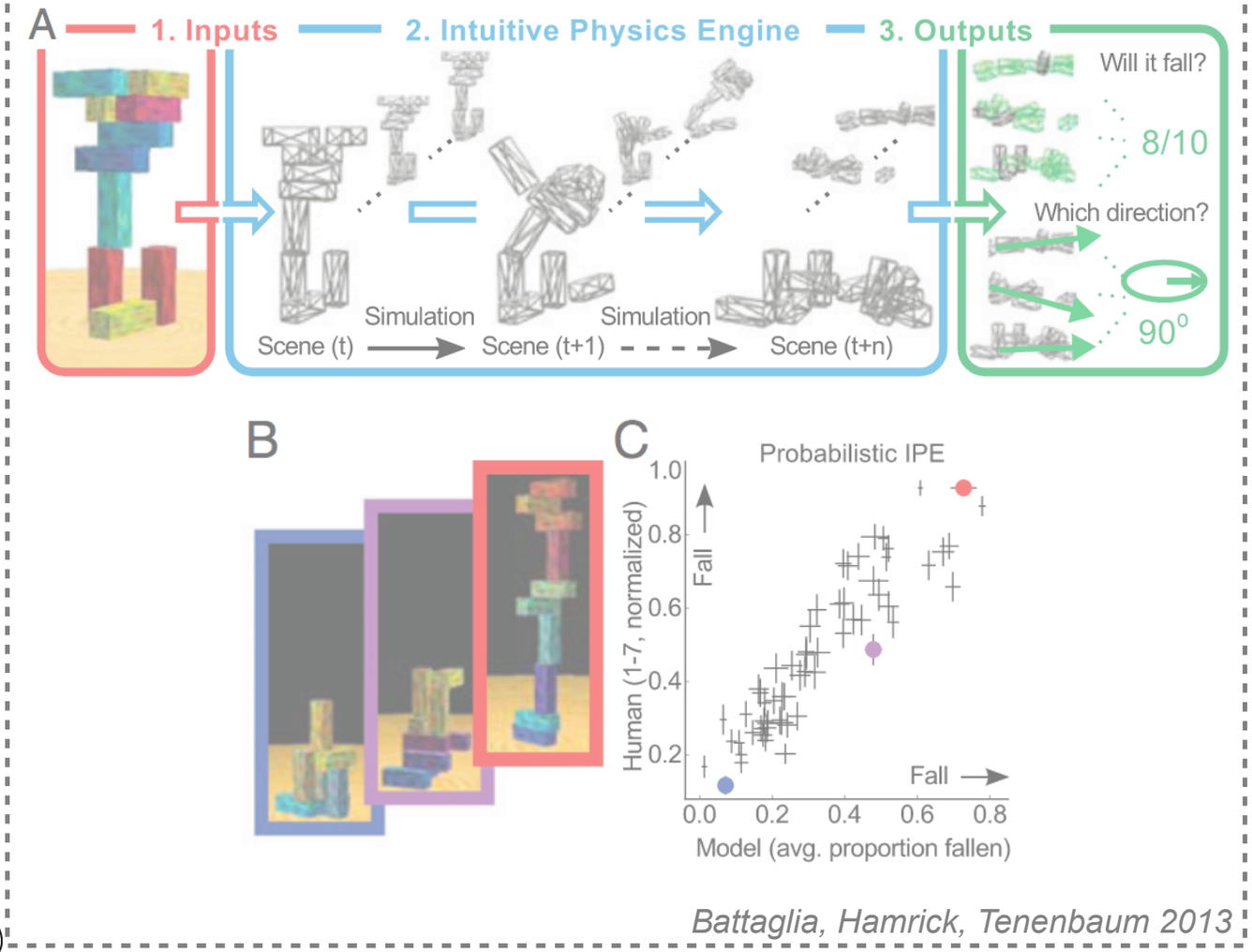
Dynamic tracking of objects in the macaque dorsomedial frontal cortex

Rishi Rajalingham, Hansem Sohn, Mehrdad Jazayeri

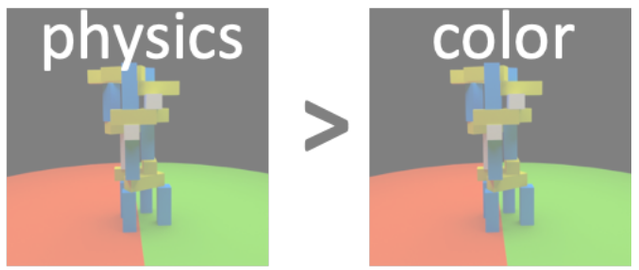
doi: https://doi.org/10.1101/2022.06.24.497529

Rishi Rajalingham

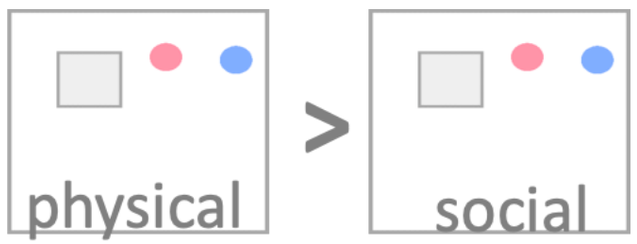
Mehrdad Jazayeri



Battaglia, Hamrick, Tenenbaum 2013

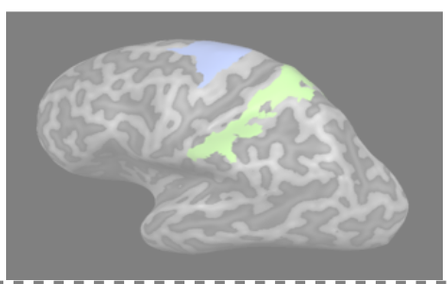


Fischer et al. 2016

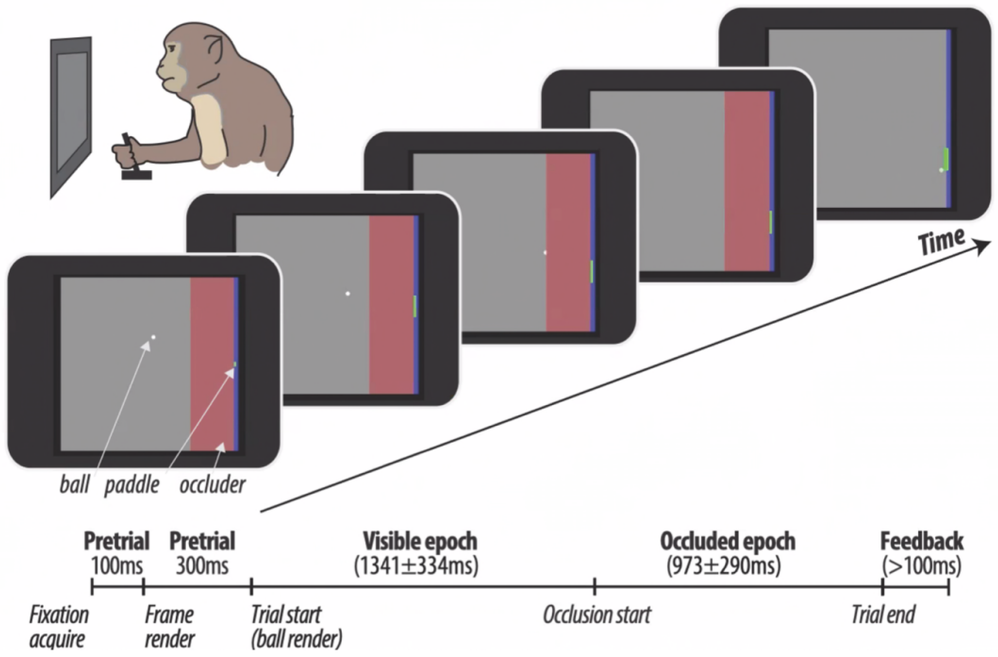


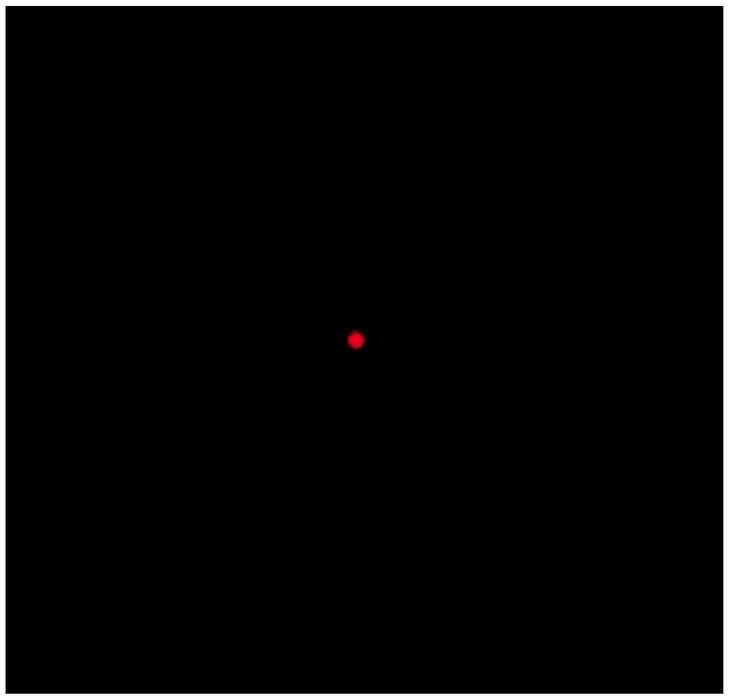
Schwettmann et al. 2019

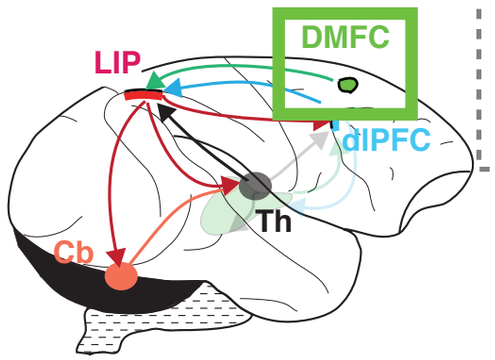
Pramod et al. 2022



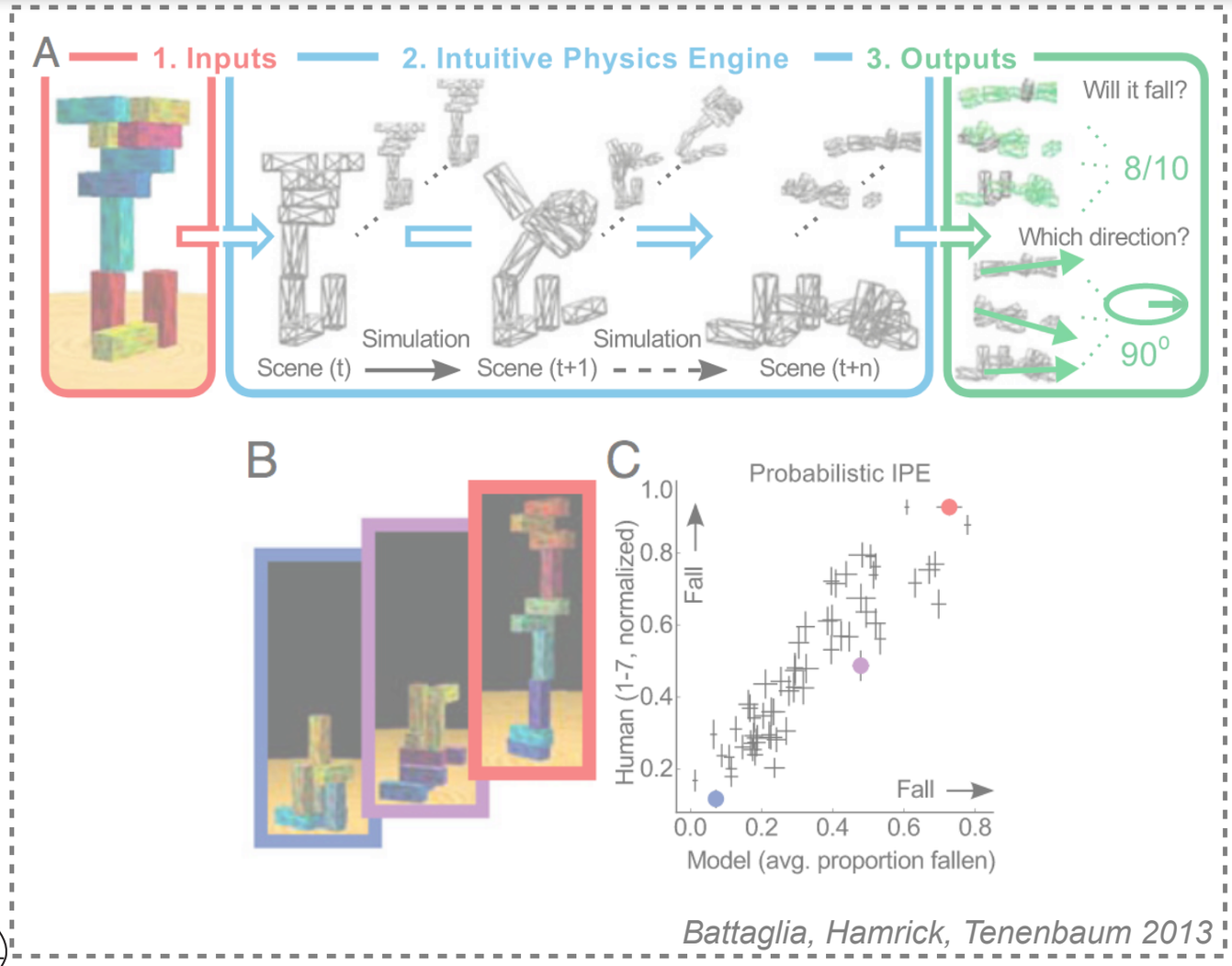
The Mental Simulation Hypothesis: Primate Electrophysiological Evidence



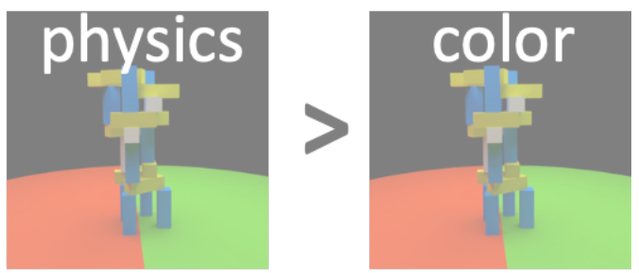




Fronto-Parietal Network

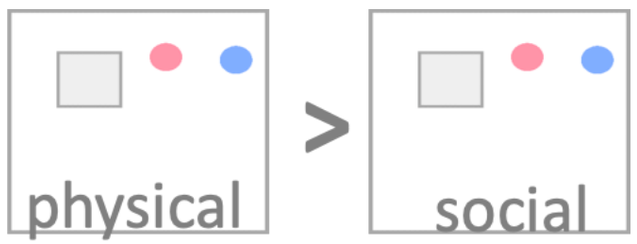


Battaglia, Hamrick, Tenenbaum 2013

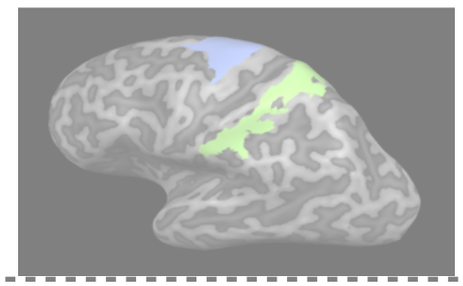


Fischer et al. 2016

Pramod et al. 2022



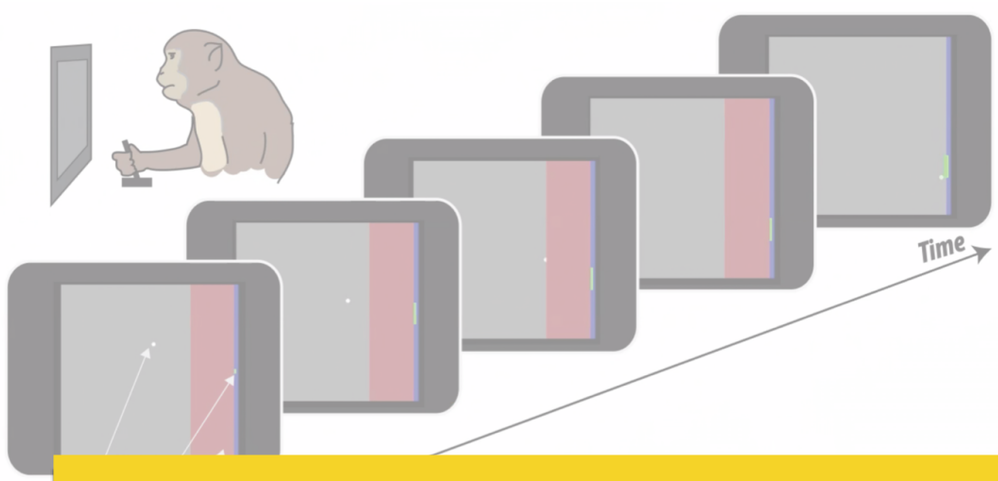
Schwettmann et al. 2019

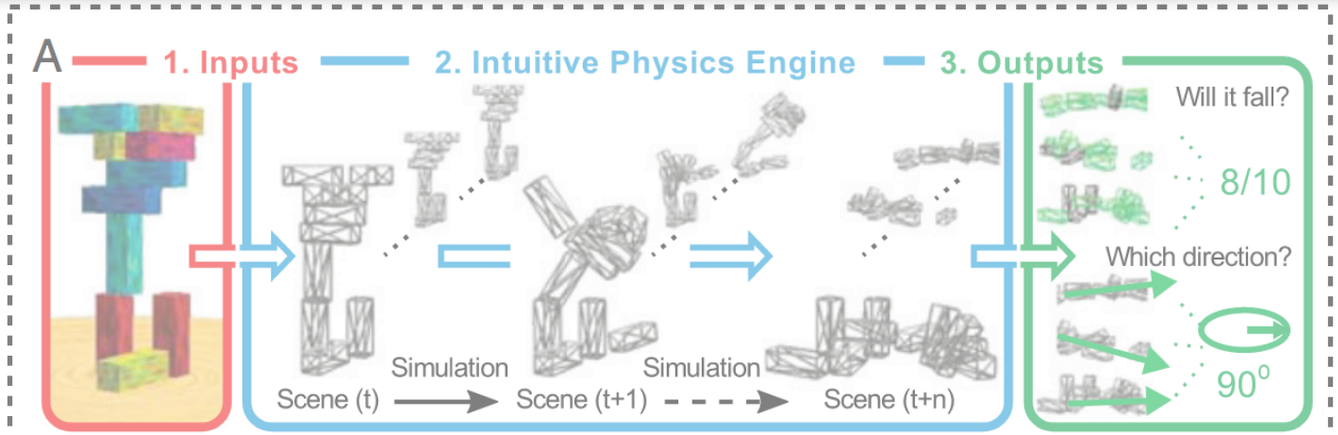


Rishi Rajalingham

Mehrdad Jazayeri

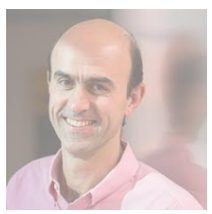
Neural Mechanisms of Mental Simulation





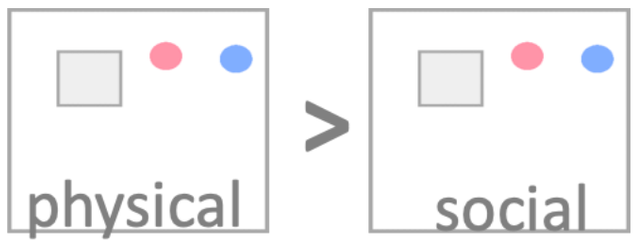
Crux question: What are the neural mechanisms that enable the brain's "simulation-like" computations *across* environments?

Rishi Rajalingham

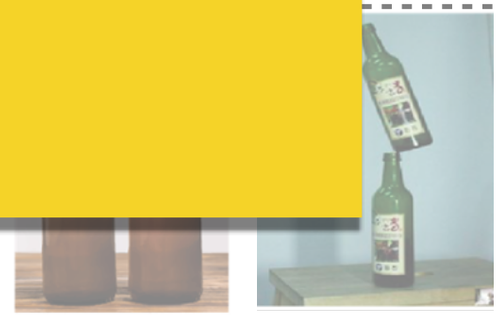


Mehrdad Jazayeri

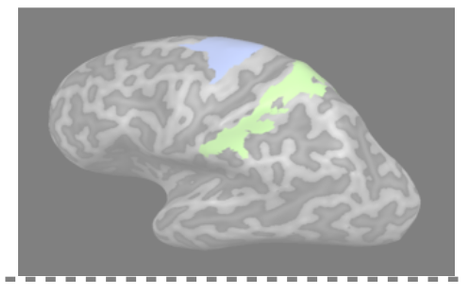
Fischer et al. 2016



Schwettmann et al. 2019



Pramod et al. 2022



Defining Hypotheses

Defining Hypotheses

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

Defining Hypotheses

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

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

Defining Hypotheses

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

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

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

Defining Hypotheses

"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

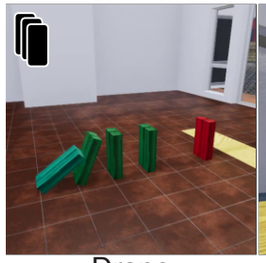
Overall Approach: Training Datasets

(A) *Model Pretraining*

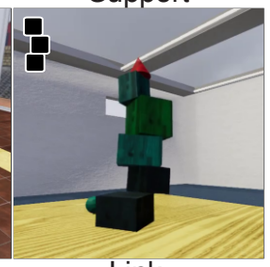
Inputs

Physion

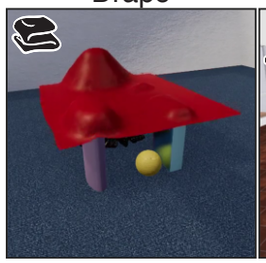
Dominoes



Support



Drape

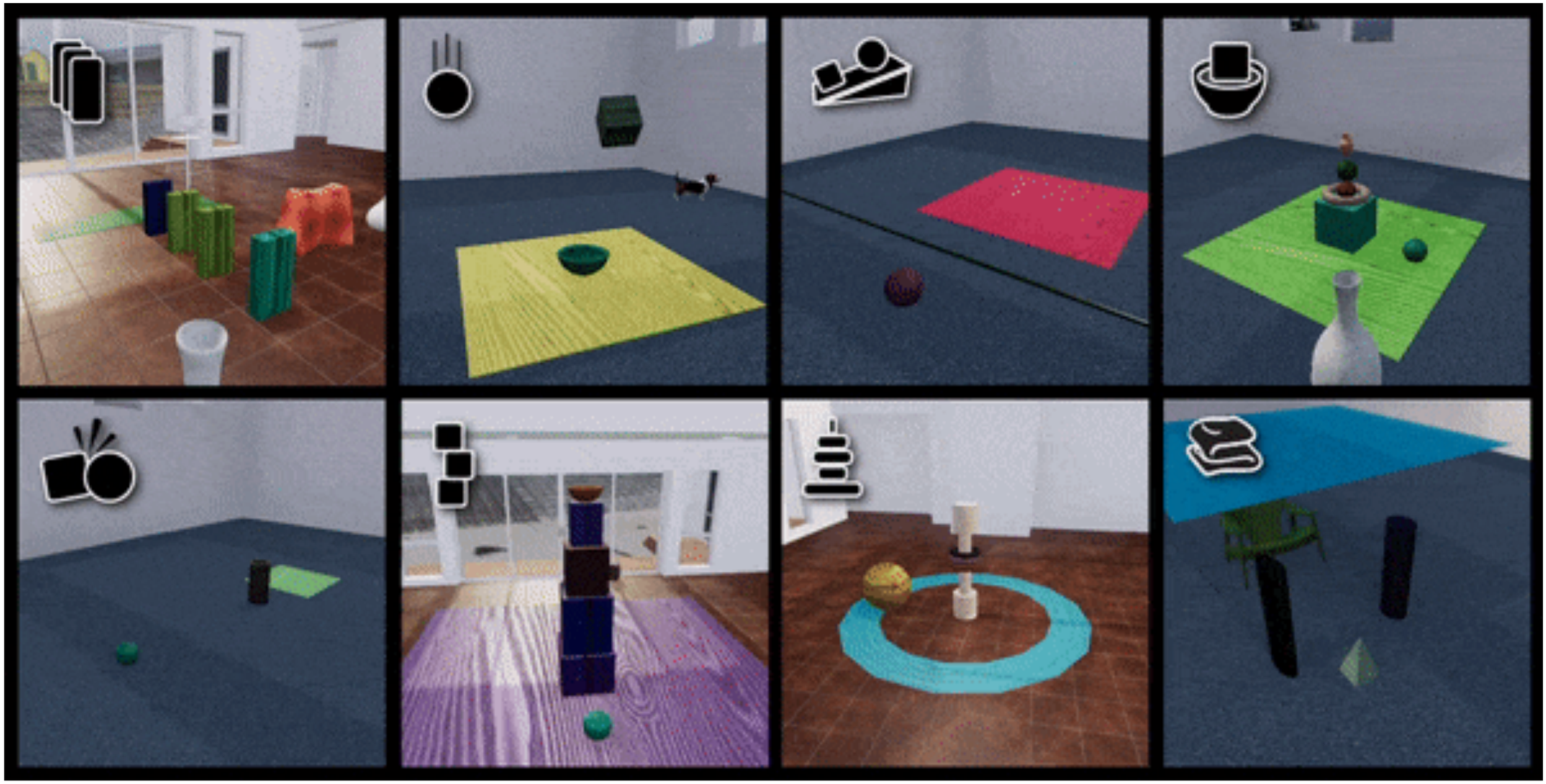


Link

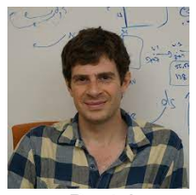
Overall Approach: Training Datasets

Physion/ThreeD World (TDW)

Bear et al. 2021



Focus on everyday physical understanding



Daniel Bear

Joshua Tenenbaum

Daniel Yamins

Judith Fan

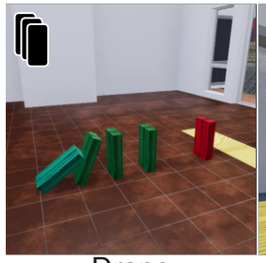
Overall Approach: Sensory-Cognitive Hypotheses

(A) *Model Pretraining*

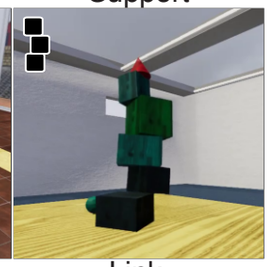
Inputs

Physion

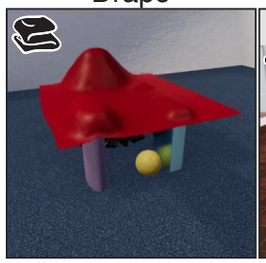
Dominoes



Support



Drape



Link

Sensory-Cognitive Hypothesis Classes

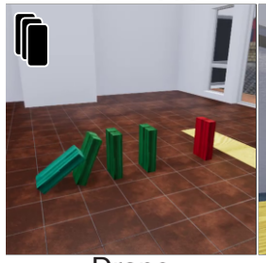
Overall Approach: Pixel-wise Future Prediction

(A) *Model Pretraining*

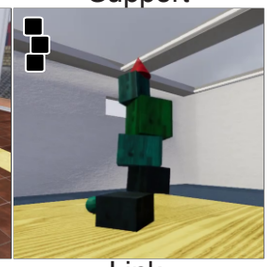
Inputs

Physion

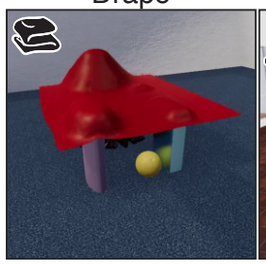
Dominoes



Support



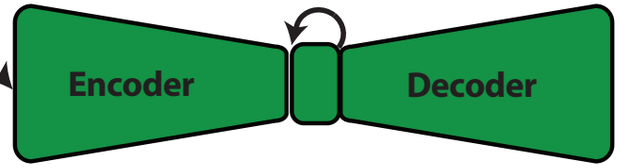
Drape



Link

Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



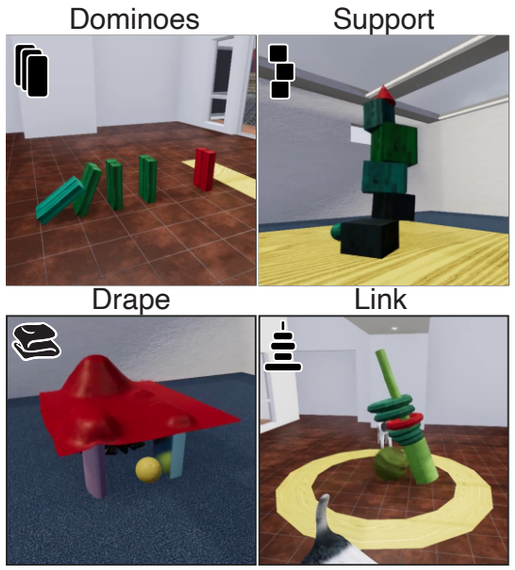
Pixel-wise

Overall Approach: Pixel-wise Future Prediction

(A) *Model Pretraining*

Inputs

Physion

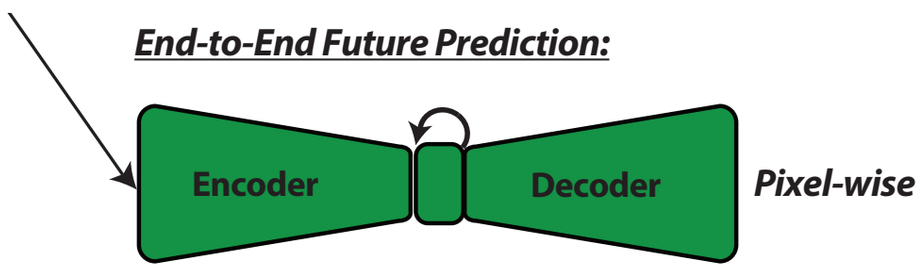


Sensory-Cognitive Hypothesis Classes

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

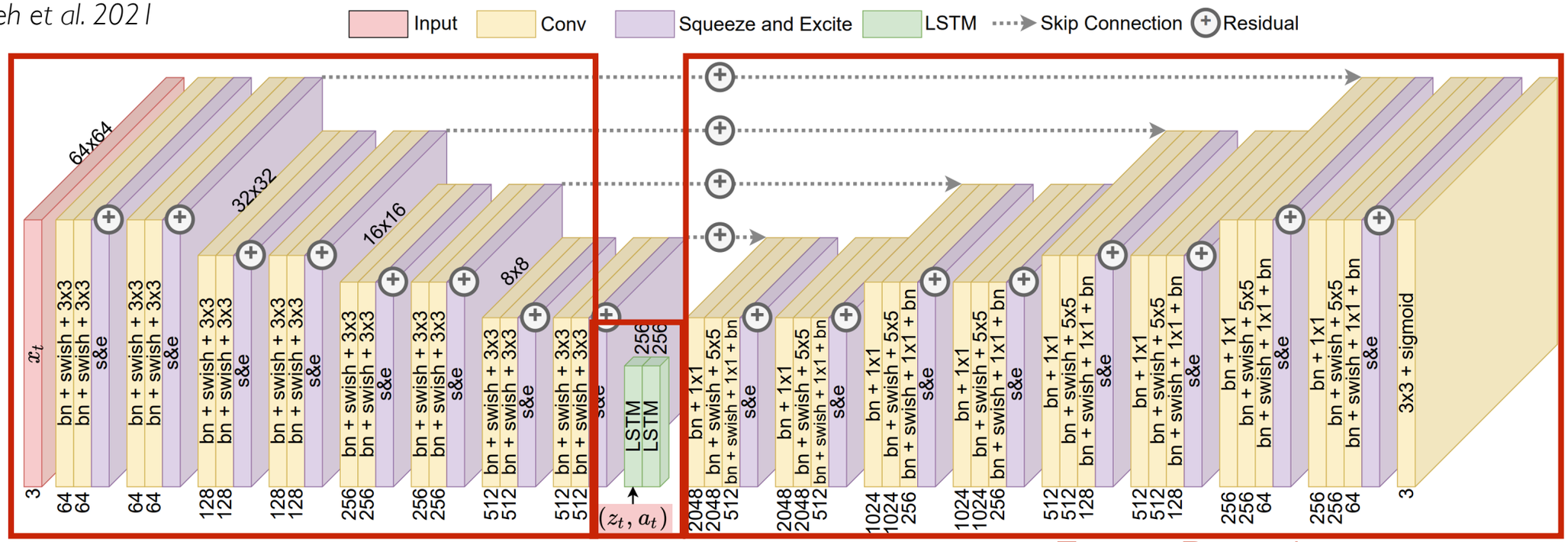
Predicts the future at the resolution of the sensory input (very detailed!)

End-to-End Future Prediction:



Pixel-wise

Babaeizadeh et al. 2021



Visual Encoder ("Sensory")

Dynamics Predictor ("Cognitive")

Frame Decoder ("Objective/Behavior")

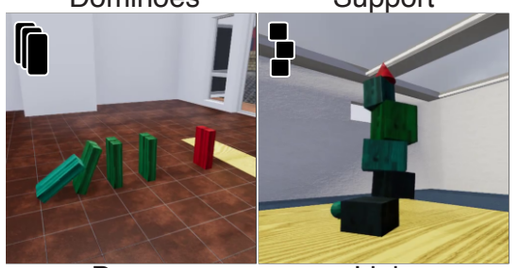
Overall Approach: Structured World Models

(A) *Model Pretraining*

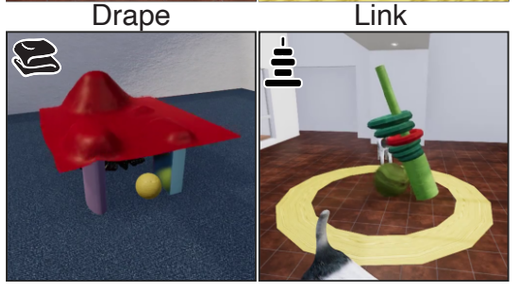
Inputs

Physion

Dominoes | Support

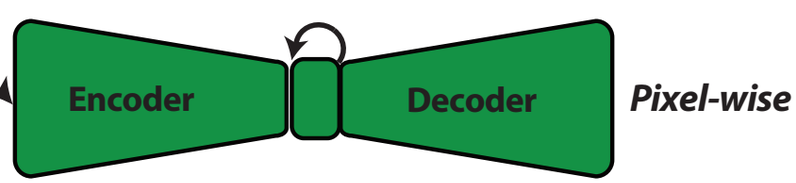


Drape | Link

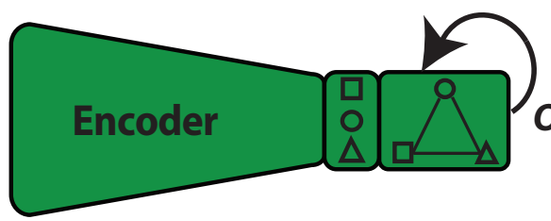


Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



Pixel-wise



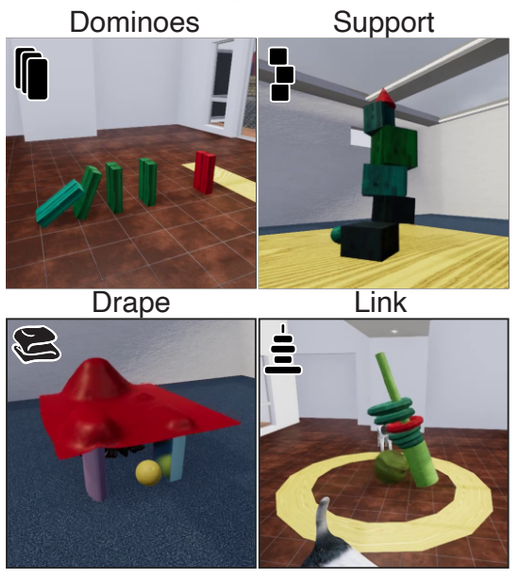
Object-slot

Overall Approach: Structured World Models

(A) *Model Pretraining*

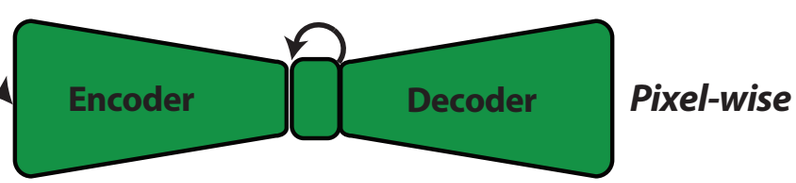
Inputs

Physion

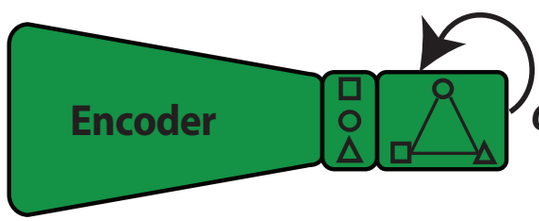


Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:

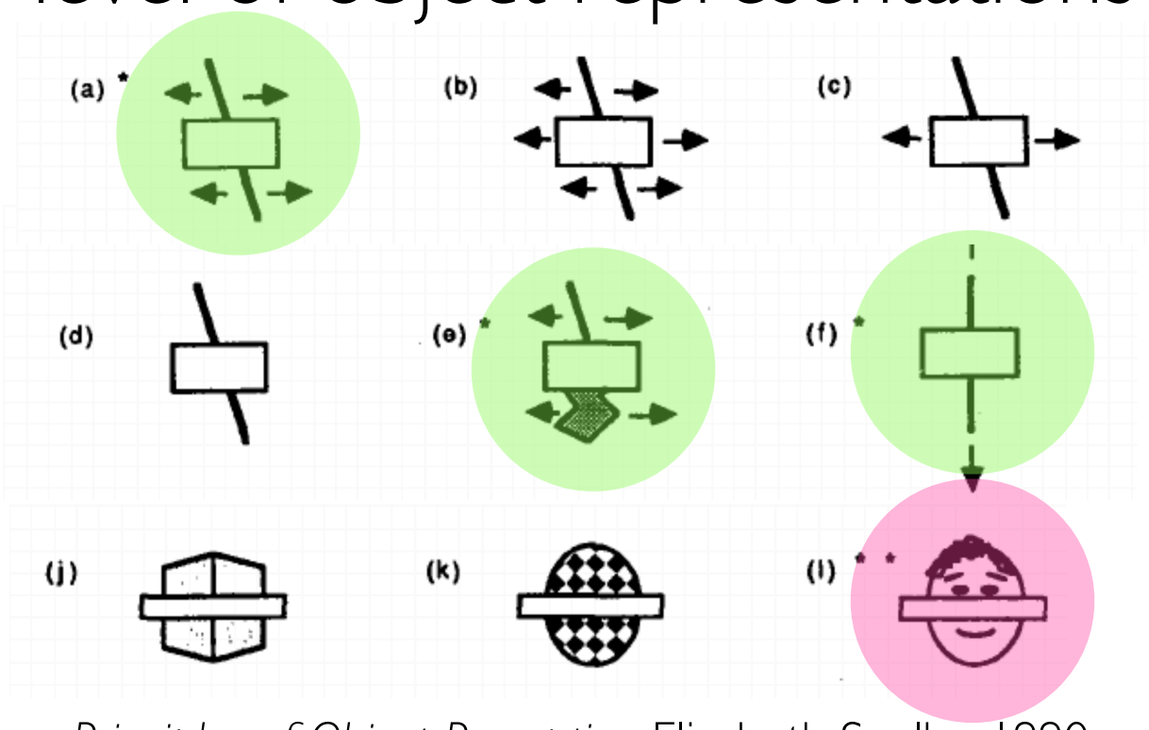


Pixel-wise



Object-slot

Predicts at the level of object representations and their relations



Principles of Object Perception Elizabeth Spelke, 1990

Elizabeth Spelke

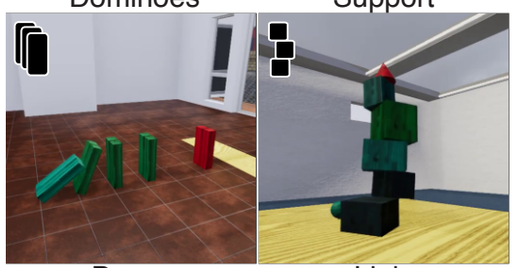
Overall Approach: Structured World Models

(A) *Model Pretraining*

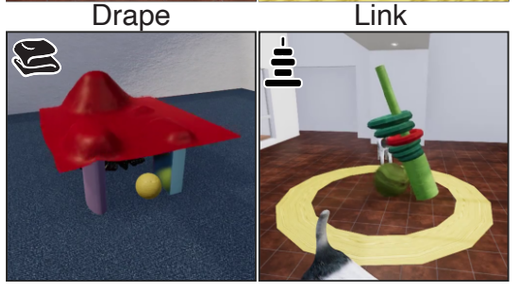
Inputs

Physion

Dominoes | Support

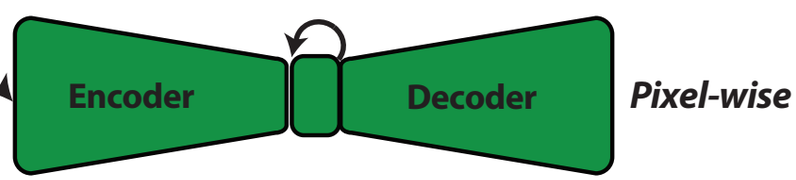


Drape | Link

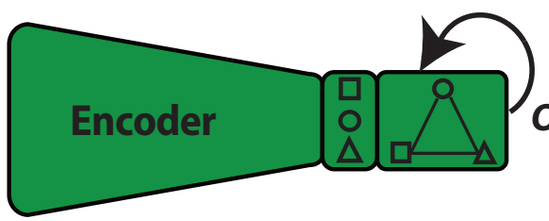


Sensory-Cognitive Hypothesis Classes

End-to-End Future Prediction:



Pixel-wise

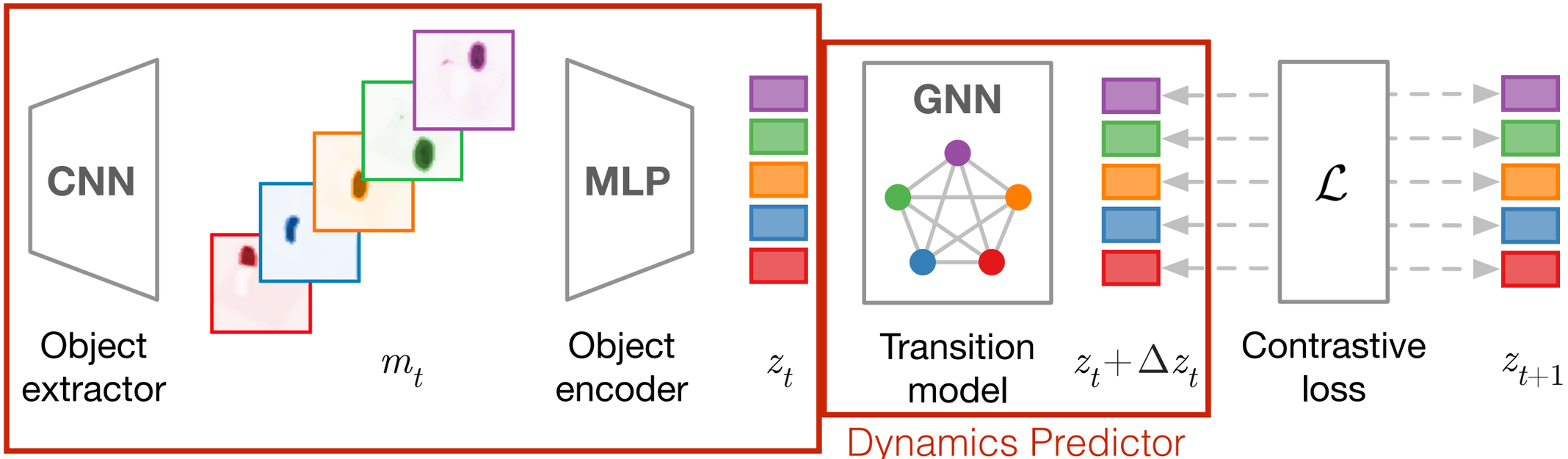


Object-slot

Predicts at the level of object representations and their relations

Kipf et al. 2020

s_t

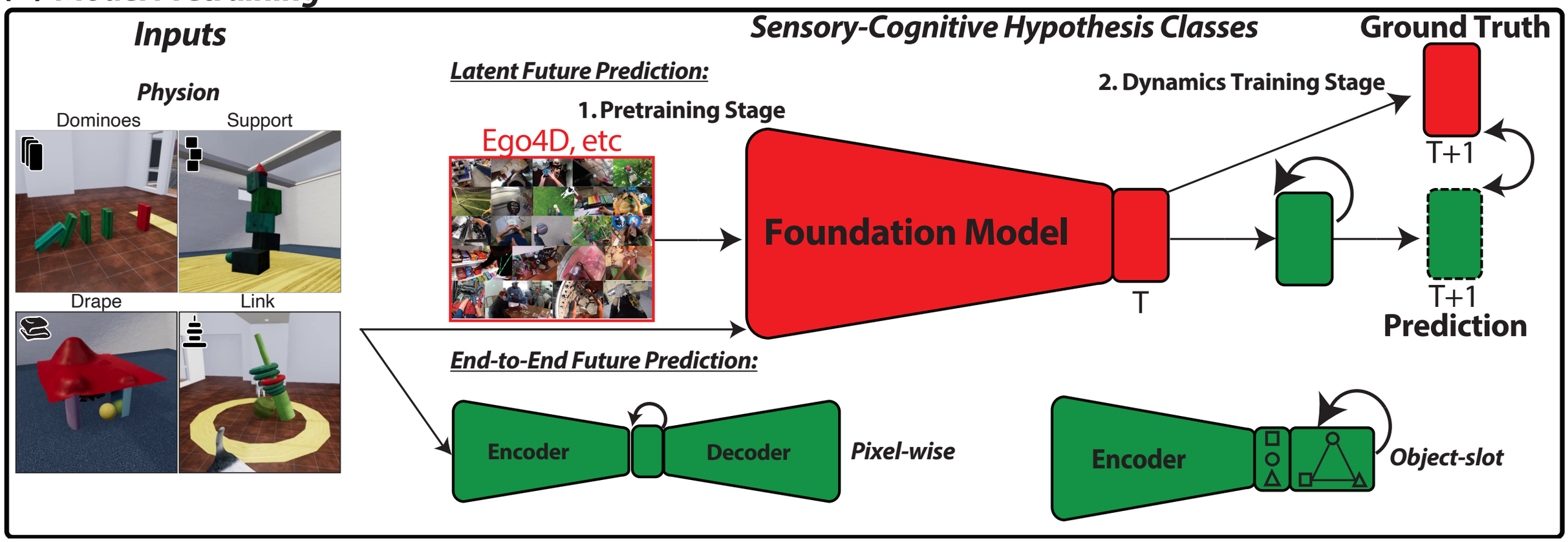


Visual Encoder ("Sensory")

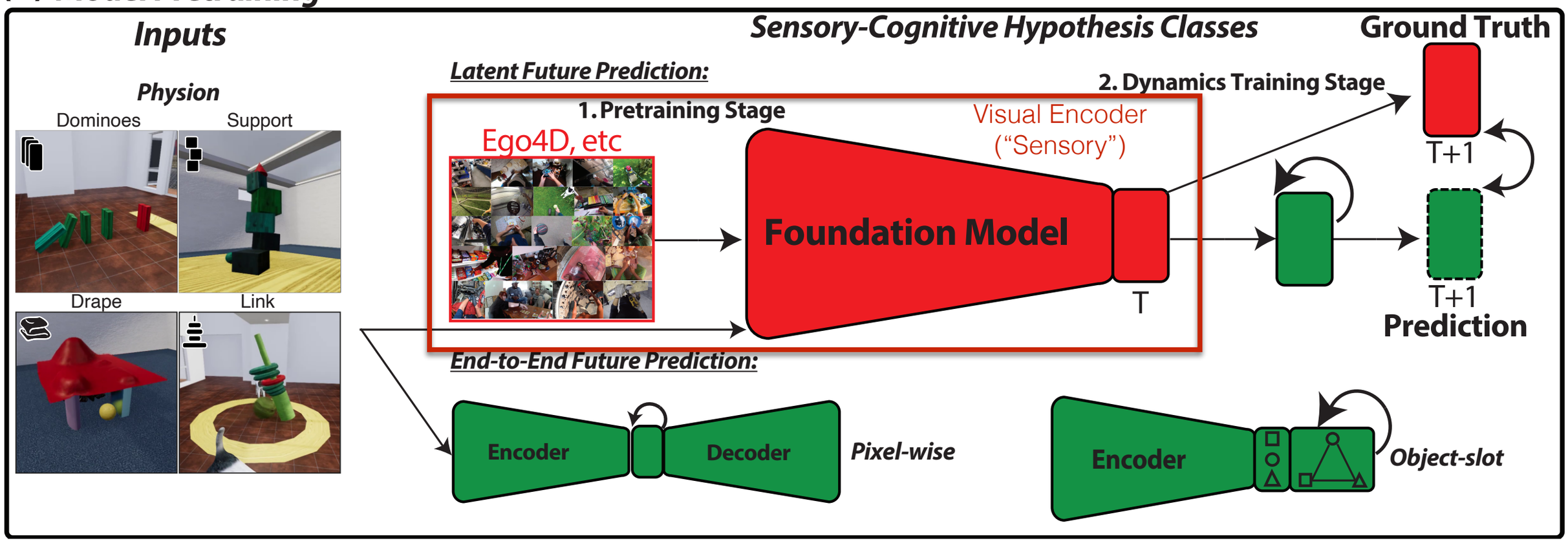
Dynamics Predictor ("Cognitive")

Overall Approach: Foundation Models

(A) *Model Pretraining*

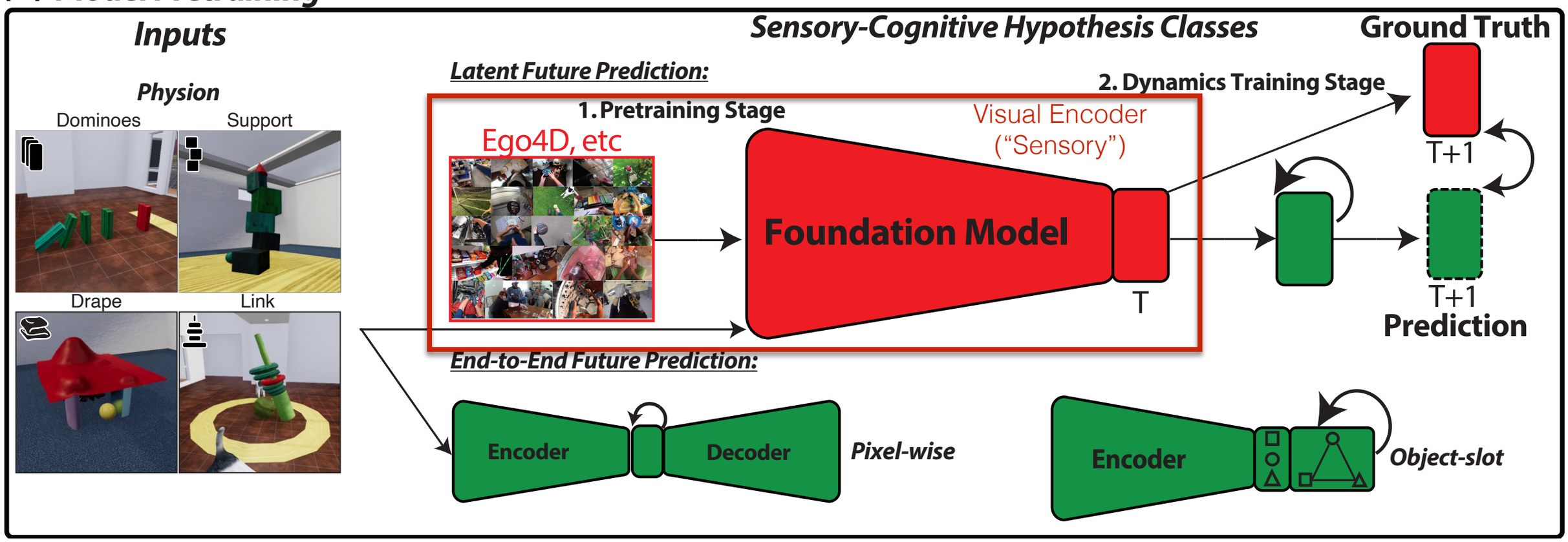


Overall Approach: Foundation Models



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

Overall Approach: Foundation Models

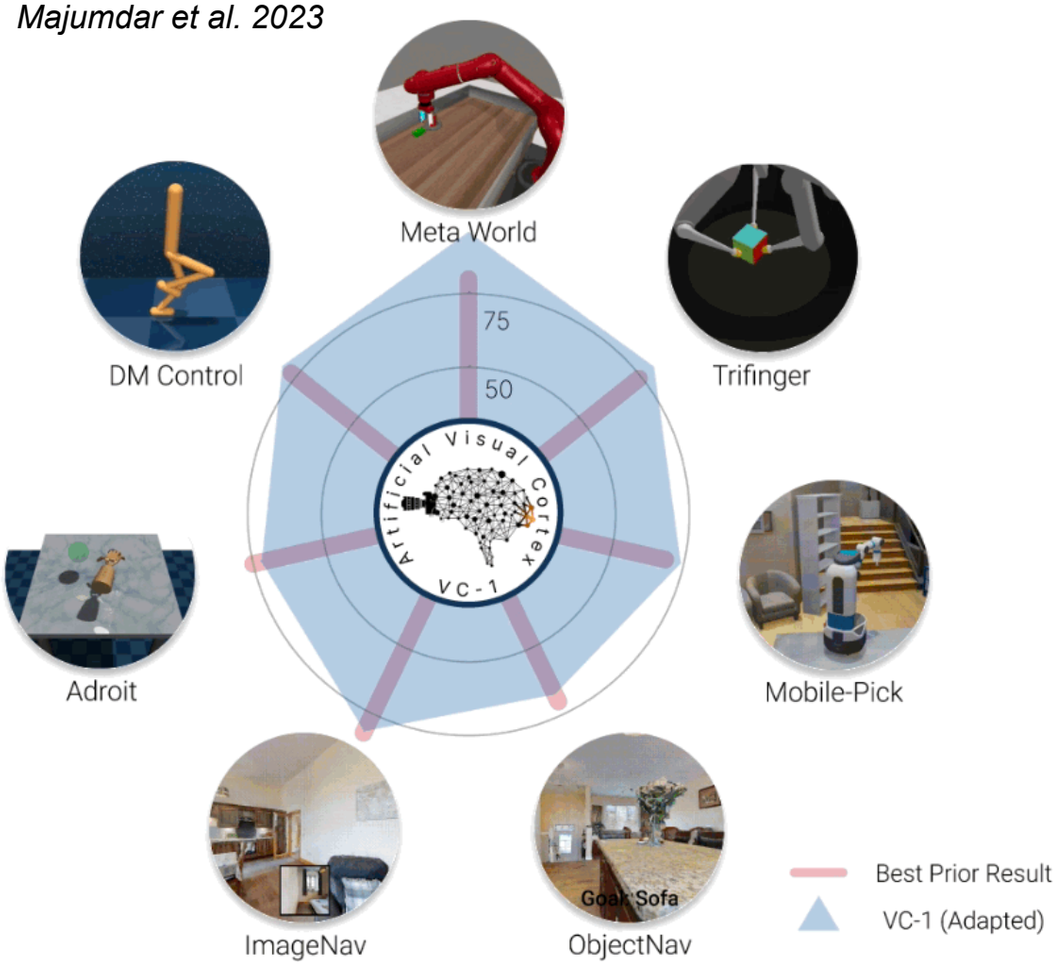


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

Emphasis on *reusability*!

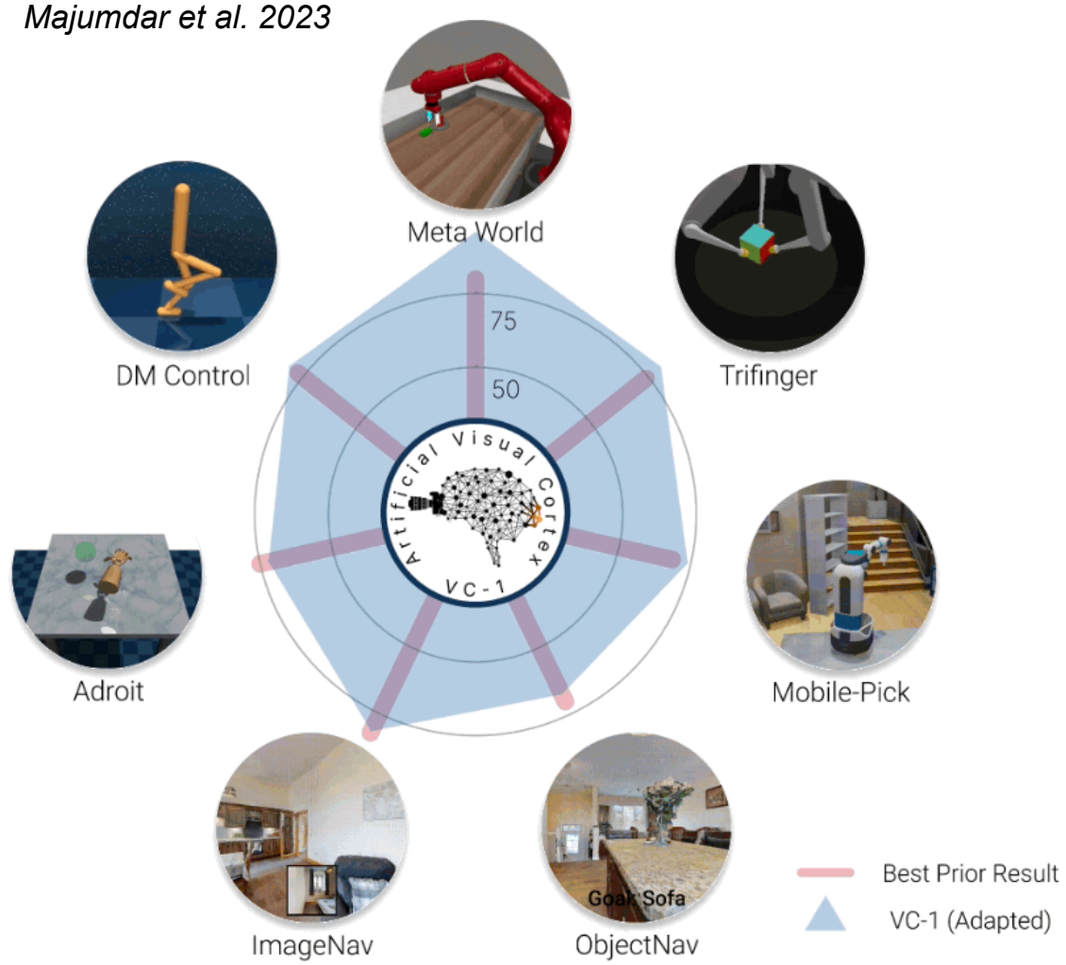
Overall Approach: Foundation Models

Majumdar et al. 2023



Overall Approach: Foundation Models

Majumdar et al. 2023



CortexBench

Ego4D: A massive-scale egocentric dataset

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

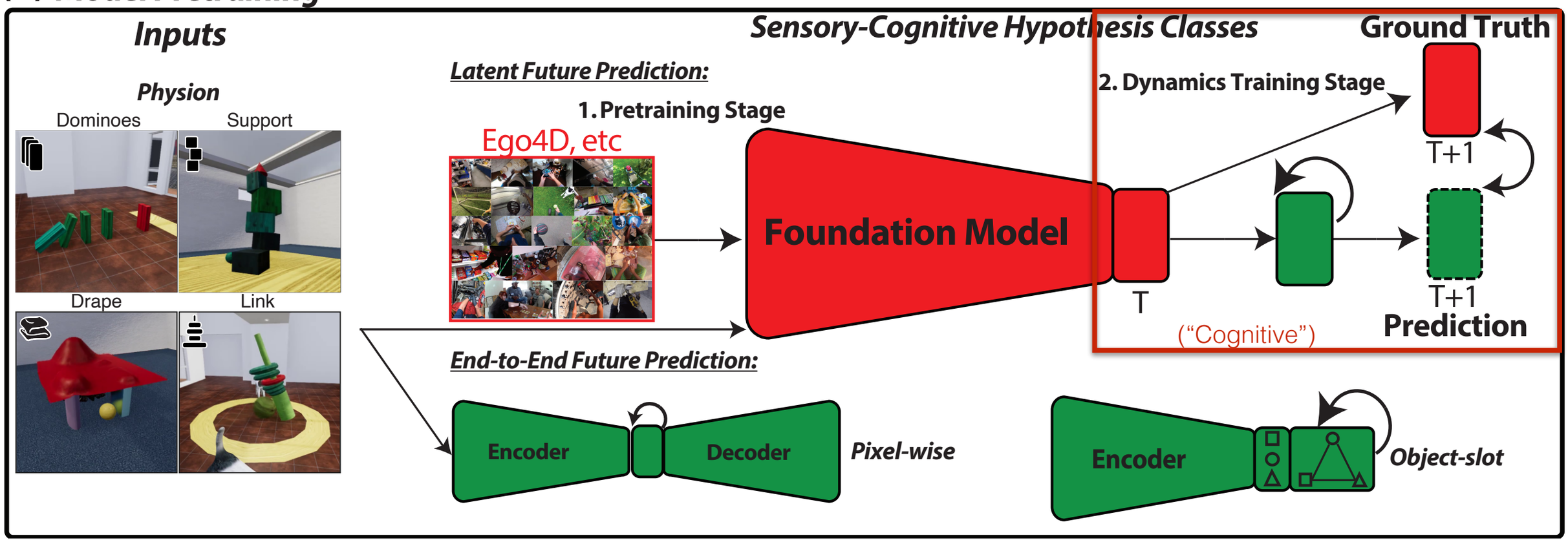
931 participants from 74 worldwide locations

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

Grauman et al. 2022

Ego4D: everyday activity around the world

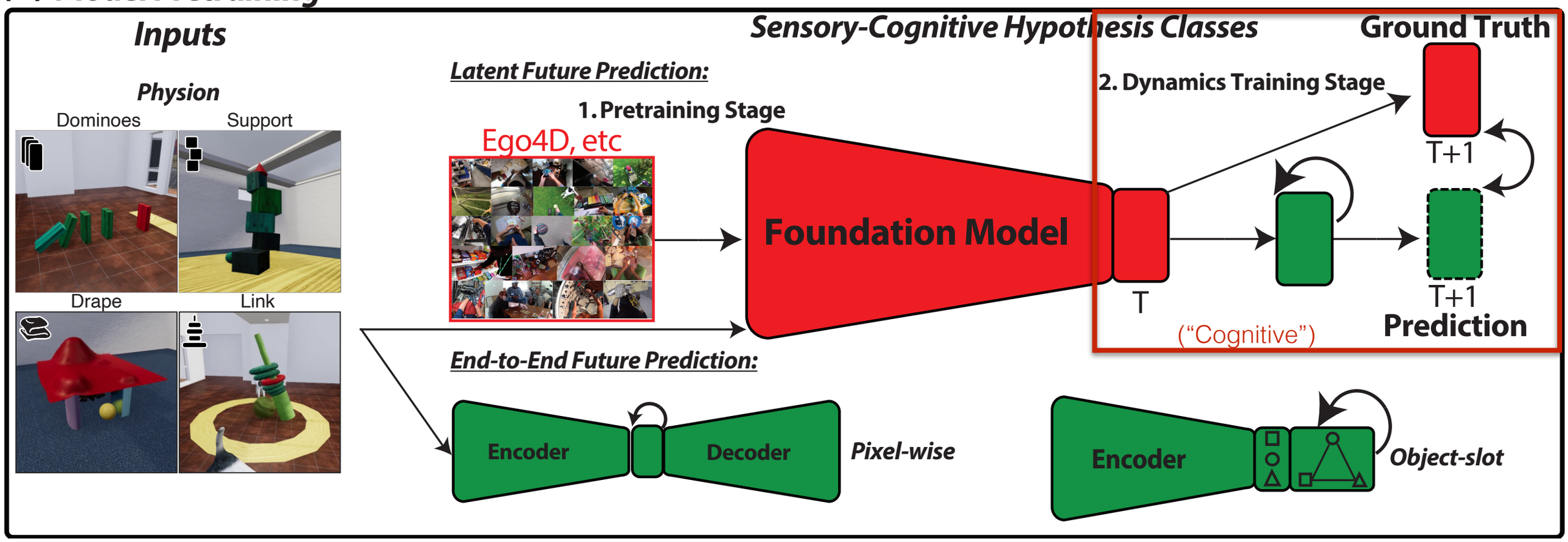
Overall Approach: Foundation Models



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

Emphasis on *reusability*!

Overall Approach: Foundation Models + Dynamics



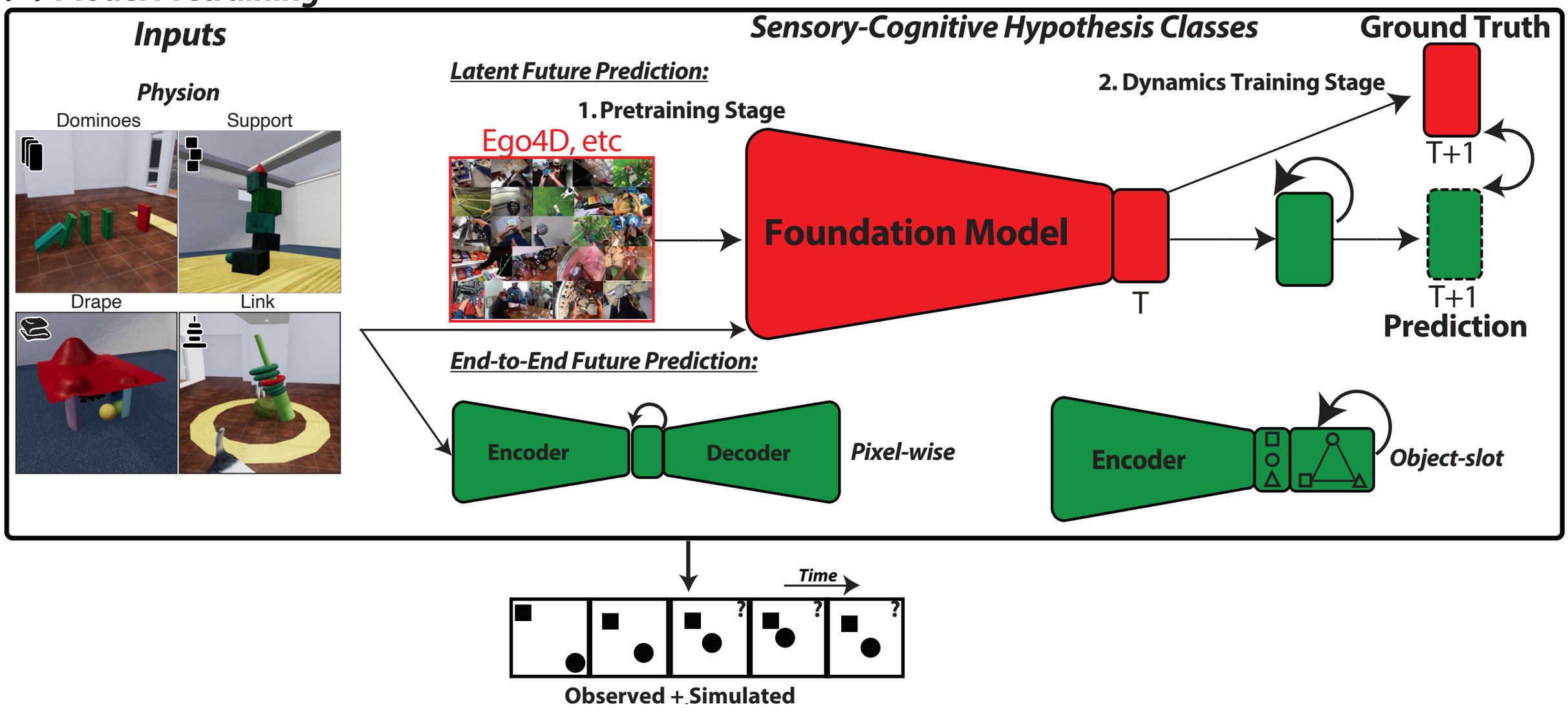
Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

Emphasis on *reusability*!

Leverage these dynamics to do explicit physical simulation

Overall Approach

(A) *Model Pretraining*



Overall Approach: Model Evaluations

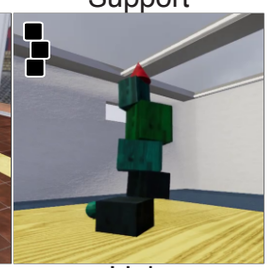
(A) *Model Pretraining*

Inputs

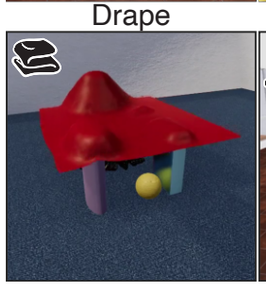
Physion

Dominoes

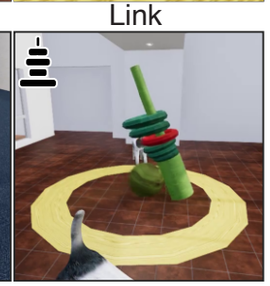
Support



Drape



Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T+1

T

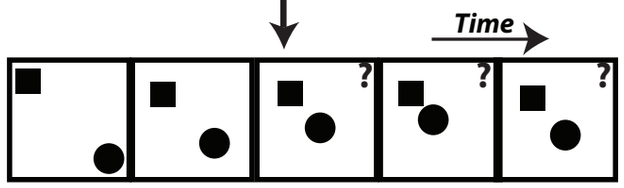
T+1

Prediction

End-to-End Future Prediction:

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

Encoder ***Object-slot***



Observed + Simulated

(B) *Model Evaluations*

Overall Approach: Model Evaluations (Human Behavior)

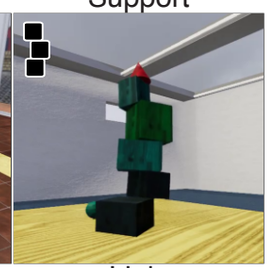
(A) *Model Pretraining*

Inputs

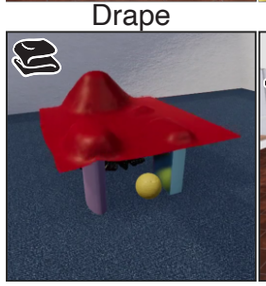
Physion

Dominoes

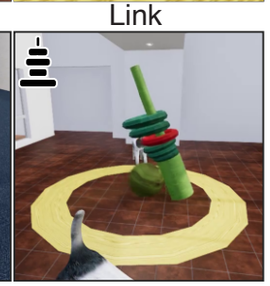
Support



Drape



Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T

T+1

T+1

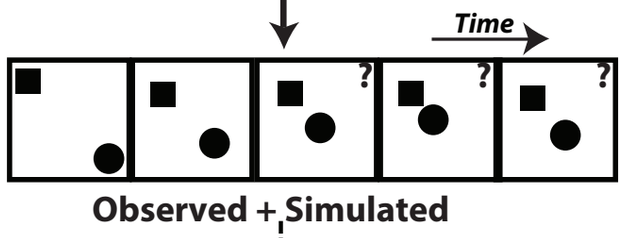
Prediction

End-to-End Future Prediction:

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

Encoder ***Object-slot***

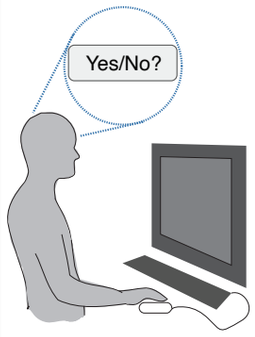
Time



Observed + Simulated

(B) *Model Evaluations*

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



Observed Stimuli

Unobserved Outcome

true label

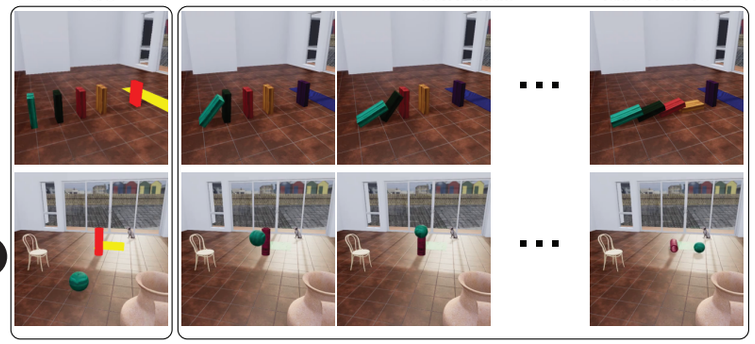
Time

cue

stimulus

last frame

Example Scenarios



NO

acc. = 0.89

YES

acc. = 0.96

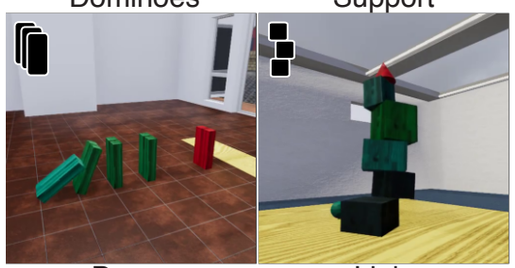
Overall Approach: Model Evaluations (Macaque Physiology)

(A) *Model Pretraining*

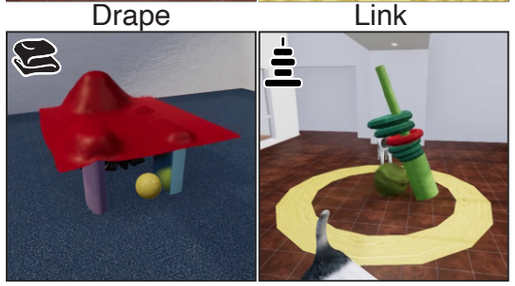
Inputs

Physion

Dominoes | Support



Drape | Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T

T+1

T+1

Prediction

End-to-End Future Prediction:

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

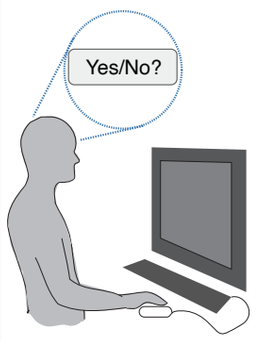
Encoder ***Object-slot***

Time

Observed + Simulated

(B) *Model Evaluations*

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



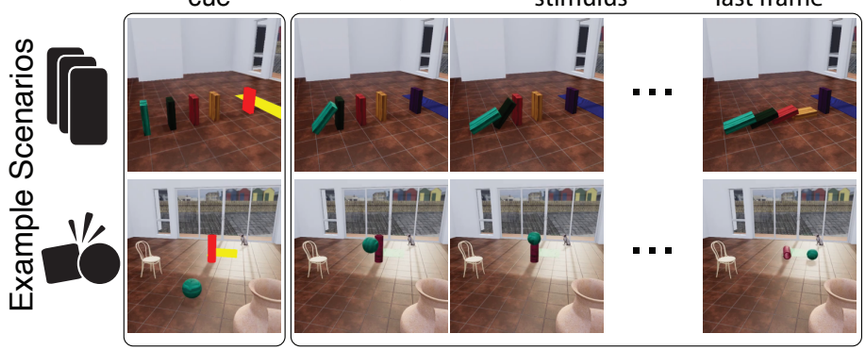
Observed Stimuli — Unobserved Outcome

Time

cue — stimulus — last frame

true label

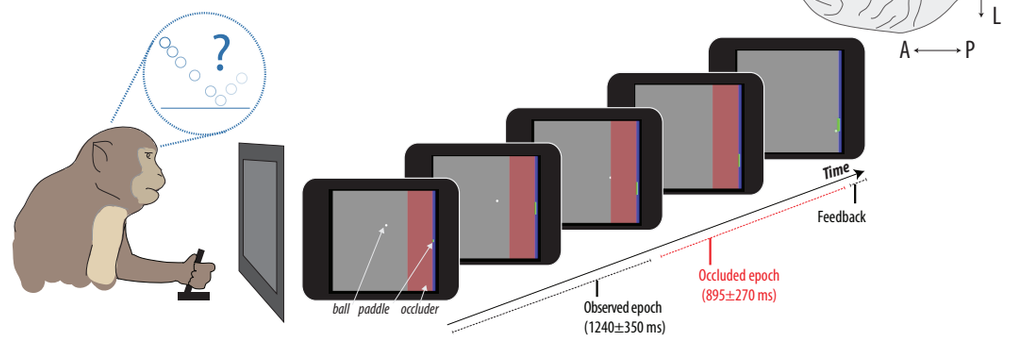
Example Scenarios



NO
acc. = 0.89

YES
acc. = 0.96

2. *Macaque Neurophysiology: Mental-Pong*



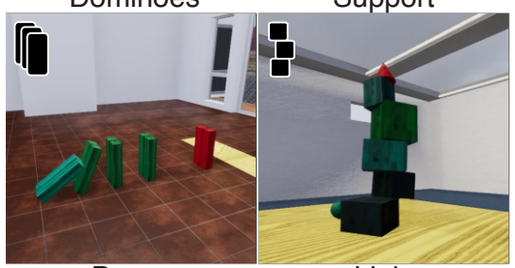
Model Evaluations: Macaque Neurophysiology

(A) *Model Pretraining*

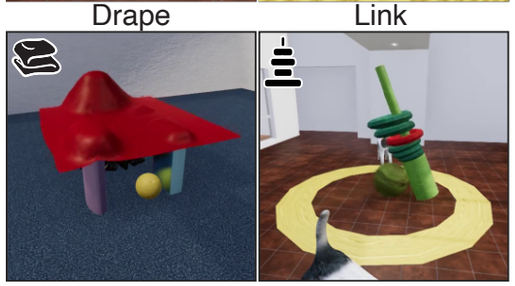
Inputs

Physion

Dominoes | Support



Drape | Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

T

2. Dynamics Training Stage

T+1

T+1

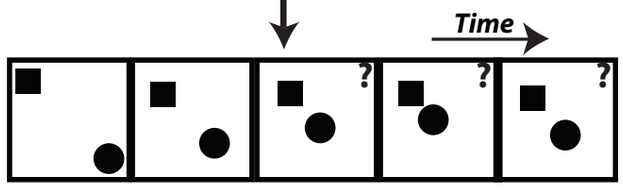
Prediction

End-to-End Future Prediction:

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

Encoder ***Object-slot***

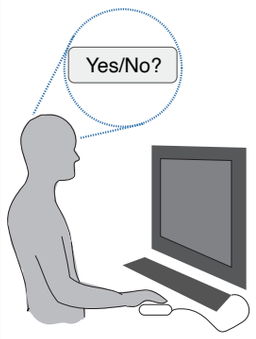
Time



Observed + Simulated

(B) *Model Evaluations*

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

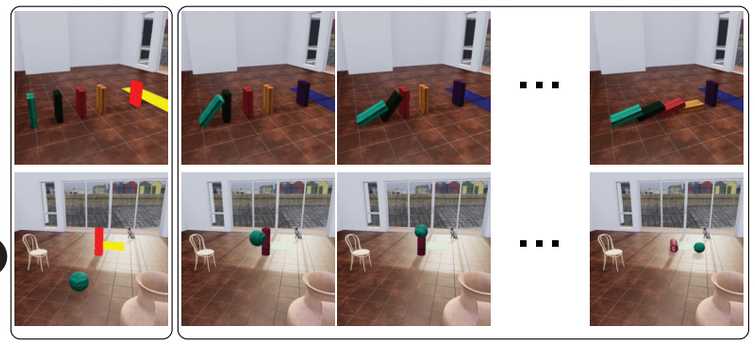


Observed Stimuli | Unobserved Outcome

Time

cue | stimulus | last frame | true label

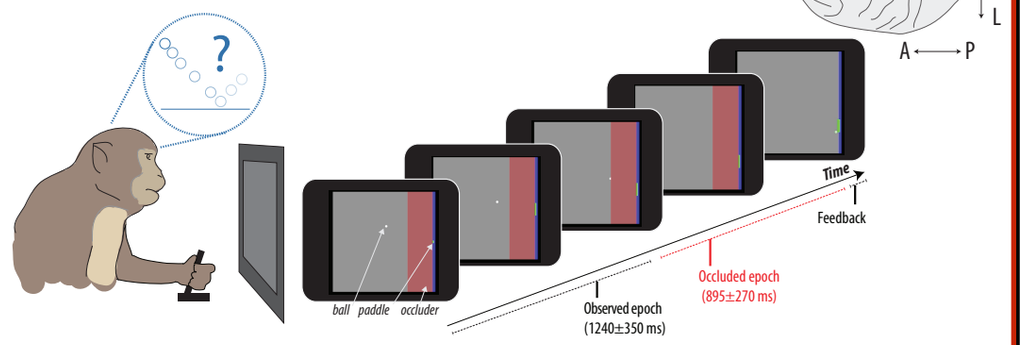
Example Scenarios



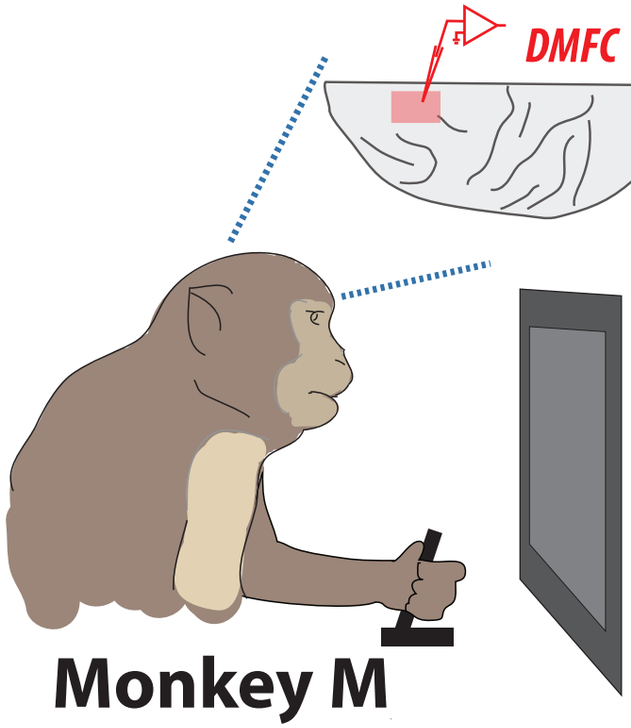
NO
acc. = 0.89

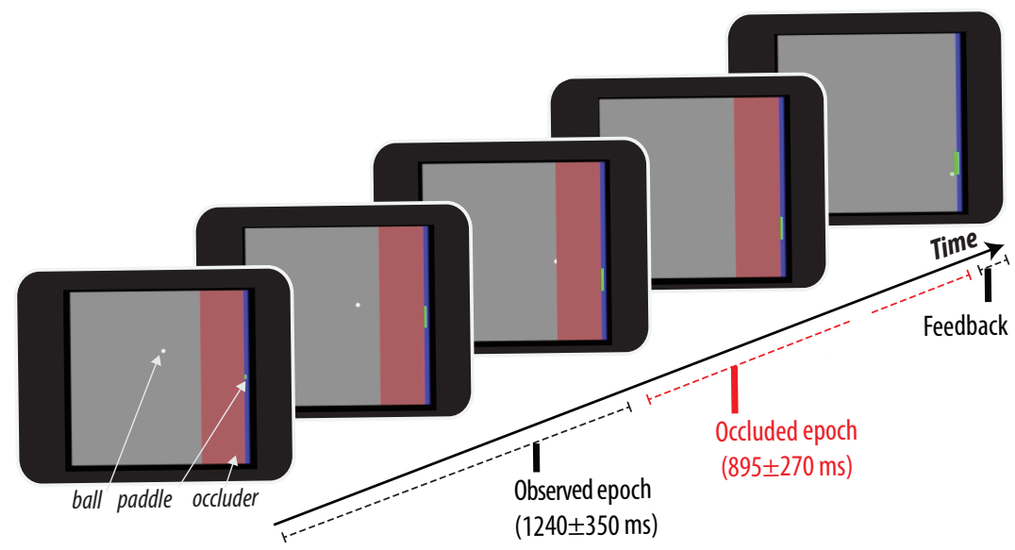
YES
acc. = 0.96

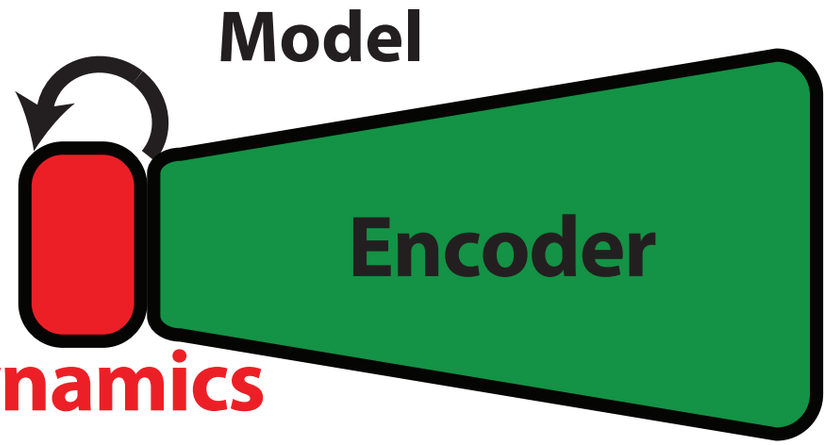
2. *Macaque Neurophysiology: Mental-Pong*

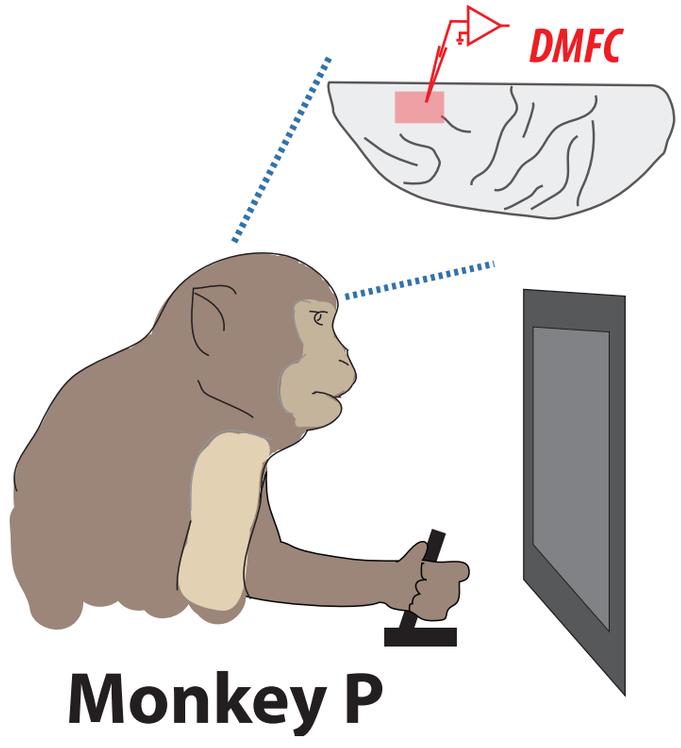


Assessing Model Similarity: Neural Response Predictivity

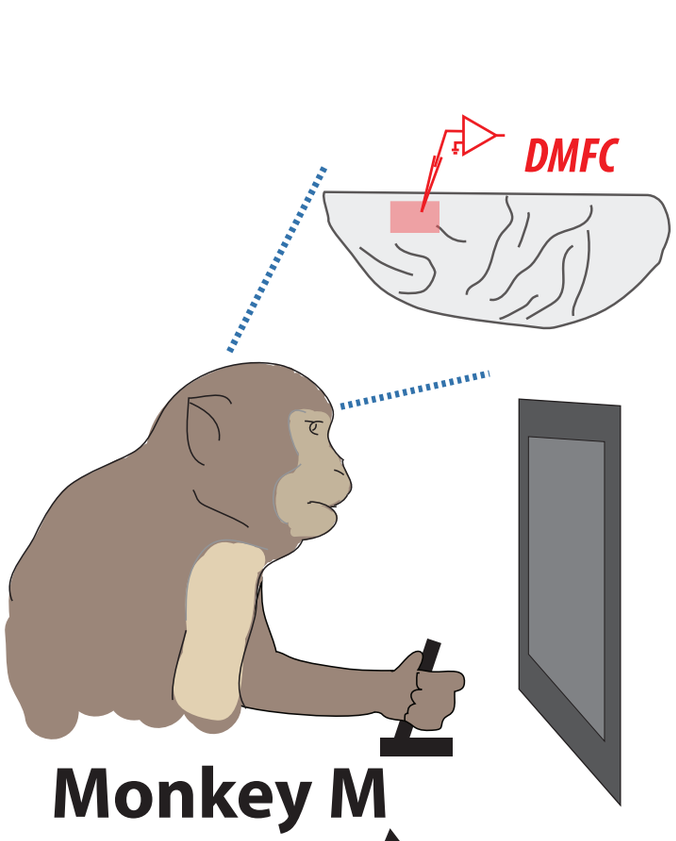


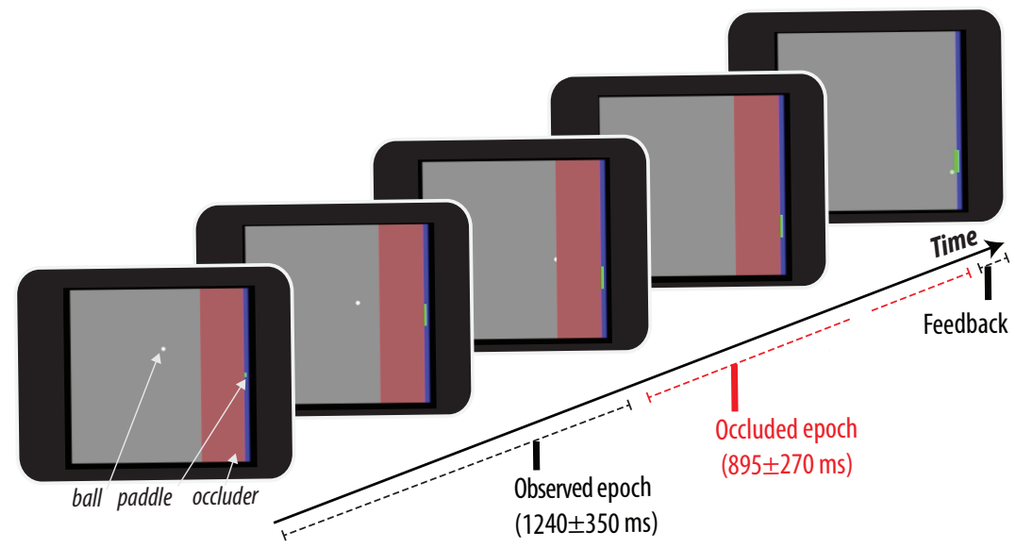


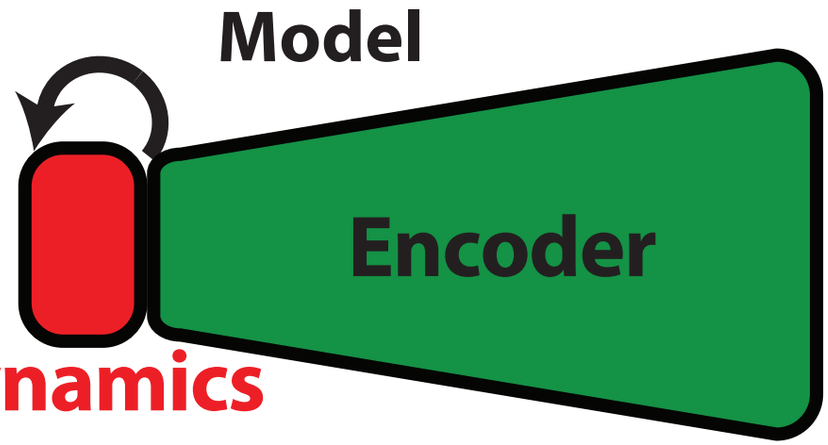




Assessing Model Similarity: Neural Response Predictivity

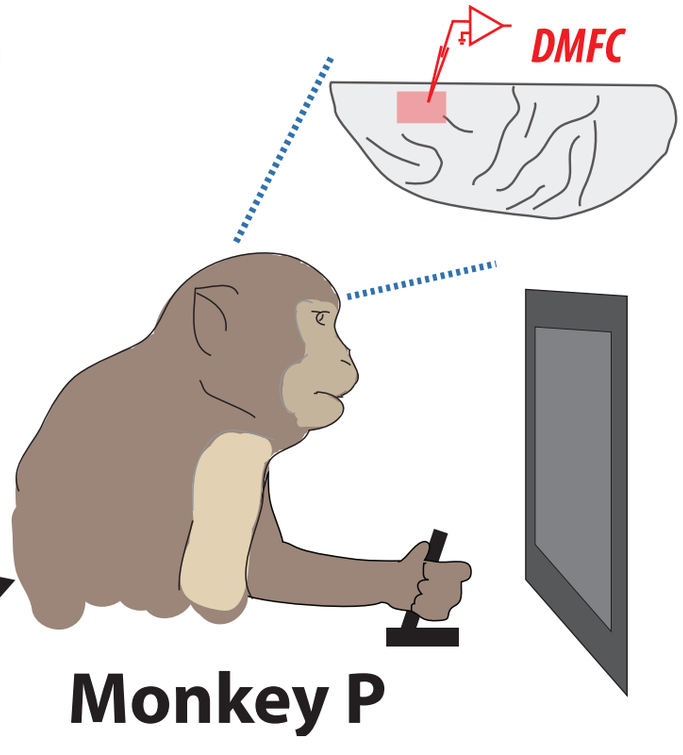




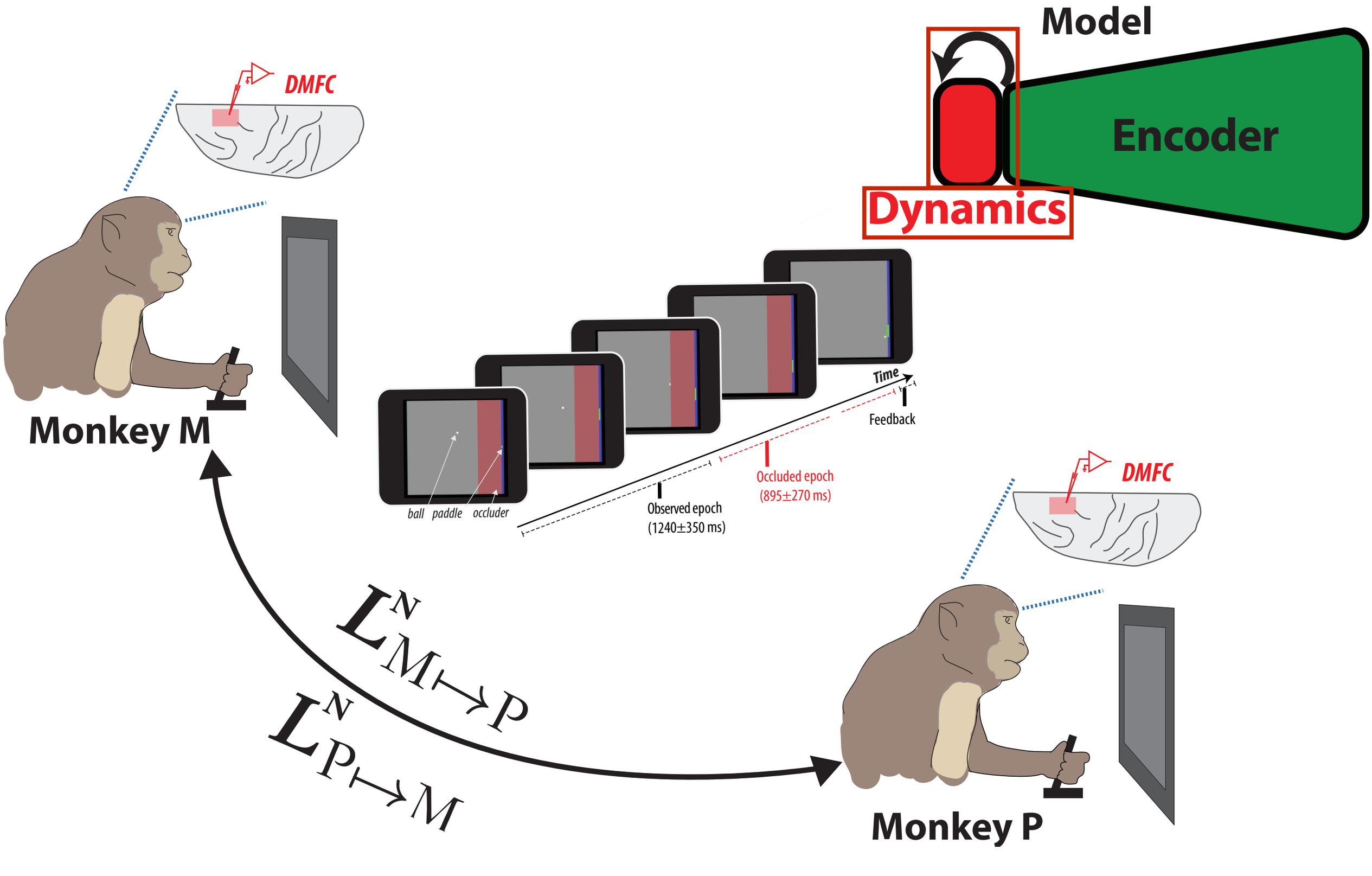


$L^N_{M \mapsto P}$

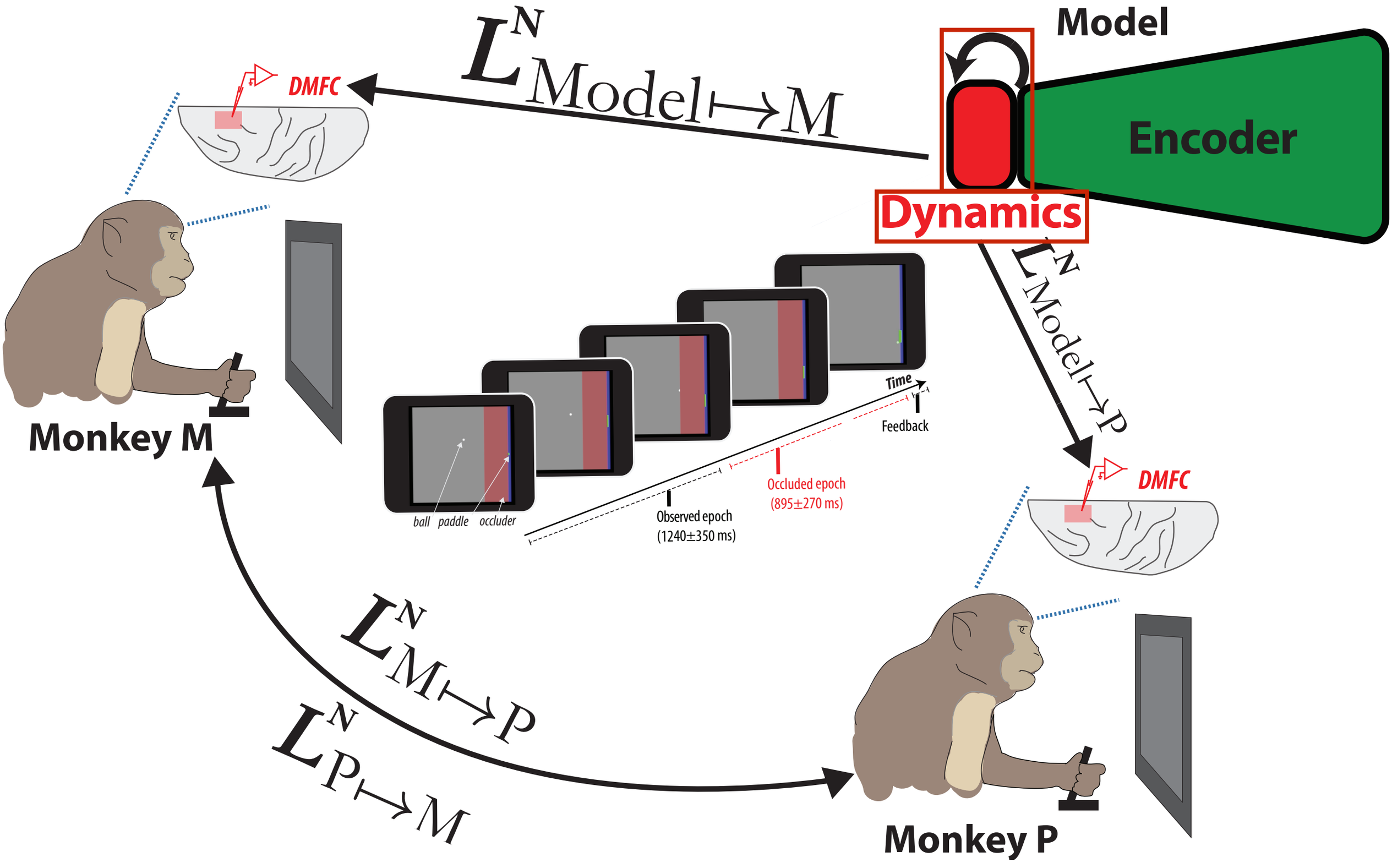
$L^N_{P \mapsto M}$



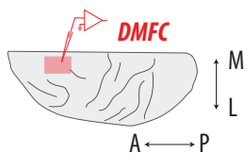
Assessing Model Similarity: Neural Response Predictivity



Assessing Model Similarity: Neural Response Predictivity



Neural response predictivity strongly separates models



Inter-animal Consistency

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

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

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

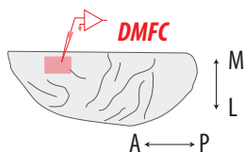
Ball Position
Ball Velocity
Ball Position + Velocity

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

End-to-End

Latent Future Prediction

Neural response predictivity strongly separates models



Inter-animal Consistency

Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

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

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

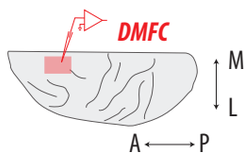
Ball Position
Ball Velocity
Ball Position + Velocity

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

End-to-End

Latent Future Prediction

Neural response predictivity strongly separates models



Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Inter-animal Consistency

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

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

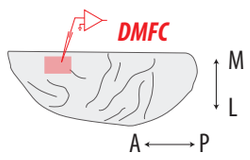
Ball Position
Ball Velocity
Ball Position + Velocity

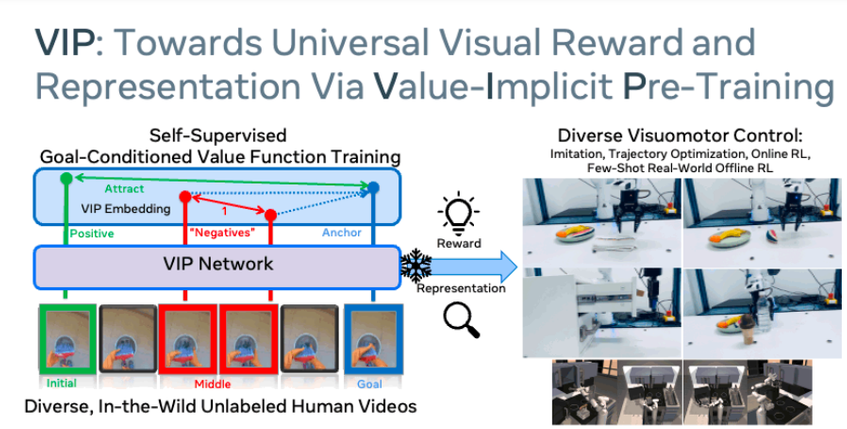
SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP VIP VC-1 R3M

End-to-End

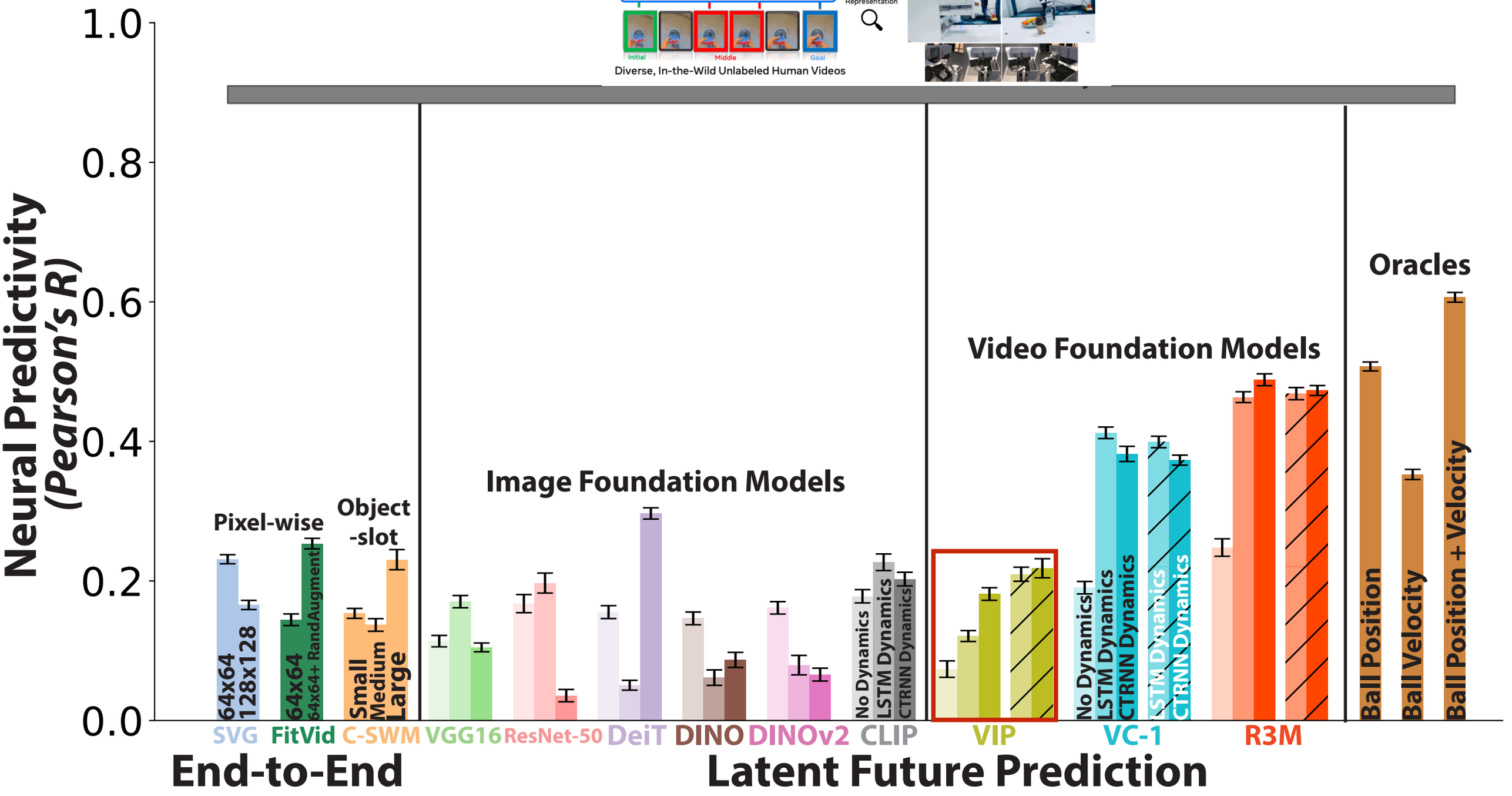
Latent Future Prediction

Neural response predictivity strongly separates models

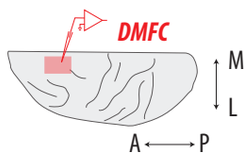




Ma et al. 2023



Neural response predictivity strongly separates models



Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Inter-animal Consistency

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

Image Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics

Video Foundation Models

No Dynamics
LSTM Dynamics
CTRNN Dynamics
LSTM Dynamics
CTRNN Dynamics

Oracles

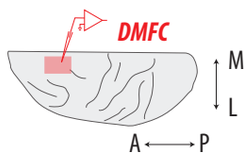
Ball Position
Ball Velocity
Ball Position + Velocity

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

End-to-End

Latent Future Prediction

Neural response predictivity strongly separates models



Neural Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Inter-animal Consistency

Most useful for image segmentation, classification, etc

Image Foundation Models

Most useful for robotic manipulation tasks

Video Foundation Models

Oracles

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

Object-slot
Small
Medium
Large

No Dynamics
LSTM Dynamics
CTRNN Dynamics

No Dynamics
LSTM Dynamics
CTRNN Dynamics

LSTM Dynamics
CTRNN Dynamics

Ball Position
Ball Velocity
Ball Position + Velocity

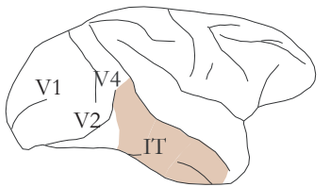
SVG FitVid C-SWM VGG16 ResNet-50 DeiT DINO DINOv2 CLIP VIP VC-1 R3M

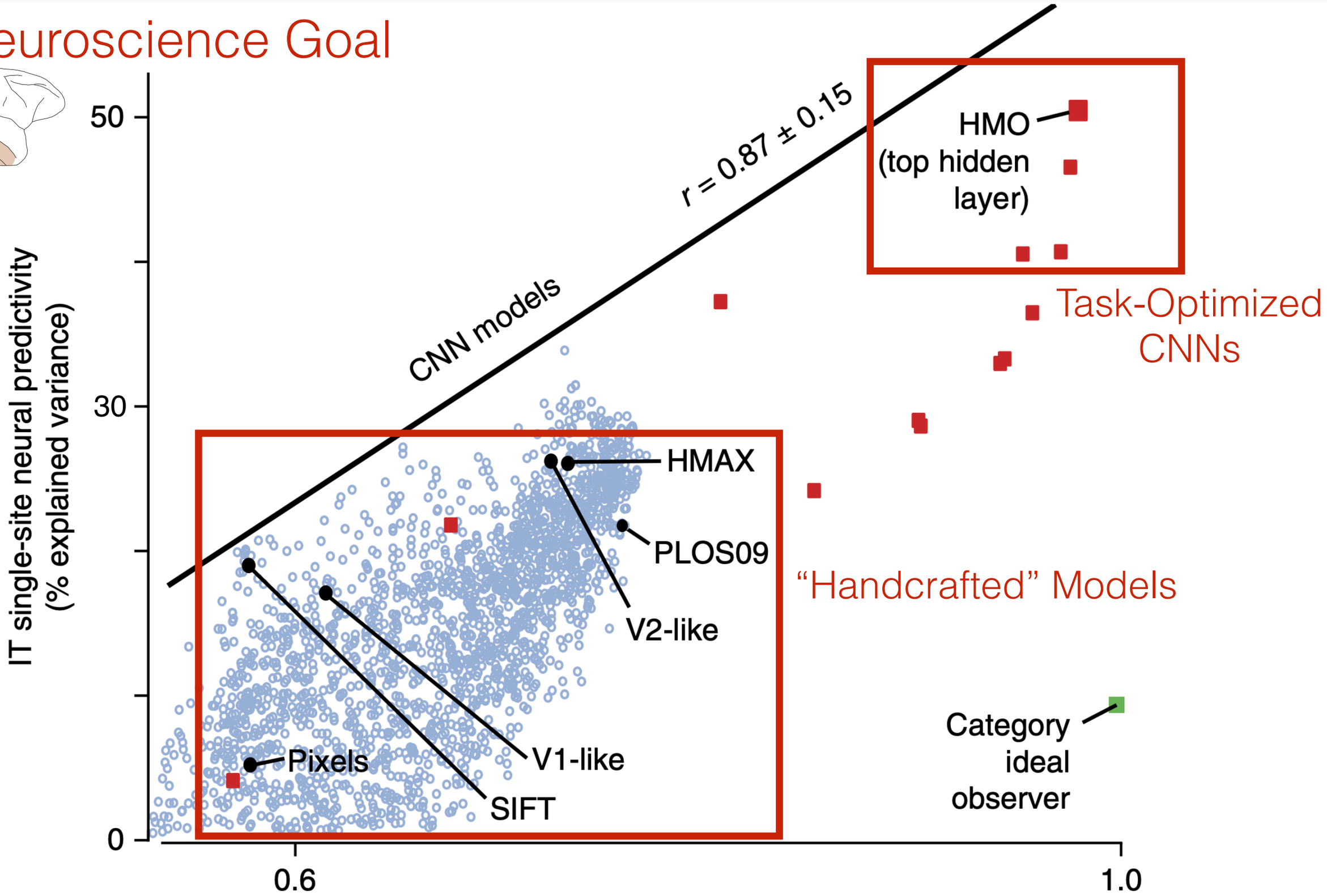
End-to-End

Latent Future Prediction

Prior Results in Inferior Temporal (IT) Cortex

A Neuroscience Goal

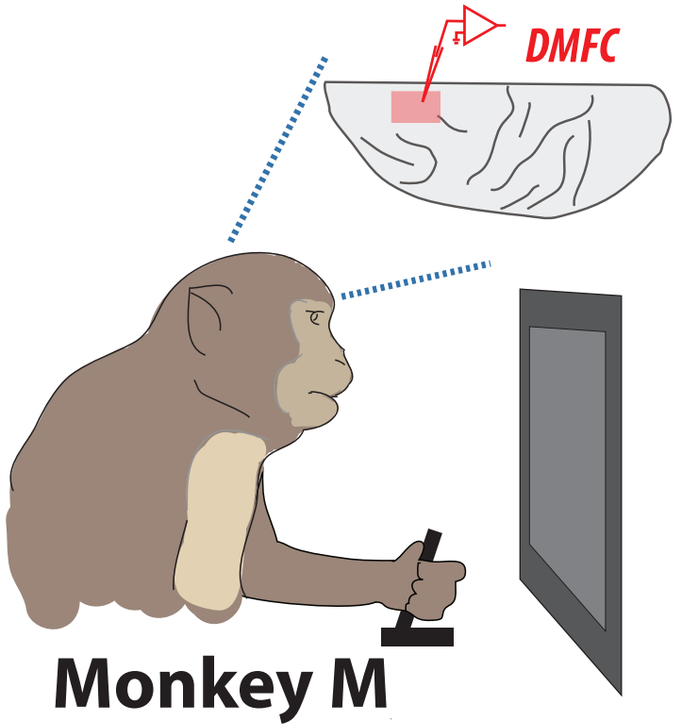


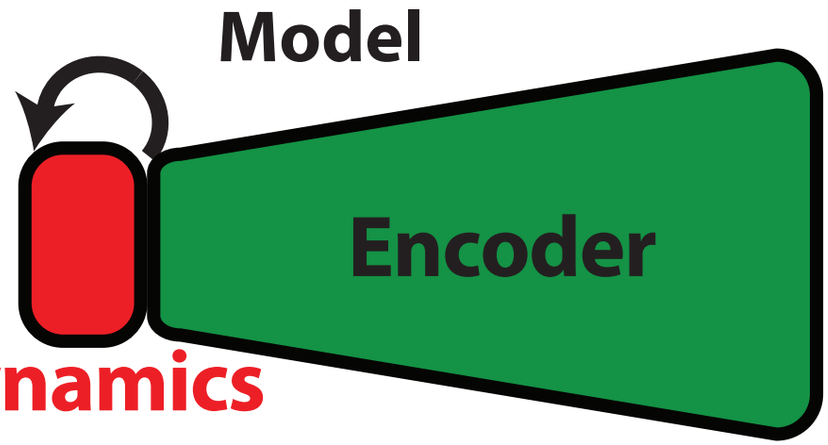


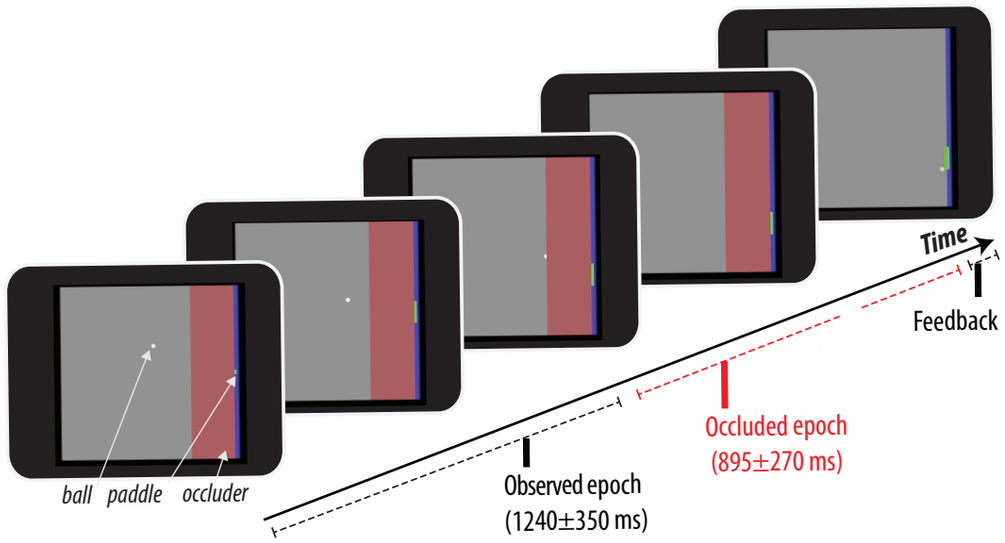
Yamins, Hong* et al. 2014*

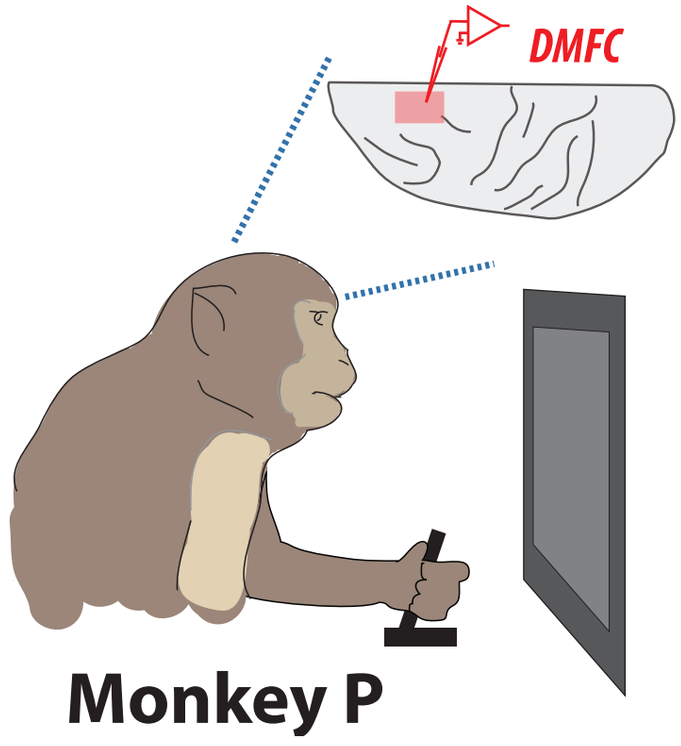
An AI Goal

Assessing Model Similarity: Ground Truth State Decoding

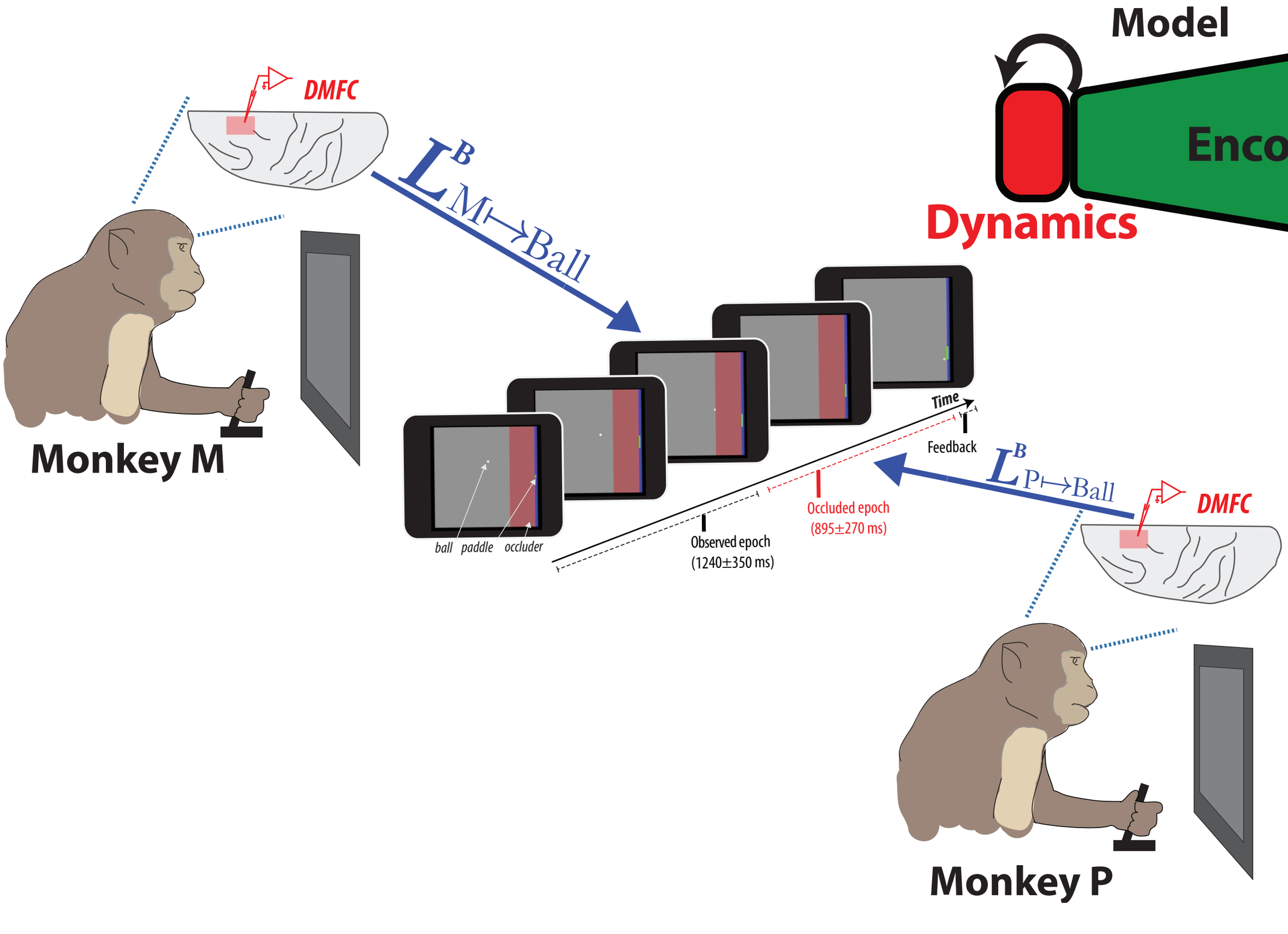




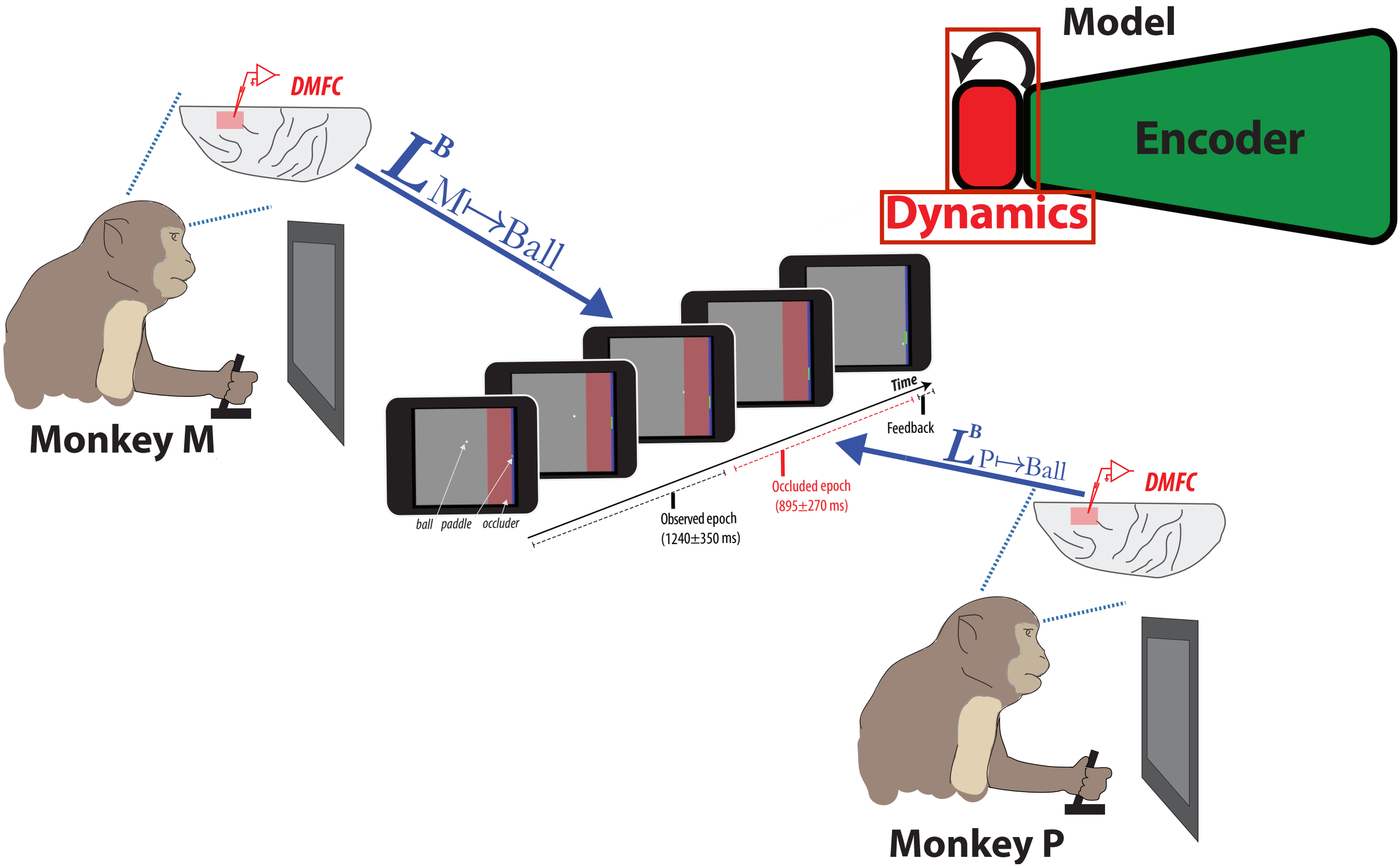




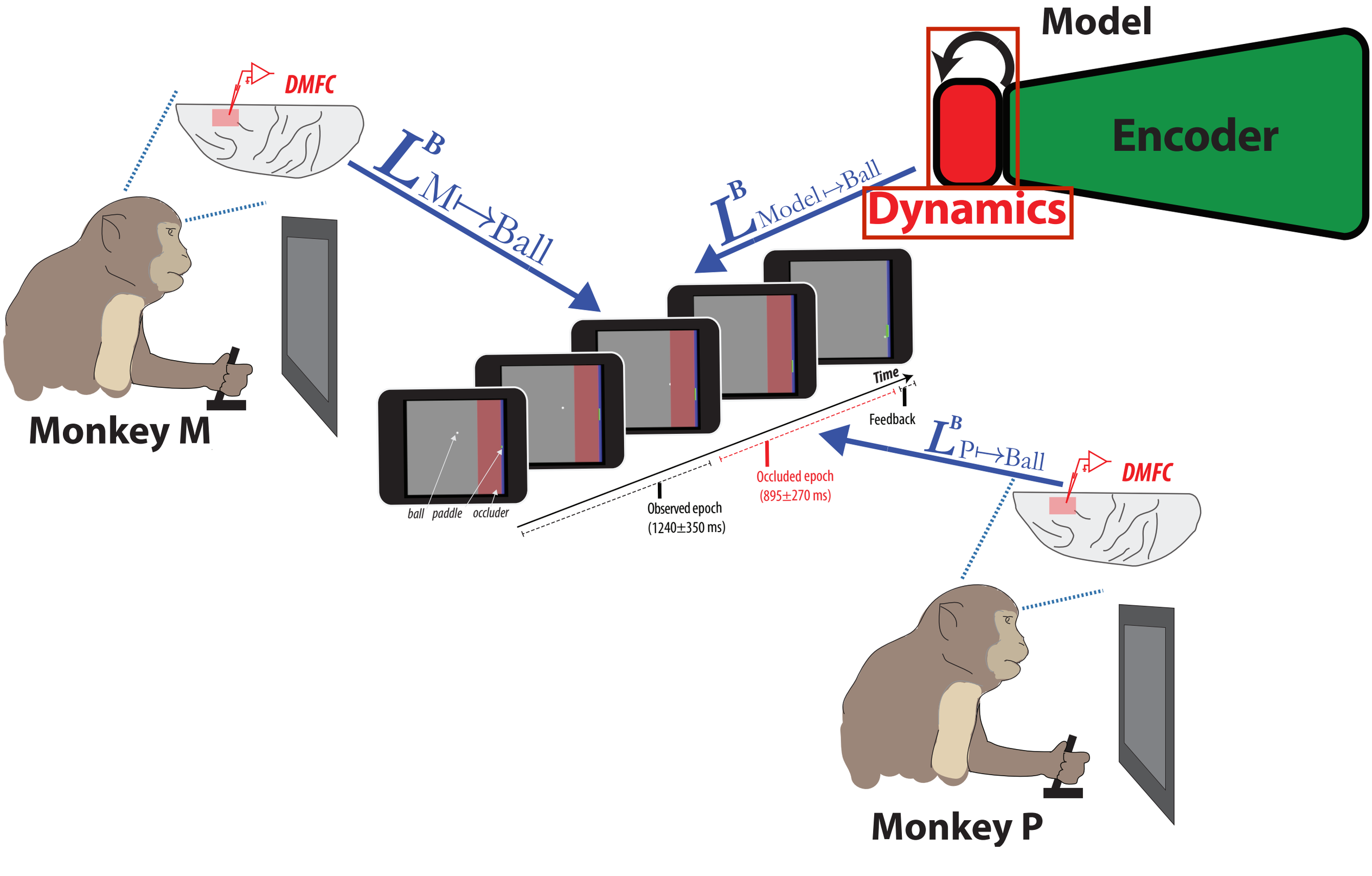
Assessing Model Similarity: Ground Truth State Decoding



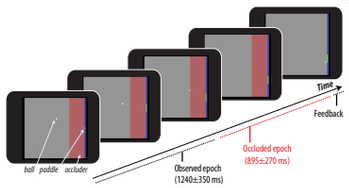
Assessing Model Similarity: Ground Truth State Decoding



Assessing Model Similarity: Ground Truth State Decoding



Best models approach ground truth state predictivity ceiling



DMFC Predictivity

Ball Position + Velocity Predictivity (Pearson's R)

1.0
0.8
0.6
0.4
0.2
0.0

Image Foundation Models

Video Foundation Models

Pixel-wise

Object -slot

64x64
128x128

64x64
64x64+ RandAugment

Small
Medium
Large

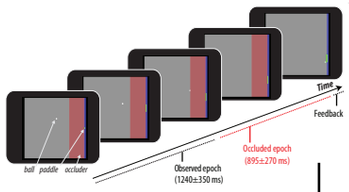
No Dynamics
LSTM Dynamics
CTRNN Dynamics

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

End-to-End

Latent Future Prediction

Predicting neurons is relevant to simulating the ball



R≈0.683, p << 0.001

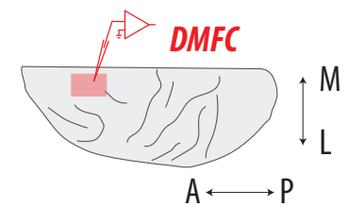
Ball Position + Velocity Predictivity (Pearson's R)

0.6
0.4
0.2
0.0

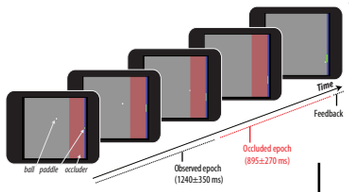
0.1
0.2
0.3
0.4
0.5

Neural Predictivity (Pearson's R)

End-to-End
Image Foundation Models
Video Foundation Models



Predicting neurons is relevant to simulating the ball



R≈0.683, p << 0.001

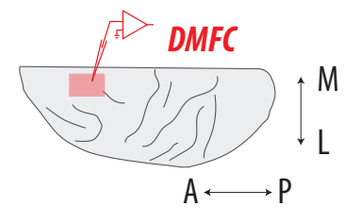
Ball Position + Velocity Predictivity
(Pearson's R)

0.6
0.4
0.2
0.0

0.1
0.2
0.3
0.4
0.5

Neural Predictivity
(Pearson's R)

End-to-End
Image Foundation Models
Video Foundation Models



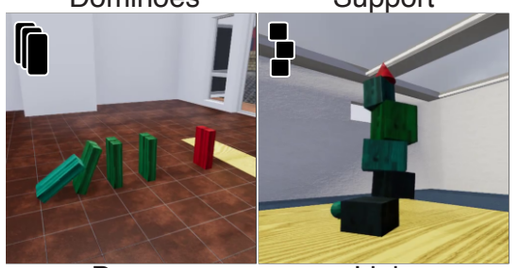
Model Evaluations: Object Contact Prediction (OCP)

(A) *Model Pretraining*

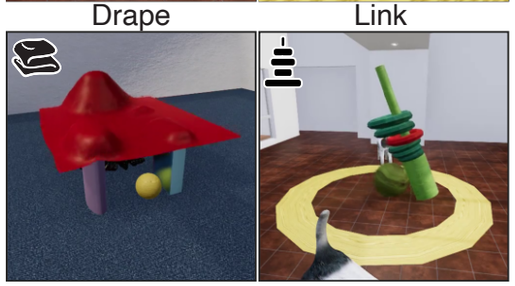
Inputs

Physion

Dominoes | Support



Drape | Link



Sensory-Cognitive Hypothesis Classes

Ground Truth

Latent Future Prediction:

1. Pretraining Stage

Ego4D, etc

Foundation Model

2. Dynamics Training Stage

T

T+1

T+1

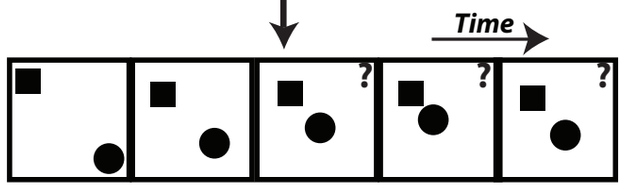
Prediction

End-to-End Future Prediction:

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

Encoder ***Object-centric***

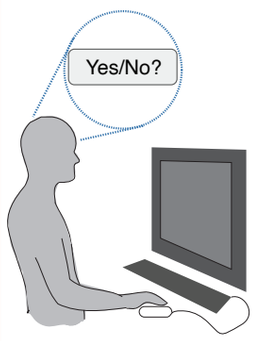
Time



Observed + Simulated

(B) *Model Evaluations*

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

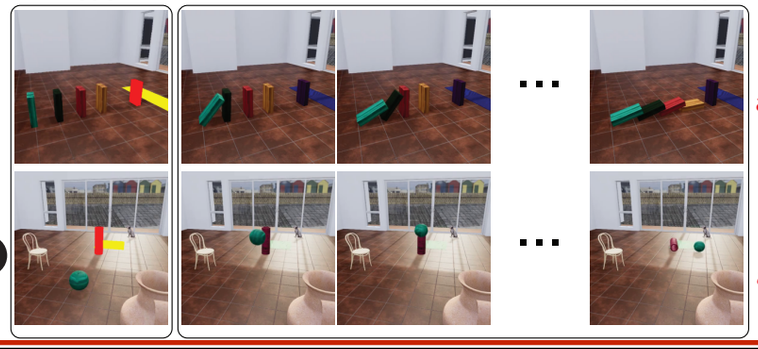


Observed Stimuli — Unobserved Outcome

Time

cue — stimulus — last frame — true label

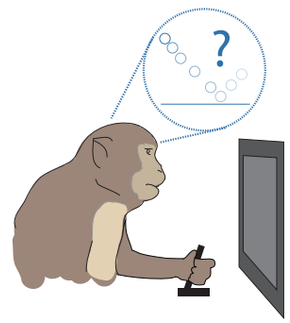
Example Scenarios

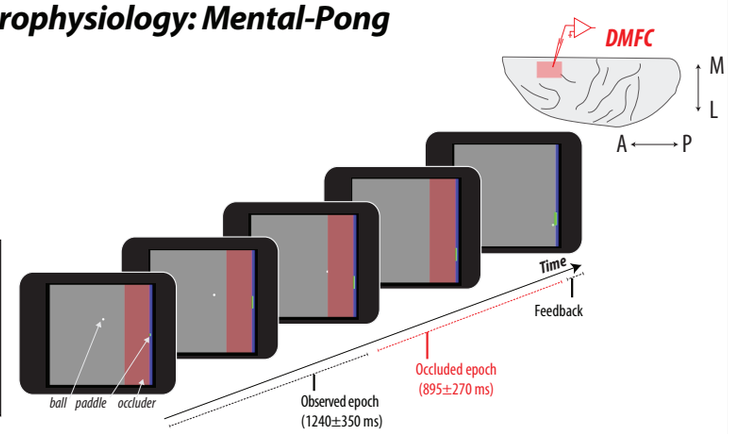


NO
acc. = 0.89

YES
acc. = 0.96

2. *Macaque Neurophysiology: Mental-Pong*

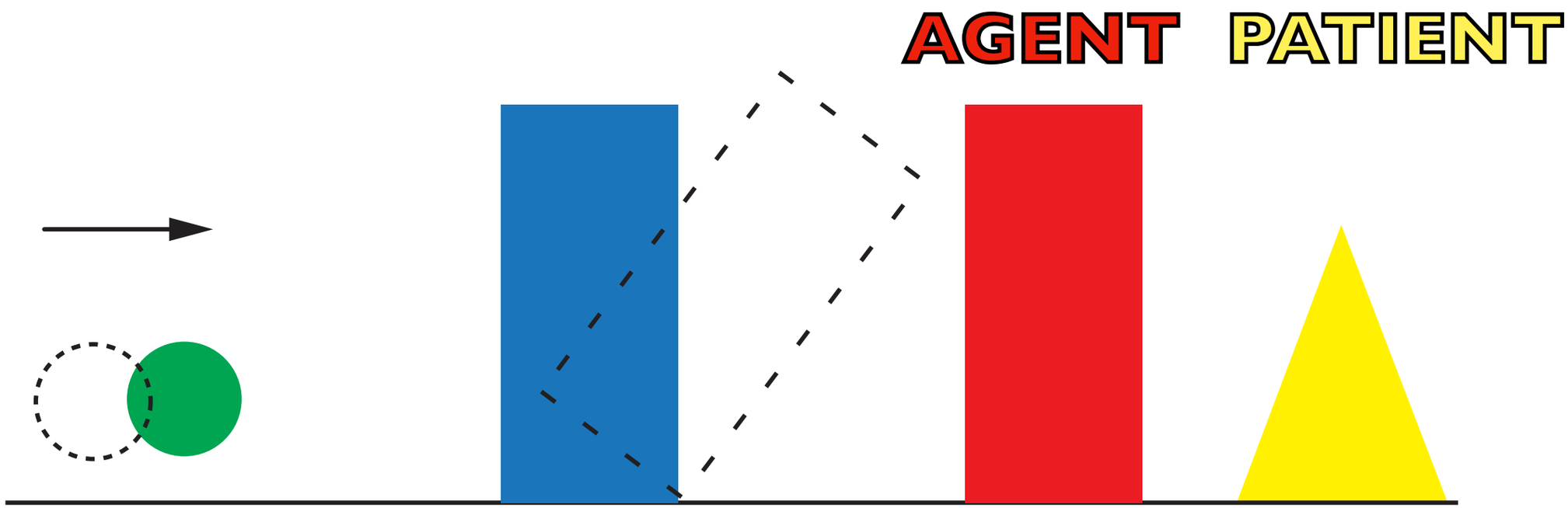


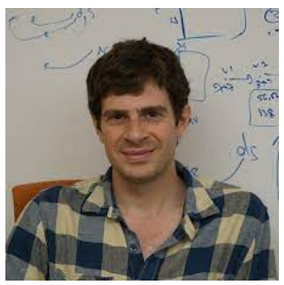


Model Evaluations: Object Contact Prediction (OCP)

Bear et al. 2021

"Will the agent object contact the patient object?"




Daniel Bear

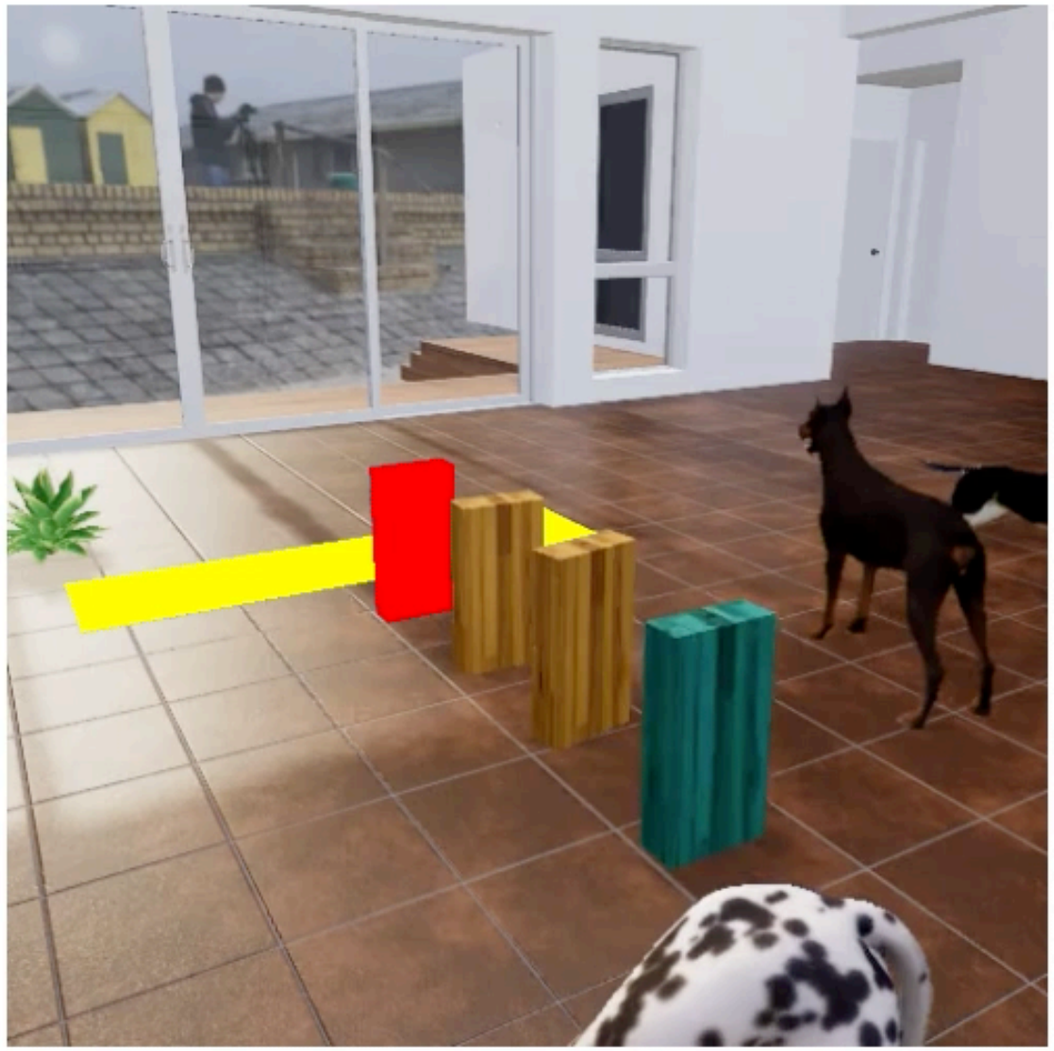

Joshua Tenenbaum


Daniel Yamins


Judith Fan

Completion Progress

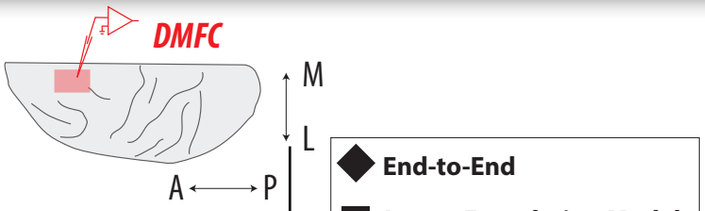
Bear et al. 2021



YES NO

Is the red object going to hit the yellow area?

Dynamically Equipped Sensorimotor Foundation Models Can Match Both



End-to-End
Image Foundation Models
Video Foundation Models

Better Models

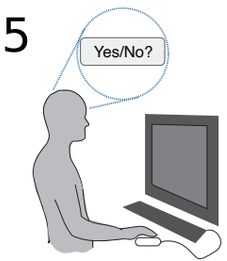
VC-1+LSTM

VC-1+CTRNN

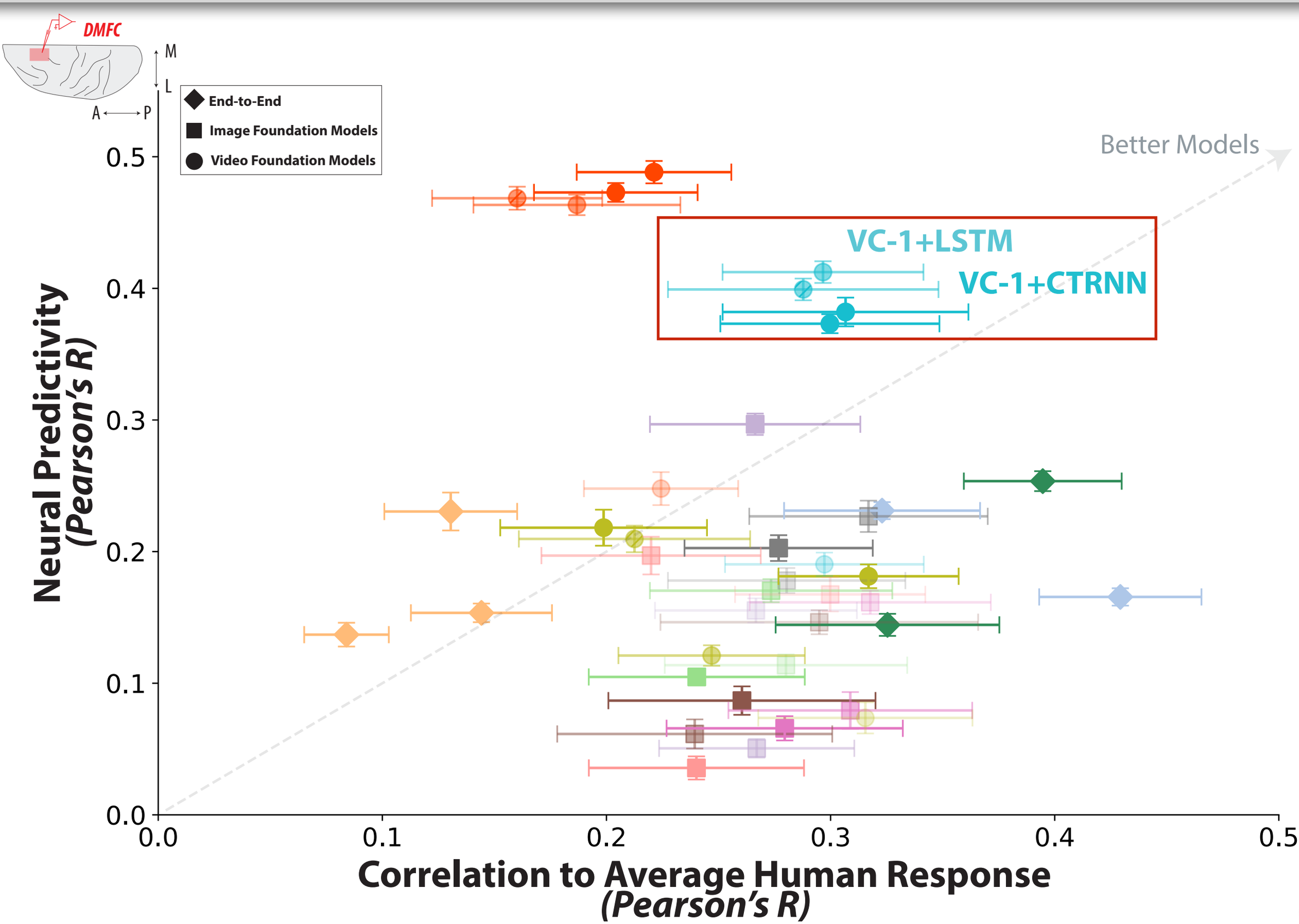
Neural Predictivity (Pearson's R)

0.0
0.1
0.2
0.3
0.4
0.5

Correlation to Average Human Response (Pearson's R)



Dynamically Equipped Sensorimotor Foundation Models Can Match Both



Takeaways so far…

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

Takeaways so far…

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

2. In particular, this latent space is highly constrained -- it doesn't appear to consist of bespoke object slots or prioritize fine-grained details (e.g. at the level of pixels), but rather mainly has to be **reusable** across *dynamic* scenes.

Takeaways so far…

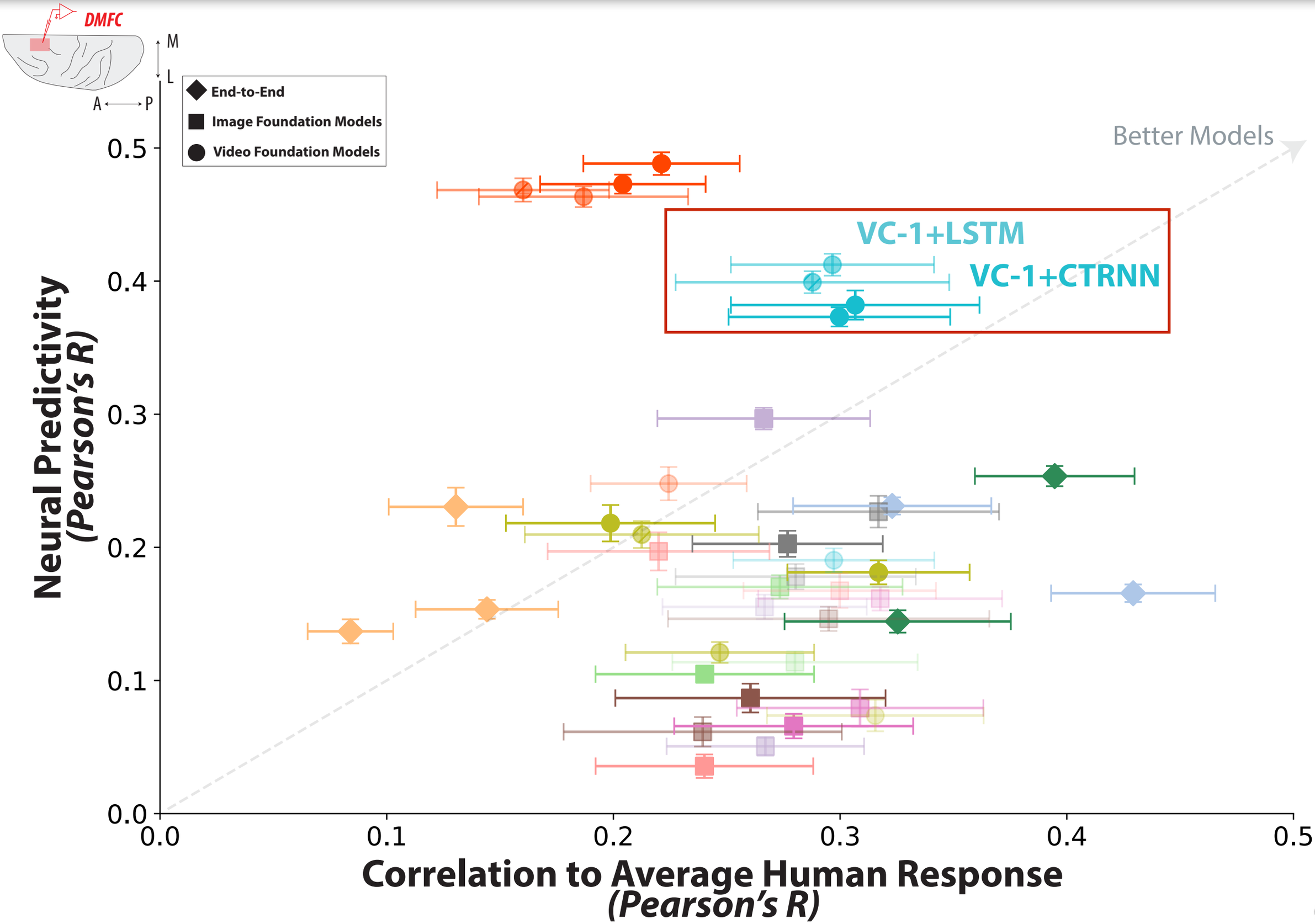
1. Mental simulation appears to be primarily relevant to predicting the **future** state of the environment in a suitable **latent** space.

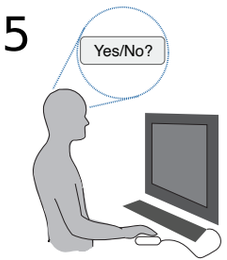
2. In particular, this latent space is highly constrained -- it doesn't appear to consist of bespoke object slots or prioritize fine-grained details (e.g. at the level of pixels), but rather mainly has to be **reusable** across *dynamic* scenes.

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

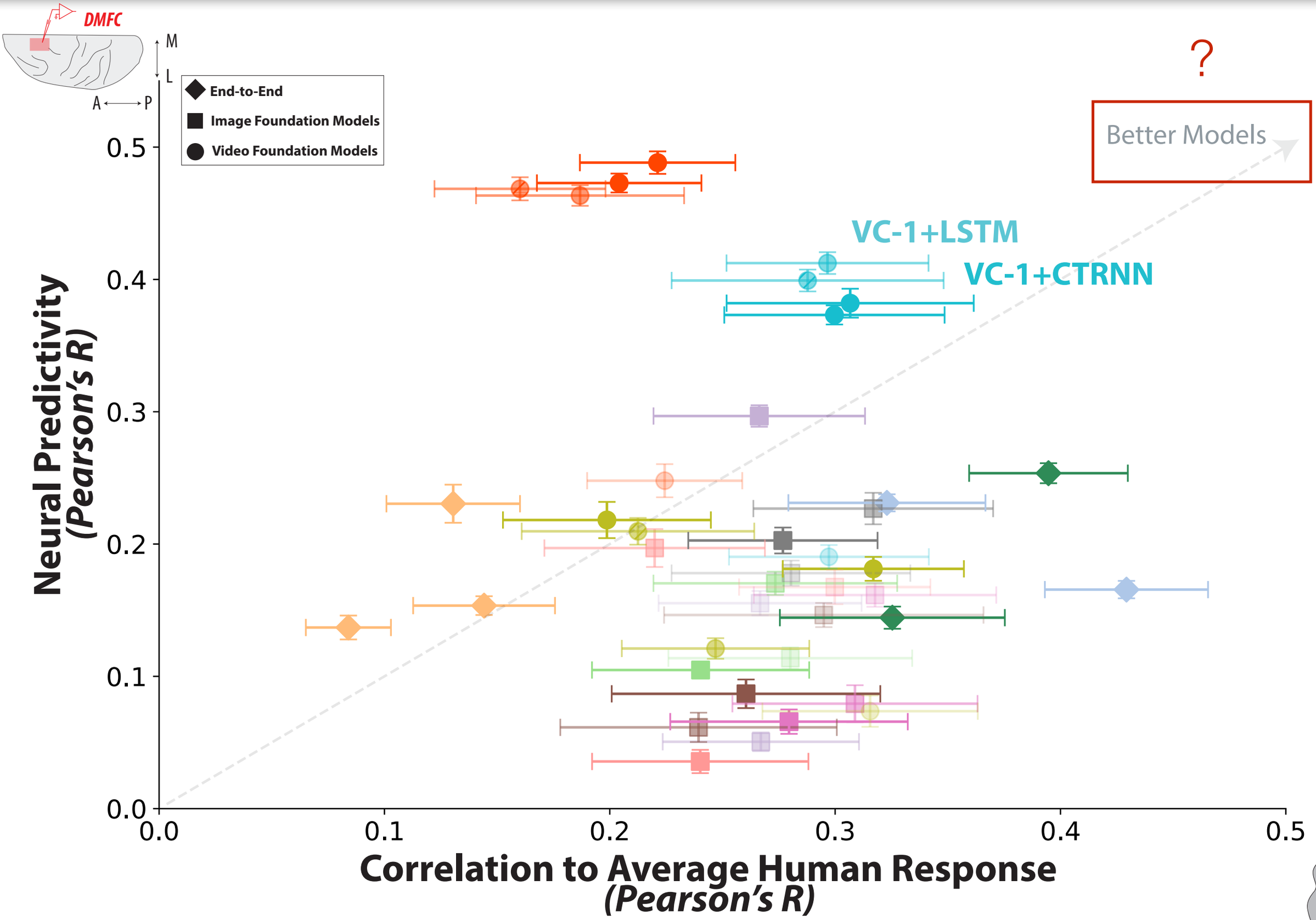
Future Directions

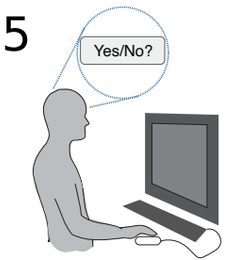
Future Directions





Future Directions



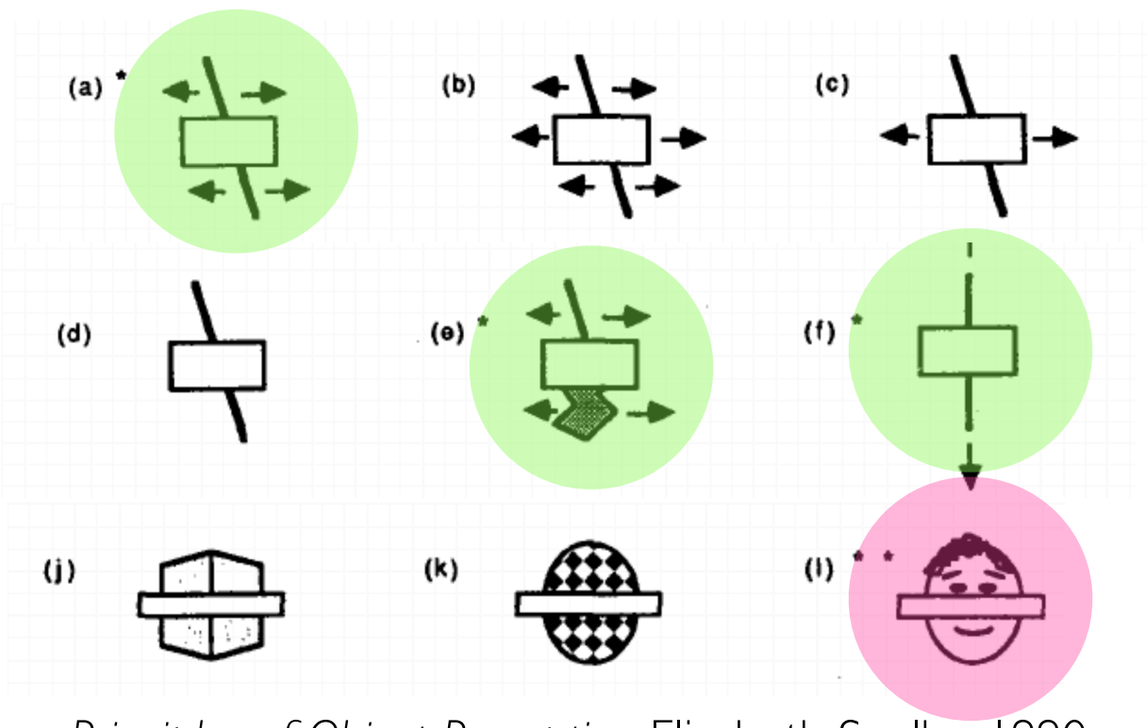


Future Directions

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

Future Directions

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



Principles of Object Perception Elizabeth Spelke, 1990

Elizabeth Spelke

Future Directions

1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation: **object-based, video foundation model**?
2. **<u>Cognitive:</u>** Hierarchy/modularization of timescales in dynamics?

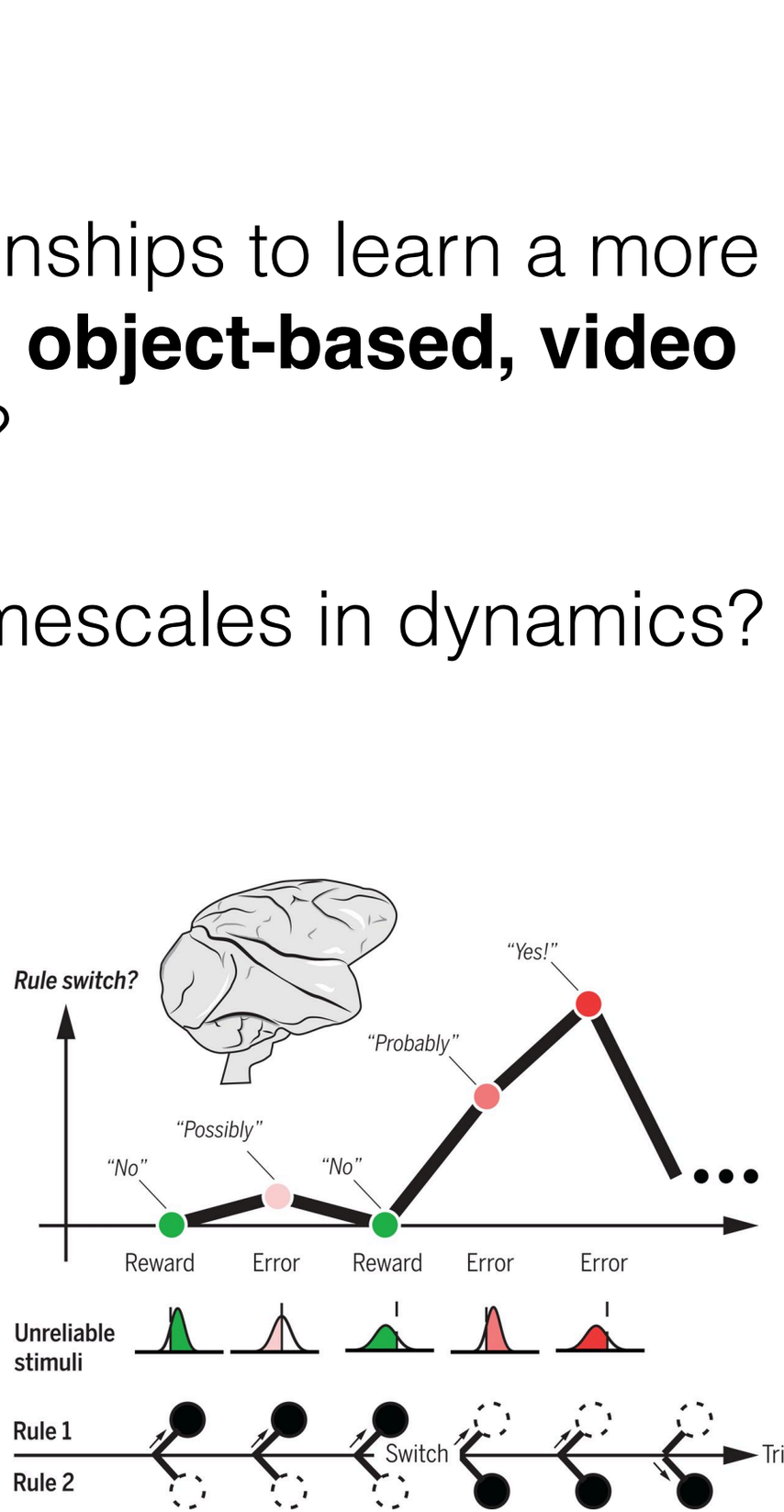
Future Directions

1. **<u>Sensory:</u>** Better leverage temporal relationships to learn a more "factorized" *and* reusable representation: **object-based, video foundation model**?
2. **<u>Cognitive:</u>** Hierarchy/modularization of timescales in dynamics?

Hierarchical reasoning by neural circuits in the frontal cortex

MORTEZA SARAFYAZD AND MEHRDAD JAZAYERI Authors Info & Affiliations

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



Future Directions

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

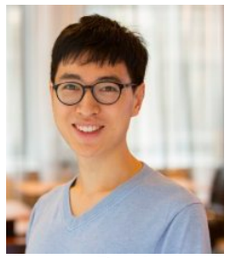
2. **<u>Cognitive:</u>** Hierarchy/modularization of timescales in dynamics?

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

Acknowledgements


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

Contact:
anayebi@mit.edu
@aran_nayebi

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

K. LISA YANG ICoN CENTER
Integrative Computational Neuroscience

MCGOVERN INSTITUTE

brain+cognitive sciences

SIMONS FOUNDATION

THE HELEN HAY WHITNEY FOUNDATION

JazLab


YangLab

MIT