Pixels over Symbols

Sensory Realism Improves Behavioral Alignment in Models of Cognition

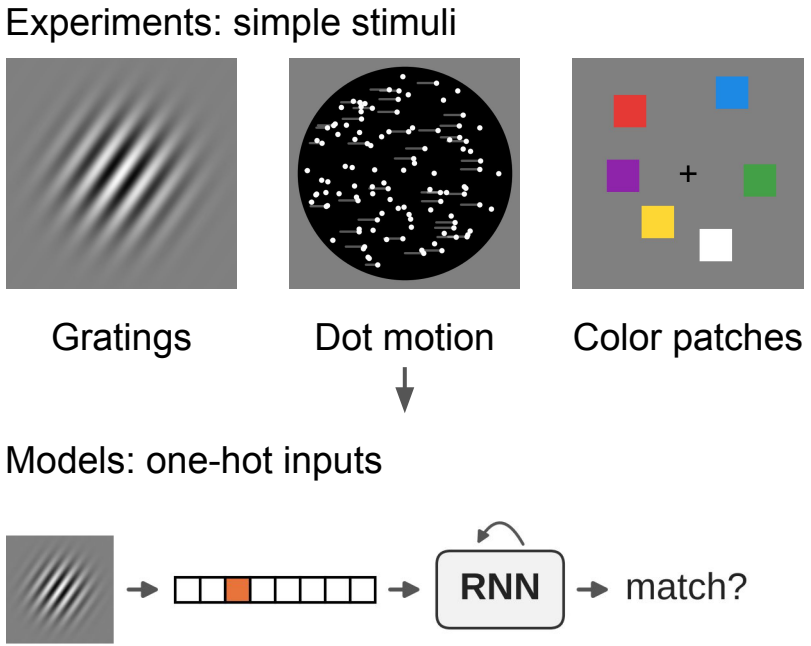
Hao Yuan Bai, Xiaoxuan Lei, Zihan Weng, Motahareh Pourrahimi, Lucas Gomez, Pouya Bashivan





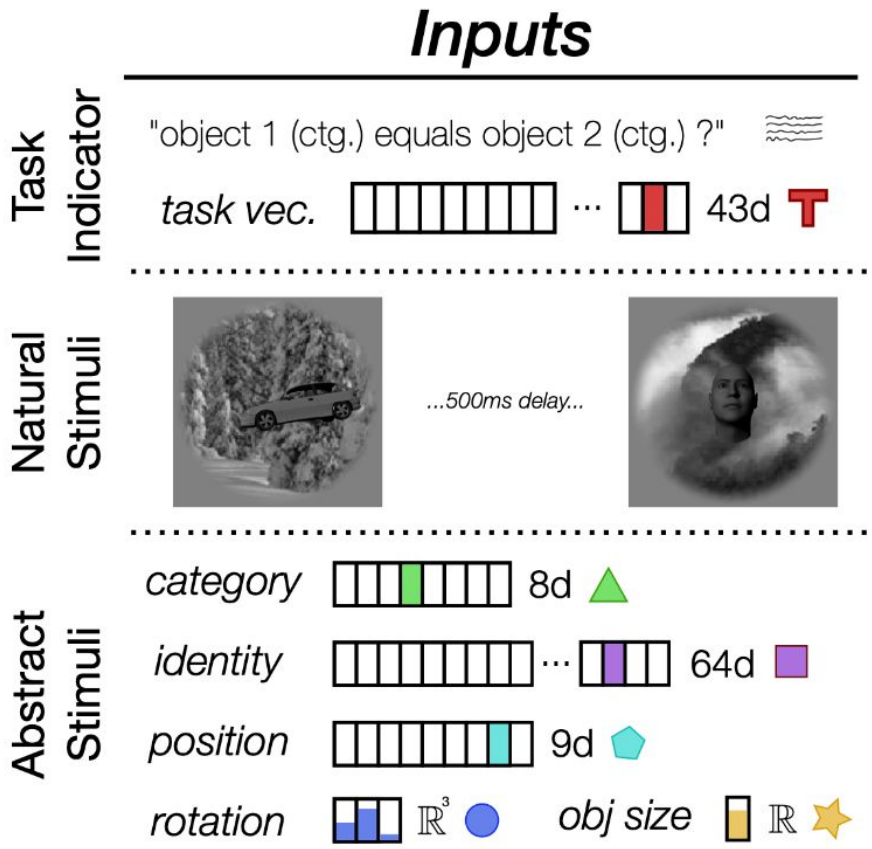
Motivation

- Models of cognition were trained on one-hot or scalar task features
  - Compute infeasibility
  - Assumption that the brain encodes stimuli categorically
- Many claims about circuit dynamics and representational geometry are built upon these models

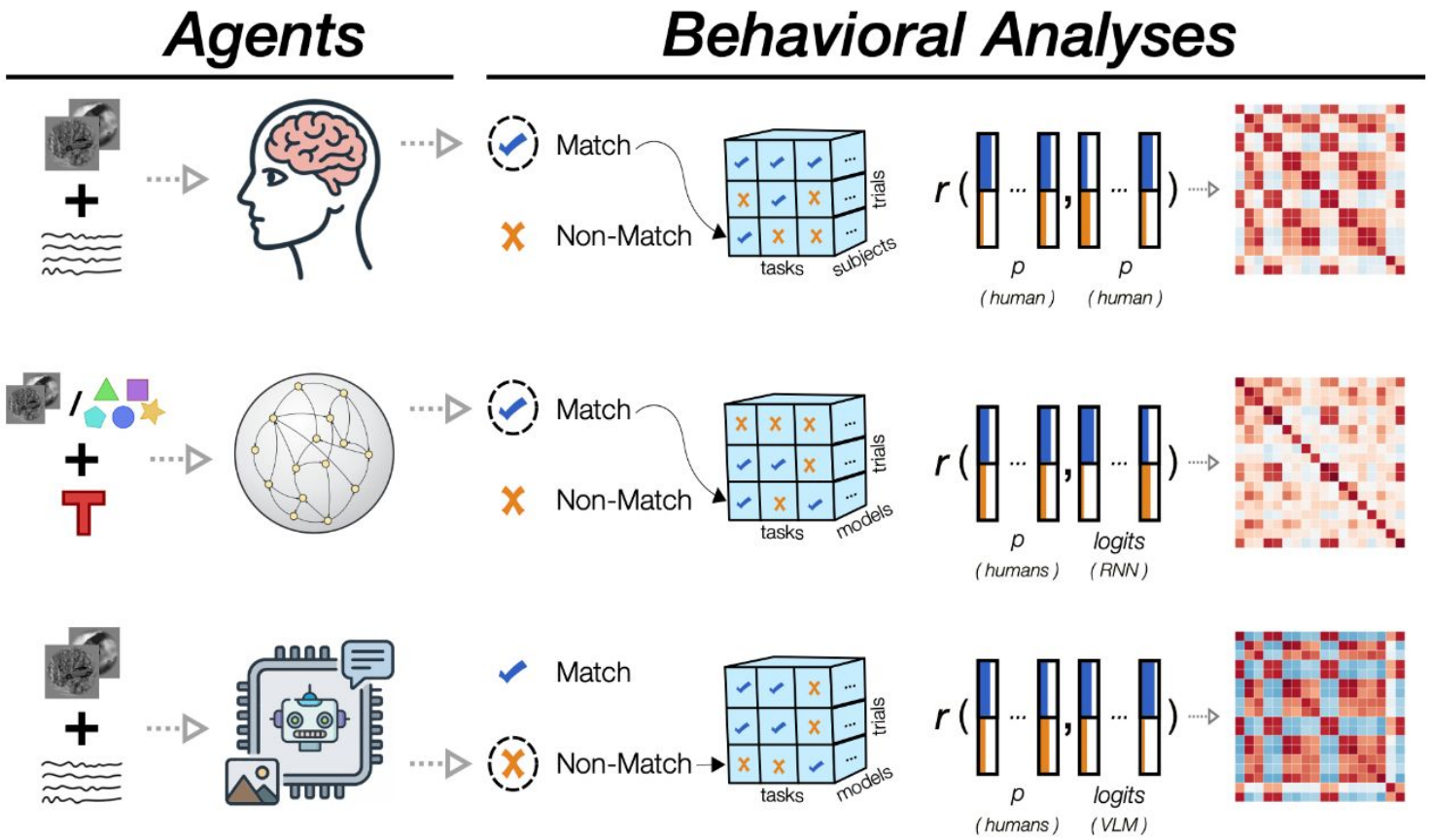


One-hot models: Mante et al., Nature 2013; Yang et al., Nat. Neurosci. 2019. Categorical coding: Freedman et al., Science 2001.
Stimuli: Campbell & Robson, J. Physiol. 1968; Newsome et al., Nature 1989; Luck & Vogel, Nature 1997.

1. Does sensory realism affect behavioral patterns?
2. Does sensory realism affect internal solutions?

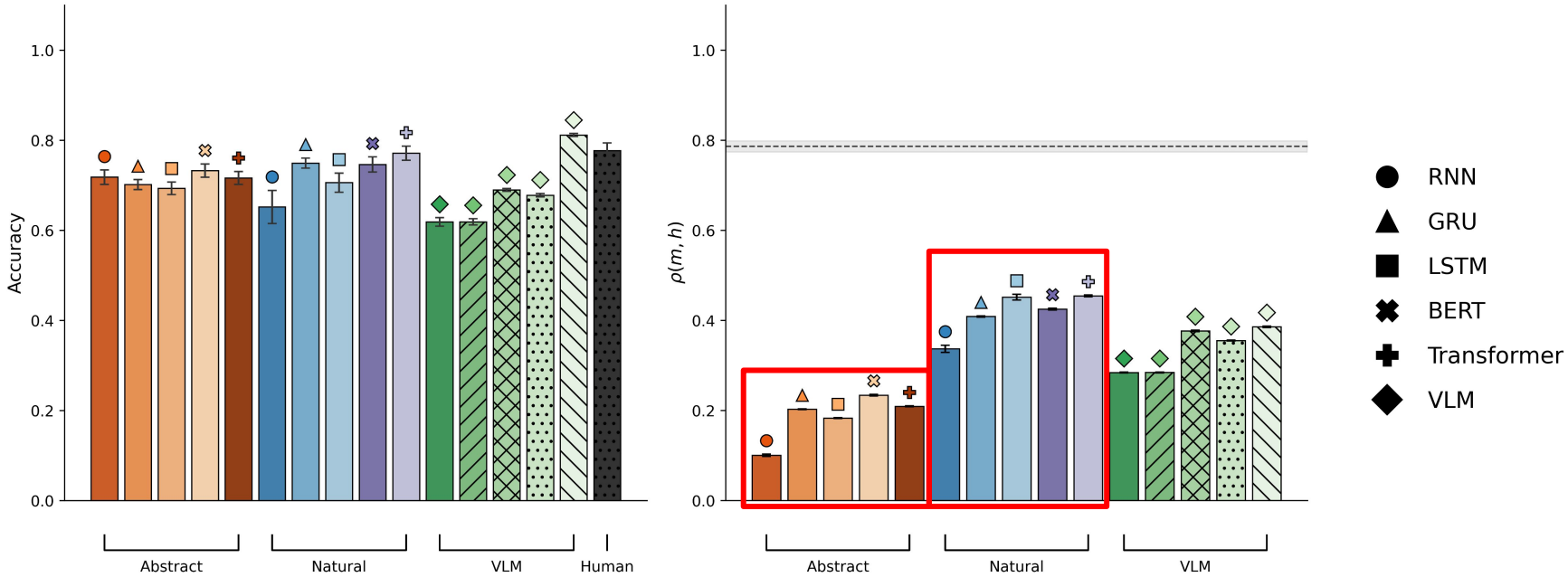


HvM stimuli: Majaj et al., J. Neurosci. 2015. iWISDM tasks: Lei et al., CoLLAs 2025.



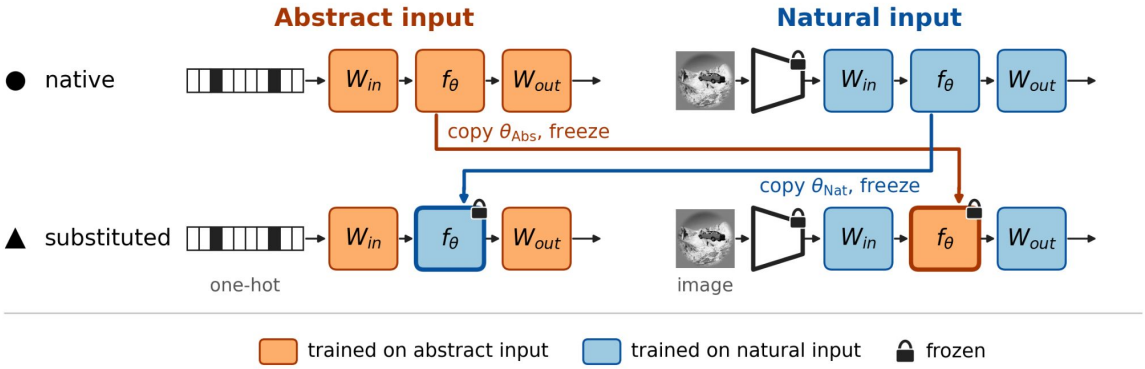
Human consistency: Rajalingham et al., J. Neurosci. 2018.

Natural models are more human-consistent



Abstract models: checkpoints with validation accuracy closest to 80%, to match Natural models.
Vision encoder: He et al., CVPR 2016 (ResNet-50); Radford et al., ICML 2021 (CLIP).

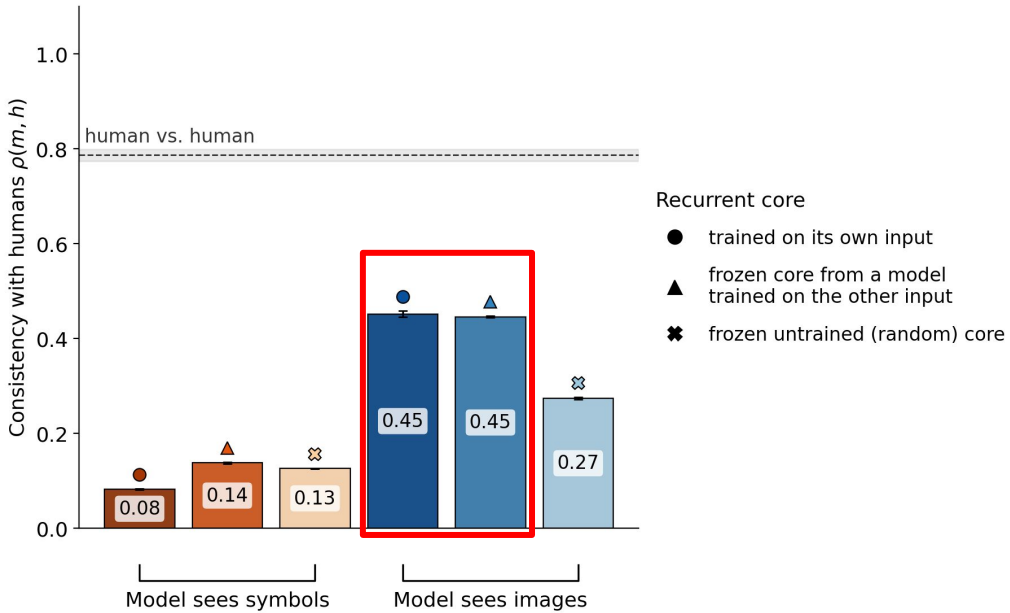
Where does the gap come from?



- Input vs. recurrent weights
- **Substituted experiment:**
  Take the recurrent core trained on one modality (e.g. abstract), and retrain the input and output projections on the other modality (e.g. natural)
- Substituted models help us separate the two mechanisms

Input modality determines behavioral consistency

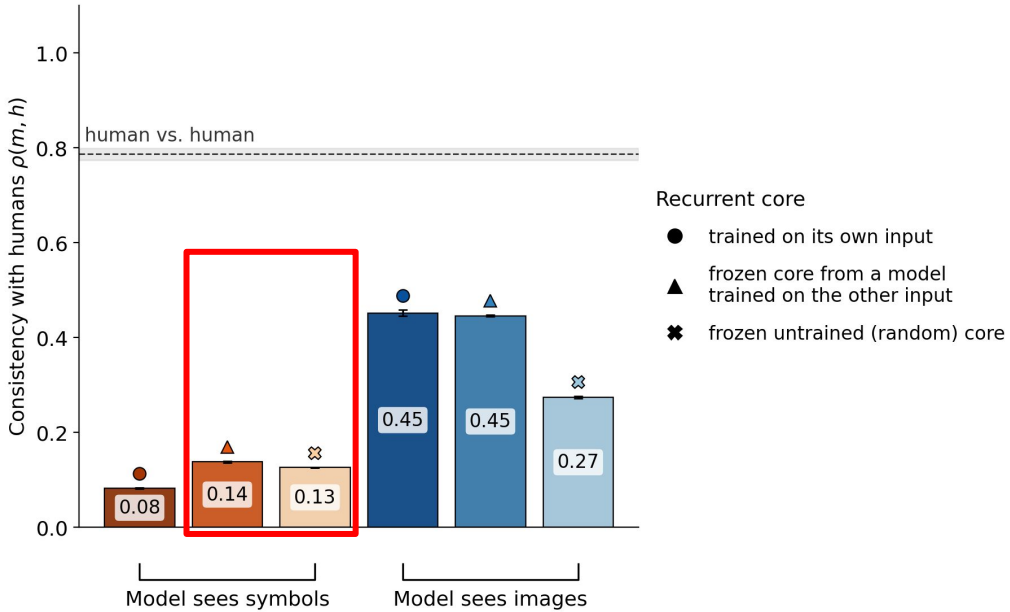
- A model retrained on **natural** input is just as human-consistent as a model trained on natural input from scratch
- However, a model retrained on abstract input barely beats a model with randomly initialized recurrent weights



LSTM models; images encoded by ResNet-50 CLIP. Abstract models use final checkpoints.

Input modality determines behavioral consistency

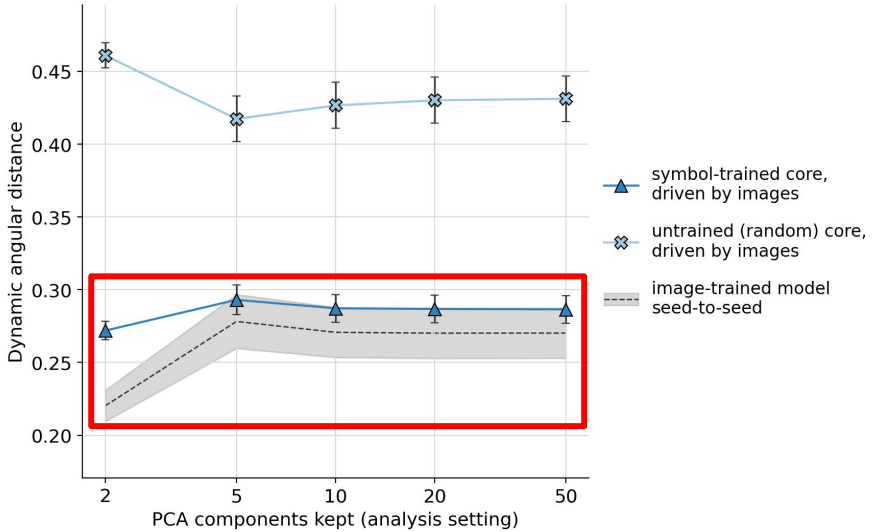
- A model retrained on **natural** input is just as human-consistent as a model trained on natural input from scratch
- However, a model retrained on abstract input barely beats a model with randomly initialized recurrent weights



LSTM models; images encoded by ResNet-50 CLIP. Abstract models use final checkpoints.

Abstract weights produce natural-like dynamics

- Do the retrained and native models converge to similar internal solutions?
- DSA measures the distance between the models' internal dynamics
- At PC ≧ 5, Natural retrained model produces dynamics that lie within the seed-to-seed variability of native natural models



LSTM models; images encoded by ResNet-50 CLIP. DSA: Ostrow et al., NeurIPS 2023.

Takeaways

- Sensory-realistic models track human behavior more closely than abstract-input models
- Input realism matters more than sequence architecture
- The gap lives in the input representation, not in the recurrent weights
- Findings obtained from abstract task models should be more carefully interpreted in light of their potential behavioral limitations







Thank you!