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

Aran Nayebi[1], Rishi Rajalingham[1,4], Mehrdad Jazayeri[1,2], Guangyu Robert Yang[1,2,3]

[1]McGovern Institute for Brain Research, MIT; Cambridge, MA 02139
[2]Department of Brain and Cognitive Sciences, MIT; Cambridge, MA 02139
[3]Department of Electrical Engineering and Computer Science, MIT; Cambridge, MA 02139
[4]Reality Labs, Meta; 390 9th Ave, New York, NY 10001

1. Overall Approach

(A) Model Pretraining

Inputs

Sensory-Cognitive Hypothesis Classes

Latent Future Prediction:
1. Pretraining Stage — Ego4D, etc
Foundation Model
2. Dynamics Training Stage
Ground Truth — T+1
Prediction — T+1

End-to-End Future Prediction:
Encoder → Decoder — Pixel-wise
Encoder → Object-slot

Observed + Simulated

(B) Model Evaluations

1. Human Behavior: Physion Object Contact Prediction (OCP)
Observed Stimuli — Time — Unobserved Outcome
NO acc. = 0.89
YES acc. = 0.96

2. Macaque Neurophysiology: Mental-Pong

2. Comparison to Human Physical Judgements

(A)

Human Accuracy — Heldout OCP Accuracy

(B) Task Accuracy & Matching Human Error Patterns are Related

R≈0.865, p << 0.001

3. Neural Response Predictivity Strongly Separates Models

(A) Neural Comparison Procedure

(B)

Inter-animal Consistency

(C) DMFC Predictivity

(D)

R≈0.701, p << 0.001

4. Dynamically-Equipped Video Foundation Models Better Match *Both* Human Behavioral & Neural Response Patterns *Across* Environments

(A)

Better Models

VC-1+LSTM
VC-1+CTRNN

(B)

Soft-body interactions could be improved uniformly across models

Drape

Support