A Benchmark Dataset for Perceiving Object Physical Relationships in Human and Machine Vision

Georgina Woo, Nancy Kanwisher, RT Pramod
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.

Humans perceive physical relationships between objects automatically and accurately. **What computational mechanisms underlie this ability?**

We introduce **ObjectRel**, a benchmark of 14,400 synthetic images spanning five physical relationships, and test various **deep learning models** as computational hypotheses of **object relation understanding in humans**.

ObjectRel

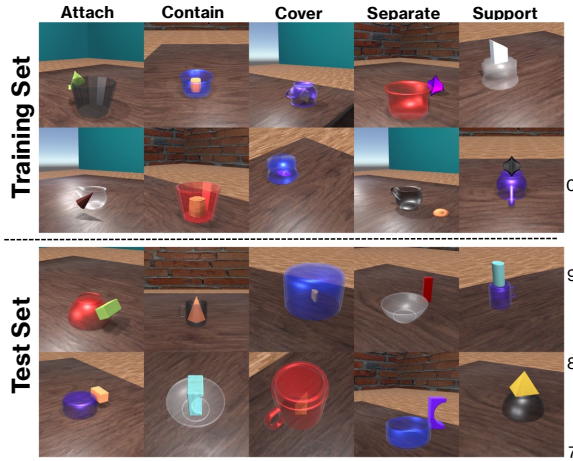
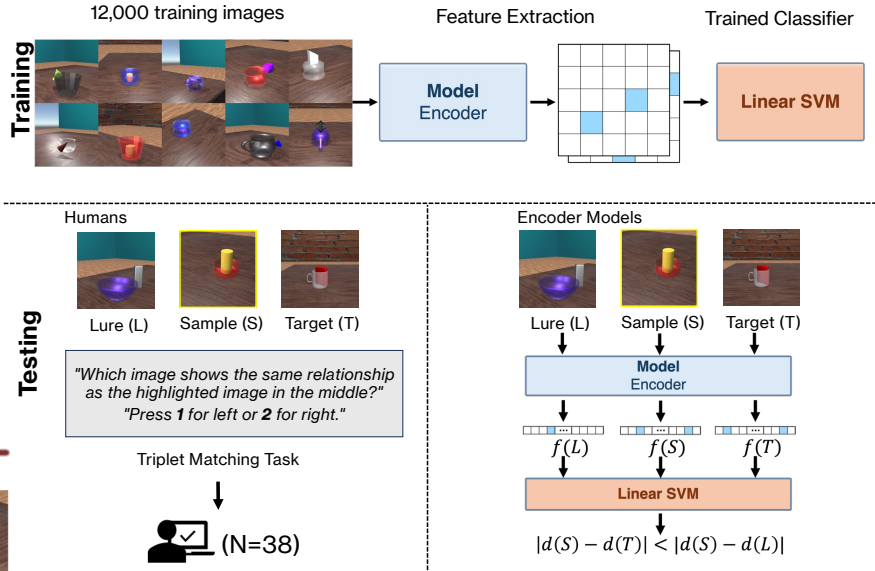


Image Dataset Design

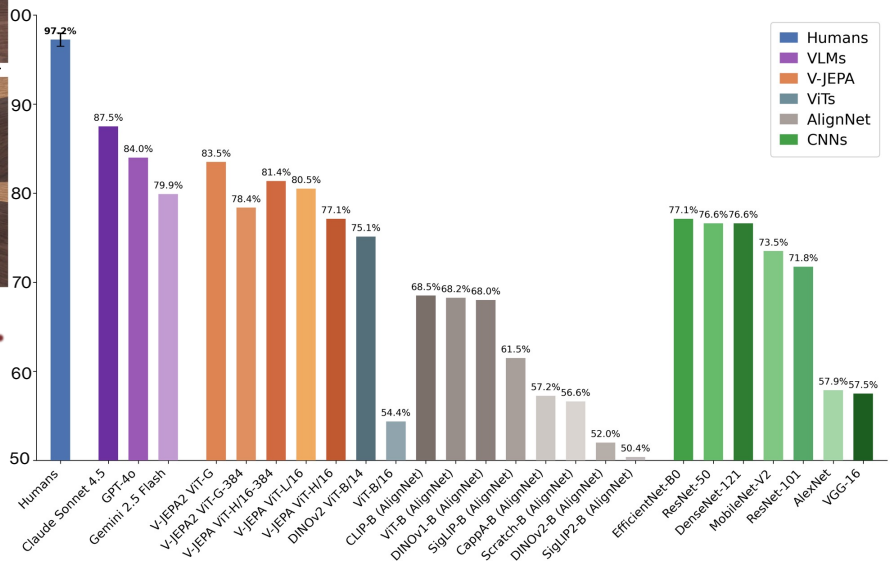
Training and test sets use entirely non-overlapping objects.

The 3 vessel types and 8 solid shapes in the test set never appear during training, requiring models to generalize to novel objects rather than memorize specific object-relationship pairings.

Methods

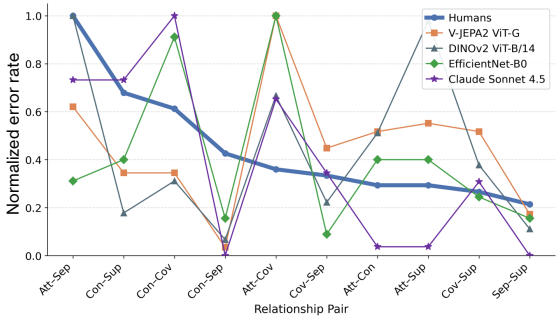


Match to sample: All models fall short of human accuracy



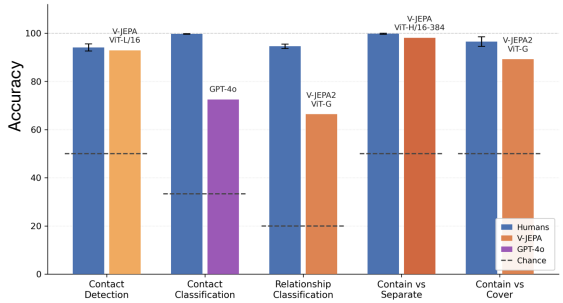
Triplet task accuracy for humans and all models.

Model and human error patterns are uncorrelated



Per-pair error rates for each model are normalized by that model's maximum error and sorted by human difficulty

Other classification tasks show similar gaps



Conclusions and further study

All models **fall short of human accuracy** despite task simplicity and access to far more training data.

Model and human error patterns are uncorrelated, suggesting that the models solve the task via different computational strategies than humans. The models also struggle on other object relationship classification tasks.

Future work will examine whether fine-tuning improves model-human alignment and develop a large-scale video dataset for dynamic relational reasoning.