Evaluating Object-Object Relationships in Artificial Vision Models

Georgina Woo, RT Pramod, Nancy Kanwisher
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Introduction

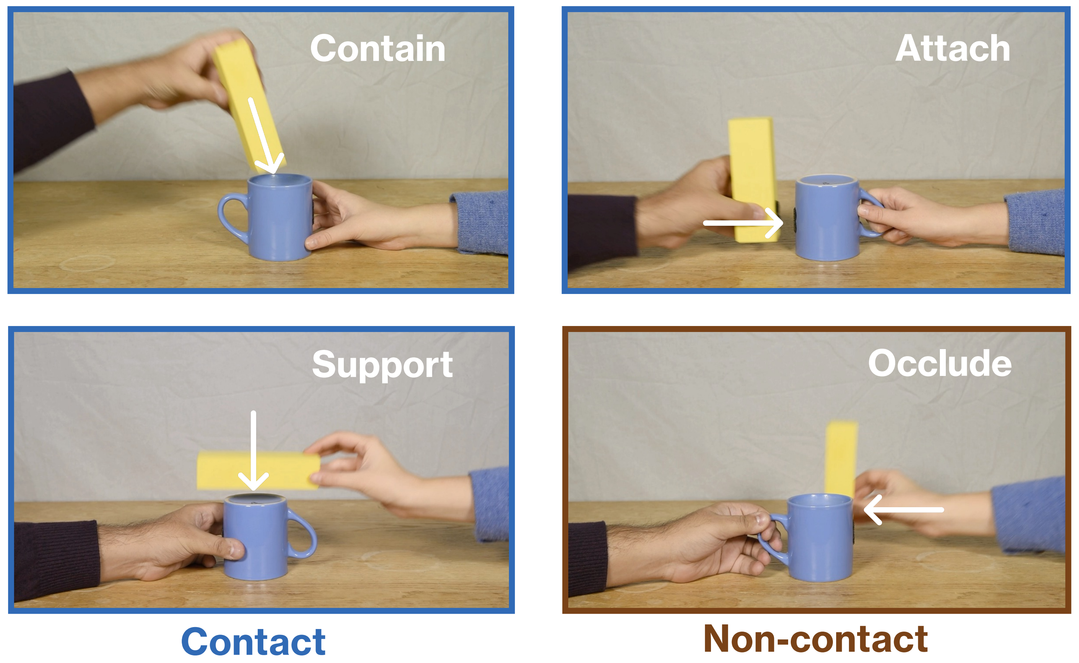
Successful interaction with the world requires **understanding the physical structure** of the scene in front of us.

This includes extracting the relationships between objects like containment, support, attachment, and occlusion. While people are quite accurate at recognizing these relationships and using them to make predictions, we still don't understand the computational mechanisms underlying such behavior.

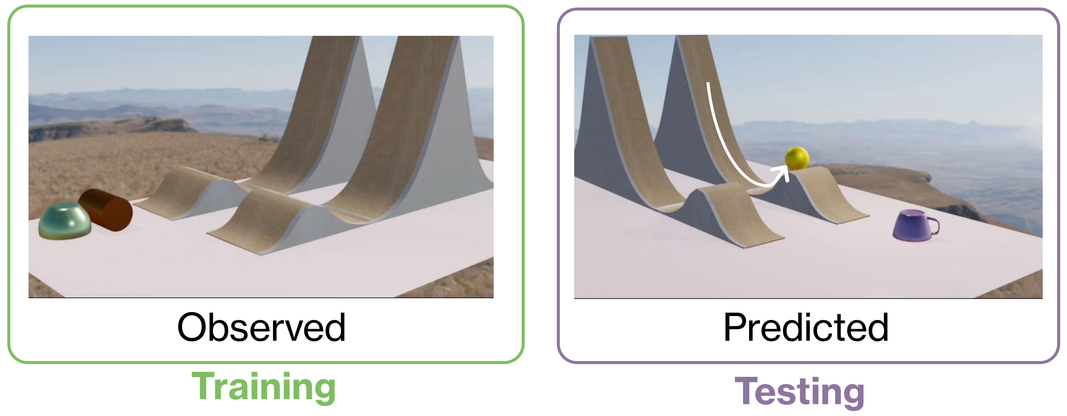
In this study, we test various **deep learning models** as computational hypotheses of **object relation understanding**.

Methods

Experiment 1 & 2: Contact Classification and Detection



Experiment 3: Contact Prediction



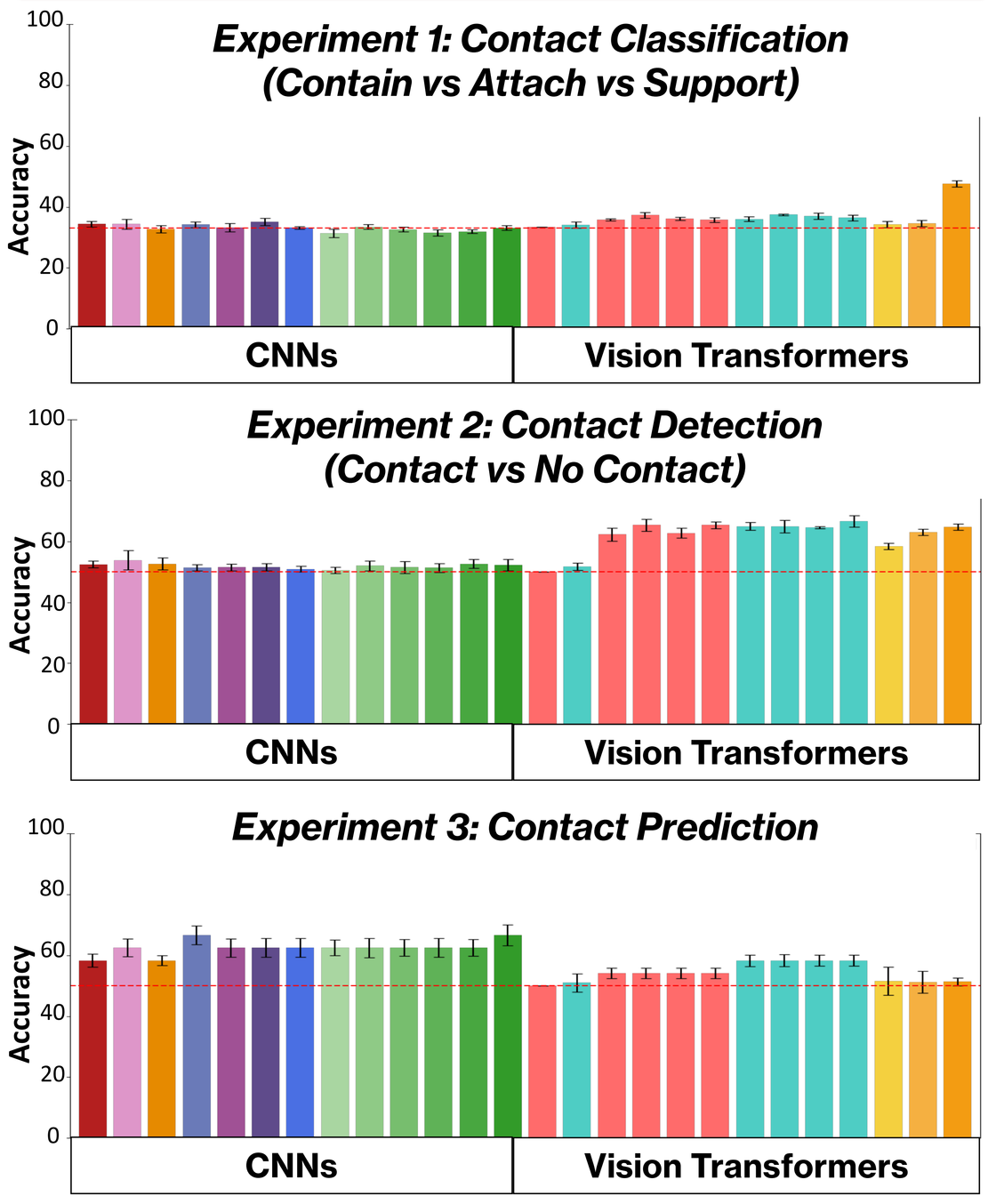
Models

- 13 CNNs trained on *Kinetics-400*
- 11 Vision Transformers trained on *Kinetics-700, Physion,* and *VideoMix2M*

We trained linear SVM classifiers on model features for **contact classification, contact detection and future contact prediction**, and tested the classifiers on different scenarios.

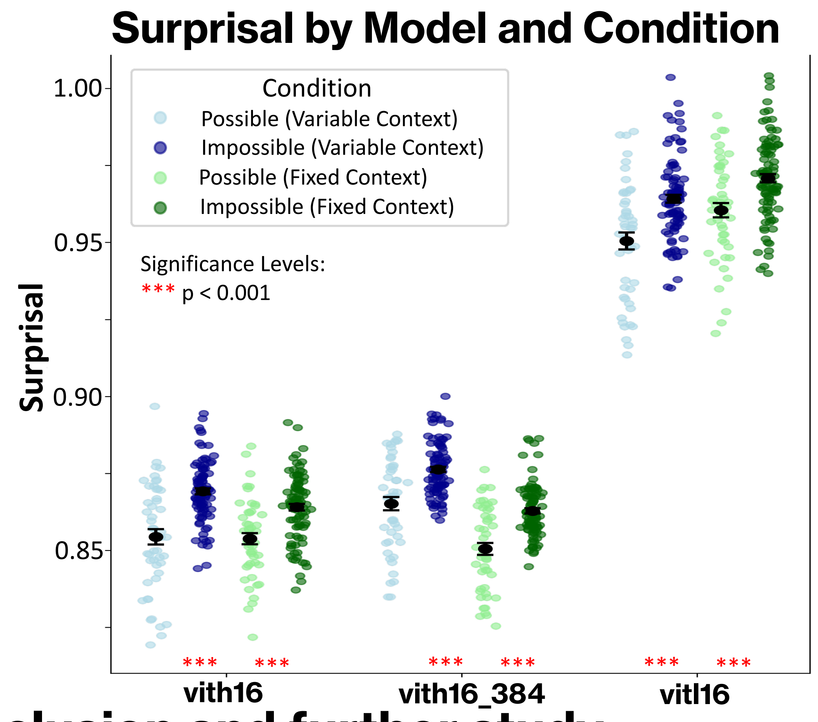


Results



Surprisal Based Evaluation

Another way to assess physical reasoning is the **violation-of-expectation** approach, which assumes physically intuitive models show **higher surprisal** for impossible events. Consistent with prior work, models assign significantly higher surprisal to impossible versus possible videos ($p < 0.001$). Yet, their inability to predict object contact despite surprisal extraction suggests reliance on **pixel-level patterns** rather than genuine **object-level reasoning**.



Conclusion and further study

Although some computational models perform above chance on relational tasks, they fail on many scenarios obvious to humans, suggesting that standard deep learning methods alone may be **insufficient to fully capture human physical scene understanding**. Future work will compare human behavior and model performance using a new static-image benchmark.



This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.