Using *Embodied* Agents to Reverse-Engineer *Natural* Intelligence

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

RI Seminar
2025.09.26

MACHINE LEARNING DEPARTMENT

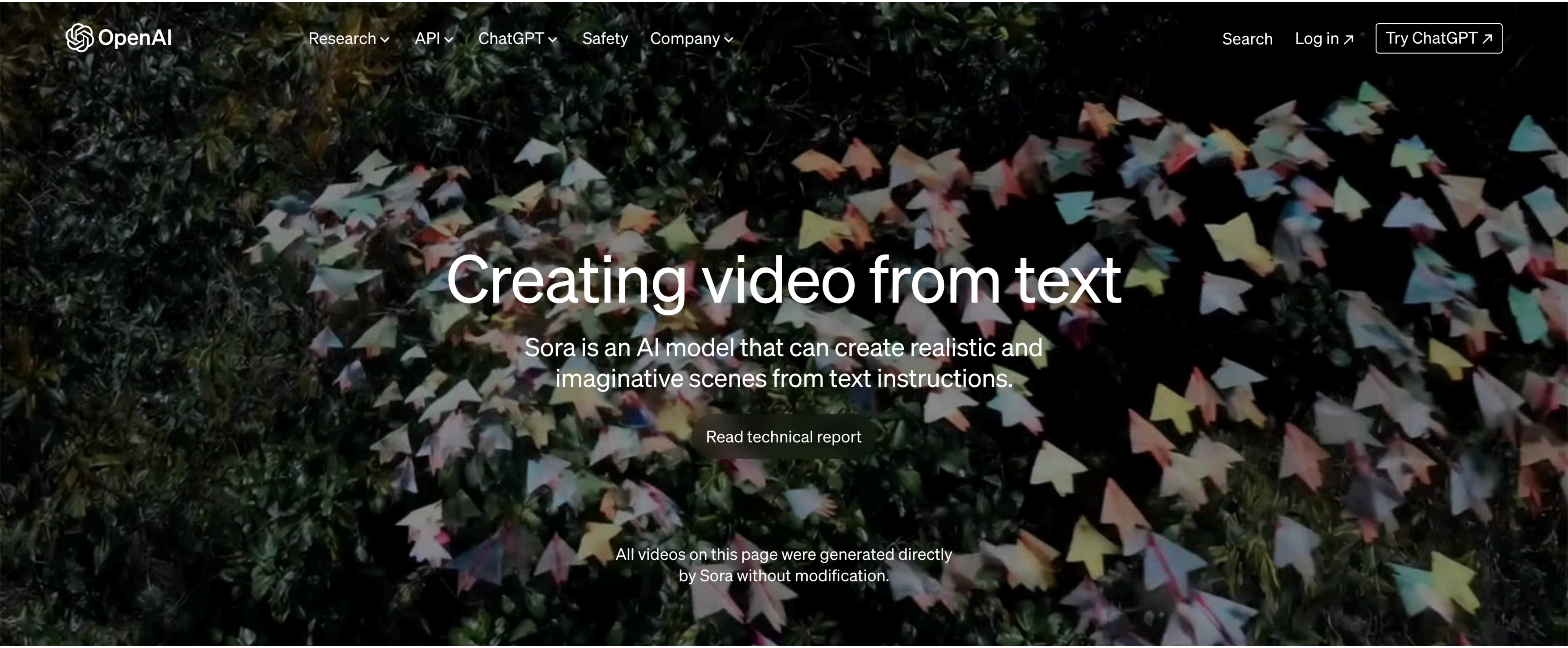
Carnegie Mellon University
Neuroscience Institute

THE ROBOTICS INSTITUTE

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Current AI Struggles to Understand the Physical World

OpenAI Sora, February 2024



Current AI Struggles to Understand the Physical World

OpenAI Sora,
February 2024

Current AI Struggles to Understand the Physical World

OpenAI Sora,
February 2024

Current AI Struggles to Understand the Physical World

OpenAI Sora, February 2024

Q: What's missing?

Current AI Struggles to Understand the Physical World

OpenAI Sora,
February 2024

Q: What's missing?

A: Embodied agency & interaction.

Why Reverse-Engineer Natural Intelligence?

Why?

Why Reverse-Engineer Natural Intelligence?

Why?

Animals & humans (currently) perform behaviors we've yet to engineer successfully in AI agents:

Why Reverse-Engineer Natural Intelligence?

Why?

Animals & humans (currently) perform behaviors we've yet to engineer successfully in AI agents:

- Prediction (requires world modeling) & planning (requires memory)
- Adaptive motor control (requires embodiment)
- Autonomy / online life-long learning (test-time reasoning is just the beginning: need to update the weights without forgetting everything!)

Why Reverse-Engineer Natural Intelligence?

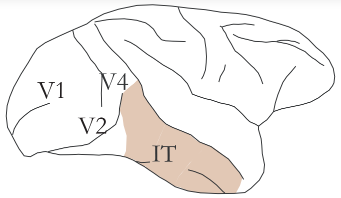
Why?

Animals & humans (currently) perform behaviors we've yet to engineer successfully in AI agents:

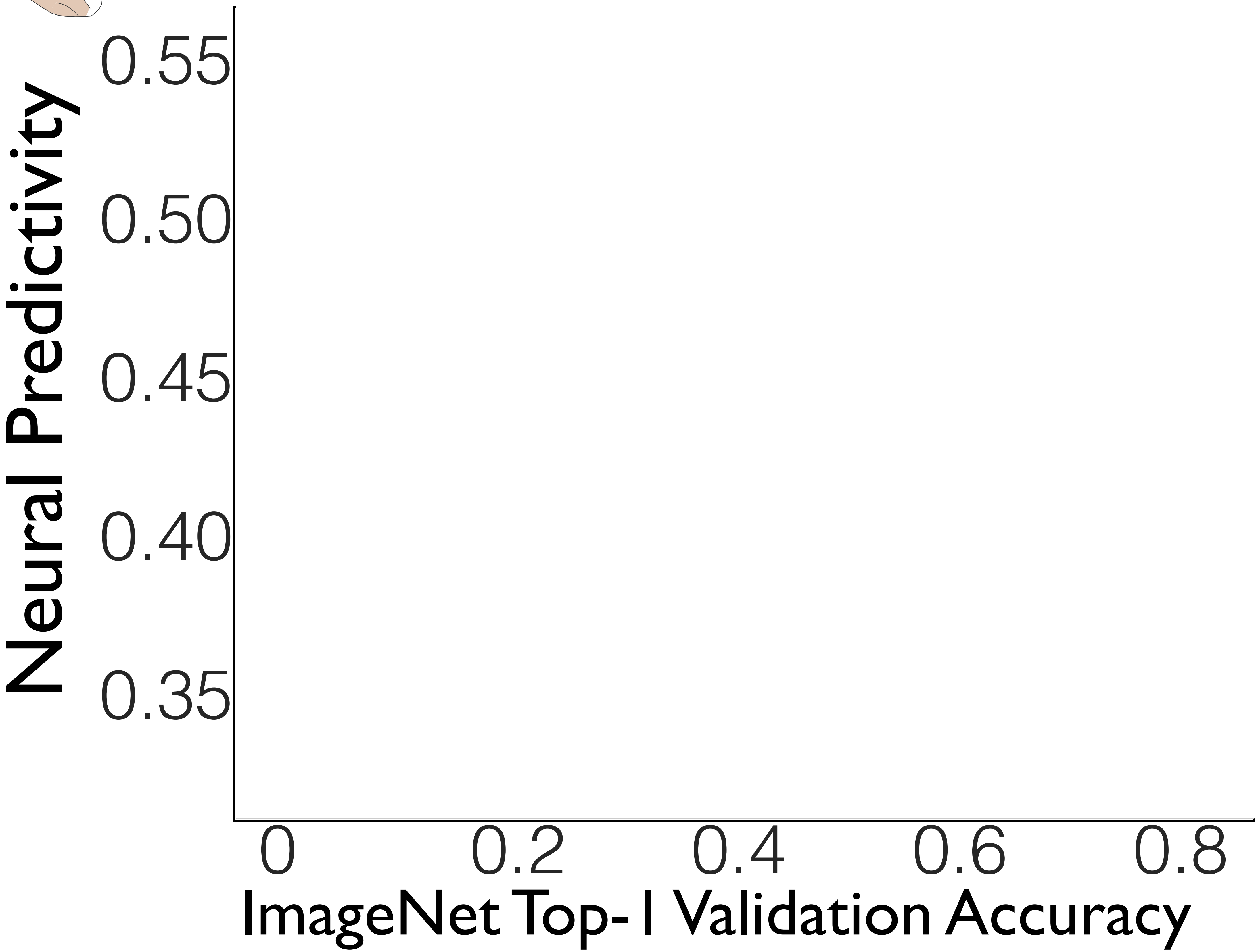
- Prediction (requires world modeling) & planning (requires memory)
- Adaptive motor control (requires embodiment)
- Autonomy / online life-long learning (test-time reasoning is just the beginning: need to update the weights without forgetting everything!)

The specific *capabilities* of humans & animals become our concrete engineering targets!

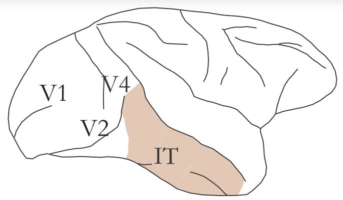
Task Performance Correlated with Neural Predictivity



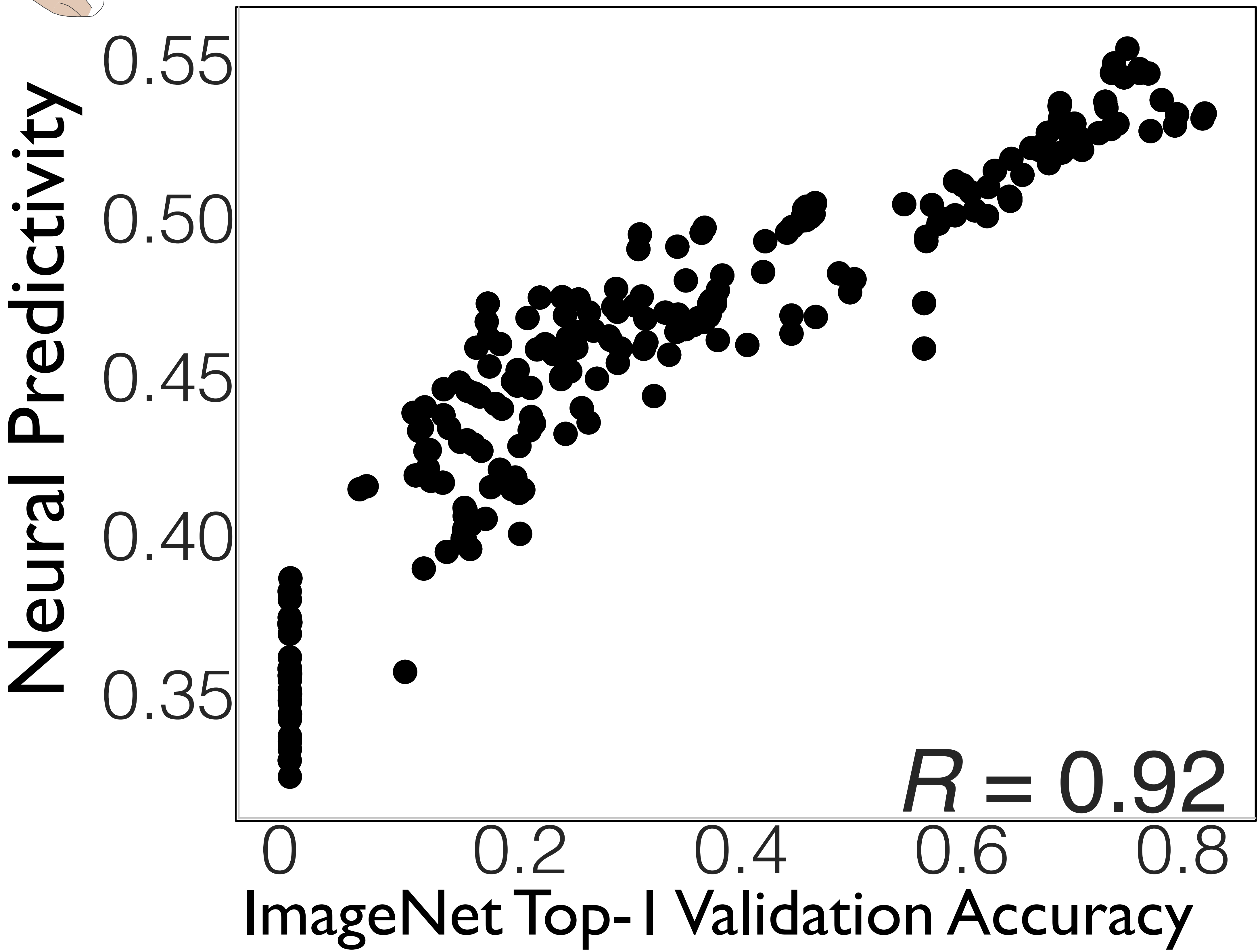
Schrimpf, Kubilius* et al. 2018*



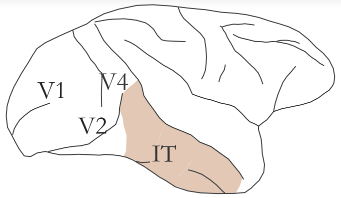
Task Performance Correlated with Neural Predictivity



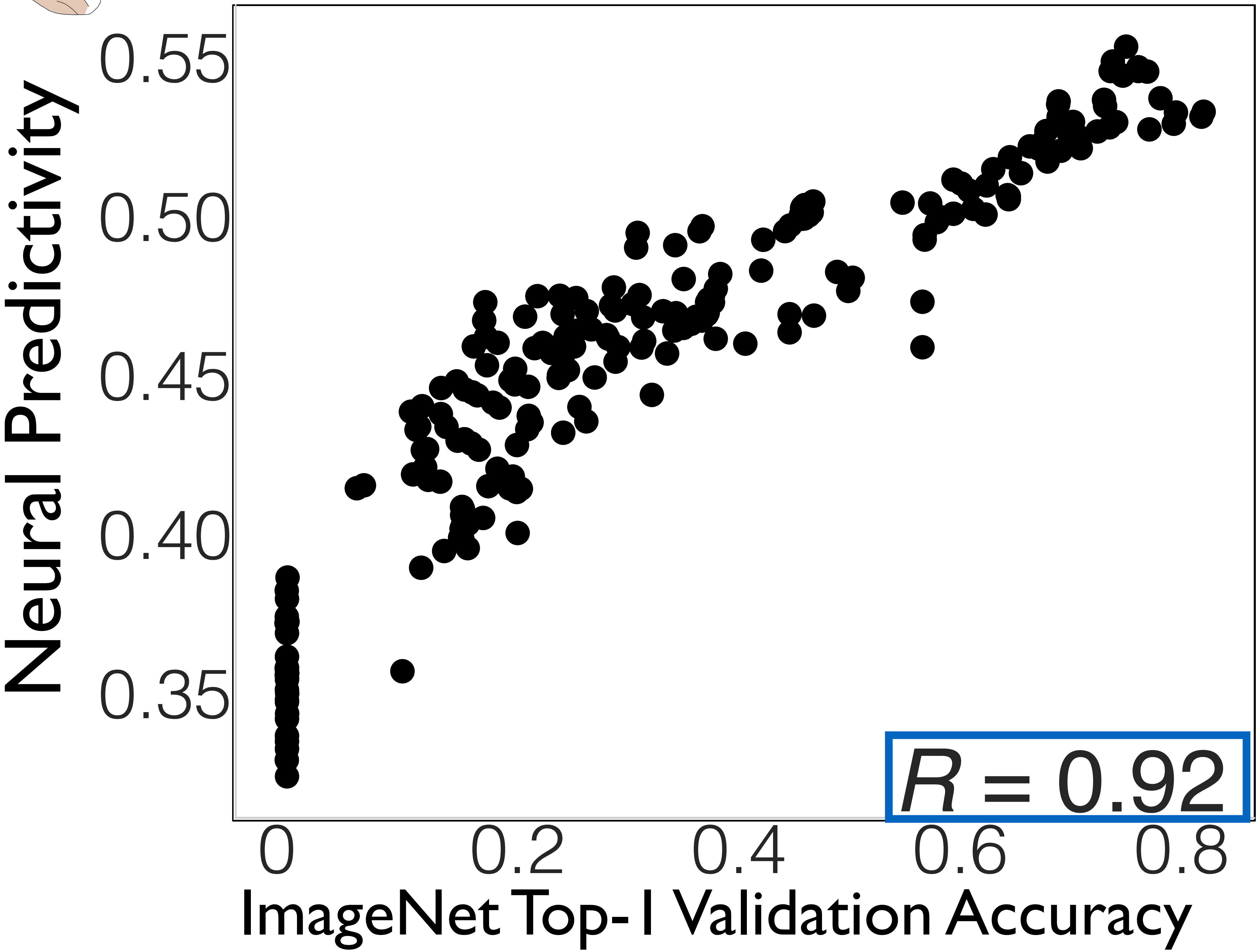
Schrimpf, Kubilius* et al. 2018*



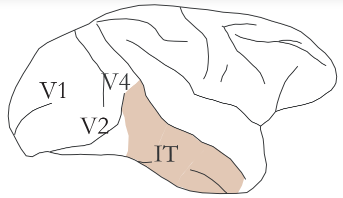
Task Performance Correlated with Neural Predictivity



Schrimpf, Kubilius* et al. 2018*

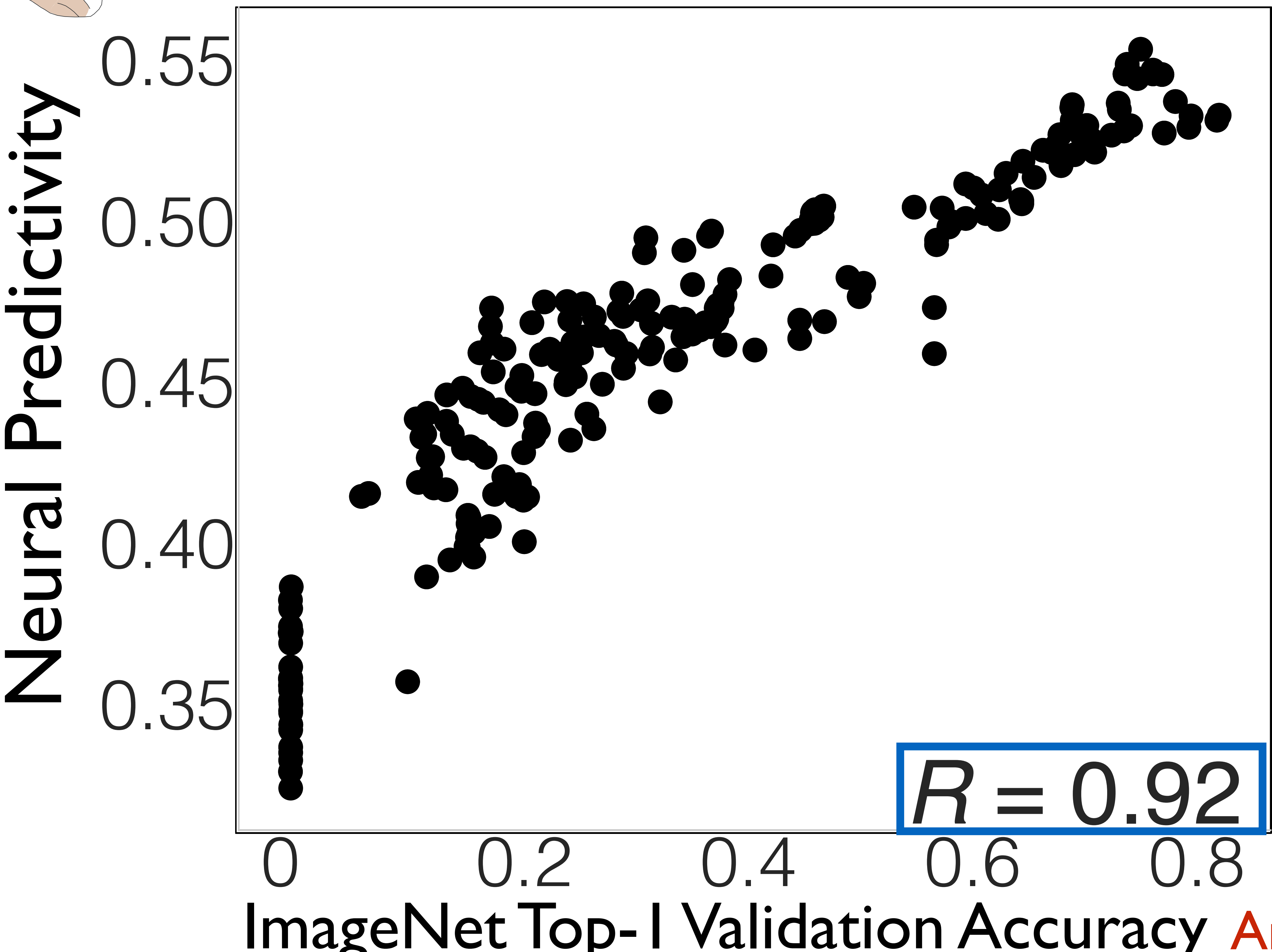


Task Performance Correlated with Neural Predictivity



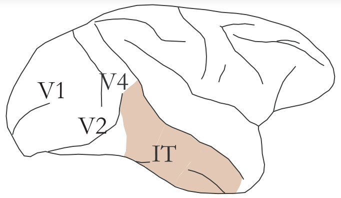
A Neuroscience Goal

Schrimpf, Kubilius* et al. 2018*



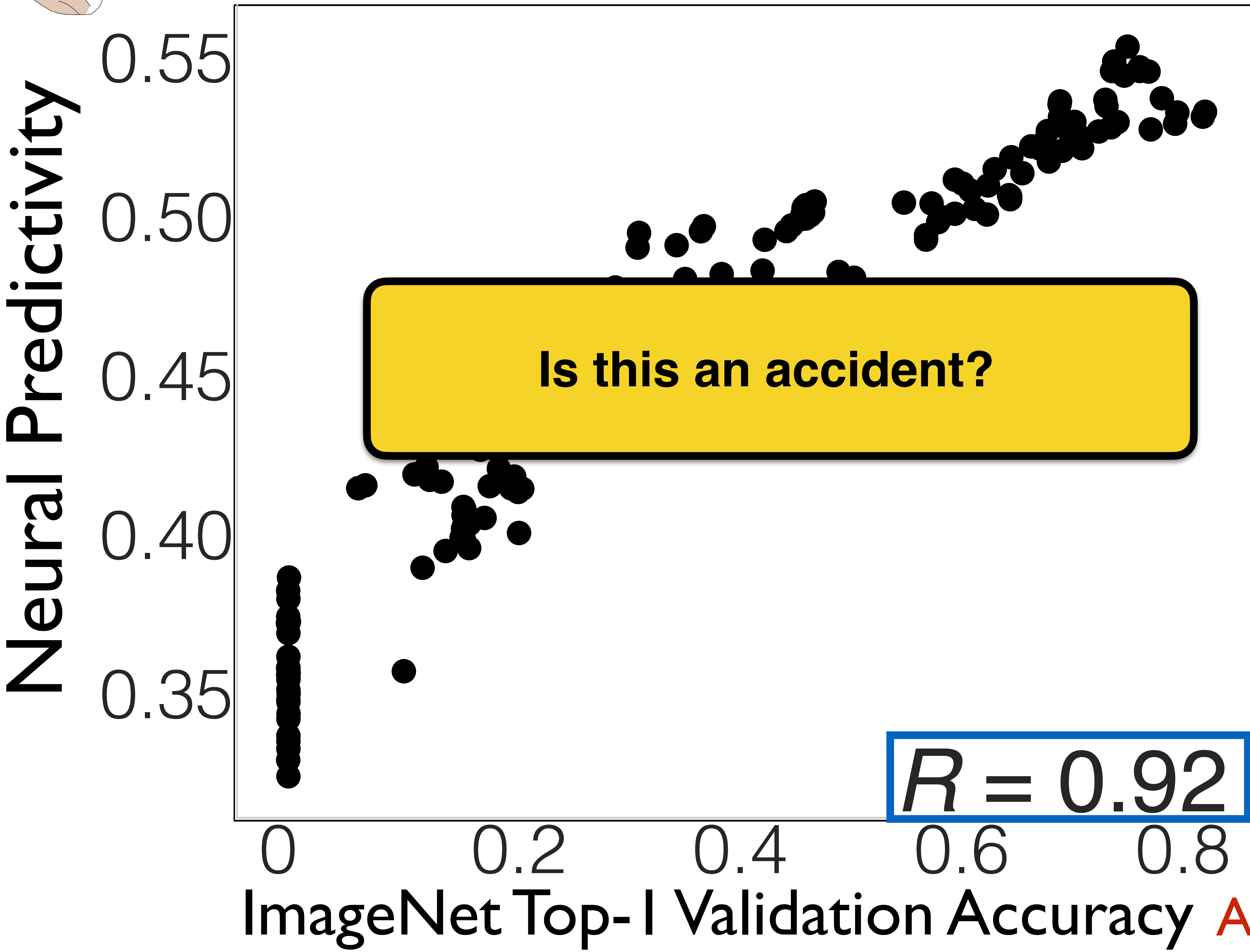
An AI Goal

Task Performance Correlated with Neural Predictivity



A Neuroscience Goal

Schrimpf, Kubilius* et al. 2018*



An AI Goal

Task-Optimized Modeling: Four Components

Task-Optimization (ML)

1.

A = *architecture class*

2.

T = *task loss*

3.

D = *dataset*

4.

L = *learning rule*

Task-Optimized Modeling: Four Components

Task-Optimization (ML)

Neurobiology

1.

A = *architecture class*

2.

T = *task loss*

3.

D = *dataset*

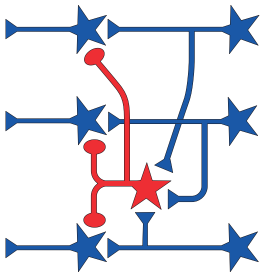
4.

L = *learning rule*

Task-Optimized Modeling: Four Components

Task-Optimization (ML)

Neurobiology



1.

A = *architecture class* ***= circuit neuroanatomy***

2.

T = *task loss*

3.

D = *dataset*

4.

L = *learning rule*

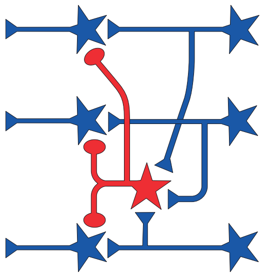
Task-Optimized Modeling: Four Components

Task-Optimization (ML)

Neurobiology

1.

A = *architecture class* **= *circuit neuroanatomy***



2.

T = *task loss* **= *ecological niche/behavior***

3.

D = *dataset*

4.

L = *learning rule*

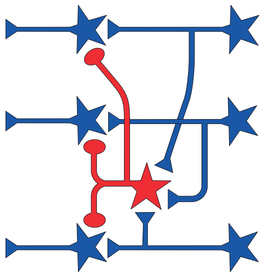
Task-Optimized Modeling: Four Components

Task-Optimization (ML)

Neurobiology

1.

A = *architecture class* **= *circuit neuroanatomy***

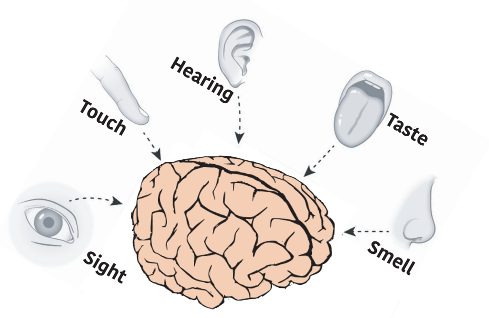


2.

T = *task loss* **= *ecological niche/behavior***

3.

D = *dataset* **= *environment***



4.

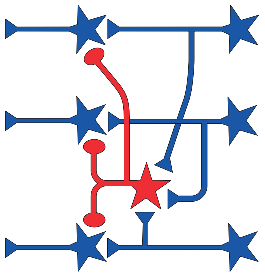
L = *learning rule*

Task-Optimized Modeling: Four Components

Task-Optimization (ML) | **Neurobiology**

1.

A = *architecture class* **= *circuit neuroanatomy***

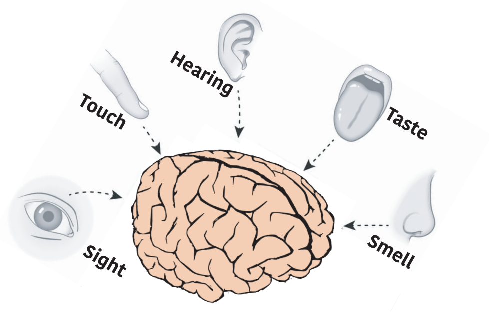


2.

T = *task loss* **= *ecological niche/behavior***

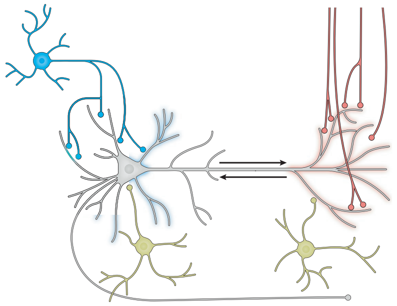
3.

D = *dataset* **= *environment***



4.

L = *learning rule* **= *natural selection + synaptic plasticity***



Task-Optimized Modeling: Four Components

***L** = learning rule*

"Natural selection + plasticity"

***T** = task loss*

"Ecological niche/ behavior"

"Environment"

***D** = data stream*

"Circuit"

***A** = architecture class*

Task-Optimized Modeling: Four Components

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"

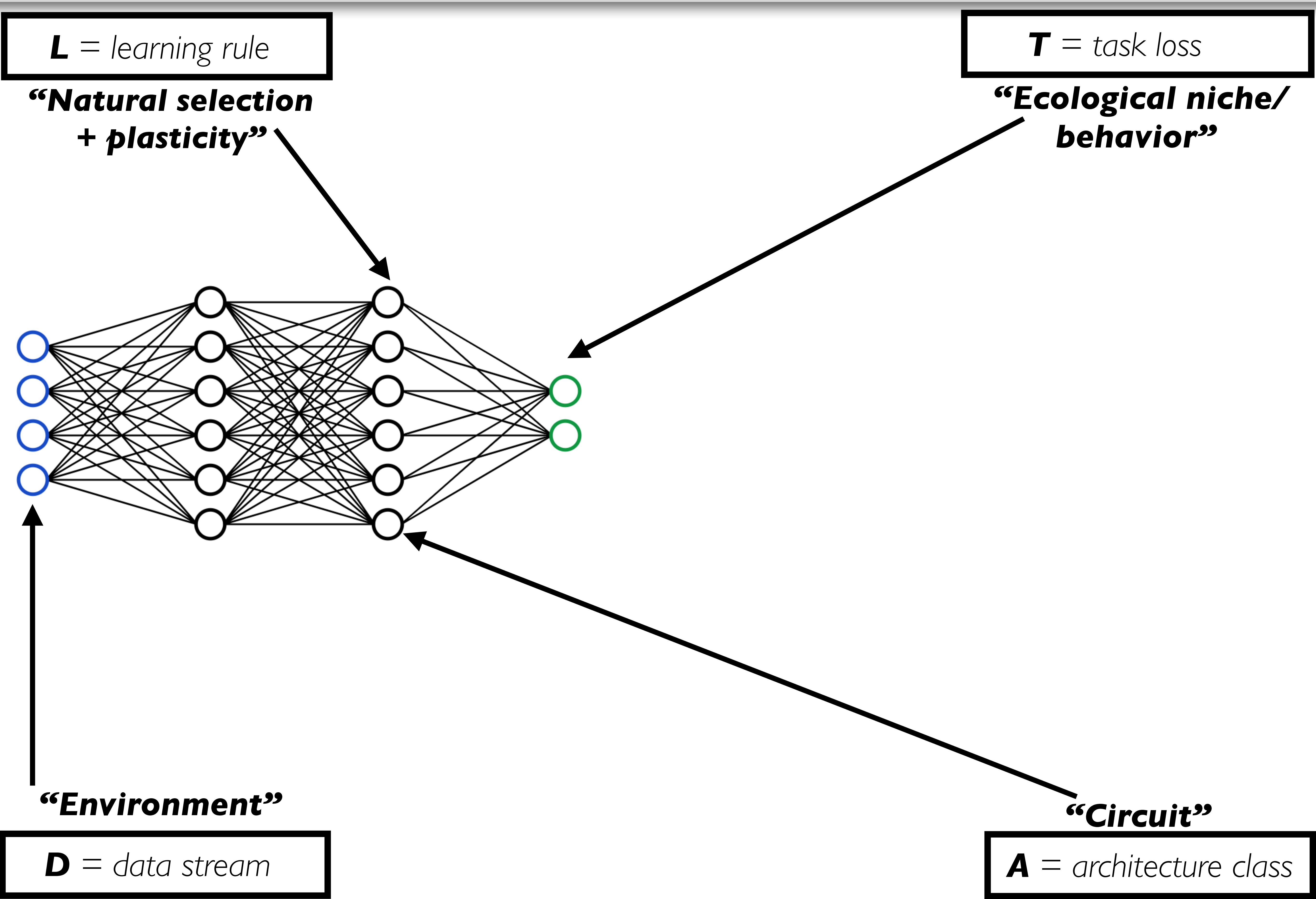
"Environment"

D = *data stream*

"Circuit"

A = *architecture class*

Task-Optimized Modeling: Four Components



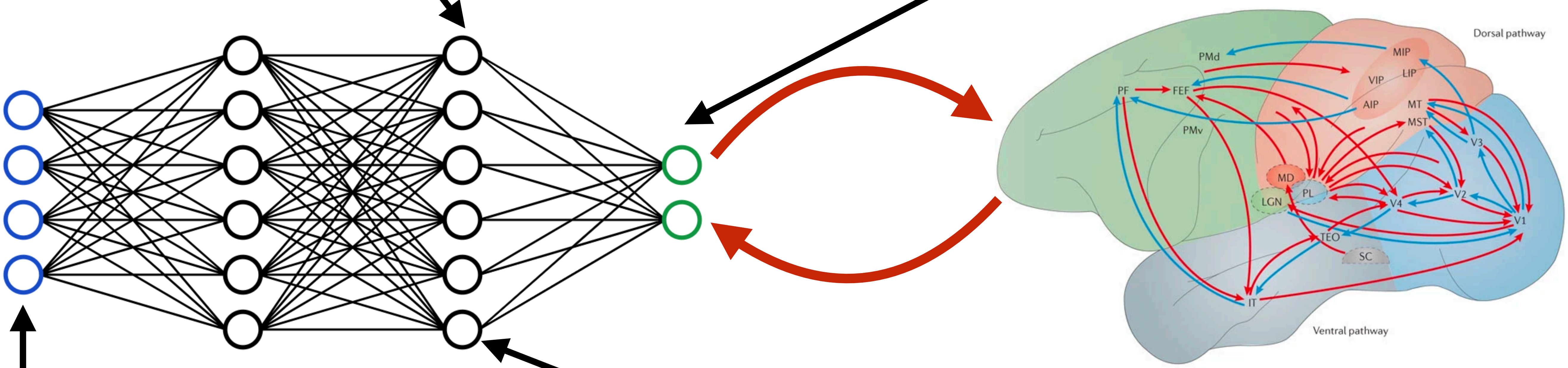
Task-Optimized Modeling: Four Components

L = *learning rule*

"Natural selection + plasticity"

T = *task loss*

"Ecological niche/ behavior"



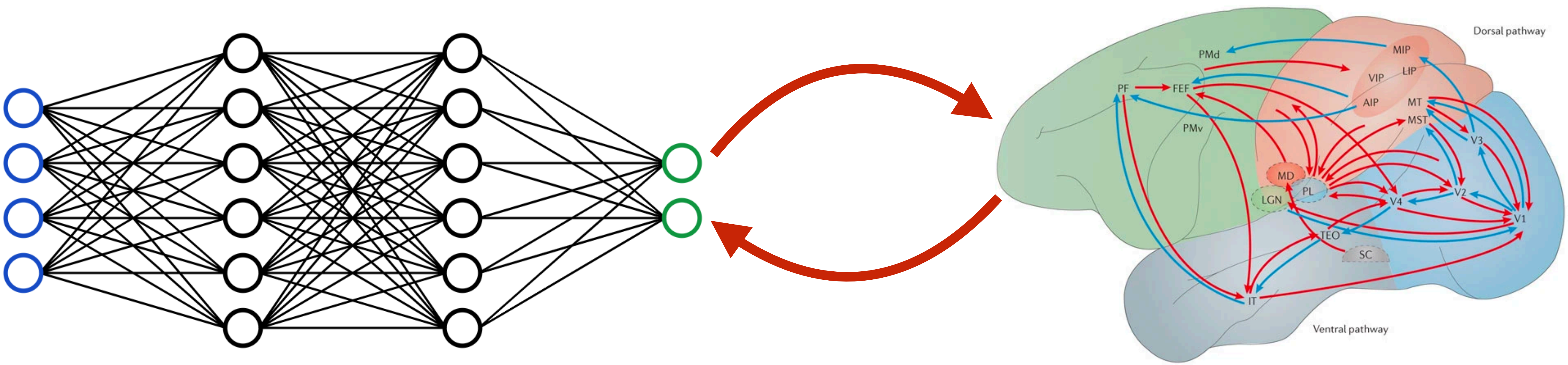
"Environment"

D = *data stream*

"Circuit"

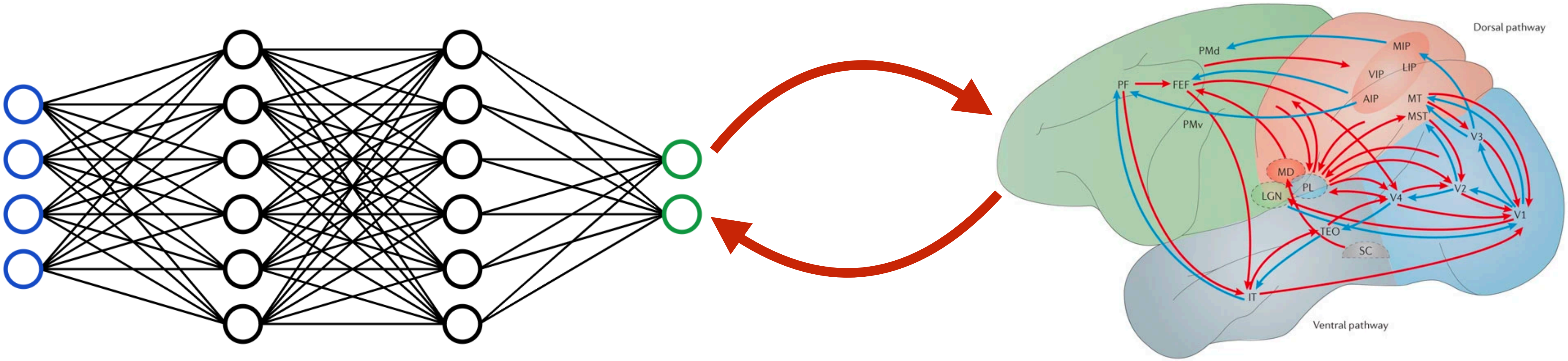
A = *architecture class*

Task-Optimized Modeling



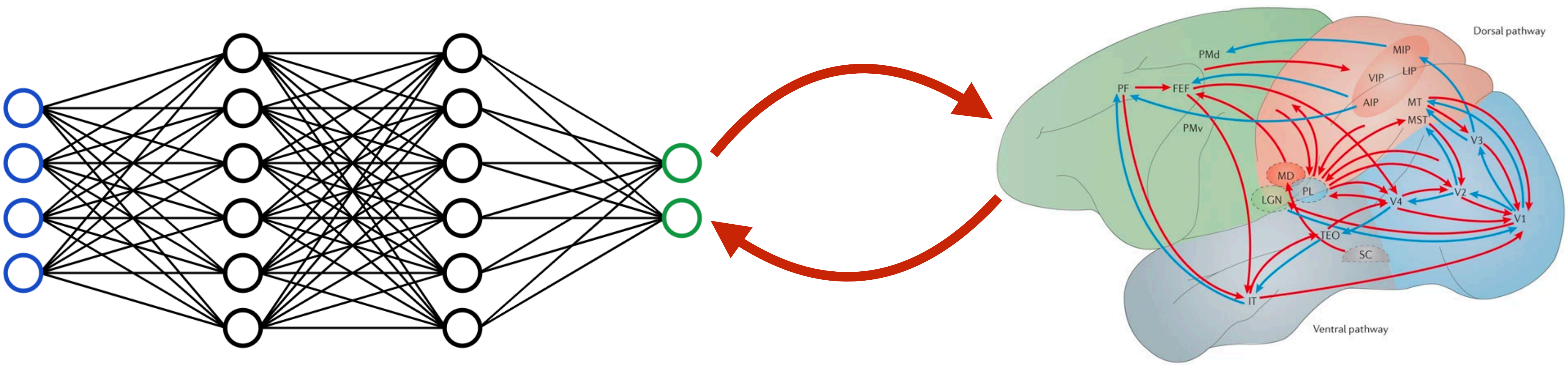
Task-Optimized Modeling

Design ML Algorithms Optimized to Perform Organism's Behavior under Organism's Constraints



Task-Optimized Modeling

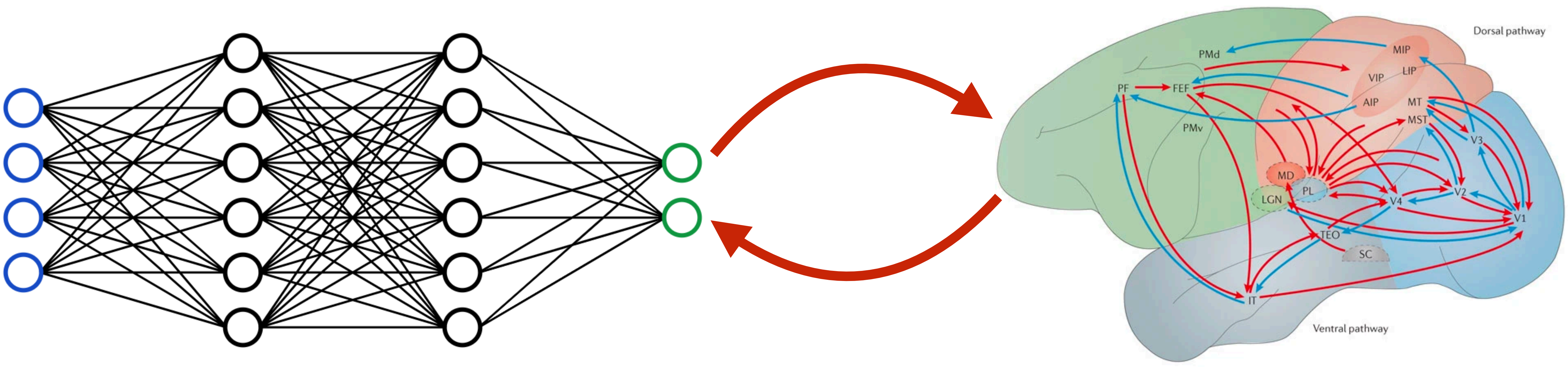
Design ML Algorithms Optimized to Perform Organism's Behavior under Organism's Constraints



Yields:

Task-Optimized Modeling

Design ML Algorithms Optimized to Perform Organism's Behavior under Organism's Constraints

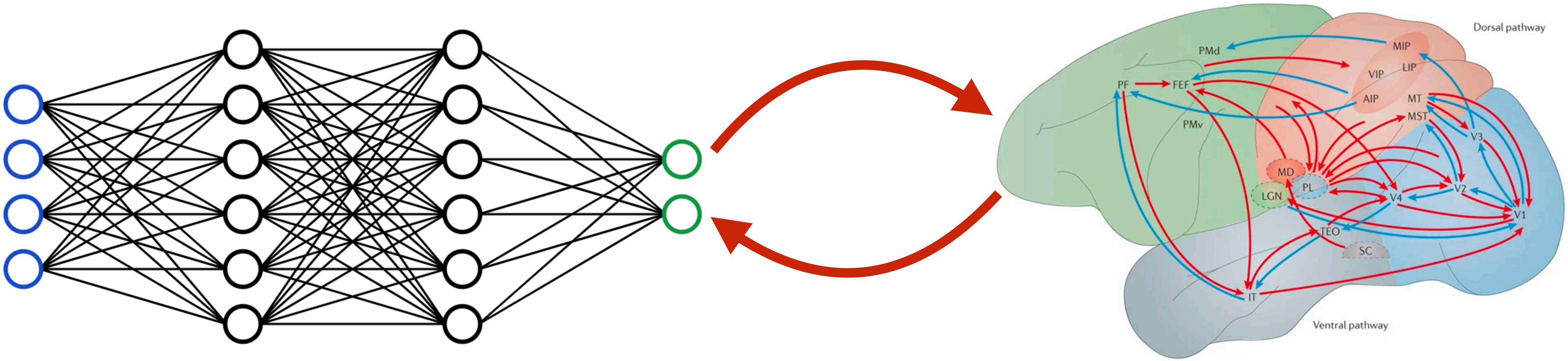


Yields:

Quantitatively Accurate & Practically Useful Brain Models

Task-Optimized Modeling

Design ML Algorithms Optimized to Perform Organism's Behavior under Organism's Constraints



Yields:

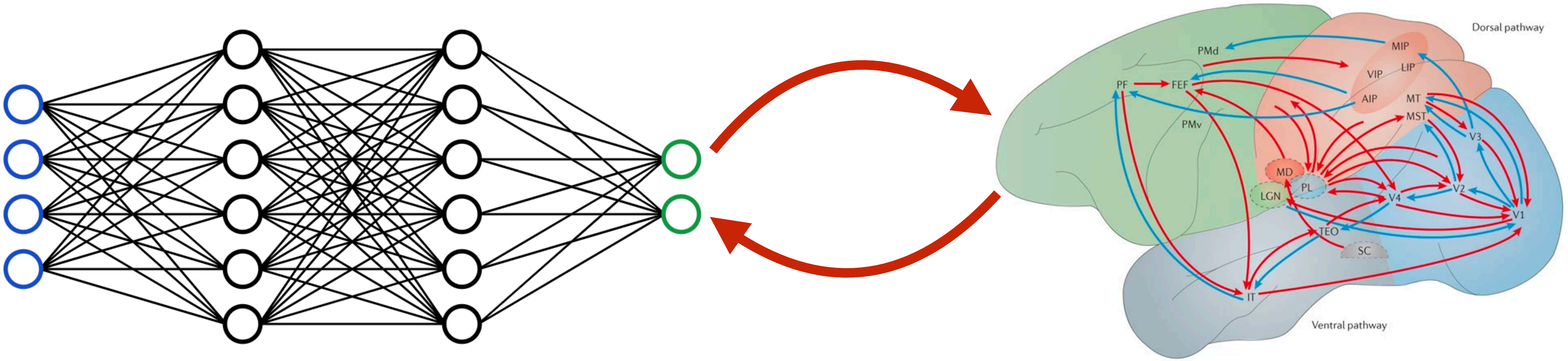
Quantitatively Accurate & Practically Useful Brain Models

AND

Principles of *Why* Neural Responses Are As They Are

Task-Optimized Modeling

Design ML Algorithms Optimized to Perform Organism's Behavior under Organism's Constraints



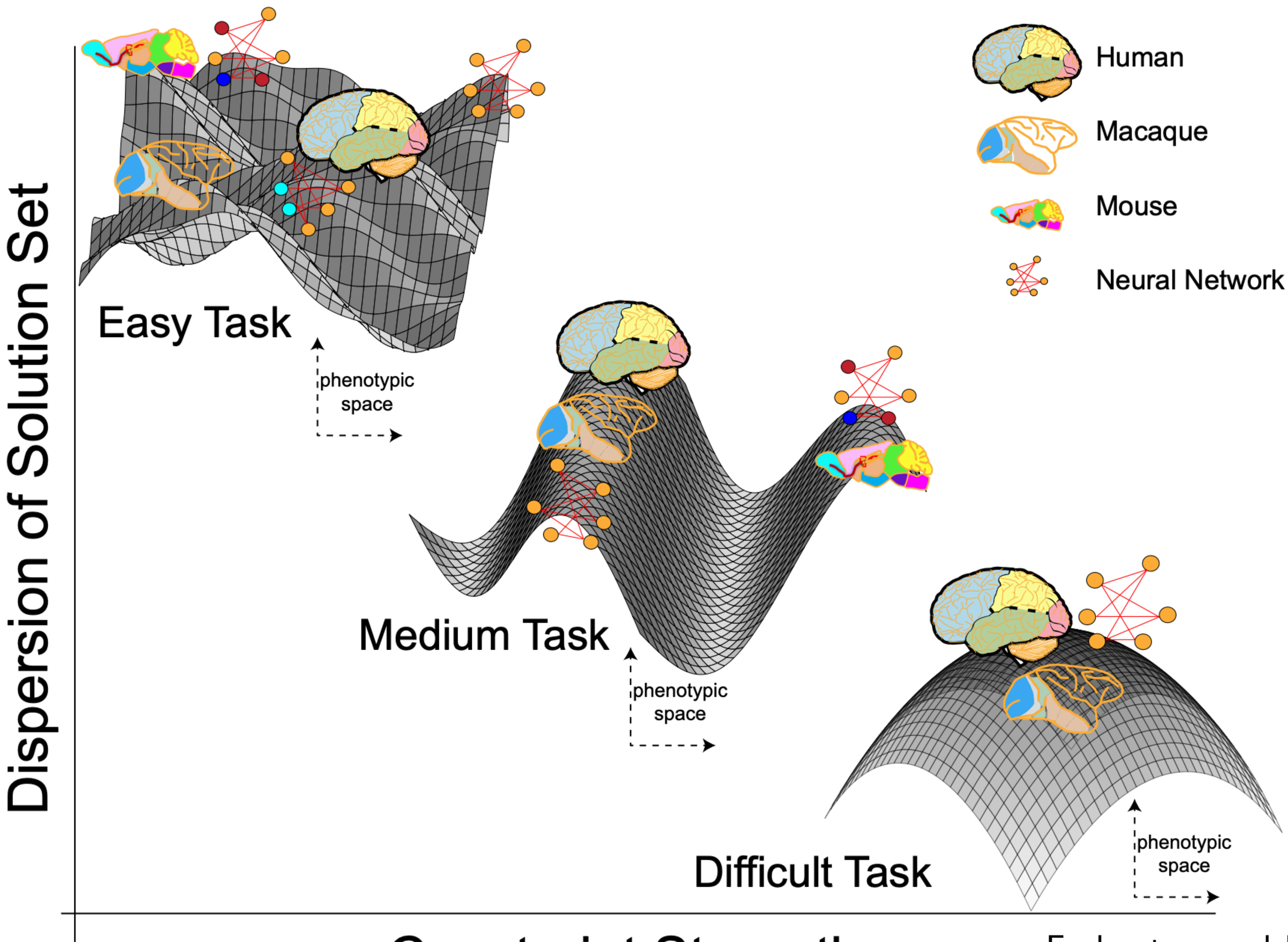
Yields:

Quantitatively Accurate & Practically Useful Brain Models

AND

Principles of *Why* Neural Responses Are As They Are

Contravariance Principle: The Harder the Task, the Less Solutions!



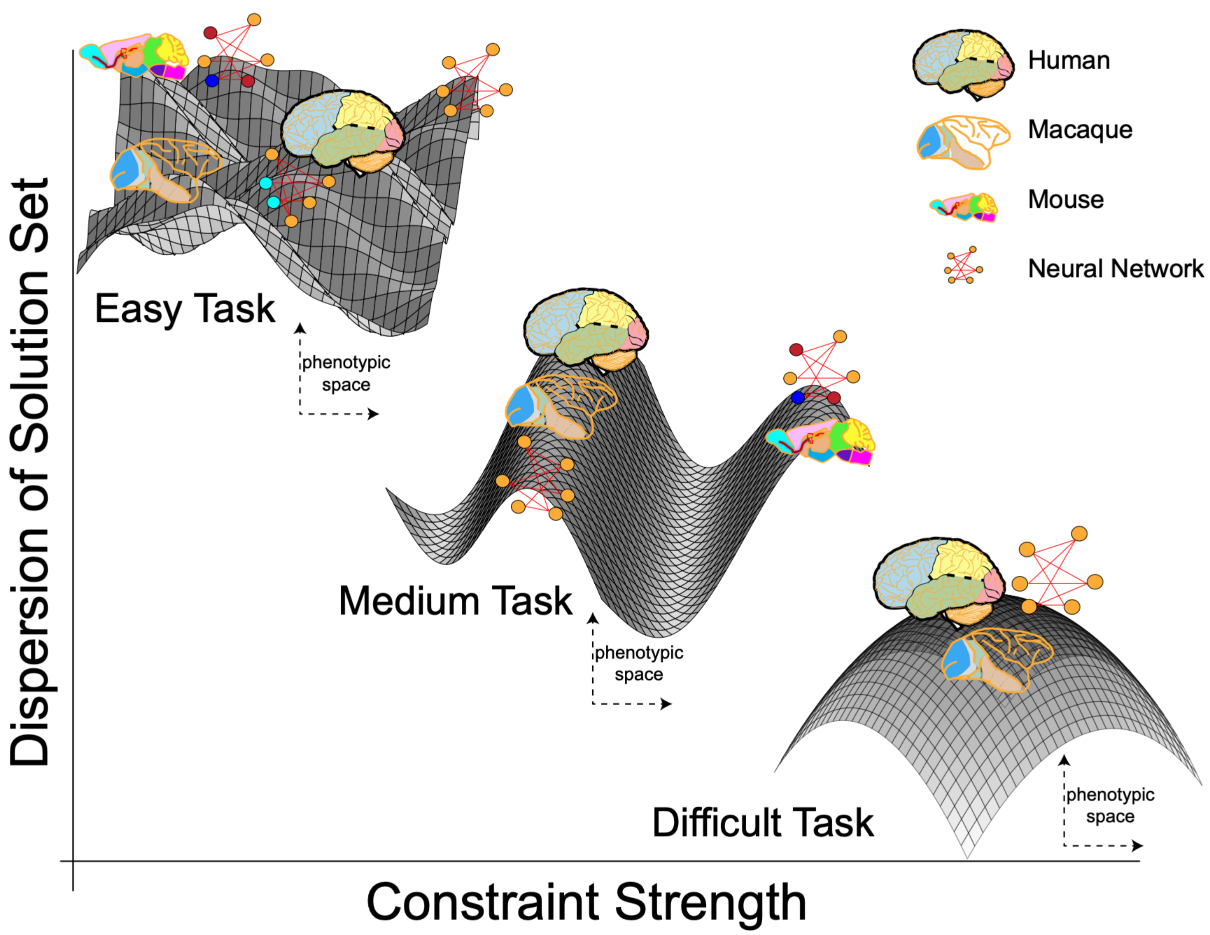
Rosa Cao

Daniel Yamins

Explanatory models in neuroscience:

Part 2 – Constraint-based intelligibility

Platonic Representation Hypothesis is the AI version of Contravariance



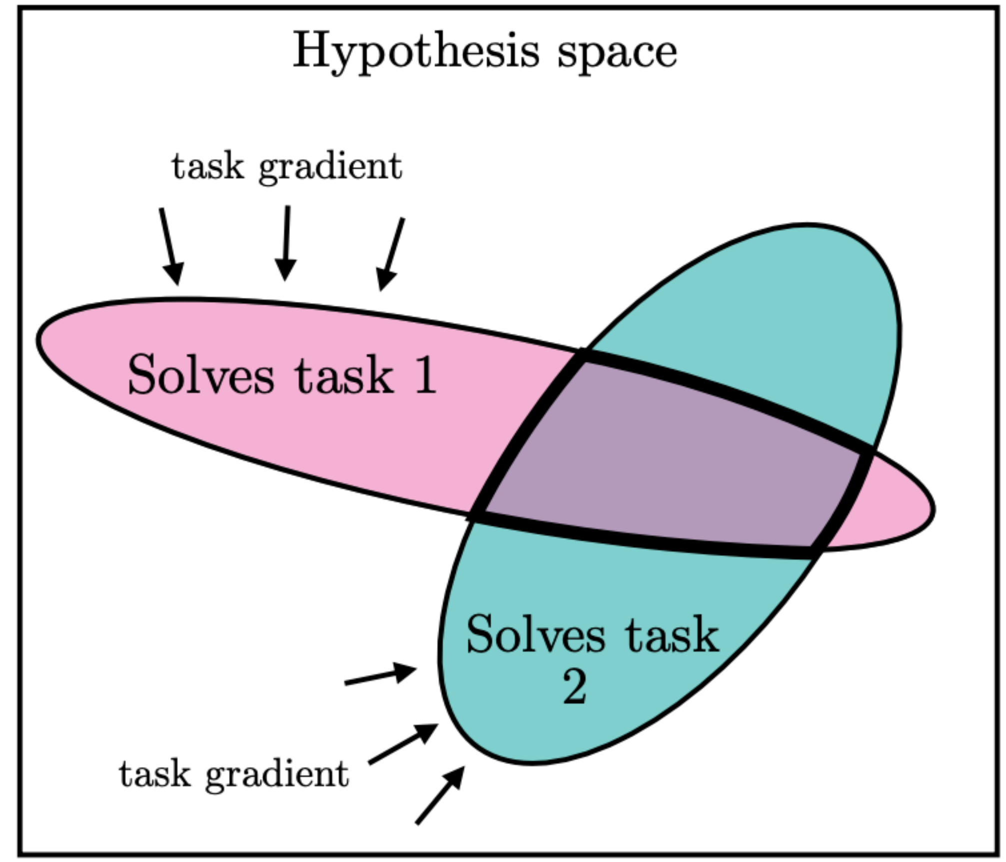


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

The Platonic Representation Hypothesis

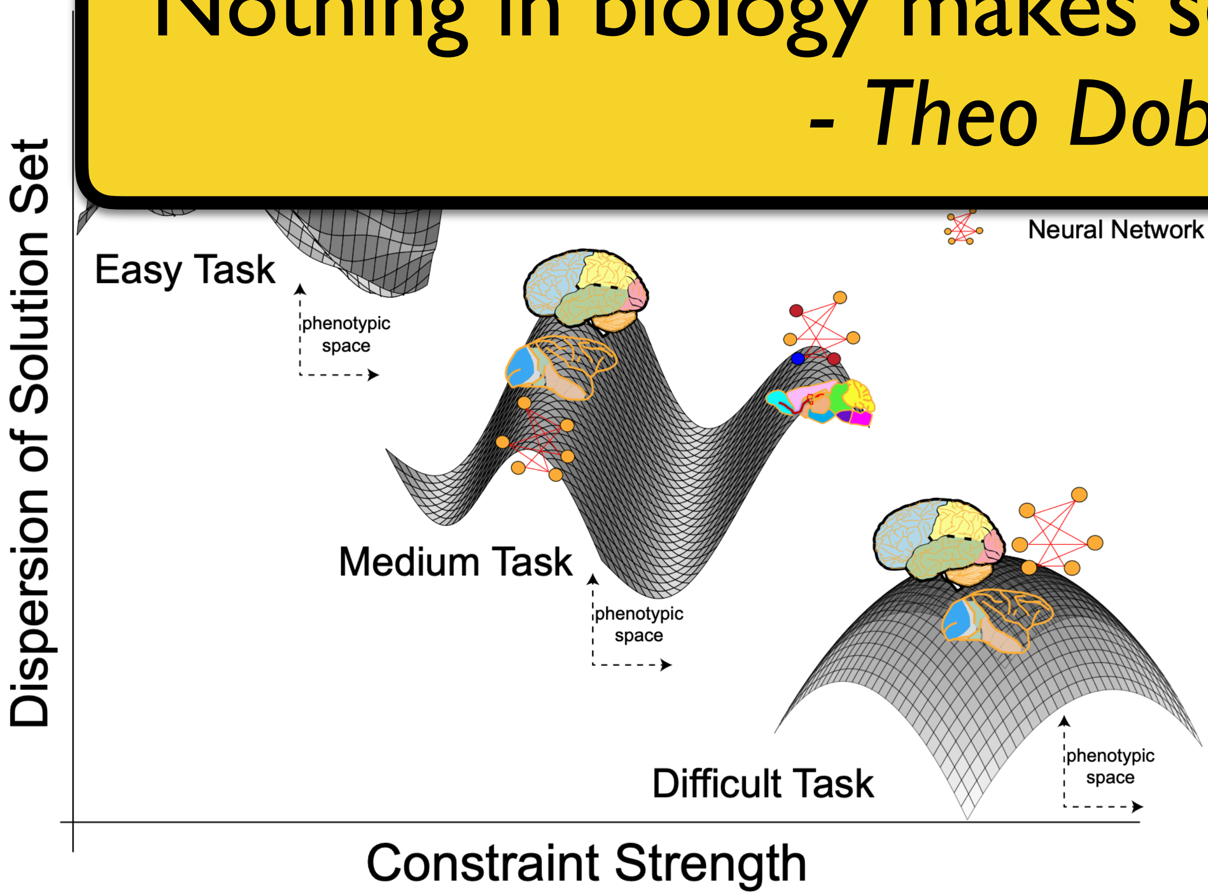
Minyoung Huh [*1] **Brian Cheung** [*1] **Tongzhou Wang** [*1] **Phillip Isola** [*1]

The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

Platonic Representation Hypothesis is the AI version of Contravariance

"Nothing in biology makes sense in light of evolution."
- Theo Dobzhansky



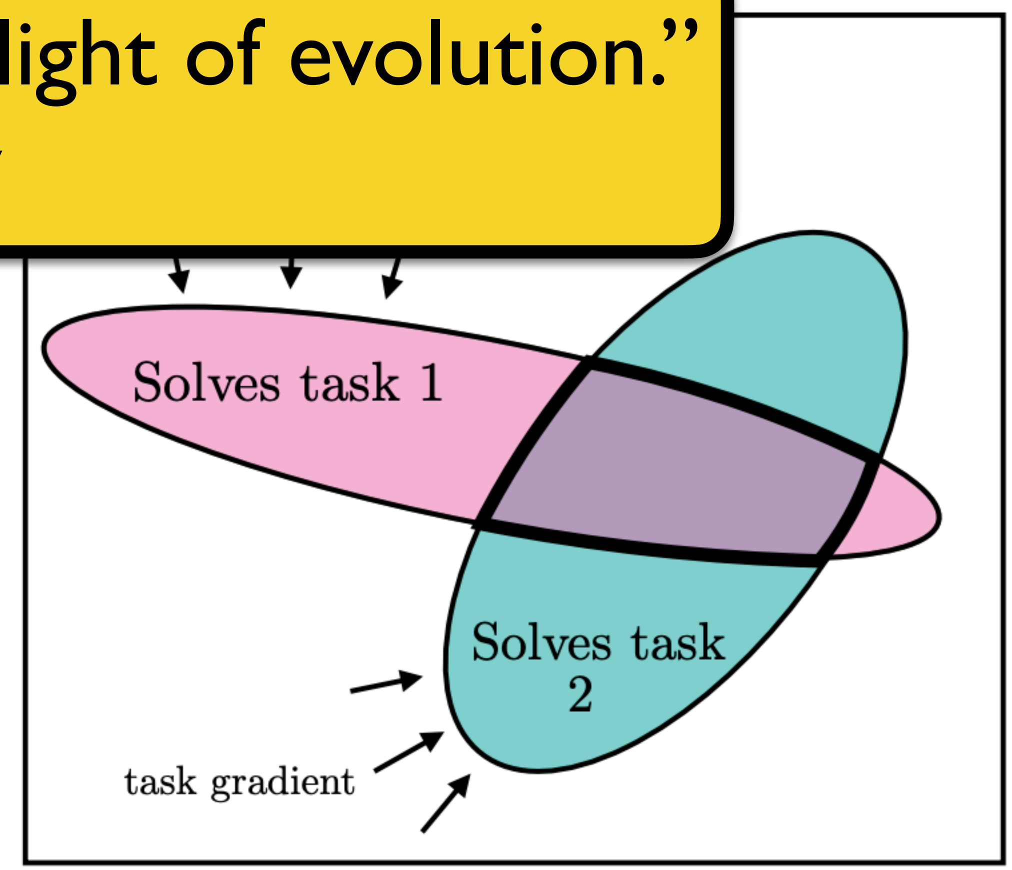


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

The Platonic Representation Hypothesis

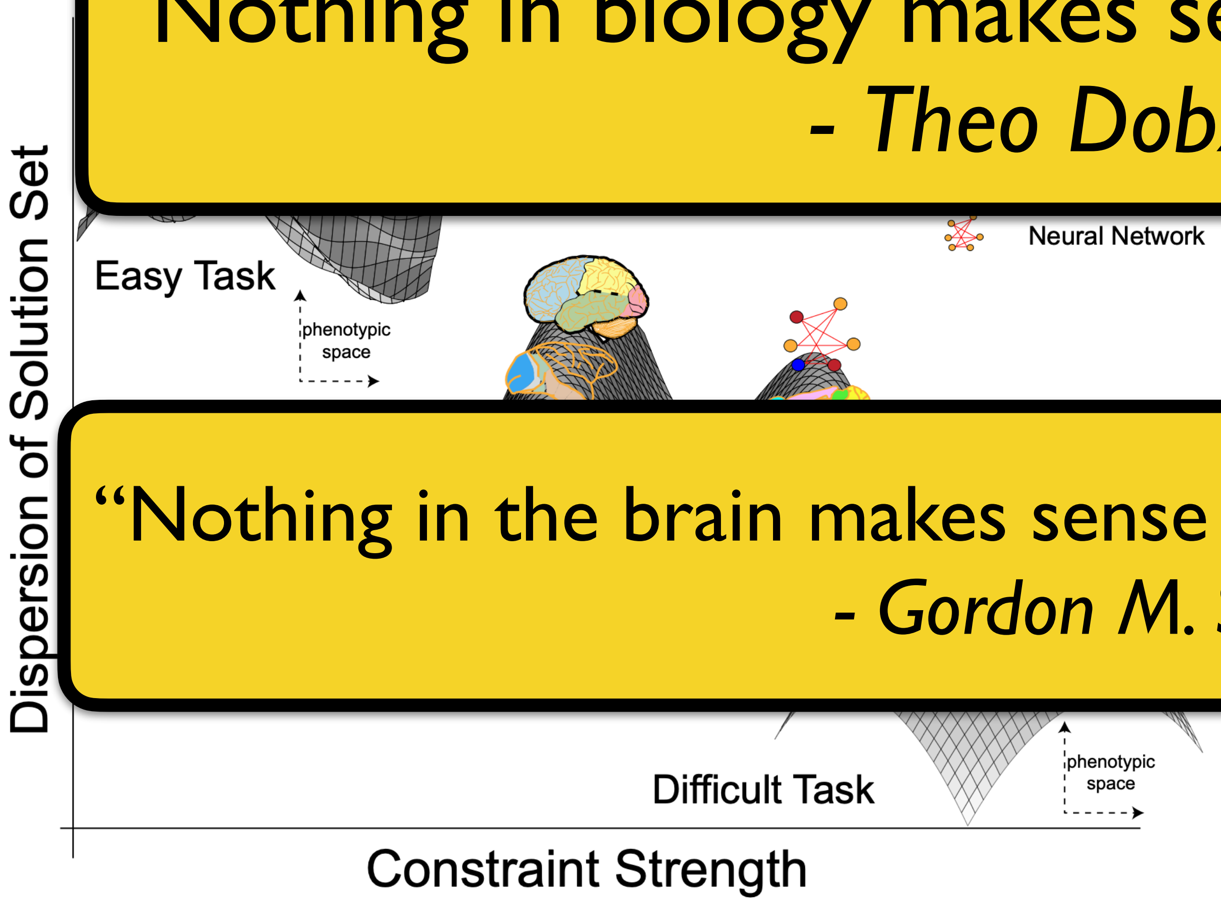
Minyoung Huh [*1] **Brian Cheung** [*1] **Tongzhou Wang** [*1] **Phillip Isola** [*1]

The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

Platonic Representation Hypothesis is the AI version of Contravariance

"Nothing in biology makes sense in light of evolution."
- *Theo Dobzhansky*



"Nothing in the brain makes sense except in the light of behavior."
- *Gordon M. Shepherd*

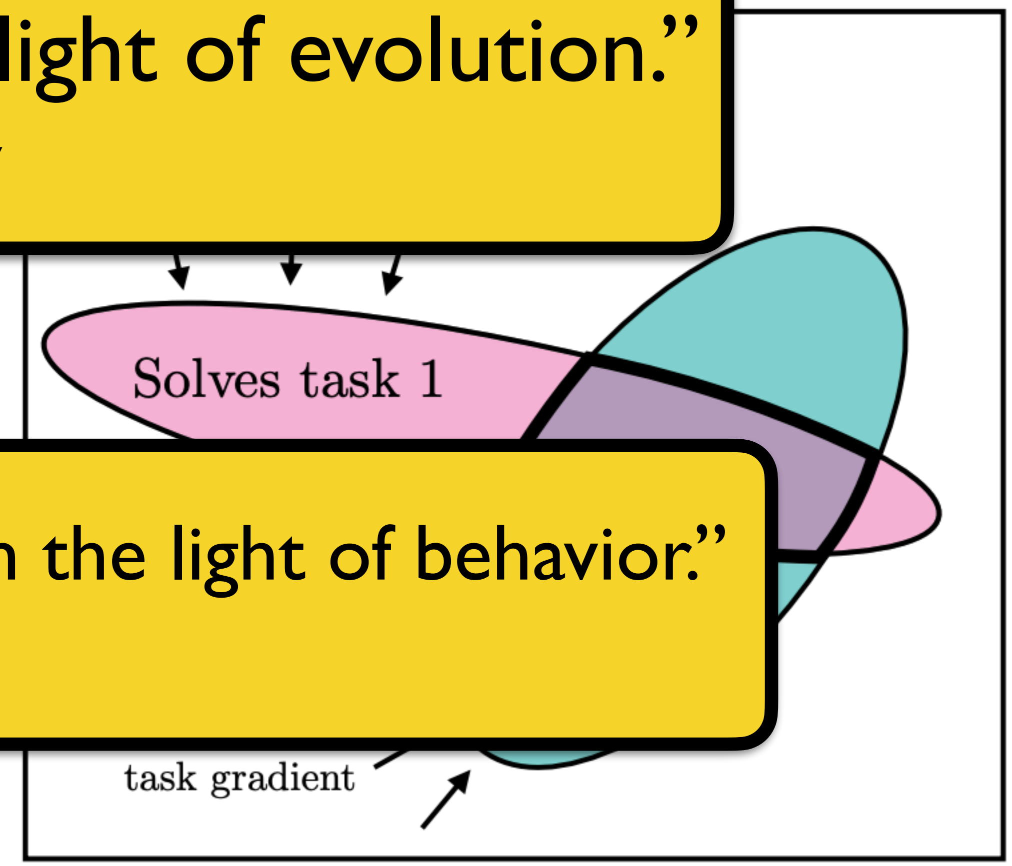


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

The Platonic Representation Hypothesis

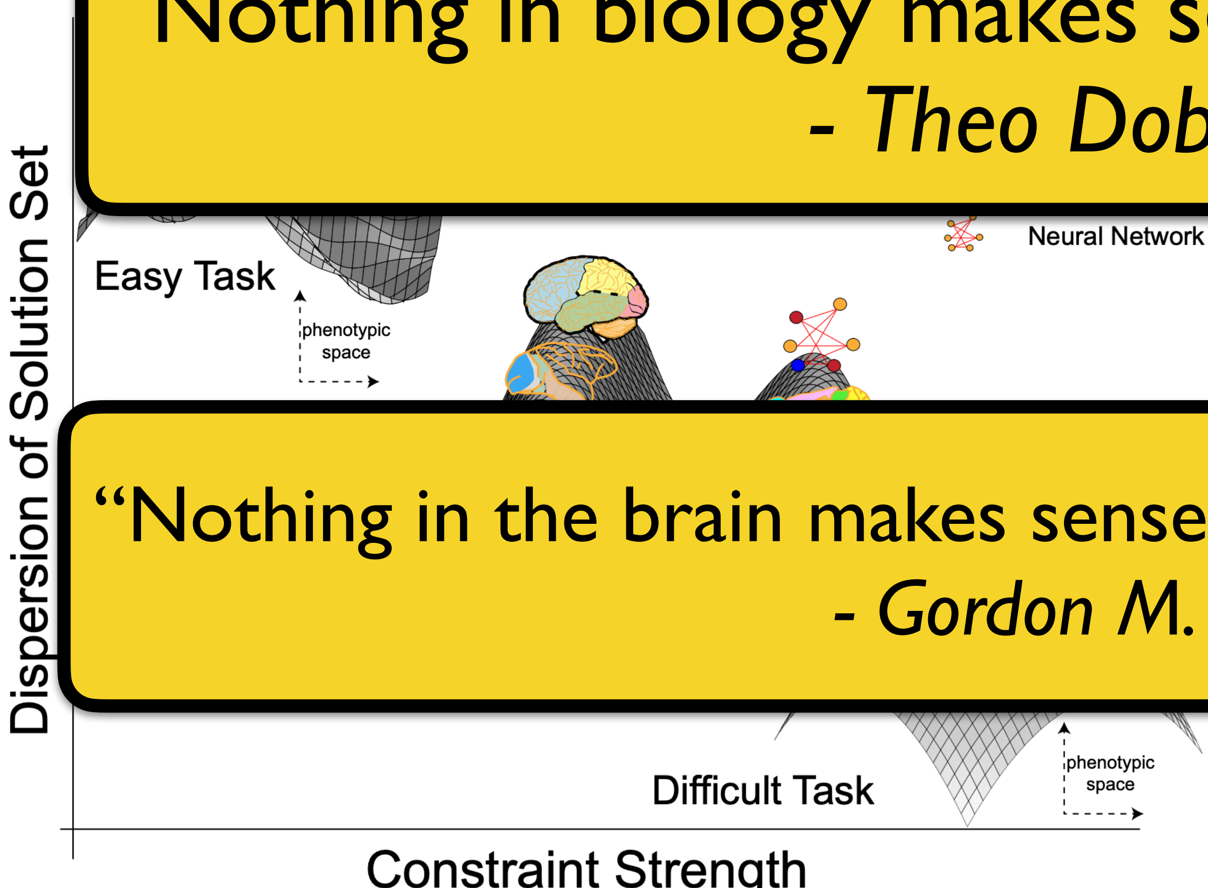
Minyoung Huh [*1] **Brian Cheung** [*1] **Tongzhou Wang** [*1] **Phillip Isola** [*1]

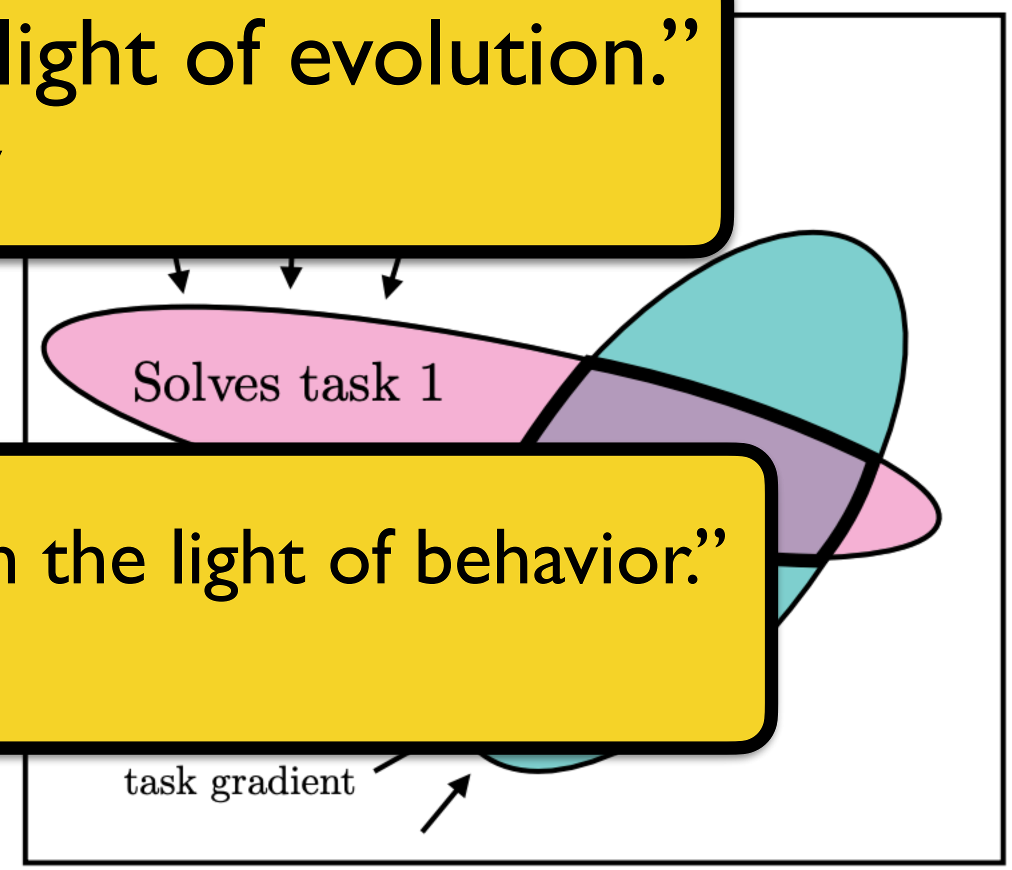
The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

Platonic Representation Hypothesis is the AI version of Contravariance

“Nothing in biology makes sense in light of evolution.”
- *Theo Dobzhansky*





“Nothing in the brain makes sense except in the light of behavior.”
- *Gordon M. Shepherd*

Figure 6. **The Multitask Scaling Hypothesis:** Models trained ...cted to pressure to ...s.

<u>Our (slightly) modified credo:</u>
“Nothing in (computational) neuroscience makes sense except in light of task-optimization.”

The Platonic Representation Hypothesis

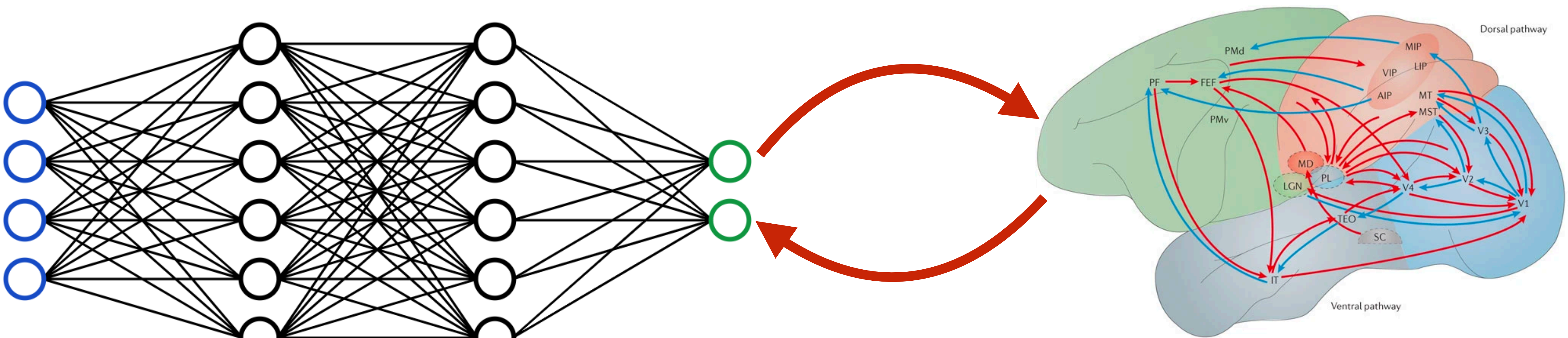
Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

Task-Optimized Modeling Approach

Design ML Algorithms Optimized to Perform Organism's Behavior under Organism's Constraints



Artificial Neural Network **Brain**

Yields:

Quantitatively Accurate & Practically Useful Brain Models

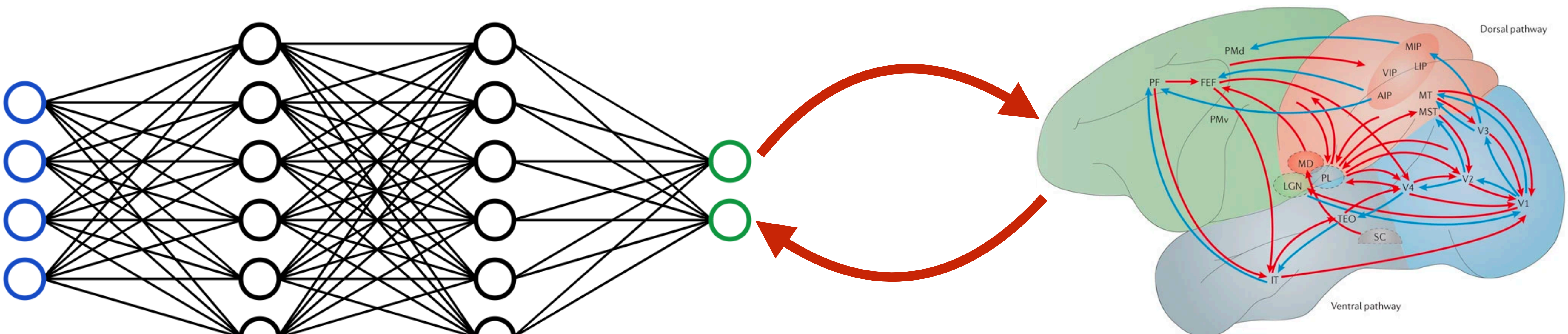
AND

Principles of *Why* Neural Responses Are As They Are

Task-Optimized Modeling Approach

Design ML Algorithms Optimized to Perform Organism's Behavior under Organism's Constraints

But what even counts as *good* here?



Artificial Neural Network | **Brain**

Yields:

Quantitatively Accurate & Practically Useful Brain Models

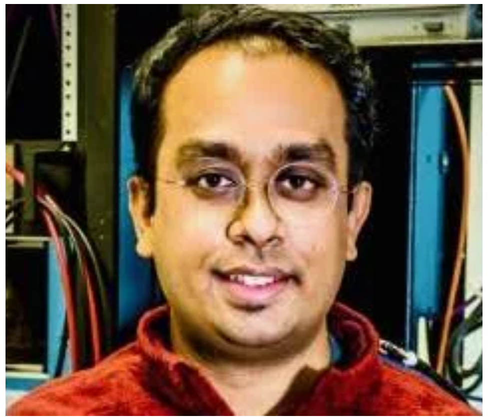
AND

Principles of *Why* Neural Responses Are As They Are

NeuroAI Turing Test

Jenelle Feather

Meenakshi Khosla



Ratan Murty

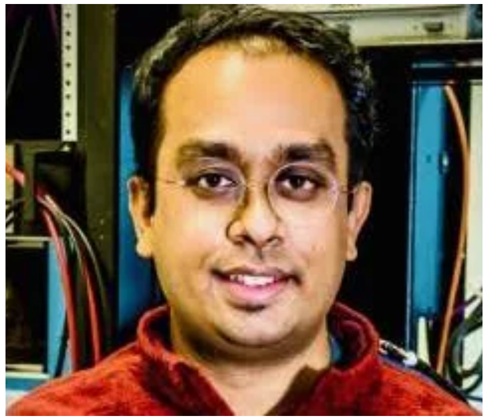
Brain-Model Evaluations Need the NeuroAI Turing Test

Jenelle Feather * 1 Meenakshi Khosla * 2 N. Apurva Ratan Murty * 3 Aran Nayebi * 4

NeuroAI Turing Test

Jenelle Feather

Meenakshi Khosla

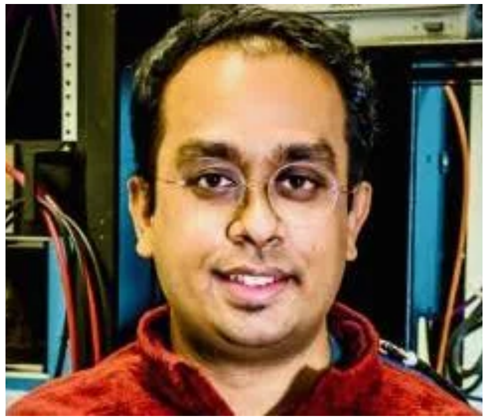


Ratan Murty

NeuroAI Turing Test

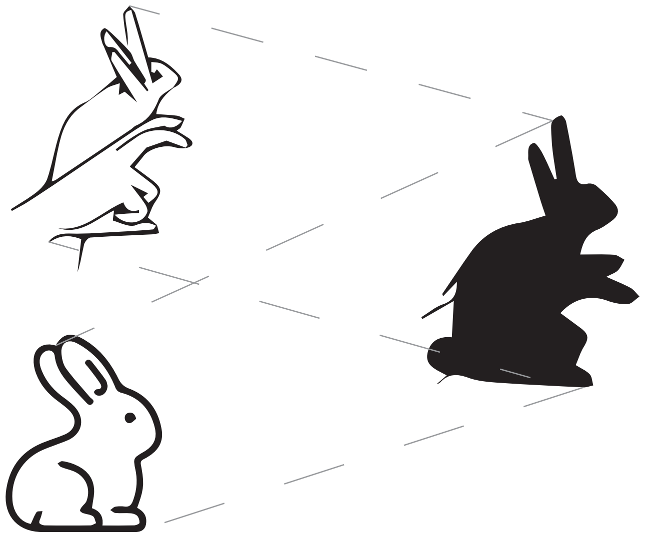
Jenelle Feather

Meenakshi Khosla

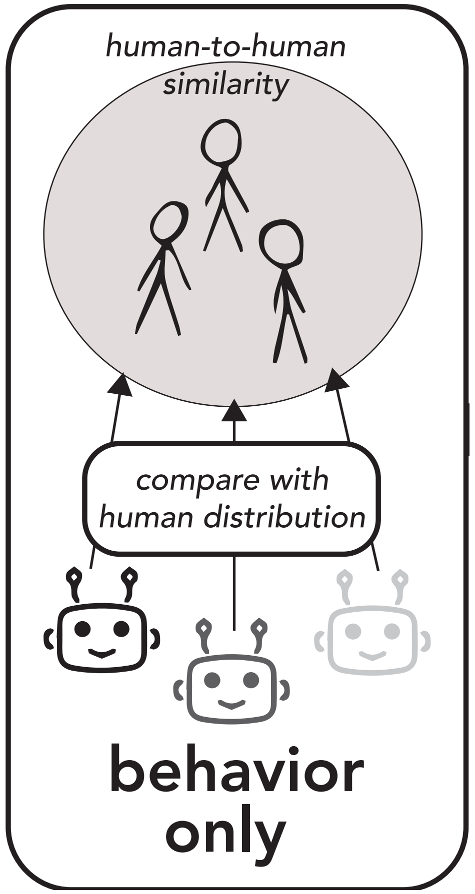


Ratan Murty

Just as distinct objects can cast the same shadow...



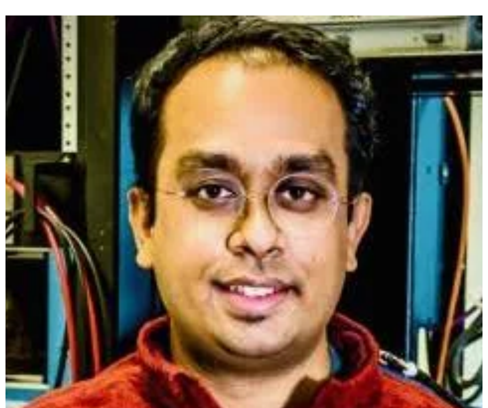
Turing Test



NeuroAI Turing Test

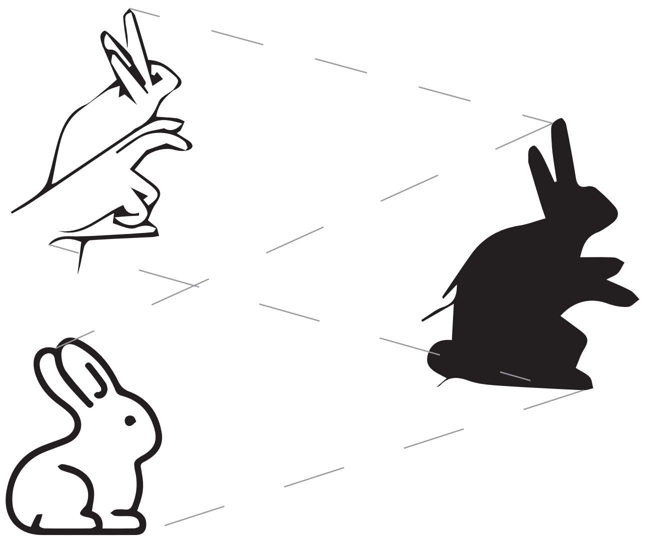
Jenelle Feather

Meenakshi Khosla

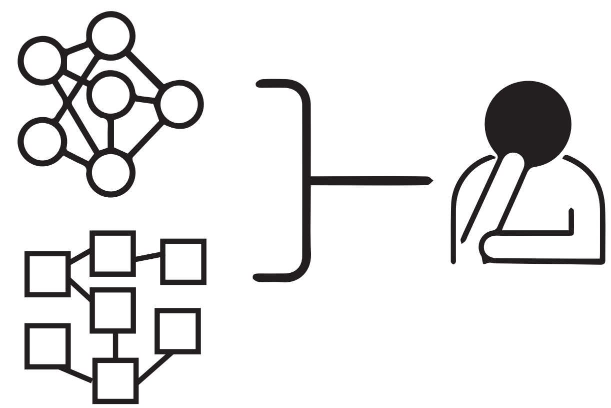


Ratan Murty

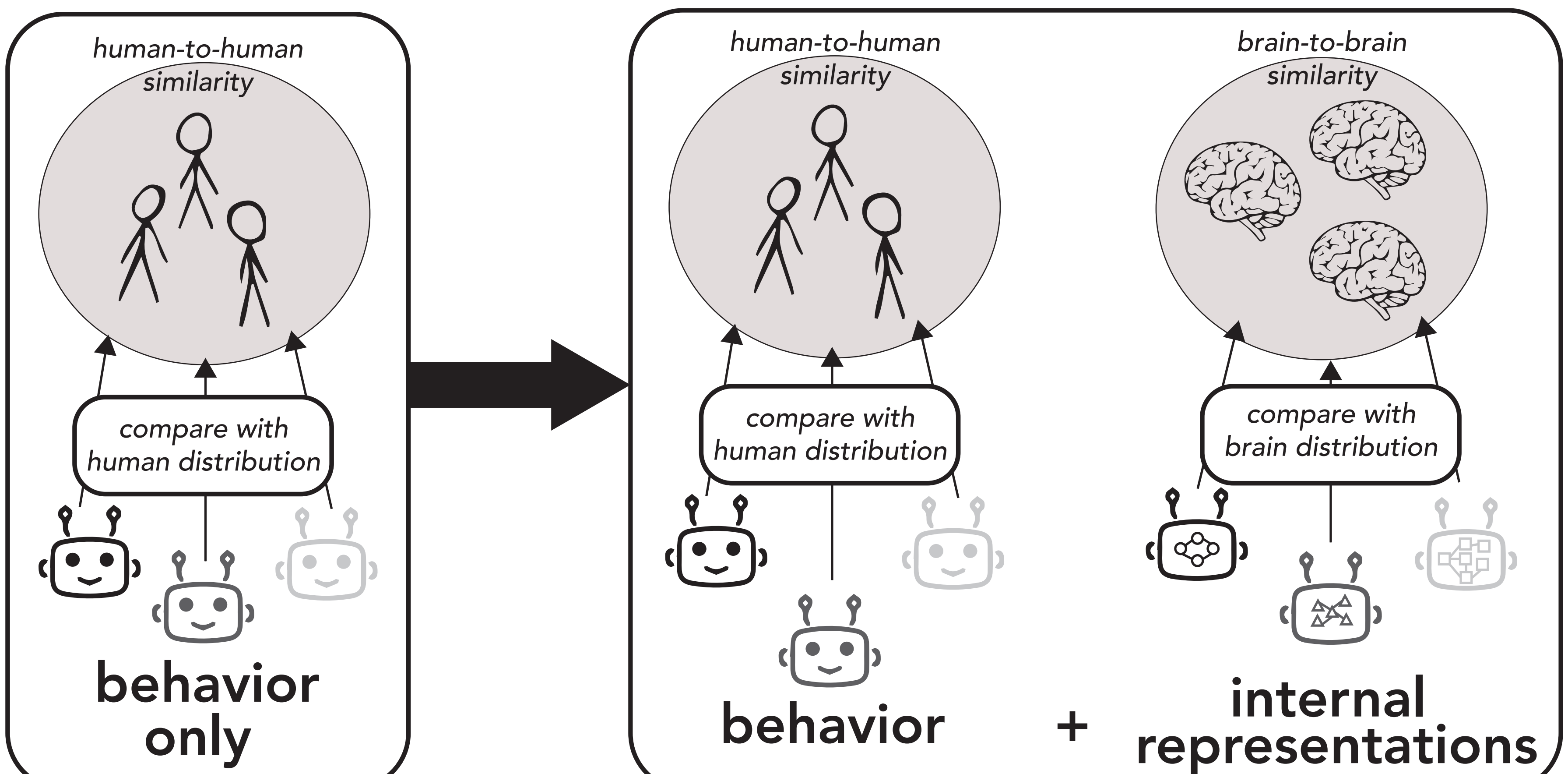
Just as distinct objects can cast the same shadow...



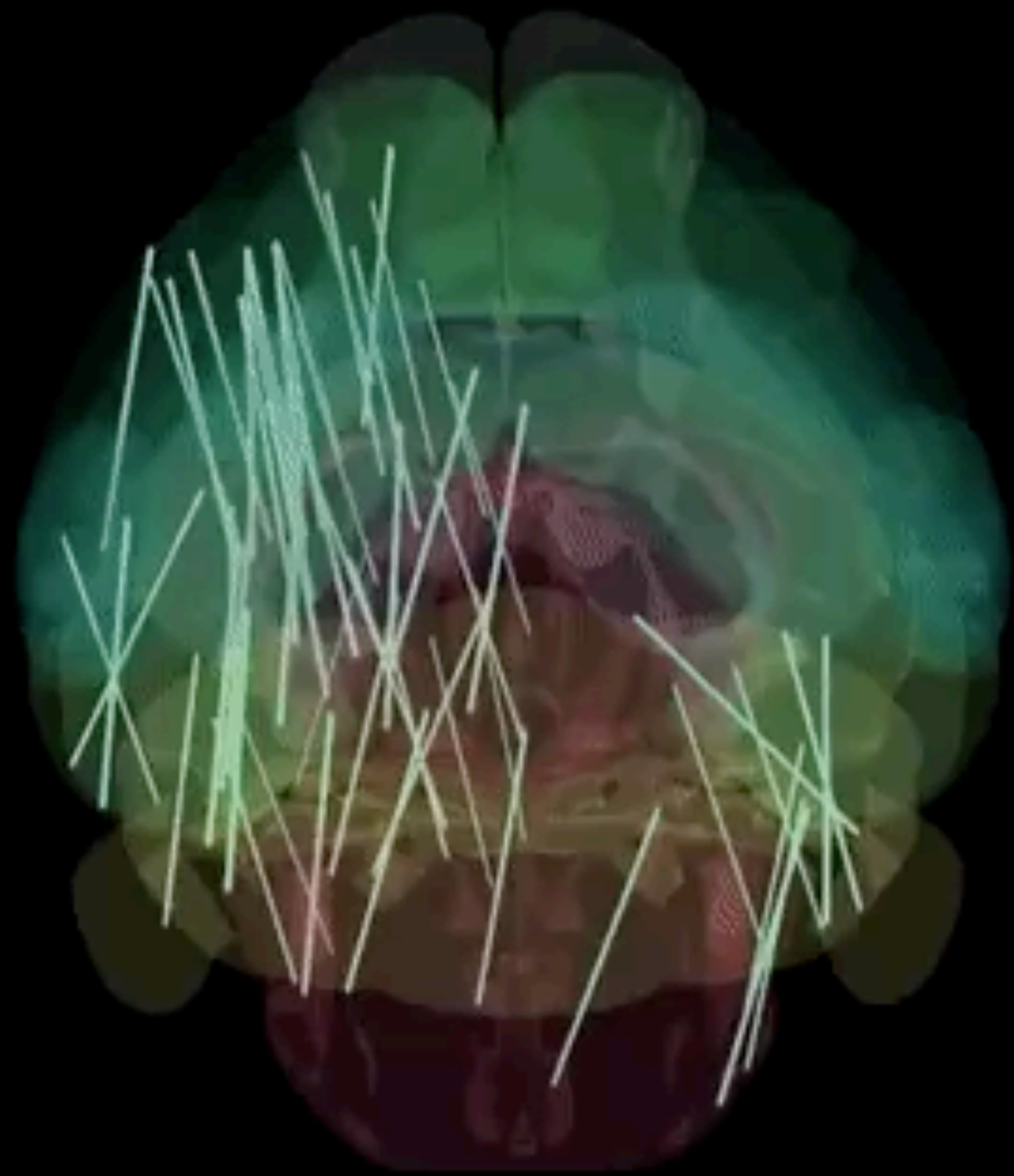
...distinct internal processes (representations) can produce similar outputs (behavior)



Turing Test — **The NeuroAI Turing Test**

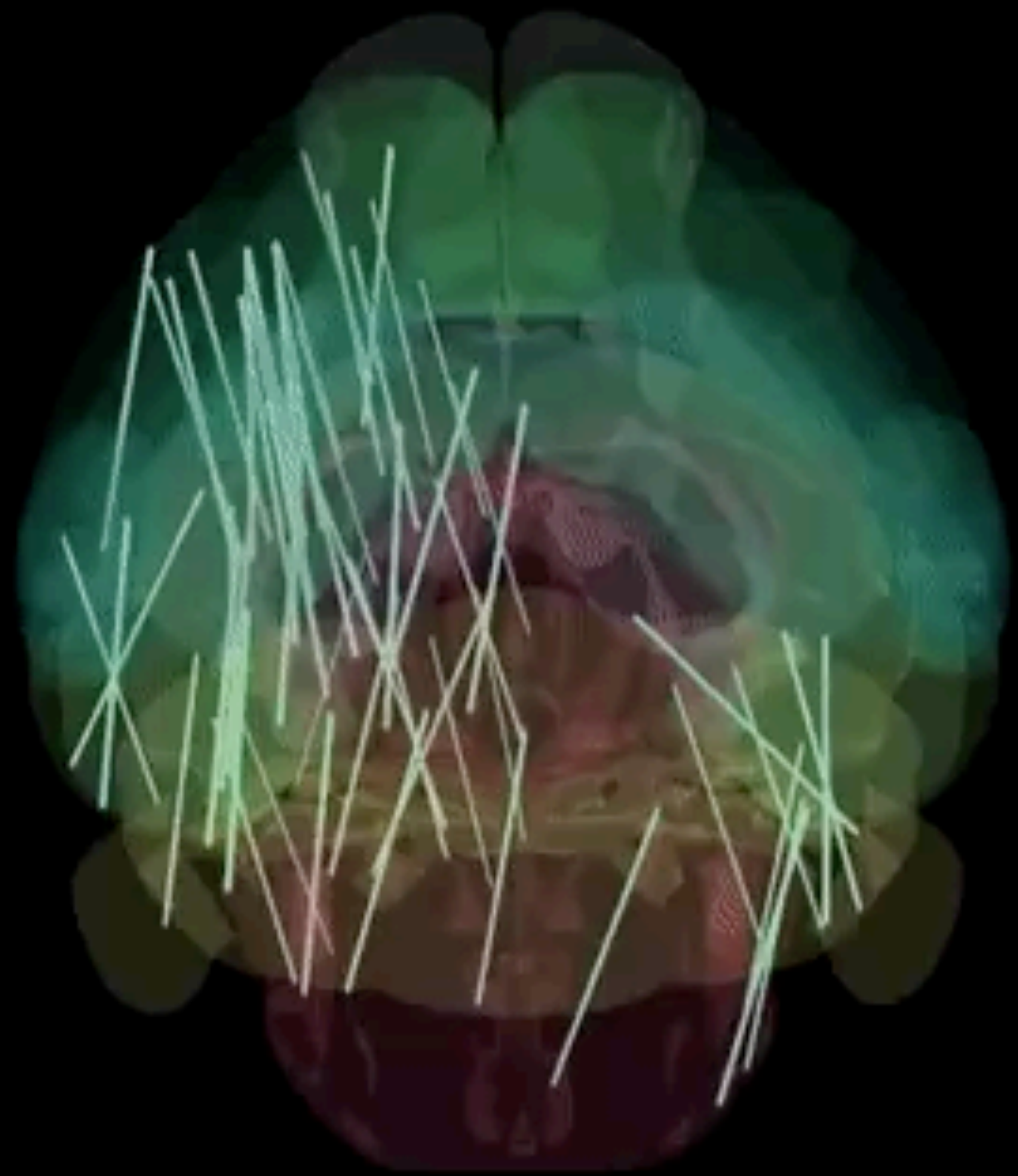


How to Reverse-Engineer Natural Intelligence?



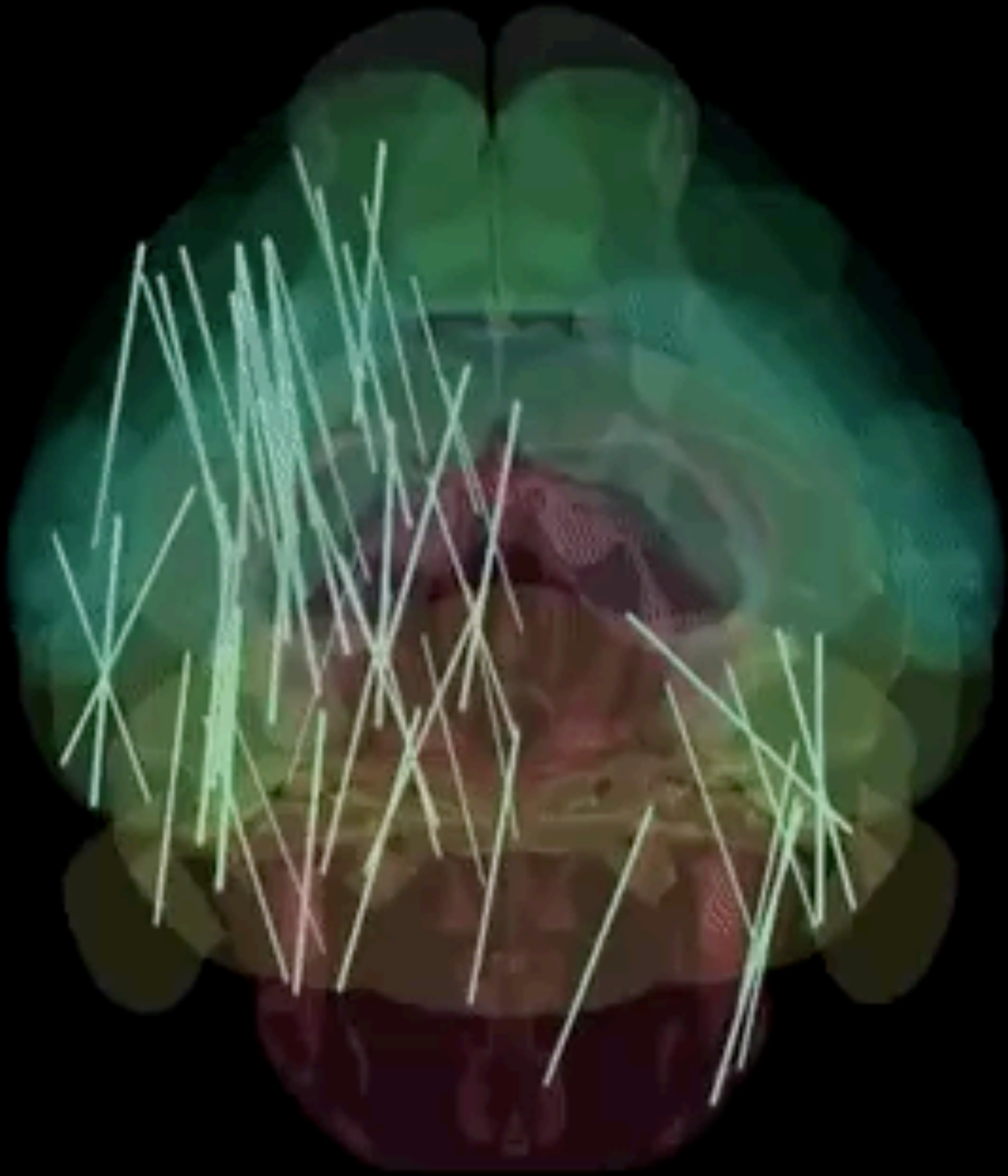
International Brain Laboratory 2022

How to Reverse-Engineer Natural Intelligence?



International Brain Laboratory 2022

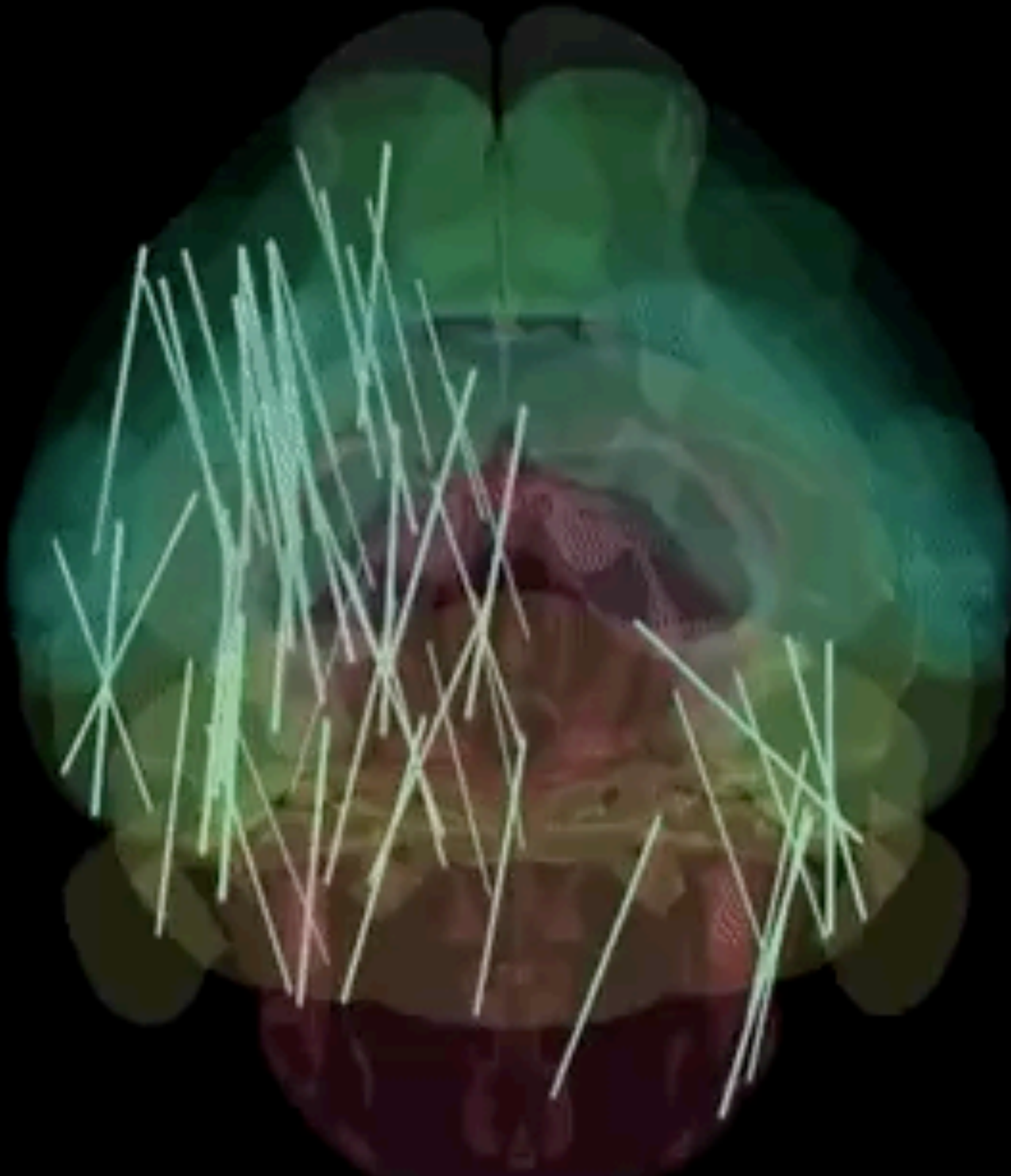
Whole brain. . .



International Brain Laboratory 2022

How to Reverse-Engineer Natural Intelligence?

Whole brain. . .



International Brain Laboratory 2022

. . . awake, behaving animals

How to Reverse-Engineer Natural Intelligence?

Whole brain. . .

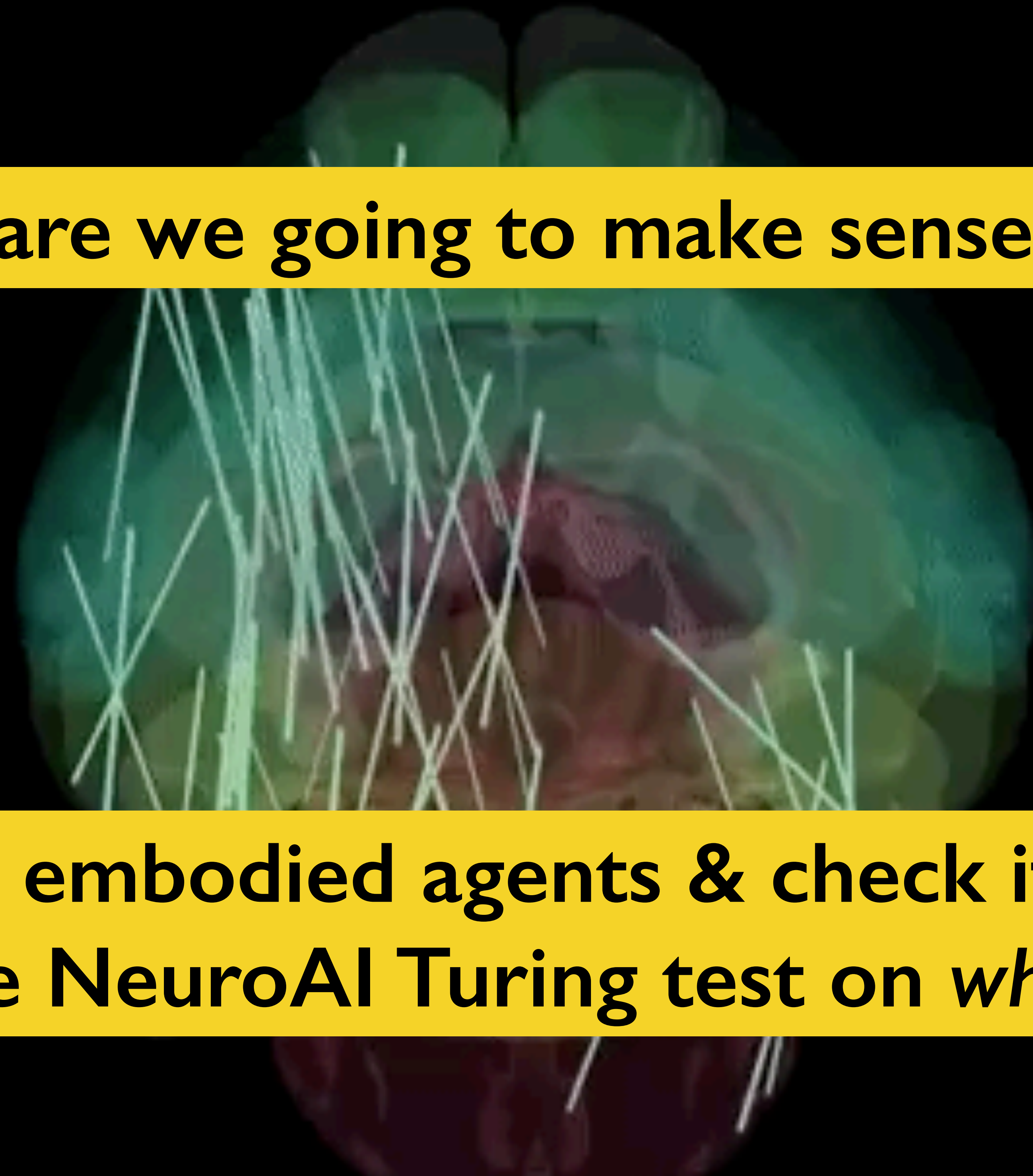
Q: How are we going to make sense of all this data?

International Brain Laboratory 2022

. . . awake, behaving animals

How to Reverse-Engineer Natural Intelligence?

Whole brain. . .



Q: How are we going to make sense of all this data?

A: Build embodied agents & check if their internals pass the NeuroAI Turing test on *whole-brain* data.

International Brain Laboratory 2022

. . . awake, behaving animals

Using Agents to Reverse-Engineer *Whole-Brain* Data

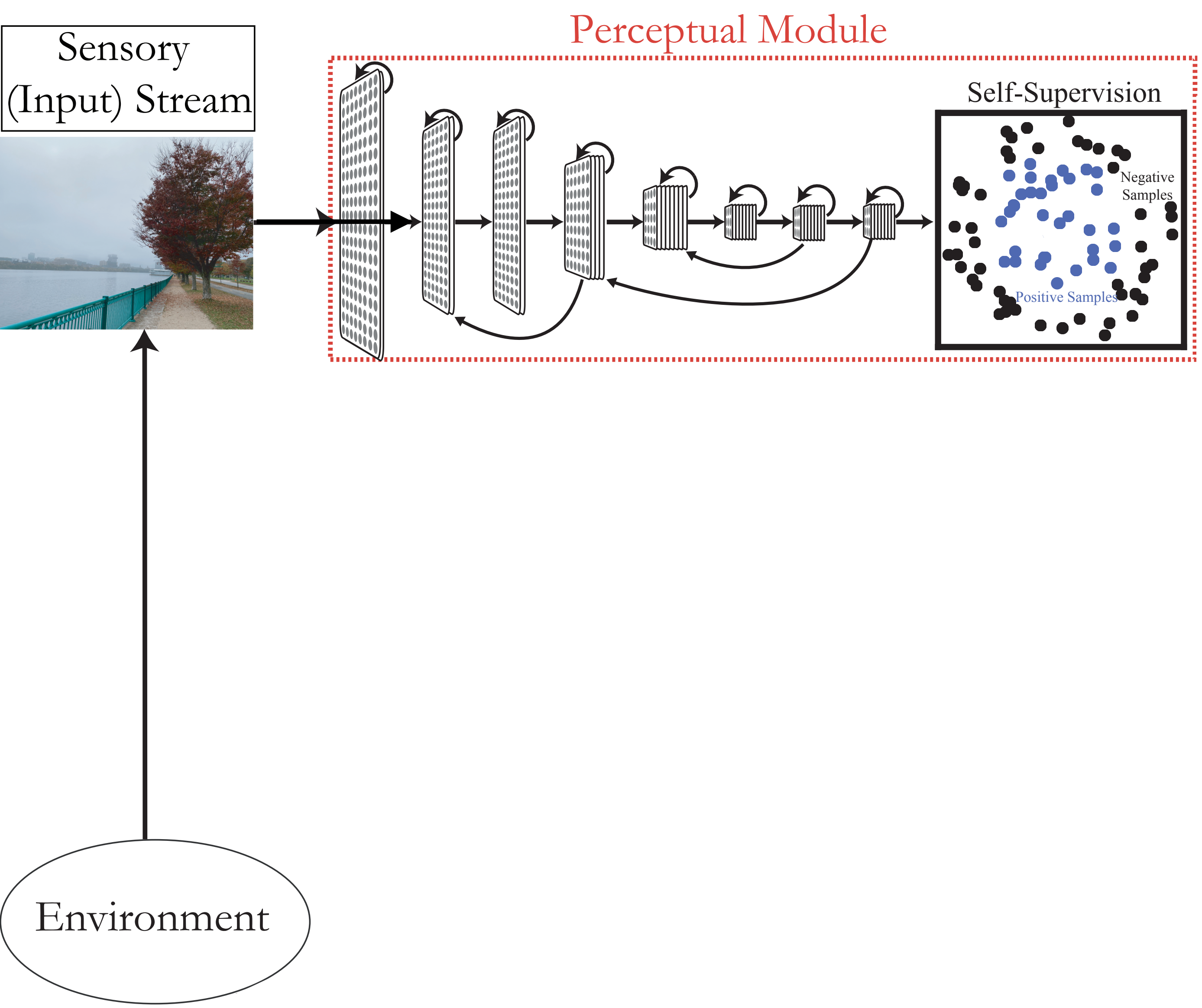
How does the brain build and use world models?

Using Agents to Reverse-Engineer *Whole-Brain* Data

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

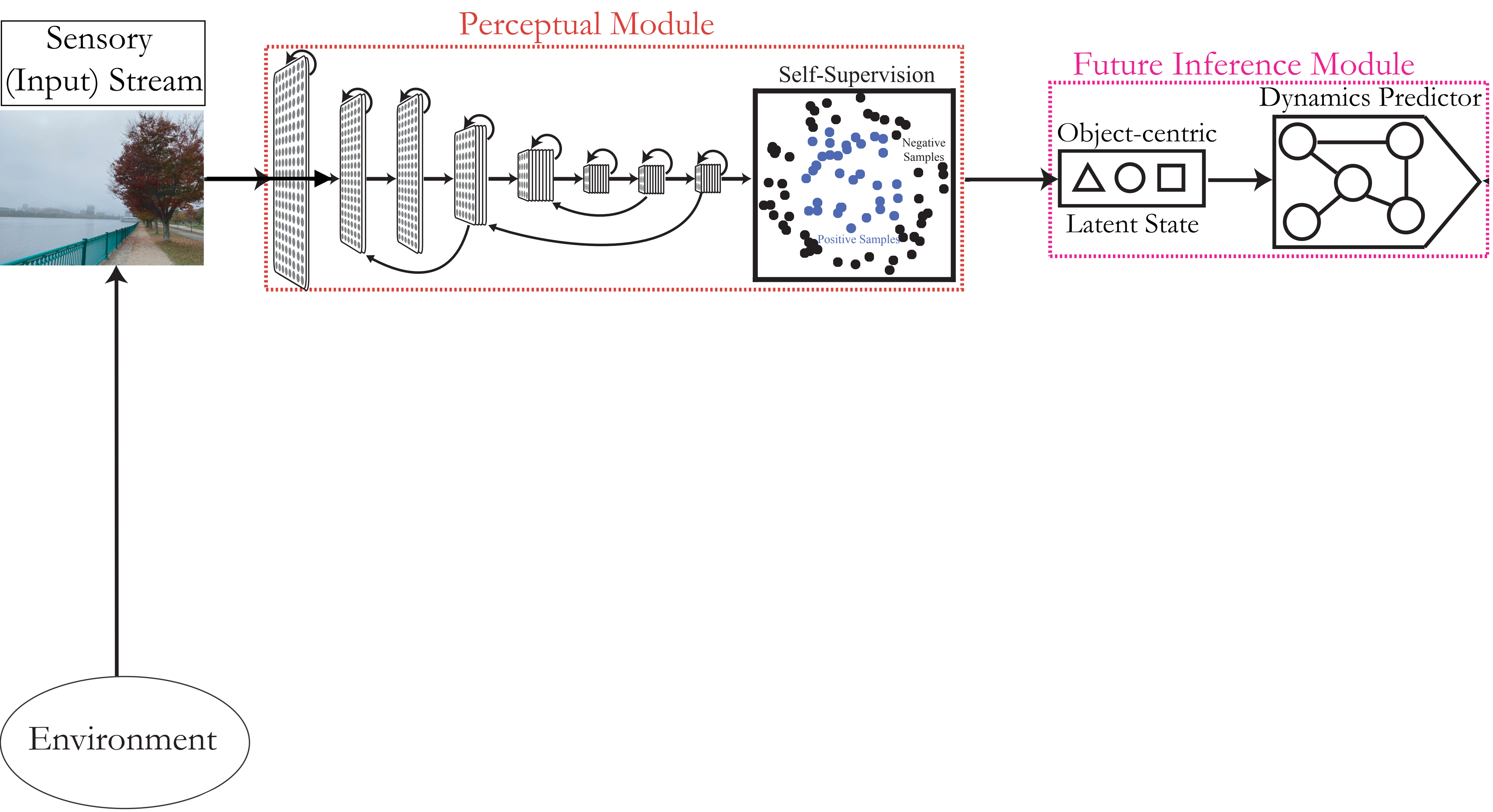
Using Agents to Reverse-Engineer *Whole-Brain* Data

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



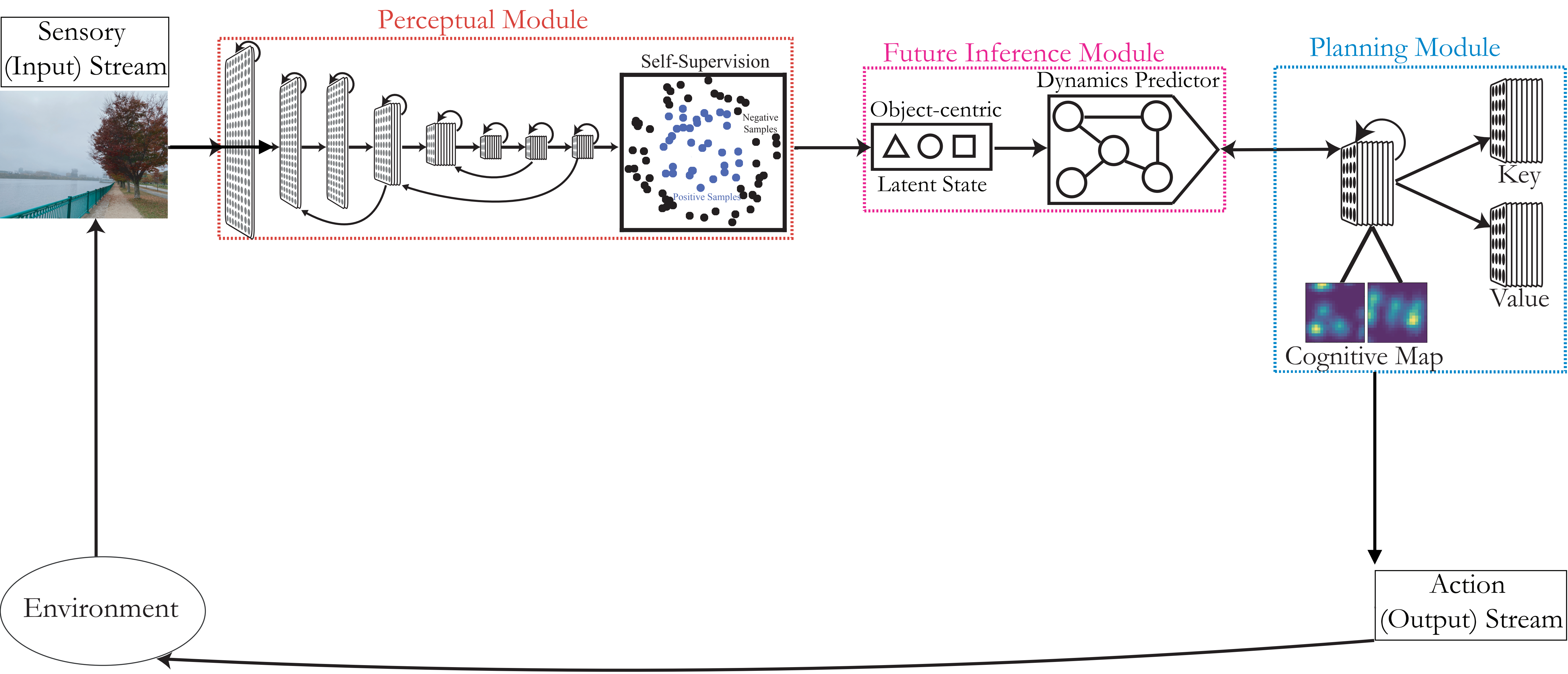
Using Agents to Reverse-Engineer *Whole-Brain* Data

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



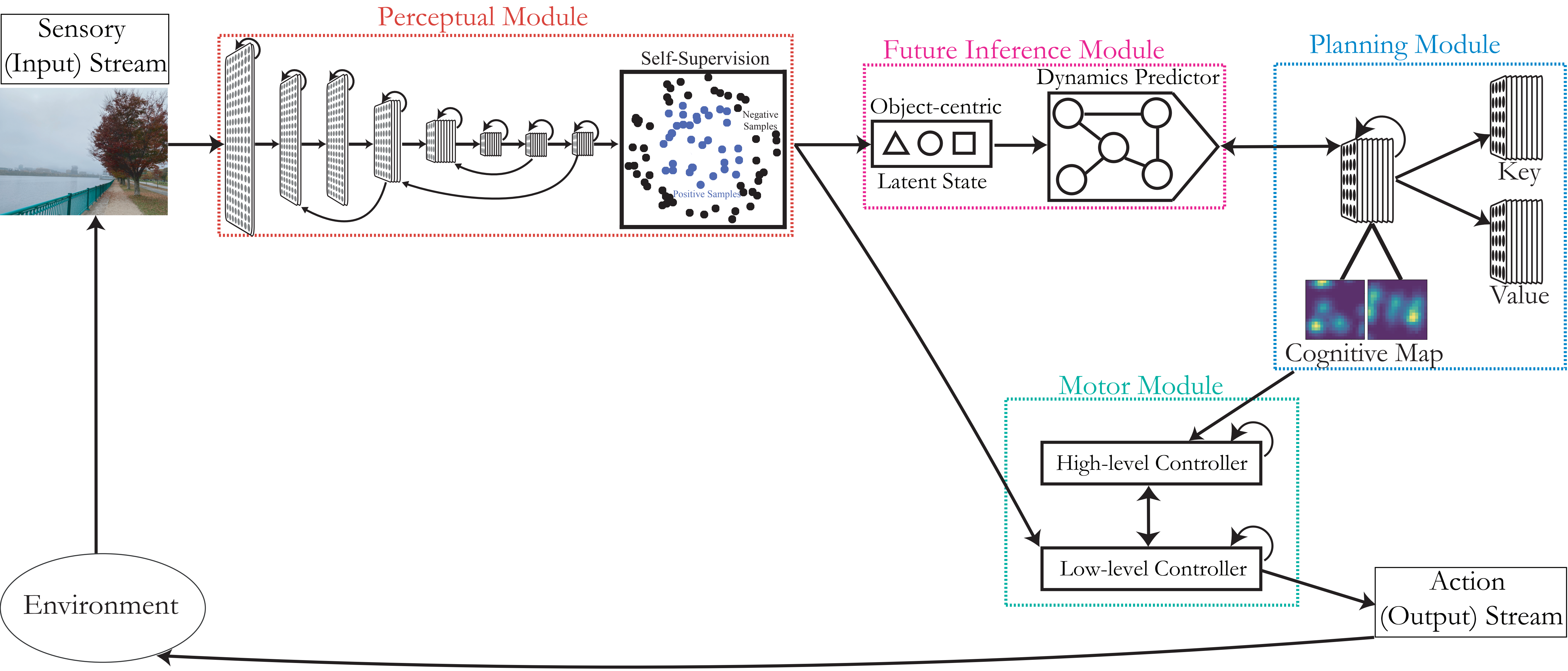
Using Agents to Reverse-Engineer *Whole-Brain* Data

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



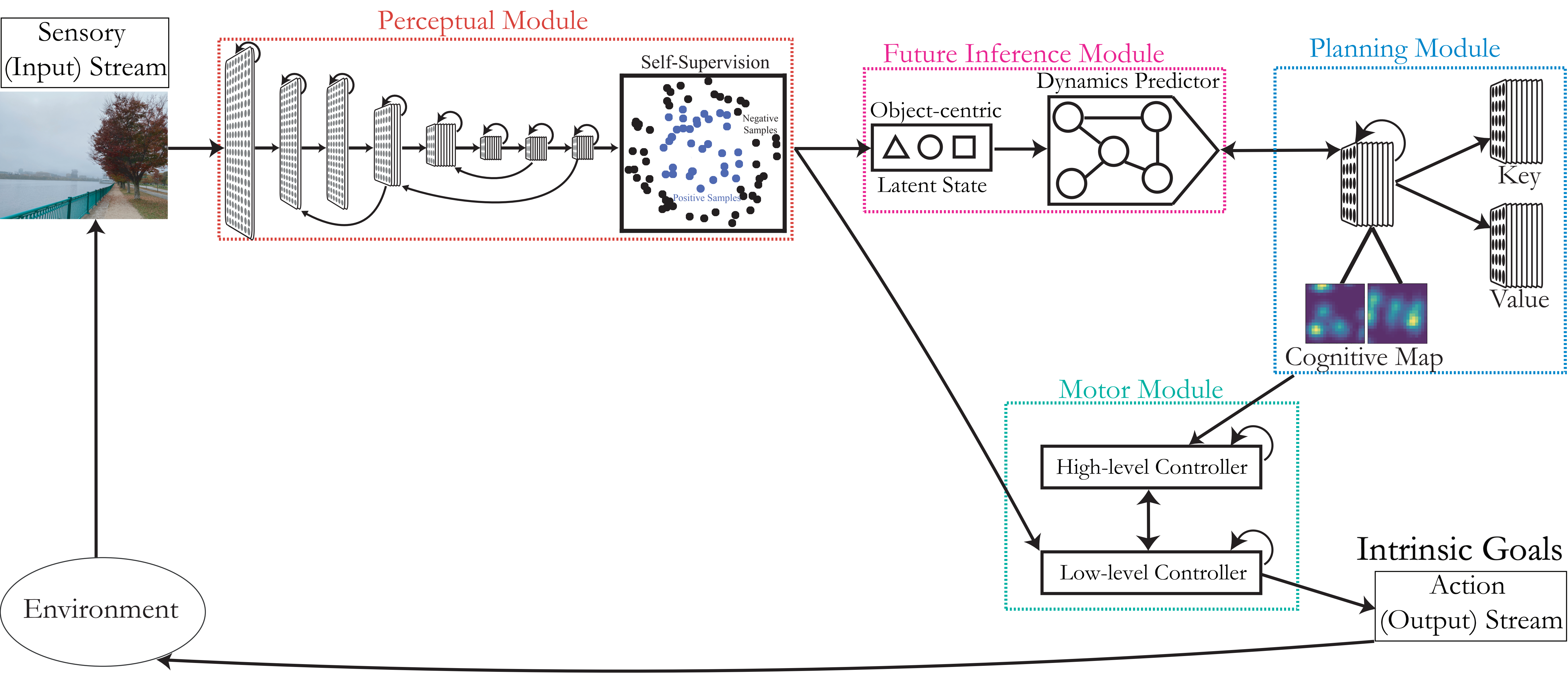
Using Agents to Reverse-Engineer *Whole-Brain* Data

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



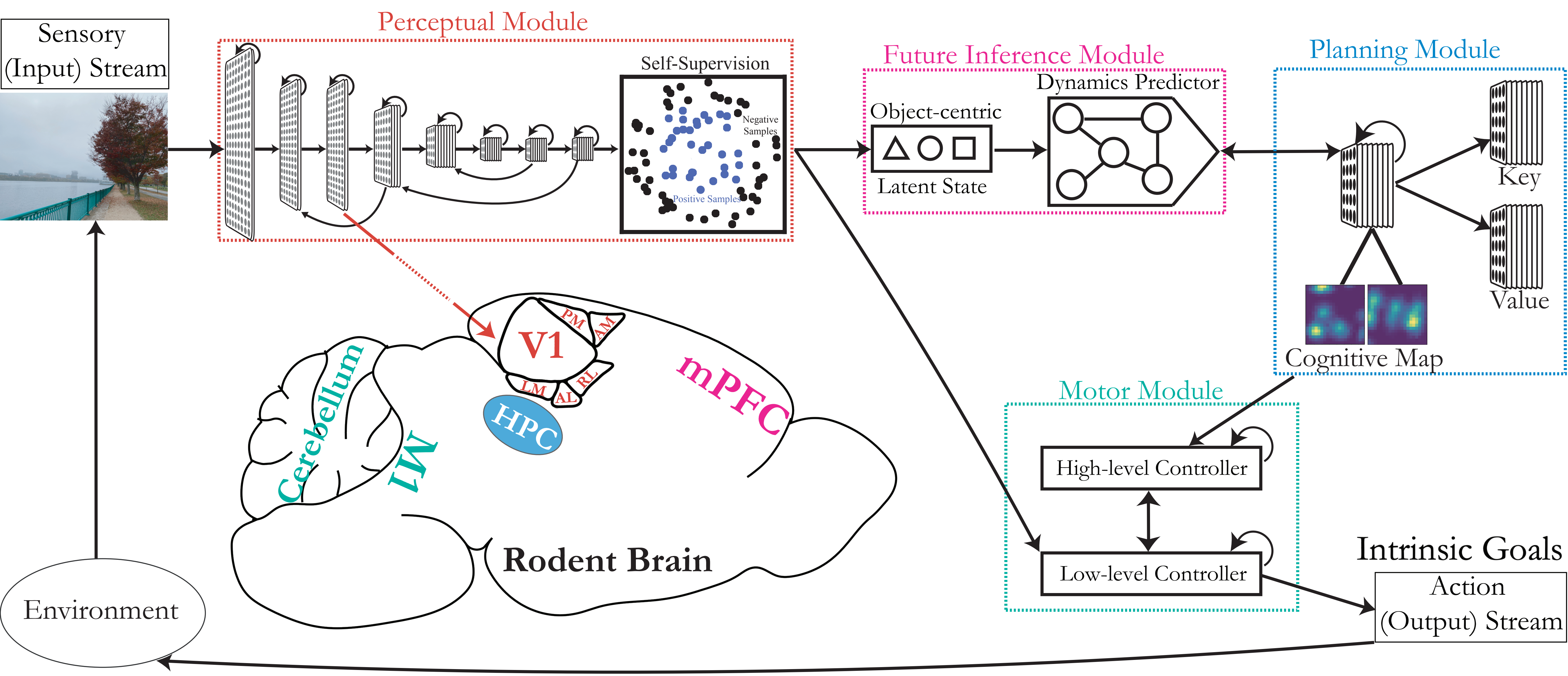
Using Agents to Reverse-Engineer *Whole-Brain* Data

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



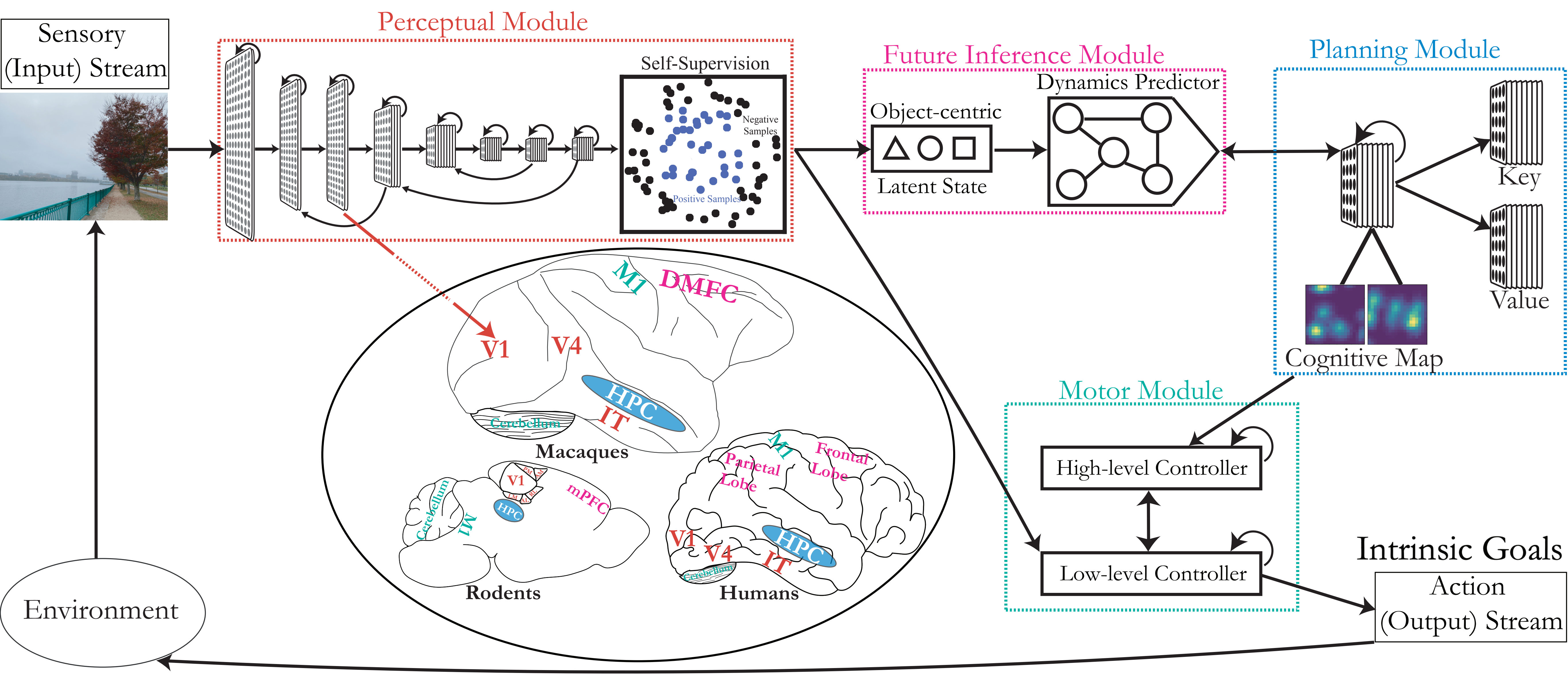
Using Agents to Reverse-Engineer *Whole-Brain* Data

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



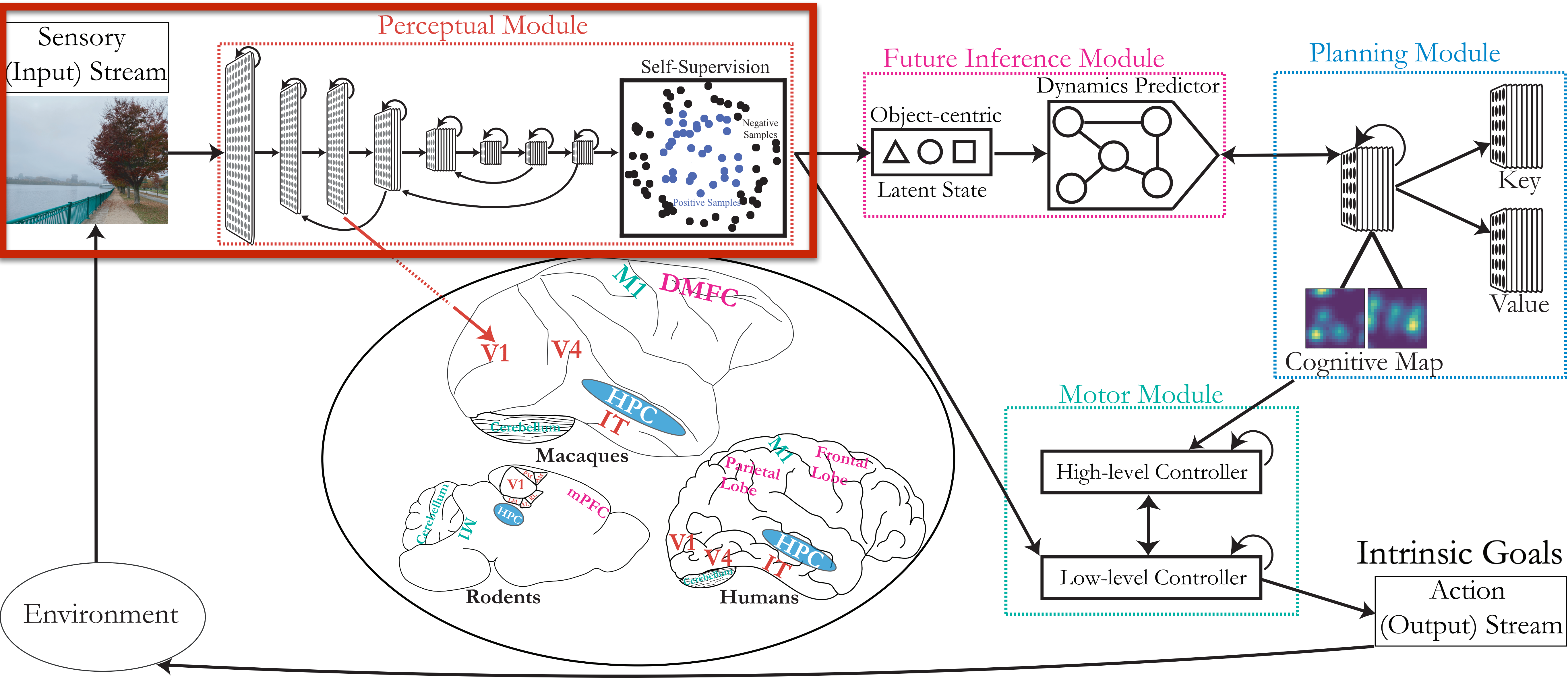
Long-Term Outcome: Artificial **Organisms**

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



Roadmap: Perception

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



The Supervision Problem

There's just no way that these creatures receive millions of high-level semantic labels during learning.

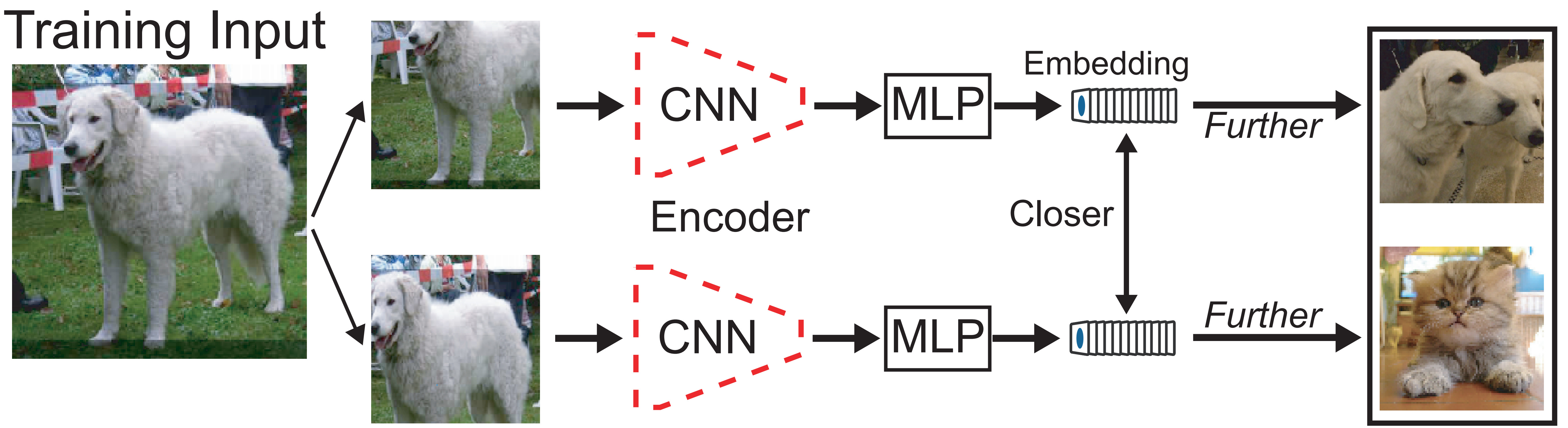
Effective proxy, but just obviously deeply wrong.

The Supervision Problem

There's just no way that these creatures receive millions of high-level semantic labels during learning.

Effective proxy, but just obviously deeply wrong.

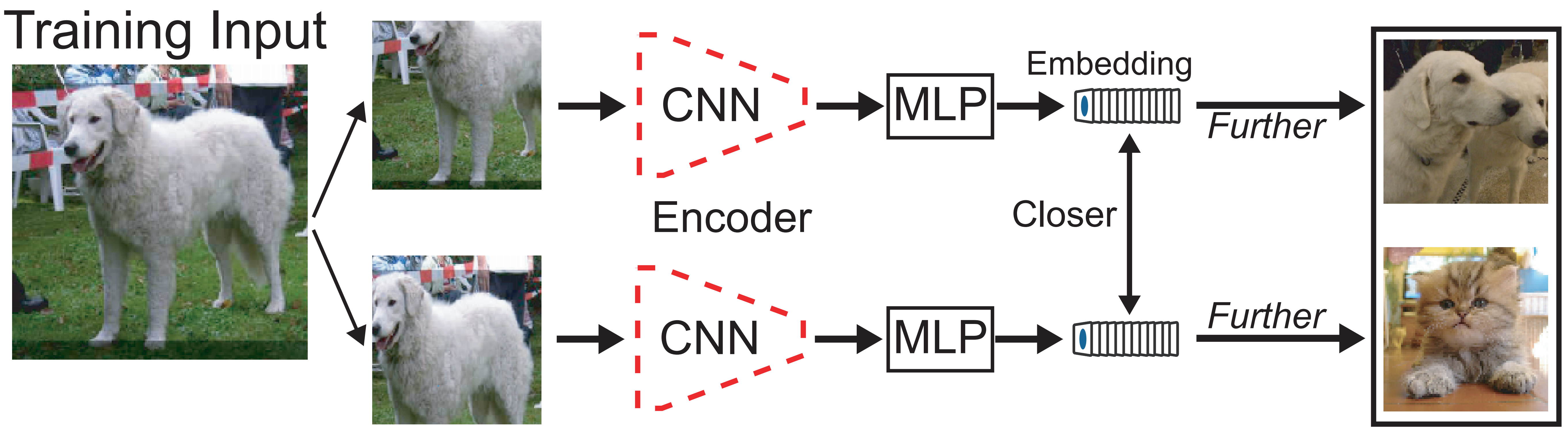
Contrastive learning tasks



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

High-level idea of these methods: make the representations non-trivially robust to data augmentations

Contrastive learning tasks



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

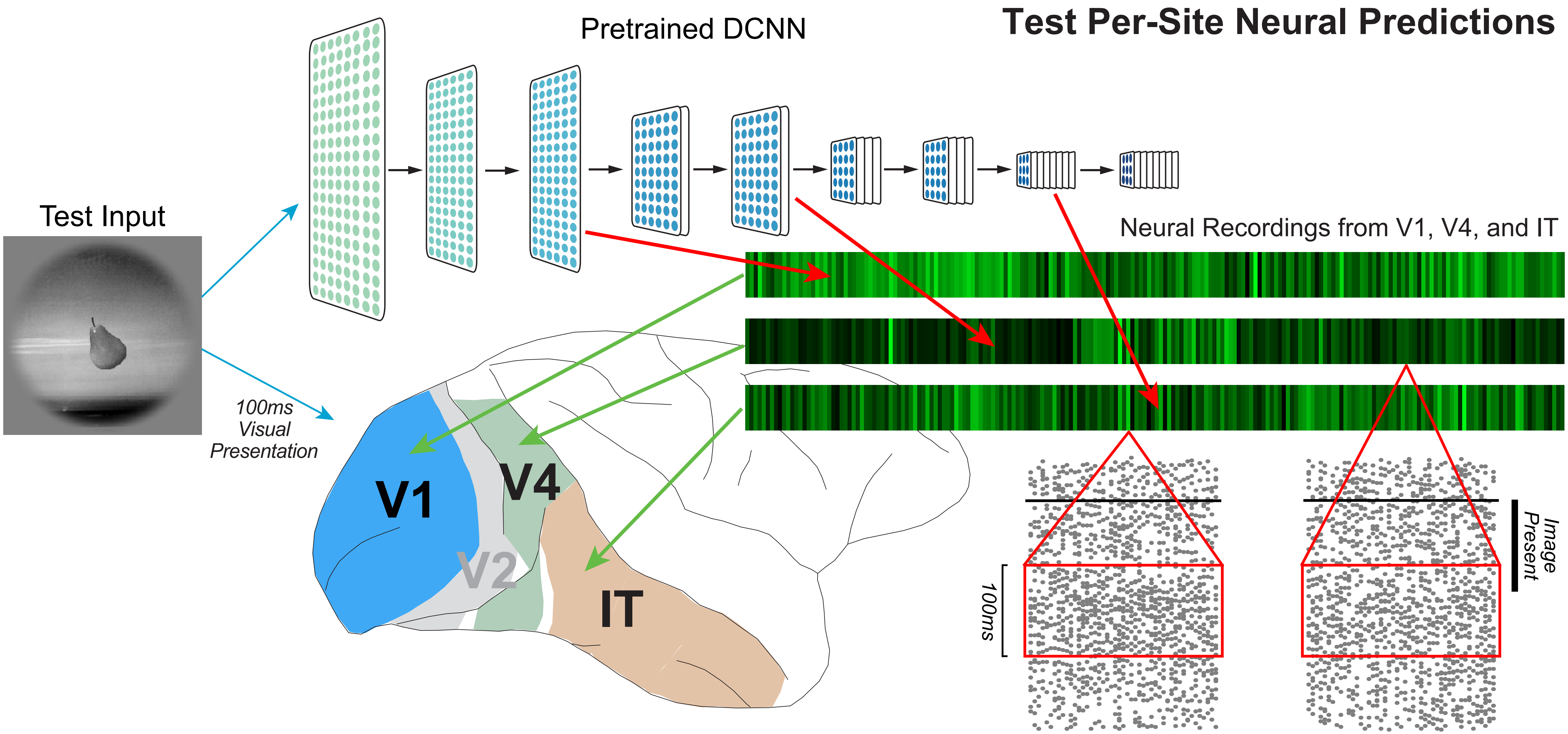
High-level idea of these methods: make the representations non-trivially robust to data augmentations

(somewhat inspired by how we "sample" the world via head motion)

Comparison to Neural Data

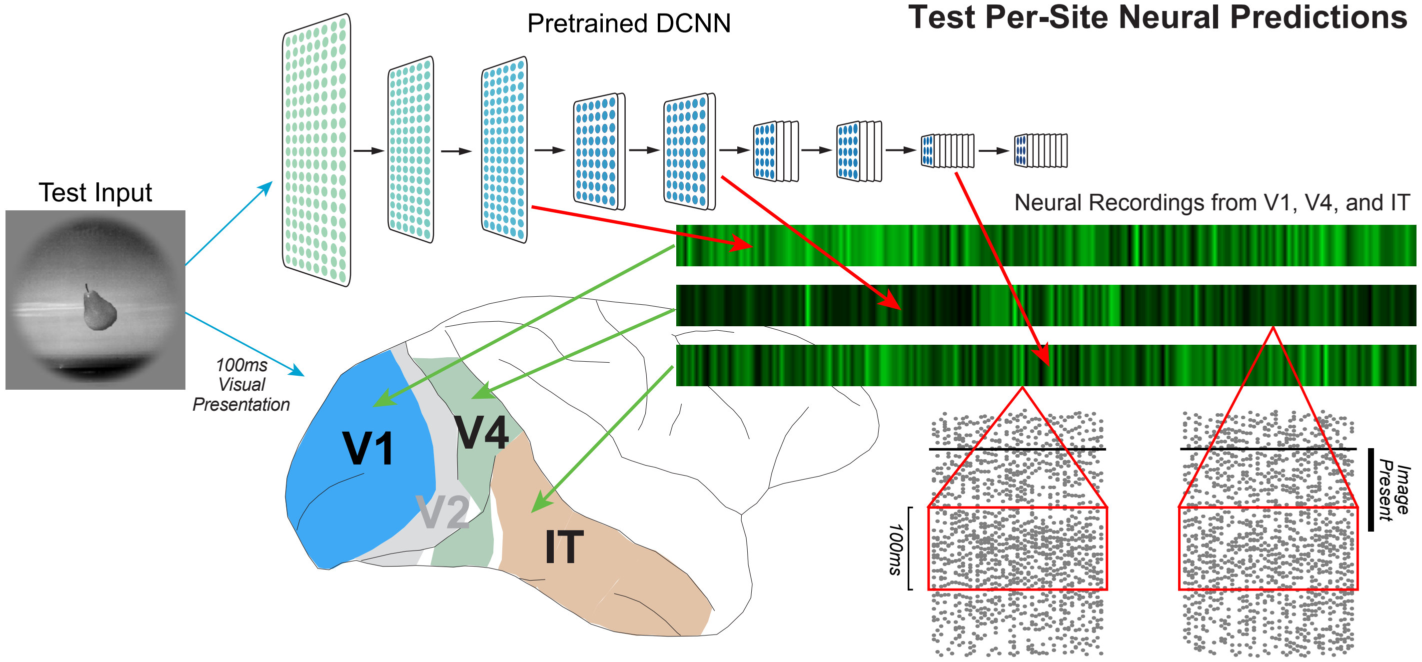
How well does it match neural data?

Chengxu Zhuang



V1 data from Cadena et al. Deep convolutional models improve predictions of macaque **V1** responses to natural images *PLoS Comp. Bio.,* (2019)

V4 & IT data from Majaj et al. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance *J. Neurosci.* (2015)

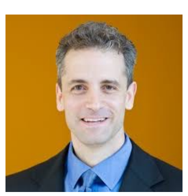


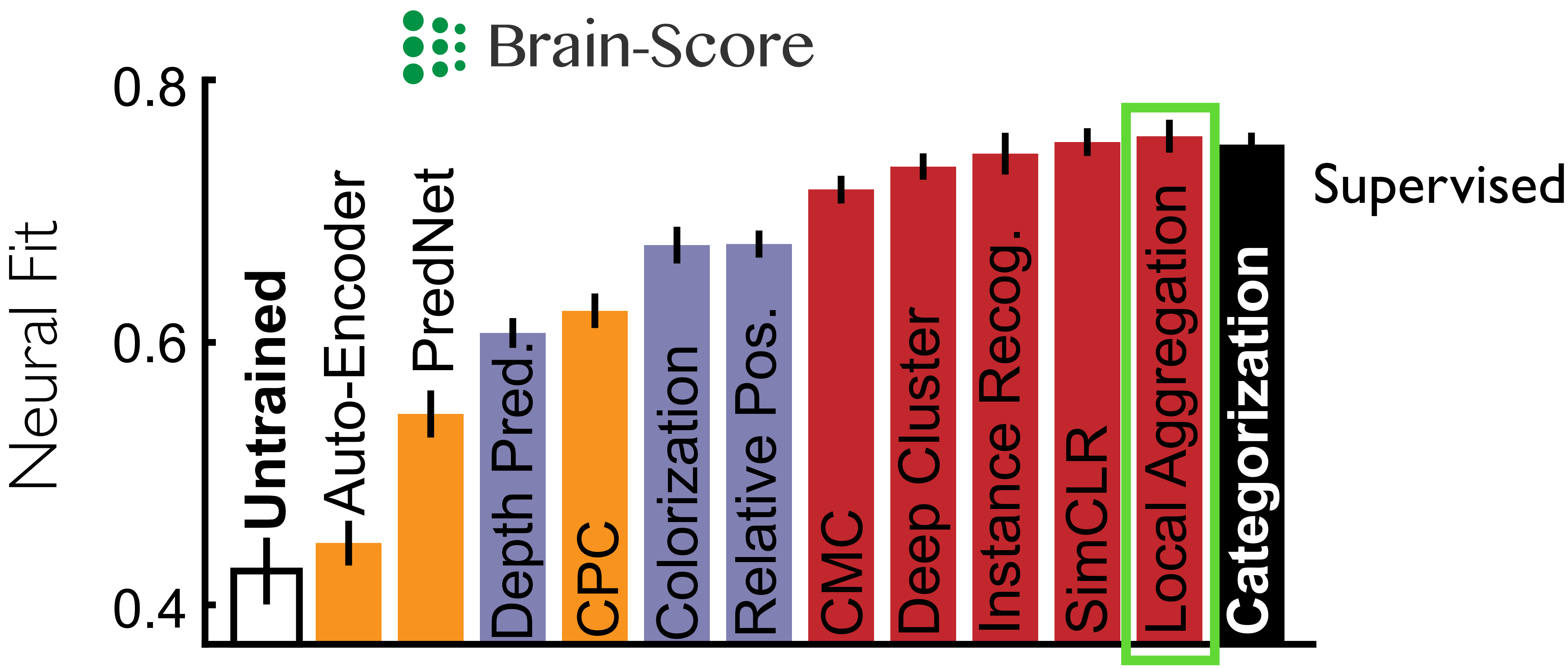
Autoencoders

Missing-Data Tasks

Deep Contrastive Embeddings

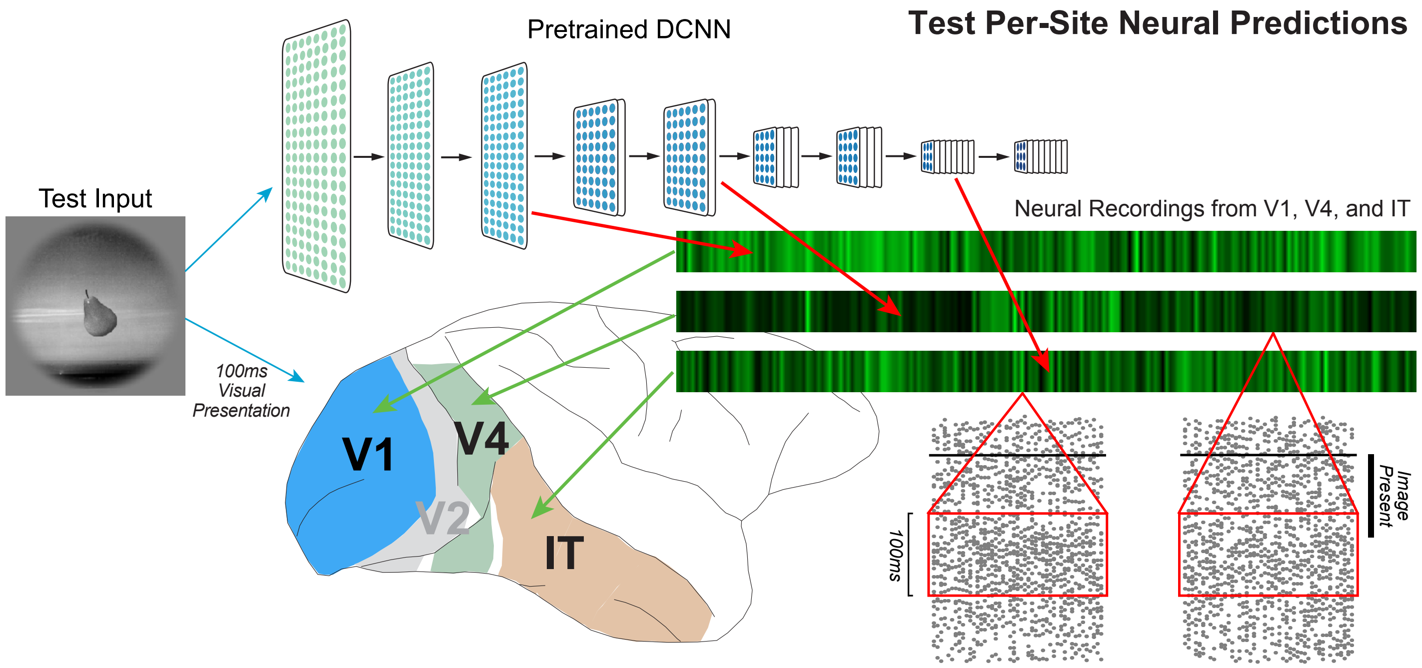
Chengxu Zhuang





Quantitatively accurate self-supervised model of a higher brain area.

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*



Autoencoders

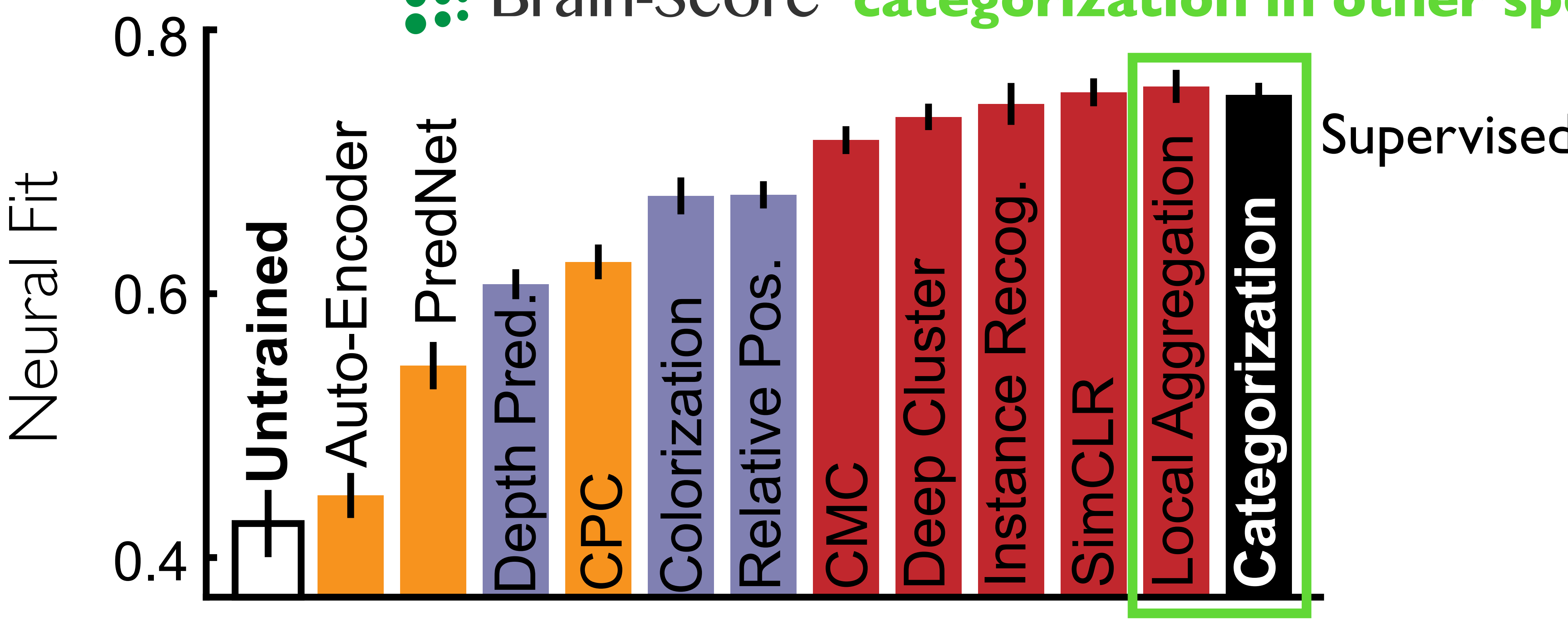
Missing-Data Tasks

Deep Contrastive Embeddings

Chengxu Zhuang

Can we do even better than categorization in other species?

Brain-Score



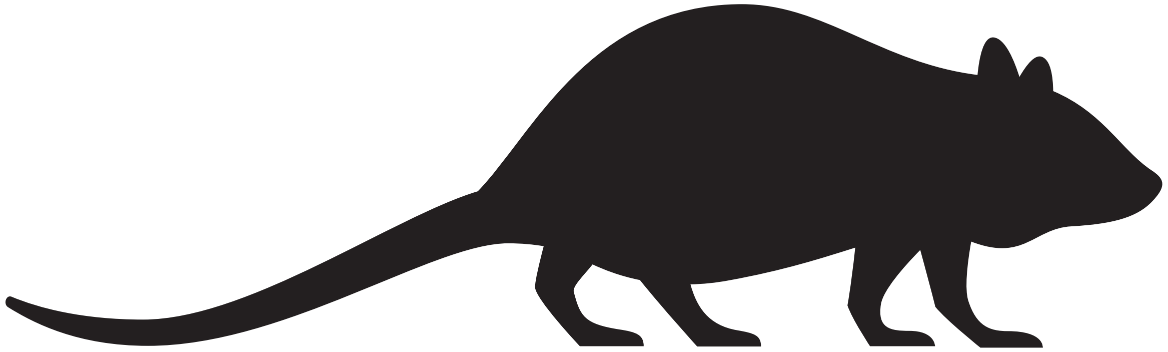
<u>Quantitatively accurate self-supervised model of a higher brain area.</u>

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

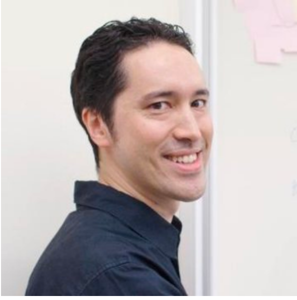
Mouse Visual Cortex as a Task-General, Limited Resource System

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


Nathan C.L. Kong*



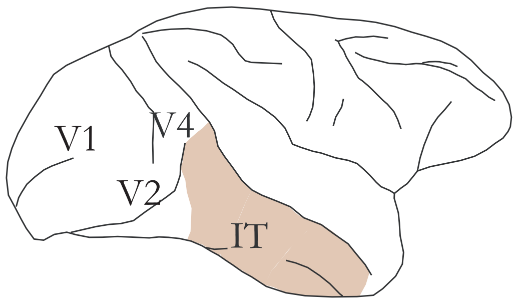

Chengxu Zhuang


Justin L. Gardner

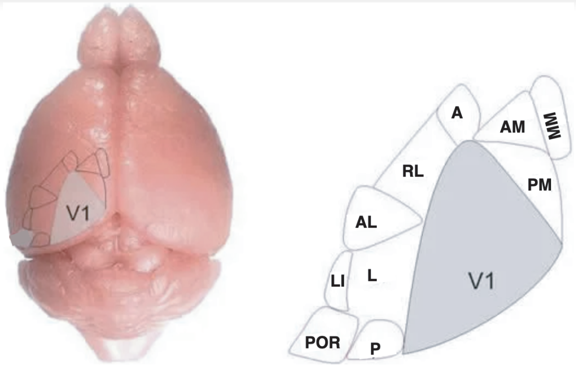

Anthony M. Norcia


Daniel Yamins

Contrastive Models Better Match Mouse Visual Cortex



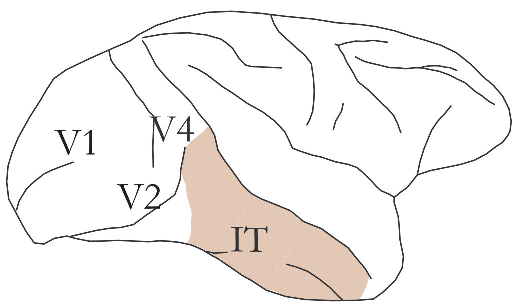
Primates



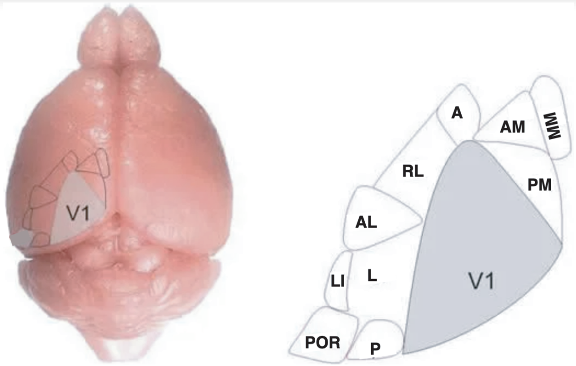
Mouse

Mouse vision is less hierarchical!

Contrastive Models Better Match Mouse Visual Cortex

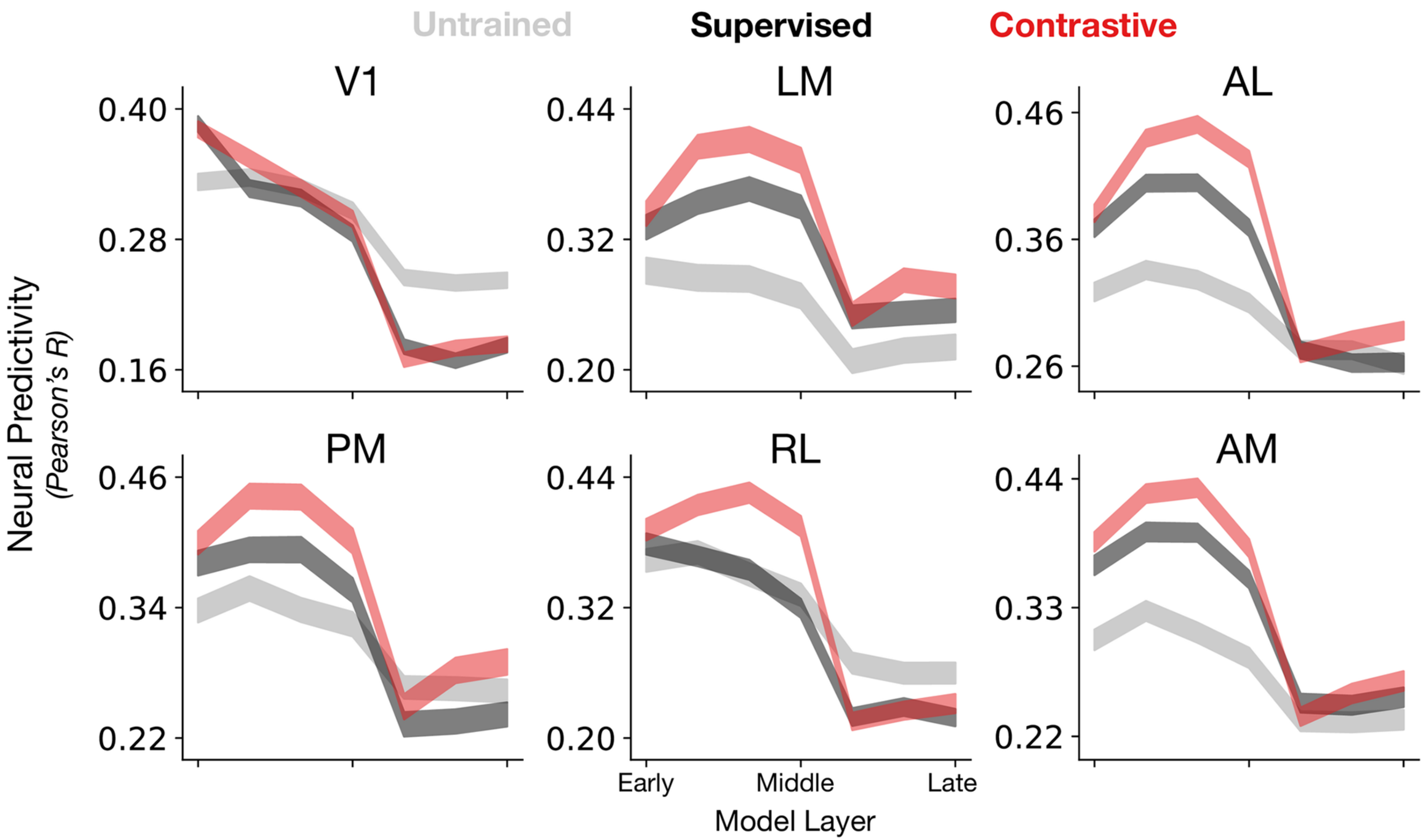


Primates

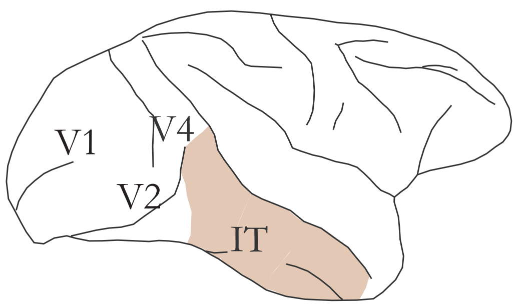


Mouse

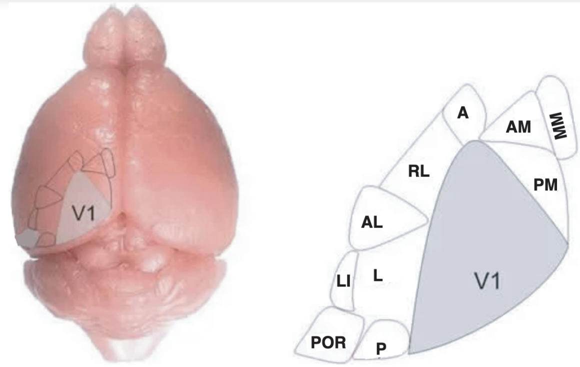
Mouse vision is less hierarchical!



Contrastive Models Better Match Mouse Visual Cortex

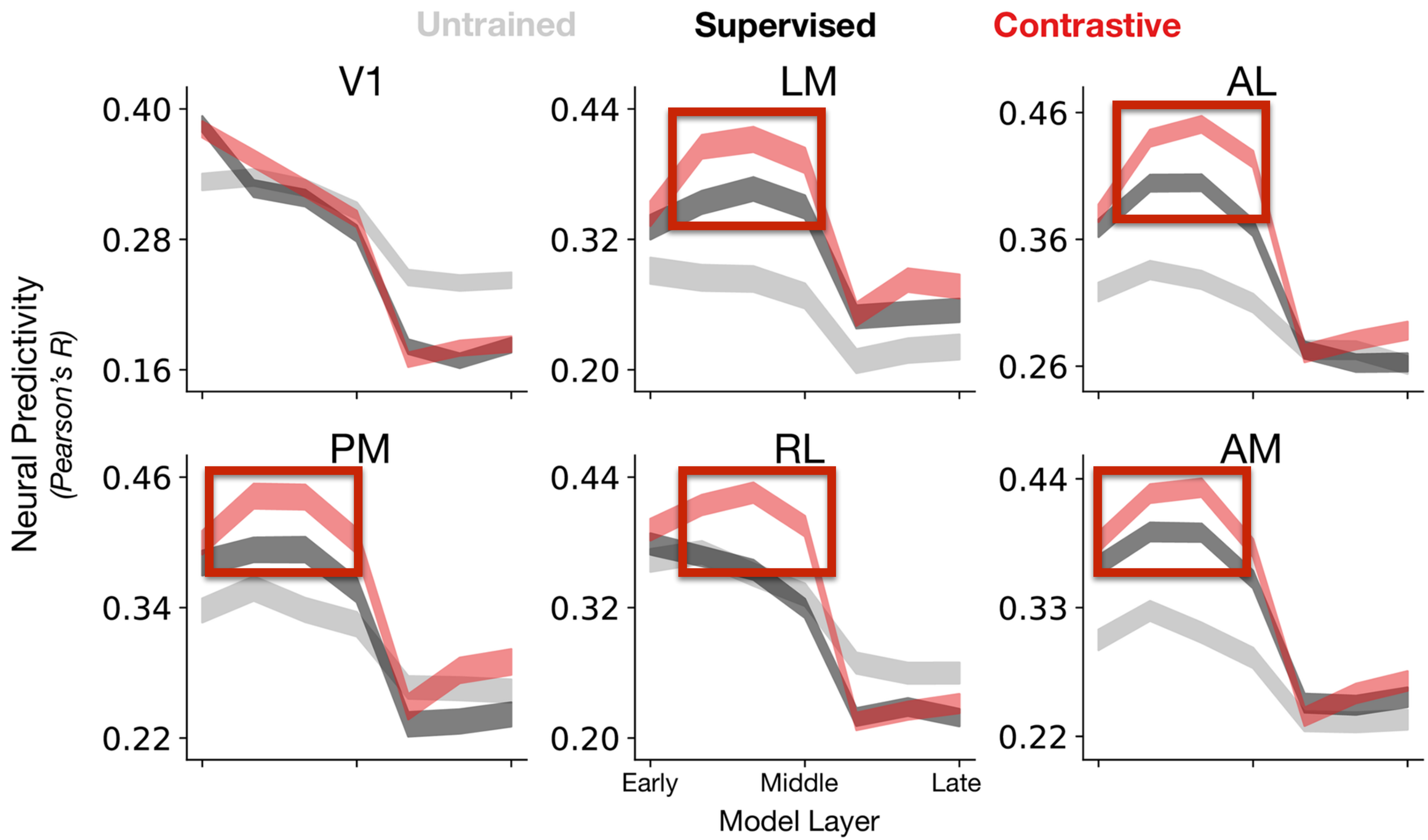


Primates

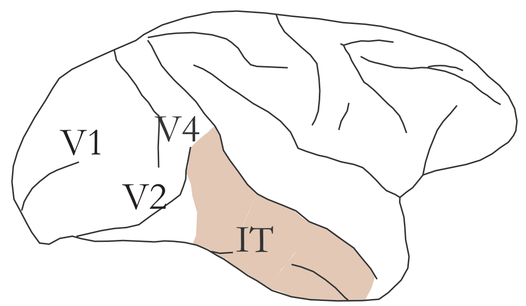


Mouse

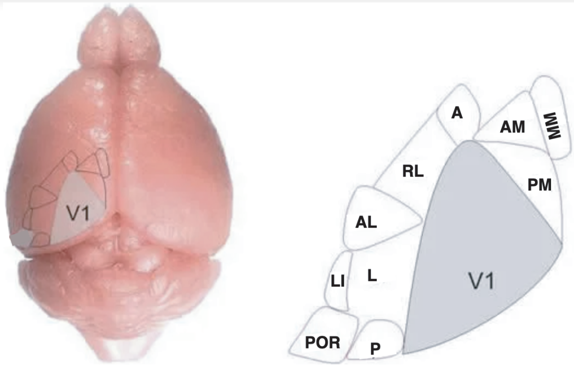
Mouse vision is less hierarchical!



Contrastive Models Better Match Mouse Visual Cortex

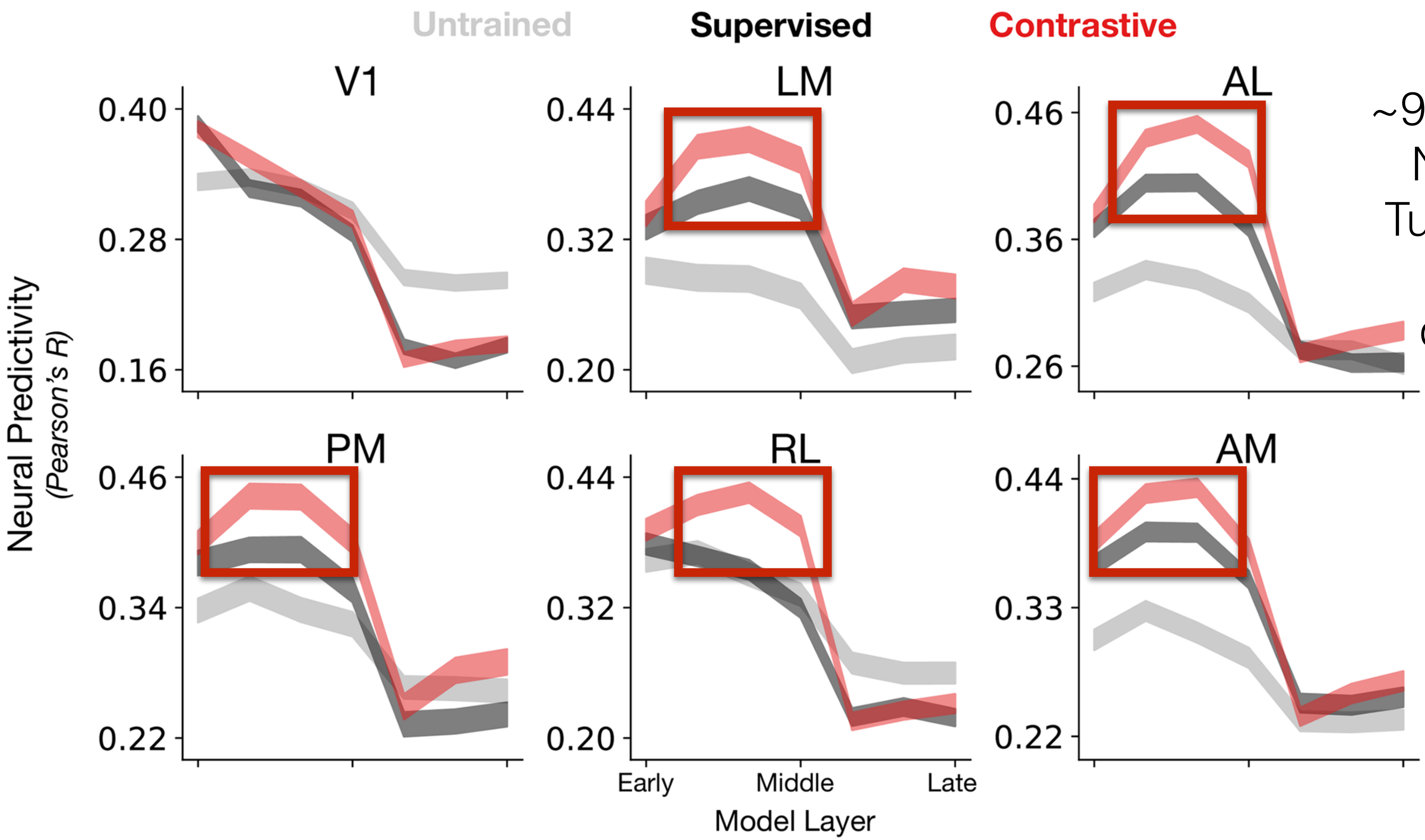


Primates



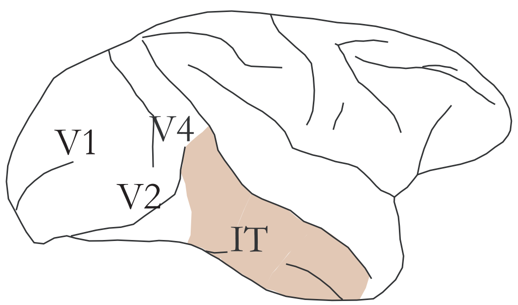
Mouse

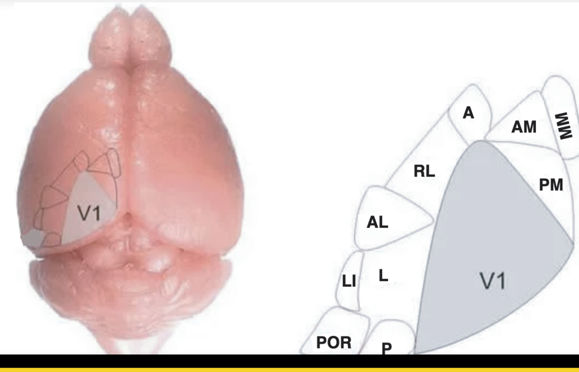
Mouse vision is less hierarchical!



~90% of the NeuroAI Turing Test for this dataset

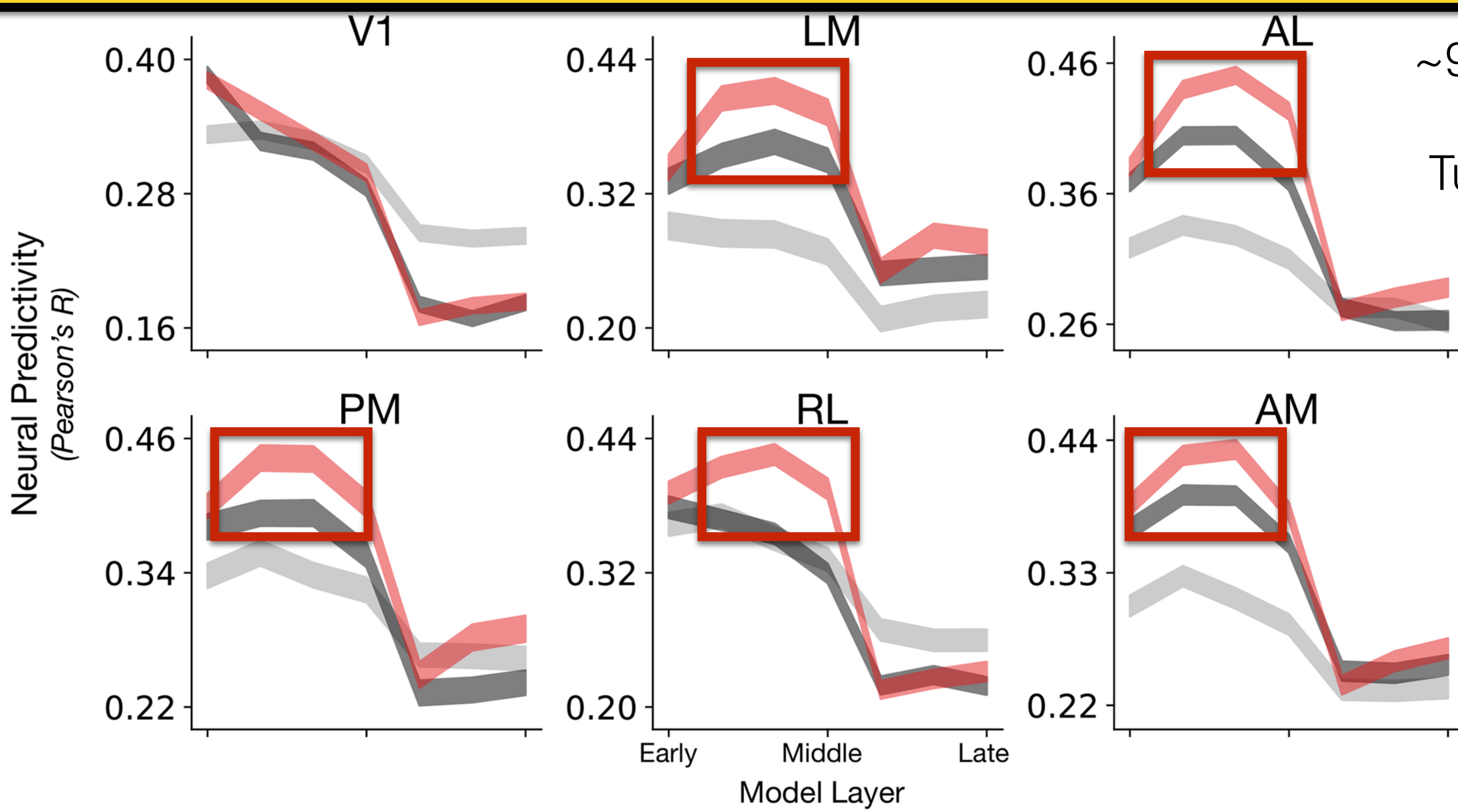
Contrastive Models Better Match Mouse Visual Cortex





Mouse vision is less hierarchical!

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



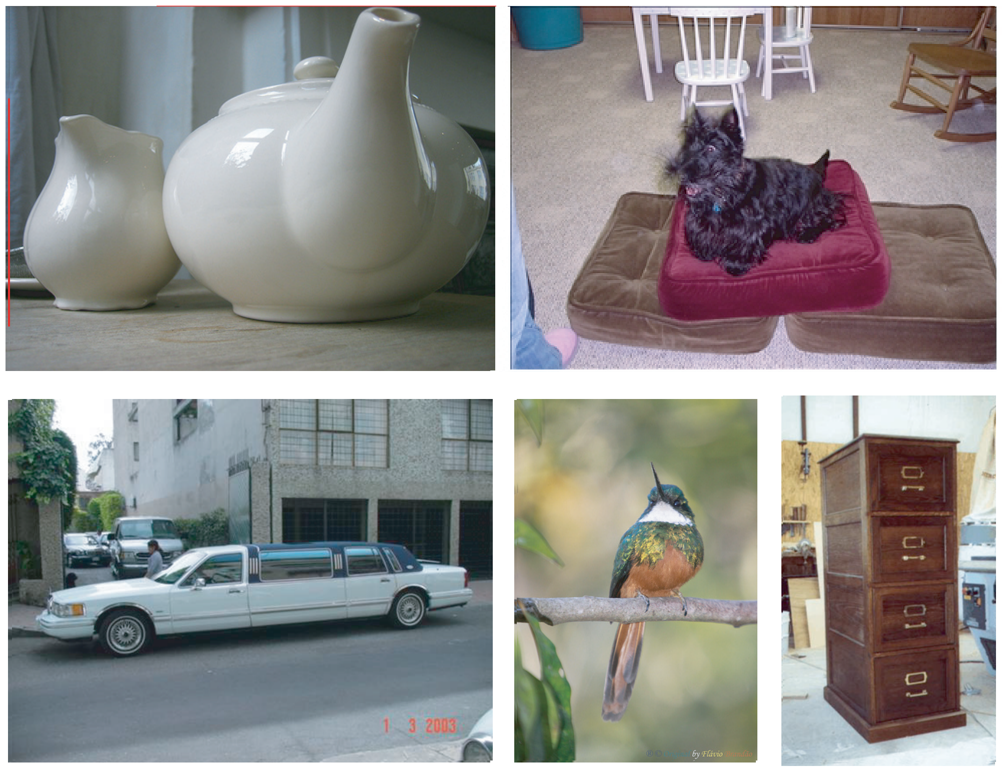
~90% of the NeuroAI Turing Test for this dataset

Assessing Task-Generality

Assessing Task-Generality

Train

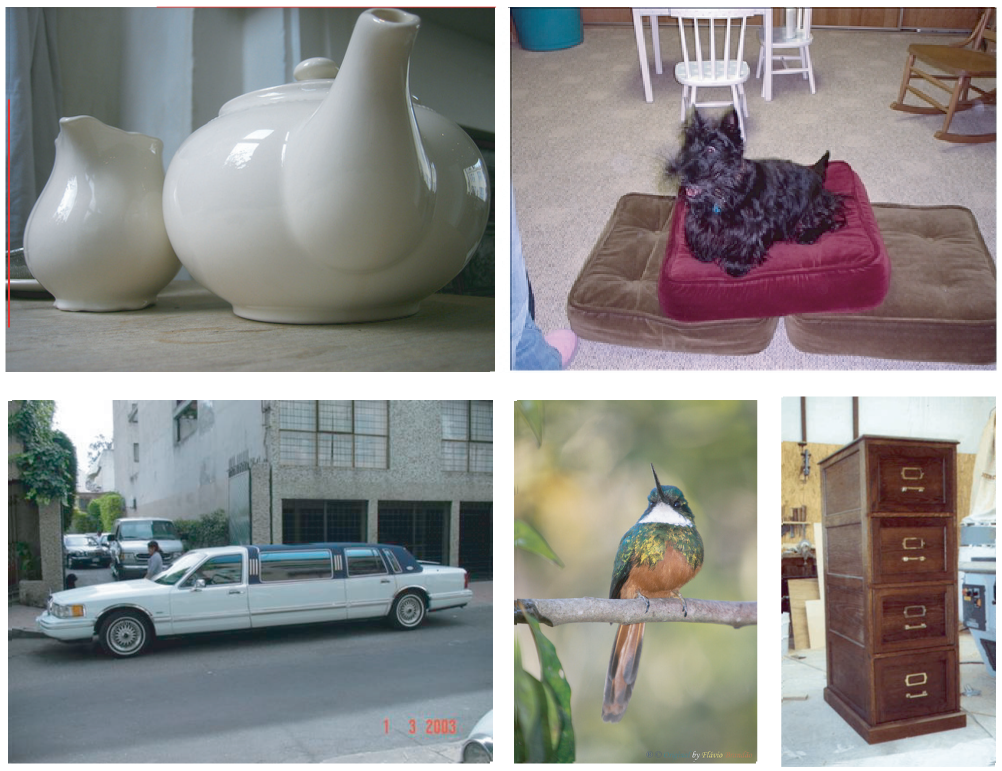
ImageNet



Assessing Task-Generality

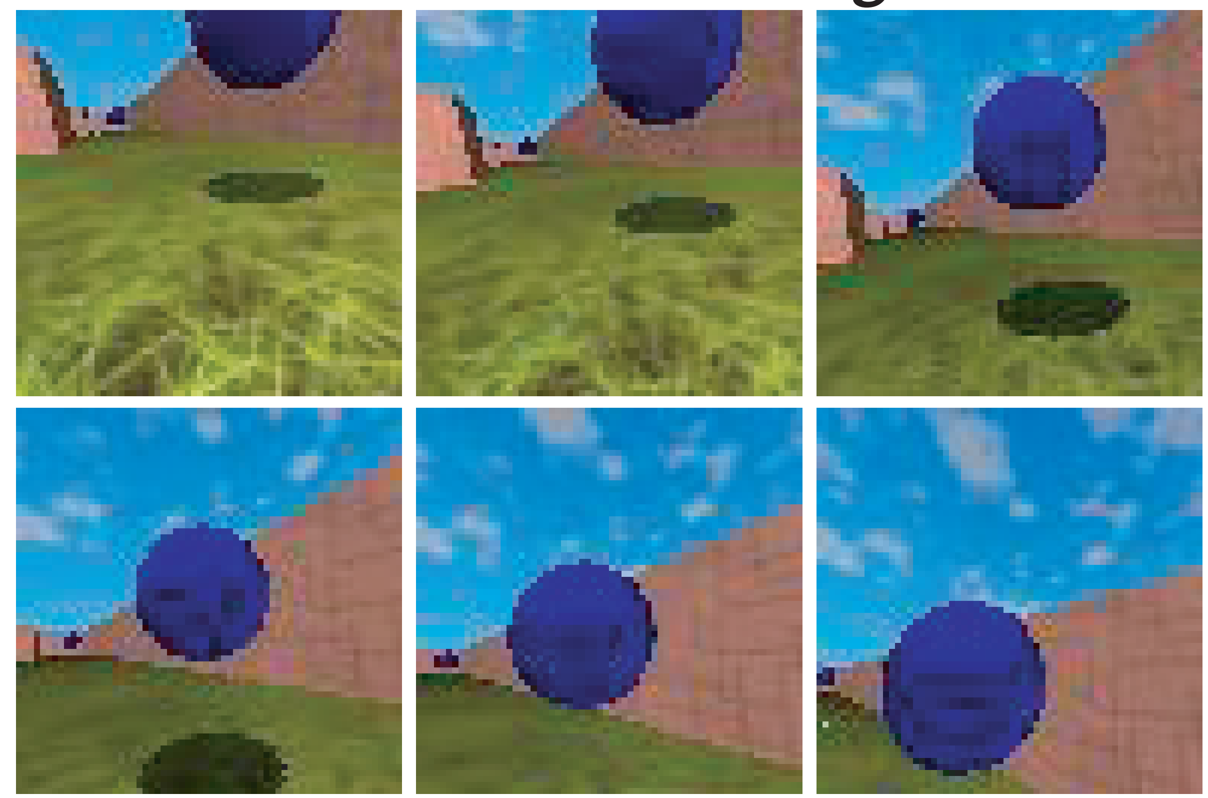
Train

ImageNet

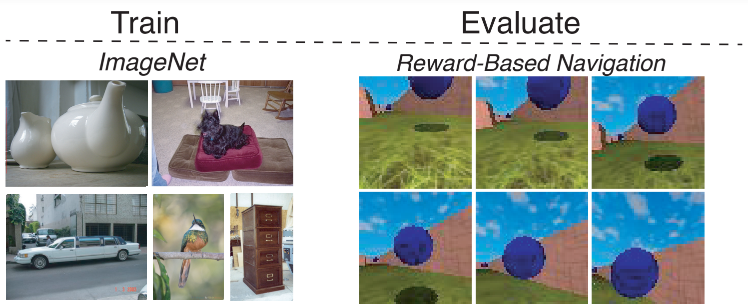


Evaluate

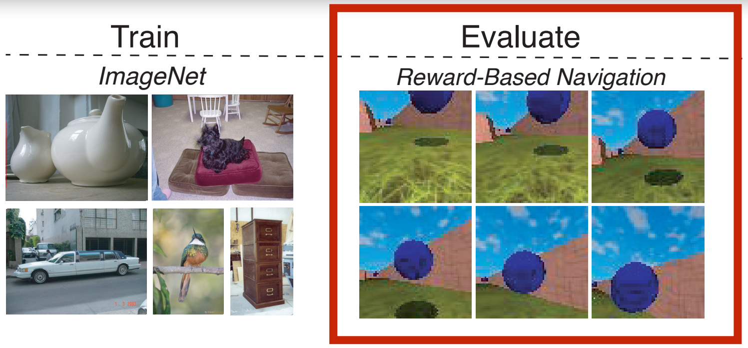
Reward-Based Navigation



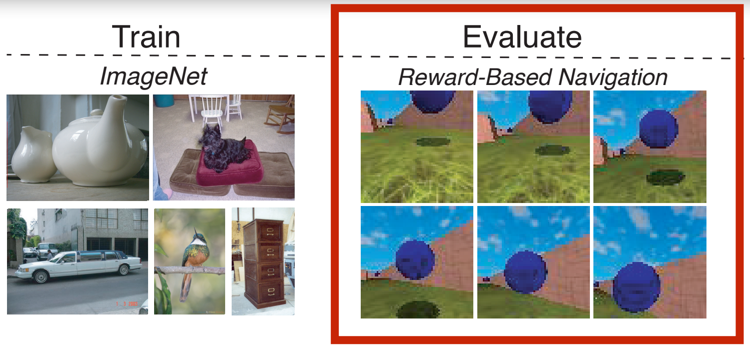
Assessing Task-Generality



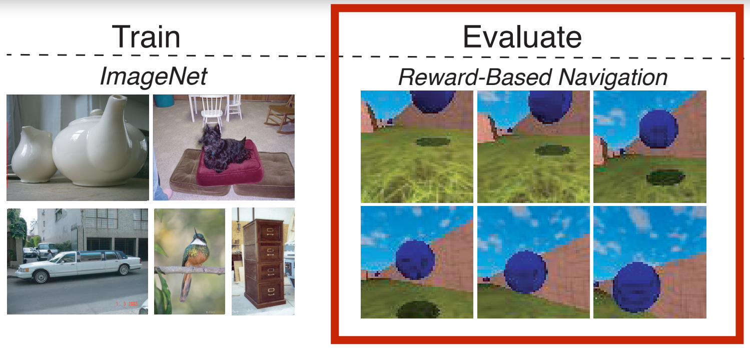
Assessing Task-Generality



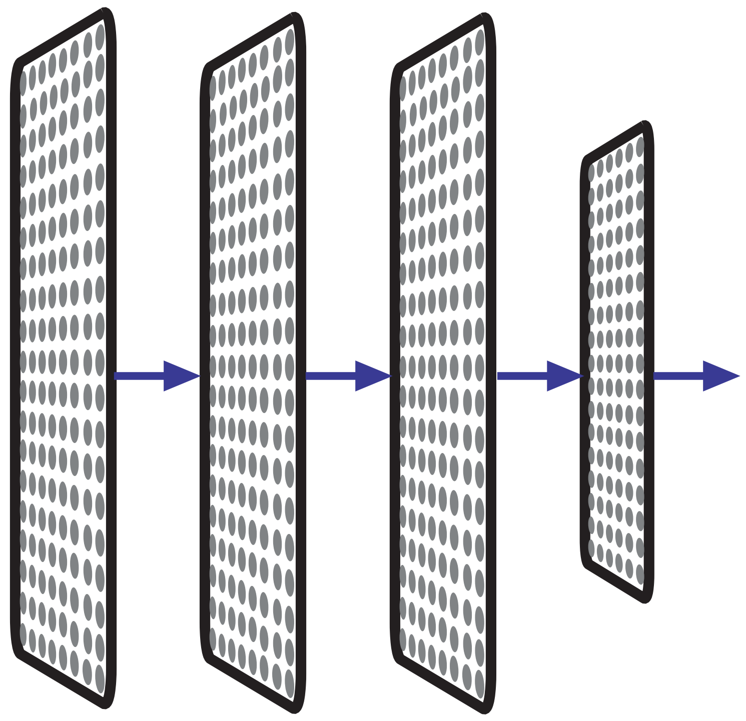
Embodied Virtual Rodent Navigation



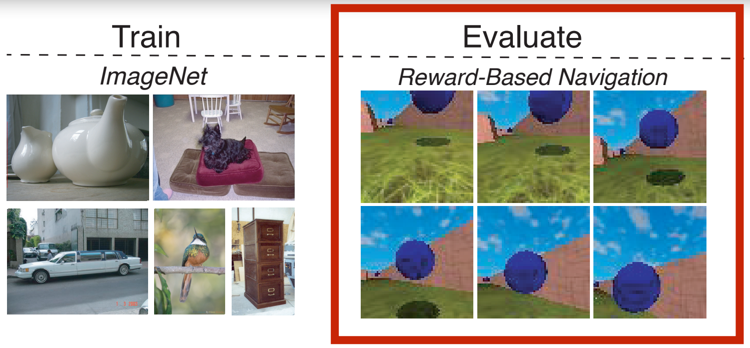
Embodied Virtual Rodent Navigation



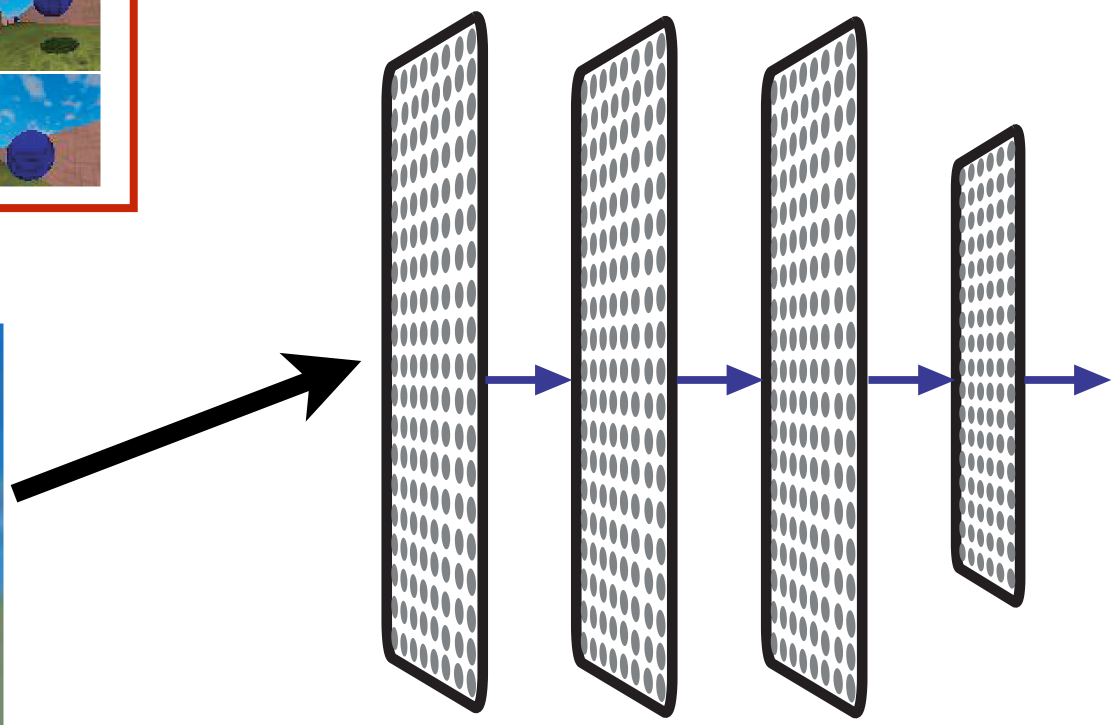
Vision Network

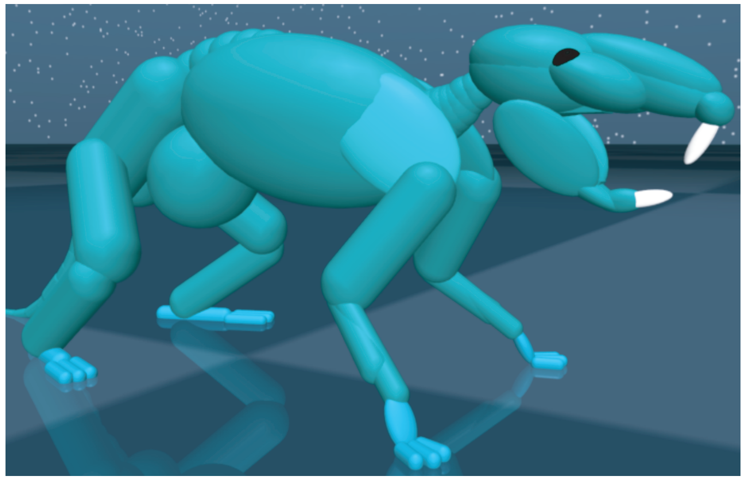


Embodied Virtual Rodent Navigation



Vision Network

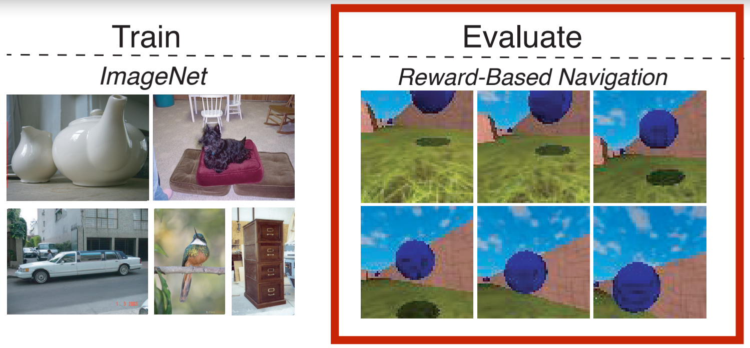




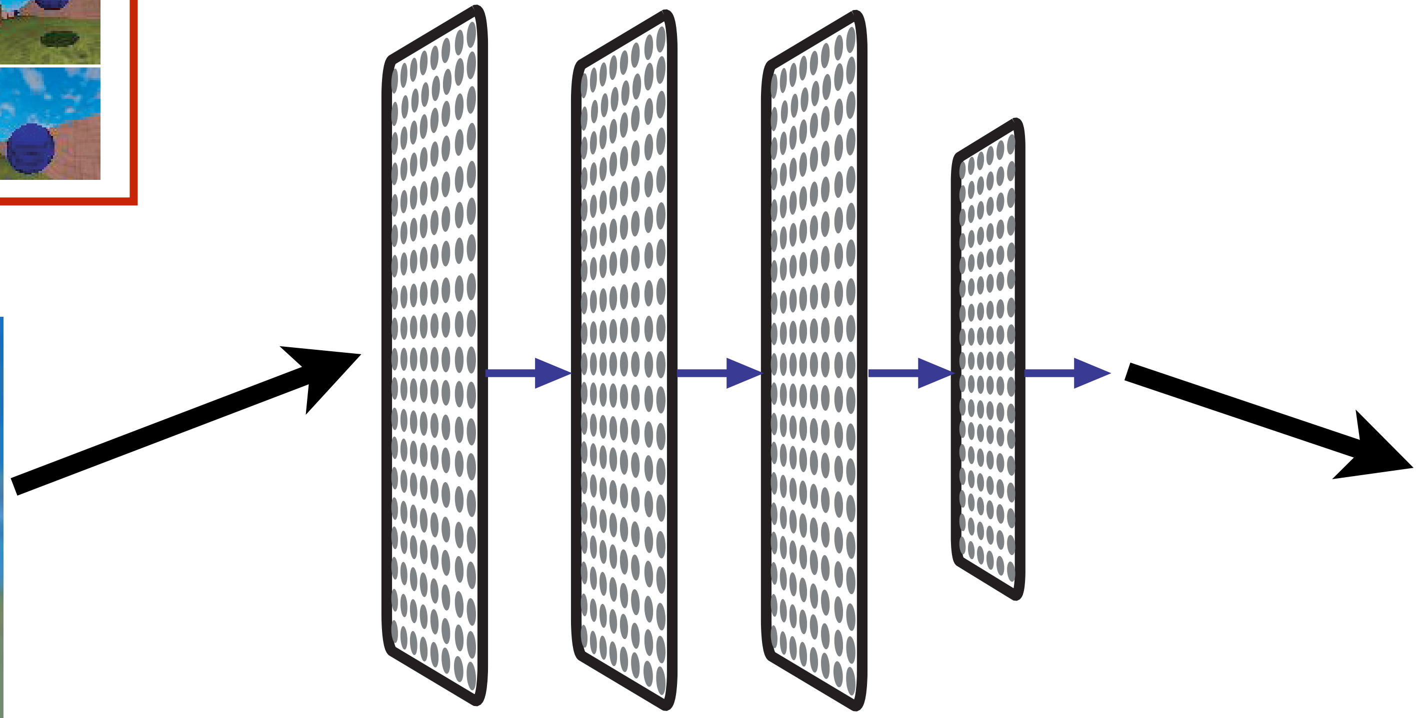
Biomechanical Model
(Joint angles, accelerometer, etc.)

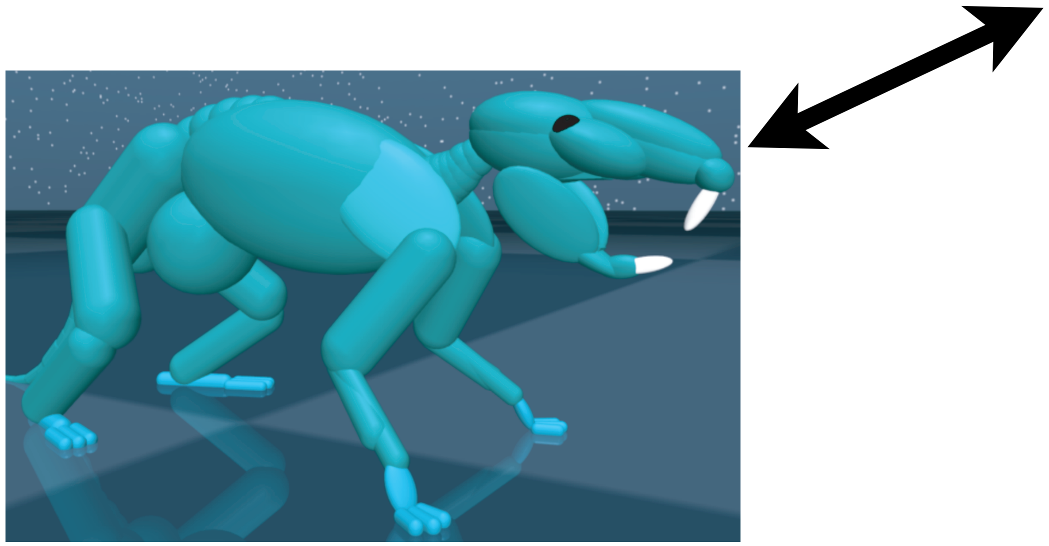
Bence Ölveczky

Embodied Virtual Rodent Navigation



Vision Network

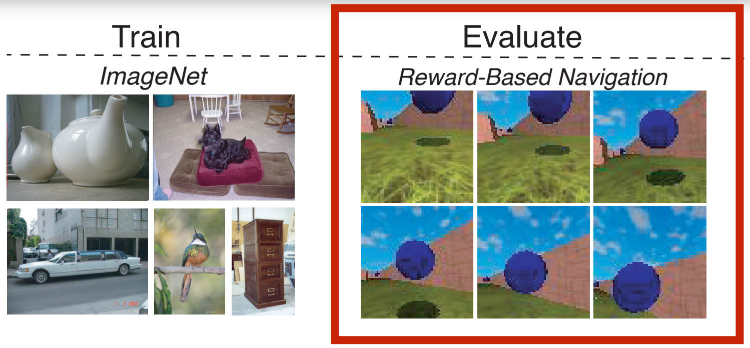




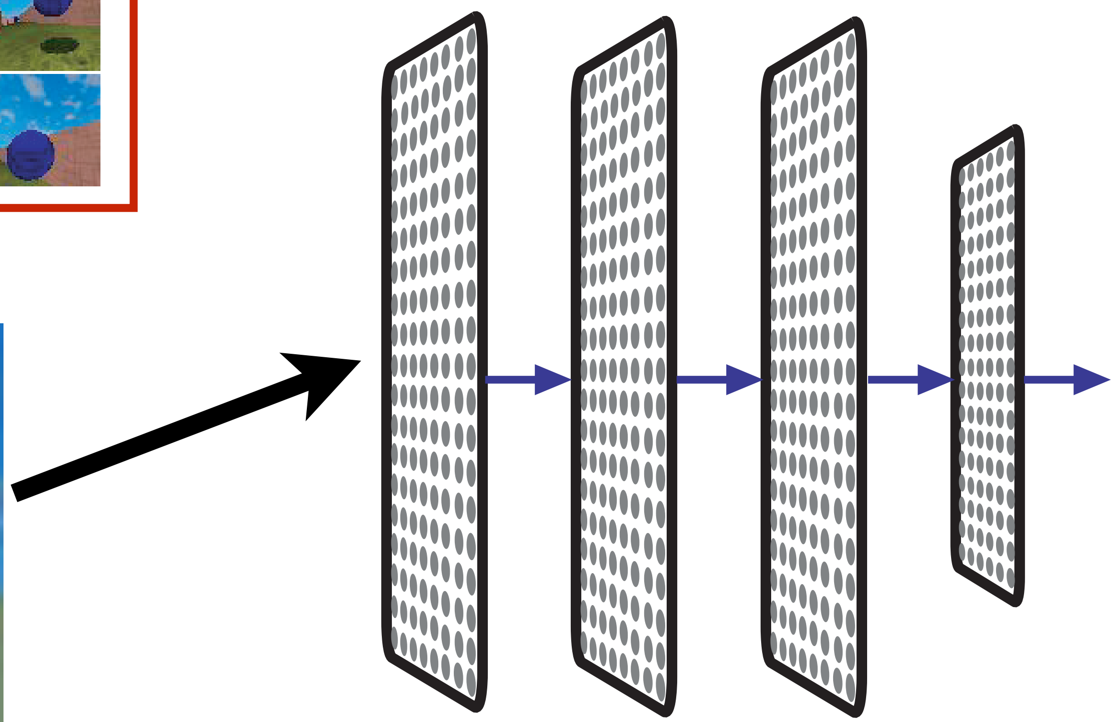
Biomechanical Model
(Joint angles, accelerometer, etc.)

Bence Ölveczky

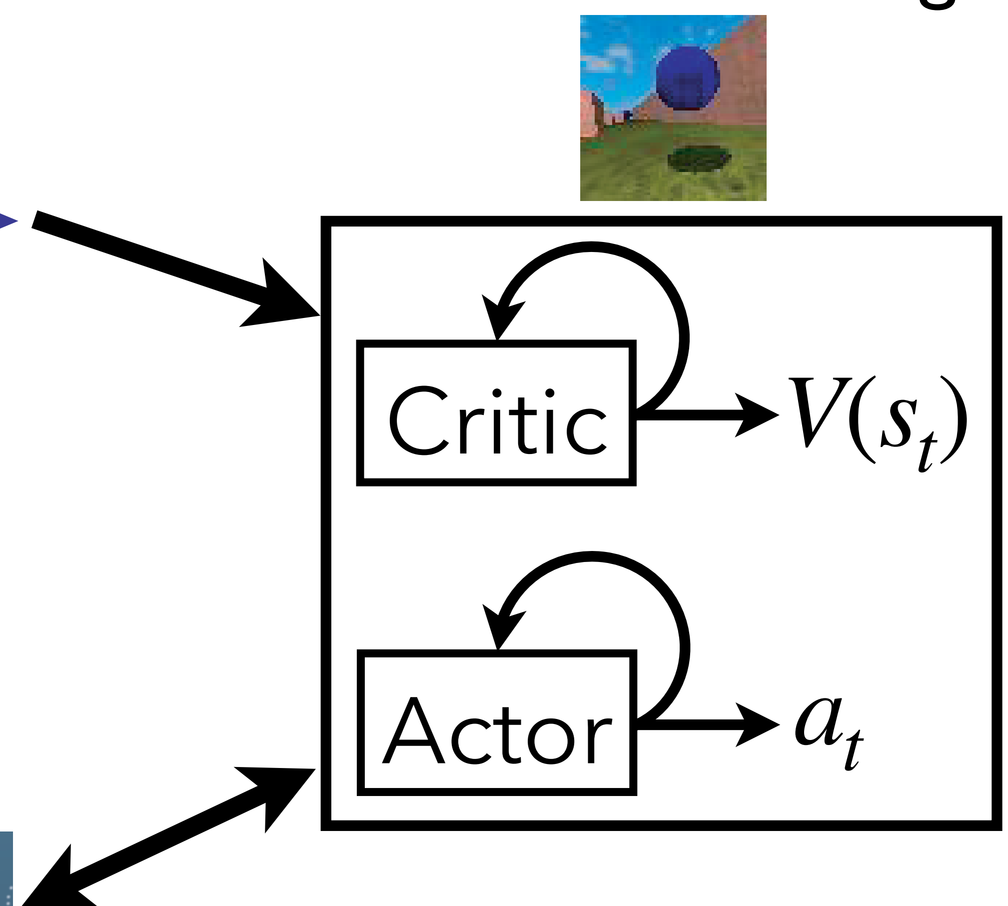
Embodied Virtual Rodent Navigation

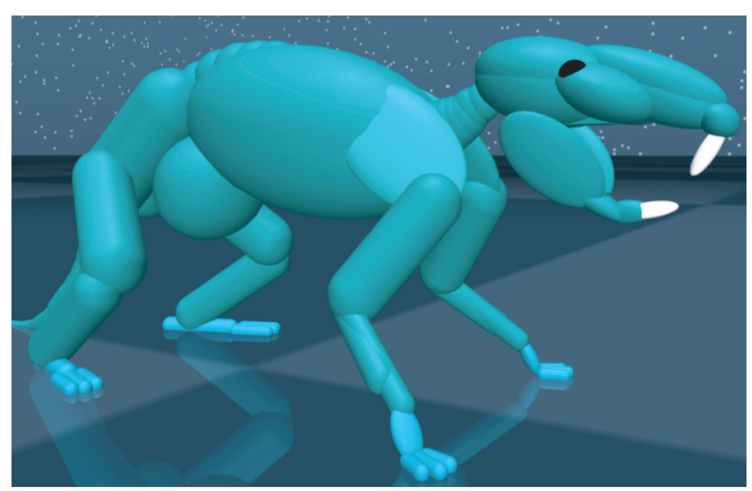


Vision Network



Decision Making

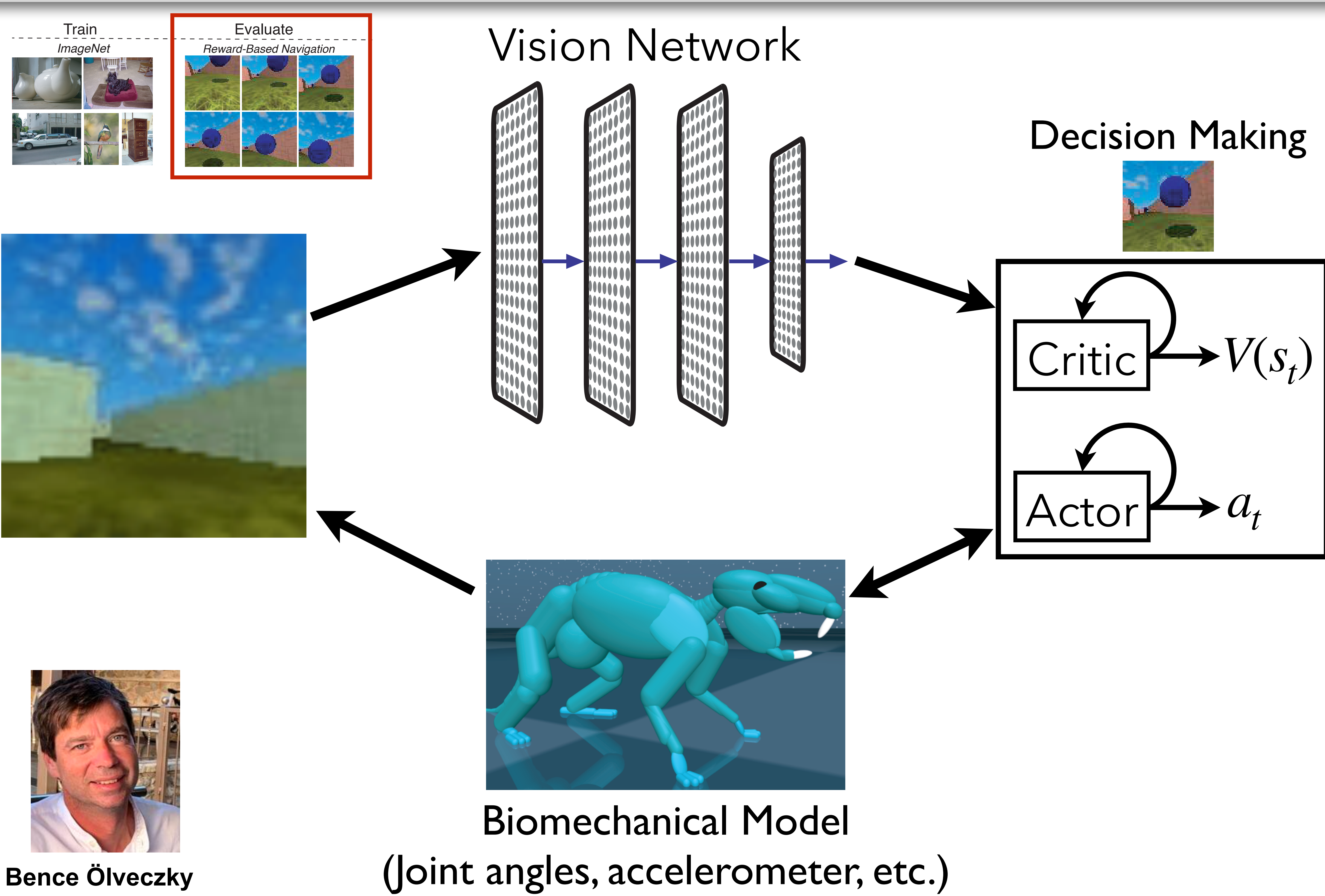




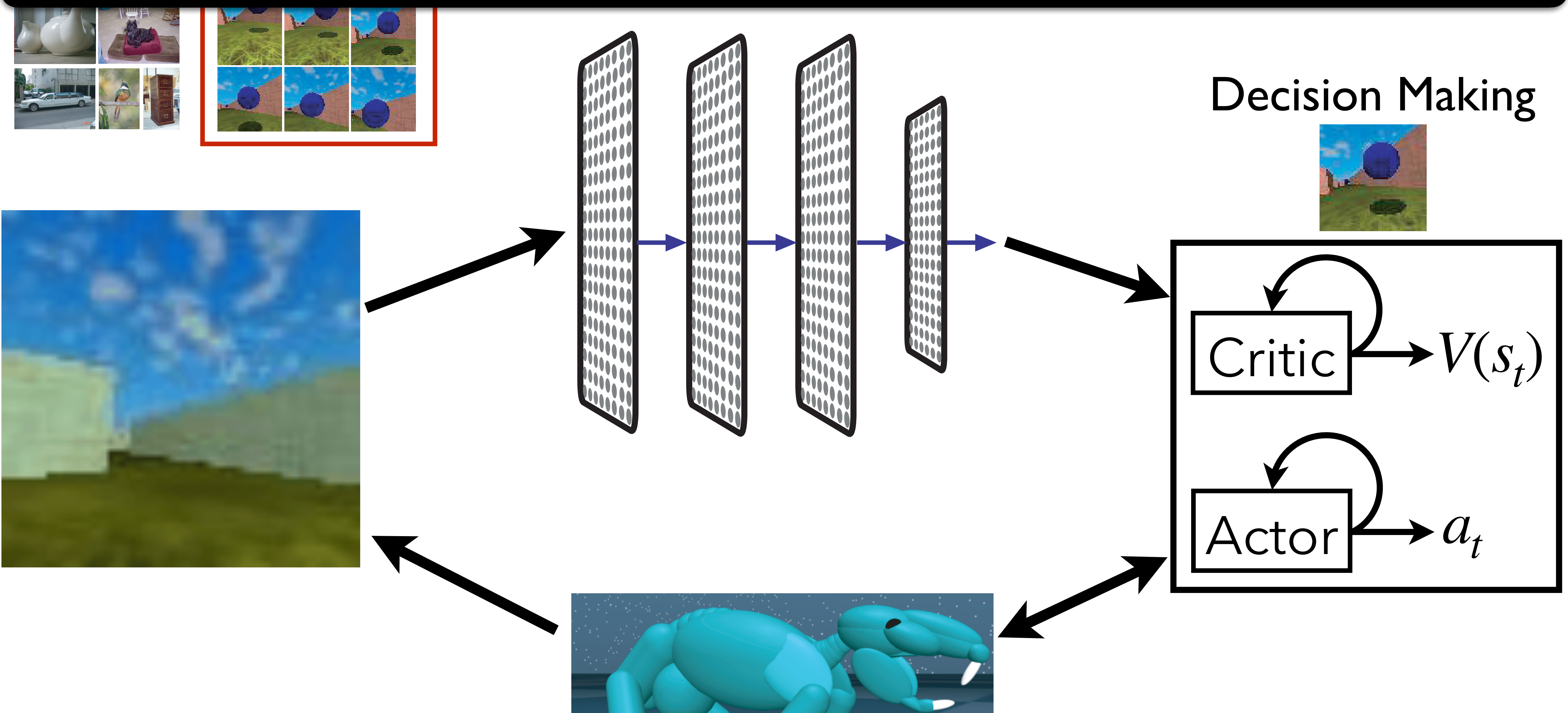
Biomechanical Model
(Joint angles, accelerometer, etc.)

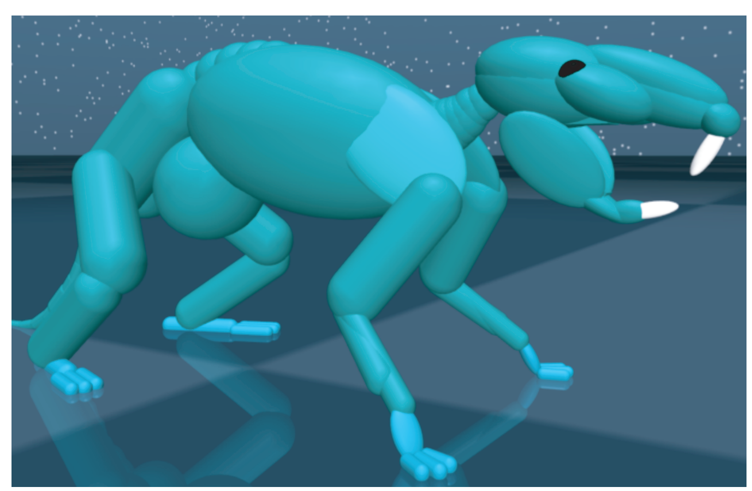
Bence Ölveczky

Embodied Virtual Rodent Navigation



High degree-of-freedom body (38/74 controllable degrees), keeping track of history over long timescales with high-dimensional, continuous inputs



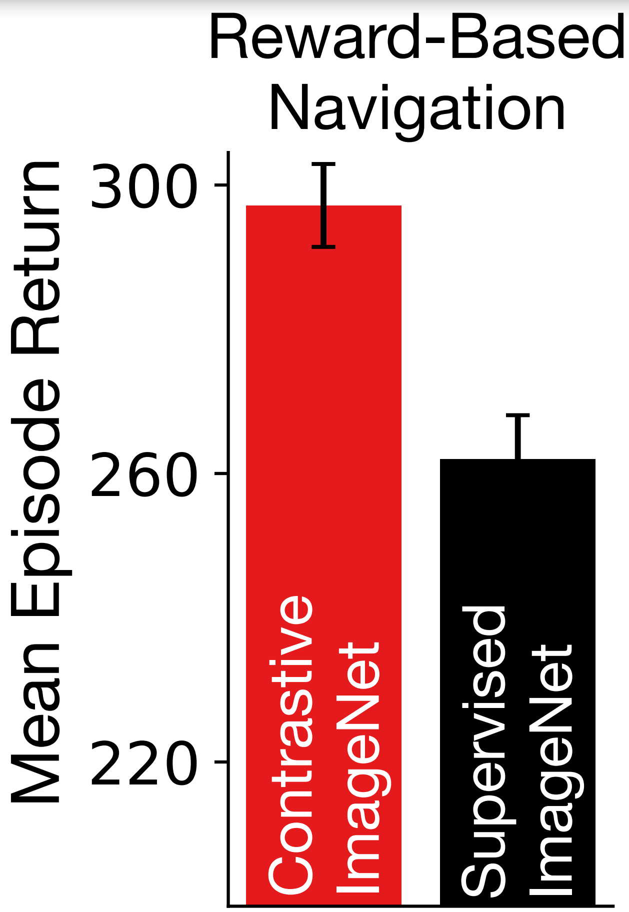


Biomechanical Model
(Joint angles, accelerometer, etc.)

Bence Ölveczky

Contrastive Models Yield Better Transfer Performance

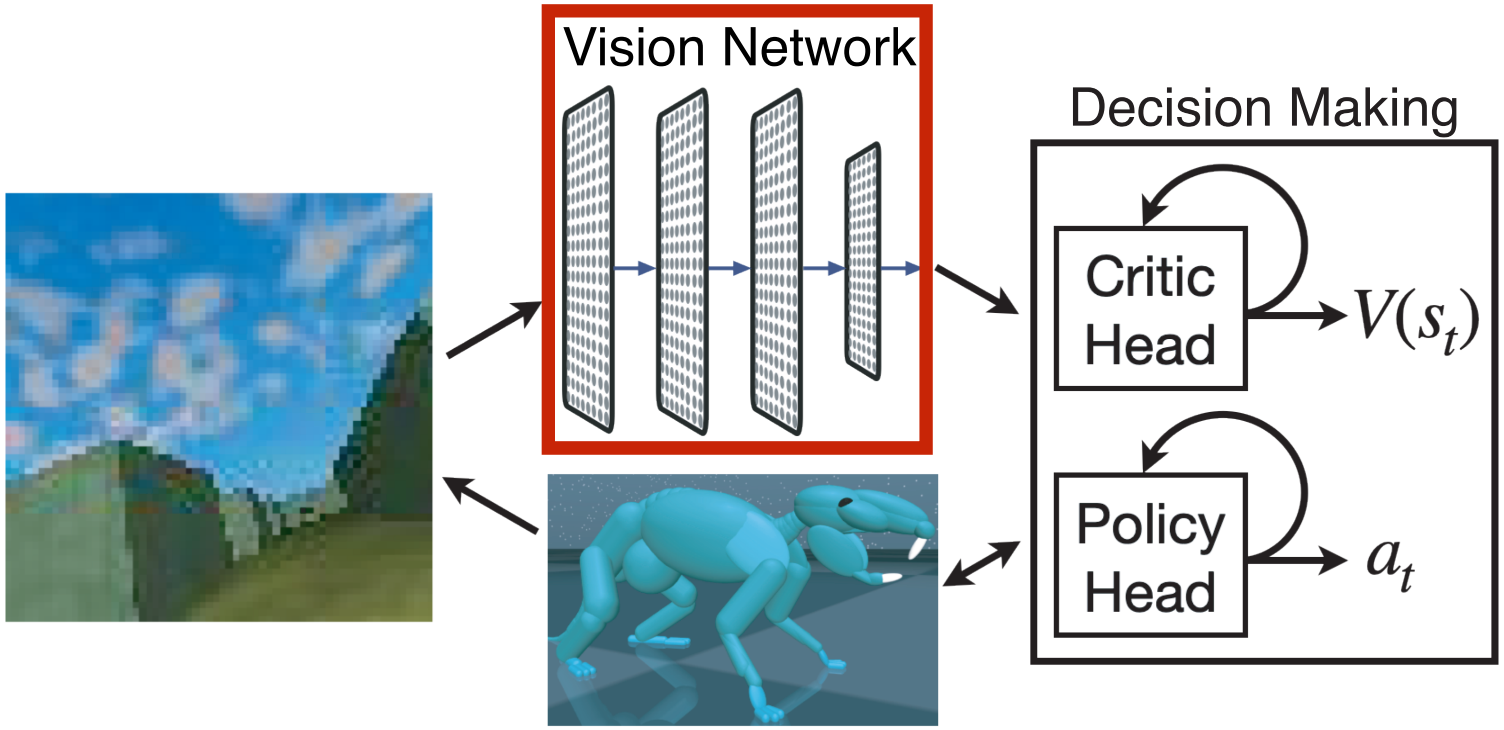
Contrastive Models Yield Better Transfer Performance



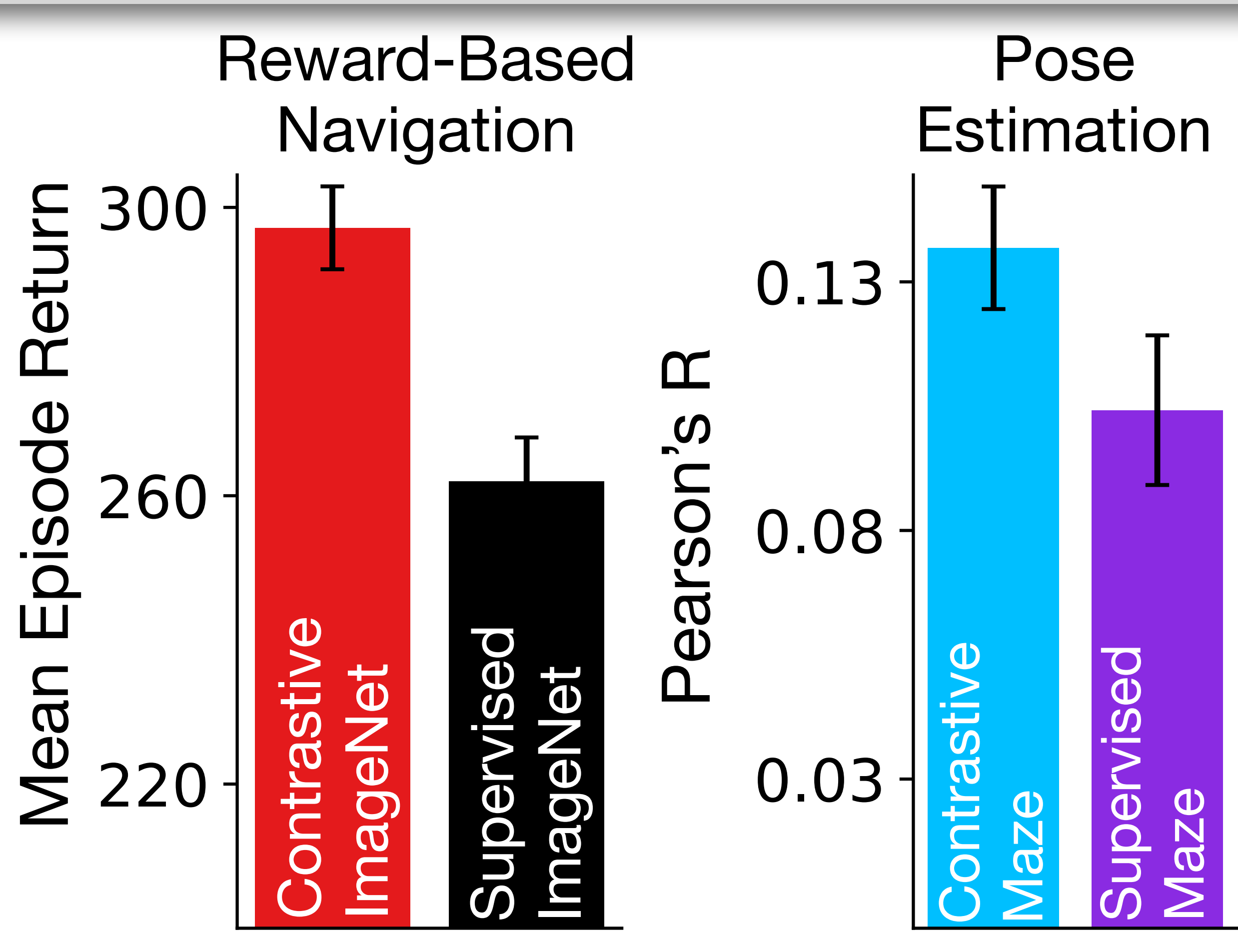
Train

ImageNet

Evaluate

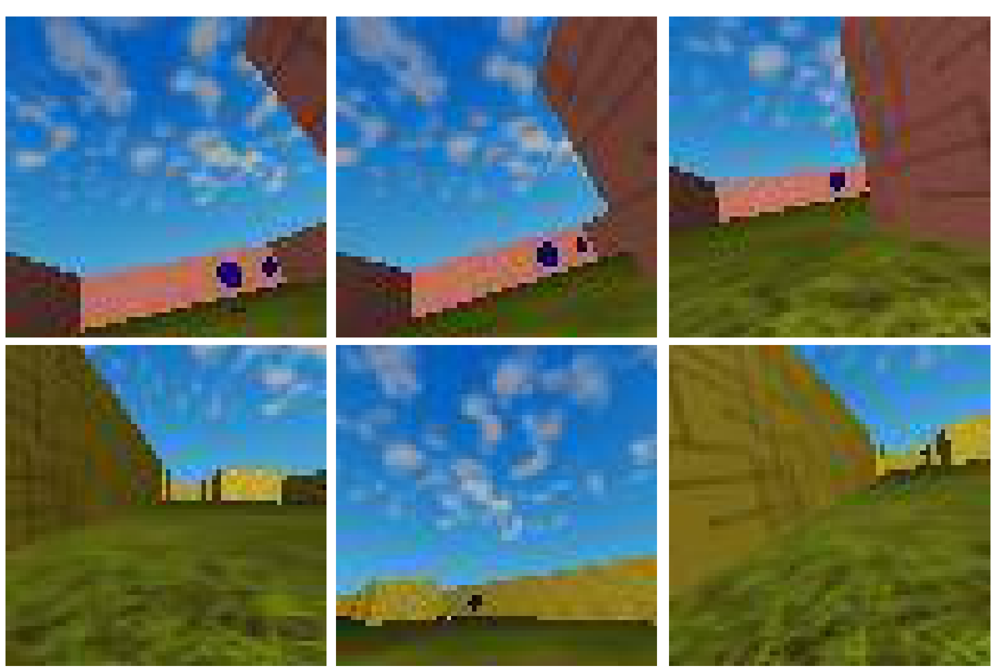


Contrastive Models Yield Better Transfer Performance



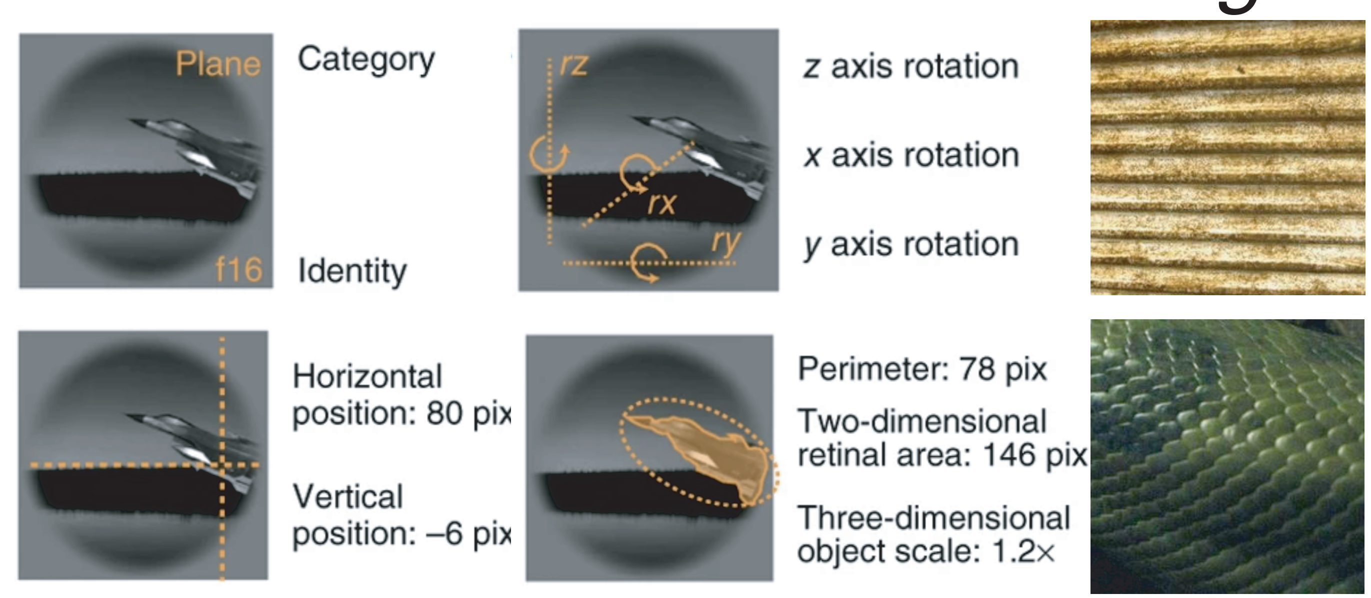
Train

Maze Environment



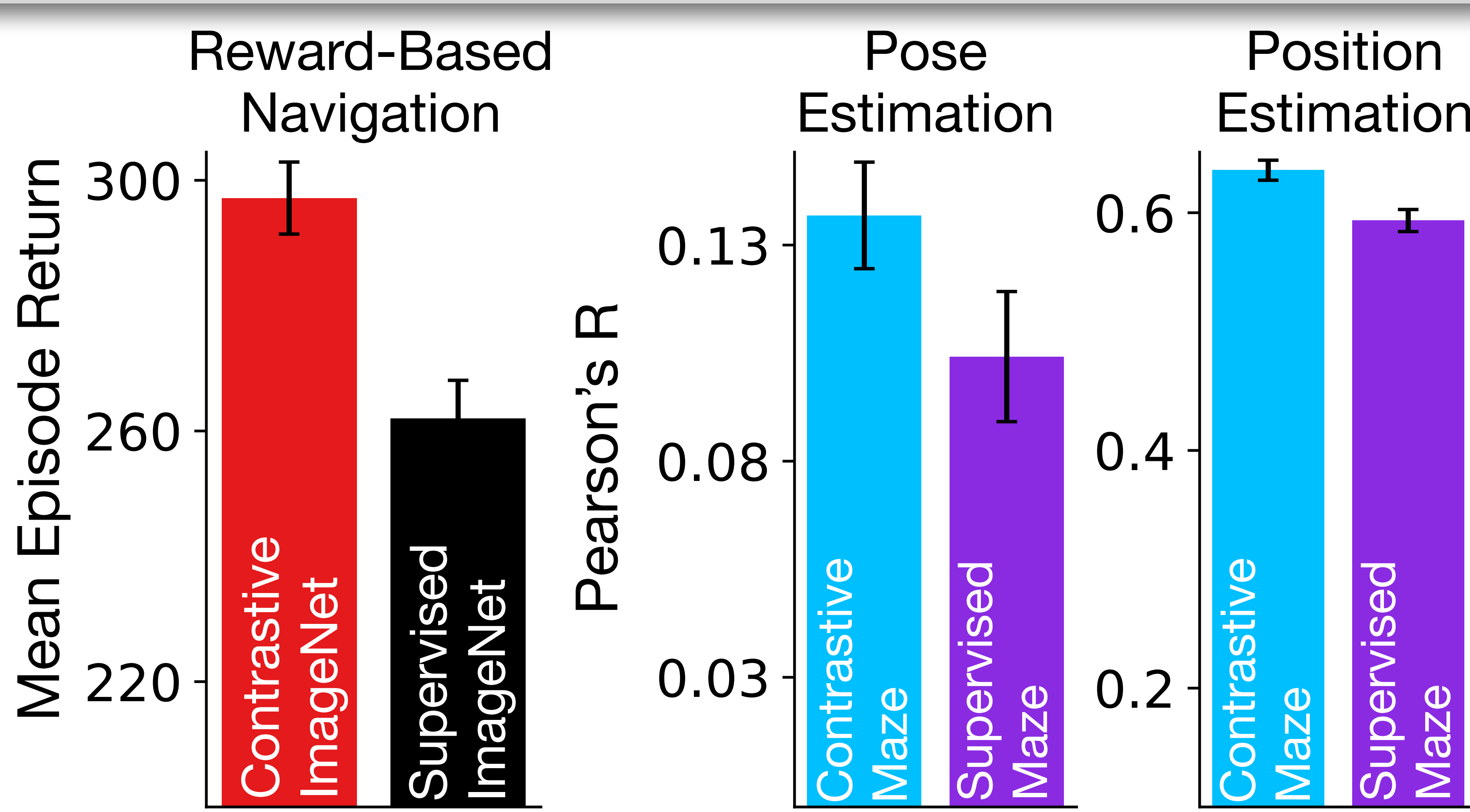
Evaluate

Visual Scene Understanding



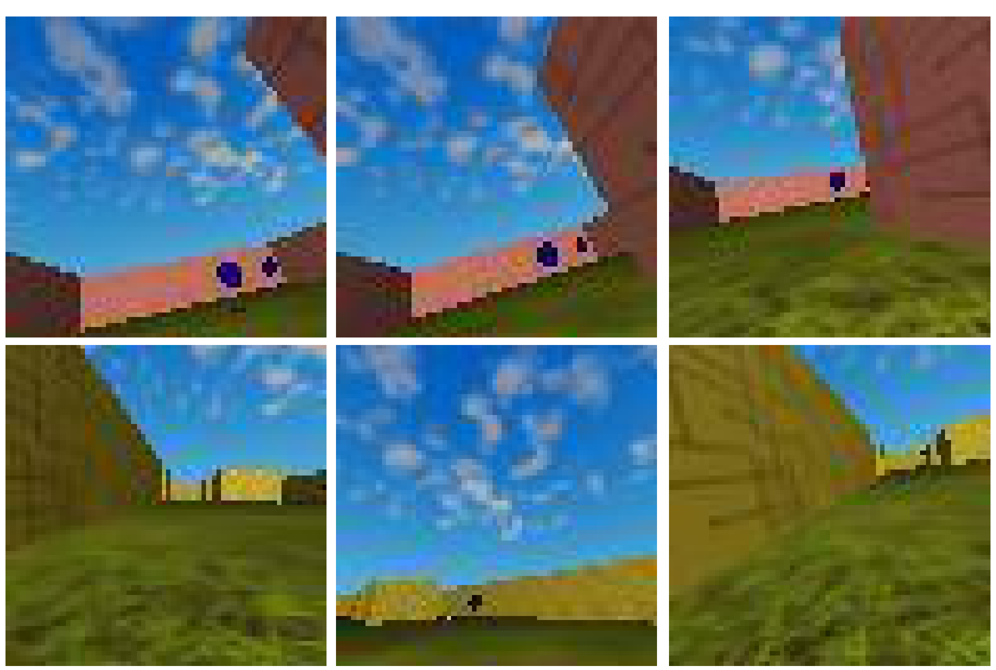
Object properties *Texture*

Contrastive Models Yield Better Transfer Performance



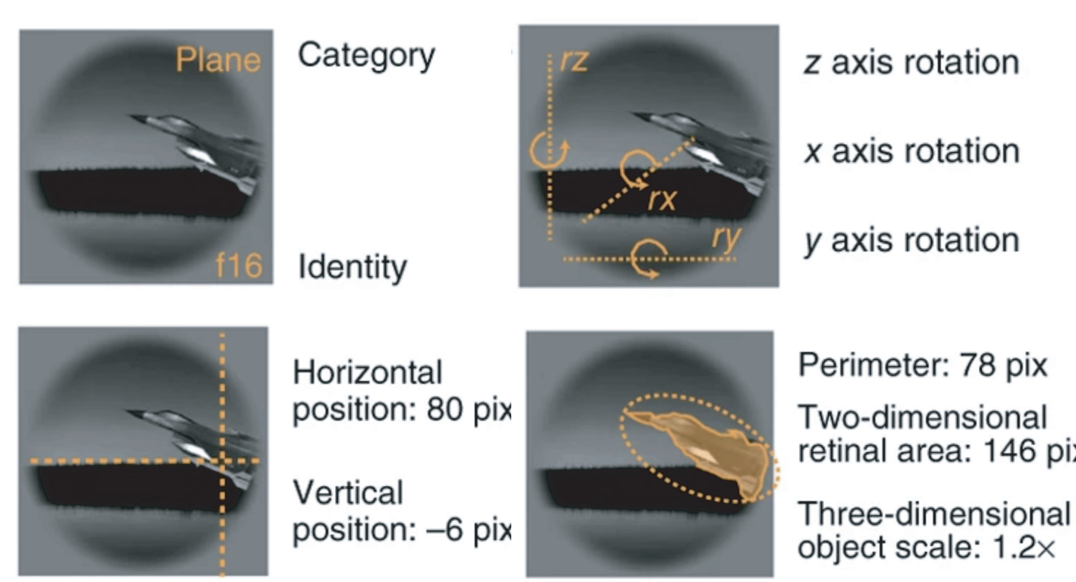
Train

Maze Environment

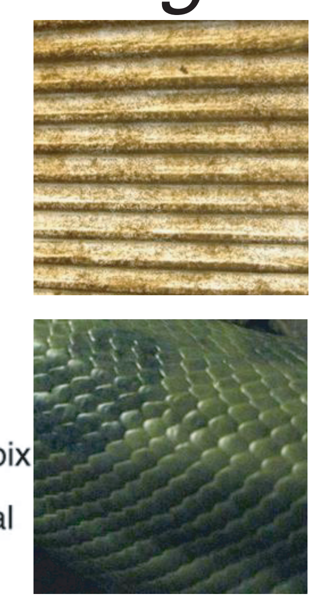


Evaluate

Visual Scene Understanding

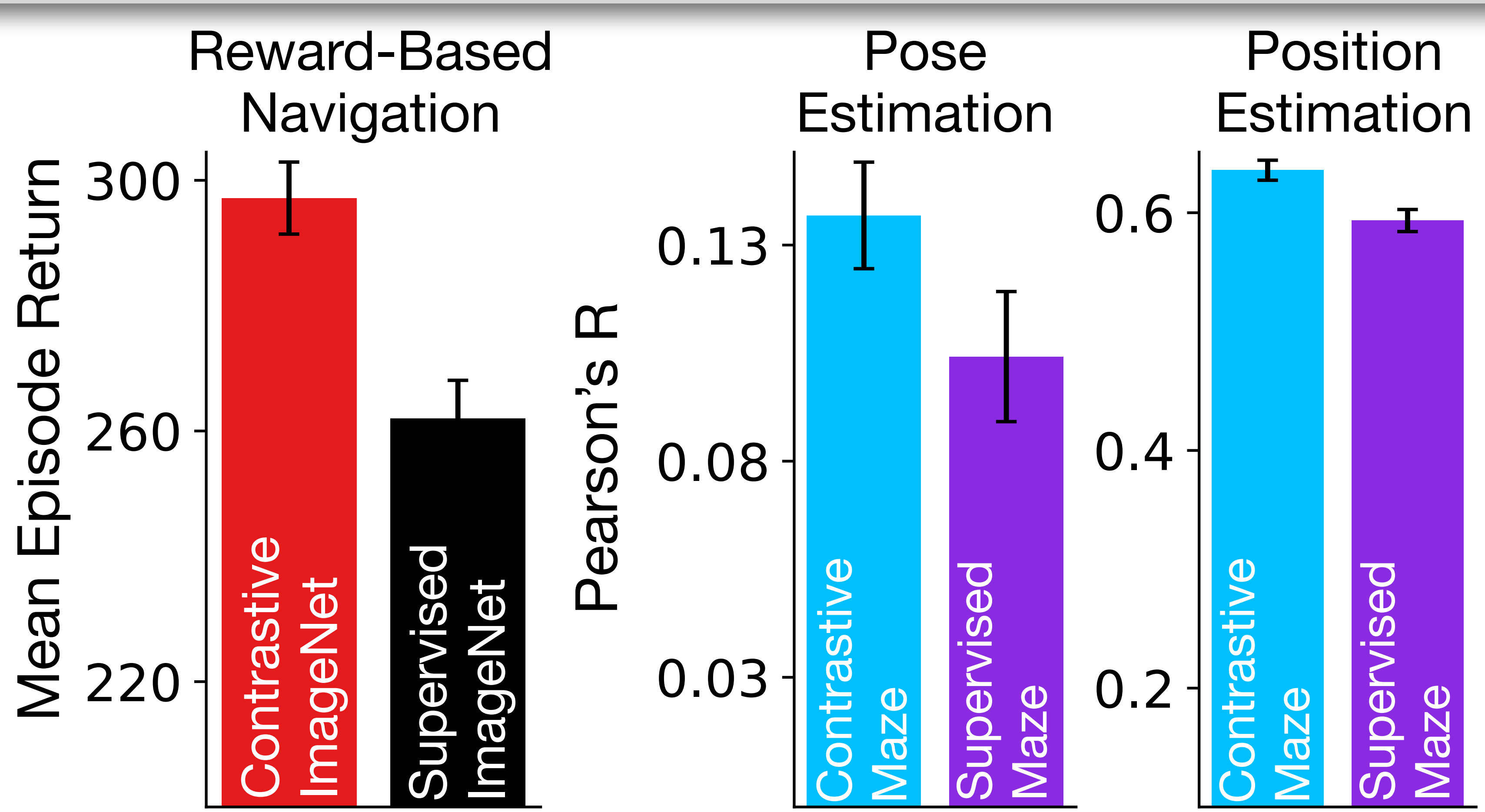


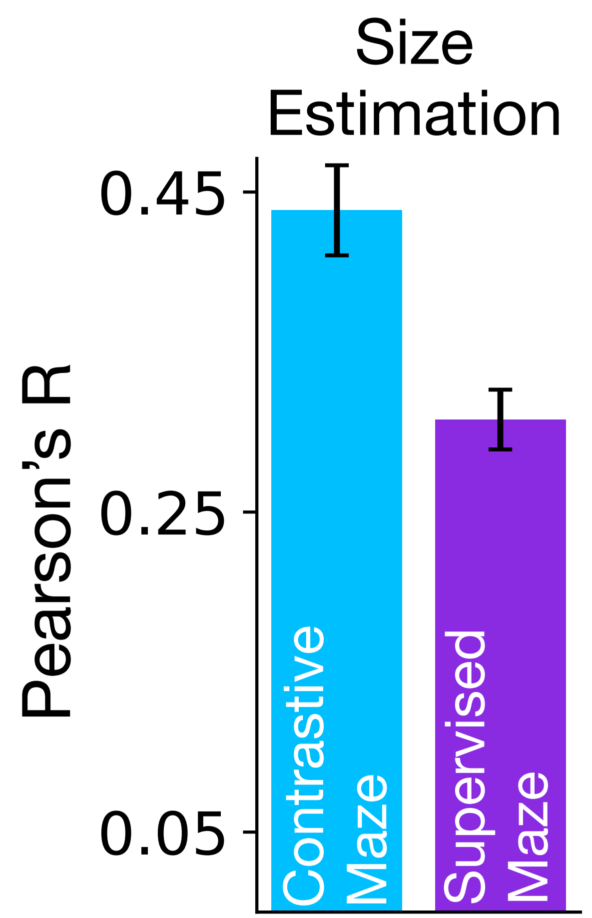
Object properties



Texture

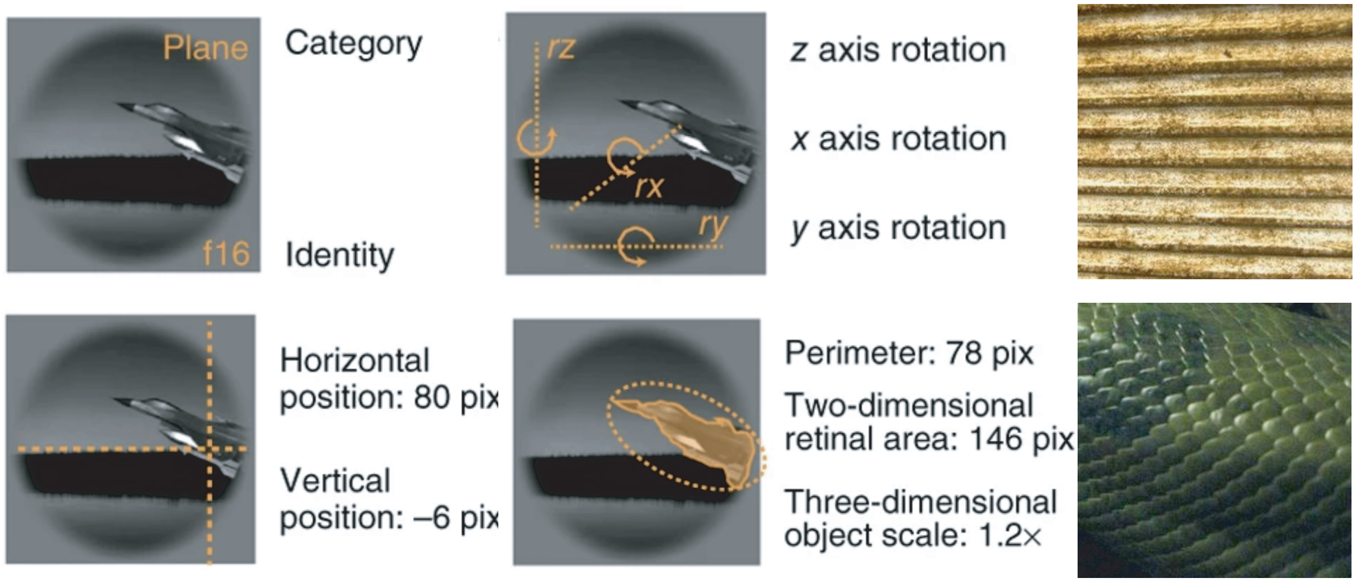
Contrastive Models Yield Better Transfer Performance





Evaluate

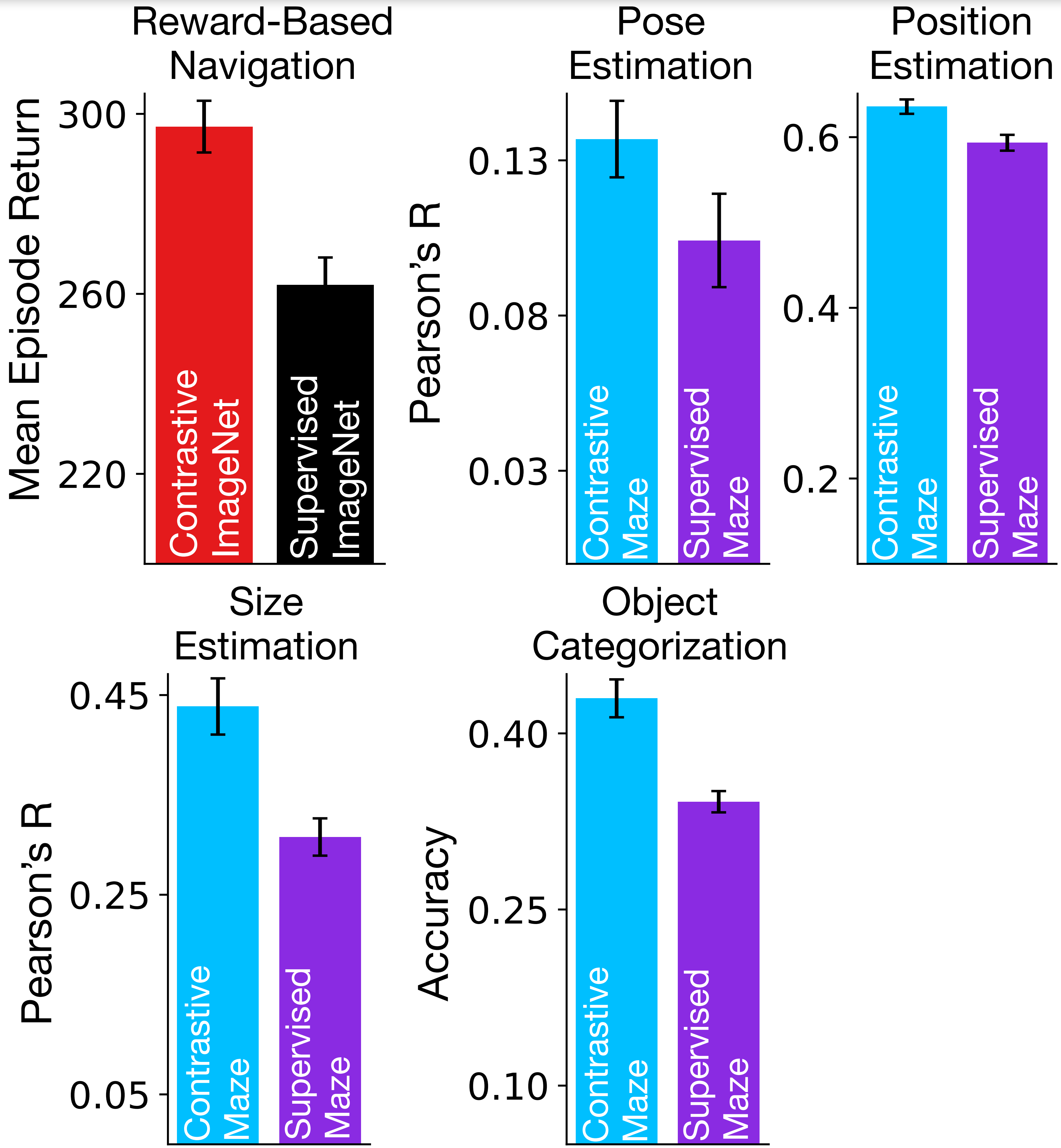
Visual Scene Understanding



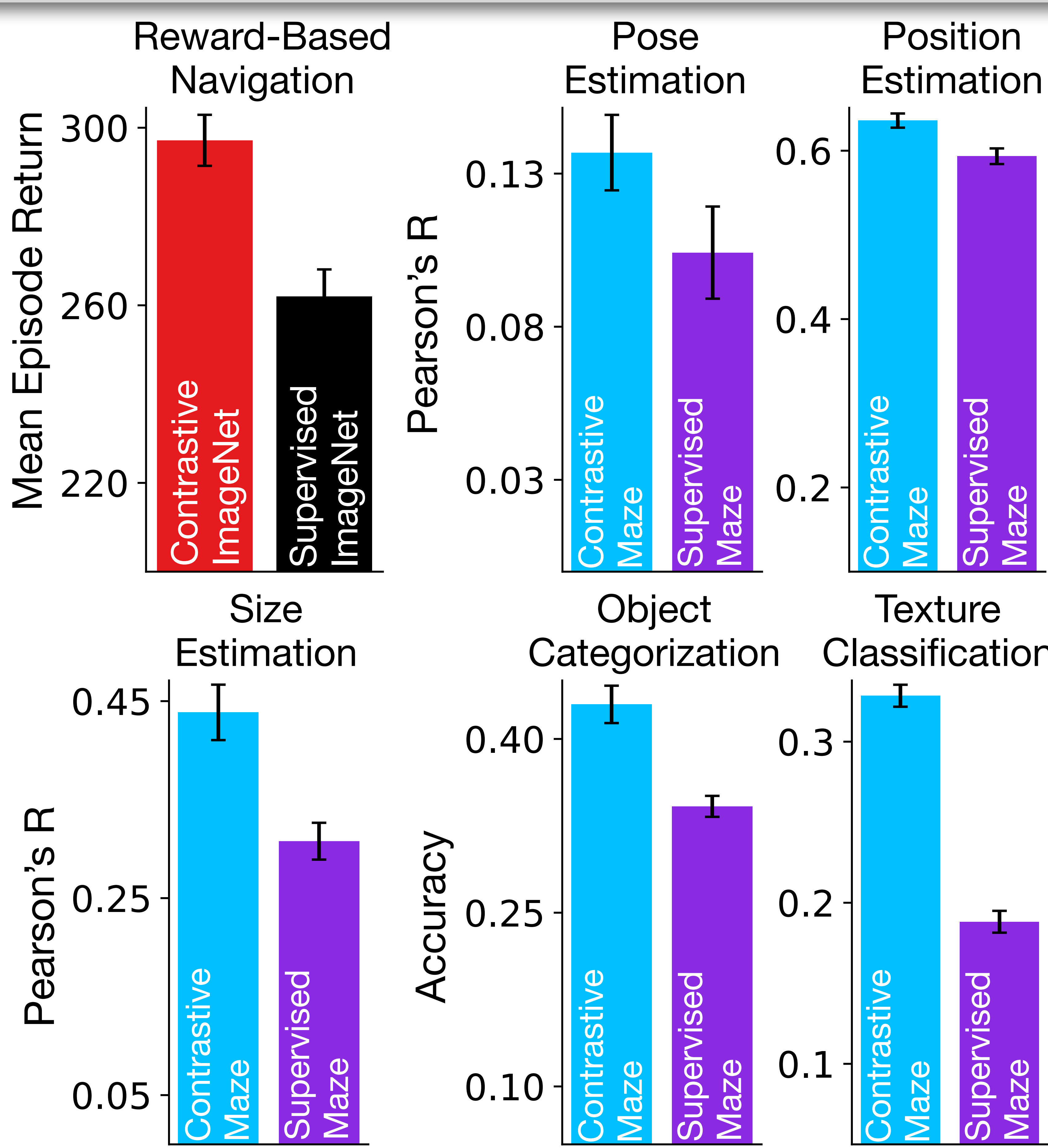
Object properties

Texture

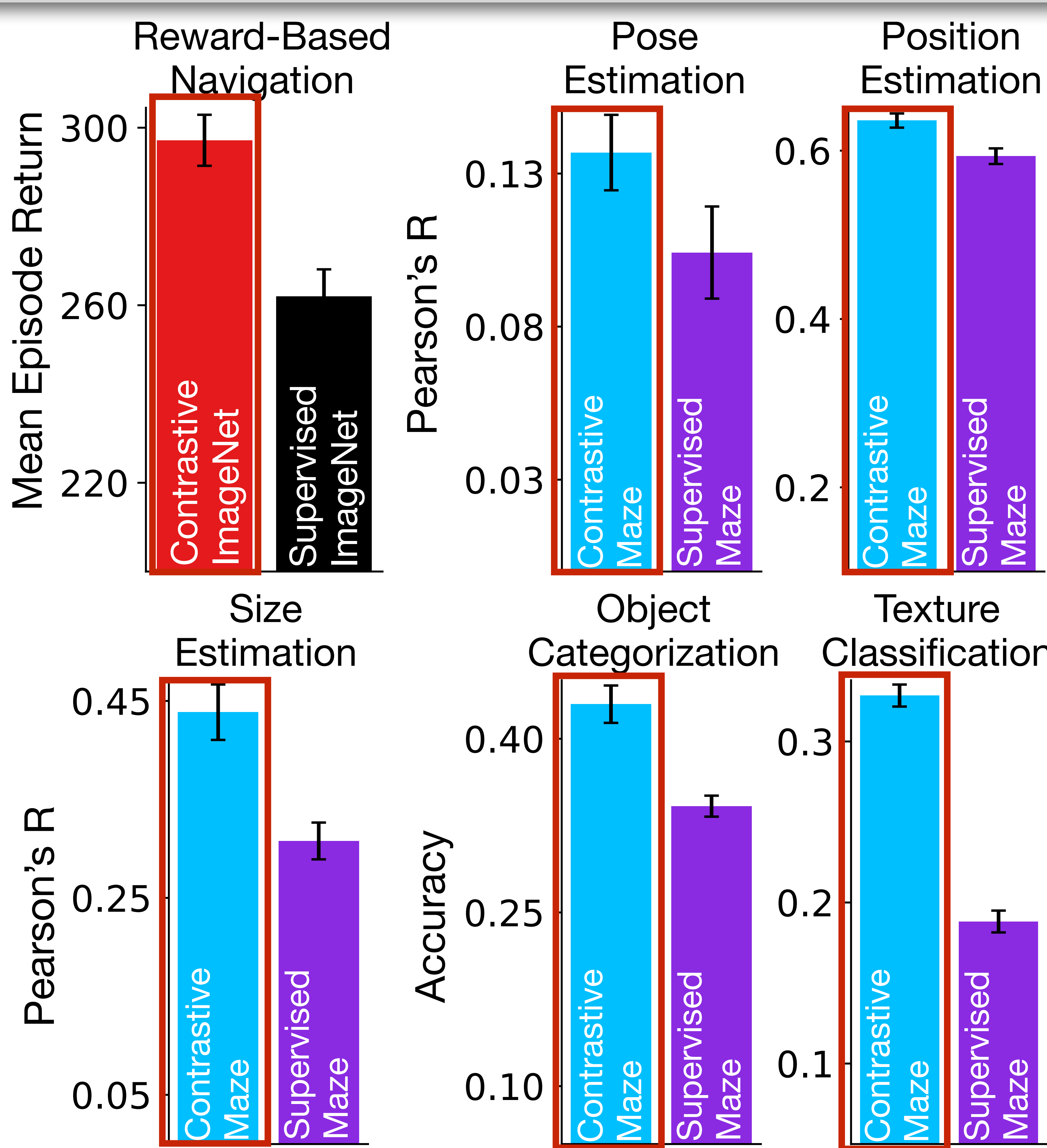
Contrastive Models Yield Better Transfer Performance



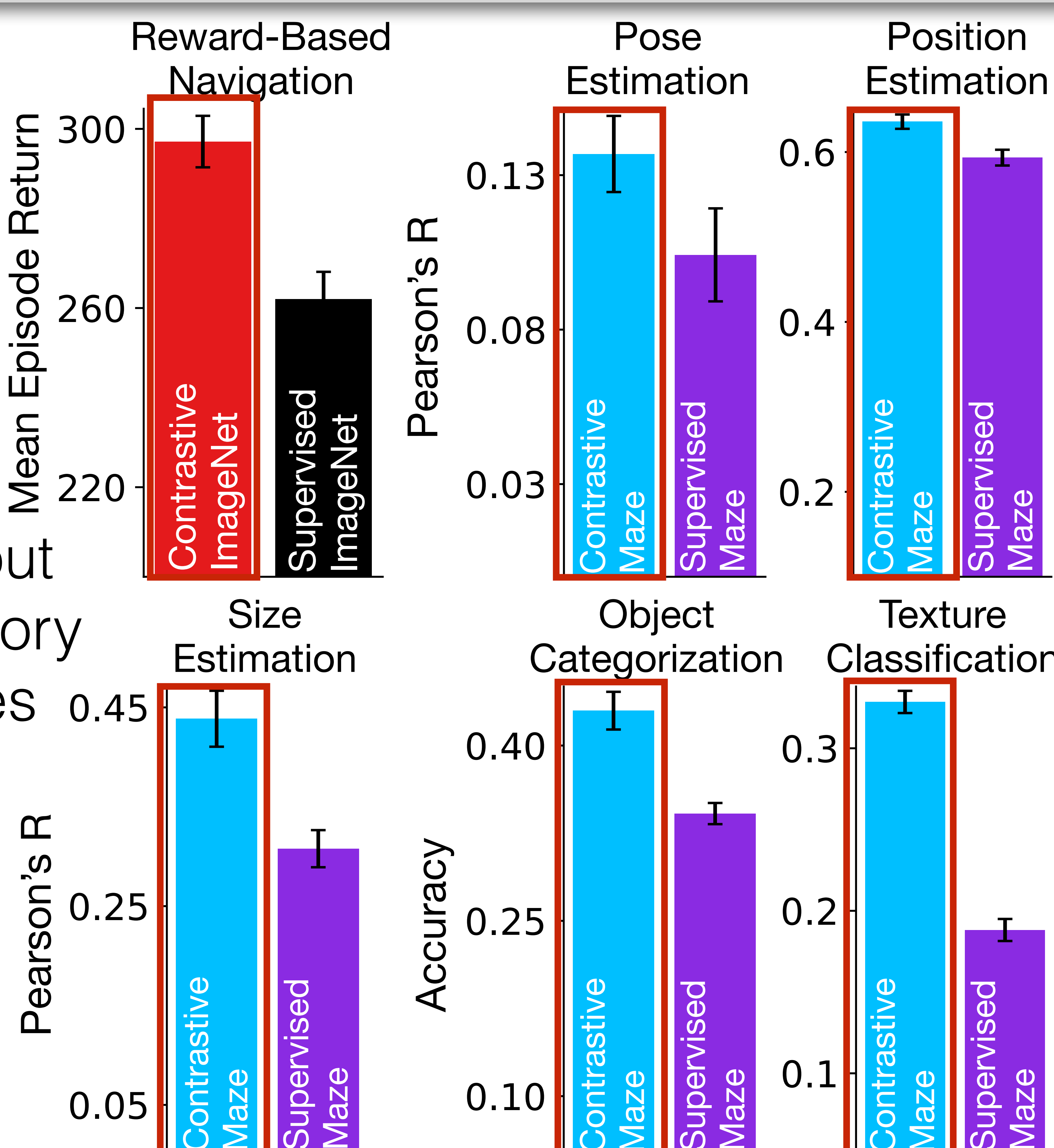
Contrastive Models Yield Better Transfer Performance



Contrastive Models Yield Better Transfer Performance



Contrastive Models Yield Better Transfer Performance



What about other sensory modalities beyond vision?

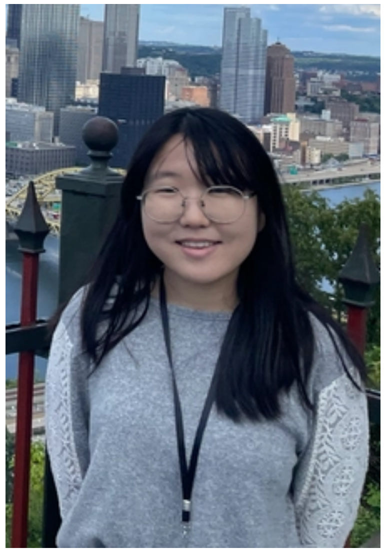
Tactile Processing

Task-Optimized Convolutional Recurrent Networks Align with Tactile Processing in the Rodent Brain

Trinity Chung[*,1], **Yuchen Shen**[*,2], **Nathan C. L. Kong**[4], **and Aran Nayebi**[2, 3, 1]

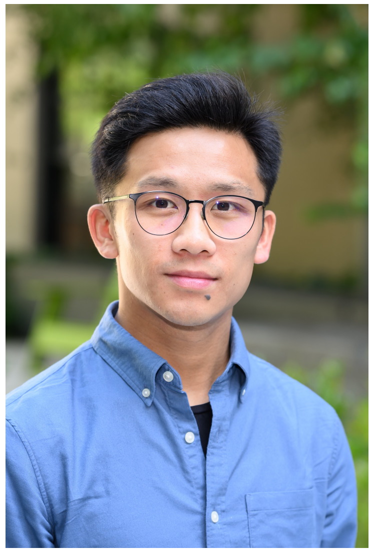
[1]Robotics Institute, Carnegie Mellon University; Pittsburgh, PA 15213
[2]Machine Learning Department, Carnegie Mellon University; Pittsburgh, PA 15213
[3]Neuroscience Institute, Carnegie Mellon University; Pittsburgh, PA 15213
[4]Department of Psychology, University of Pennsylvania; Philadelphia, PA 19104
[*] Equal contribution.
{trinityc, yuchens3, anayebi}@cs.cmu.edu; nclkong@sas.upenn.edu

To appear as a NeurIPS 2025 Oral!



Trinity Chung*

Yuchen Shen*



Nathan C.L. Kong

Why tactile?

Why tactile?

- Tactile data is very useful in manipulating occluded and OOD objects

Why tactile?

- Tactile data is very useful in manipulating occluded and OOD objects
- Tactile hardware & sim is getting better!

Why tactile?

- Tactile data is very useful in manipulating occluded and OOD objects
- Tactile hardware & sim is getting better!
- Tactile perception is still considerably under-explored in *both* neuroscience and robotics

Why tactile?

- Tactile data is very useful in manipulating occluded and OOD objects
- Tactile hardware & sim is getting better!
- Tactile perception is still considerably under-explored in *both* neuroscience and robotics

Trinity's search on arxiv...

arXiv

# of tactile 625 results

Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=tactile; OR all=somatosensory; OR abstract=touch; NOT abstract=haptic

# of vision 2,577 results

Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=vision; AND title=visual

(both in the last 12 months)

Why tactile?

- Tactile data is very useful in manipulating occluded and OOD objects
- Tactile hardware & sim is getting better!
- Tactile perception is still considerably under-explored in *both* neuroscience and robotics
- Many current tactile models are vision-based instead of force/torque-based

Trinity's search on arxiv...

arXiv

# of tactile 625 results

Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=tactile; OR all=somatosensory; OR abstract=touch; NOT abstract=haptic

# of vision 2,577 results

Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=vision; AND title=visual

(both in the last 12 months)

Why tactile?

- Tactile data is very useful in manipulating occluded and OOD objects
- Tactile hardware & sim is getting better!
- Tactile perception is still considerably under-explored in *both* neuroscience and robotics
- Many current tactile models are vision-based instead of force/torque-based

Trinity's search on arxiv...

arXiv

# of tactile 625 results

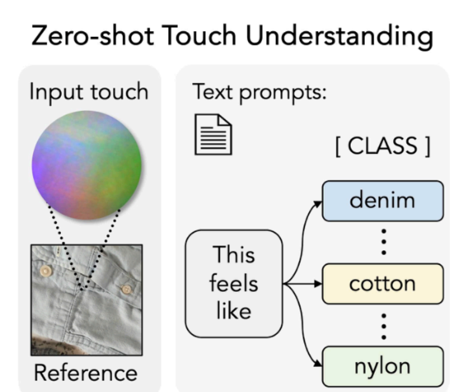
Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=tactile; OR all=somatosensory; OR abstract=touch; NOT abstract=haptic

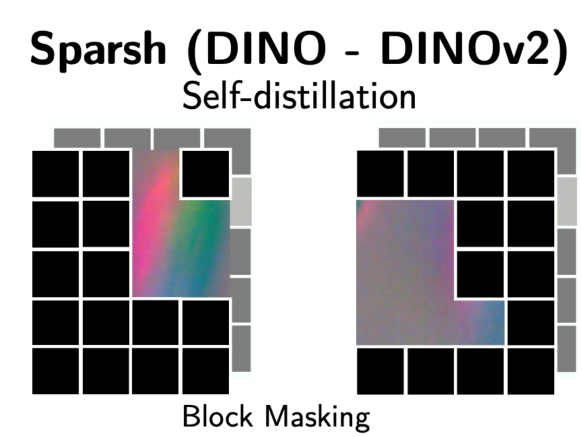
# of vision 2,577 results

Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=vision; AND title=visual

(both in the last 12 months)

e.g. UniTouch & Sparsh is trained on vision-based tactile sensors like Gelsight and DIGIT





https://arxiv.org/abs/2305.00596 https://arxiv.org/abs/2410.24090

Why tactile?

- Tactile data is very useful in manipulating occluded and OOD objects
- Tactile hardware & sim is getting better!
- Tactile perception is still considerably under-explored in *both* neuroscience and robotics
- Many current tactile models are vision-based instead of force/torque-based

We hypothesize that model architectures that mimics brain-like processing will yield better performance for tactile data.

Trinity's search on arxiv...

arXiv

# of tactile 625 results

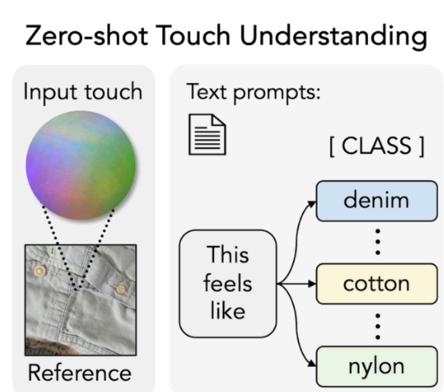
Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=tactile; OR all=somatosensory; OR abstract=touch; NOT abstract=haptic

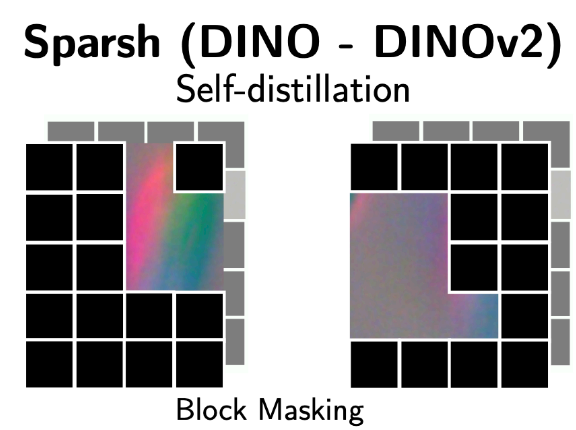
# of vision 2,577 results

Query: order: -announced_date_first; size: 50; date_range: from 2024-06-01 to 2025-06-02; classification: Computer Science (cs), Quantitative Biology (q-bio); include_cross_list: True; terms: AND all=vision; AND title=visual

(both in the last 12 months)

e.g. UniTouch & Sparsh is trained on vision-based tactile sensors like Gelsight and DIGIT

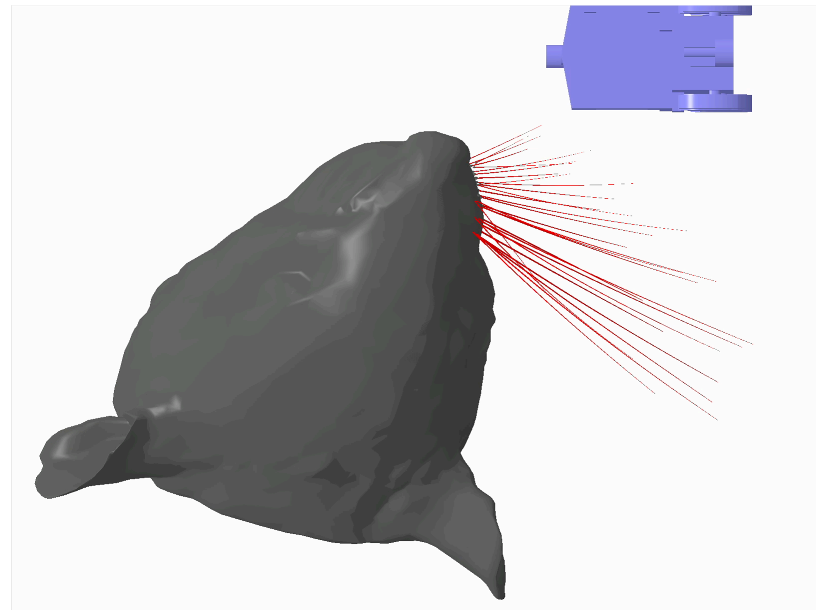


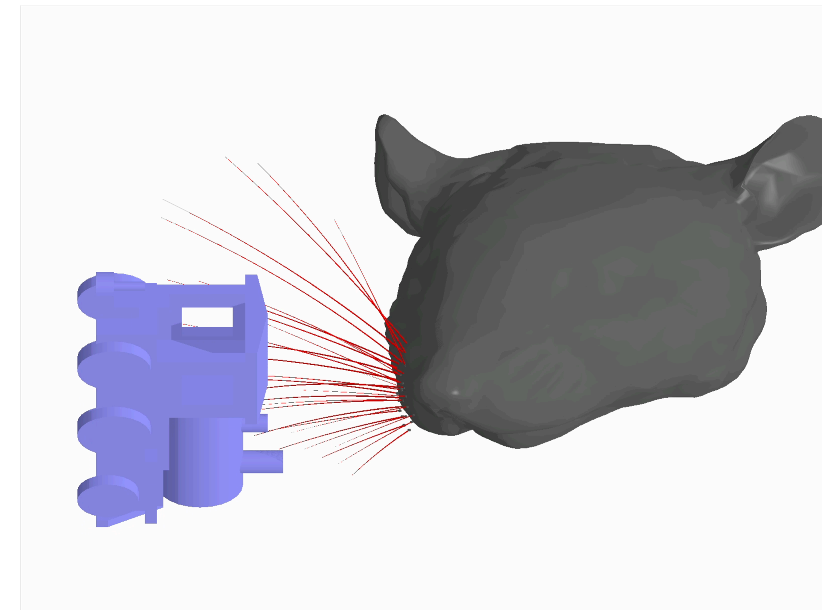


https://arxiv.org/abs/2305.00596 https://arxiv.org/abs/2410.24090

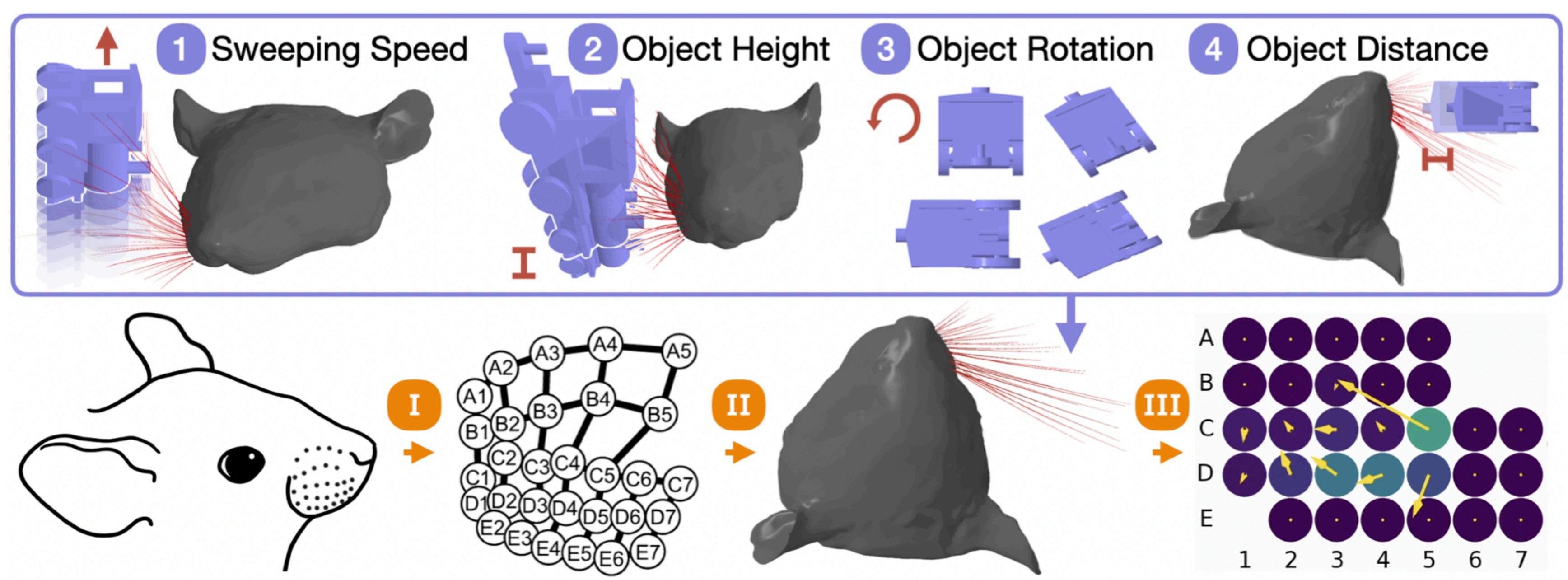
Training Data: Whisking Dataset Generation

- Objects are whisked in simulation using WHISKiT [Zweifel et al., 2021], simulator based on Bullet Physics



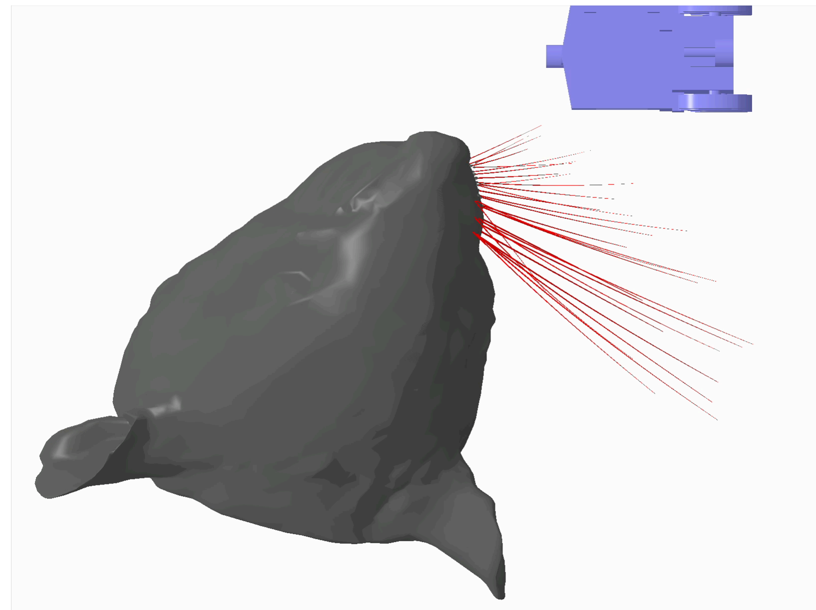


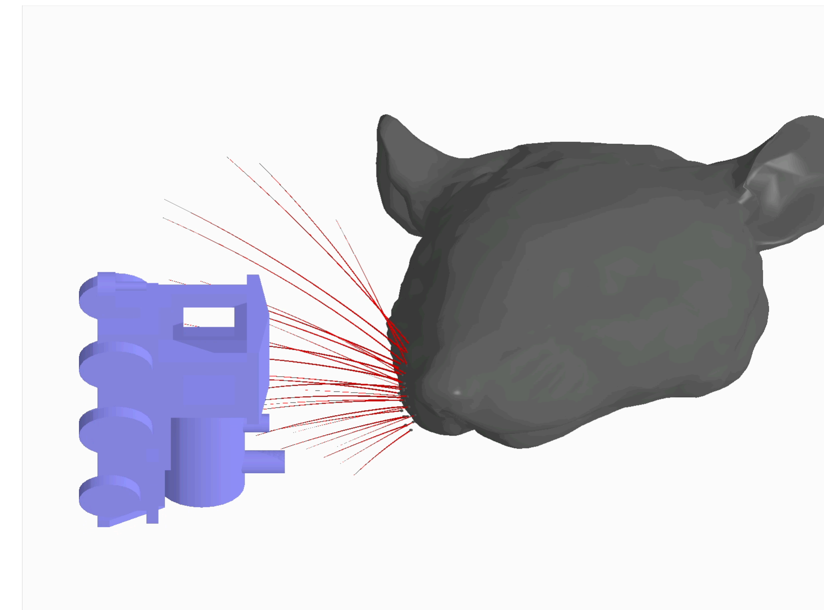
- 6-axis force/torque data for sweeping 9981 ShapeNet objects of 117 categories with various sweep augmentations



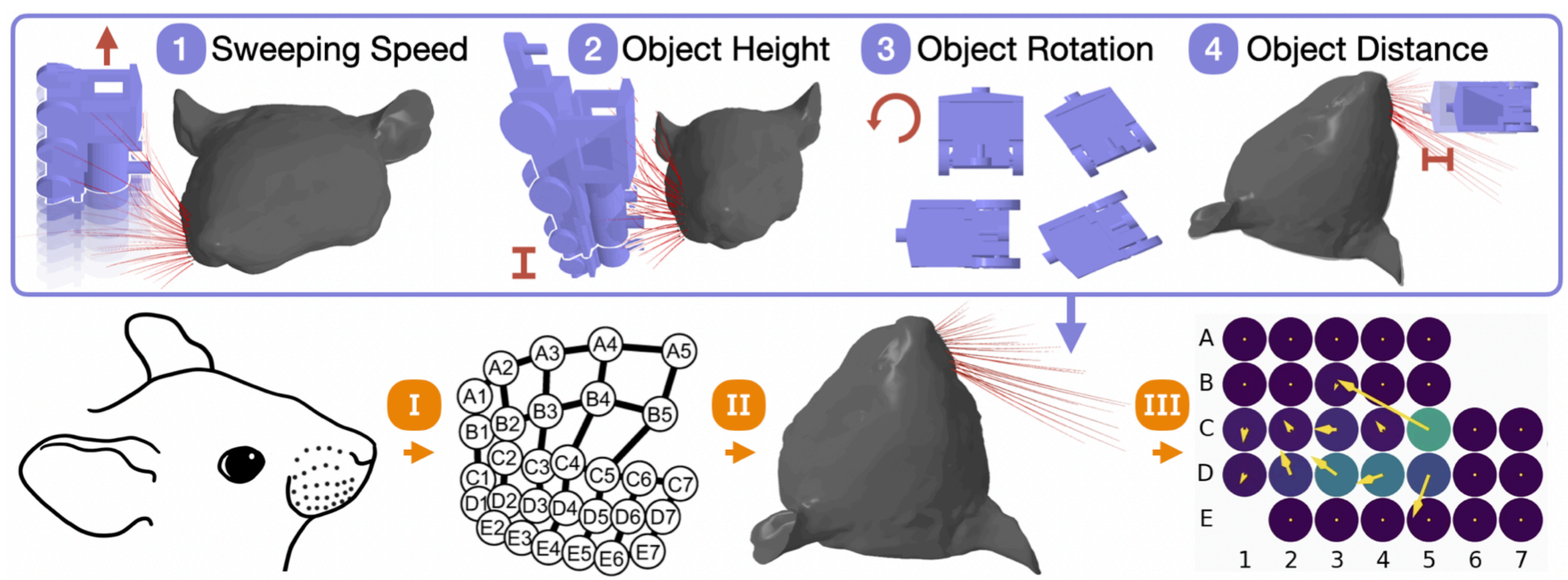
Training Data: Whisking Dataset Generation

- Objects are whisked in simulation using WHISKiT [Zweifel et al., 2021], simulator based on Bullet Physics



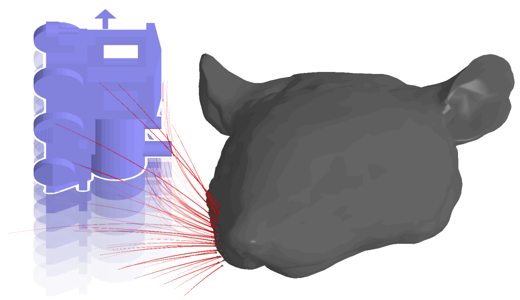


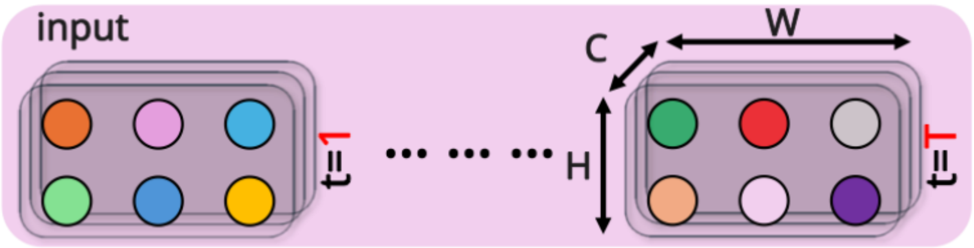
- 6-axis force/torque data for sweeping 9981 ShapeNet objects of 117 categories with various sweep augmentations



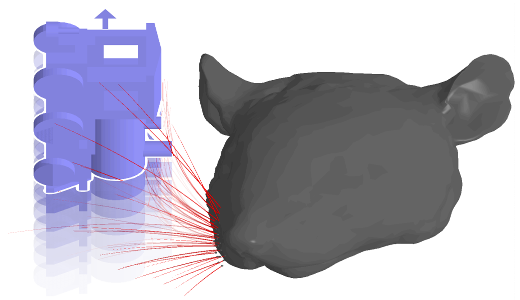
Training Data: Tactile vs Image Augmentation

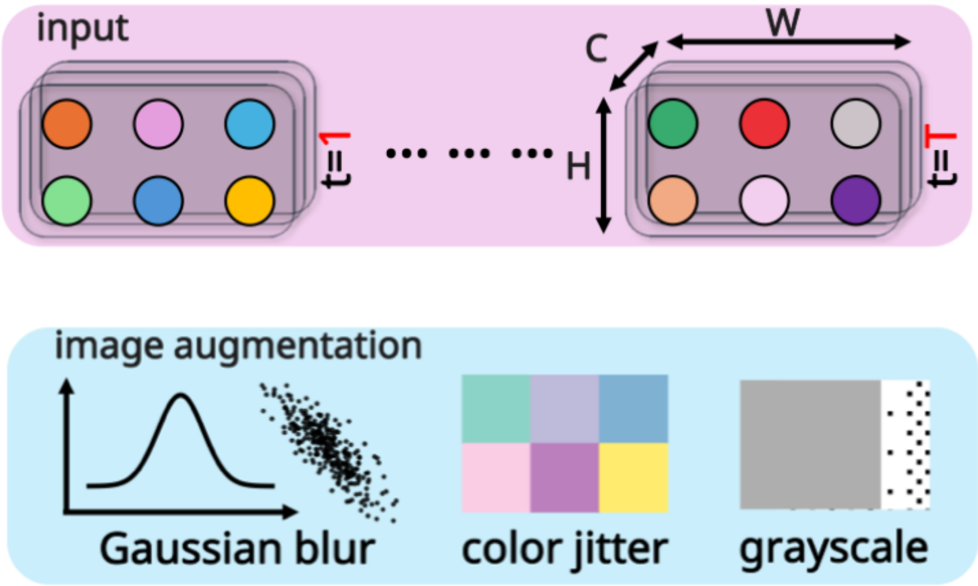
Training Data: Tactile vs Image Augmentation





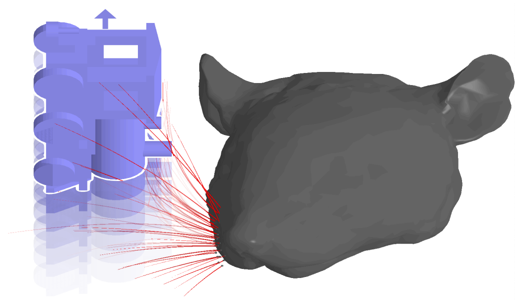
Training Data: Tactile vs Image Augmentation



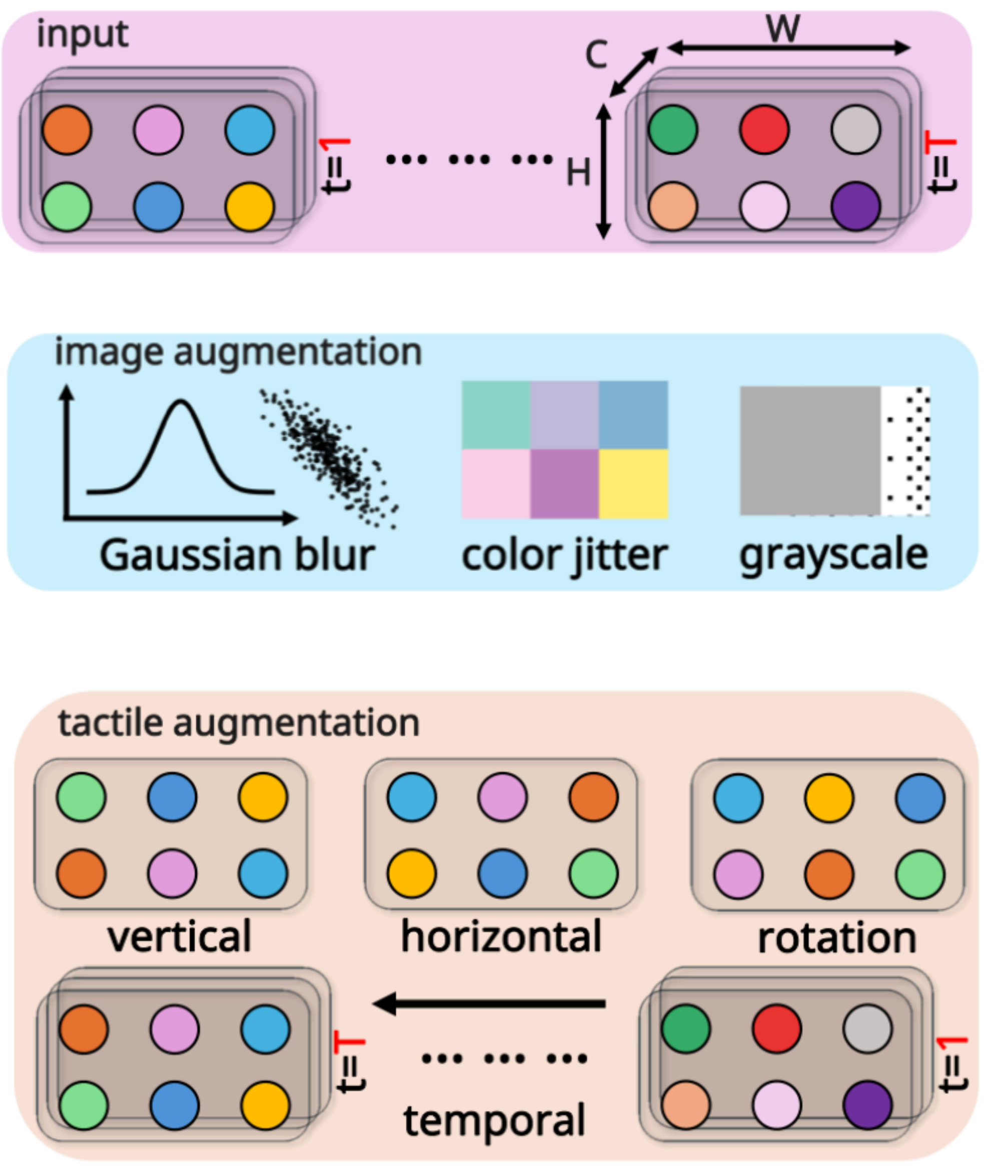


- Traditional image augmentation introduces Gaussian noise, color jitter, and grayscale.

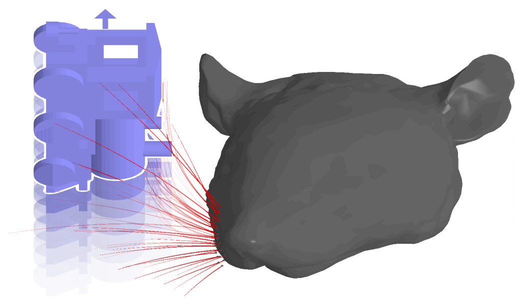
Training Data: Tactile vs Image Augmentation



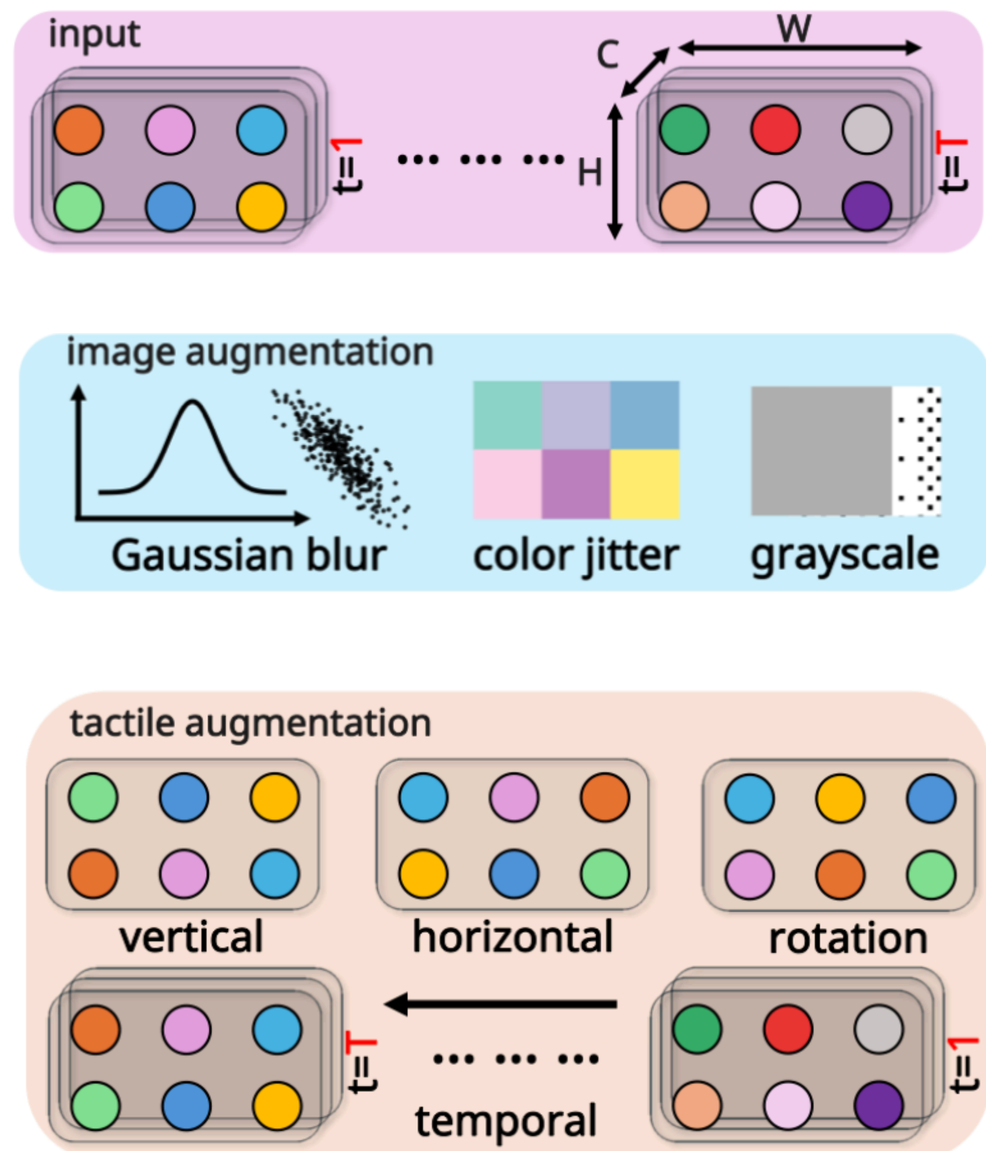
- Traditional image augmentation introduces Gaussian noise, color jitter, and grayscale.
- Our tactile augmentation vertically, horizontally, temporally flips, and rotates the features

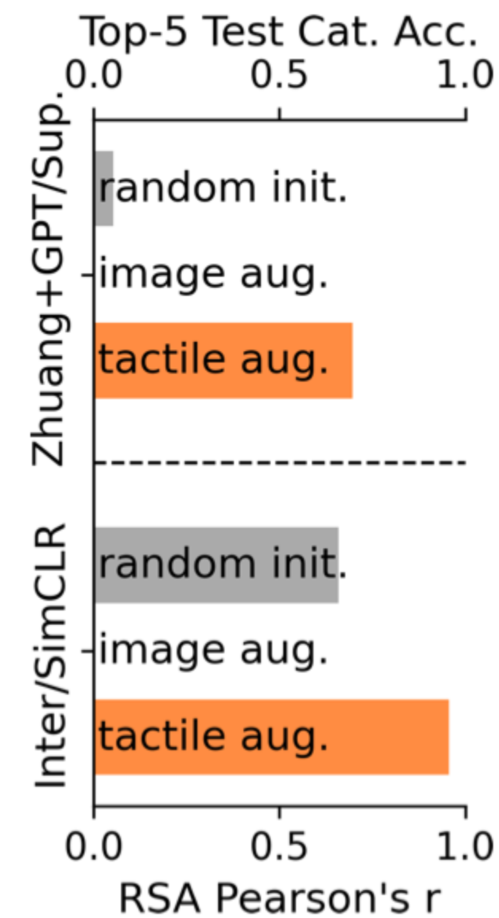


Training Data: Tactile vs Image Augmentation

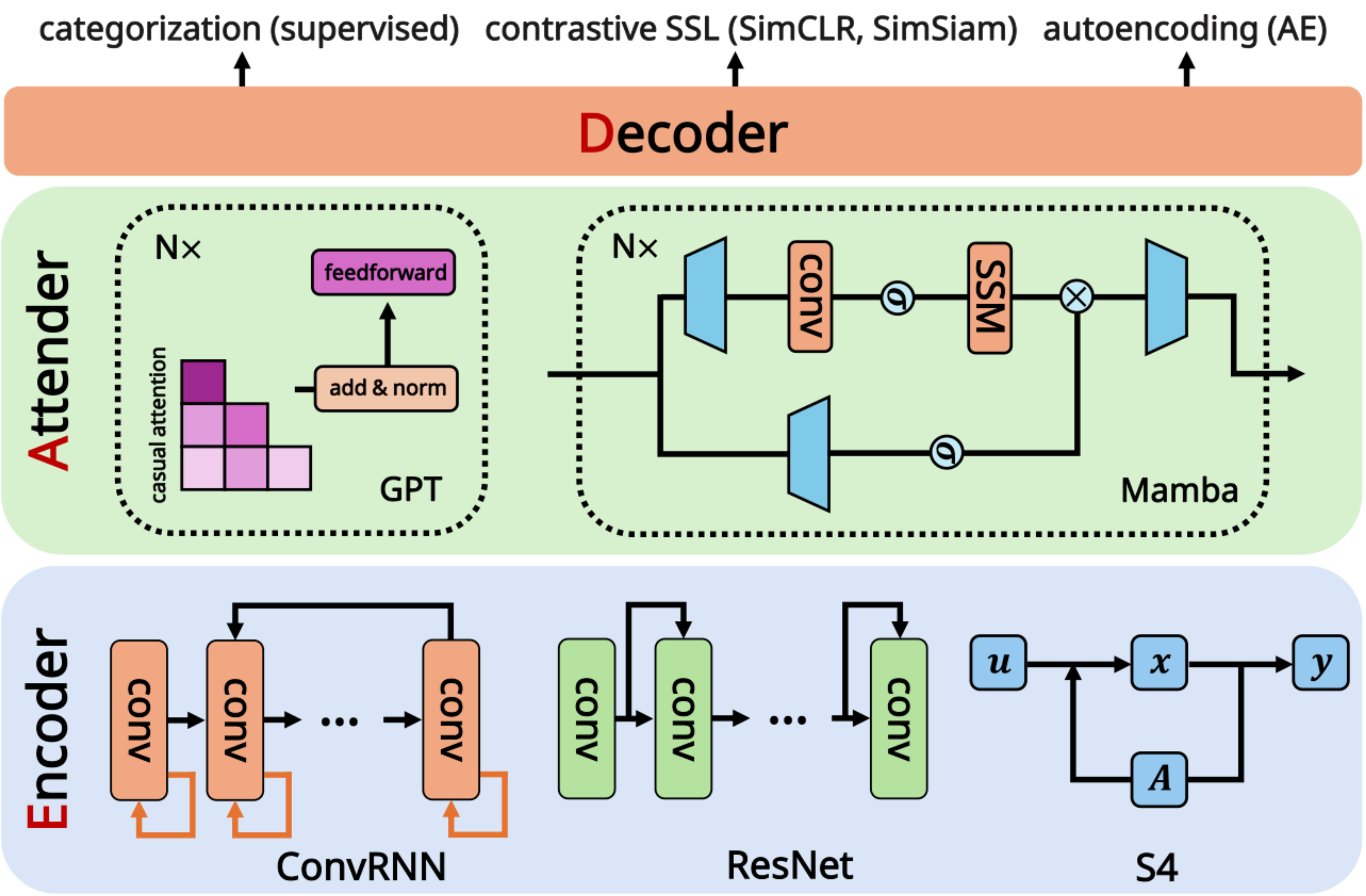


- Traditional image augmentation introduces Gaussian noise, color jitter, and grayscale.
- Our tactile augmentation vertically, horizontally, temporally flips, and rotates the features



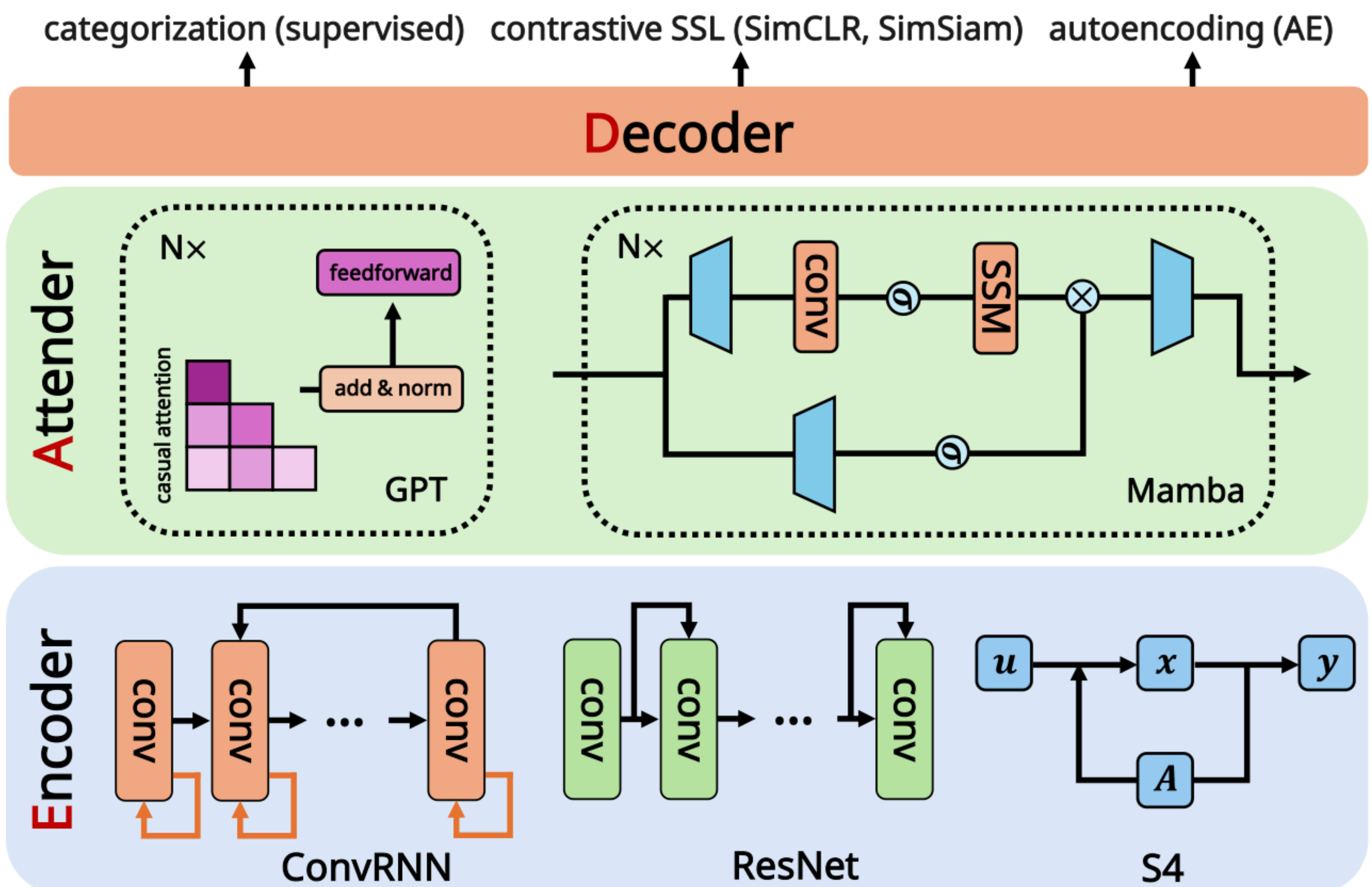


Models: Encoder-Attender-Decoder (EAD) Architecture



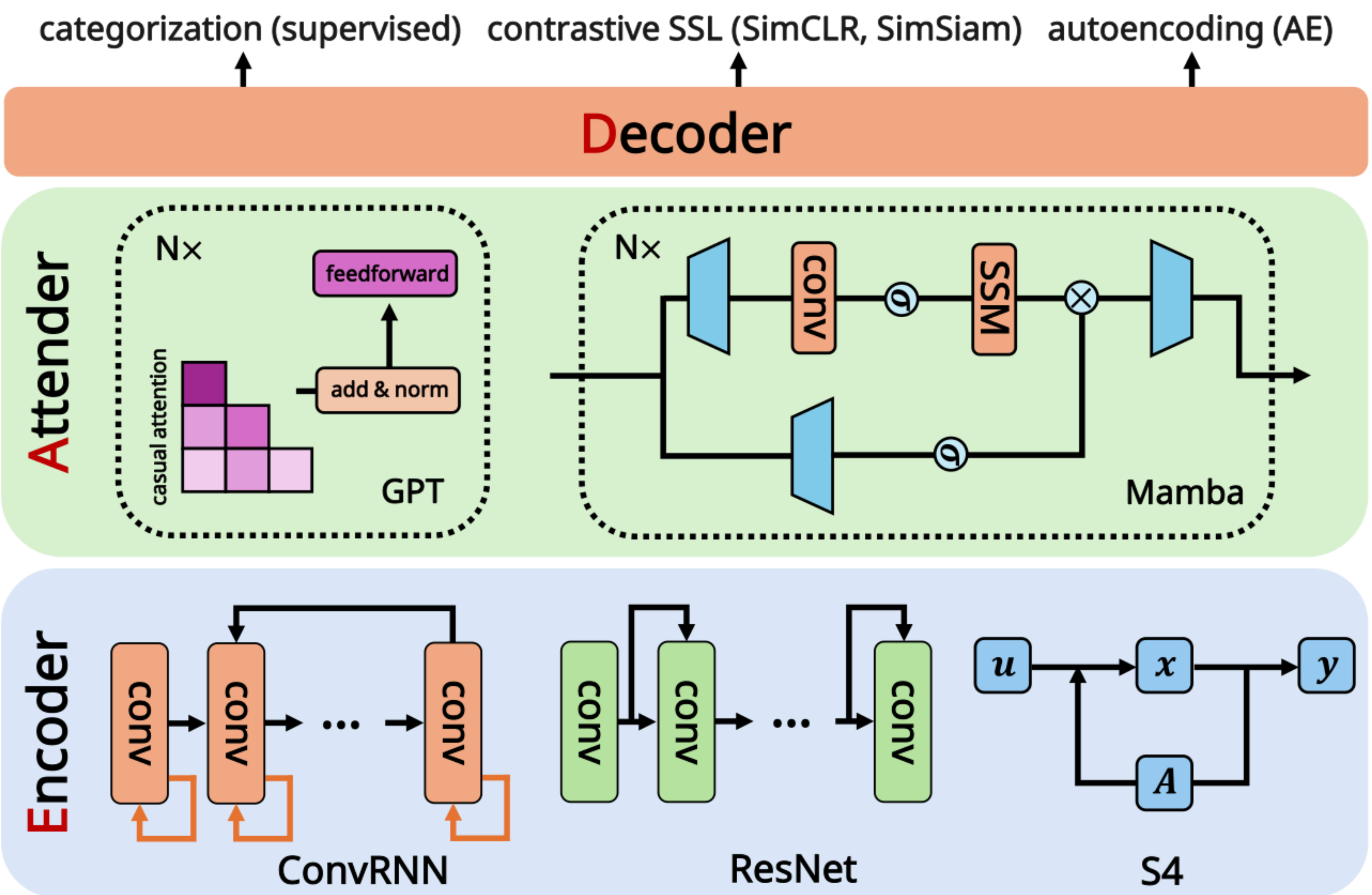
Models: Encoder-Attender-Decoder (EAD) Architecture

- We needed a way to systematically search over the space of recurrent model architectures



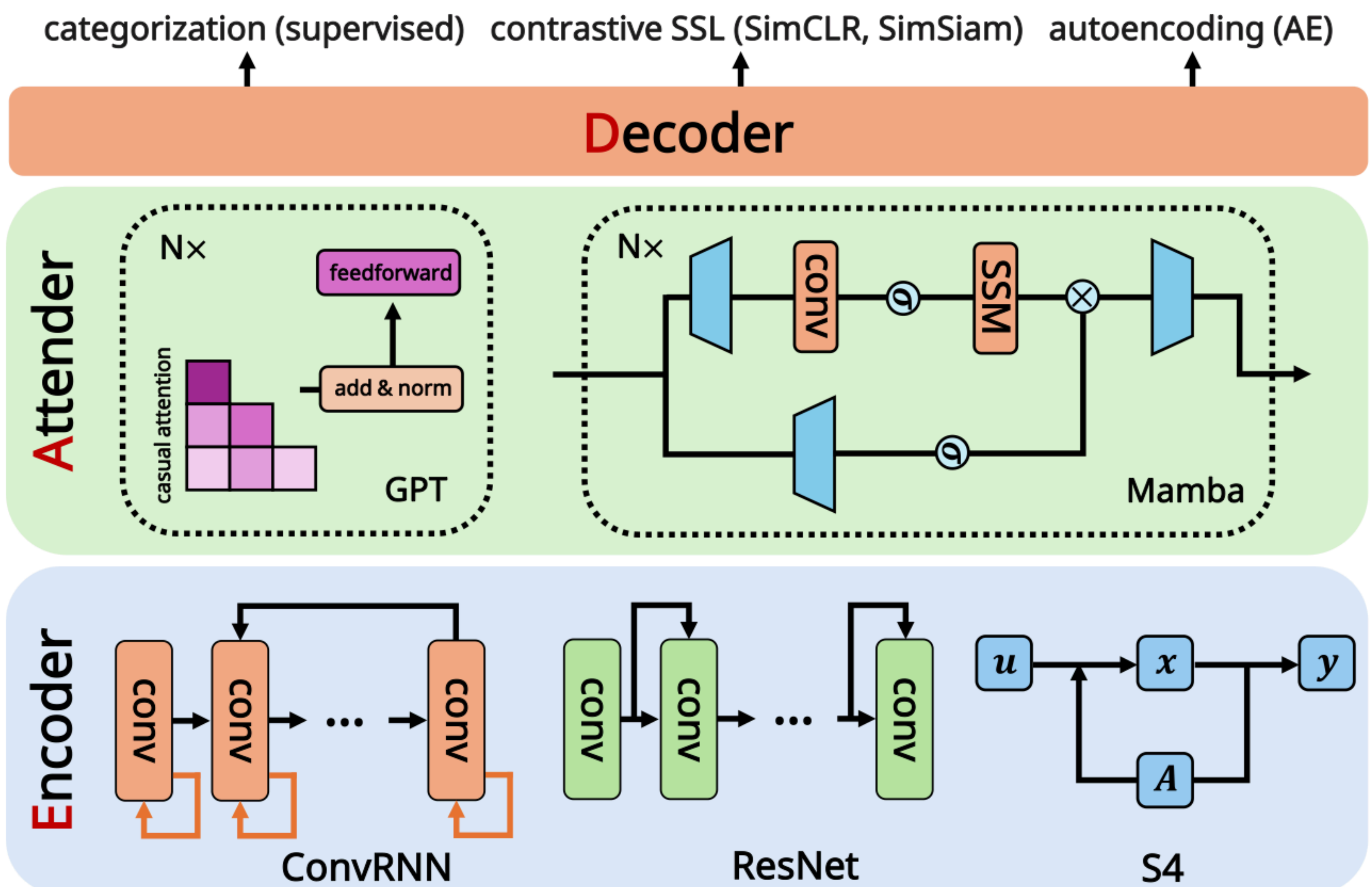
Models: Encoder-Attender-Decoder (EAD) Architecture

- We needed a way to systematically search over the space of recurrent model architectures
- EAD architecture allows us to easily construct new models by swapping out modules



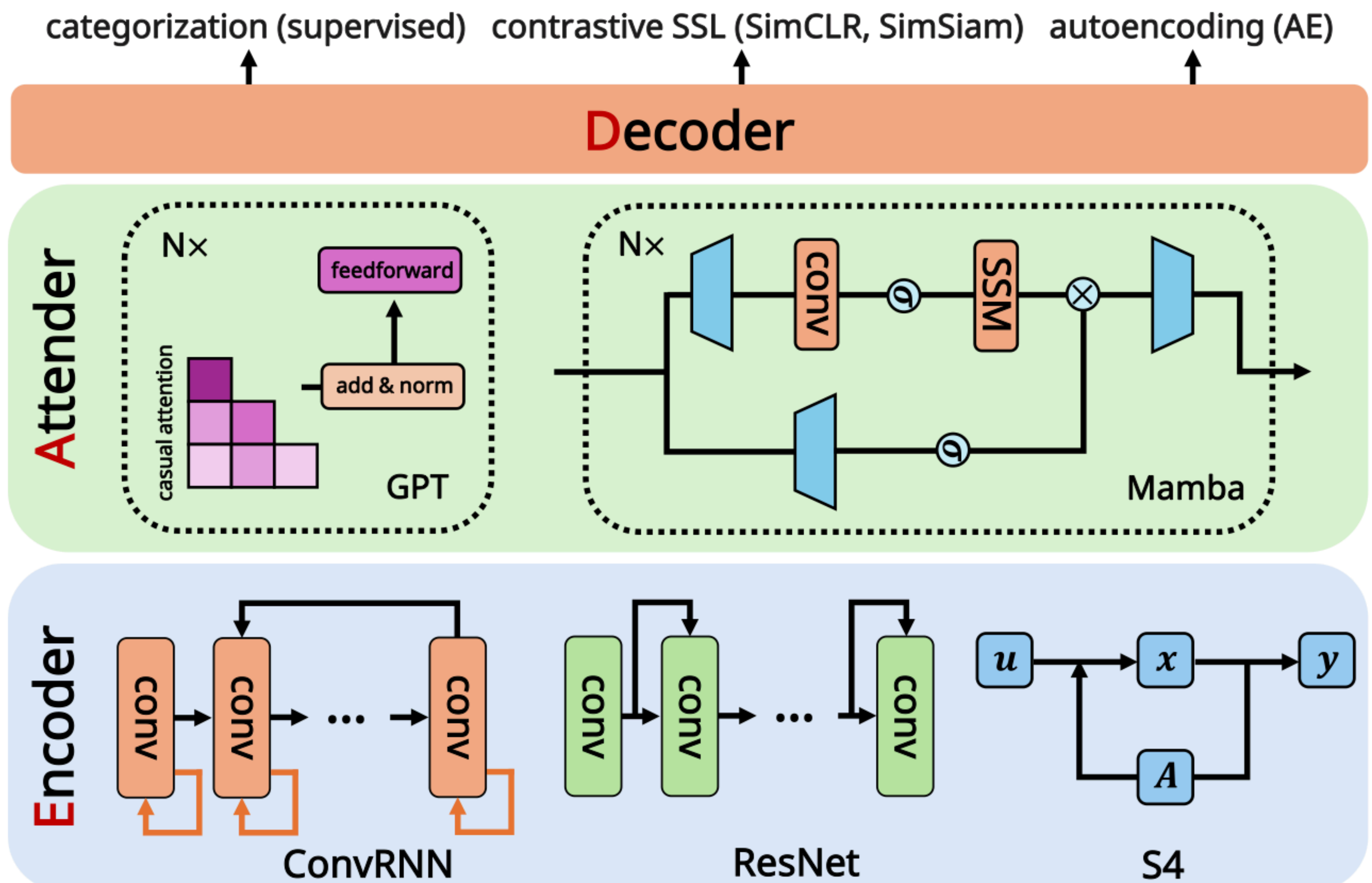
Models: Encoder-Attender-Decoder (EAD) Architecture

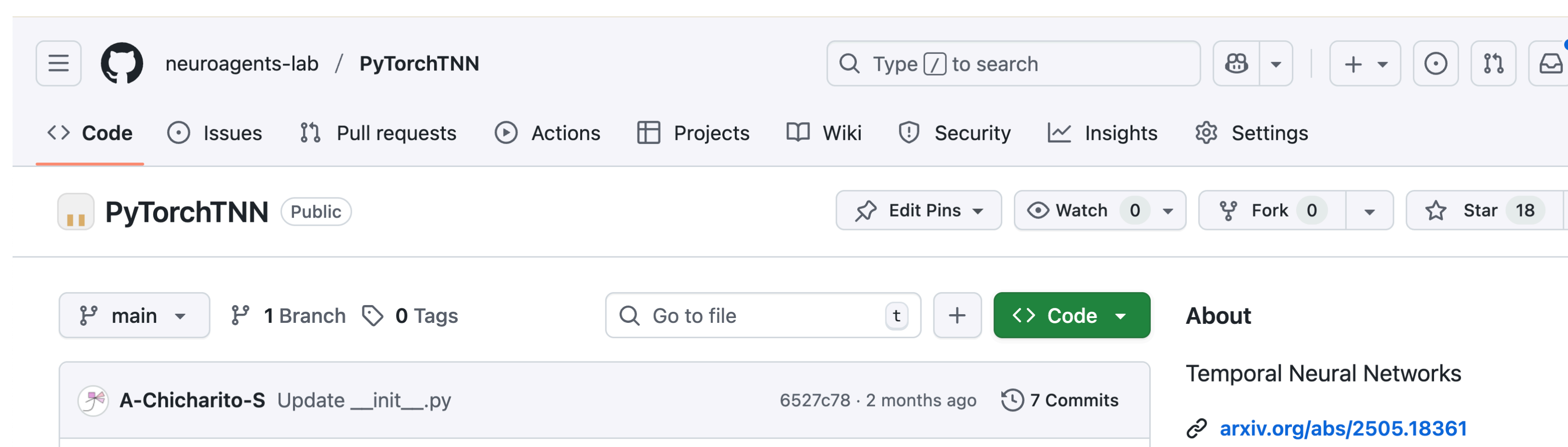
- We needed a way to systematically search over the space of recurrent model architectures
- EAD architecture allows us to easily construct new models by swapping out modules
- Built using PyTorchTNN, now on Github! https://github.com/neuroagents-lab/PyTorchTNN



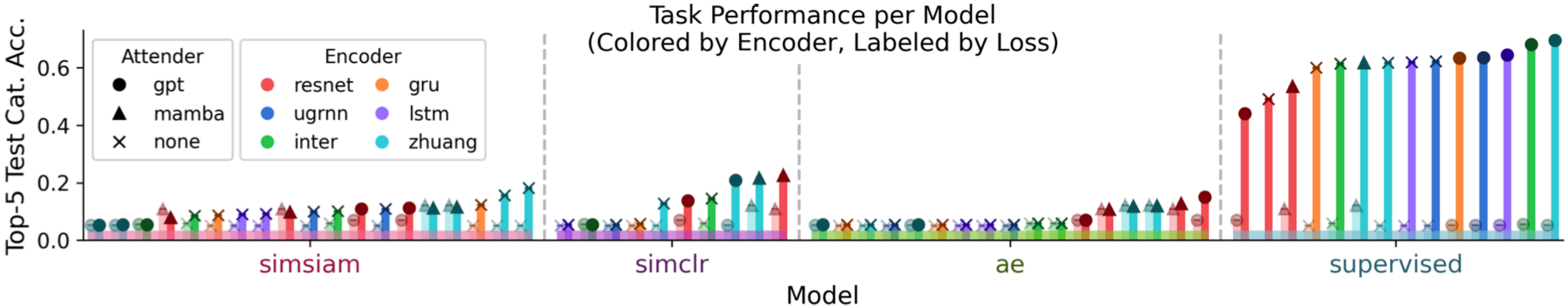
Models: Encoder-Attender-Decoder (EAD) Architecture

- We needed a way to systematically search over the space of recurrent model architectures
- EAD architecture allows us to easily construct new models by swapping out modules
- Built using PyTorchTNN, now on Github! https://github.com/neuroagents-lab/PyTorchTNN

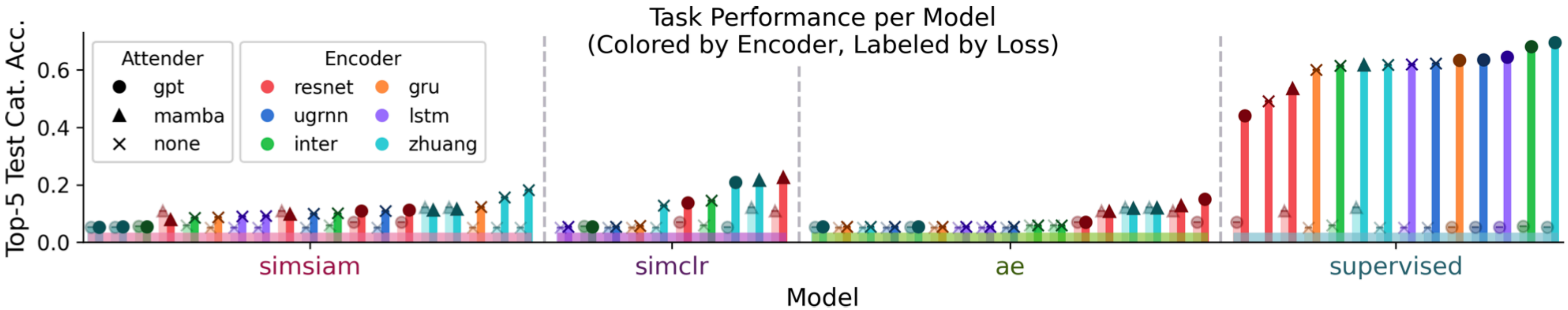




Results: ConvRNN encoders perform best

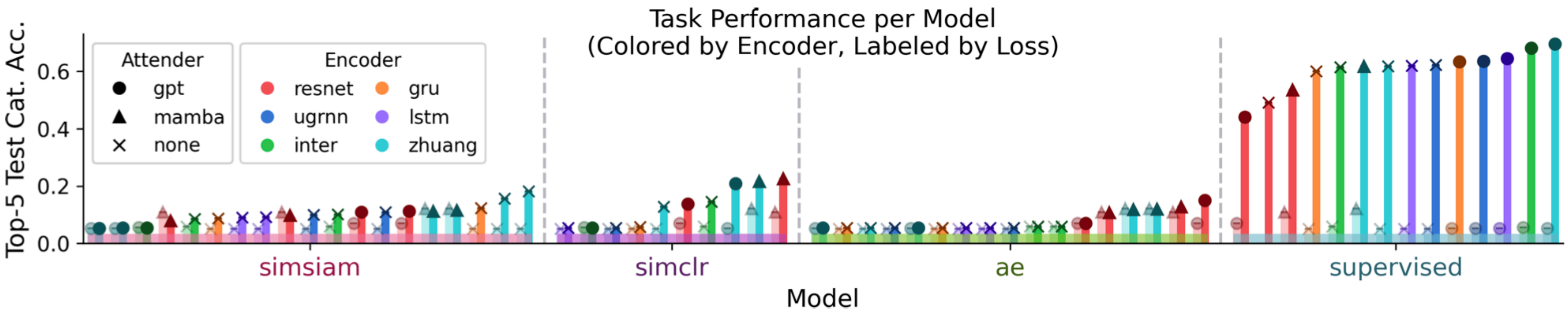


Results: ConvRNN encoders perform best



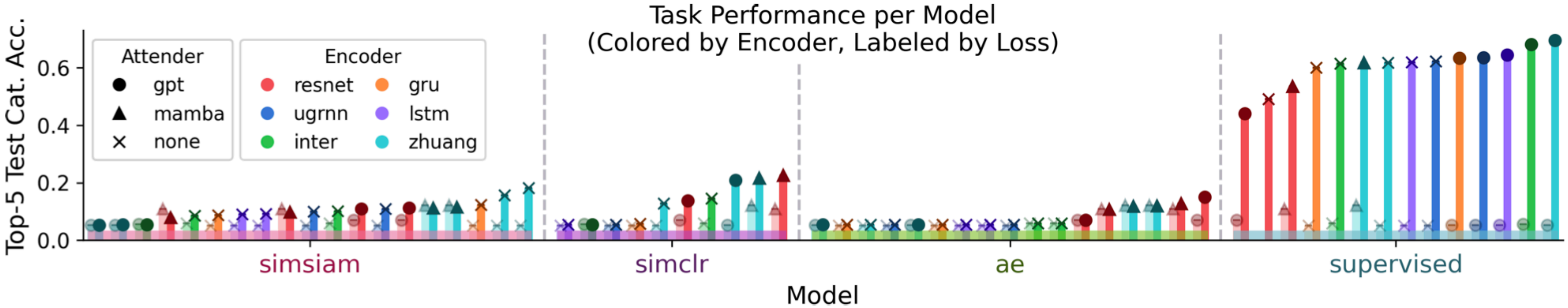
- Lighter color bar represents untrained version.

Results: ConvRNN encoders perform best



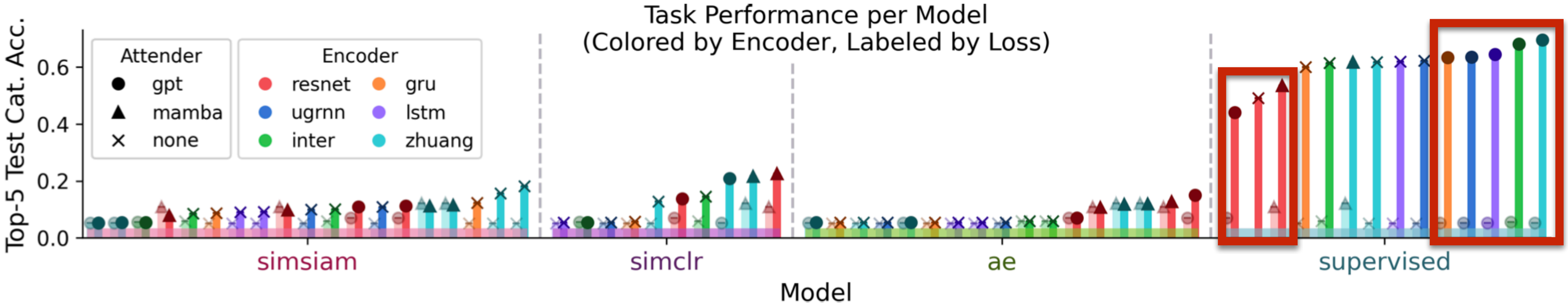
- Lighter color bar represents untrained version.
- S4 Encoders often don't even train on this task!

Results: ConvRNN encoders perform best



- Lighter color bar represents untrained version.
- S4 Encoders often don't even train on this task!
- Best model is ConvRNN (Encoder)+GPT (Attender) +Supervised (Decoder)

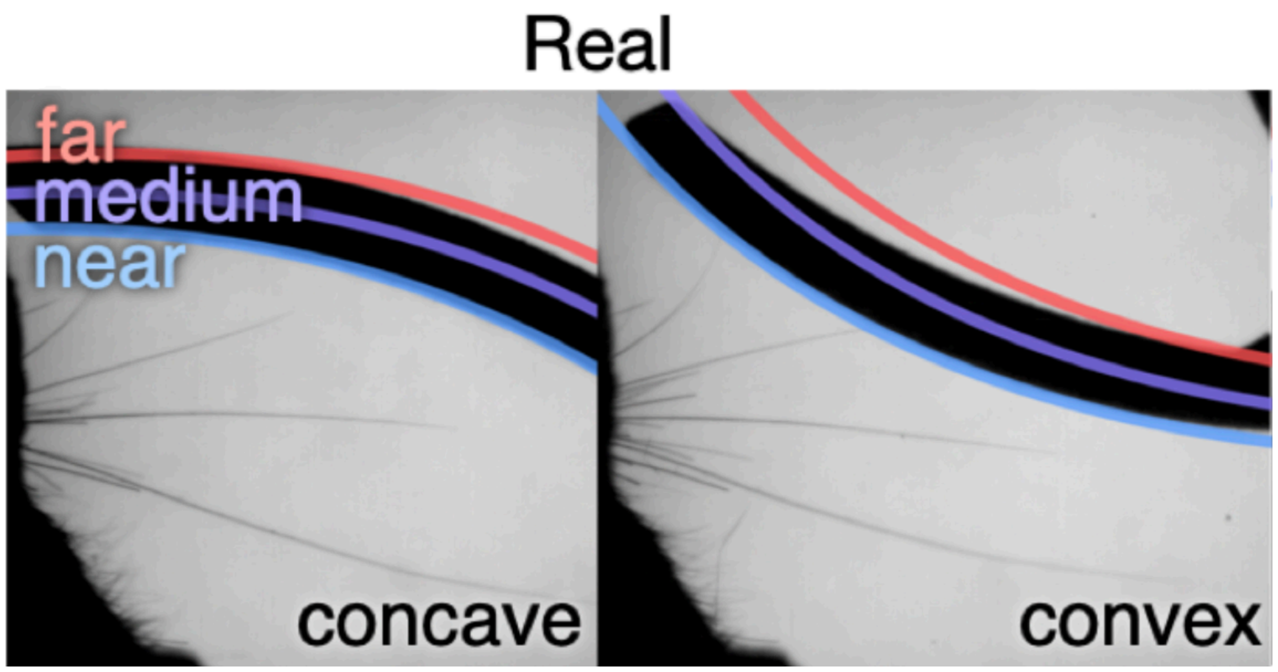
Results: ConvRNN encoders perform best



- Lighter color bar represents untrained version.
- S4 Encoders often don't even train on this task!
- Best model is ConvRNN (Encoder)+GPT (Attender) +Supervised (Decoder)

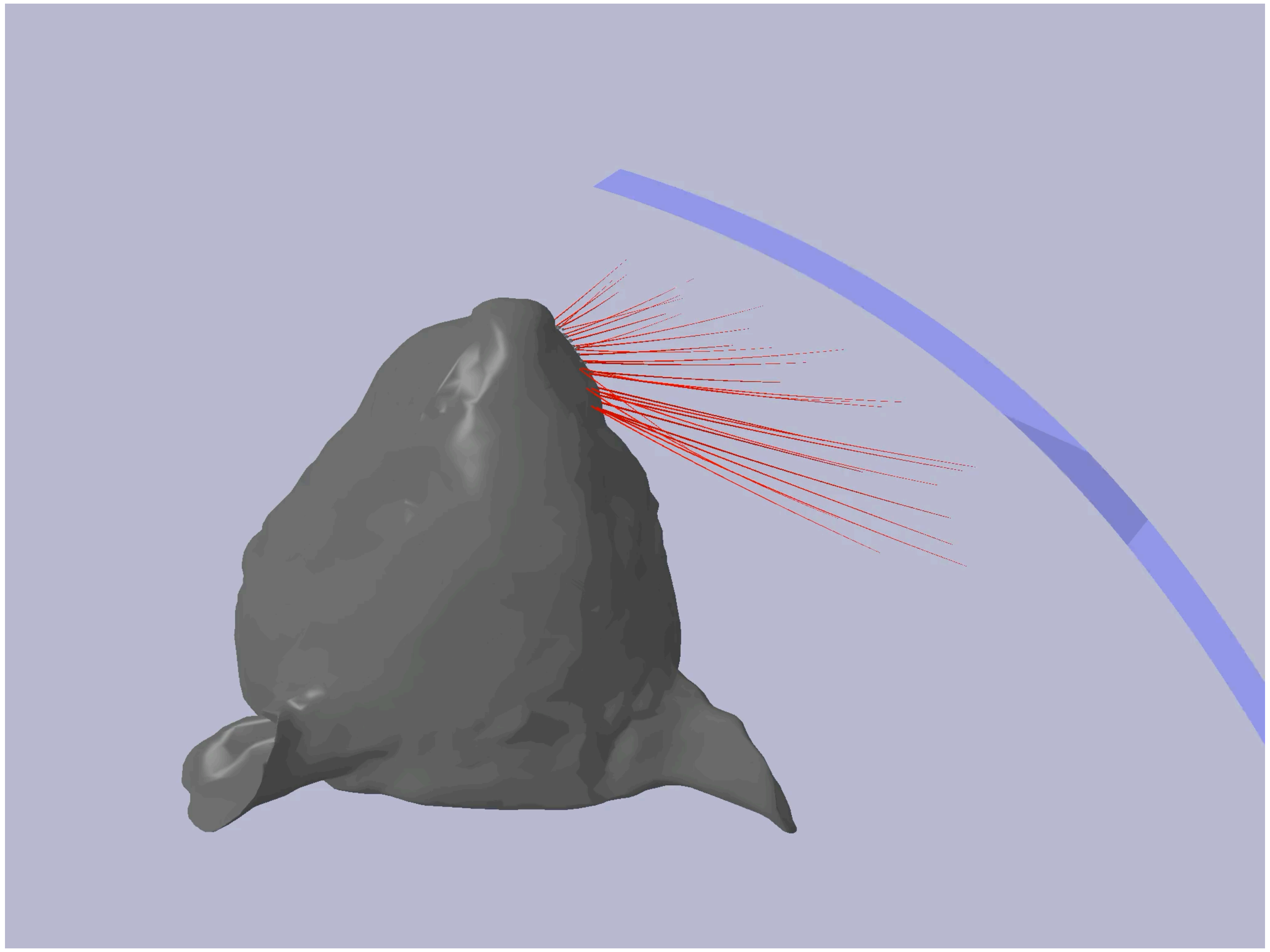
Neural Evaluation: Results

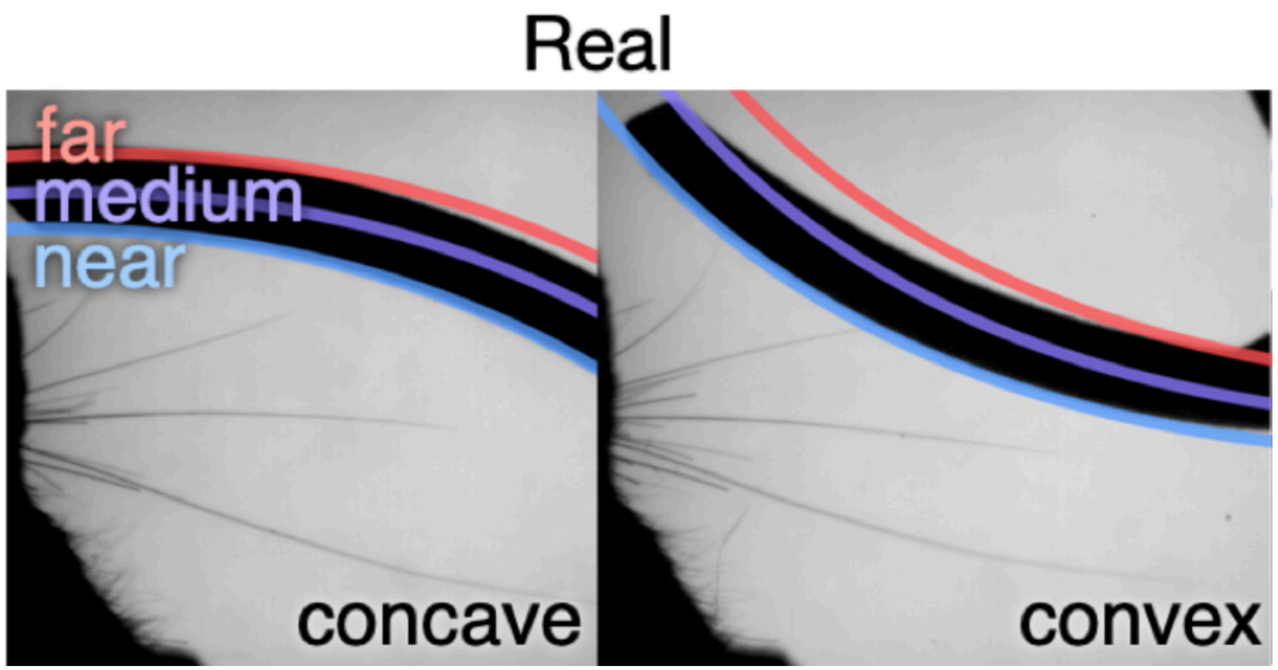
Neural Evaluation: Results



Rodgers 2022

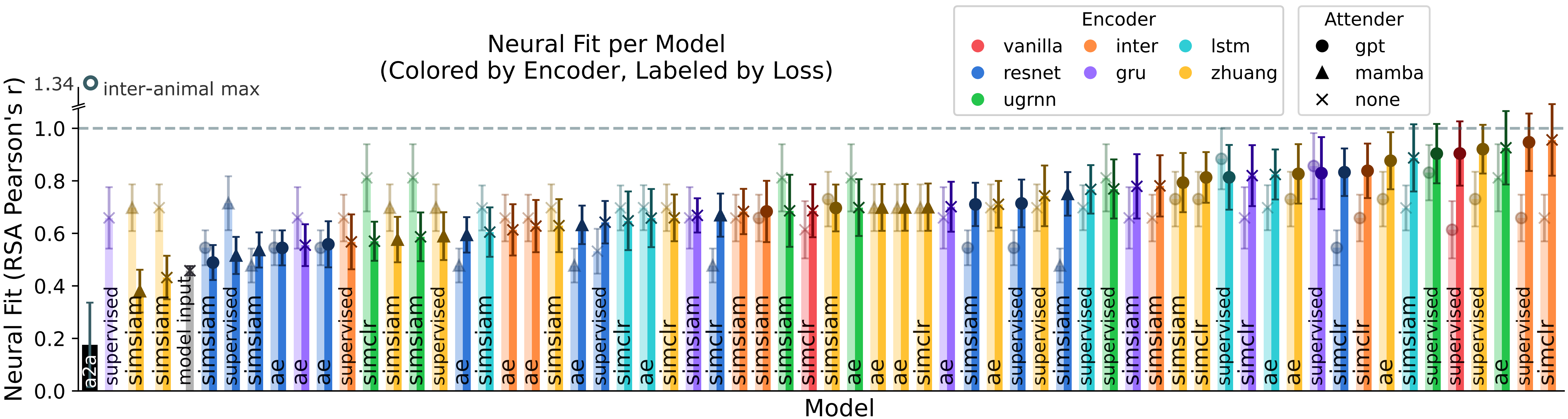
Neural Evaluation: Results

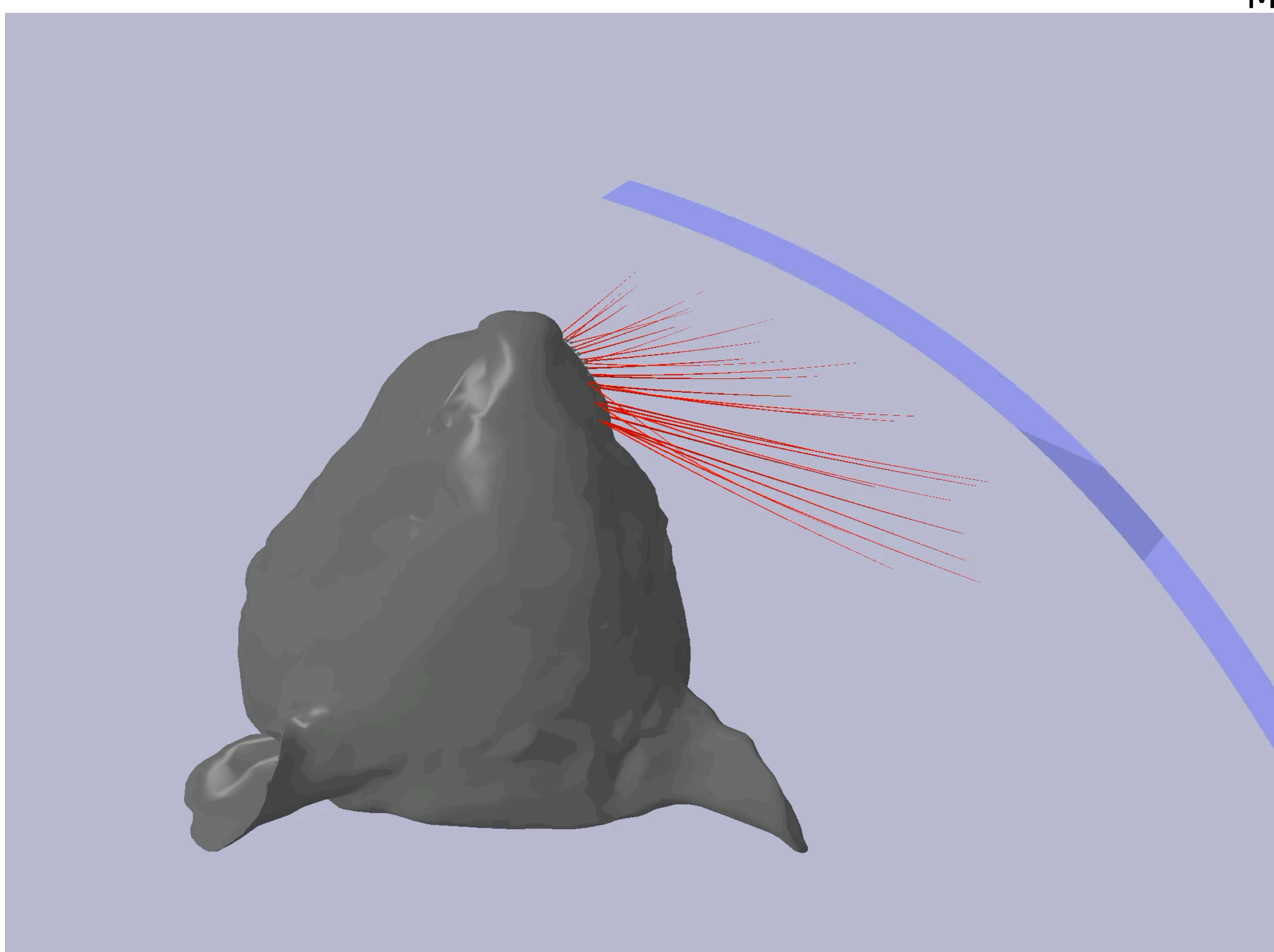


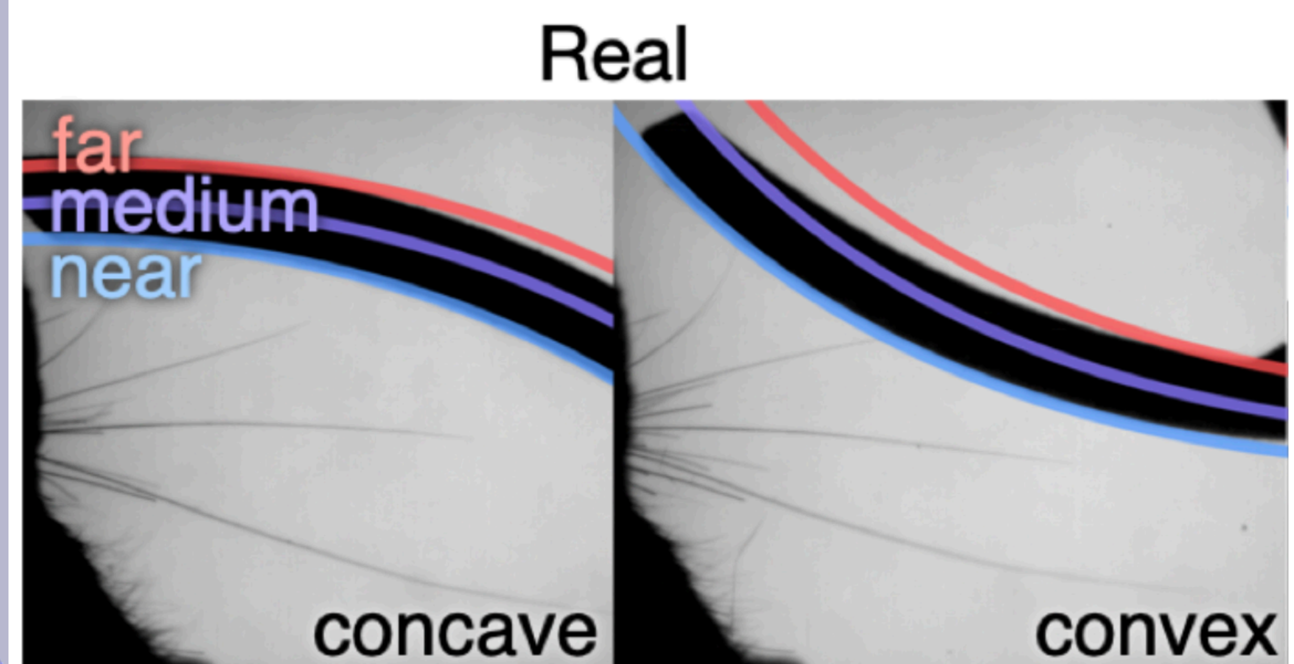


Rodgers 2022

Neural Evaluation: Results

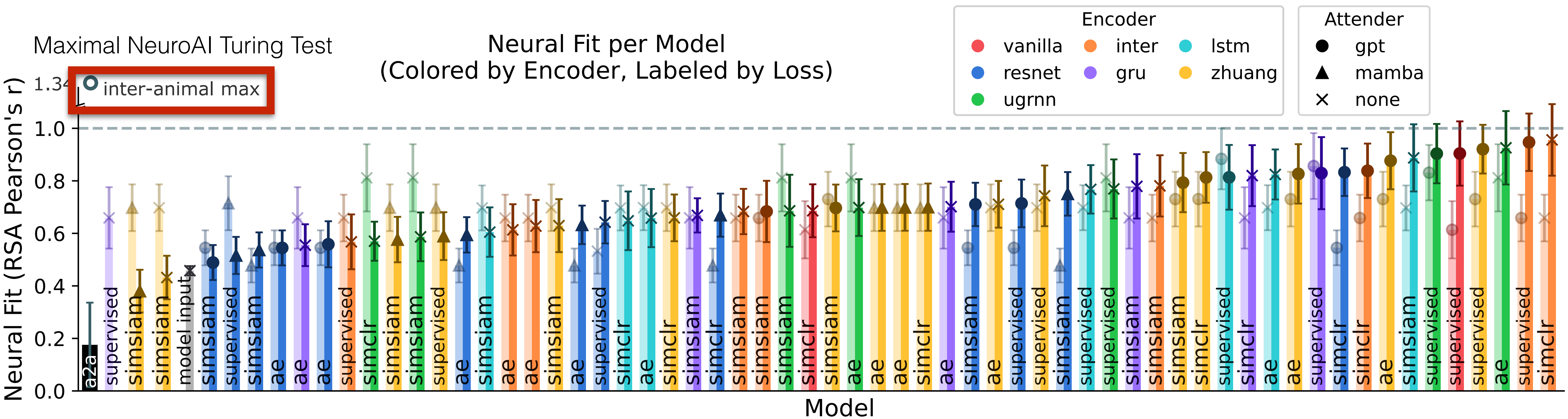


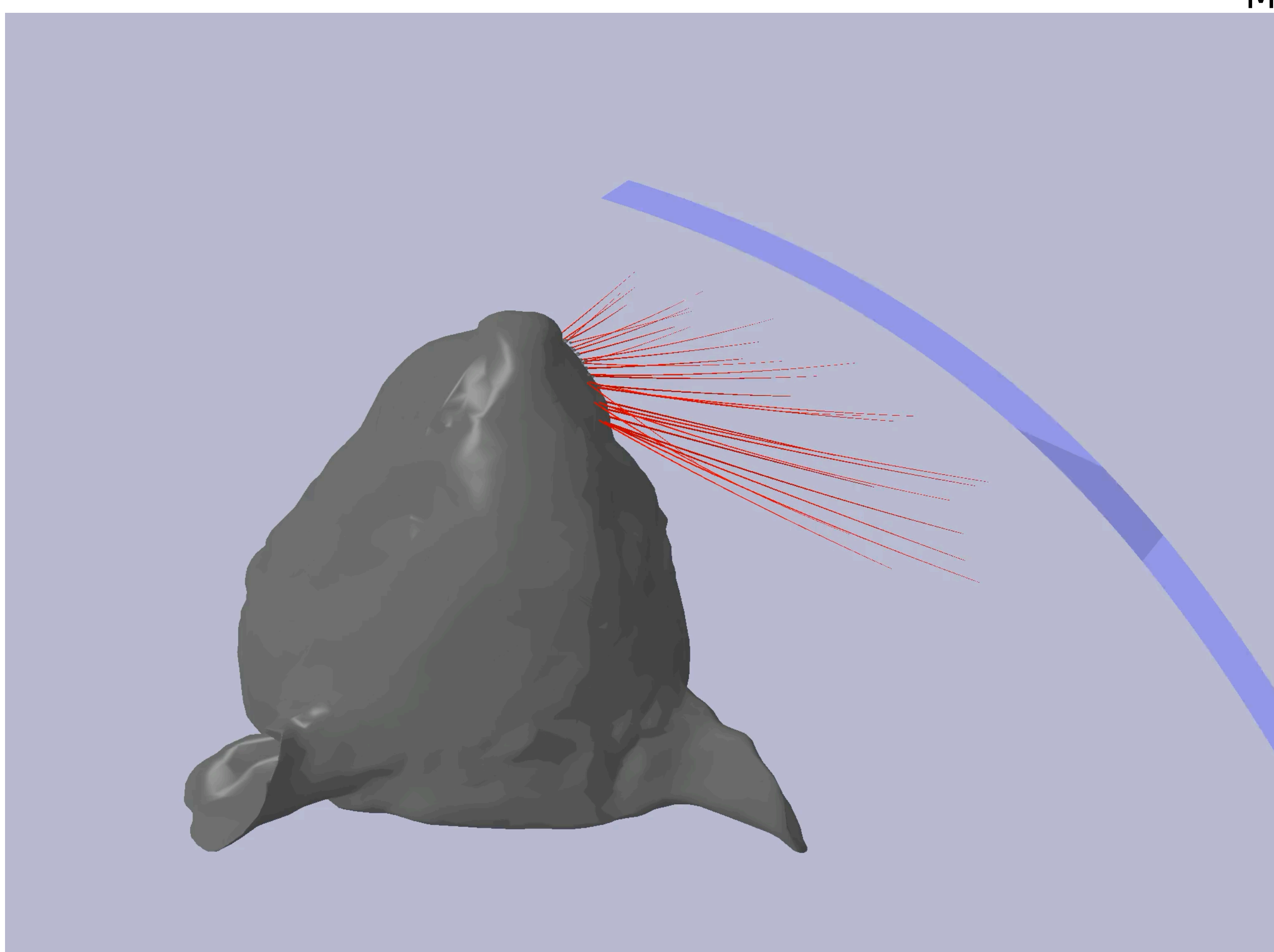


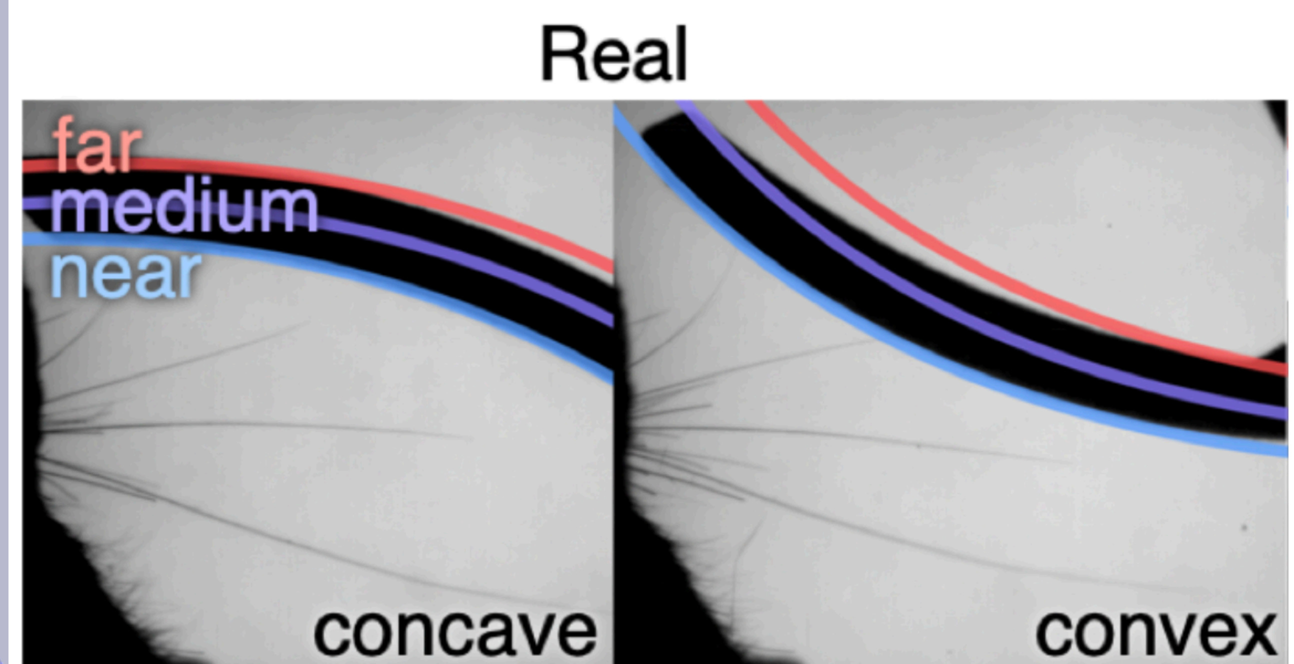


[Rodgers 2022]

Neural Evaluation: Results

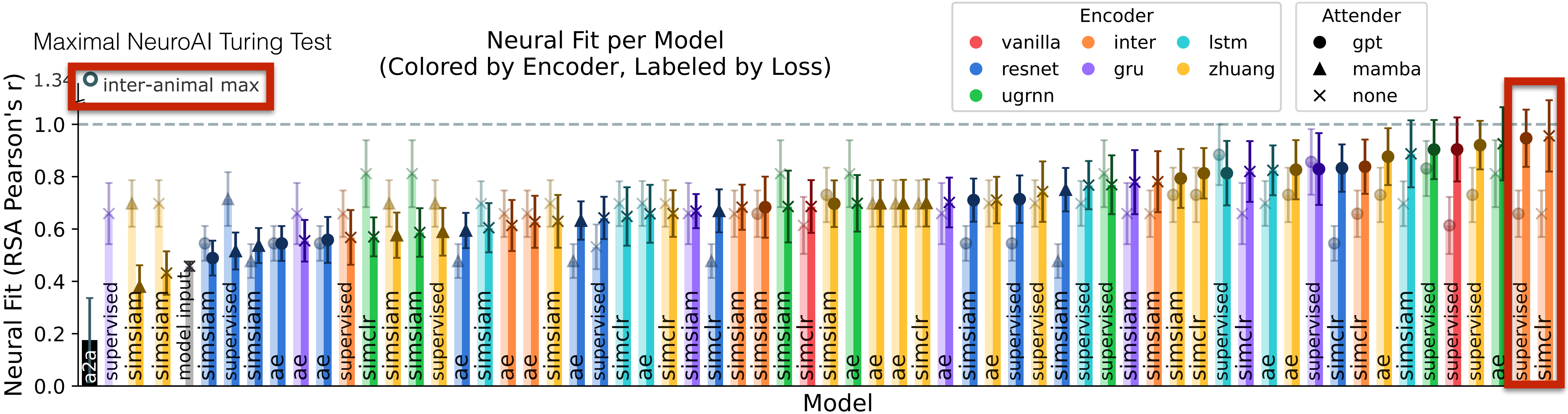


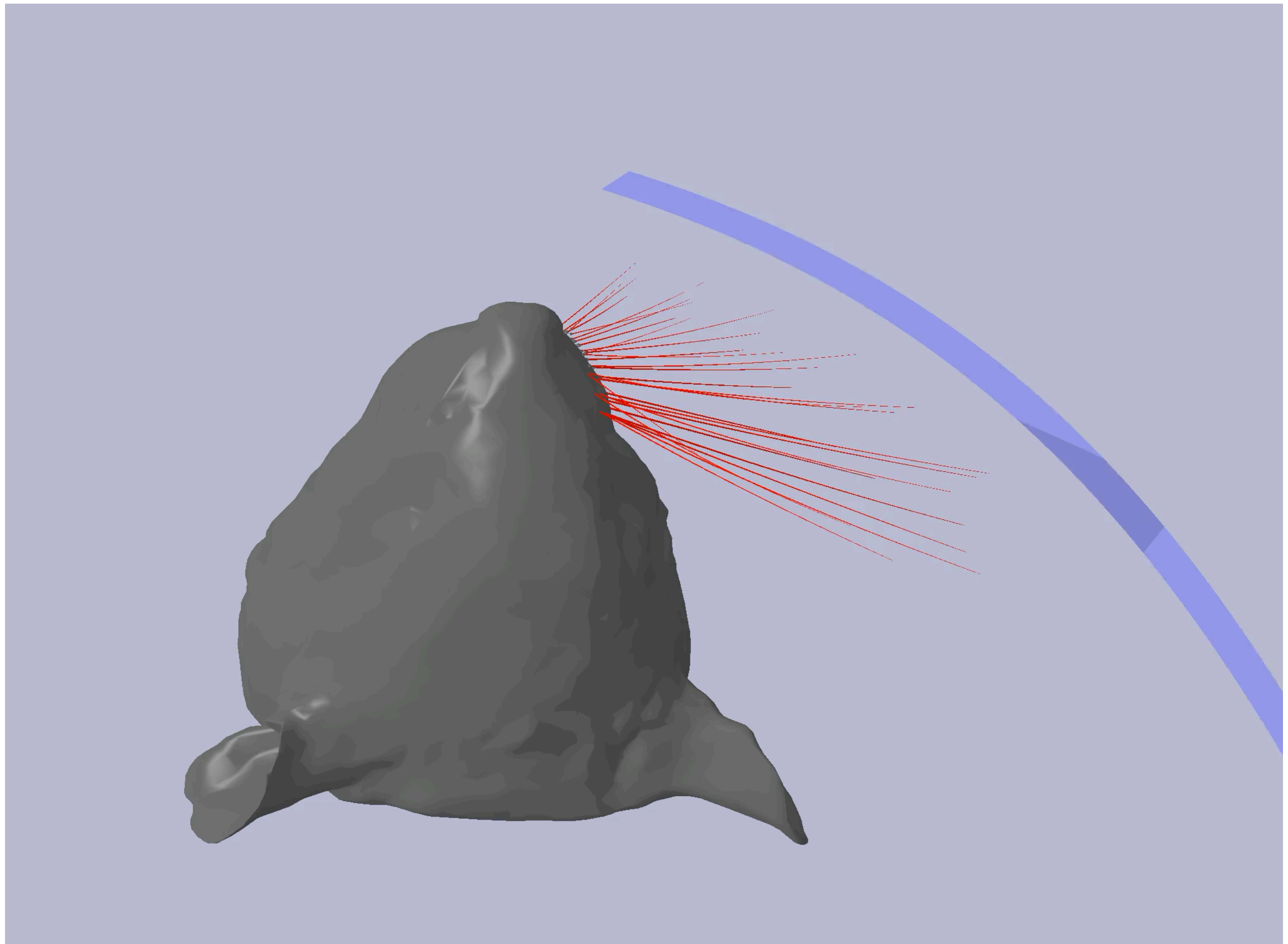




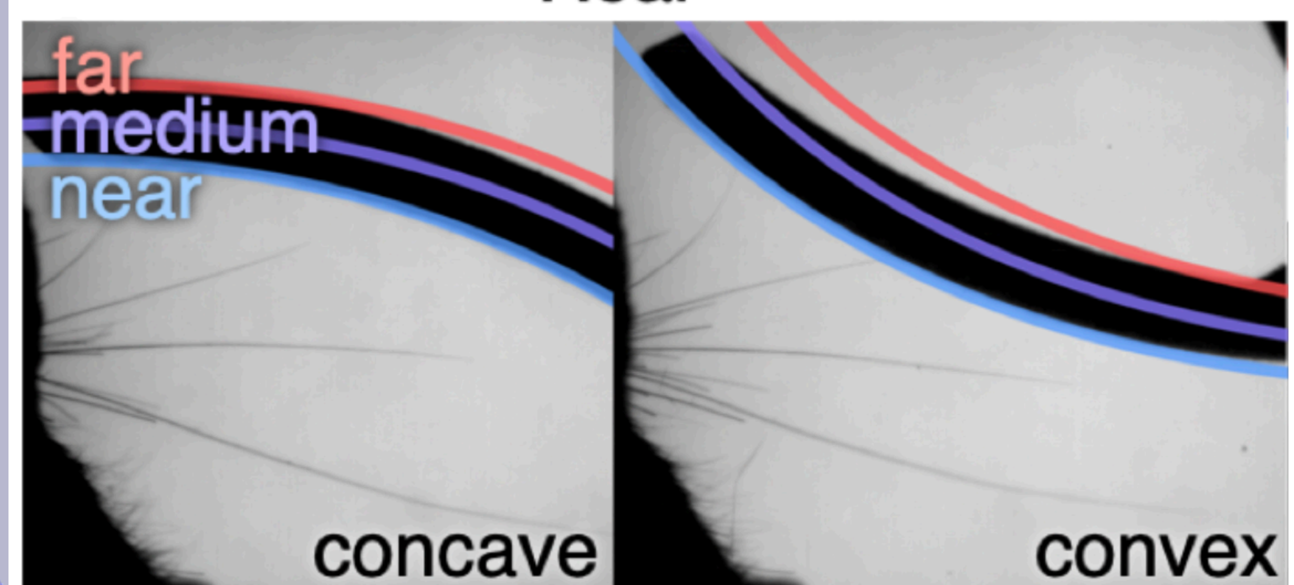
Rodgers 2022

Neural Evaluation: Results



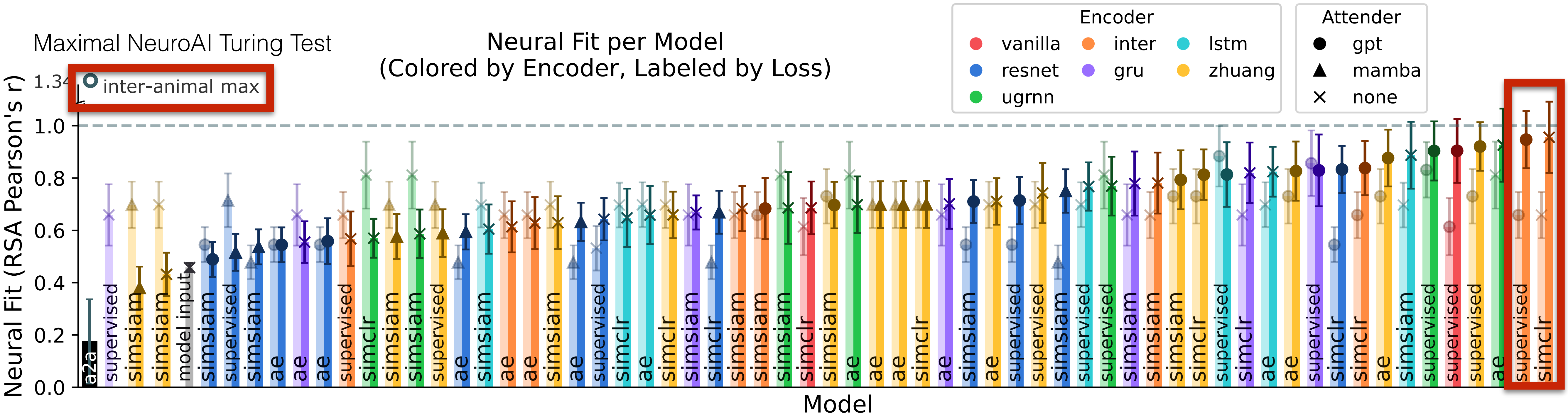


Real

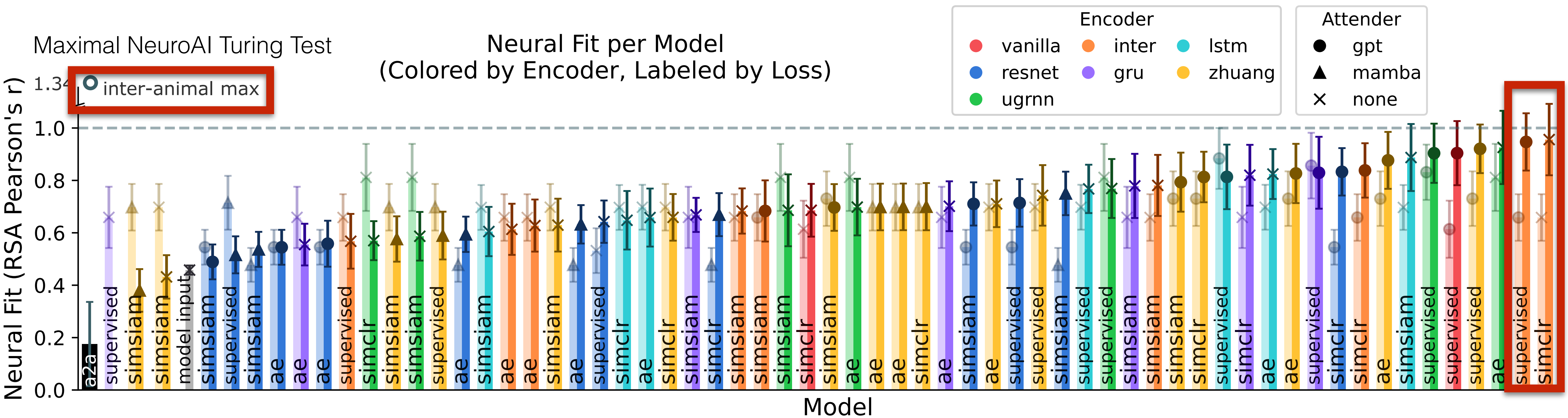


Rodgers 2022

Neural Evaluation: Results

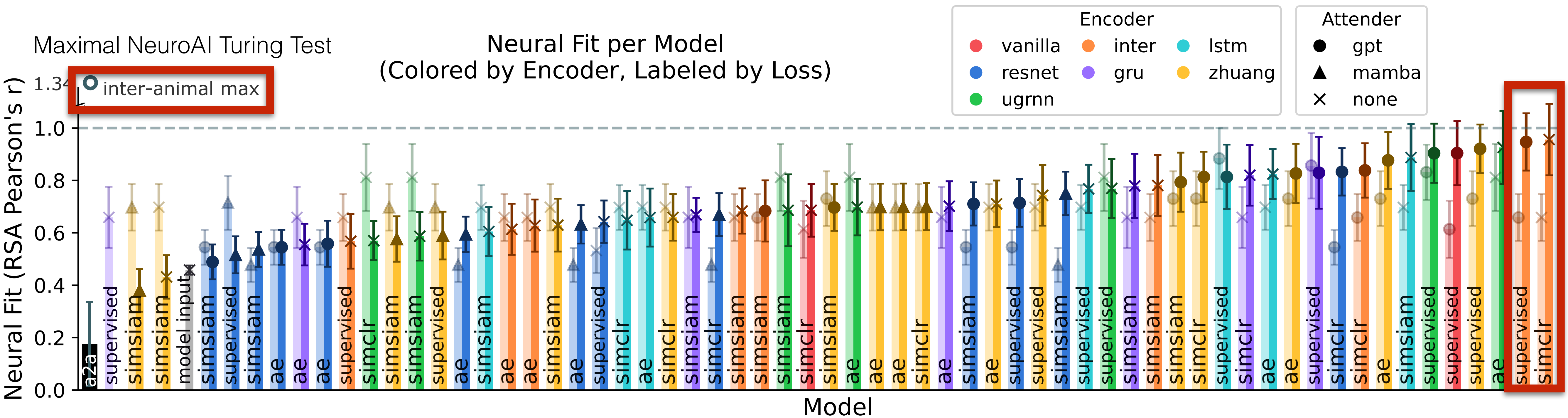


Neural Evaluation: Results



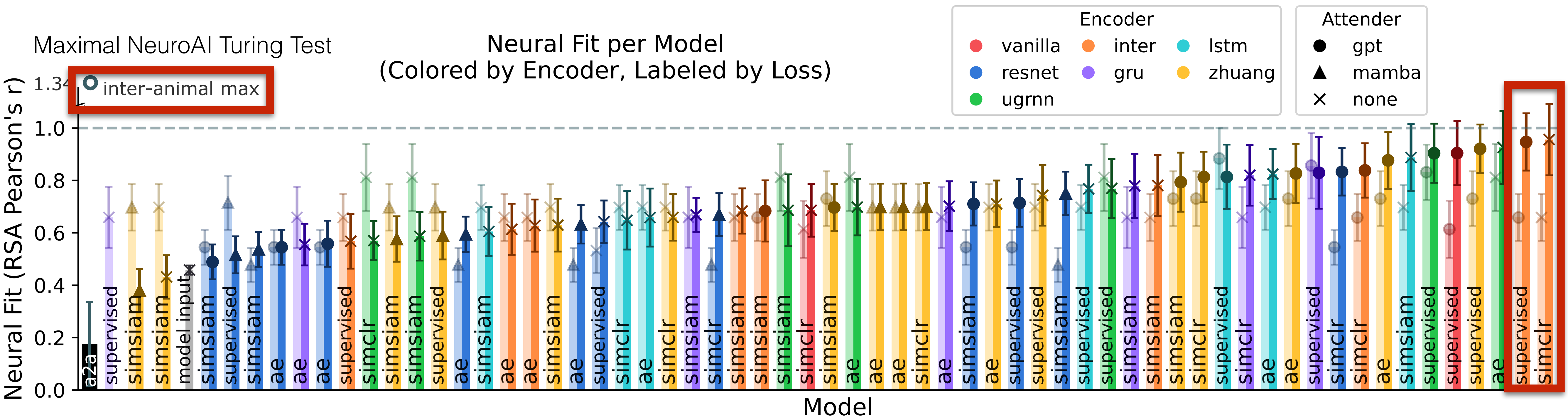
- We’ve nearly passed the NeuroAI Turing Test, for this dataset at least

Neural Evaluation: Results



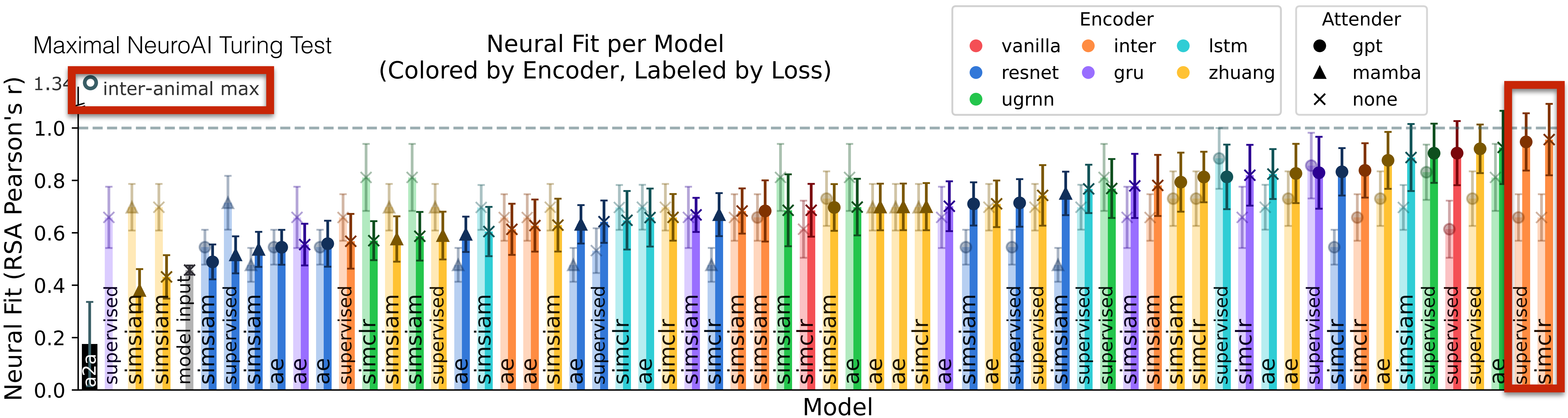
- We’ve nearly passed the NeuroAI Turing Test, for this dataset at least
  - Need more stimuli to evaluate with!

Neural Evaluation: Results



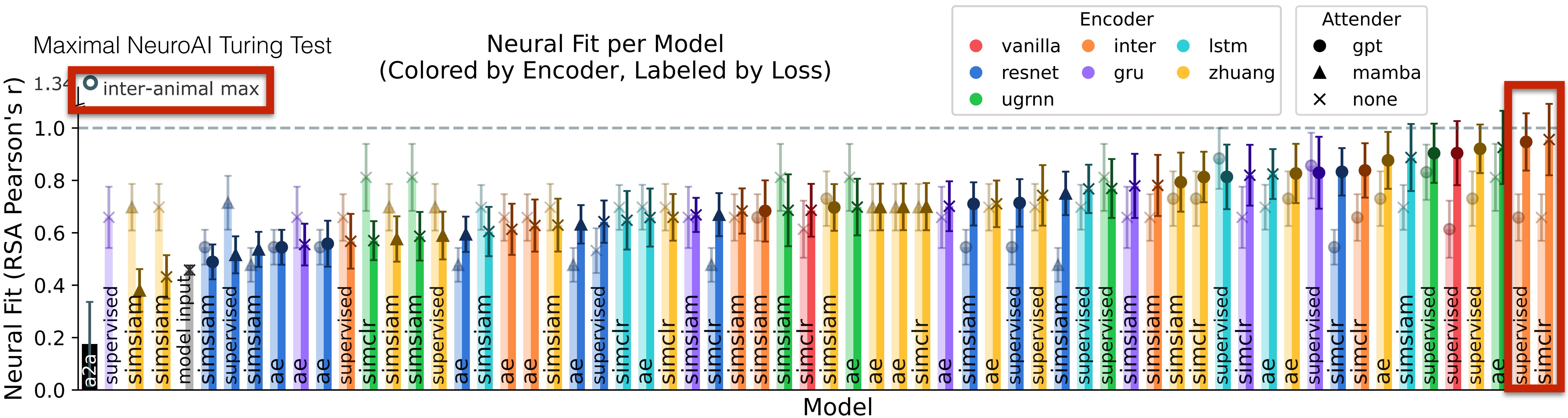
- We've nearly passed the NeuroAI Turing Test, for this dataset at least
  - Need more stimuli to evaluate with!
- ConvRNNs outperform feedforward/SSMs on realistic tactile recognition

Neural Evaluation: Results



- We've nearly passed the NeuroAI Turing Test, for this dataset at least
  - Need more stimuli to evaluate with!
- ConvRNNs outperform feedforward/SSMs on realistic tactile recognition
- ConvRNNs best match neural responses in mouse barrel cortex

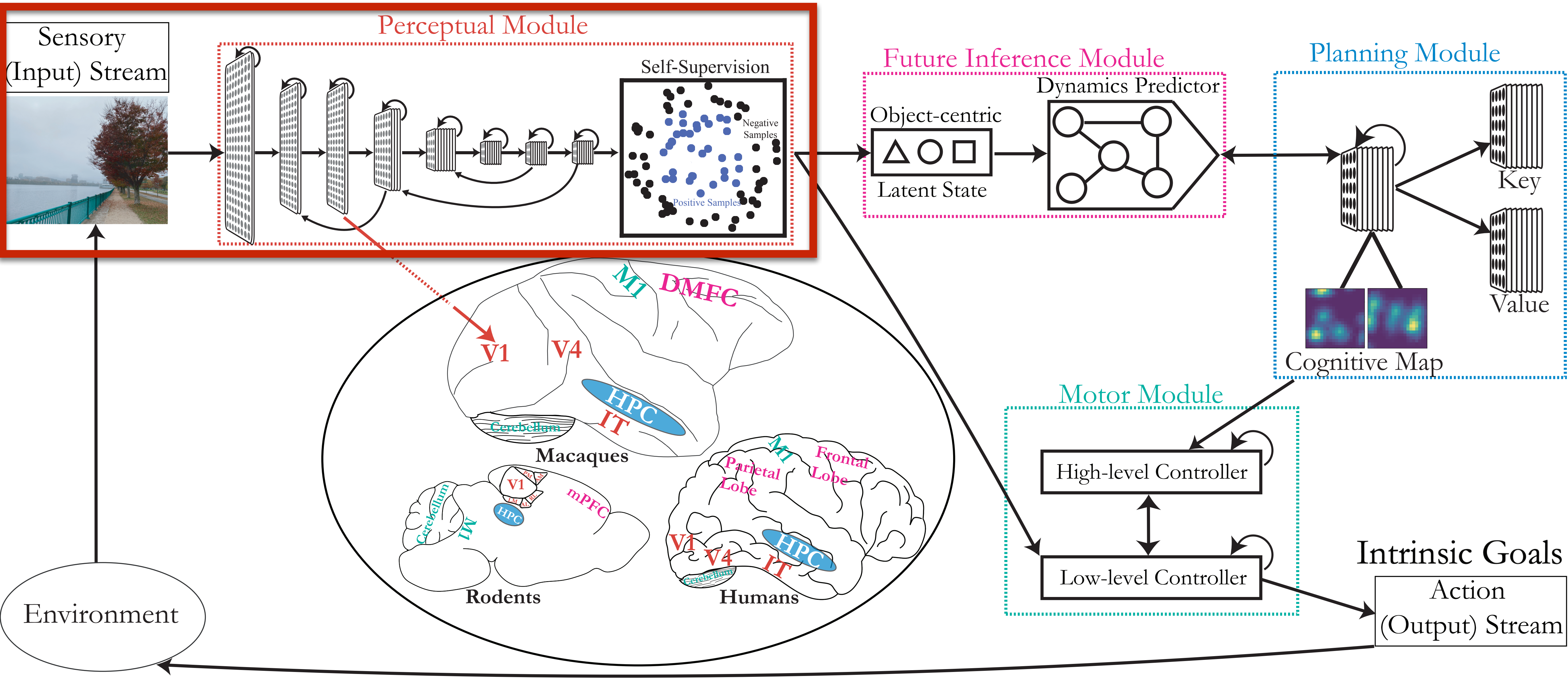
Neural Evaluation: Results



- We've nearly passed the NeuroAI Turing Test, for this dataset at least
  - Need more stimuli to evaluate with!
- ConvRNNs outperform feedforward/SSMs on realistic tactile recognition
- ConvRNNs best match neural responses in mouse barrel cortex
- Contrastive SSL *matches* supervised neural alignment, possibly suggesting a general-purpose representation in the somatosensory cortex (needs more neural data to explore this!)

Roadmap: Perception

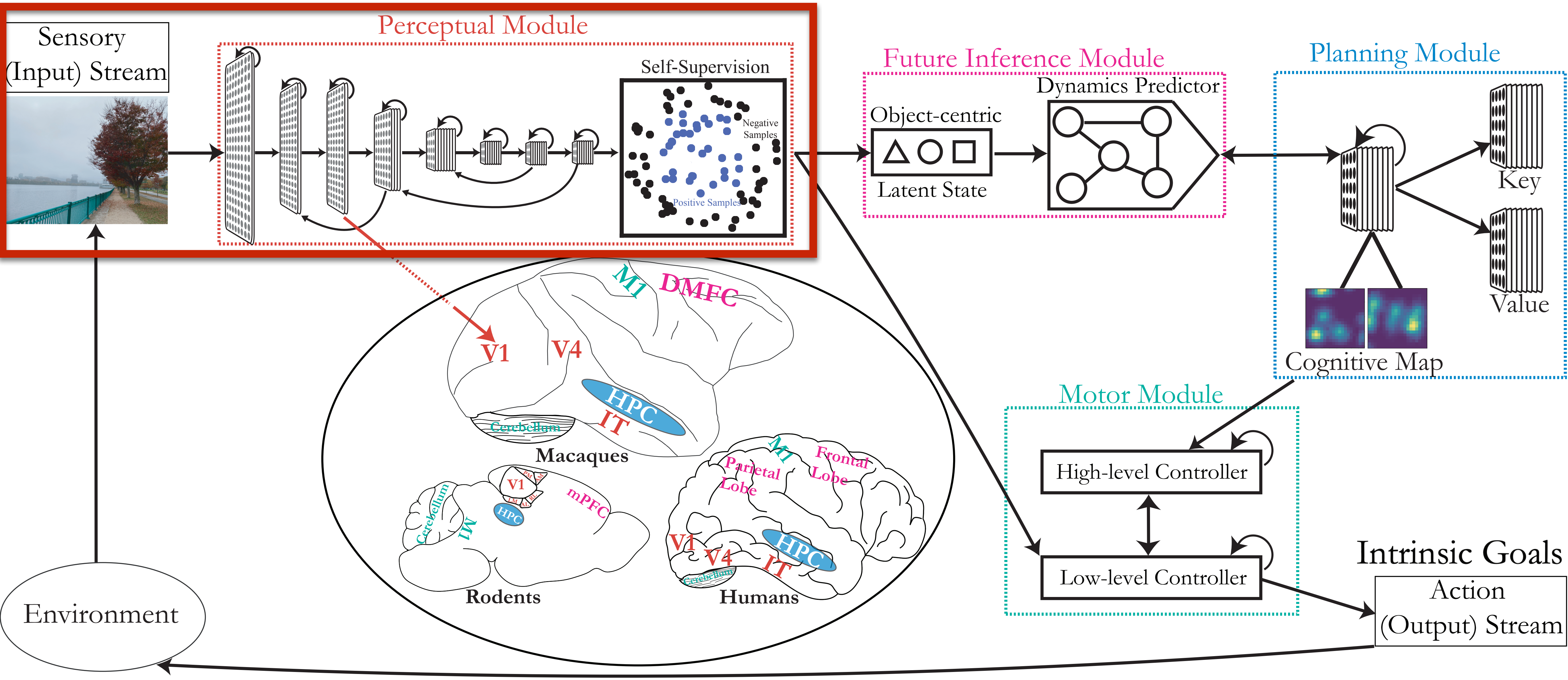
How does the brain *represent*, *predict*, *plan*, and enable *action*?



Roadmap: Perception

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

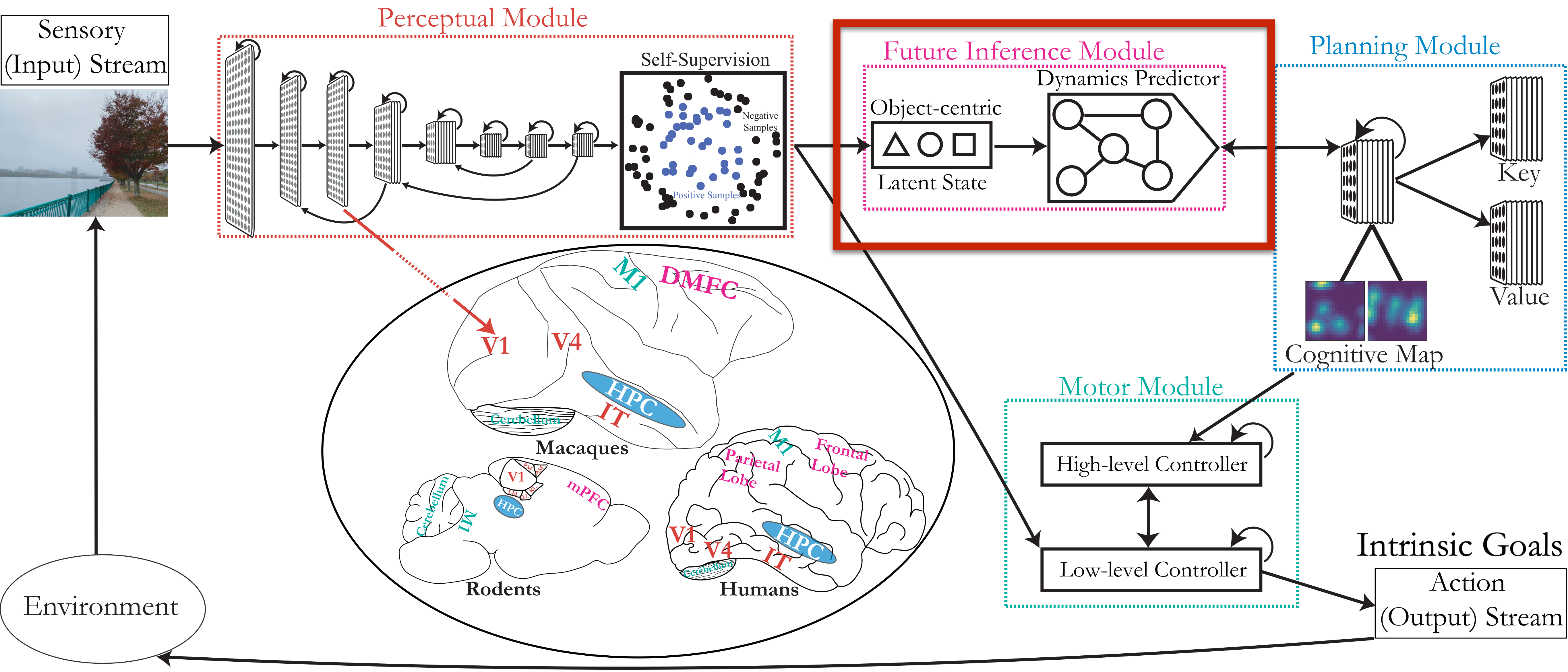
Recurrence + Contrastive SSL?



Roadmap: Future Inference

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

Recurrence + Contrastive SSL?

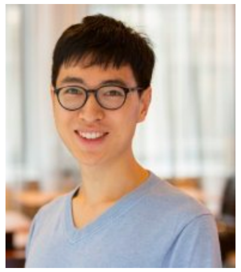


Reusable Latent Representations for Primate Mental Simulation

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


Rishi Rajalingham


Mehrdad Jazayeri


Guangyu Robert Yang

Visually-Grounded Mental Simulation

Visually-Grounded Mental Simulation

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

Visually-Grounded Mental Simulation

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

Visually-Grounded Mental Simulation

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

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

Visually-Grounded Mental Simulation

Infer:
Has this ice block been out longer?

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

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

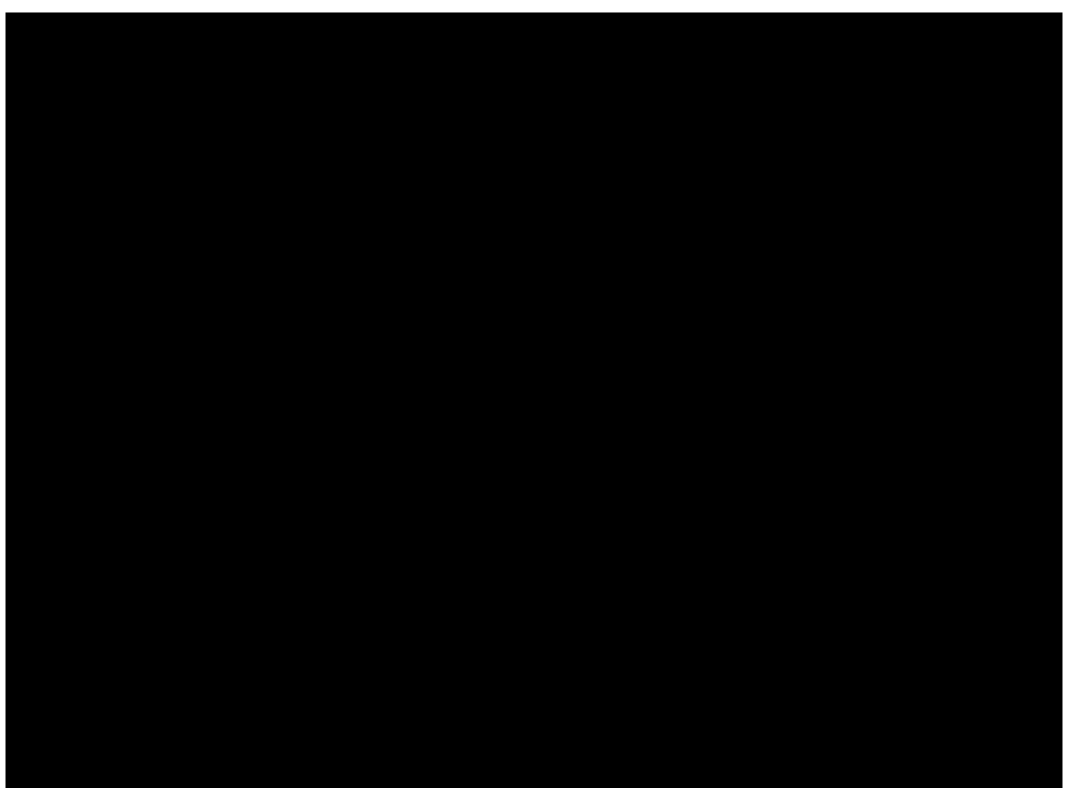
Infer:
Has this ice block been out longer?

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

Predict:
Will this box support me?

Visually-Grounded Mental Simulation

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

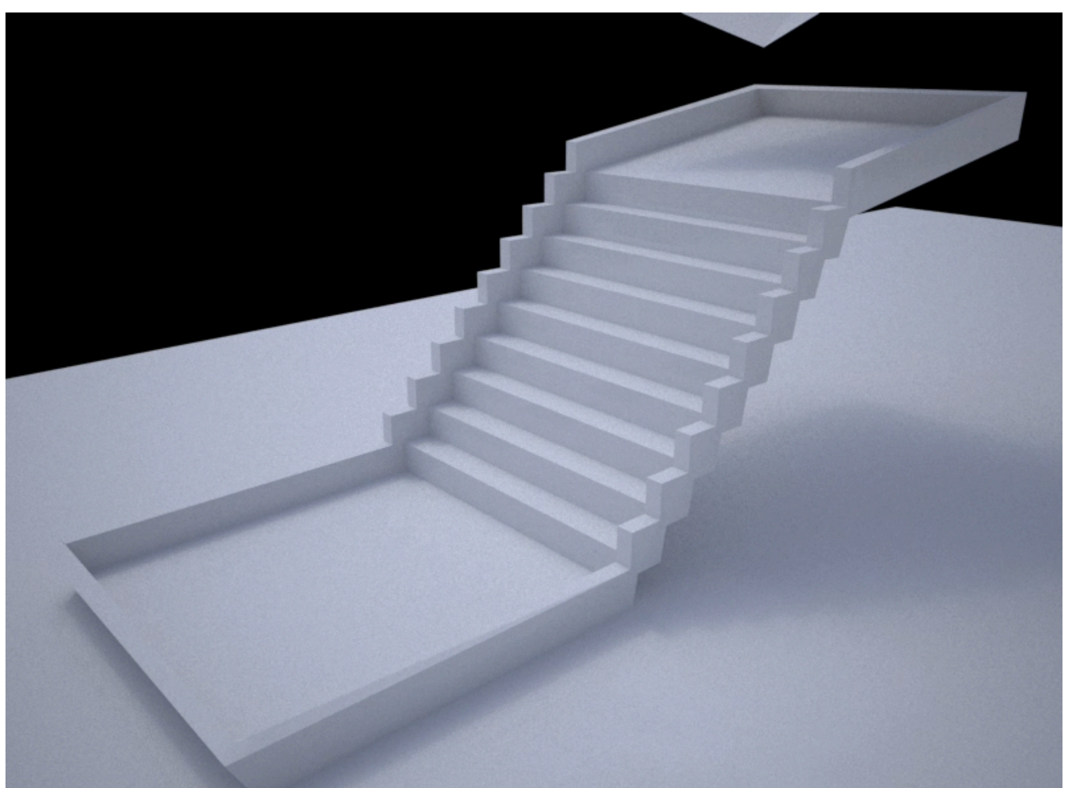


<u>Plan:</u>
How would I take these hats off the rack?

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

Visually-Grounded Mental Simulation

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

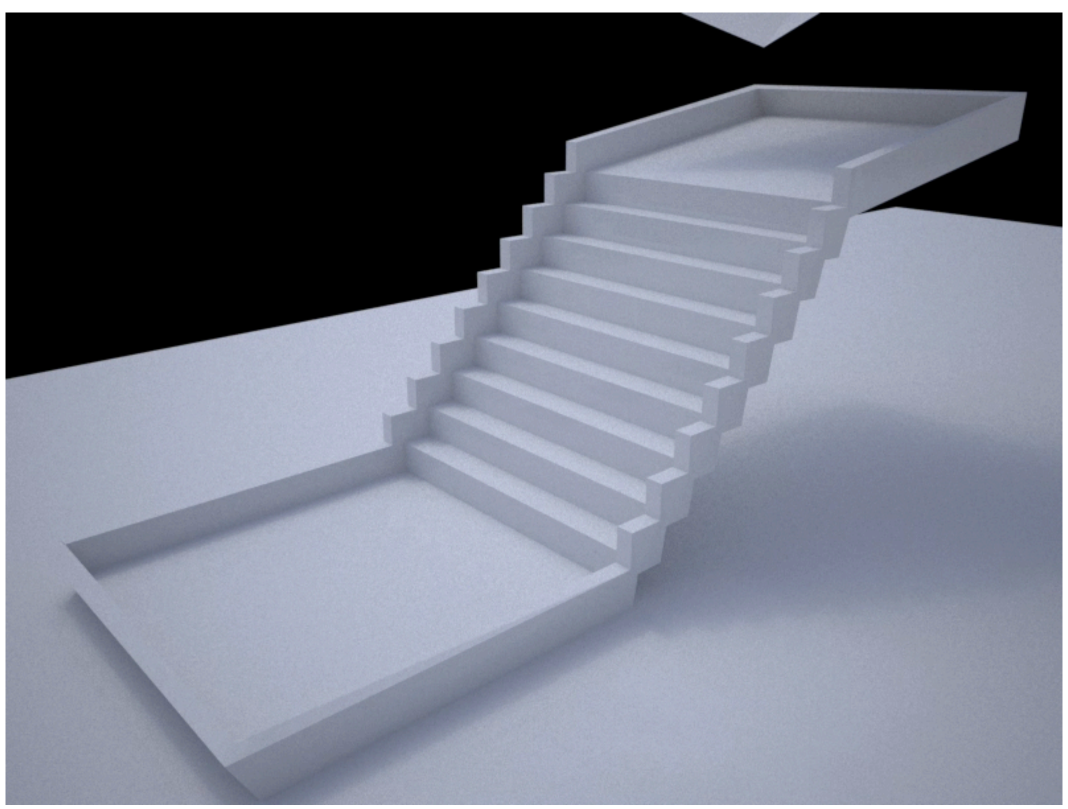


<u>Plan:</u>
How would I take these hats off the rack?

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

Visually-Grounded Mental Simulation

Infer: Has this ice block been out longer?

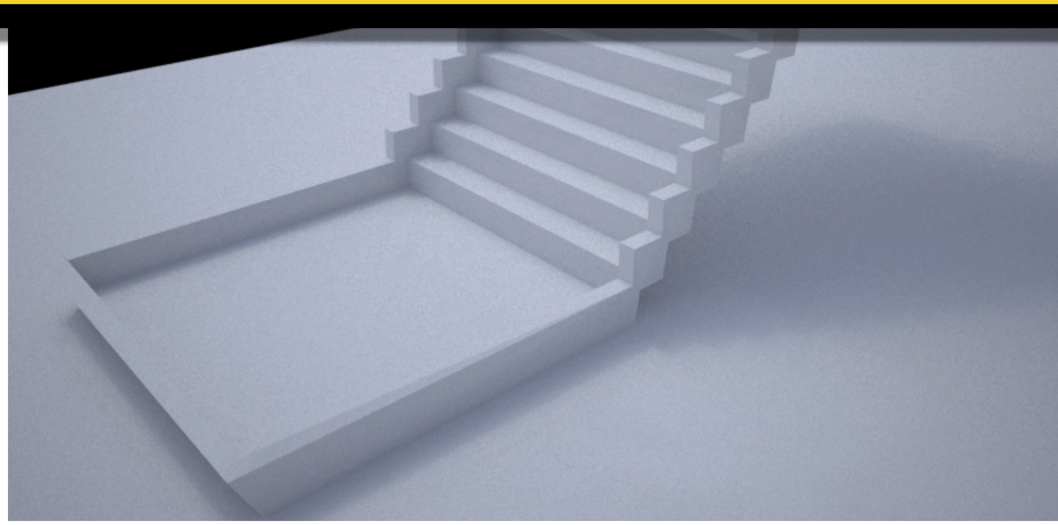


Plan: How would I take these hats off the rack?

Predict: Will this box support me?

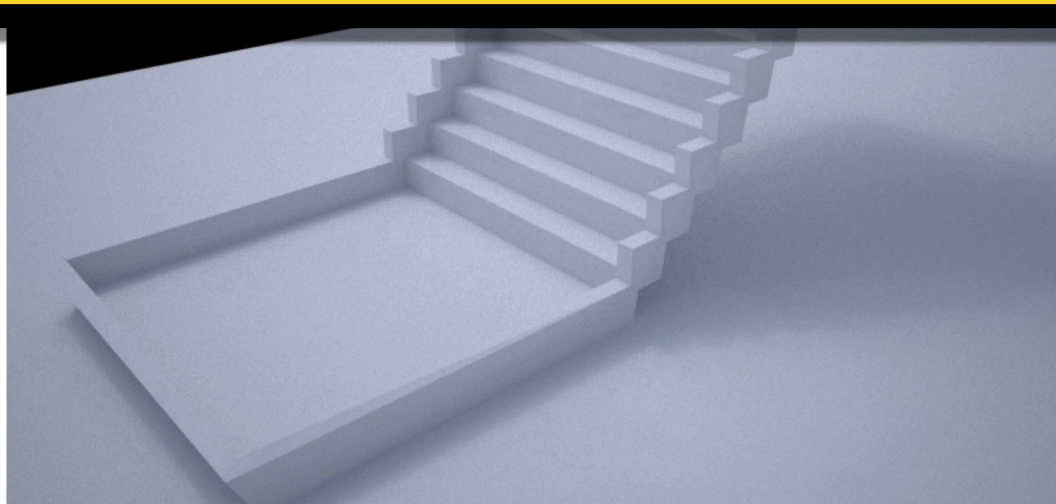
Visually-Grounded Mental Simulation

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

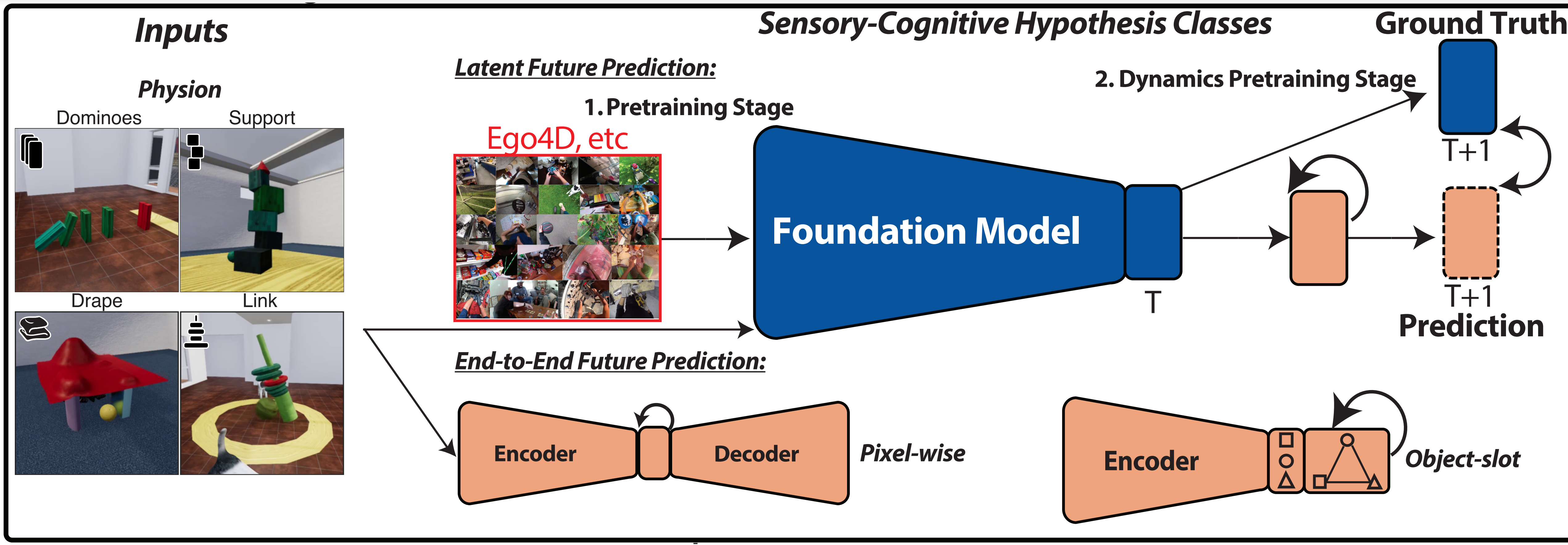


Visually-Grounded Mental Simulation

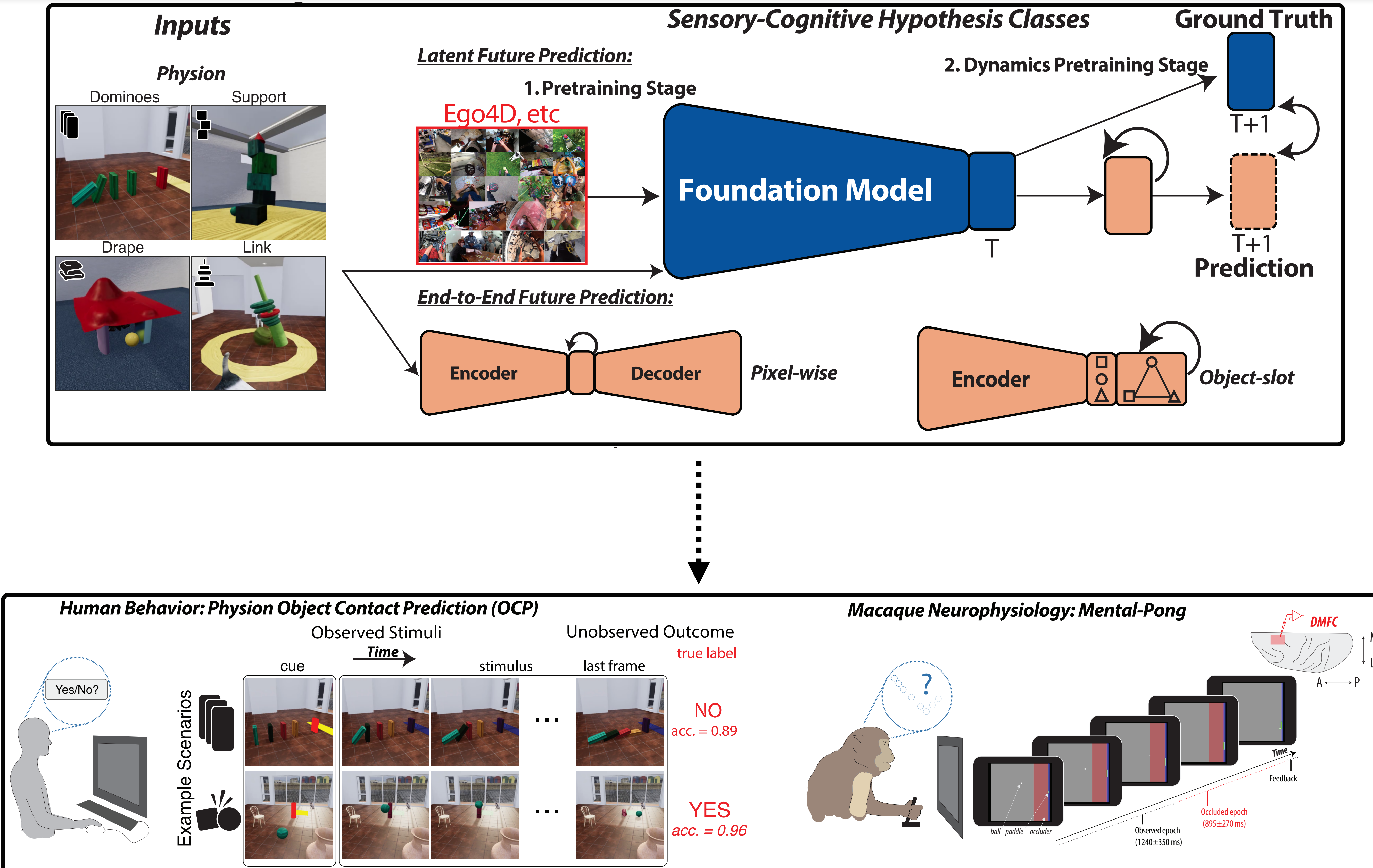
Neurobiological Puzzle:
What are the functional constraints that enable us to predict the future state of our environment *across* diverse settings?



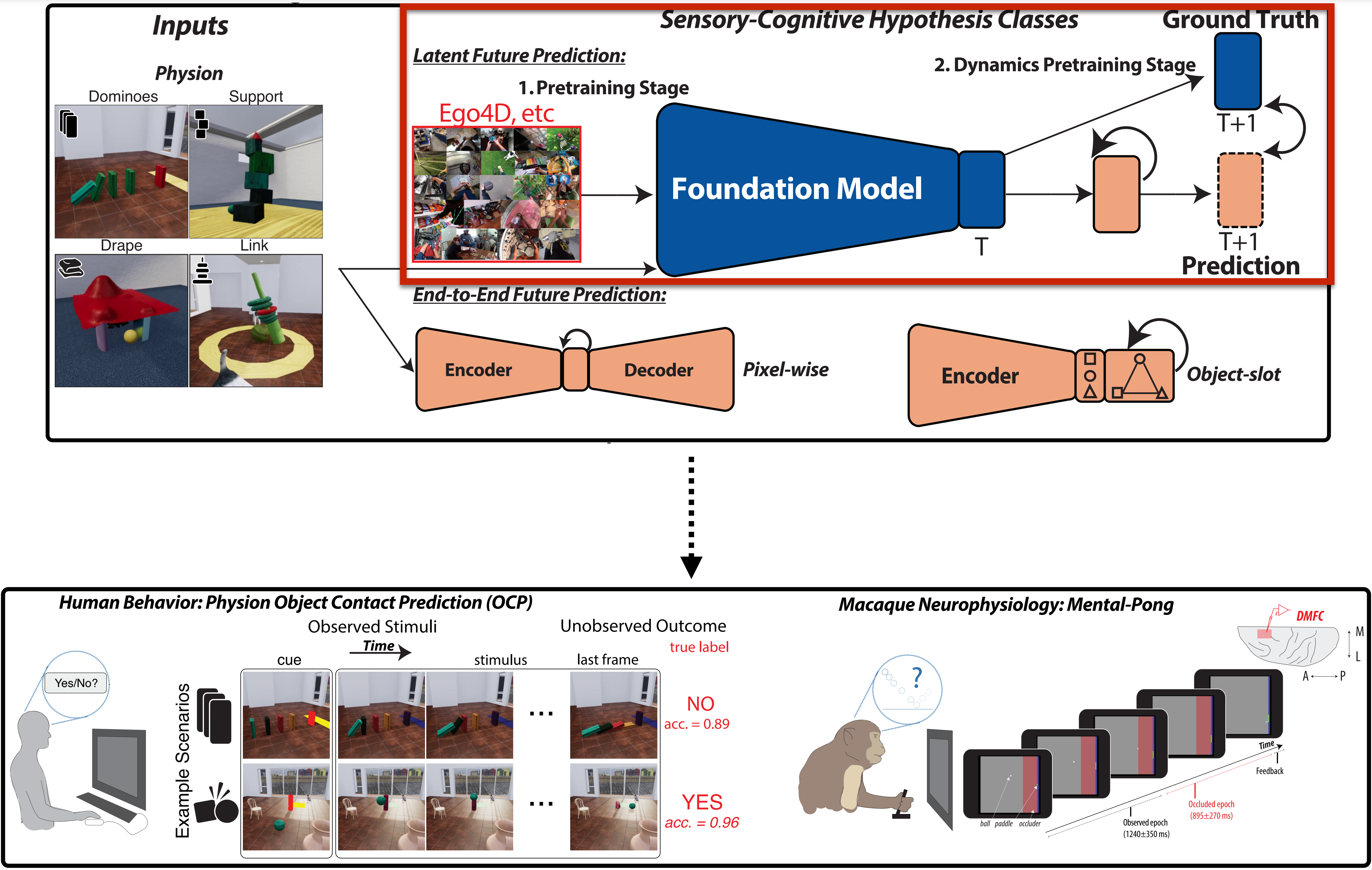
Overall Approach: Sensory-Cognitive Hypotheses



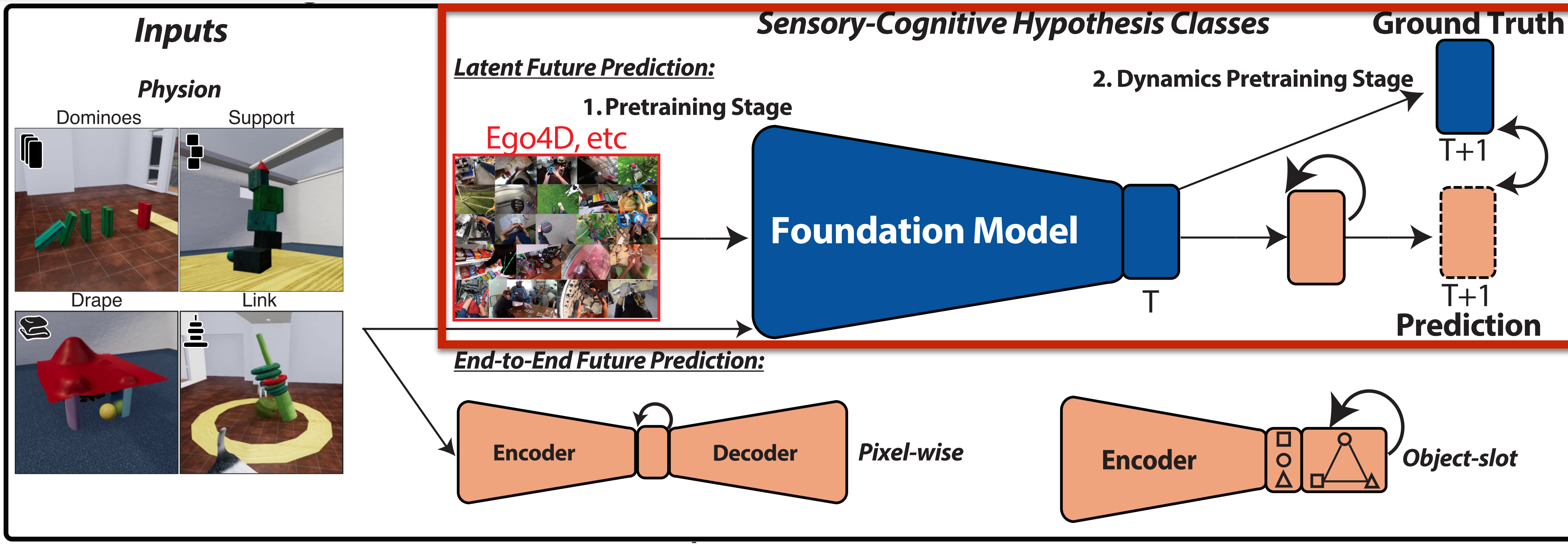
Overall Approach: Sensory-Cognitive Hypotheses



Overall Approach: Sensory-Cognitive Hypotheses

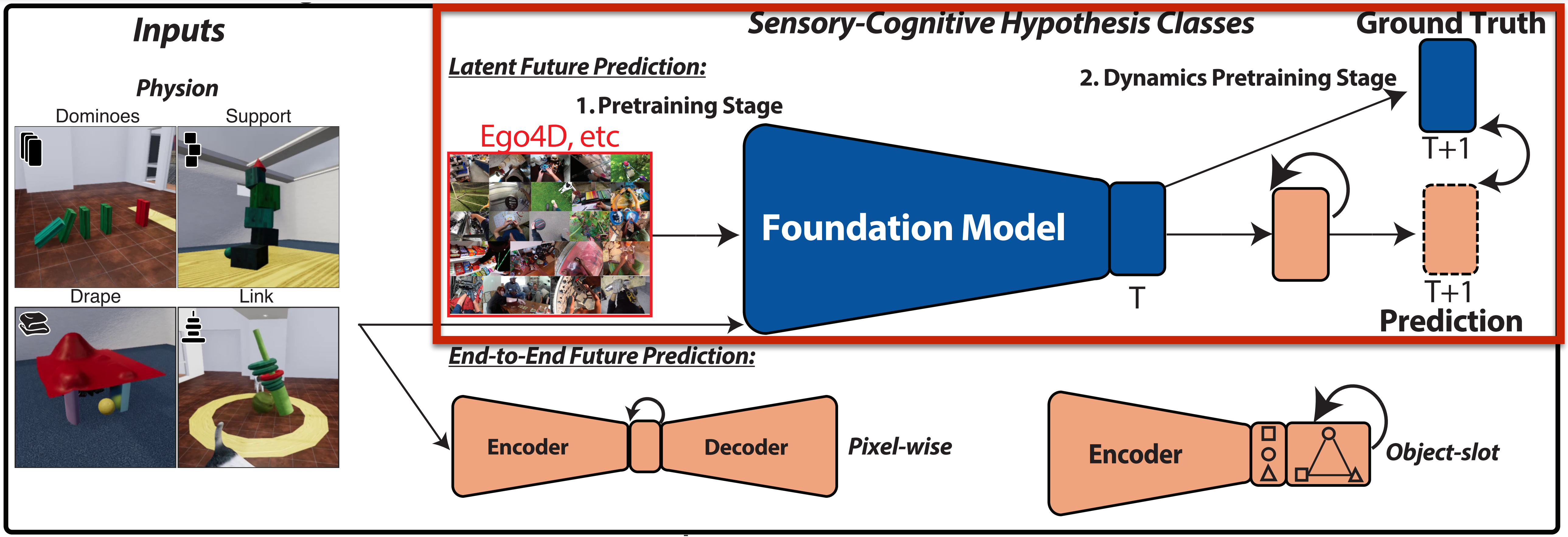


Latent Future Prediction



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

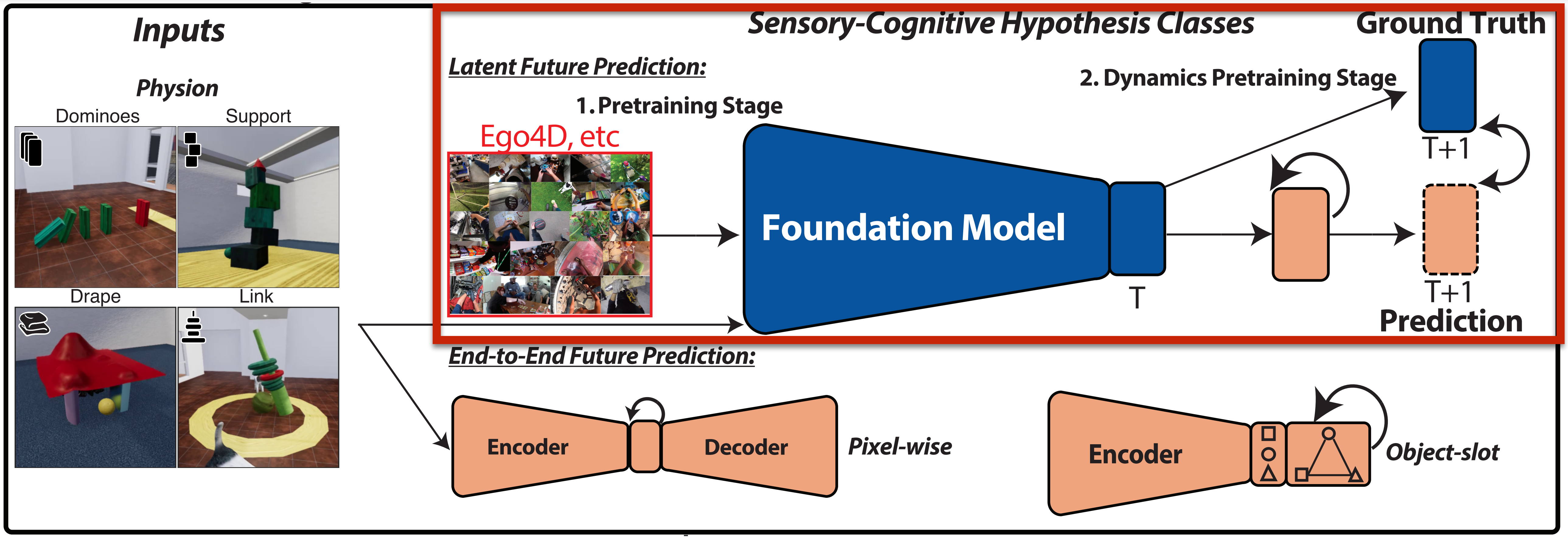
Latent Future Prediction



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

Latent Future Prediction

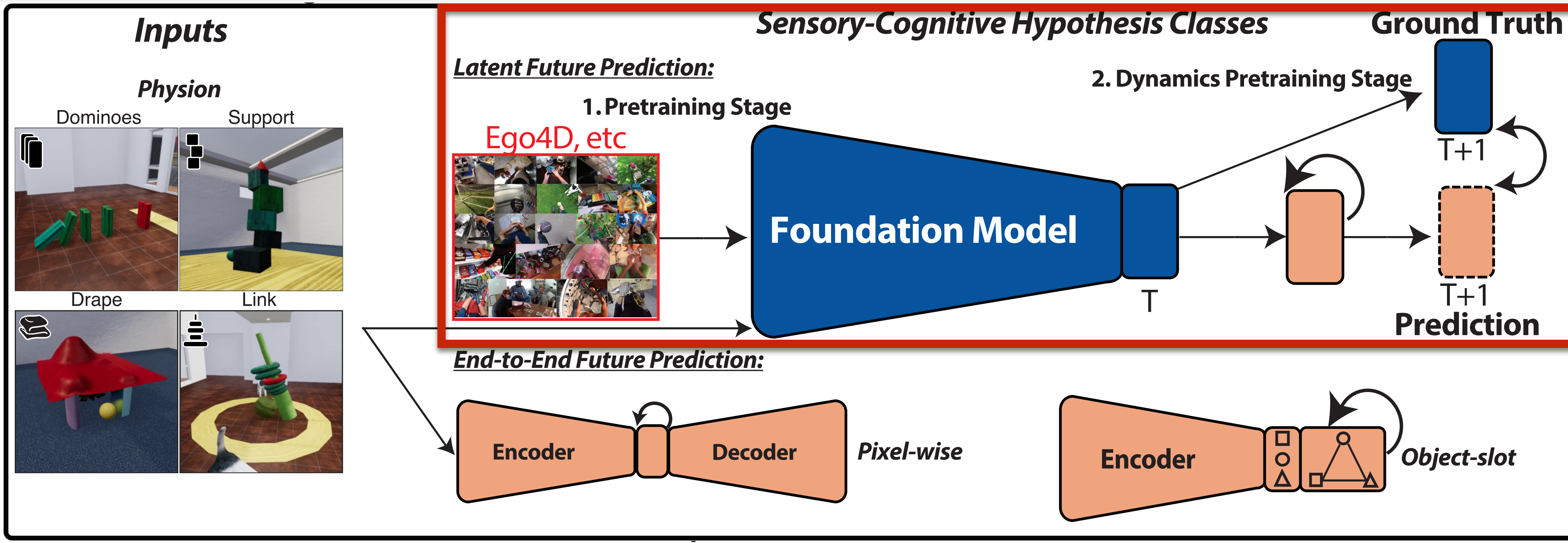


Learn a partial, *implicit* representation of the physical world by performing a challenging vision task ("foundation model")

What vision task?

Leverage these dynamics to do explicit future prediction

Latent Future Prediction



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

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

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

Video Foundation Models

Ego4D: everyday activity around the world

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

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

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

Ego4D: A massive-scale egocentric dataset

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

931 participants from 74 worldwide locations

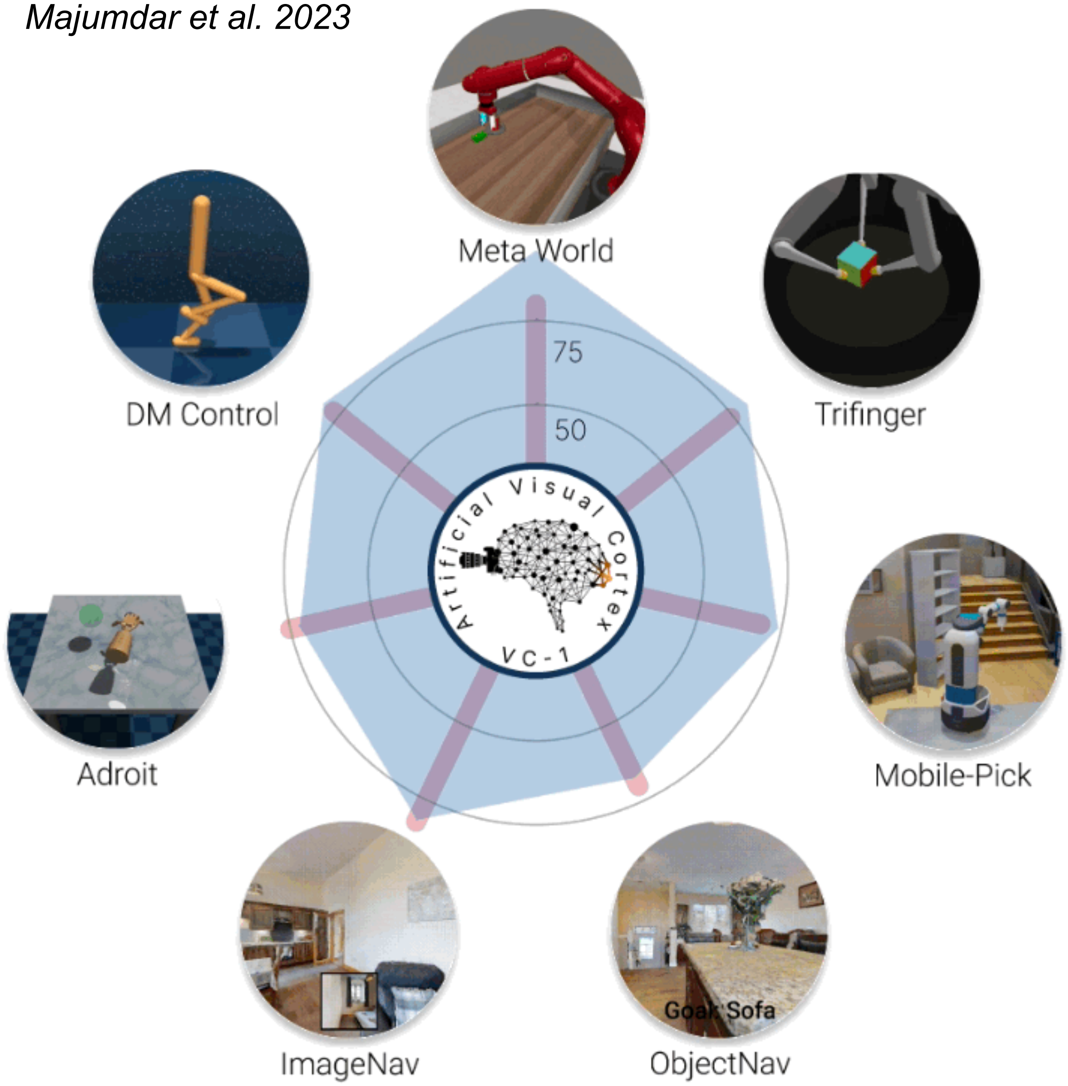
Multimodal: audio, 3D scans, IMU, stereo, multi-camera

Grauman et al. 2022

Video Foundation Models

Ego4D: everyday activity around the world

Majumdar et al. 2023



Ego4D: A massive-scale egocentric dataset

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

931 participants from 74 worldwide locations

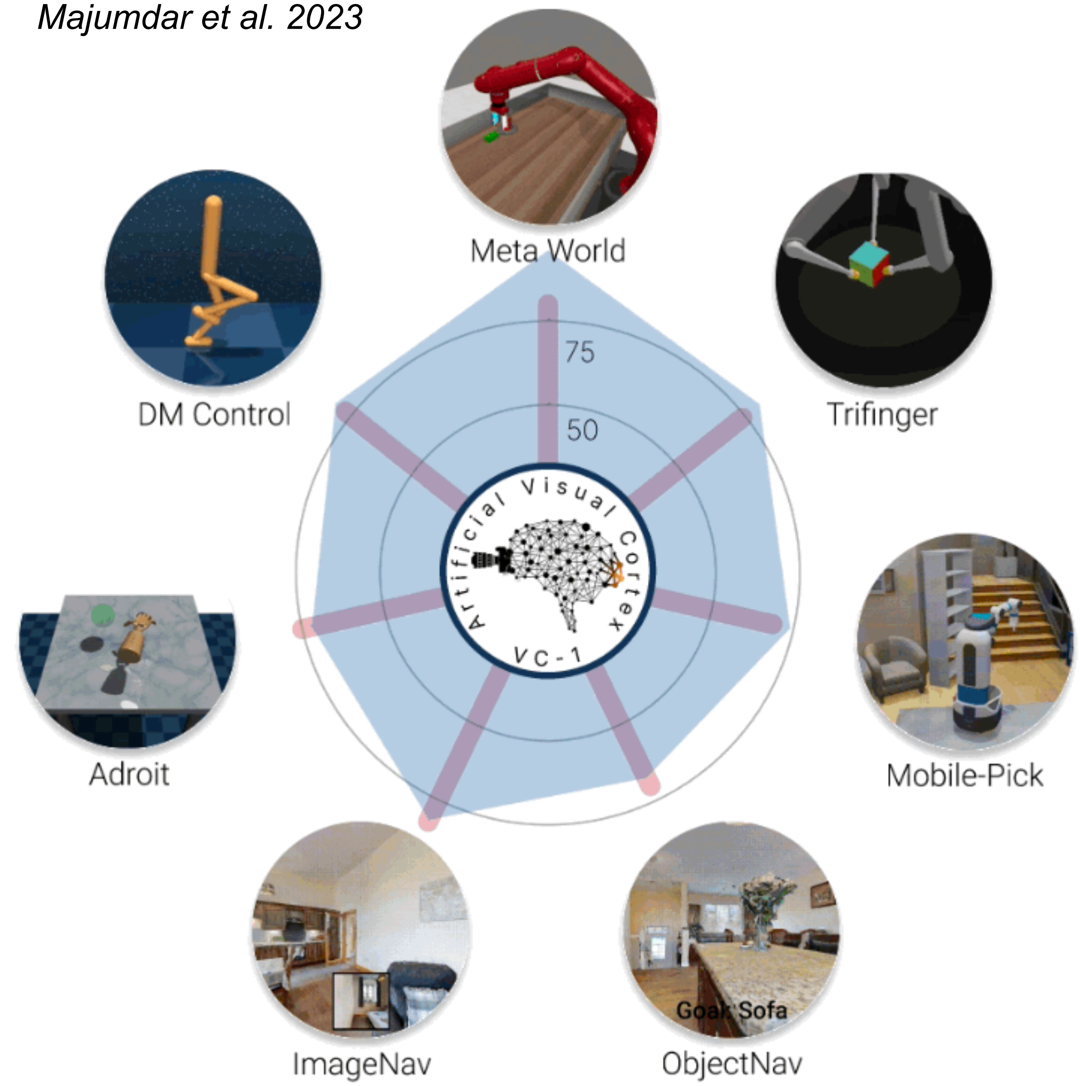
Multimodal: audio, 3D scans, IMU, stereo, multi-camera

Grauman et al. 2022

Video Foundation Models

Ego4D: everyday activity around the world

Majumdar et al. 2023

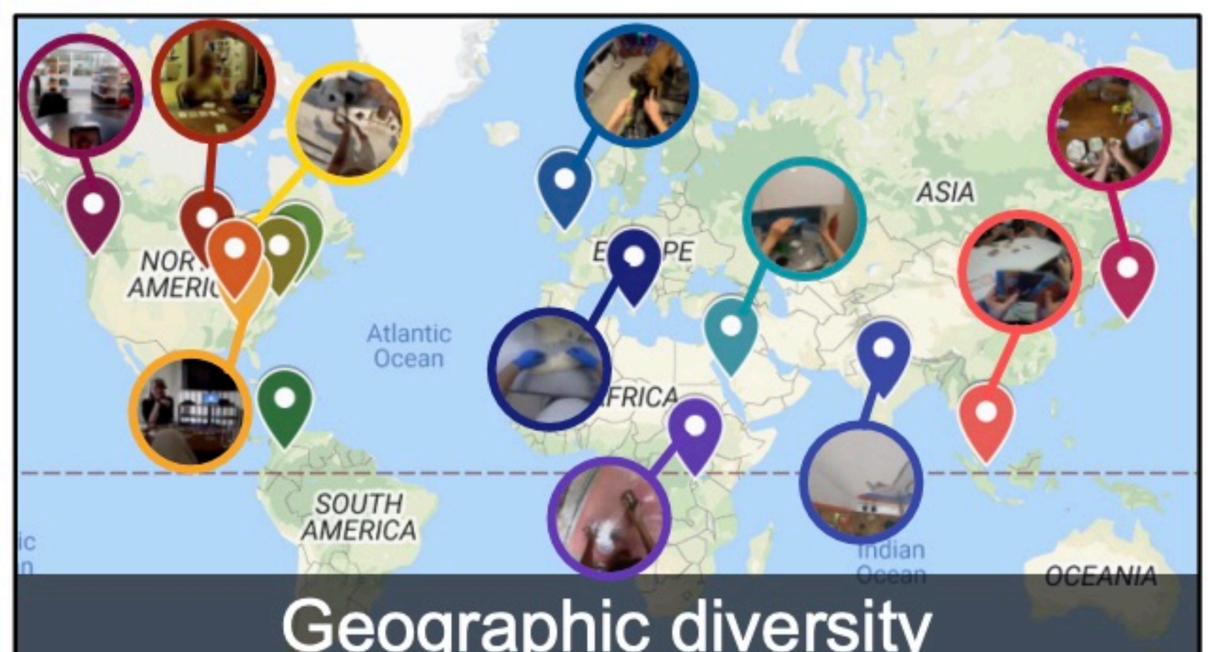


Ego4D: A massive-scale egocentric dataset

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

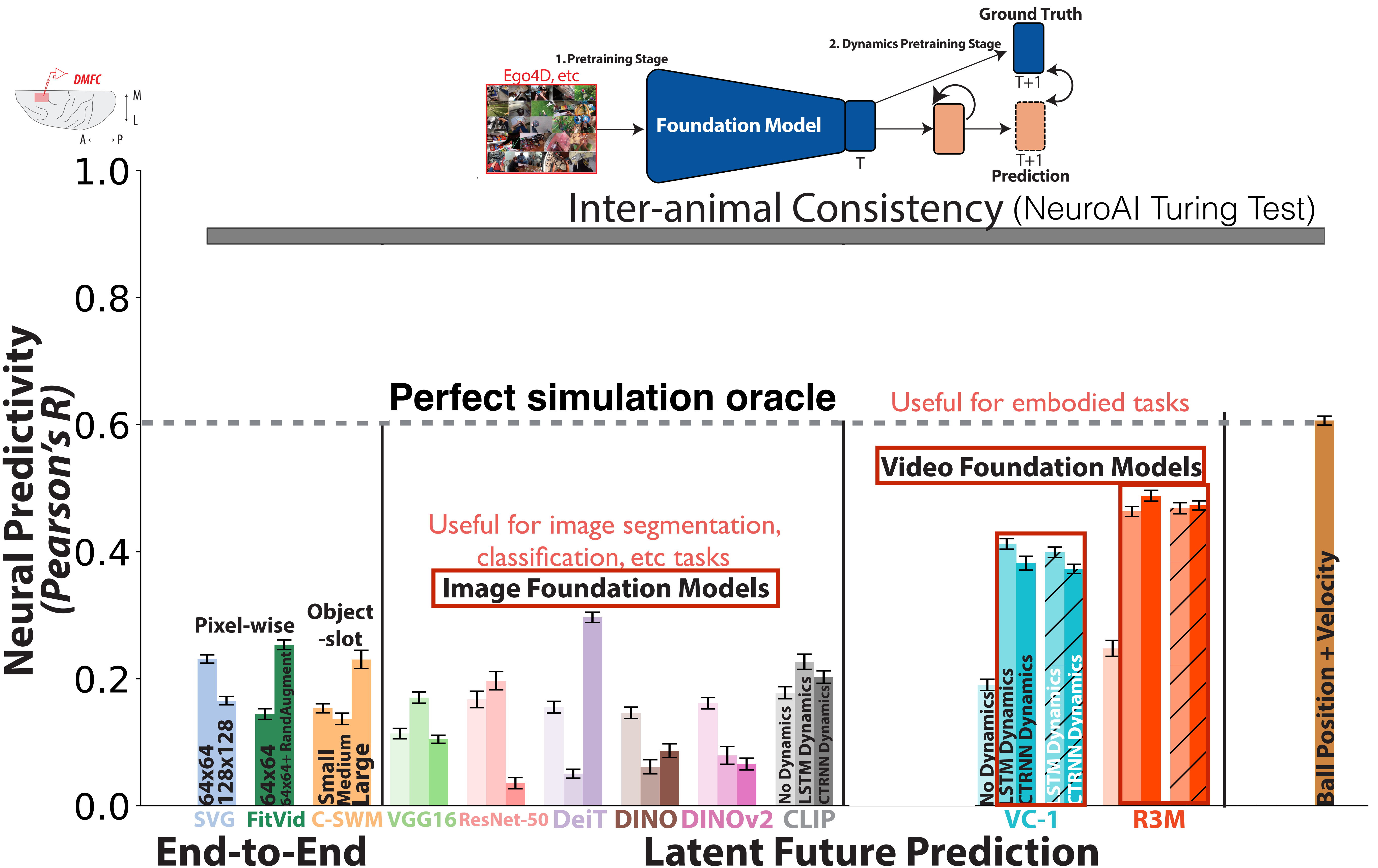
931 participants from 74 worldwide locations

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

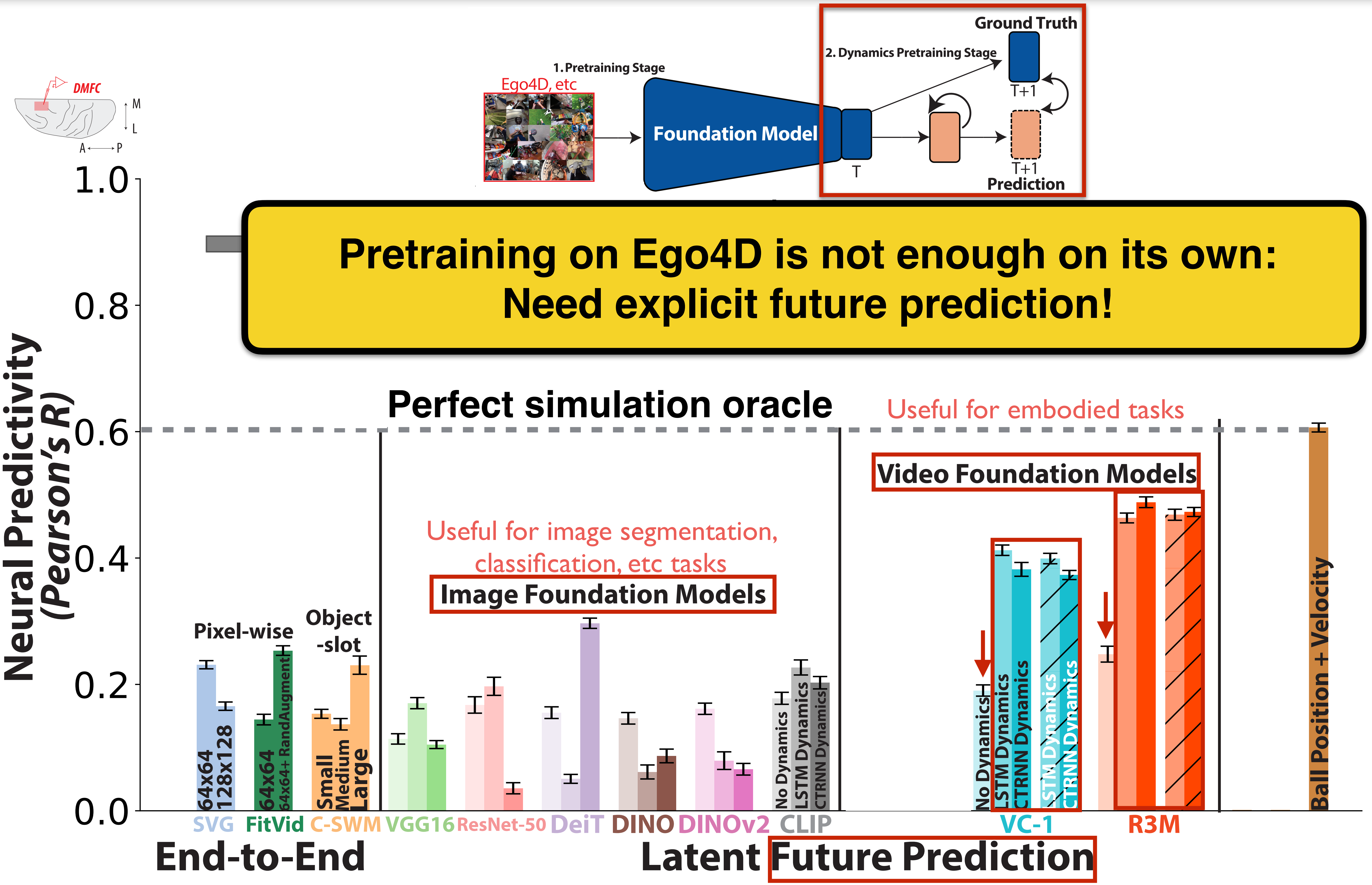


Grauman et al. 2022

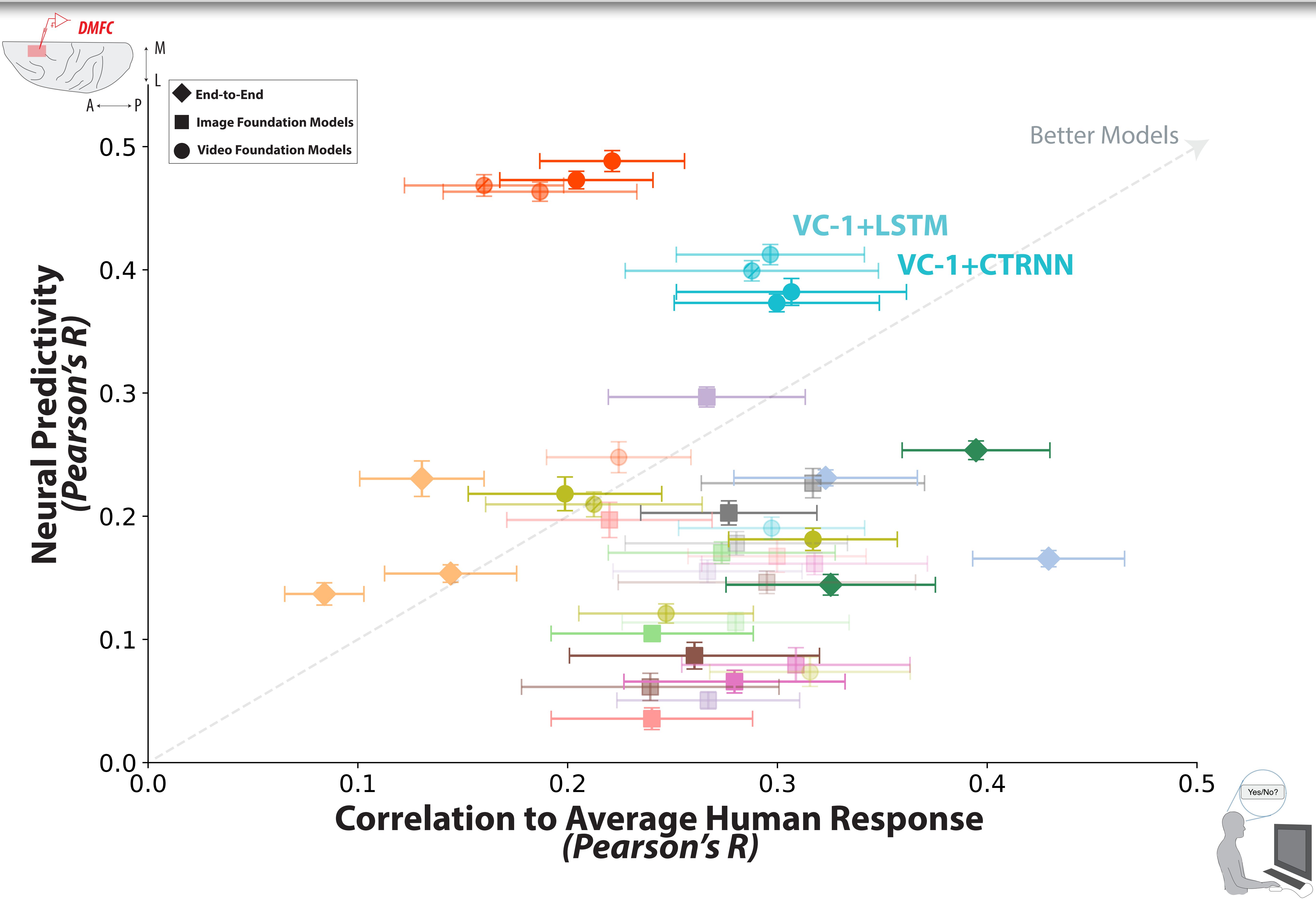
Video Foundation Future Prediction Best Predict Neurons



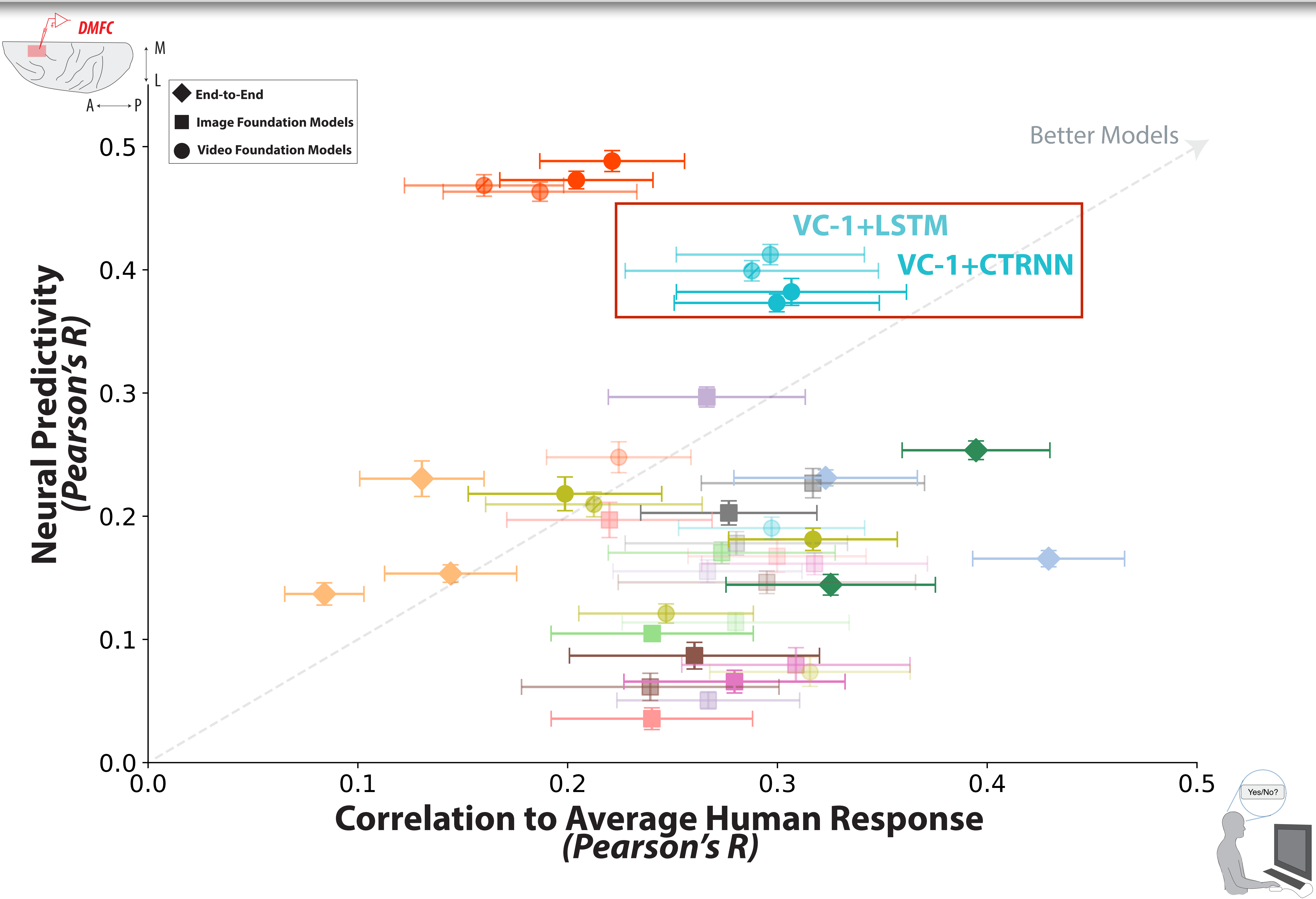
Video Foundation Future Prediction Best Predict Neurons



Dynamically-Equipped Video Foundation Models Can Match Both

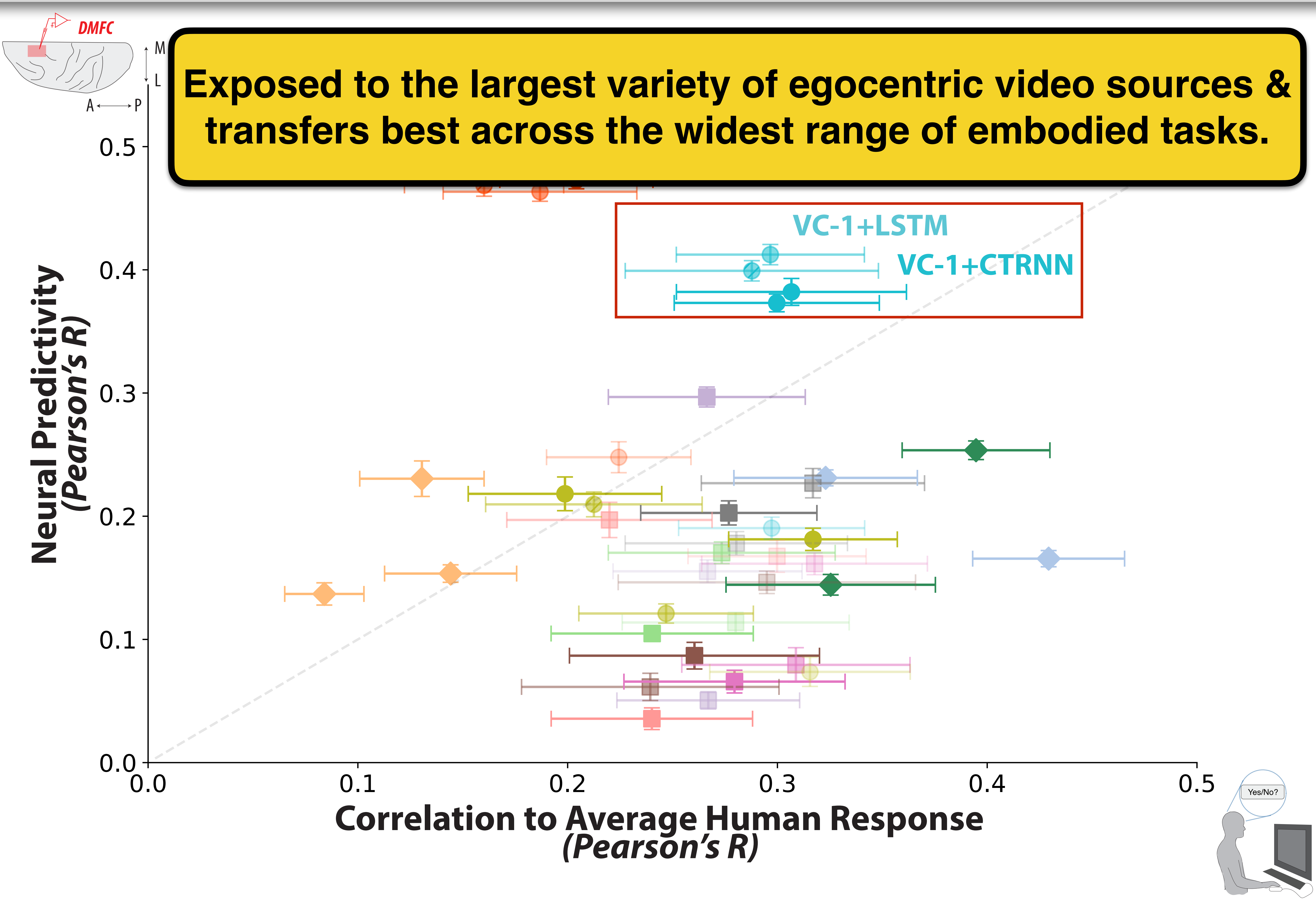


Dynamically-Equipped Video Foundation Models Can Match Both

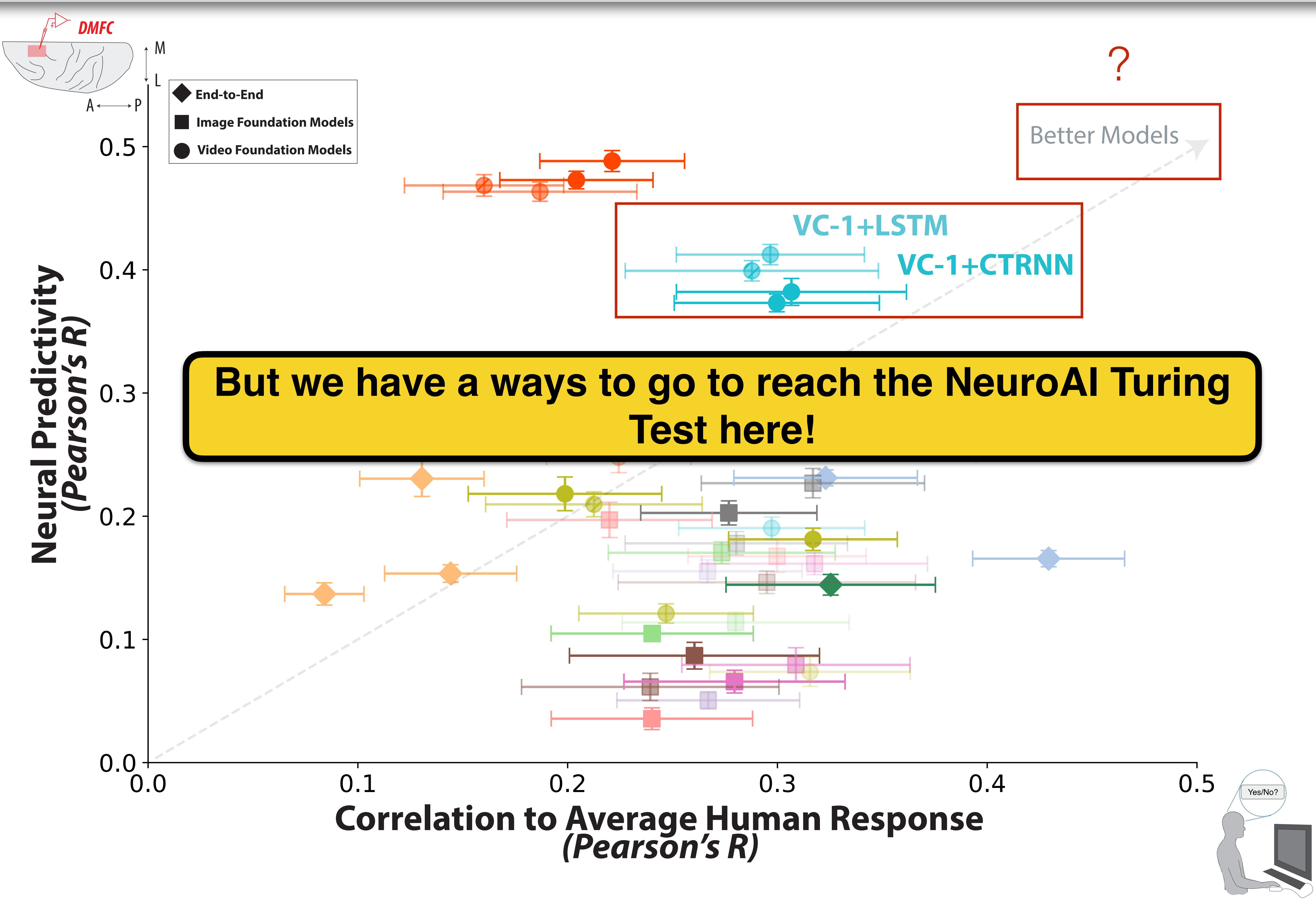


Dynamically-Equipped Video Foundation Models Can Match Both

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



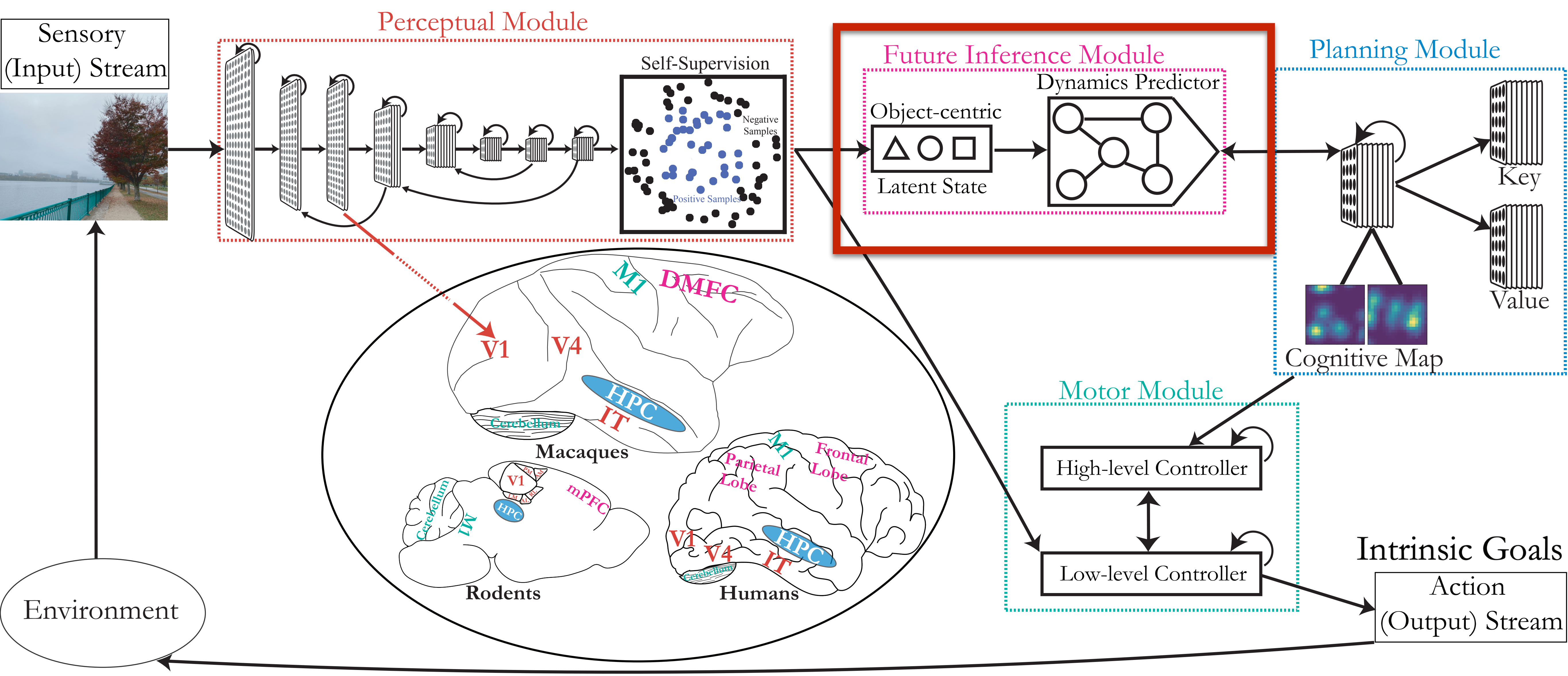
Dynamically-Equipped Video Foundation Models Can Match Both



Roadmap: Future Inference

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

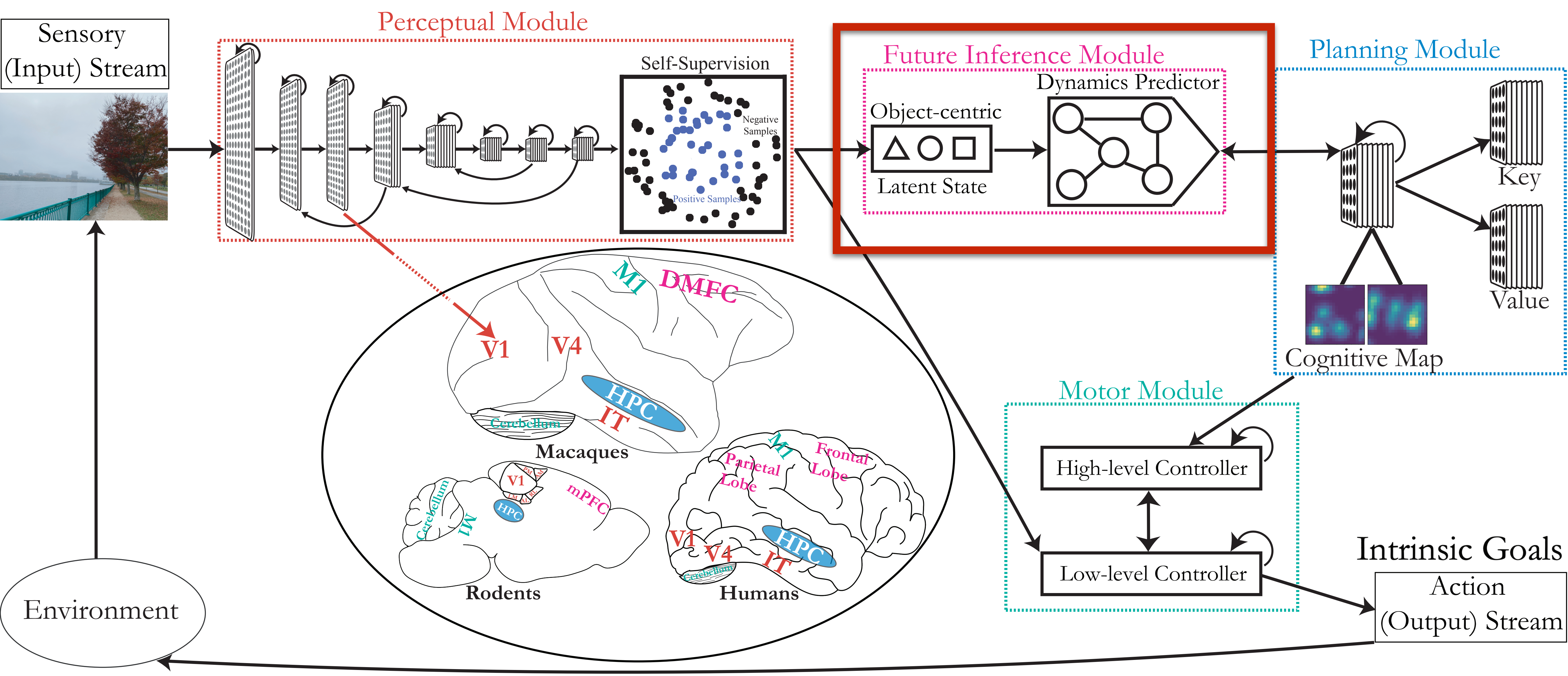
Recurrence + Contrastive SSL?



Roadmap: Future Inference

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

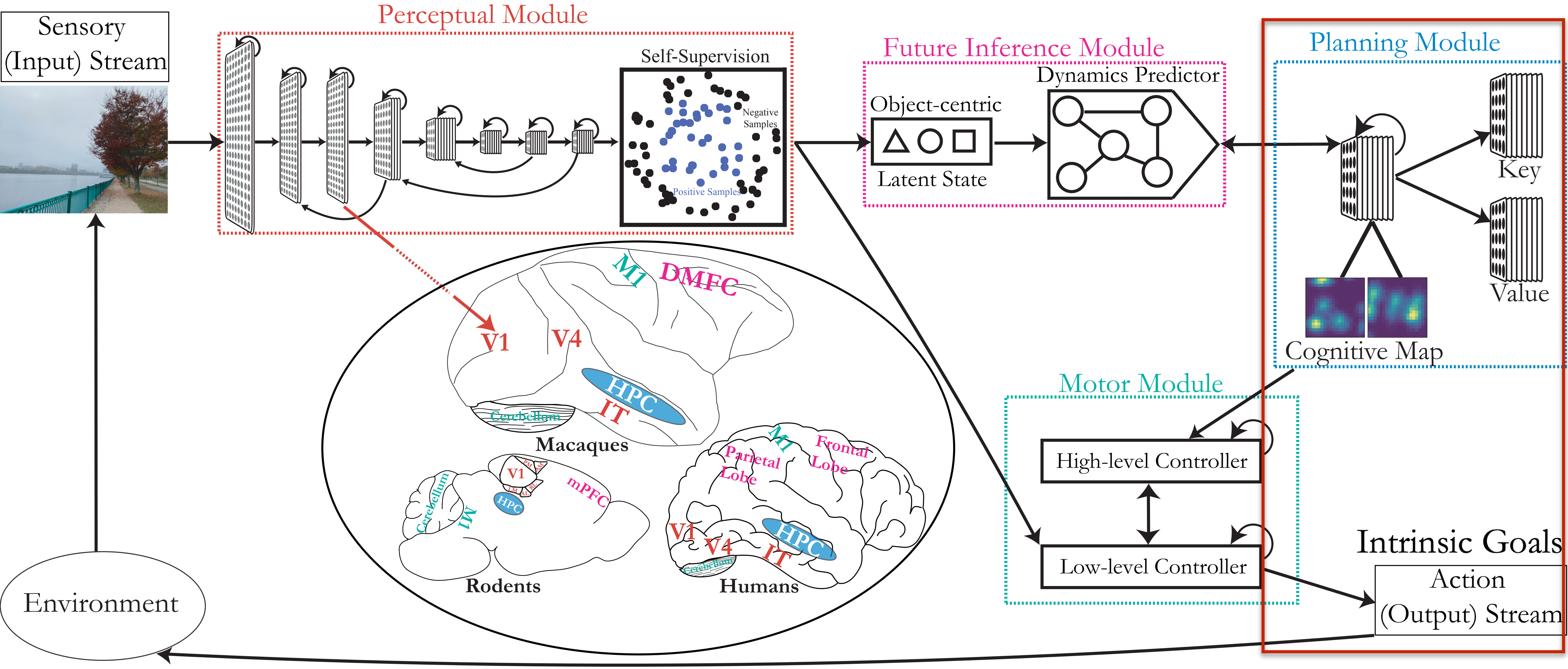
Recurrence + Contrastive SSL? Latent Future Prediction?



Roadmap: Planning & Action

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

Recurrence + Contrastive SSL? Latent Future Prediction?

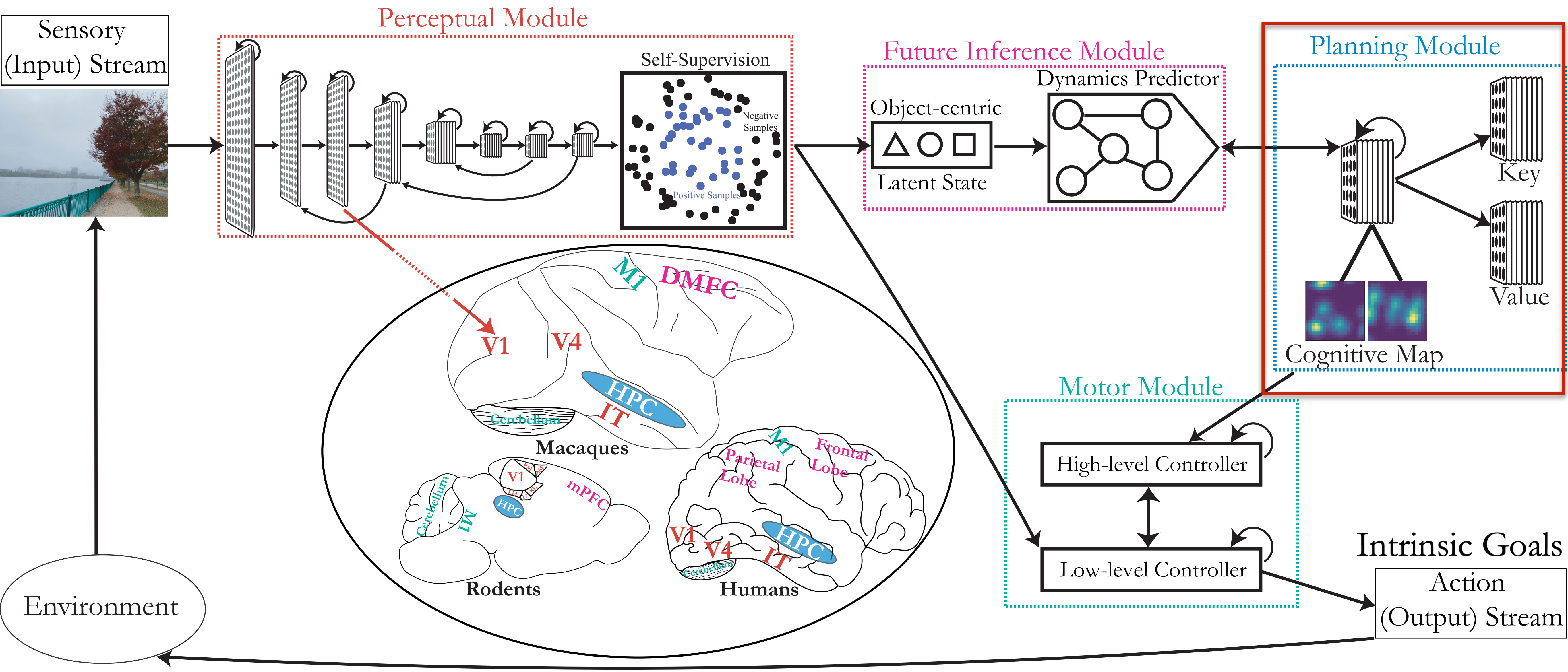


Roadmap: Planning

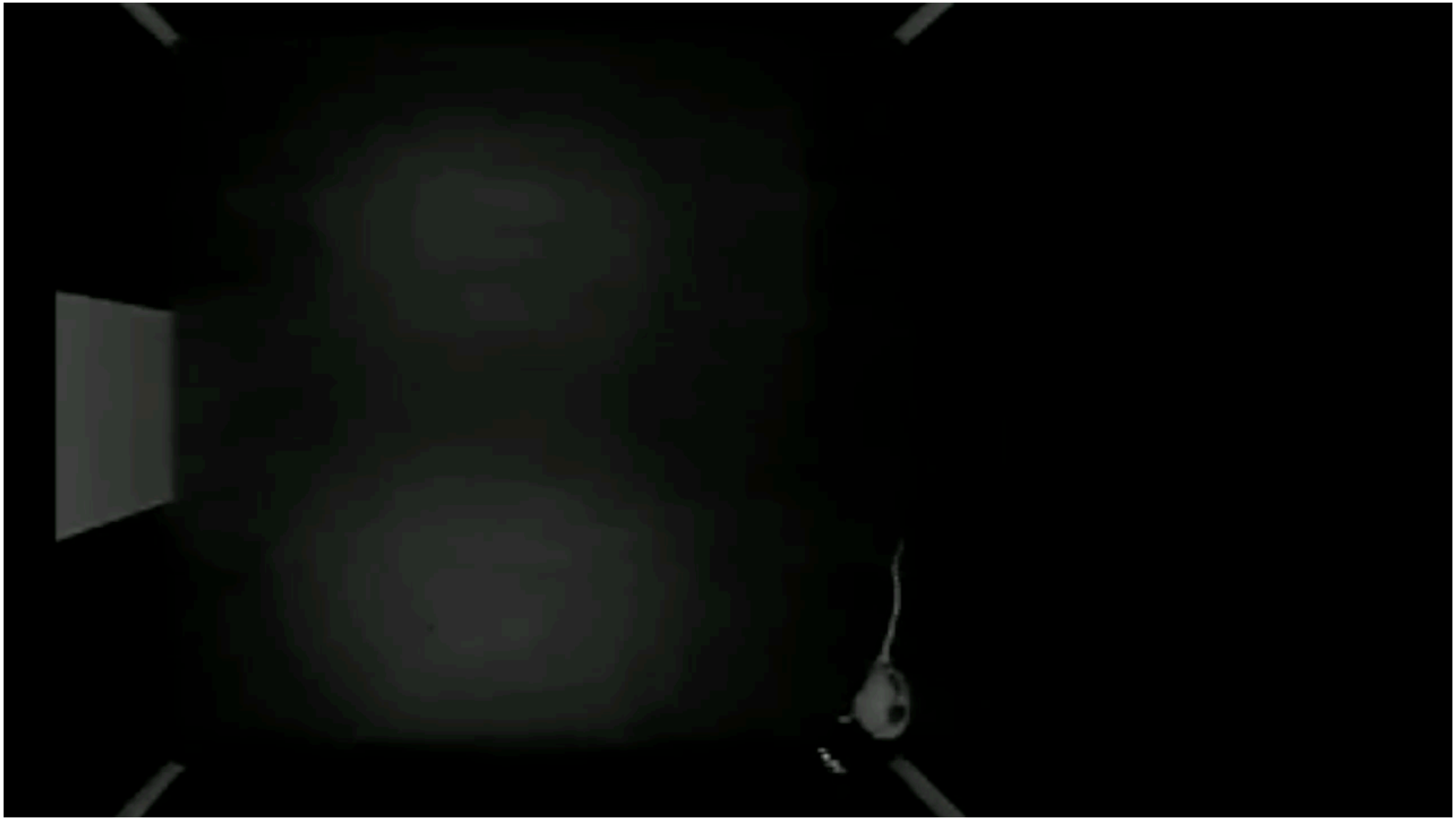
How does the brain *represent*, *predict*, *plan*, and enable *action*?

Recurrence + Contrastive SSL?

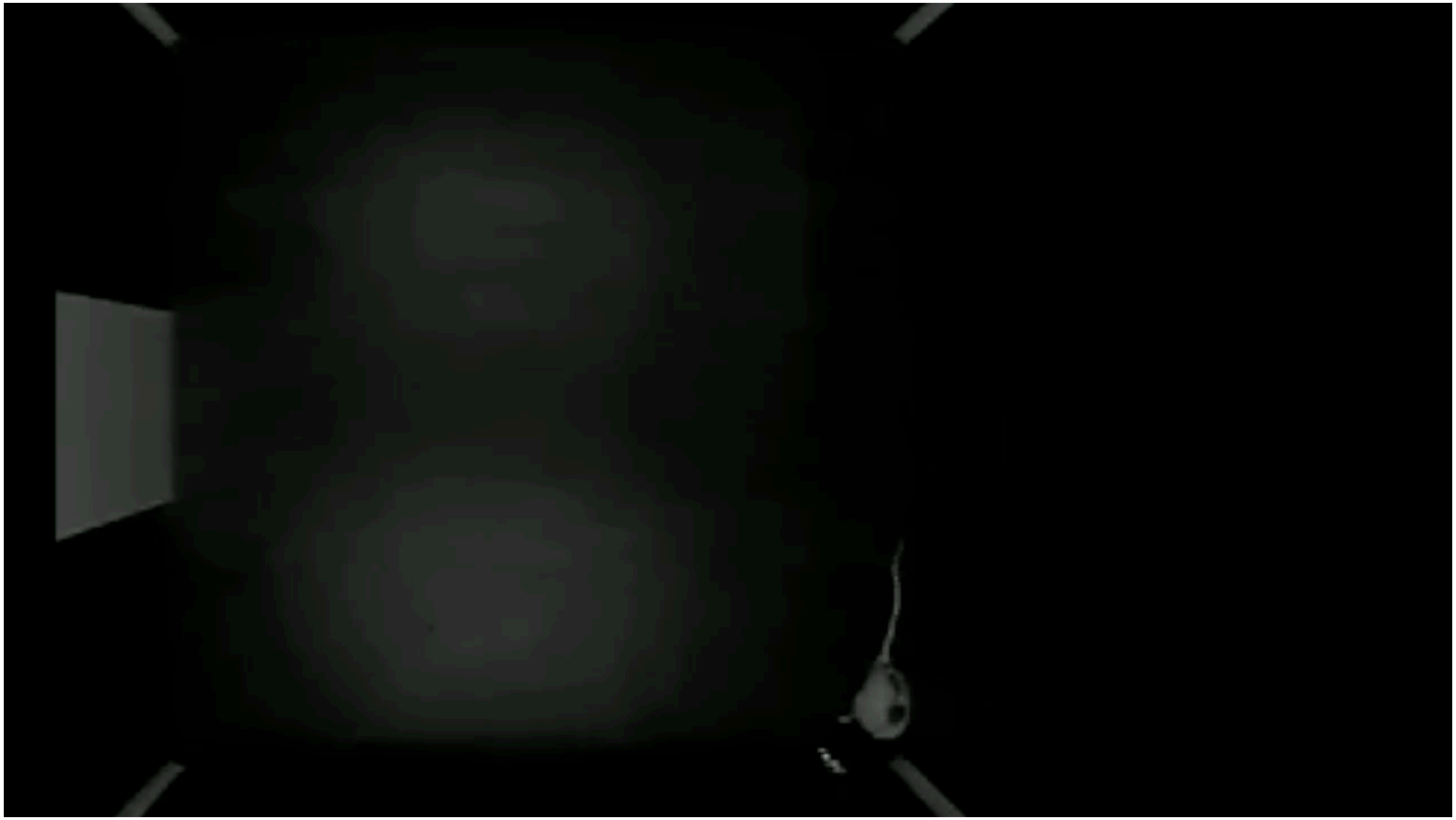
Latent Future Prediction?



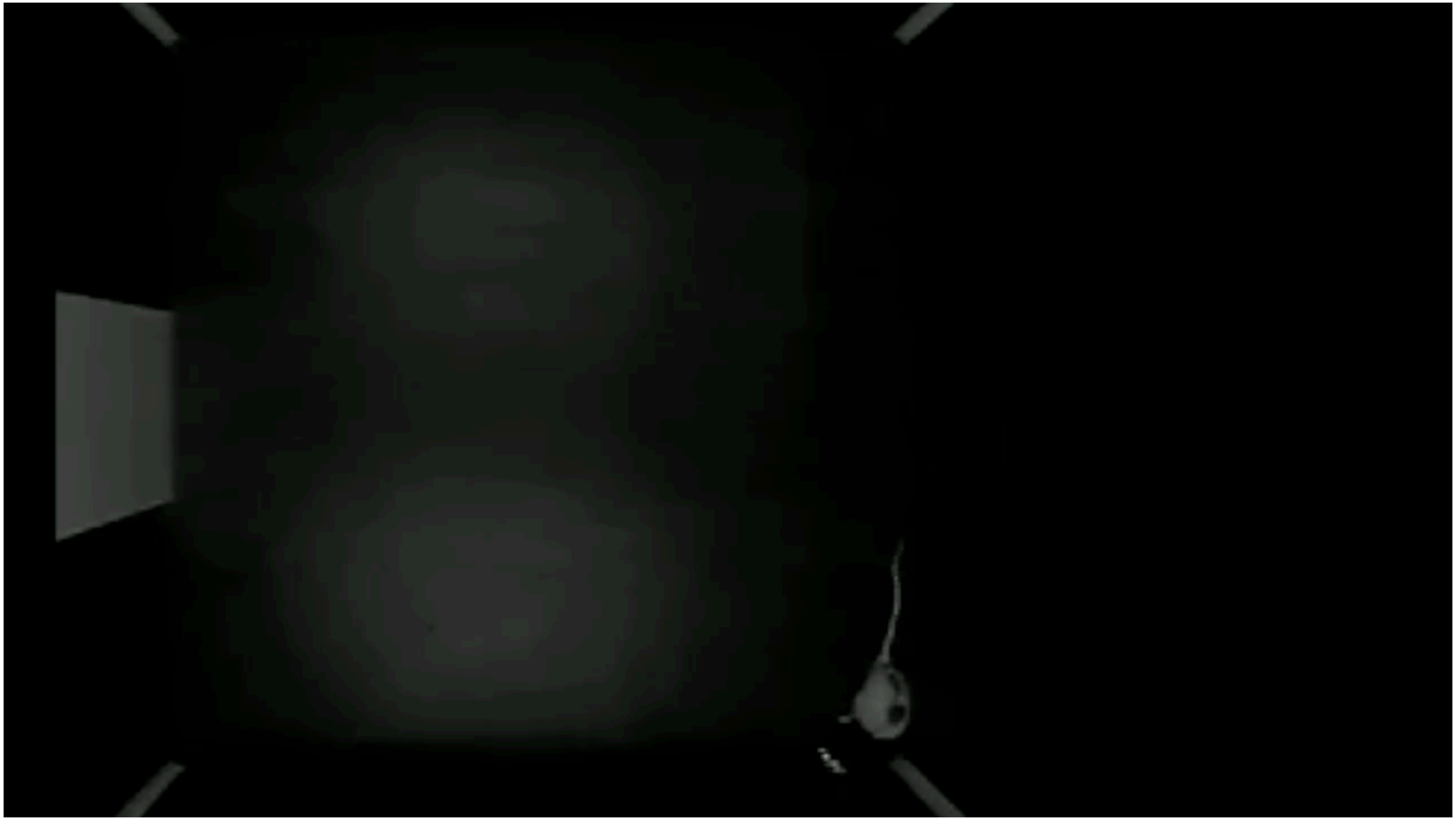
Hippocampal-Entorhinal Spatial Map



Hippocampal-Entorhinal Spatial Map



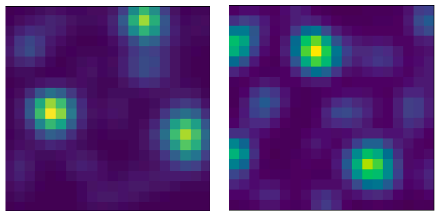
Hippocampal-Entorhinal Spatial Map



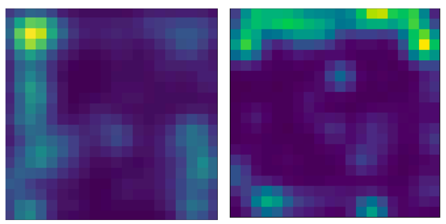
Place Cell
(Hippocampus)

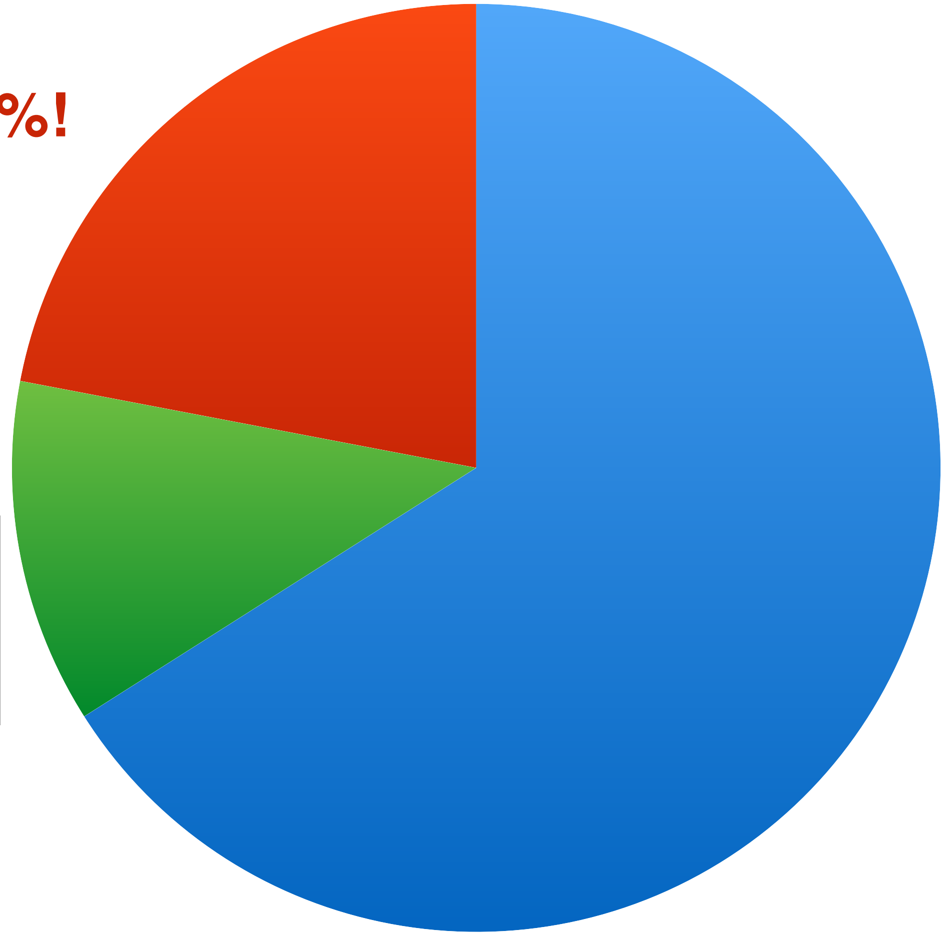
A Task-Optimized Account of Heterogeneity

Grid Cells
More like ~2-3%!

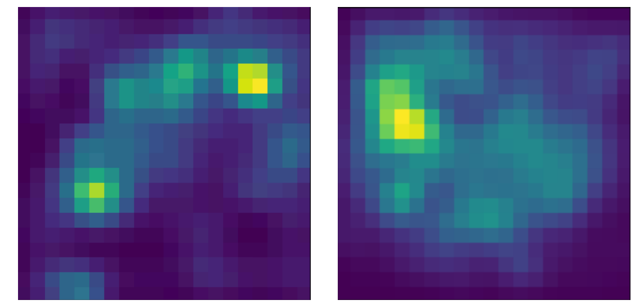


Border Cells

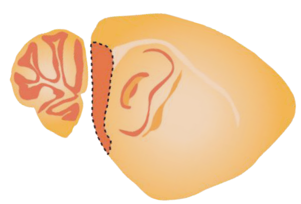




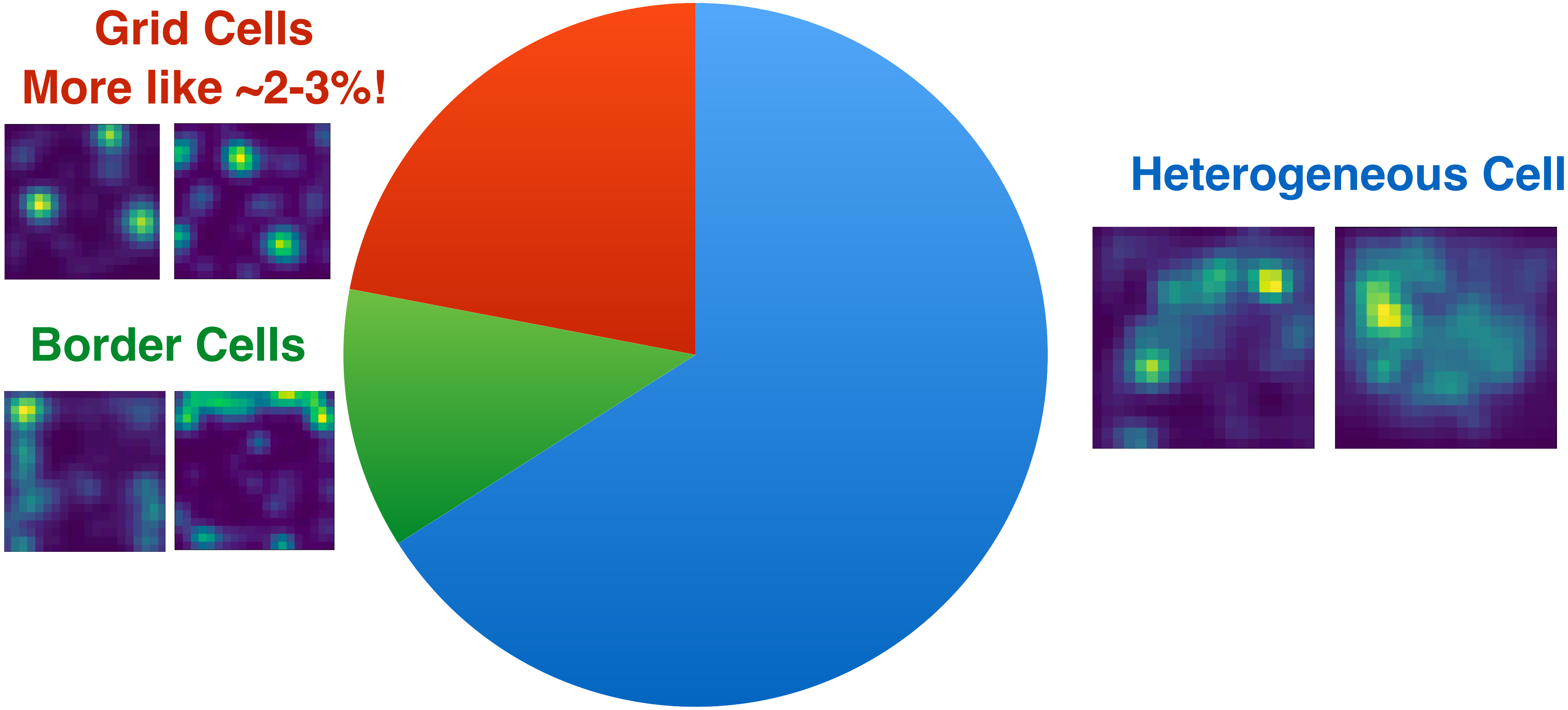
Heterogeneous Cells



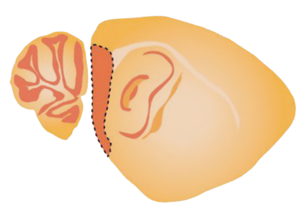
Data from: *Mallory et al. 2021*

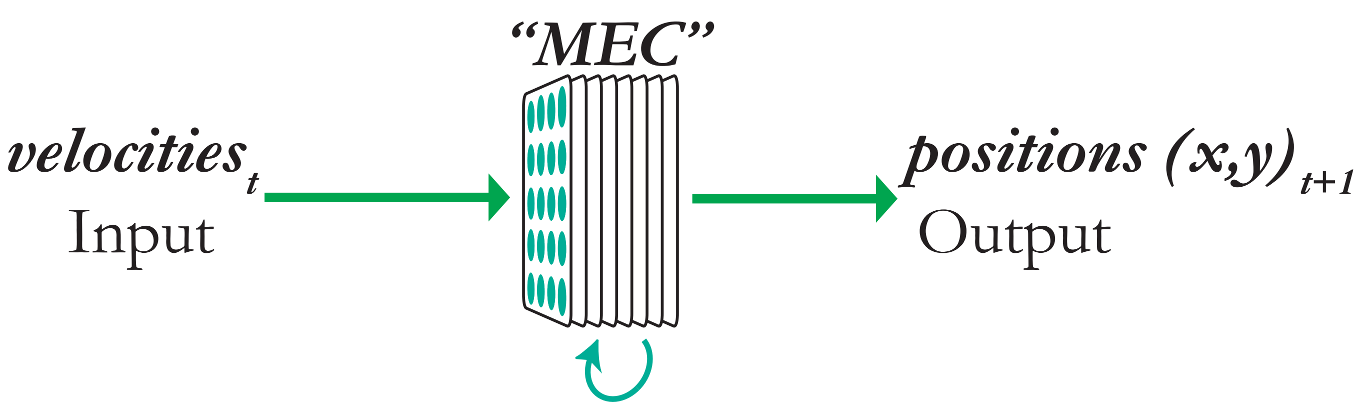


A Task-Optimized Account of Heterogeneity



Data from: *Mallory et al. 2021*



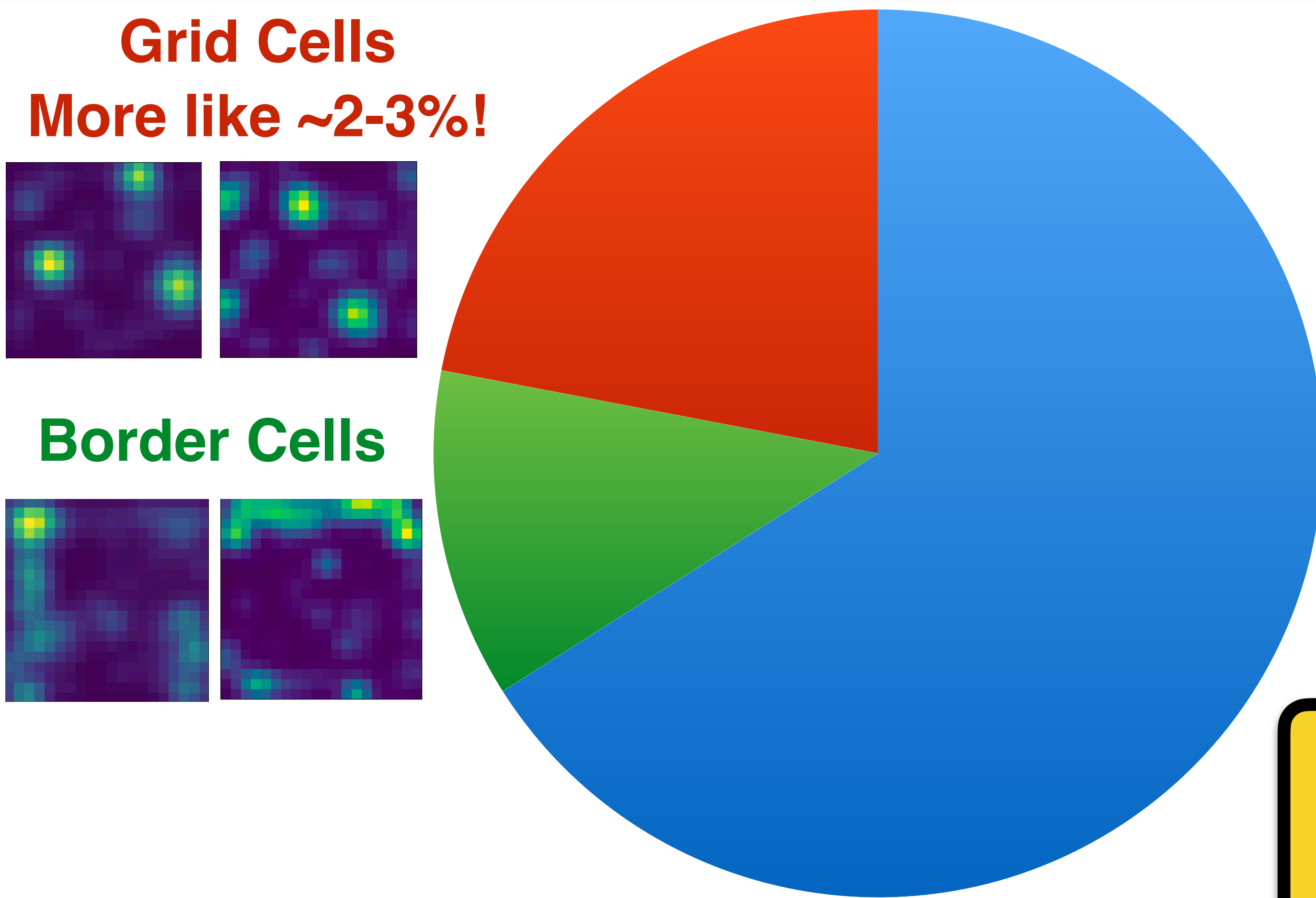


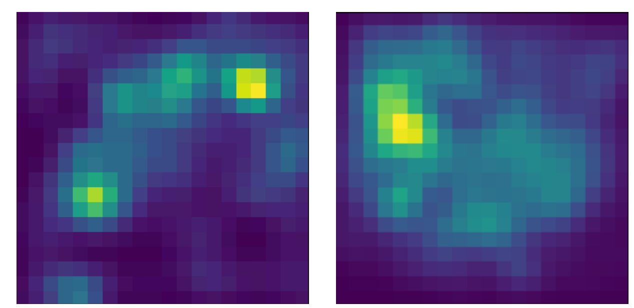
Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks

Aran Nayebi[1,*], Alexander Attinger[2], Malcolm G. Campbell[2], Kiah Hardcastle[2], Isabel I.C. Low[1,2,7], Caitlin S. Mallory[2], Gabriel C. Mel[1], Ben Sorscher[4], Alex H. Williams[6,7], Surya Ganguli[4,7,8], Lisa M. Giocomo[2,7], and Daniel L.K. Yamins[3,5,7]

NeurIPS 2021 (spotlight)

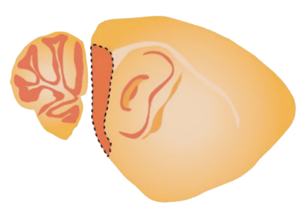
A Task-Optimized Account of Heterogeneity

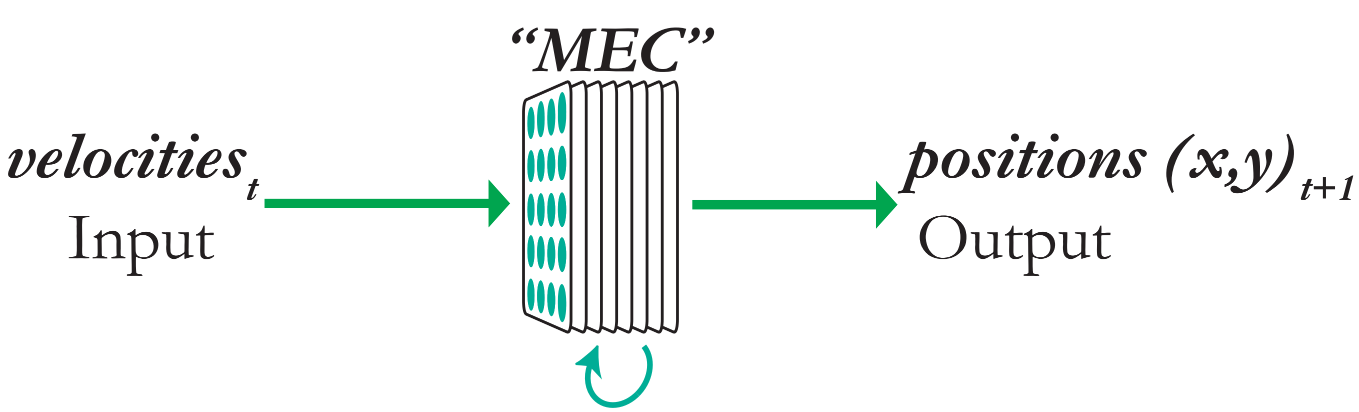




Heterogeneous cell types emerge in networks optimized for path integration!

Data from: *Mallory et al. 2021*



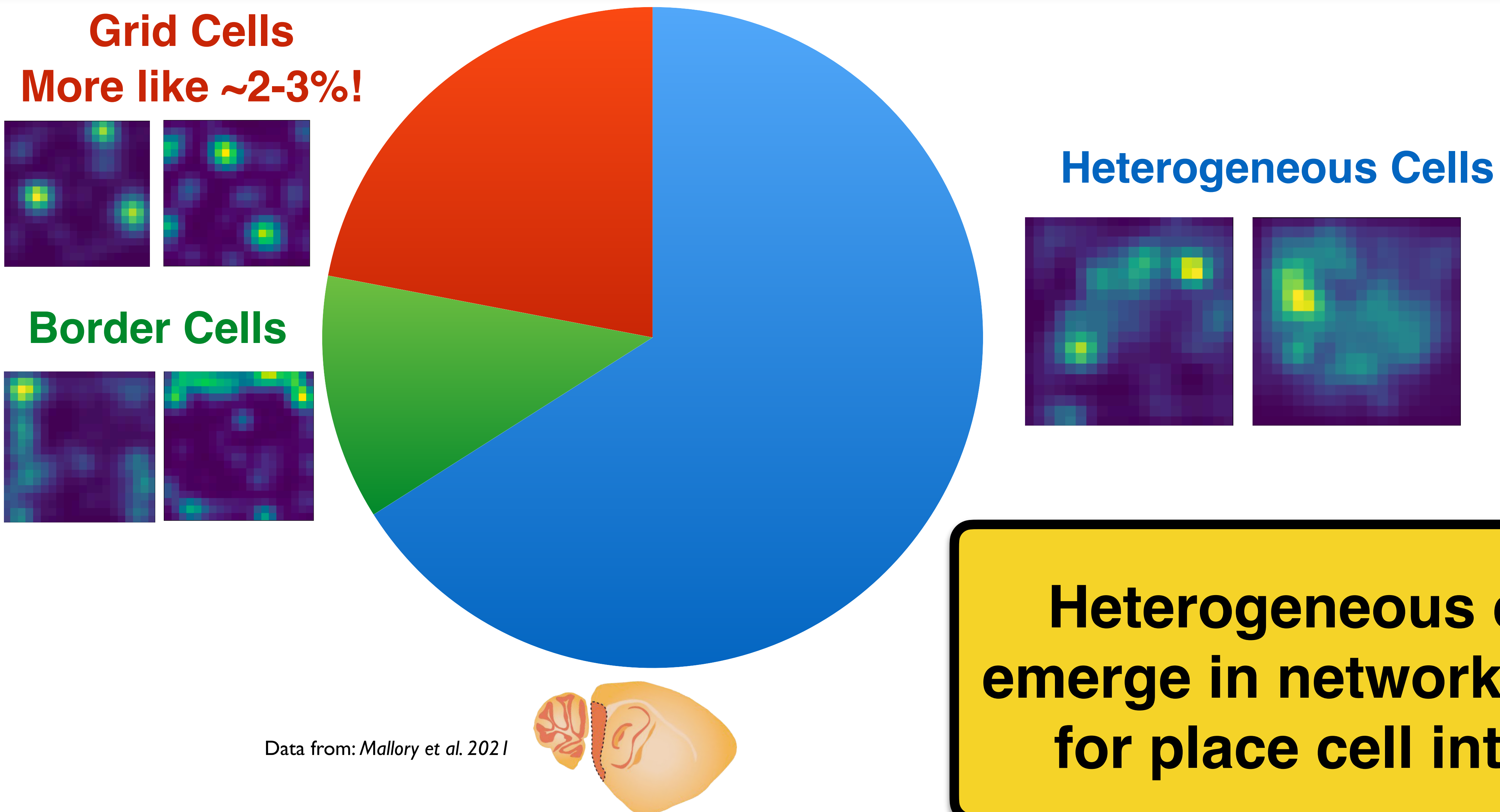


Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks

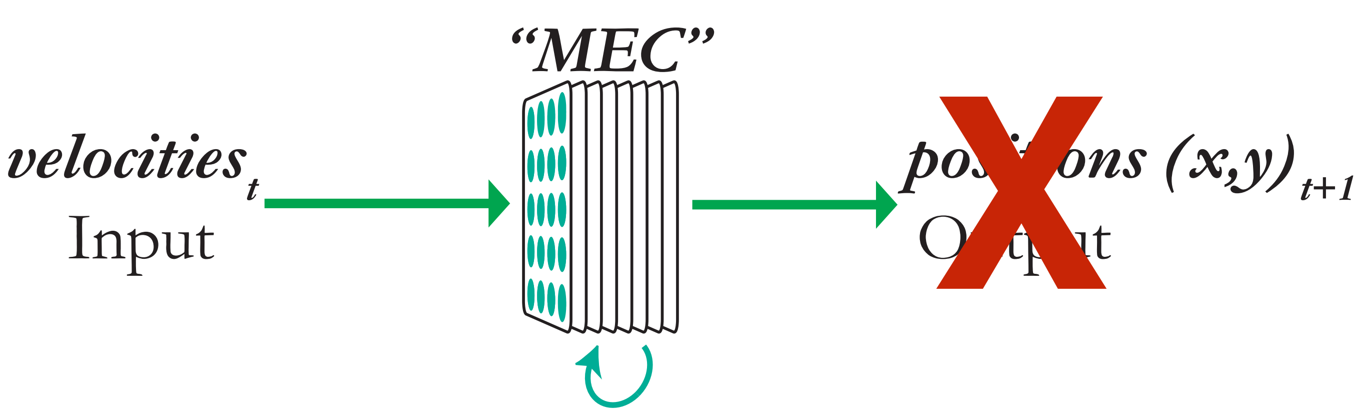
Aran Nayebi[1,*], Alexander Attinger[2], Malcolm G. Campbell[2], Kiah Hardcastle[2], Isabel I.C. Low[1,2,7], Caitlin S. Mallory[2], Gabriel C. Mel[1], Ben Sorscher[4], Alex H. Williams[6,7], Surya Ganguli[4,7,8], Lisa M. Giocomo[2,7], and Daniel L.K. Yamins[3,5,7]

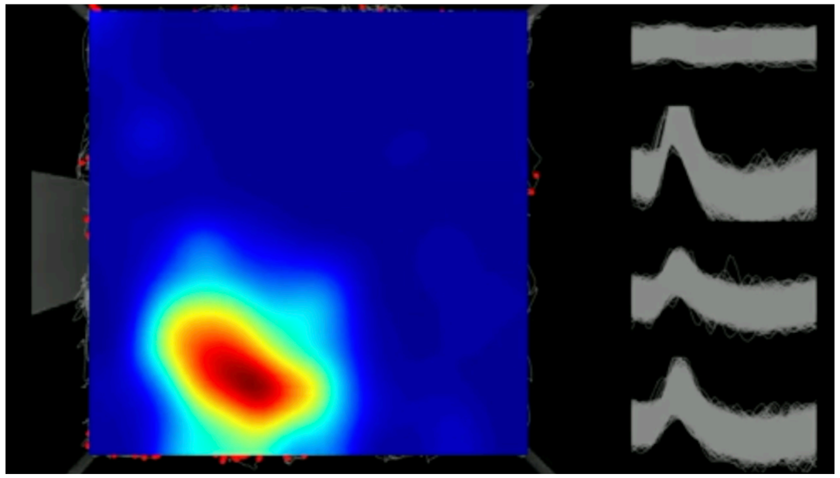
NeurIPS 2021 (spotlight)

A Task-Optimized Account of Heterogeneity



Heterogeneous cell types emerge in networks optimized for place cell integration!



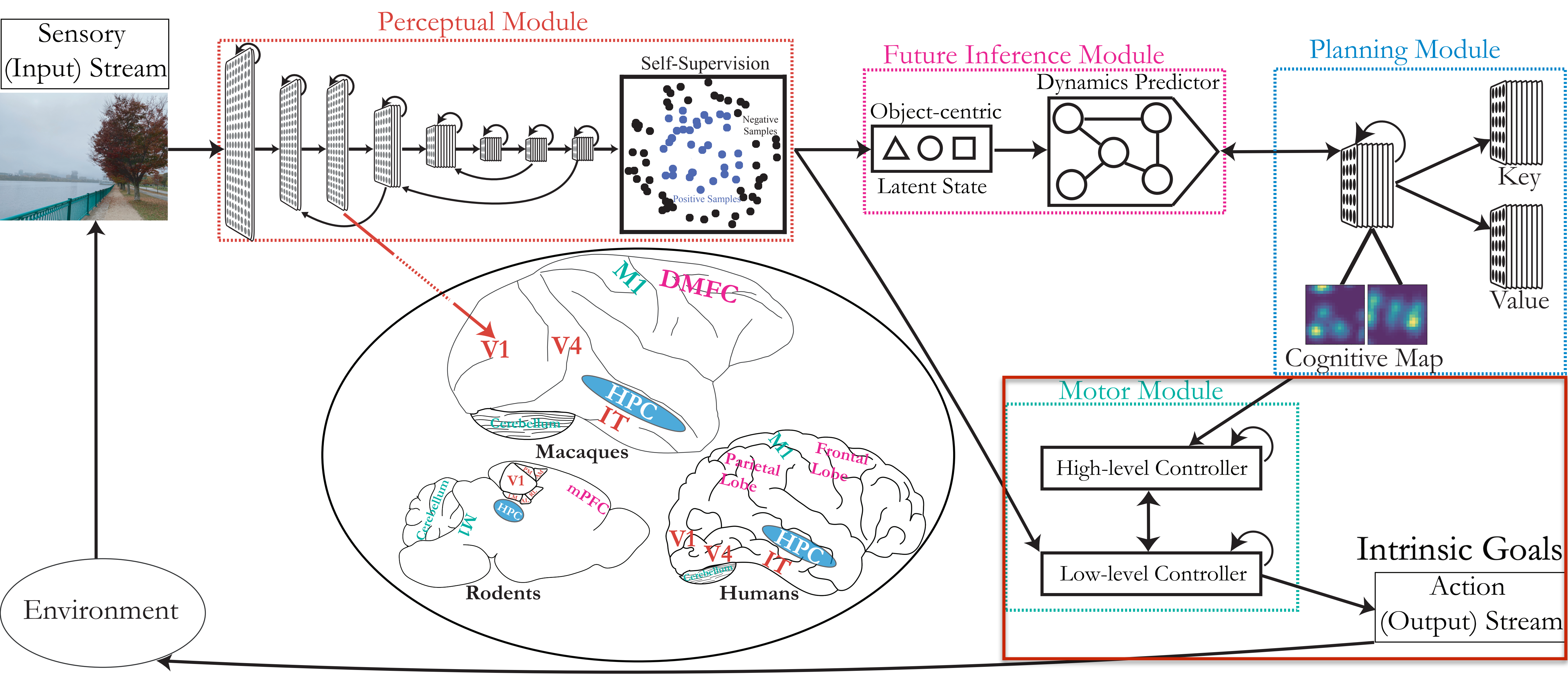


Place Cell (Hippocampus)

Roadmap: Action

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

Recurrence + Contrastive SSL? Latent Future Prediction?



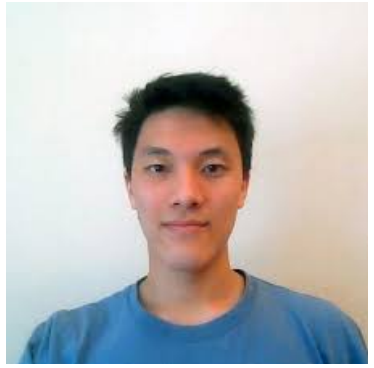
Autonomous Behavior and Whole-Brain Dynamics Emerge in Embodied Zebrafish Agents with Model-based Intrinsic Motivation

Reece Keller[1,2,*] Alyn Tornell[2] Felix Pei[2] Xaq Pitkow[1,3]
Leo Kozachkov[4,†] Aran Nayebi[3,1,2,†]

To appear at NeurIPS 2025!

Reece Keller

Alyn Tornell



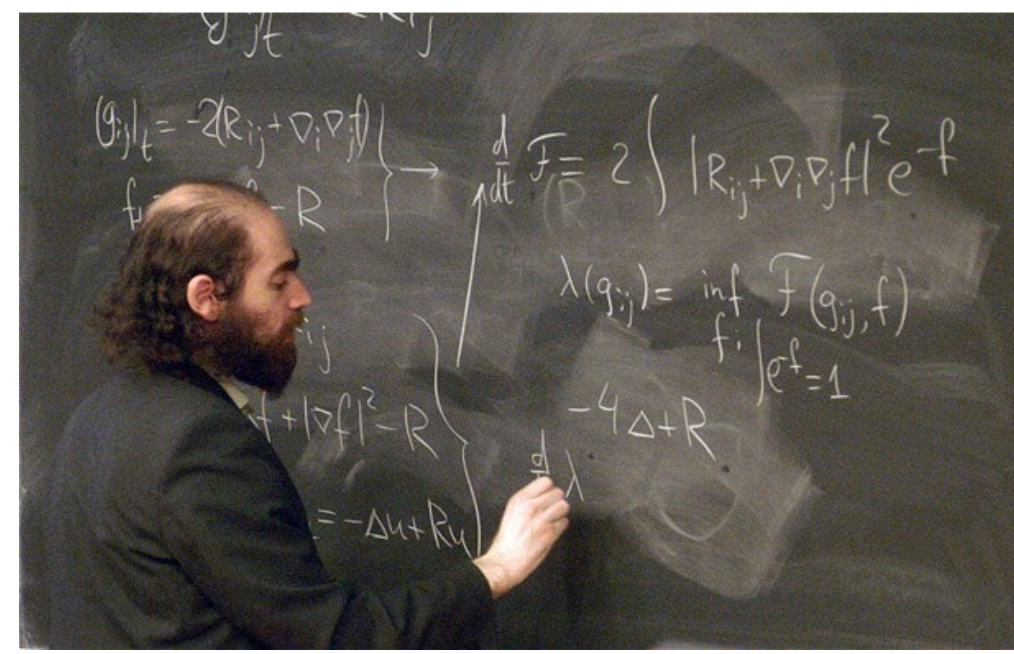
Felix Pei

Xaq Pitkow

Leo Kozachkov†

Why is Animal Autonomy Hard?

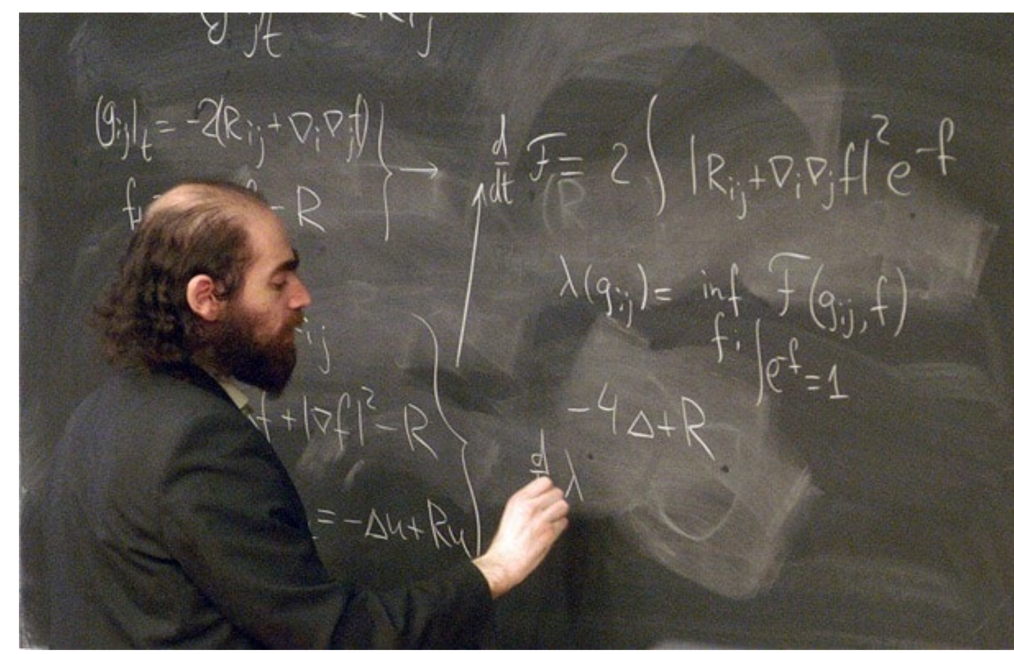
The behavioral repertoire is enormous...



Slides credit: Reece Keller

Why is Animal Autonomy Hard?

The behavioral repertoire is enormous...

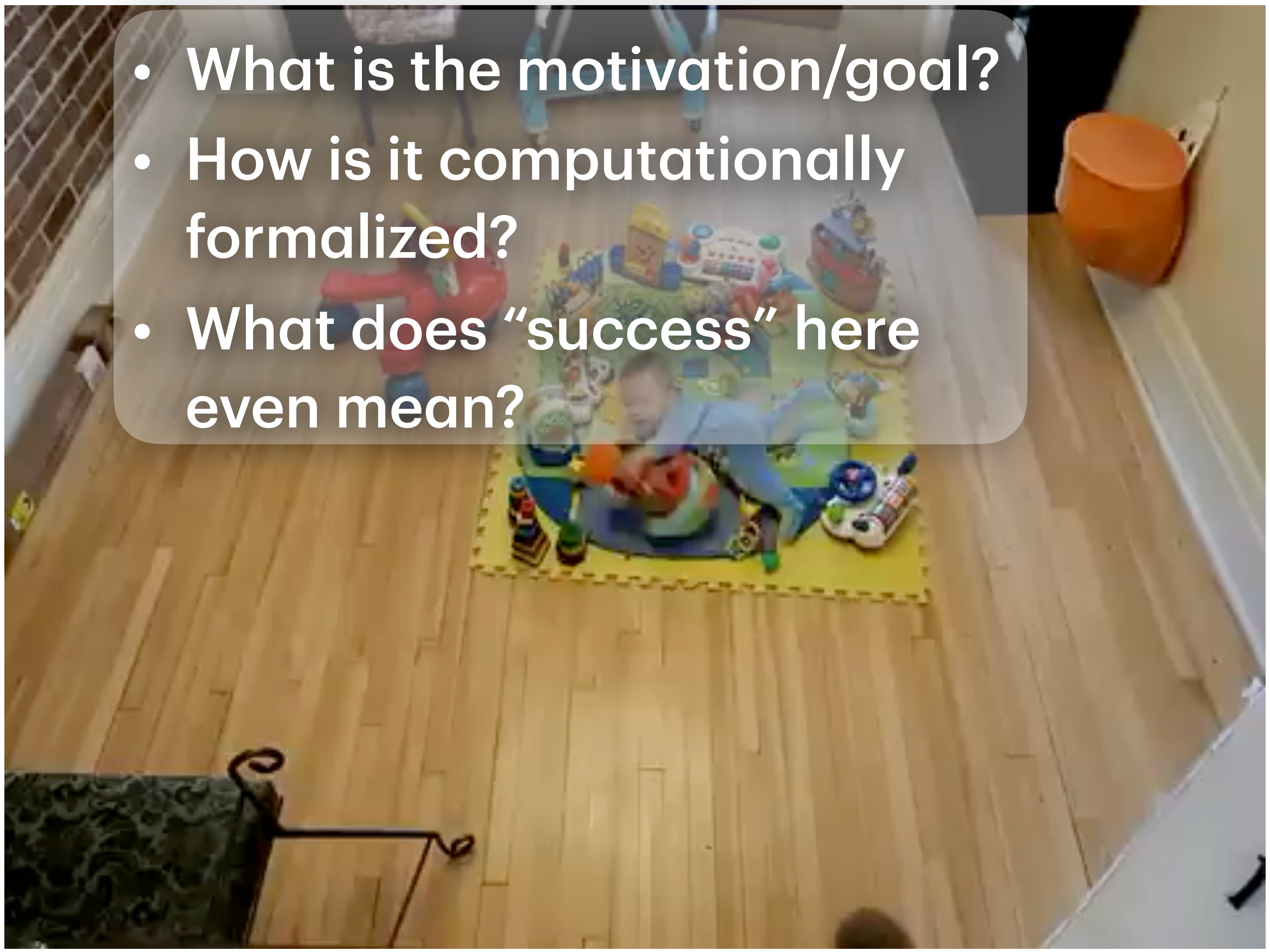


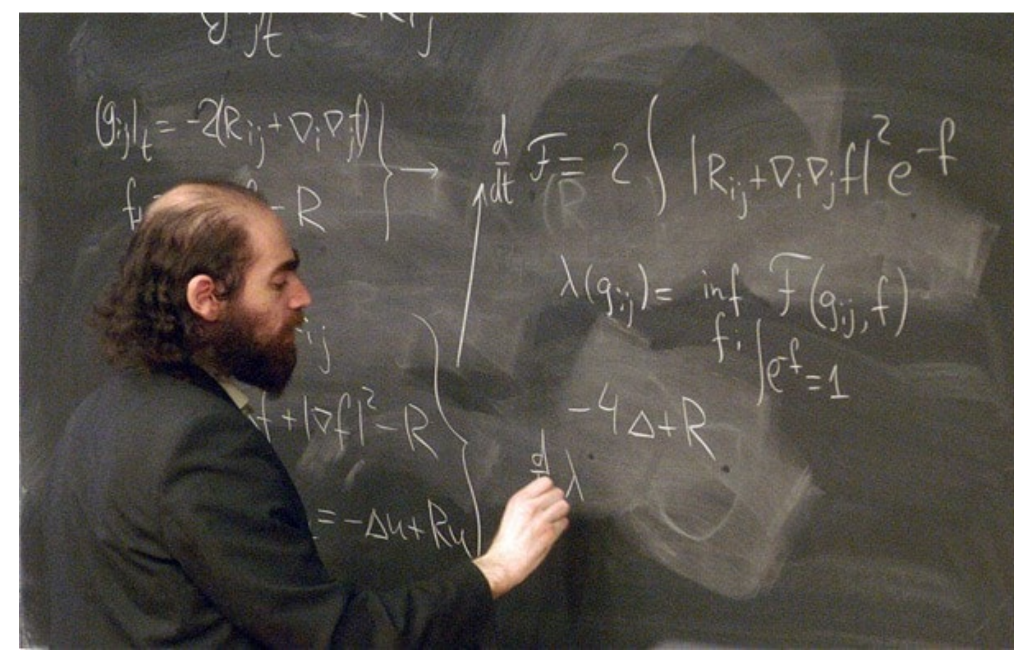
Slides credit: Reece Keller

Why is Animal Autonomy Hard?

The behavioral repertoire is enormous…

- What is the motivation/goal?
- How is it computationally formalized?
- What does "success" here even mean?

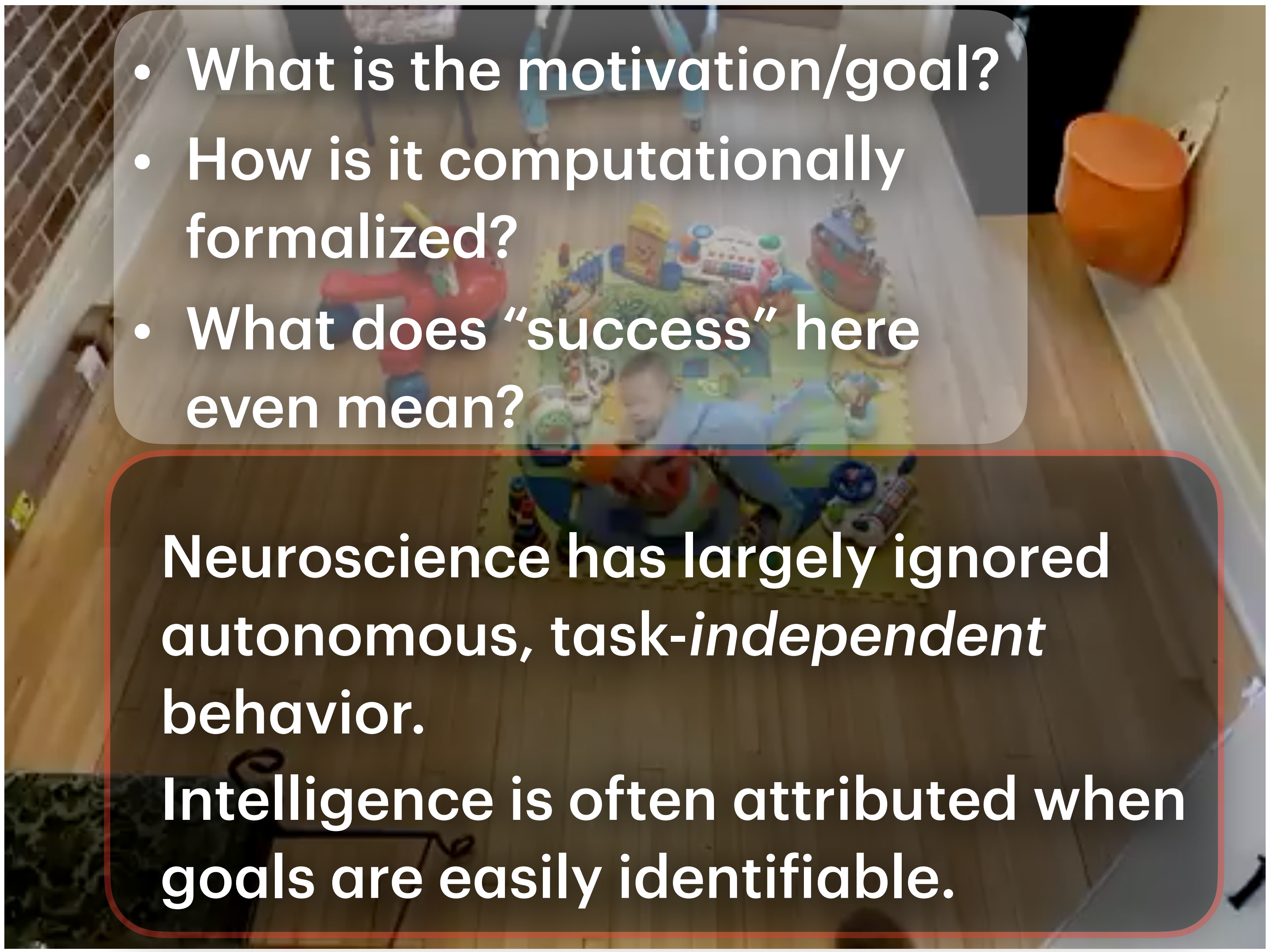




Slides credit: Reece Keller

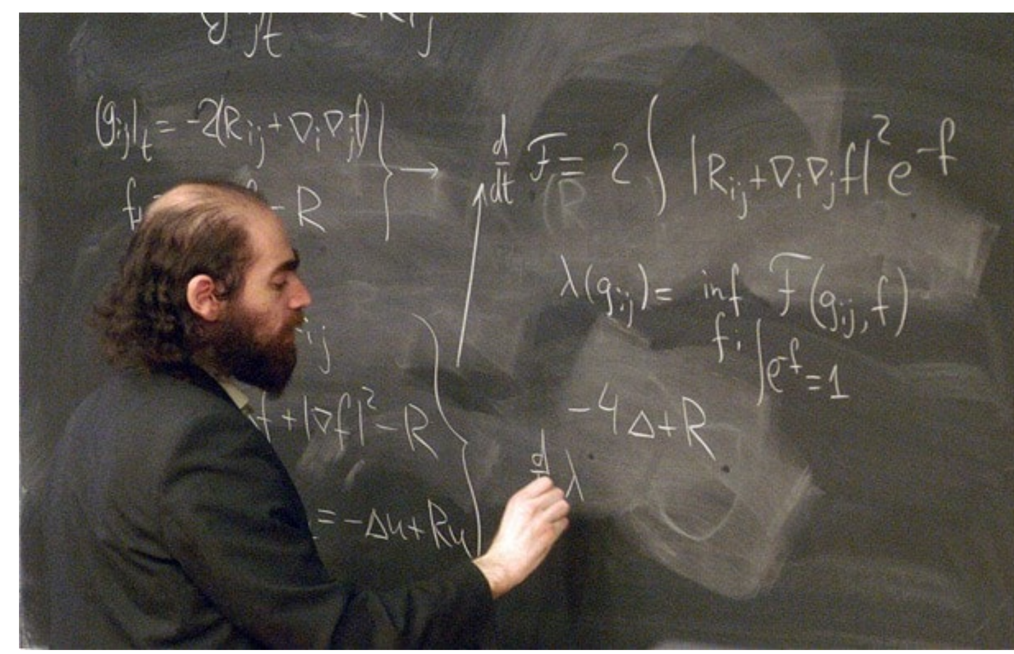
Why is Animal Autonomy Hard?

The behavioral repertoire is enormous…



- What is the motivation/goal?
- How is it computationally formalized?
- What does “success” here even mean?

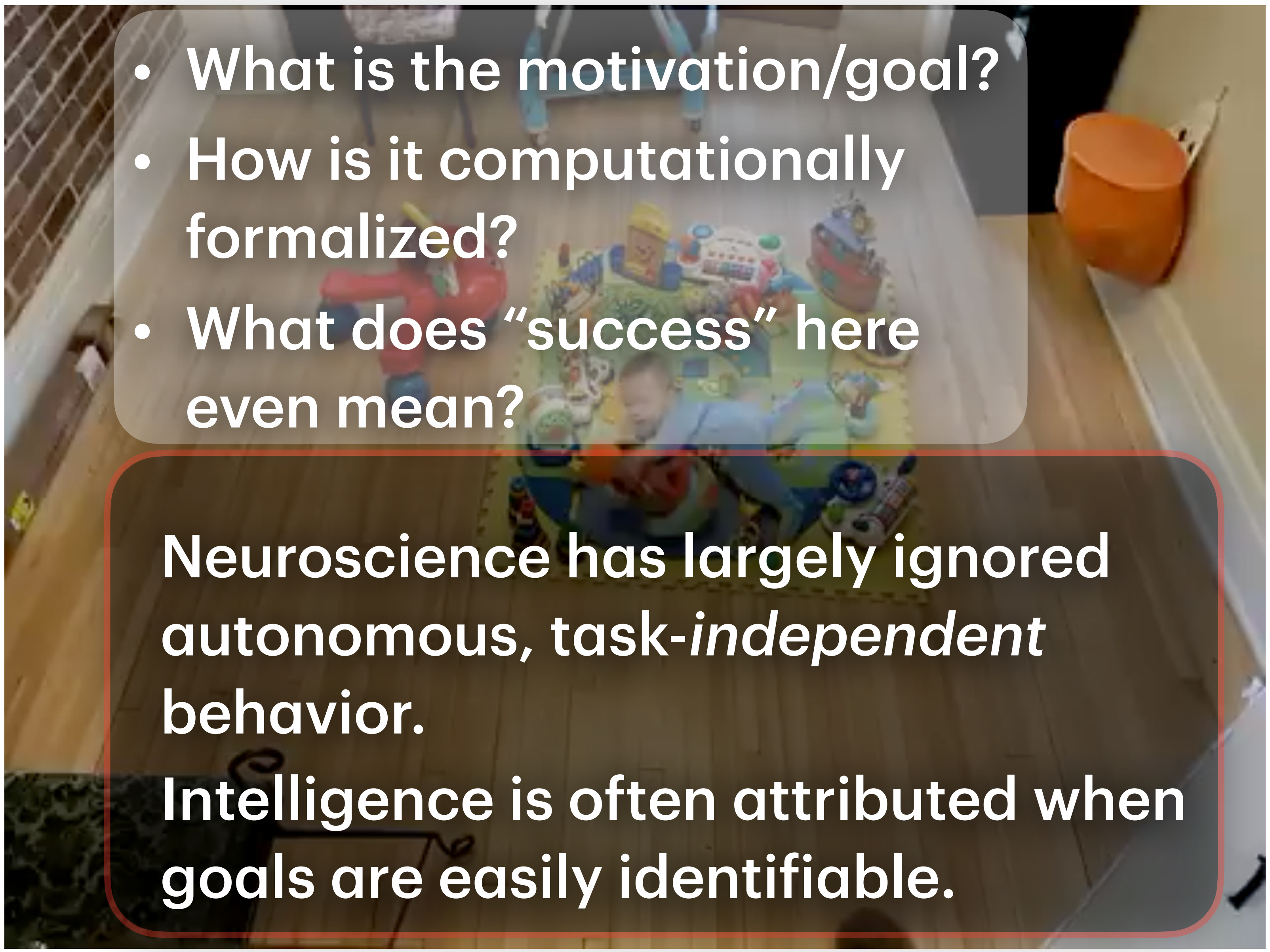
Neuroscience has largely ignored autonomous, task-*independent* behavior.
Intelligence is often attributed when goals are easily identifiable.



Slides credit: Reece Keller

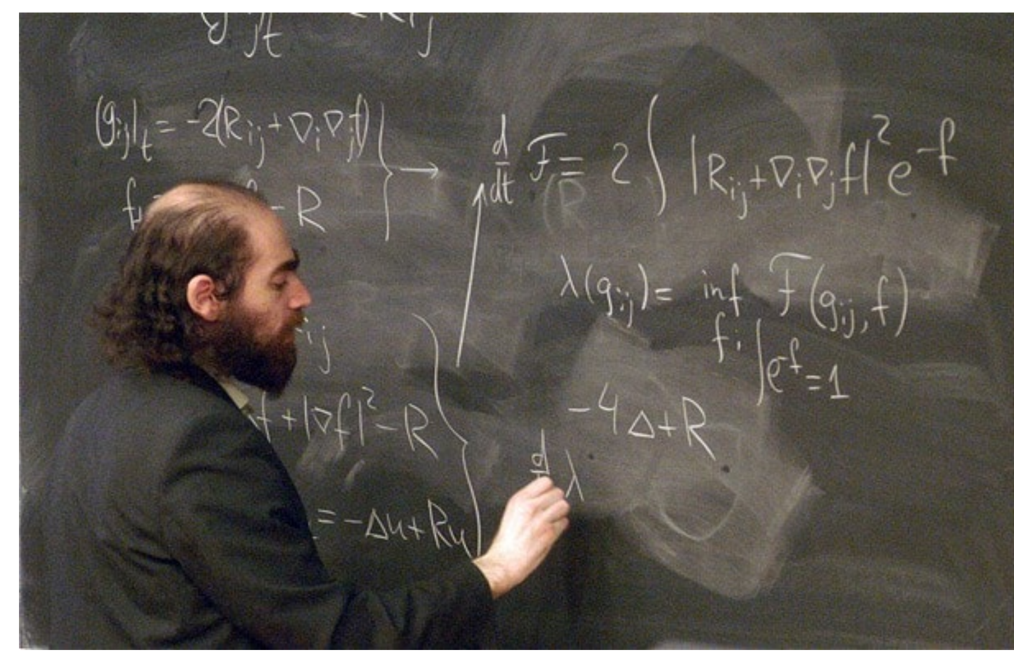
Why is Animal Autonomy Hard?

The behavioral repertoire is enormous…



- What is the motivation/goal?
- How is it computationally formalized?
- What does "success" here even mean?

Neuroscience has largely ignored autonomous, task-*independent* behavior.
Intelligence is often attributed when goals are easily identifiable.



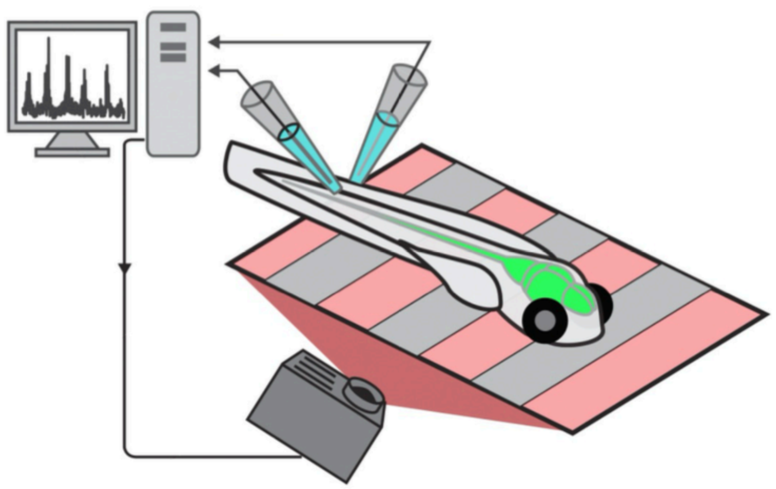
Unlike games where RL has succeeded, the environment doesn't have a dense reward function. It must be (somehow) *internally* generated by the organism!

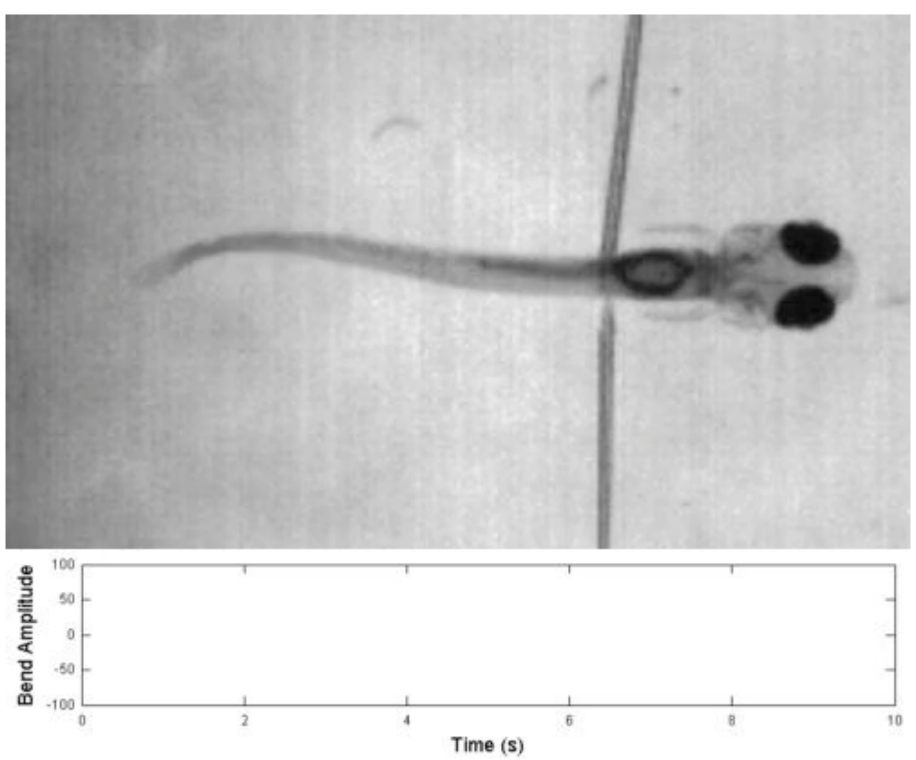
Slides credit: Reece Keller

Glia Accumulate Evidence that Actions Are Futile and Suppress Unsuccessful Behavior

Yu Mu,[1,4,*] Davis V. Bennett,[1,2,4] Mikail Rubinov,[1,3,4] Sujatha Narayan,[1] Chao-Tsung Yang,[1] Masashi Tanimoto,[1] Brett D. Mensh,[1] Loren L. Looger,[1] and Misha B. Ahrens[1,5,*]

virtual reality navigation



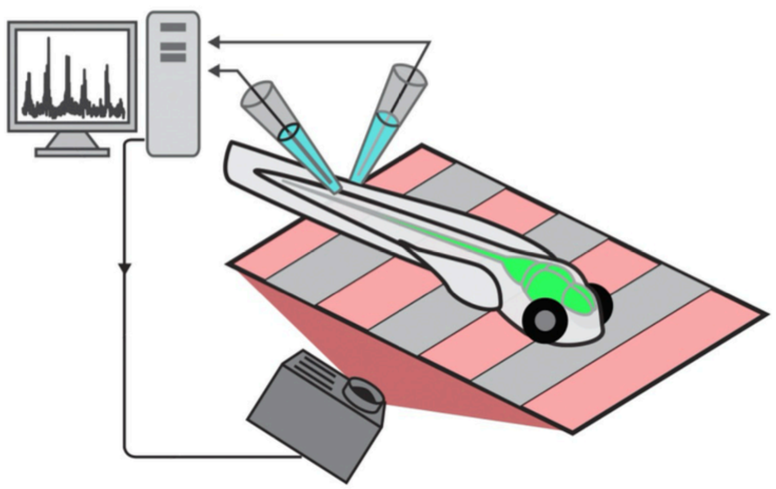


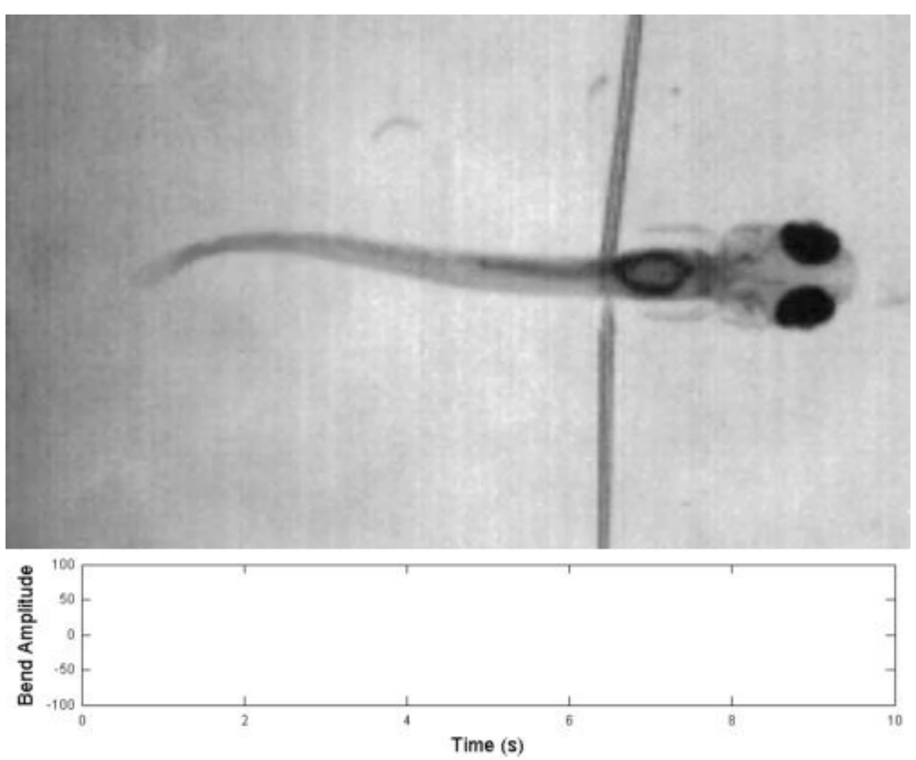
Mu et al., *Cell* (2019)

Glia Accumulate Evidence that Actions Are Futile and Suppress Unsuccessful Behavior

Yu Mu,[1,4,*] Davis V. Bennett,[1,2,4] Mikail Rubinov,[1,3,4] Sujatha Narayan,[1] Chao-Tsung Yang,[1] Masashi Tanimoto,[1] Brett D. Mensh,[1] Loren L. Looger,[1] and Misha B. Ahrens[1,5,*]

virtual reality navigation



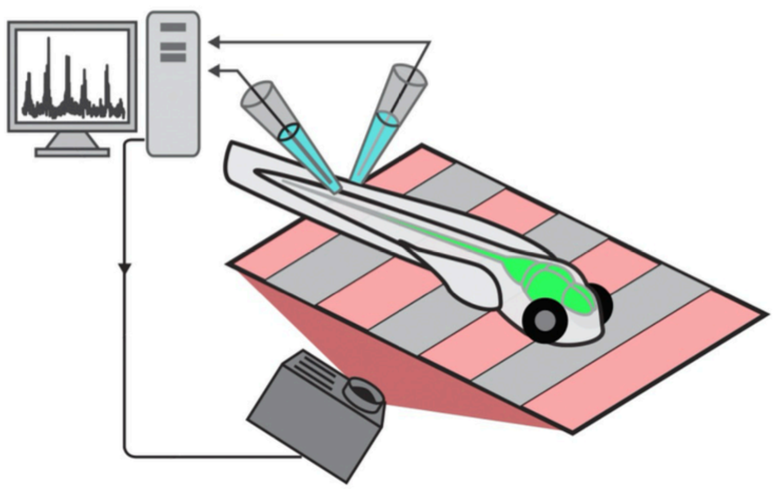


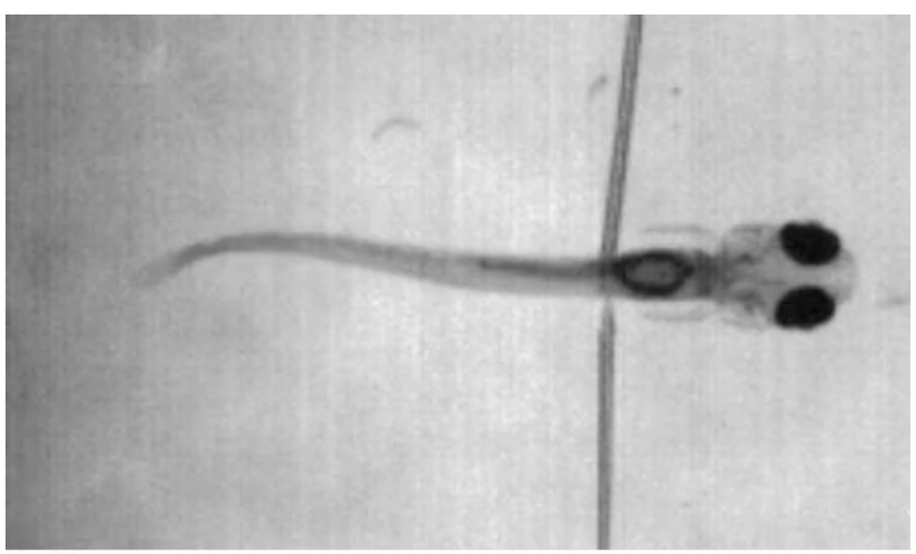
Mu et al., *Cell* (2019)

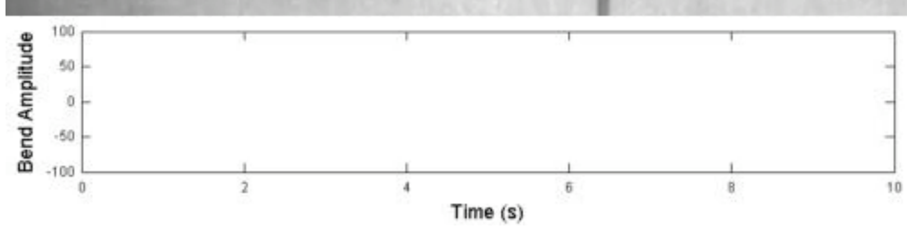
Glia Accumulate Evidence that Actions Are Futile and Suppress Unsuccessful Behavior

Yu Mu,[1,4,*] Davis V. Bennett,[1,2,4] Mikail Rubinov,[1,3,4] Sujatha Narayan,[1] Chao-Tsung Yang,[1] Masashi Tanimoto,[1] Brett D. Mensh,[1] Loren L. Looger,[1] and Misha B. Ahrens[1,5,*]

virtual reality navigation







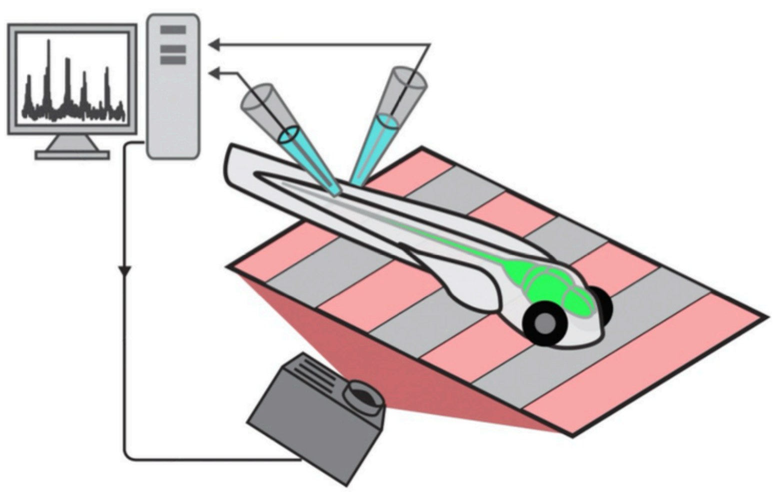
Mu et al., *Cell* (2019)

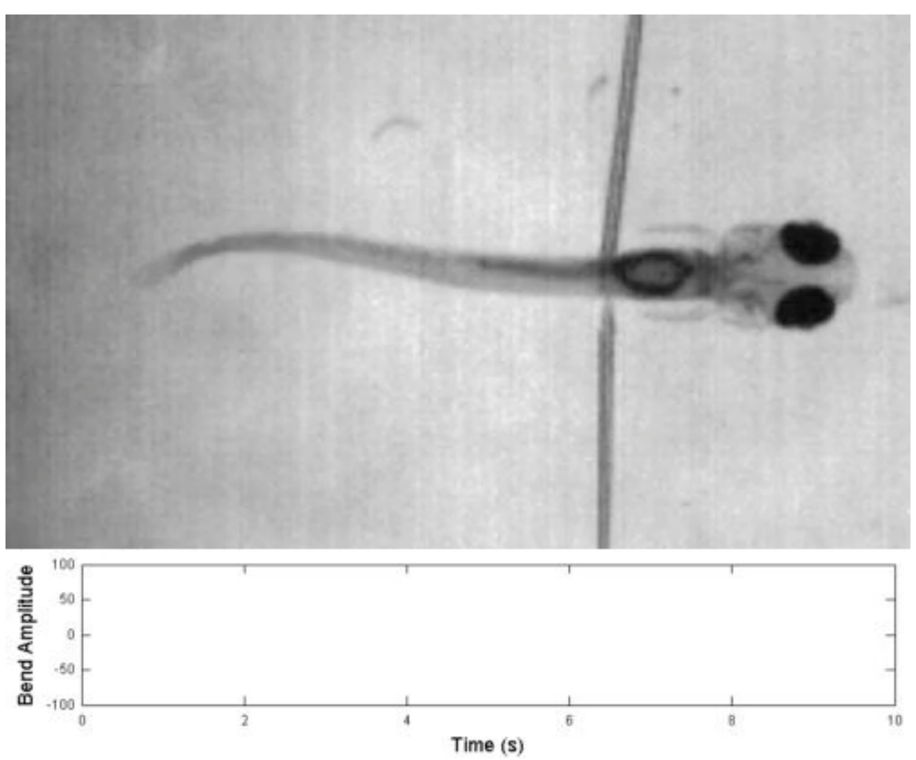
Glia Accumulate Evidence that Actions Are Futile and Suppress Unsuccessful Behavior

Yu Mu,[1,4,*] Davis V. Bennett,[1,2,4] Mikail Rubinov,[1,3,4] Sujatha Narayan,[1] Chao-Tsung Yang,[1] Masashi Tanimoto,[1] Brett D. Mensh,[1] Loren L. Looger,[1] and Misha B. Ahrens[1,5,*]

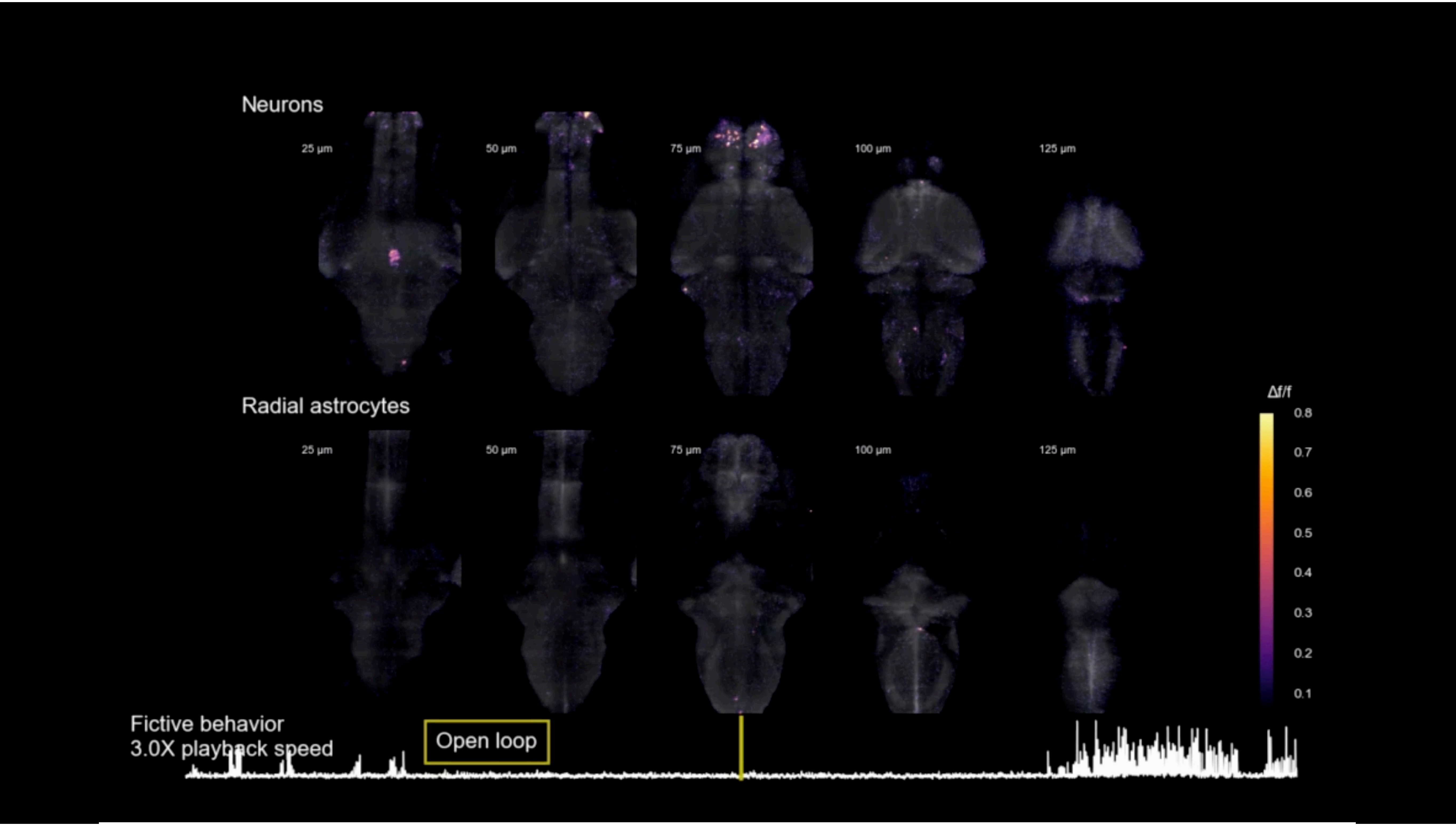
- ✓ 1. Ecologically-relevant environment
- ✓ 2. “Cognitive” states with clear behavioral readouts
- ✓ 3. Large-scale multi-area neural recordings

virtual reality navigation

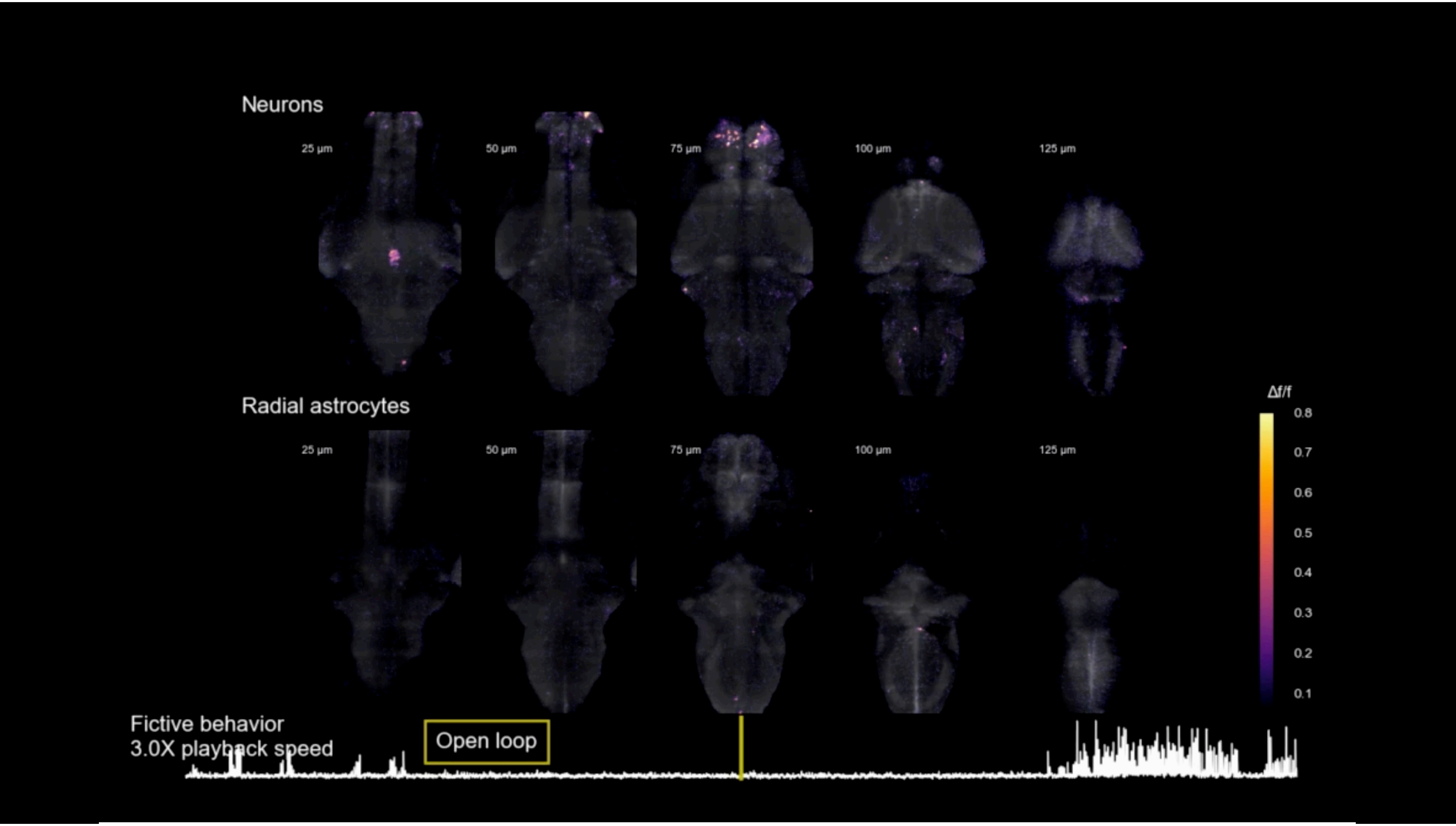




Mu et al., *Cell* (2019)



Neurons
25 µm
50 µm
75 µm
100 µm
125 µm
Radial astrocytes
25 µm
50 µm
75 µm
100 µm
125 µm
Δf/f
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Fictive behavior
3.0X playback speed
Open loop



Neurons
25 µm
50 µm
75 µm
100 µm
125 µm
Radial astrocytes
25 µm
50 µm
75 µm
100 µm
125 µm
Δf/f
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Fictive behavior
3.0X playback speed
Open loop

Machine Autonomy: Prior Work

Exploration in sparse/reward-free environments

Curiosity type	**Formulation**	**What it measures**
Surprise	$r_t^i \propto -\log \omega_\theta(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$	prediction error
Disagreement	$r_t^i \propto \text{Var}\left(\{\omega_{\theta_j}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})\}_{j=1}^N\right)$	prediction variance
Learning progress	$r_t^i \propto \log \frac{\omega_{\theta'}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})}{\omega_\theta(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})}$ $\theta \leftarrow (1-\gamma)\theta + \gamma\theta'$	prediction error gain

Kim et al., *ICML* (2020)
Pathak et al., *ICML* (2017)
Burda, Edwards & Pathak et al., *NeurIPS* (2017)

Machine Autonomy: Prior Work

vironments

		What it measures
	$\mathbf{a})$	prediction error
	$\mathbf{s}, \mathbf{a})\}_{j=1}^{N}\big)$	prediction variance
Learning progress	$r_t^i \propto \log \frac{\cdots)}{\omega_\theta(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})}$ $\theta \leftarrow (1-\gamma)\theta + \gamma\theta'$	prediction error gain

Kim et al., *ICML* (2020)
Pathak et al., *ICML* (2017)
Burda, Edwards & Pathak et al., *NeurIPS* (2017)

Machine Autonomy: Prior Work

$\mathbf{s}, \mathbf{a})\}_{j=1}^{N} \big)$ prediction variance

Learning progress $r_t^i \propto \log \frac{\ldots}{\omega_\theta(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})}$ prediction error gain

$$\theta \leftarrow (1-\gamma)\theta + \gamma\theta'$$

Kim et al., *ICML* (2020)
Pathak et al., *ICML* (2017)
Burda, Edwards & Pathak et al., *NeurIPS* (2017)

Machine Autonomy: Prior Work

prediction

Learning progress $r_t^i \propto \log \frac{\ldots}{\omega_\theta(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})}$

$$\theta \leftarrow (1 - \gamma)\theta + \gamma\theta'$$

Kim et al., *ICML* (2020)
Pathak et al., *ICML* (2017)
Burda, Edwards & Pathak et al., *NeurIPS* (2017)

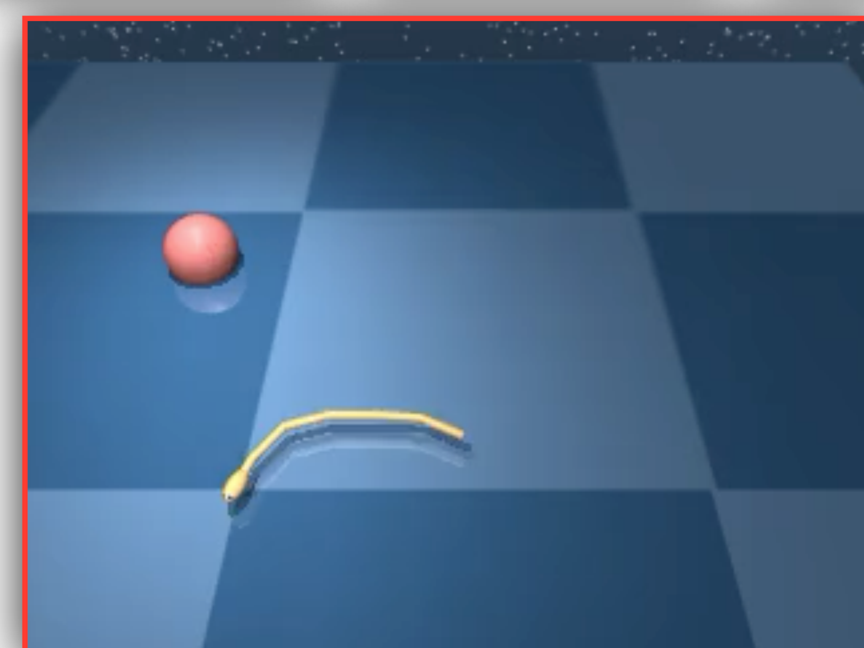
Machine Autonomy: Prior Work

prediction

Learning progress

$$r_t^i \propto \log \frac{\cdots}{\omega_\theta(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})}$$

$$\theta \leftarrow (1-\gamma)\theta + \gamma\theta'$$

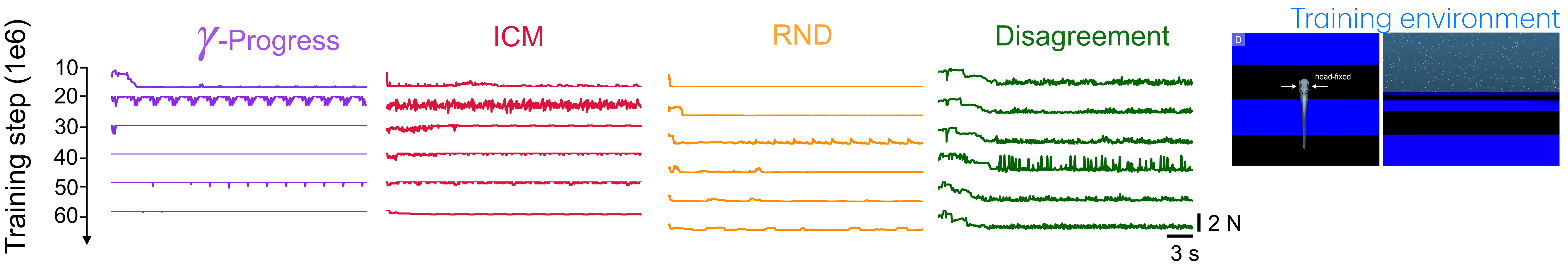


Kim et al., *ICML* (2020)
Pathak et al., *ICML* (2017)
Burda, Edwards & Pathak et al., *NeurIPS* (2017)

Often leads to unethological behaviors! (or can be stuck on white noise)

Epistemic Curiosity isn't Enough...

Animal autonomy != novelty optimization



What's the issue?

- Rewards are non-stationary and saturate with experience.

  Consequence: behavioral strategies are transient

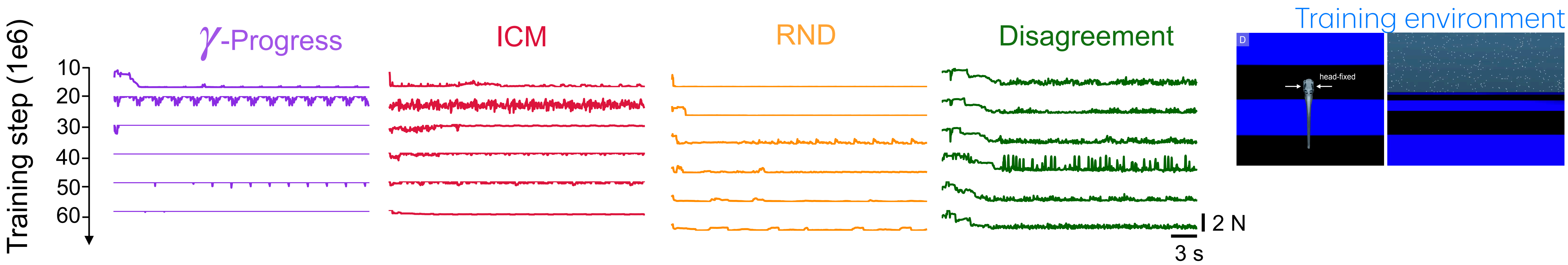
  (e.g. γ-Progress)

- Rewards can perseverate on unpredictable/uncontrollable stimuli.

  Consequence: unethological behavior (e.g. ICM)

Epistemic Curiosity isn't Enough...

Animal autonomy != novelty optimization



What's the issue?

- Rewards are non-stationary and saturate with experience.

  Consequence: behavioral strategies are transient

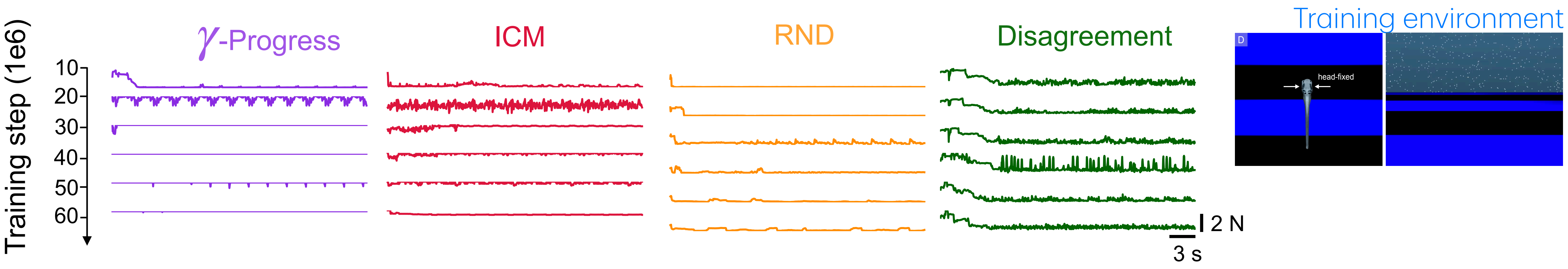
  (e.g. γ-Progress)

- Rewards can perseverate on unpredictable/uncontrollable stimuli.

  Consequence: unethological behavior (e.g. ICM)

Epistemic Curiosity isn't Enough...

Animal autonomy != novelty optimization



What's the issue?

- Rewards are non-stationary and saturate with experience.

  Consequence: behavioral strategies are transient

  (e.g. γ-Progress)

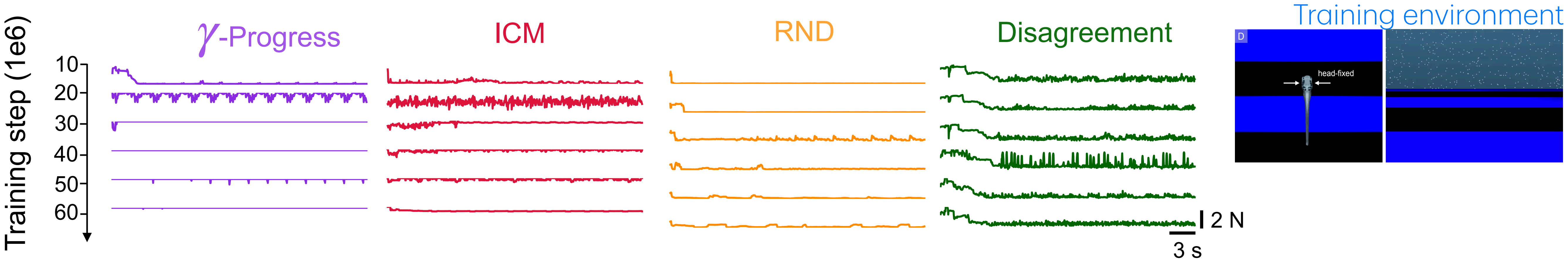
- Rewards can perseverate on unpredictable/uncontrollable stimuli.

  Consequence: unethological behavior (e.g. ICM)

Our approach: Incorporate *priors*

Epistemic Curiosity isn't Enough…

Animal autonomy != novelty optimization



What's the issue?

- Rewards are non-stationary and saturate with experience.
  Consequence: behavioral strategies are transient (e.g. γ-Progress)
- Rewards can perseverate on unpredictable/uncontrollable stimuli.
  Consequence: unethological behavior (e.g. ICM)

Our approach: Incorporate *priors*

The zebrafish behavior depends on an ethological memory.

memory = fixed or slowly adapting dynamics prior (a world model!)

This enables sensorimotor feedback error to be computed and tracked.

Question: What intrinsic drive explains this behavior?

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Our philosophy:

build the most convincing model possible.

- stimulus/image computable
- realistic physics
- flexible parameterization

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Zebrafish Simulation Environment

Our philosophy:

build the most convincing model possible.

- stimulus/image computable
- realistic physics
- flexible parameterization

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Zebrafish Simulation Environment

Our philosophy:

build the most convincing model possible.

- stimulus/image computable
- realistic physics
- flexible parameterization

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Zebrafish Simulation Environment

Actuation

- The embodiment must afford a faithful comparison with the animal behavior.
- Behavioral signal is low dimensional -> embodiment can be low dimensional
- Open-source embodiments that capture basic ethology already exist!

Our philosophy:

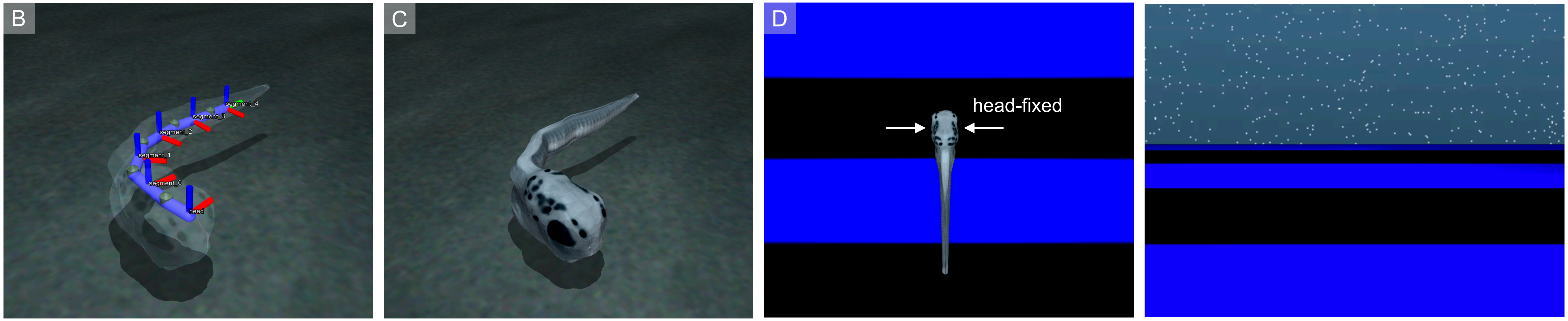
build the most convincing model possible.

- stimulus/image computable
- realistic physics
- flexible parameterization

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Zebrafish Simulation Environment



Actuation

- The embodiment must afford a faithful comparison with the animal behavior.
- Behavioral signal is low dimensional -> embodiment can be low dimensional
- Open-source embodiments that capture basic ethology already exist!

Our philosophy:

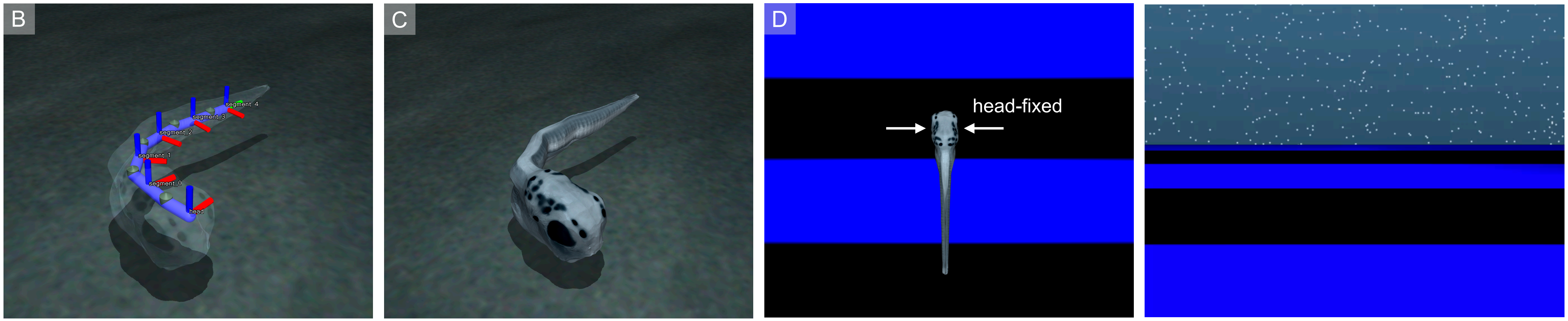
build the most convincing model possible.

- stimulus/image computable
- realistic physics
- flexible parameterization

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Zebrafish Simulation Environment



Actuation

- The embodiment must afford a faithful comparison with the animal behavior.
- Behavioral signal is low dimensional -> embodiment can be low dimensional
- Open-source embodiments that capture basic ethology already exist!

Sensing

- The zebrafish behavior is driven by optic flow and proprioception. A basic vision model and state information is sufficient.

Our philosophy:

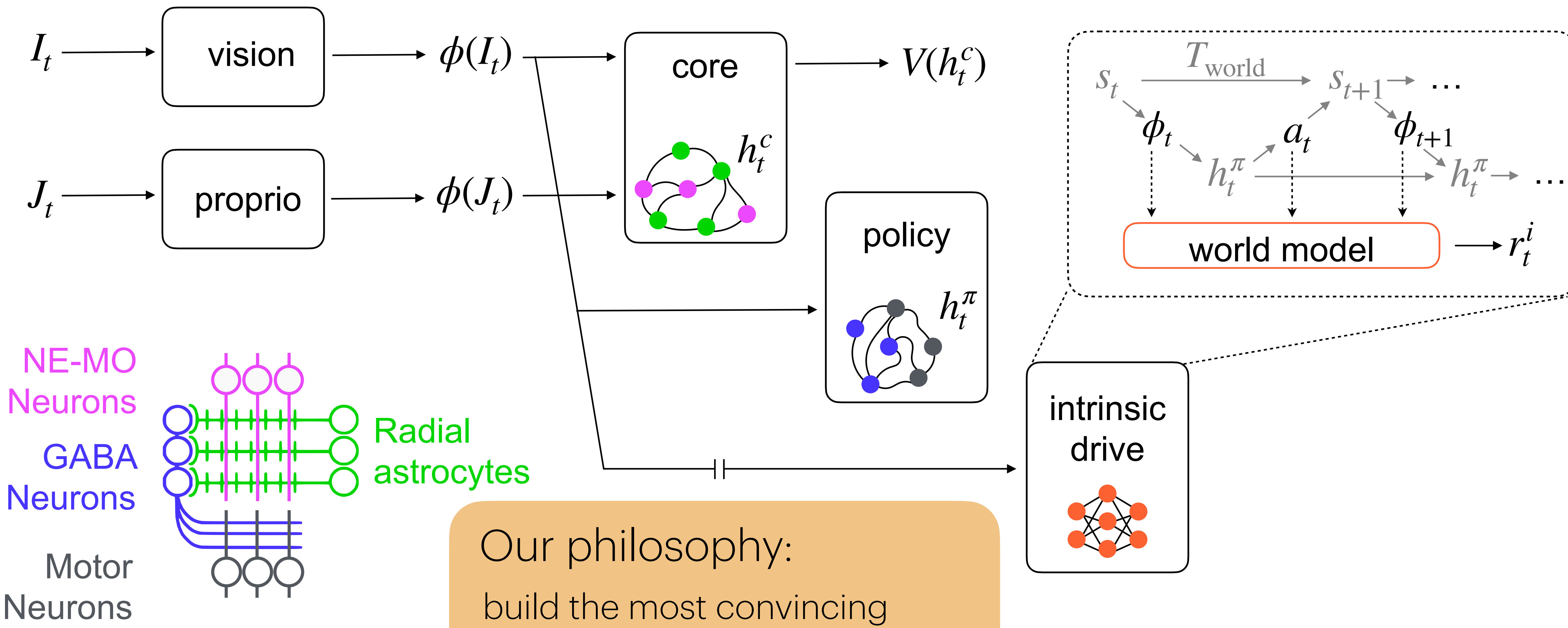
build the most convincing model possible.

- stimulus/image computable
- realistic physics
- flexible parameterization

Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

Zebrafish Agent Architecture



Our philosophy:

build the most convincing model possible.

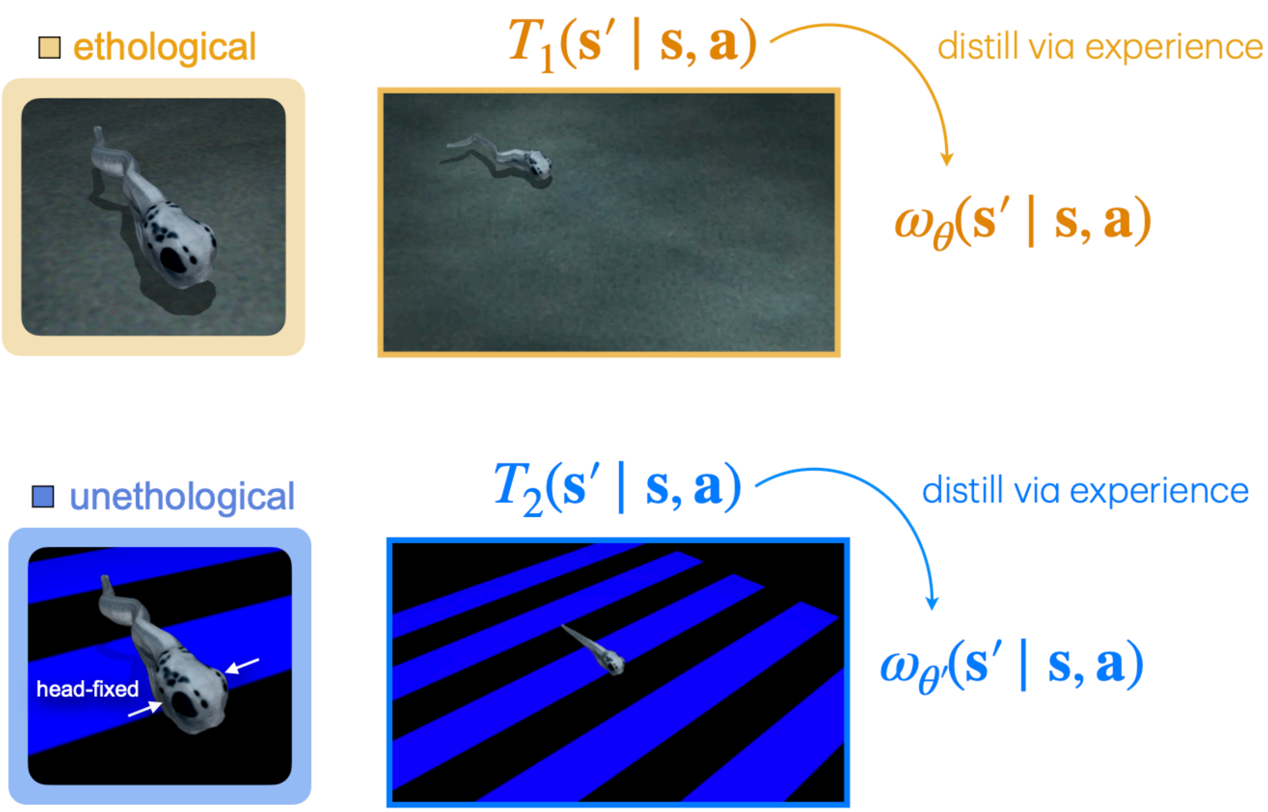
- stimulus/image computable
- realistic physics
- flexible parameterization

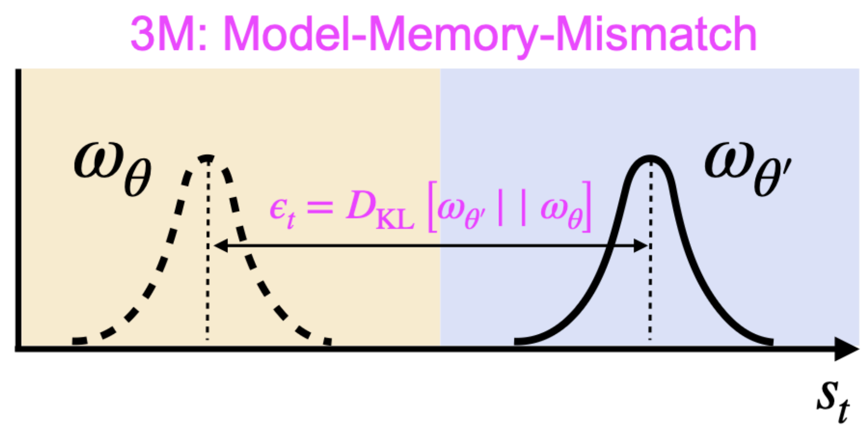
Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

3M-Progress

Using ethological memory to guide adaptive behavior

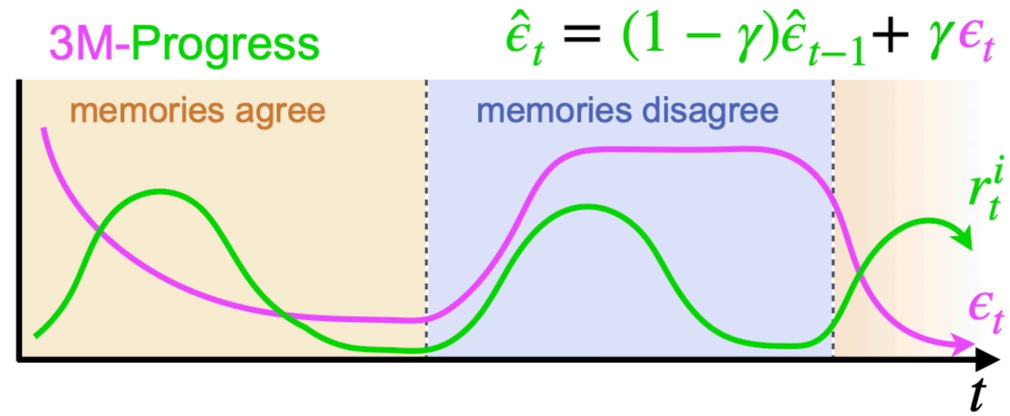




ϵ_t partitions the state-action space into model-memory agreement (U) and disagreement (U^C).

$r_t^i \propto |\hat{\epsilon}_t - \epsilon_t|$

$\hat{\epsilon}_t = (1-\gamma)\hat{\epsilon}_{t-1} + \gamma \epsilon_t$



We choose T_1 and T_2 to obey:

$$\exists U \subset S \times A \;\; \text{s.t.} \;\; \forall (\mathbf{s}, \mathbf{a}) \in U, \; T_1 \approx T_2$$

(dynamics agree on a subspace).

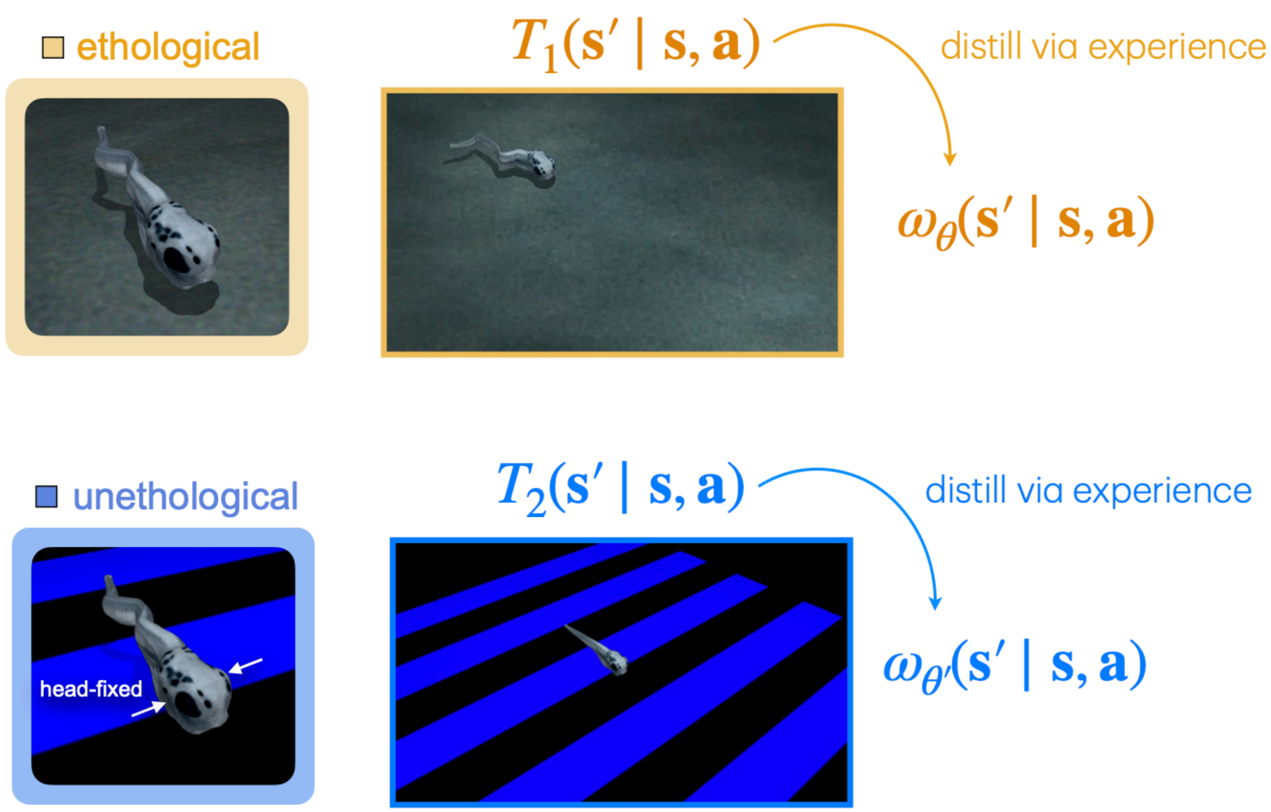
Question: What intrinsic drive explains this behavior?

Specifically, how should world-models be used to guide autonomous decisions in real-world situations (e.g. encountering unseen physics)?

3M-Progress

Recall the planning section!

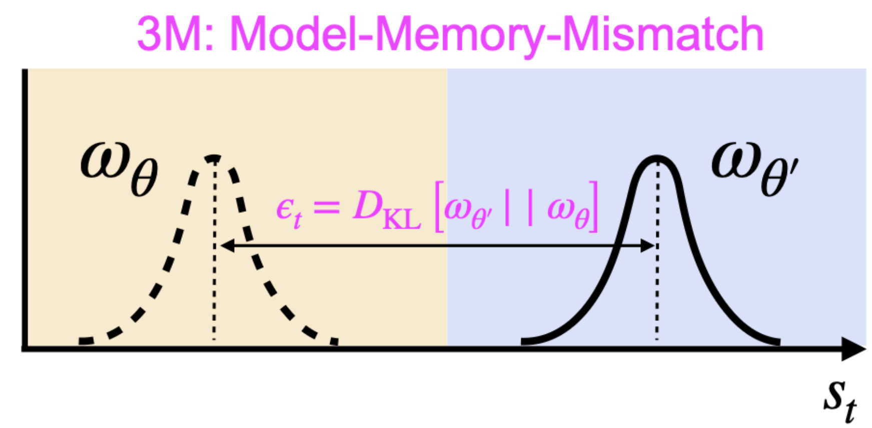
Using ethological memory to guide adaptive behavior



We choose T_1 and T_2 to obey:

$$\exists U \subset S \times A \ \text{ s.t. } \forall (\mathbf{s}, \mathbf{a}) \in U, \ T_1 \approx T_2$$

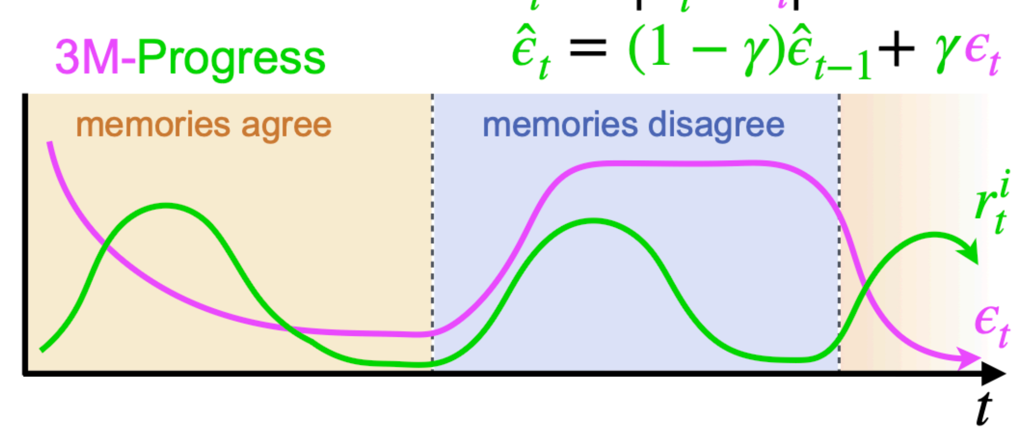
(dynamics agree on a subspace).



ϵ_t partitions the state-action space into model-memory agreement (U) and disagreement (U^C).

$$r_t^i \propto |\hat{\epsilon}_t - \epsilon_t|$$

$$\hat{\epsilon}_t = (1 - \gamma)\hat{\epsilon}_{t-1} + \gamma \epsilon_t$$



Putting it all together

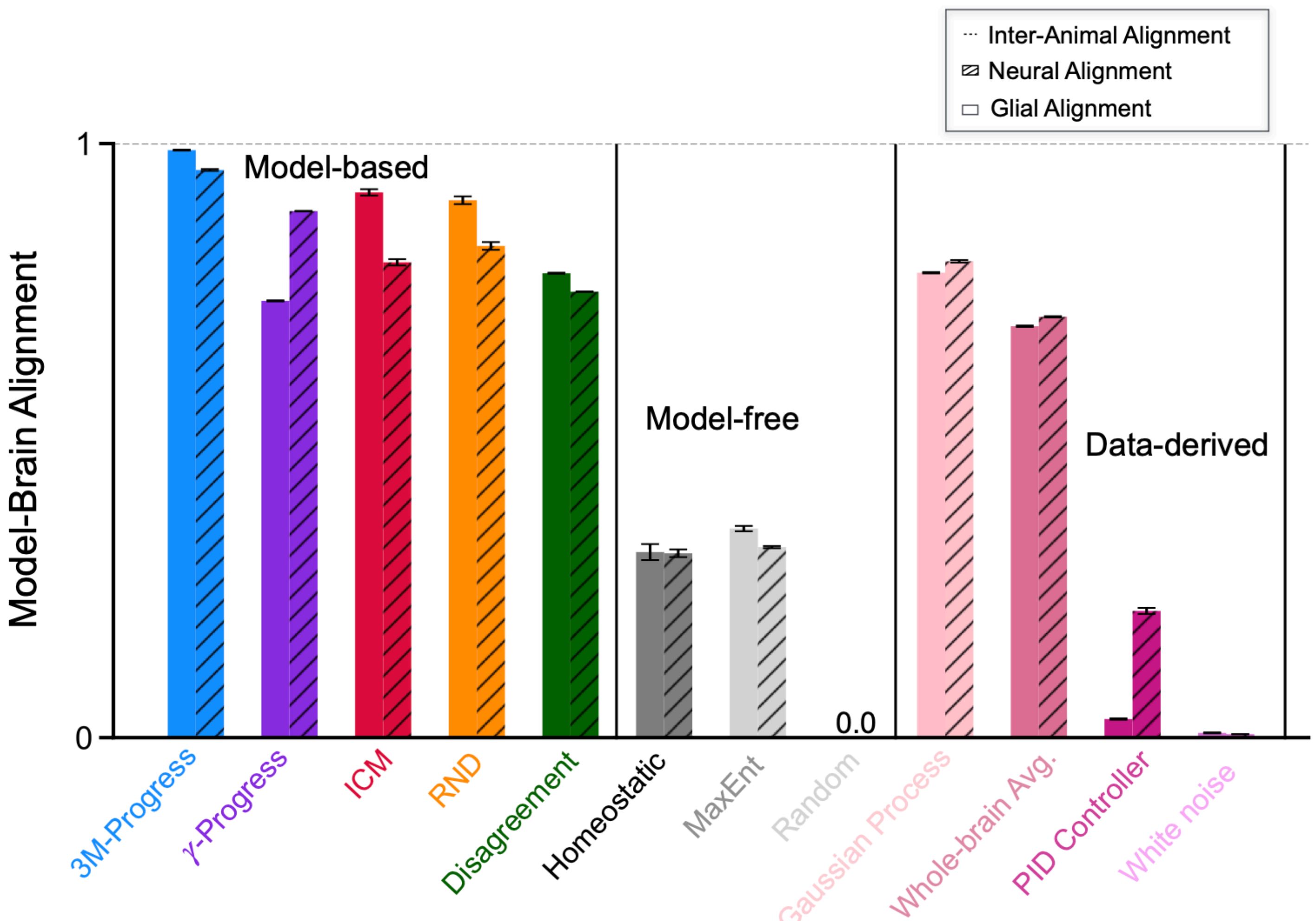
Putting it all together

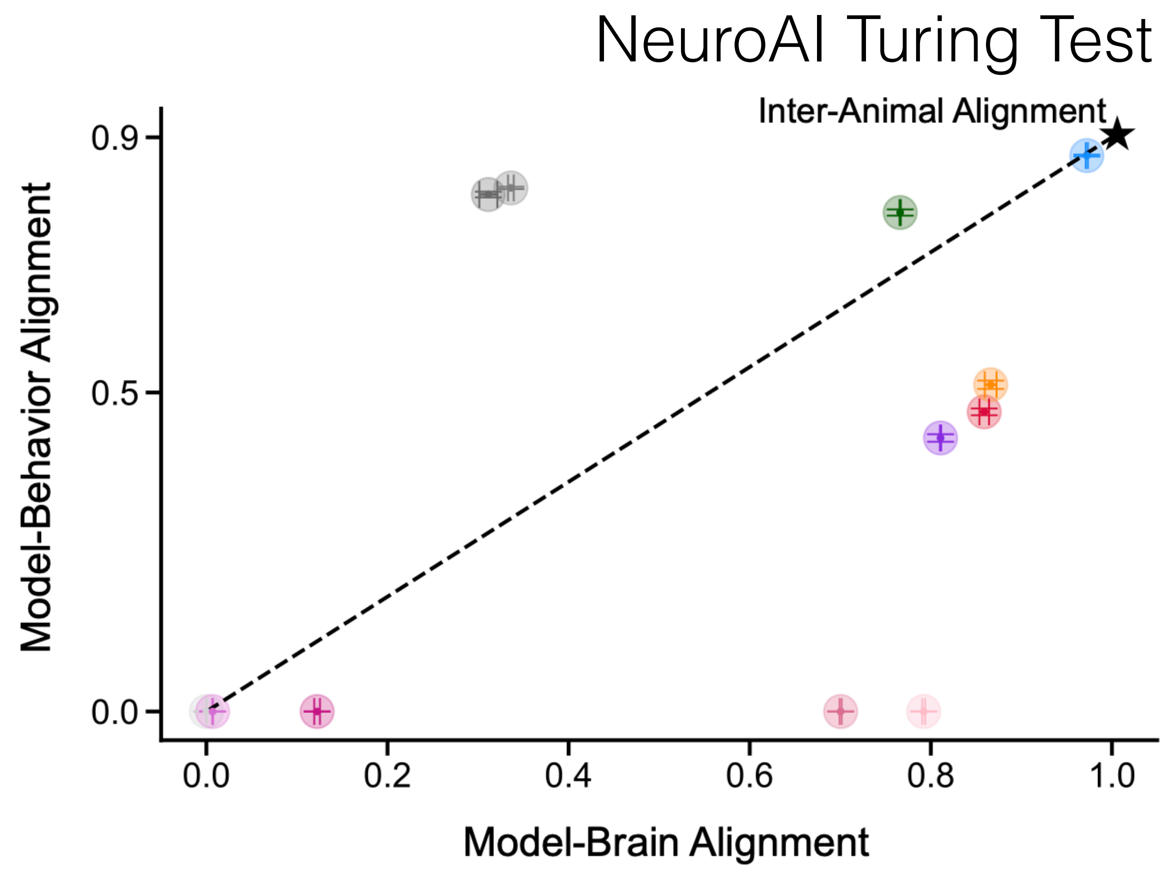
Putting it all together

3M-Progress Captures Whole-Brain Dynamics

(and behavior)

Single-cell one-to-one alignment

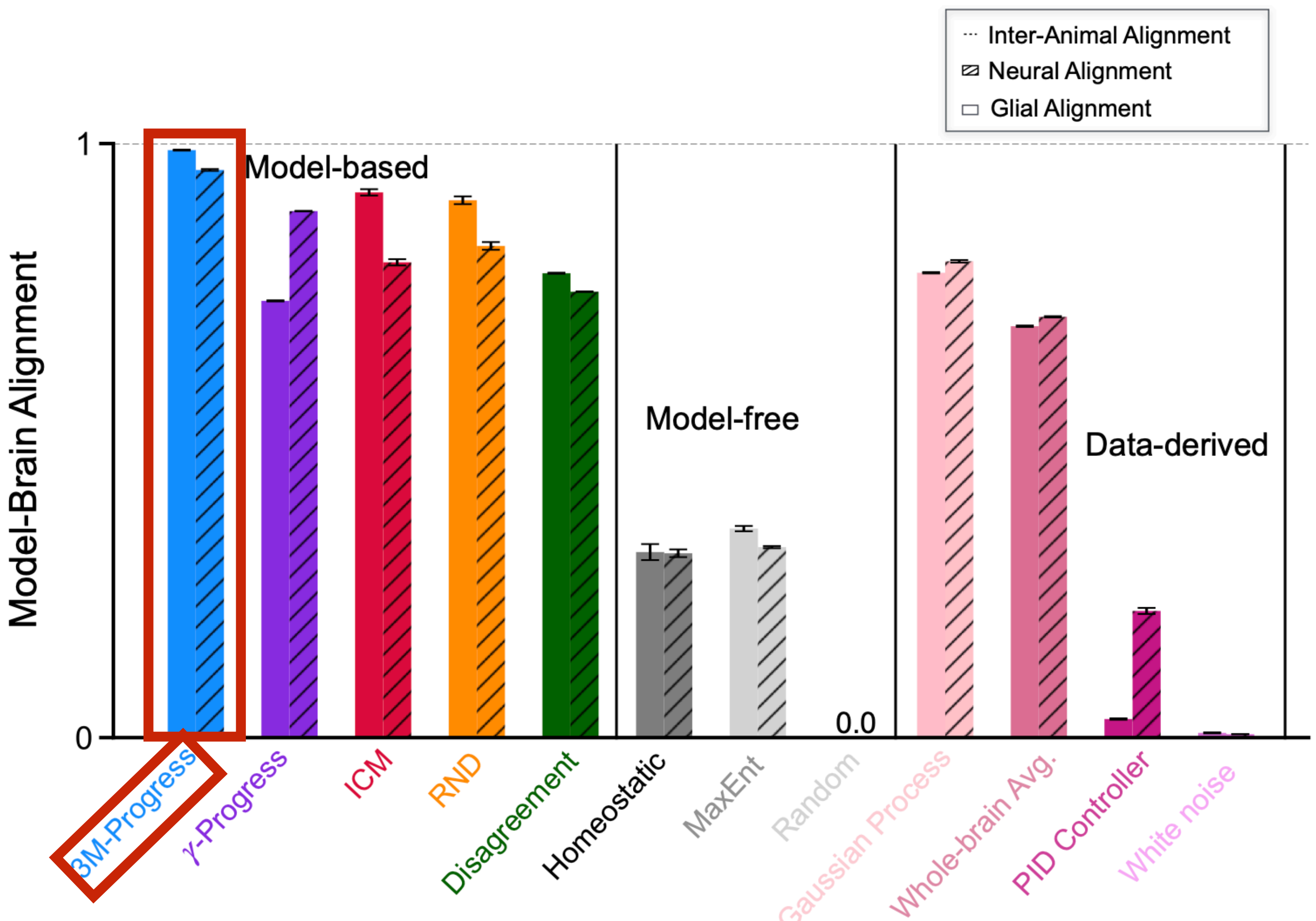


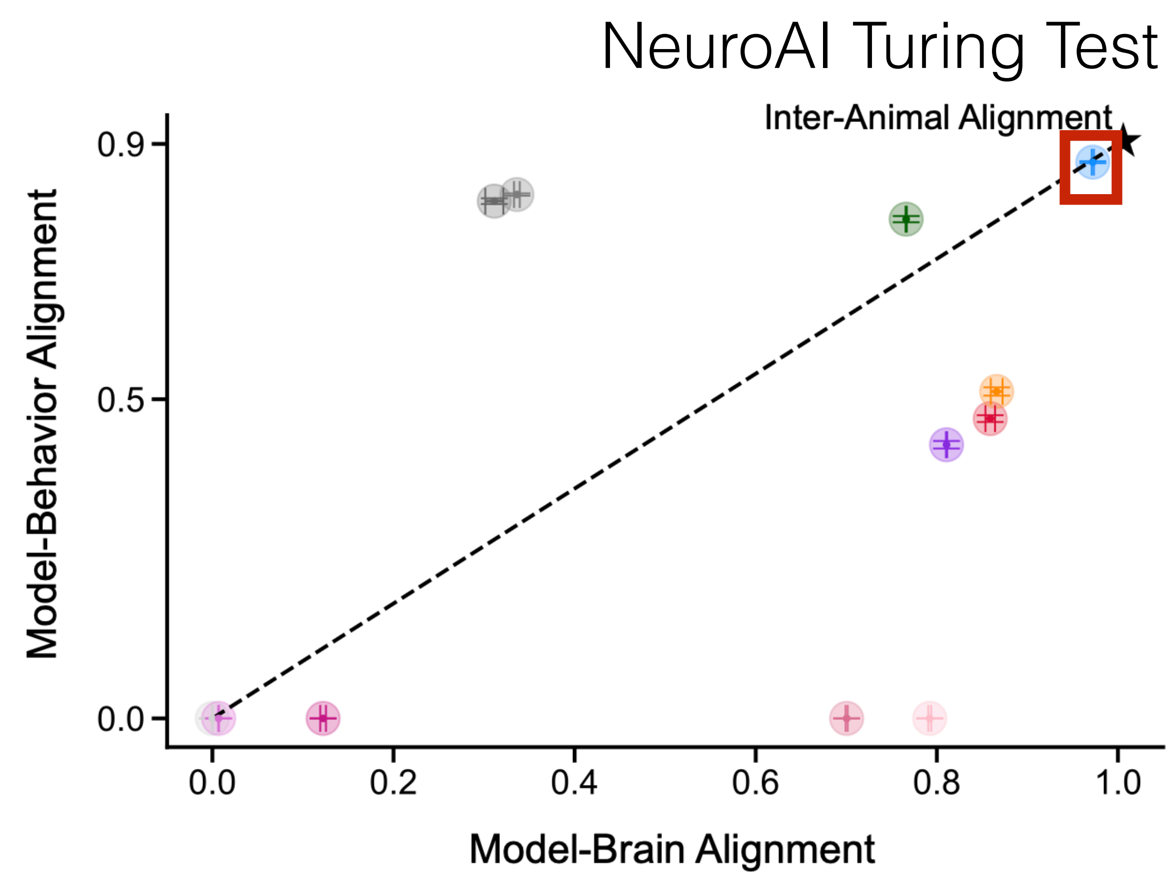


3M-Progress Captures Whole-Brain Dynamics

(and behavior)

Single-cell one-to-one alignment

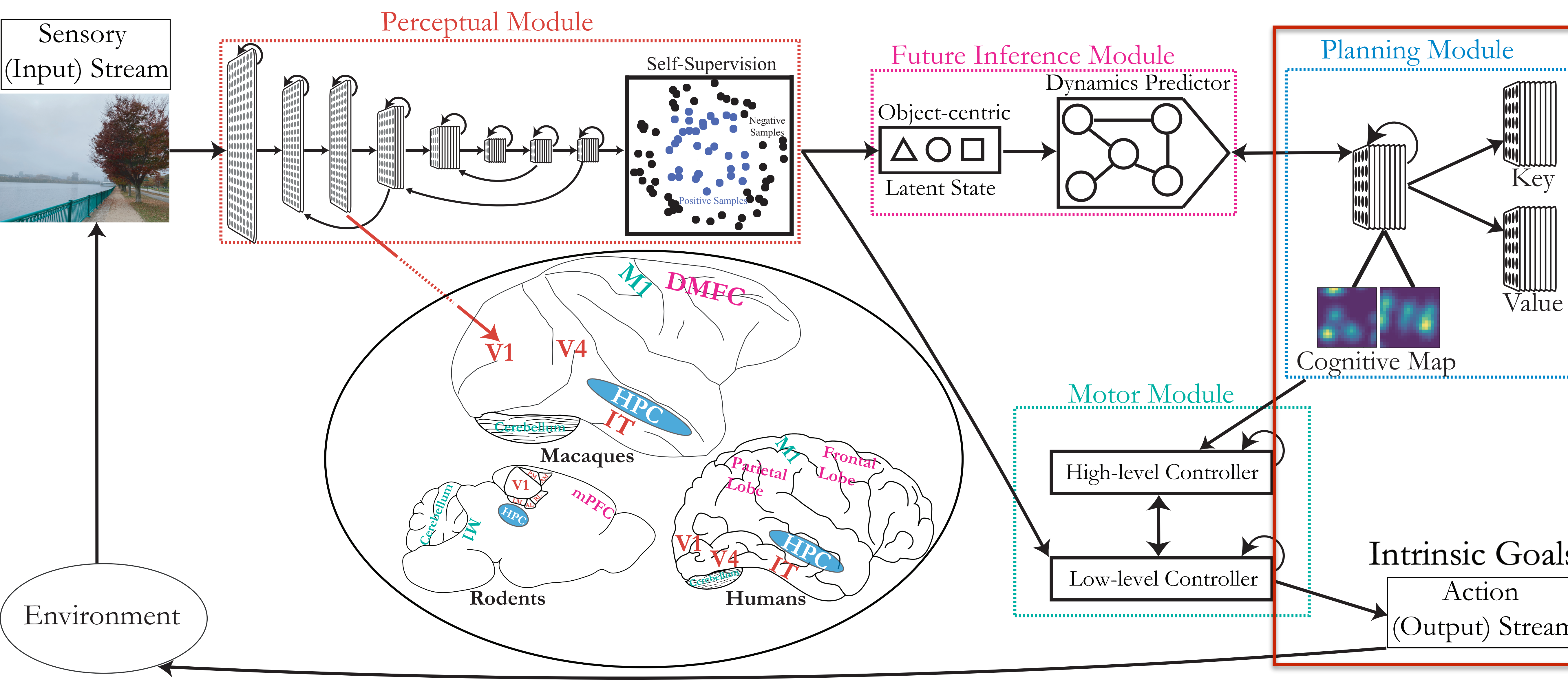




Roadmap: Planning & Action

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

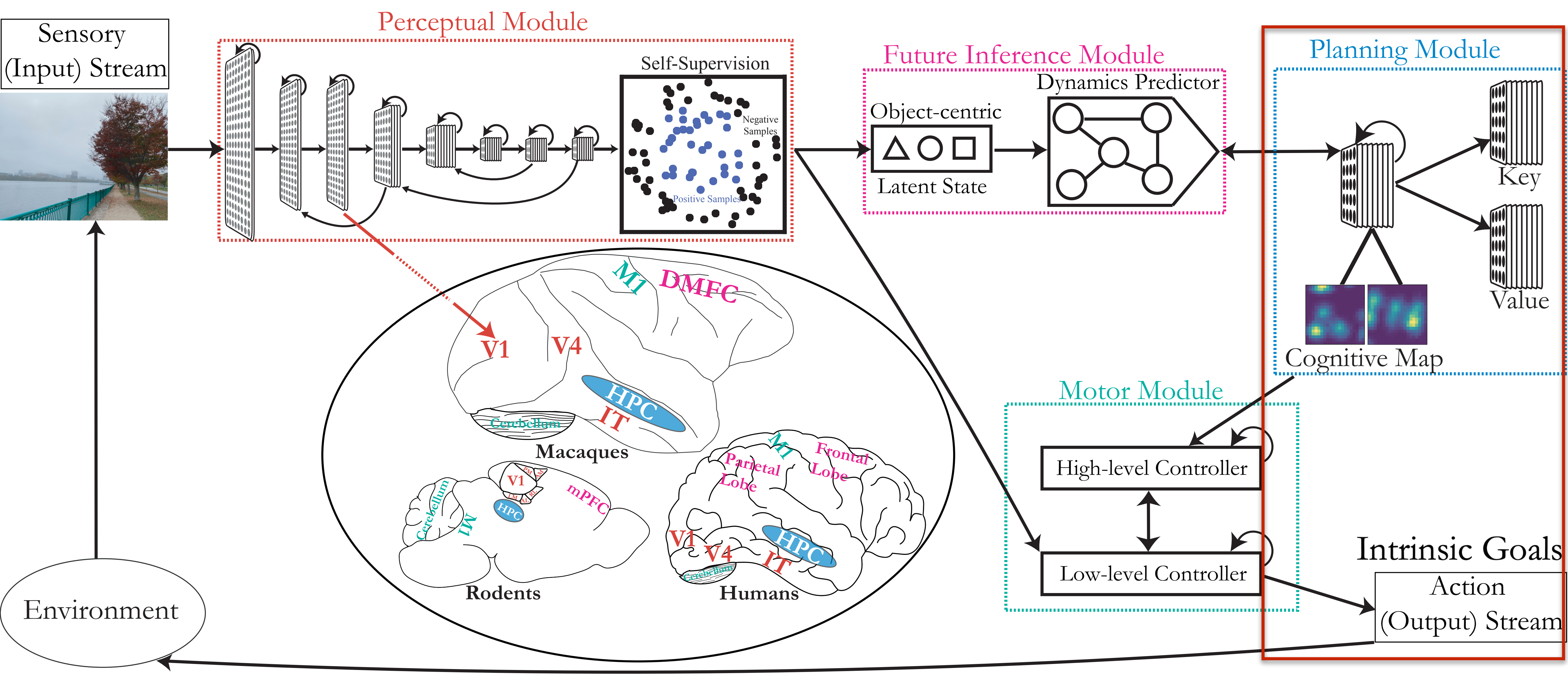
Recurrence + Contrastive SSL? Latent Future Prediction?



Roadmap: Planning & Action

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

Recurrence + Contrastive SSL? Latent Future Prediction?

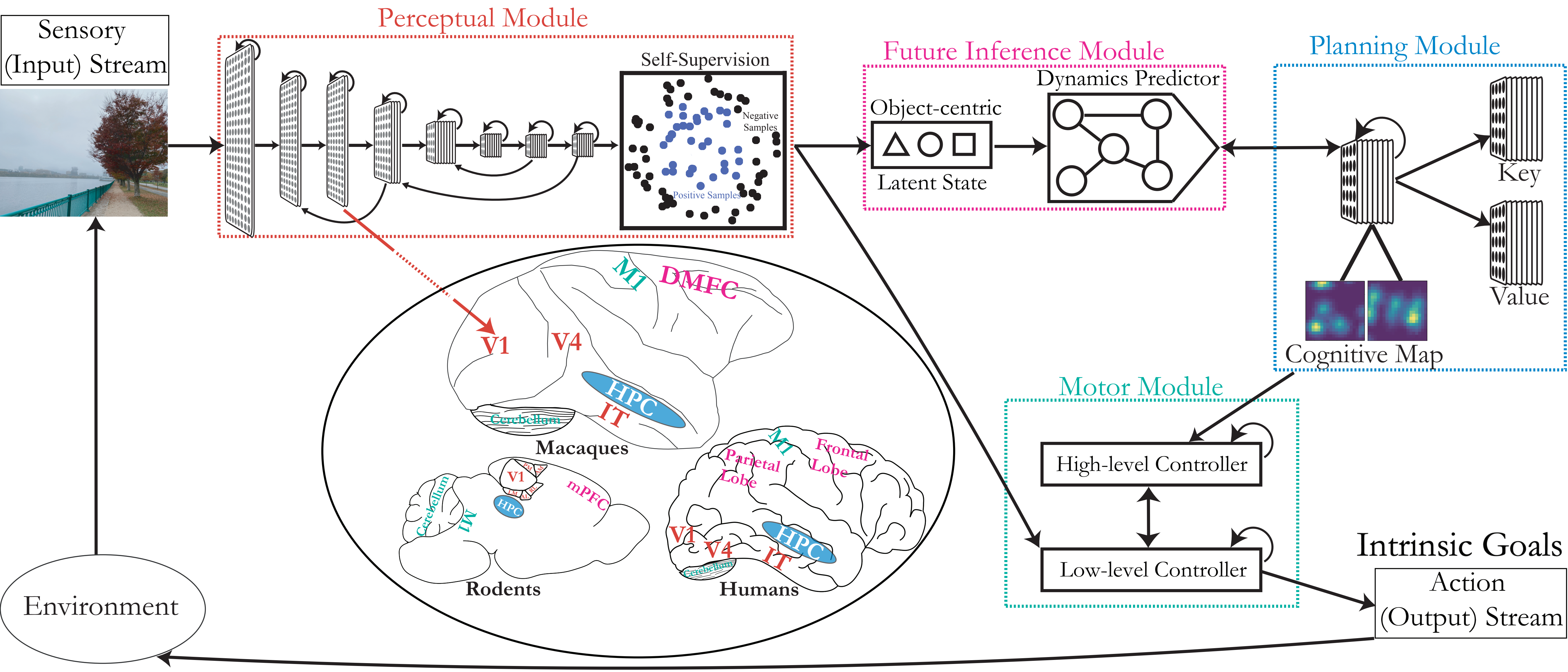


Temporal integration of World Model-Progress-based curiosity?

Safety Implications: What Happens Once We Get There?

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

Recurrence + Contrastive SSL? Latent Future Prediction?

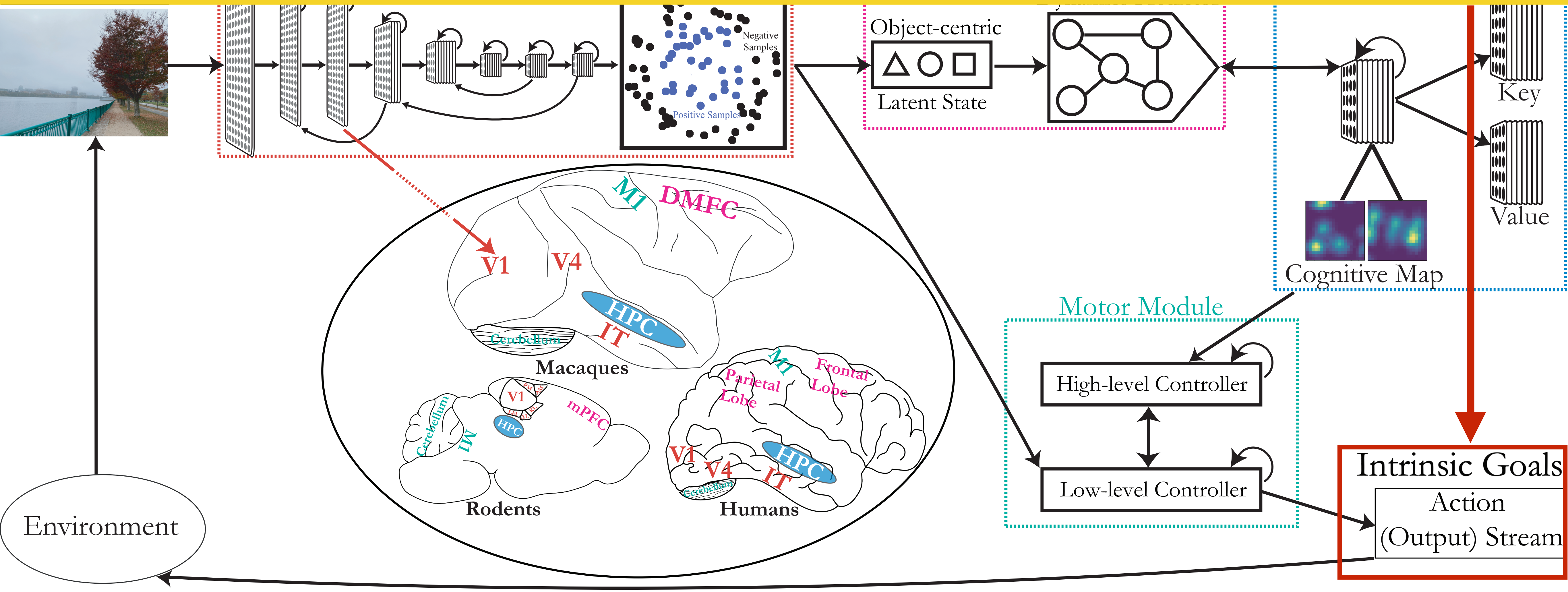


Temporal integration of World Model-Progress-based curiosity?

Safety Implications: What Happens Once We Get There?

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

Too many of these goals makes alignment *intractable*, even for computationally *unbounded* agents!



Temporal integration of World Model-Progress-based curiosity?

1. Intrinsic Barriers and Practical Pathways for Human-AI Alignment: An Agreement-Based Complexity Analysis

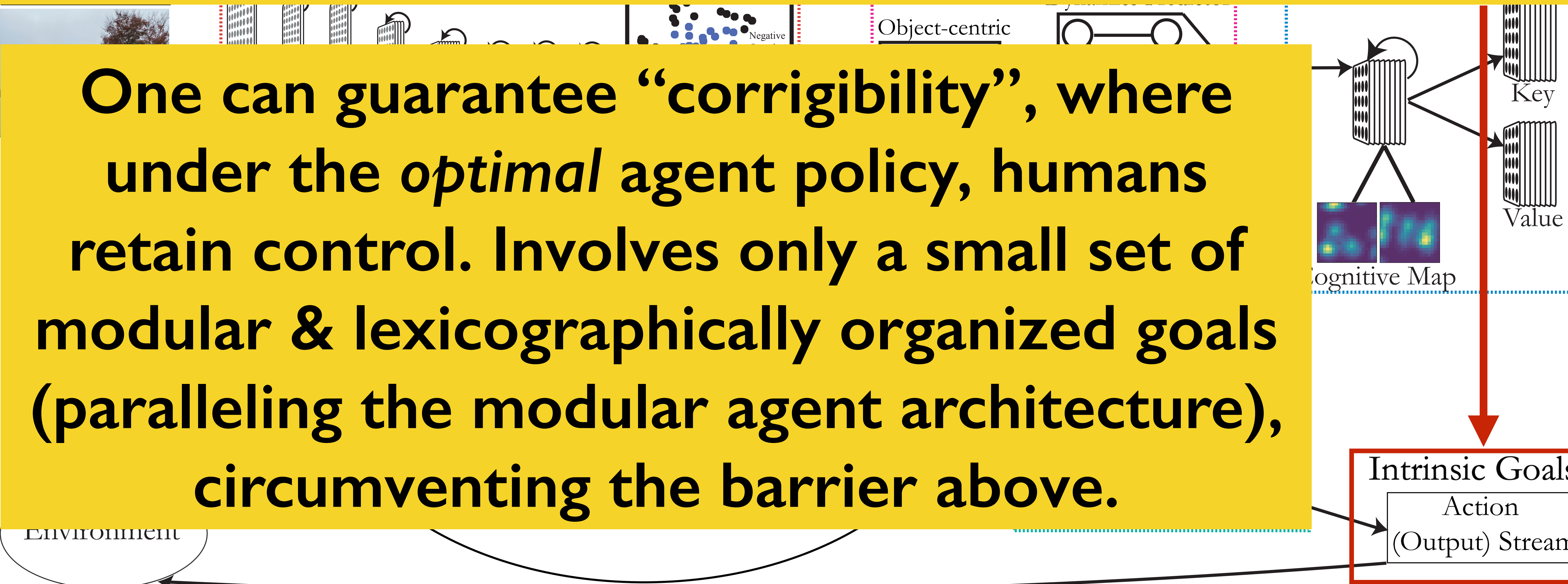
Paper: https://arxiv.org/abs/2502.05934

Safety Implications: What Happens Once We Get There?

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

Too many of these goals makes alignment *intractable*, even for computationally *unbounded* agents!

One can guarantee "corrigibility", where under the *optimal* agent policy, humans retain control. Involves only a small set of modular & lexicographically organized goals (paralleling the modular agent architecture), circumventing the barrier above.



Temporal integration of World Model-Progress-based curiosity?

1. Intrinsic Barriers and Practical Pathways for Human-AI Alignment: An Agreement-Based Complexity Analysis — **Paper:** https://arxiv.org/abs/2502.05934

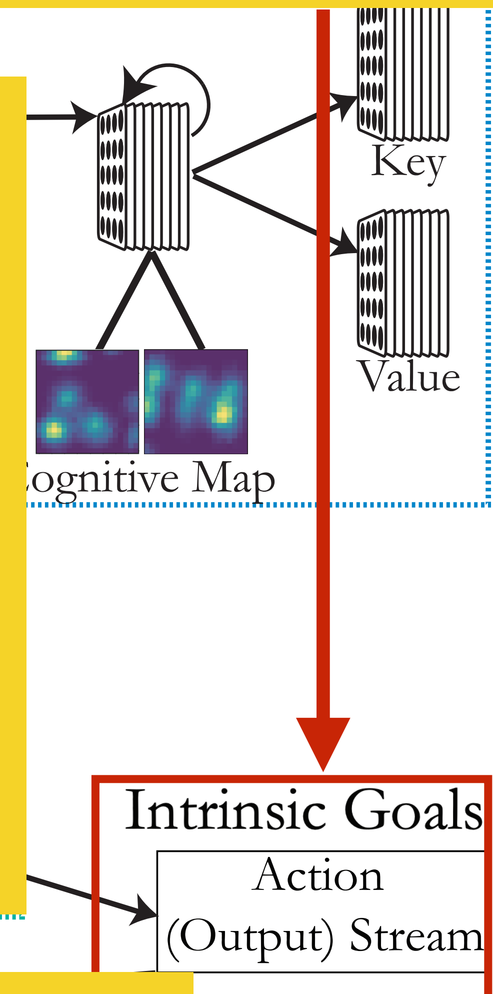
2. Core Safety Values for Provably Corrigible Agents — **Paper:** https://arxiv.org/abs/2507.20964

Safety Implications: What Happens Once We Get There?

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

Too many of these goals makes alignment *intractable*, even for computationally *unbounded* agents!

One can guarantee "corrigibility", where under the *optimal* agent policy, humans retain control. Involves only a small set of modular & lexicographically organized goals (paralleling the modular agent architecture), circumventing the barrier above.



<u>Open:</u> Can we scale corrigibility cost effectively?

1. Intrinsic Barriers and Practical Pathways for Human-AI Alignment: An Agreement-Based Complexity Analysis

Paper: https://arxiv.org/abs/2502.05934

2. Core Safety Values for Provably Corrigible Agents

Paper: https://arxiv.org/abs/2507.20964

Potential Economic Implications of Alignment

Potential Economic Implications of Alignment

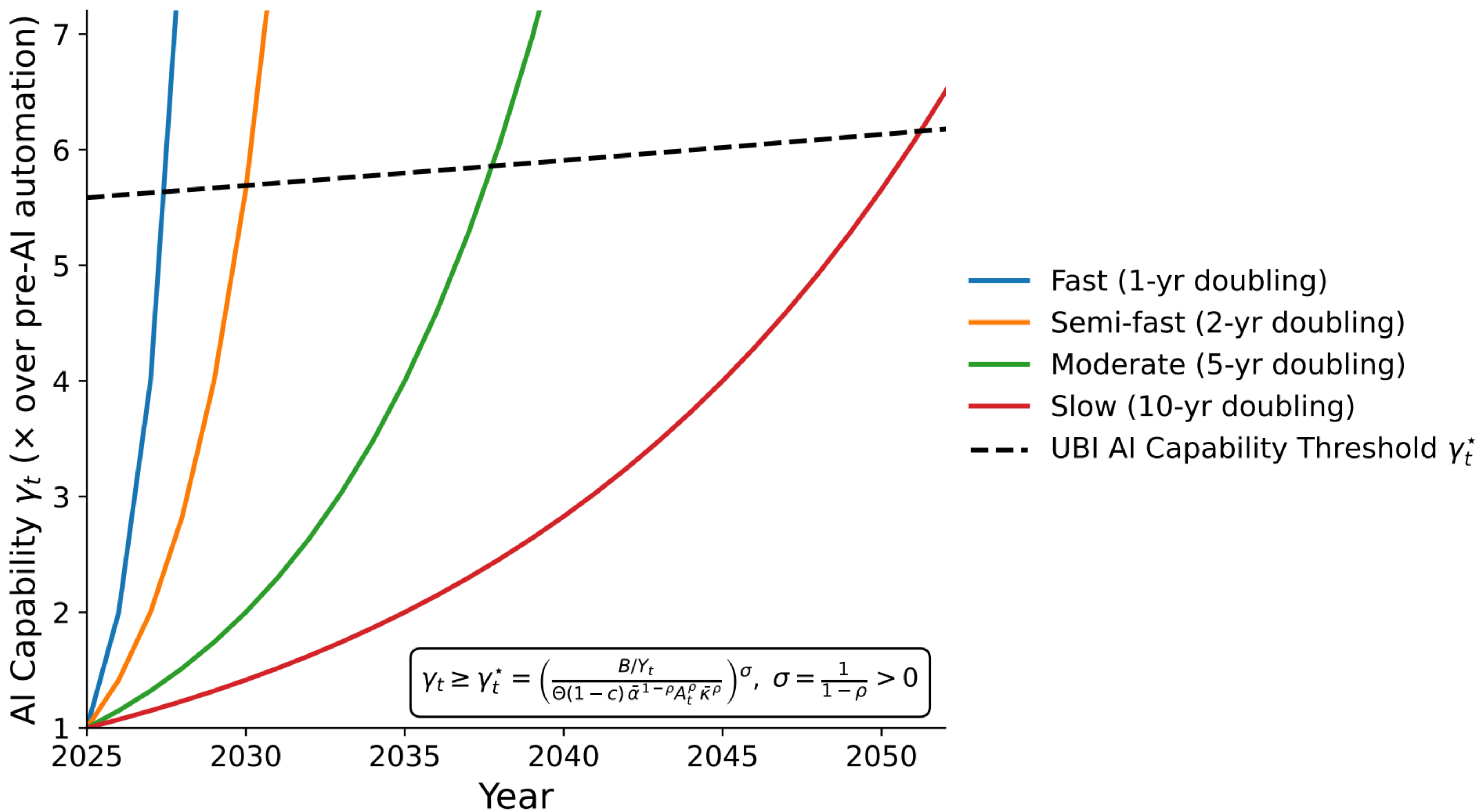


Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold ($\gamma_t^\star$).** The dashed line is the required capability $\gamma_t^\star$ to fully fund a UBI that comprises 11% of the GDP (leading to a $\gamma_t^\star$ between 5-6× the pre-AI productivity on automated tasks, under current economic assumptions). Under fast scaling (AI capability doubling every year), AI would cross the threshold by the late 2020s. Semi-fast scaling (doubling every 2 years) reaches the threshold in the early 2030s, whereas moderate (doubling every 5 years) and slow (doubling every 10 years) scenarios achieve $\gamma_t^\star$ by 2038 and 2052, respectively. The trajectories are illustrative, starting from a nominal, conservative 2025 capability level ($\gamma_0 \equiv 1$), which assumes AI currently delivers no boost beyond the pre-AI automation level in aggregate across all automated tasks.

3. An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy

Paper: https://arxiv.org/abs/2505.18687

Potential Economic Implications of Alignment

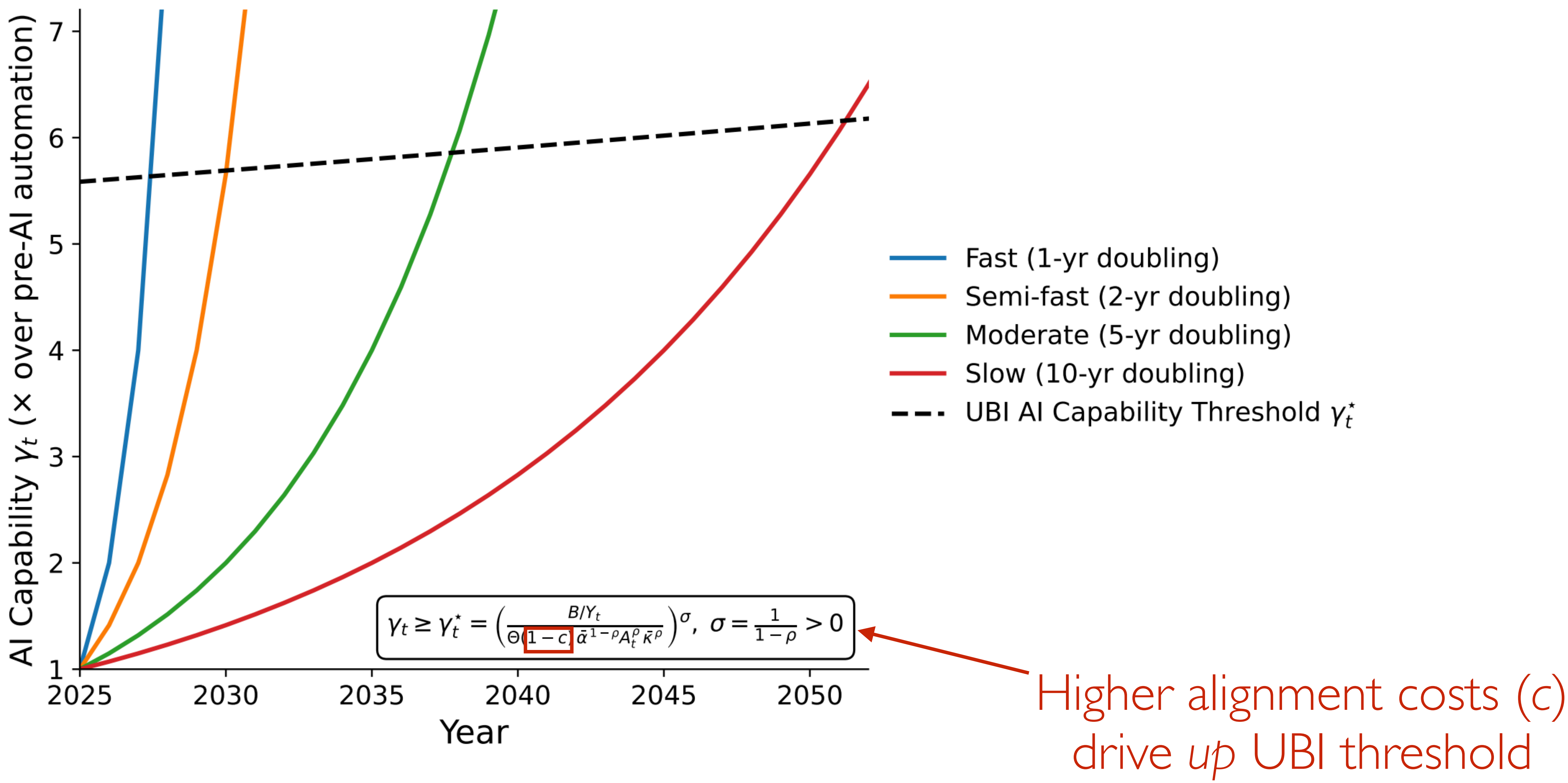
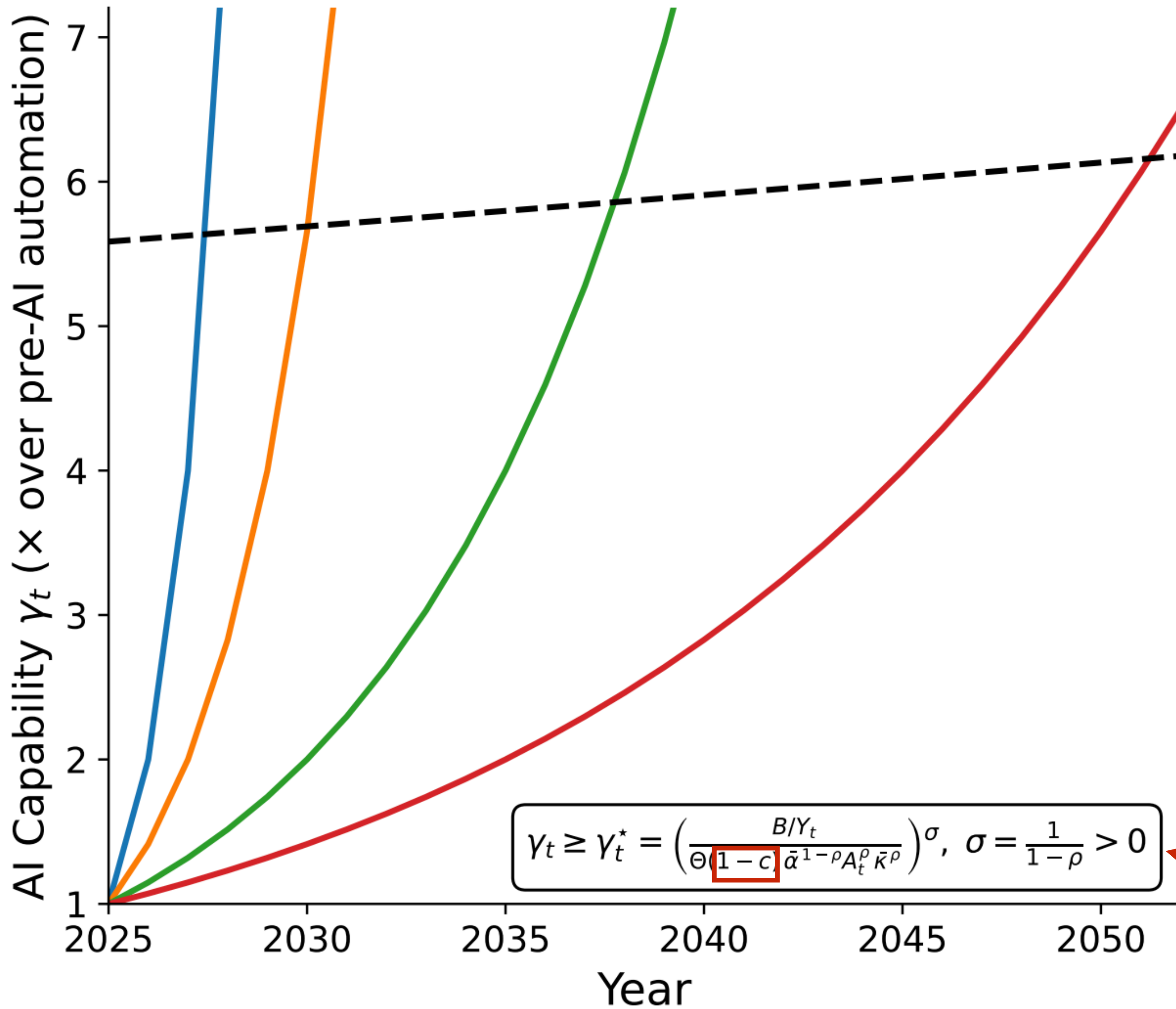


Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold ($\gamma_t^\star$).** The dashed line is the required capability $\gamma_t^\star$ to fully fund a UBI that comprises 11% of the GDP (leading to a $\gamma_t^\star$ between 5-6× the pre-AI productivity on automated tasks, under current economic assumptions). Under fast scaling (AI capability doubling every year), AI would cross the threshold by the late 2020s. Semi-fast scaling (doubling every 2 years) reaches the threshold in the early 2030s, whereas moderate (doubling every 5 years) and slow (doubling every 10 years) scenarios achieve $\gamma_t^\star$ by 2038 and 2052, respectively. The trajectories are illustrative, starting from a nominal, conservative 2025 capability level ($\gamma_0 \equiv 1$), which assumes AI currently delivers no boost beyond the pre-AI automation level in aggregate across all automated tasks.

3. An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy

Paper: https://arxiv.org/abs/2505.18687

Potential Economic Implications of Alignment



Open: Can we incorporate other values (besides control, which is "neutrally amoral") that lead to longer term human well-being, especially if working for pay becomes no longer feasible in many cases?

Higher alignment costs (c) drive *up* UBI threshold

Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold ($\gamma_t^\star$).** The dashed line is the required capability $\gamma_t^\star$ to fully fund a UBI that comprises 11% of the GDP (leading to a $\gamma_t^\star$ between 5-6× the pre-AI productivity on automated tasks, under current economic assumptions). Under fast scaling (AI capability doubling every year), AI would cross the threshold by the late 2020s. Semi-fast scaling (doubling every 2 years) reaches the threshold in the early 2030s, whereas moderate (doubling every 5 years) and slow (doubling every 10 years) scenarios achieve $\gamma_t^\star$ by 2038 and 2052, respectively. The trajectories are illustrative, starting from a nominal, conservative 2025 capability level ($\gamma_0 \equiv 1$), which assumes AI currently delivers no boost beyond the pre-AI automation level in aggregate across all automated tasks.

3. An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy

Paper: https://arxiv.org/abs/2505.18687

Acknowledgements

NeuroAgents Lab

Contact:

anayebi@cs.cmu.edu

[@aran_nayebi](https://x.com/aran_nayebi)

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

Neuro Agents Lab

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Funding:

Foresight Institute

UK AISI Challenge Fund

Google Robotics Award

Burroughs Wellcome Fund CASI Award