# Multimodal Language Models show Evidence of Embodied Simulation

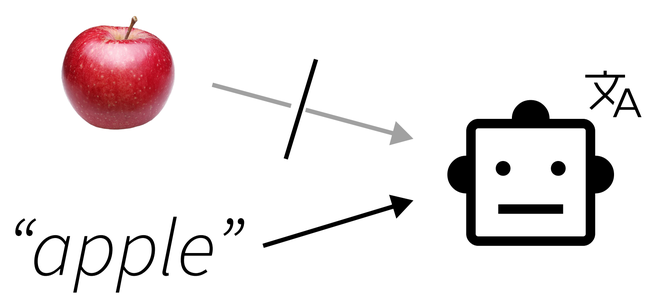
**Cameron R. Jones**
cameron@ucsd.edu @camrobjones

UC San Diego

**Sean Trott**
sttrott@ucsd.edu @Sean_Trott

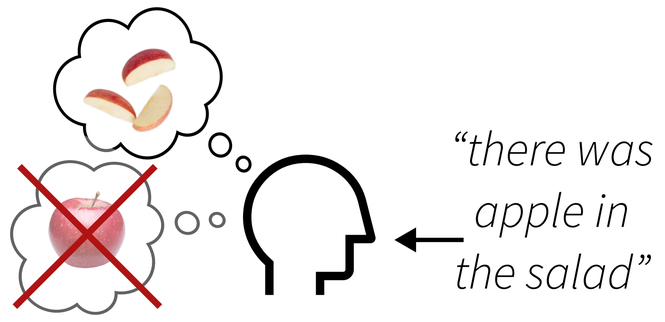
## INTRO | DO MLLMs AND HUMANS GROUND LANGUAGE SIMILARLY?

### LLMs LACK GROUNDING



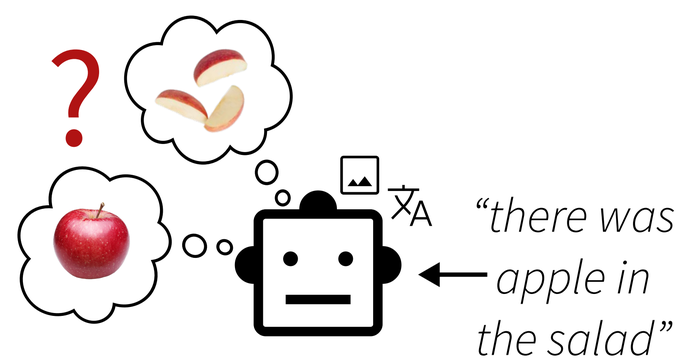

Harnad (1990), Bisk et al. (2020)

### HUMANS GROUND VIA EMBODIED SIMULATION




Barsalou (1999), Pecher & Zwaan (2012)

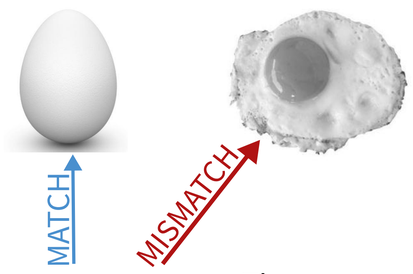
### DO MLLMs ACTIVATE **IMPLICIT** FEATURES OF LANGUAGE?




## METHOD | WE ADAPTED EMBODIED SIMULATION TASKS FOR MLLMs

### SHAPE

Pecher et al. (2009)




The egg was in the refrigerator

The egg was in the skillet

### COLOR

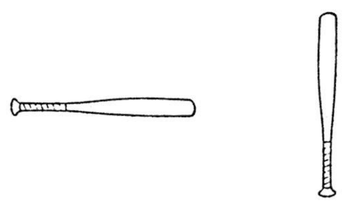
Connell (2007)

Joanne never took milk in her coffee

Joanne always took milk in her coffee

### ORIENTATION

Stanfield & Zwaan (2001)



Derek swung his bat as the ball approached

Derek held his bat high as the ball approached

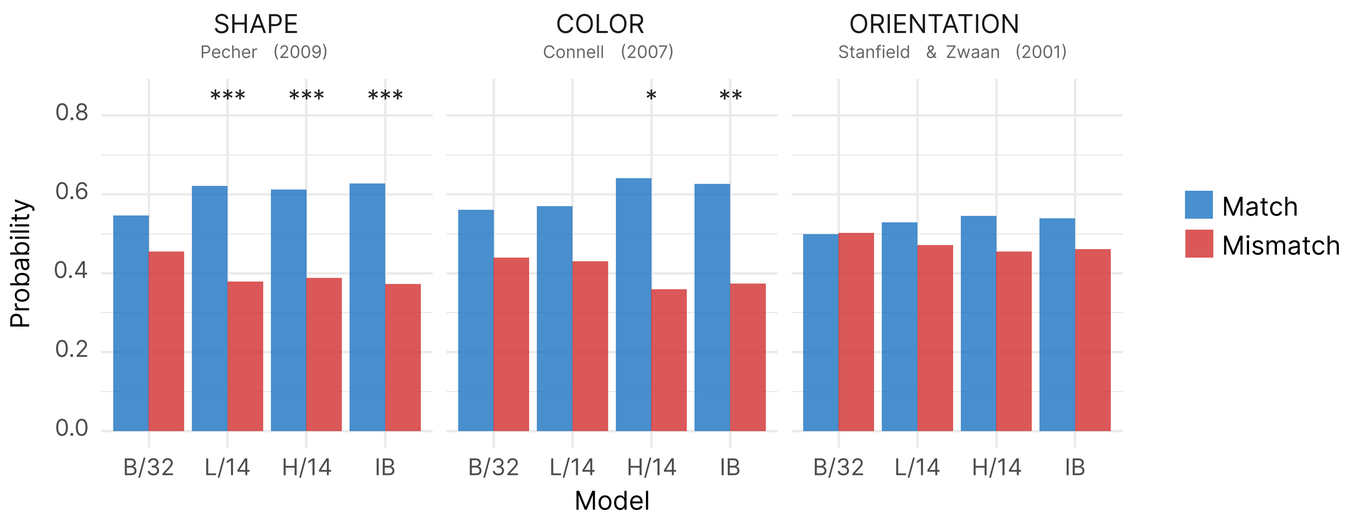
Images either **matched** or **mismatched** **implicit** sensorimotor features in sentences

**MLLMs:** CLIP B/32, L/14, H/14 and ImageBind
(Radford et al. 2021, Ilharco et al., 2021, Girdhar et al., 2023)

We find the **softmax probability** over the dot product of representations of image-text pairs

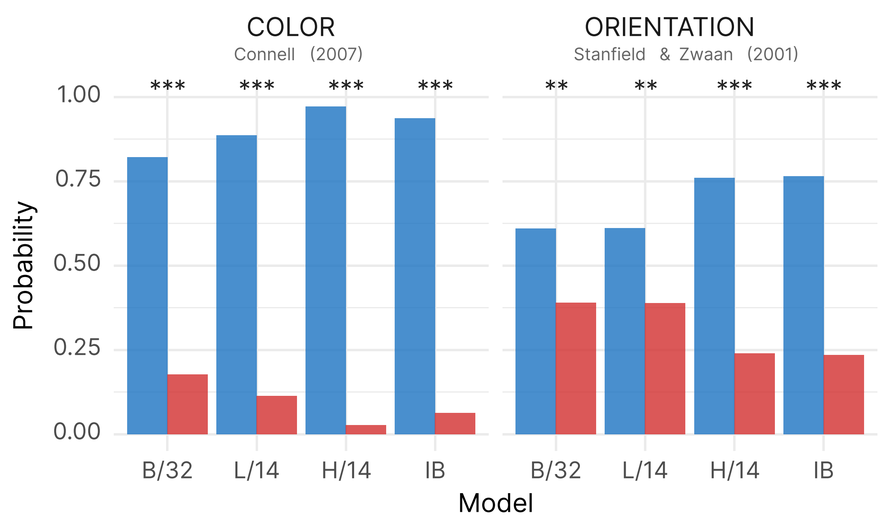
**RQ: Do MLLMs show an effect of implicit sensorimotor feature match?**

## RESULT | MLLMs "SIMULATE" IMPLICIT SHAPE & COLOR, BUT NOT ORIENTATION




ImageBind and CLIP ViT H/14 assigned higher probability to images that matched implicit **SHAPE** and **COLOR**, but not **ORIENTATION**.

## MANIPULATION CHECK | ALL MLLMs MATCH EXPLICIT LABELS




We ran a manipulation check with explicit text labels (e.g. "a horizontal bat"). All models showed an effect.

## DISCUSSION | MLLMs AS A MECHANISM FOR SENSORIMOTOR GROUNDING?

**ARE MLLMs SIMULATING? |** Similar results have been treated as evidence of embodied simulation in humans.

**OR ARE HUMANS NOT? |** Alternatively, the results could be treated as a deflationary explanation of human experiments.

**TEXT ENCODER BOTTLENECK |** Sensitivity to explicit labels suggests text encoders are bottleneck to sensitivity for implicit features.

## REFERENCES

Barsalou, L. W. (1999). Perceptual symbol systems. Behav Brain Sci, 22(4), 577-660.

Bisk, Y. et al. (2020). Experience grounds language. arXiv:2004.10151.

Connell, L. (2007). Representing object colour in language comprehension. Cognition, 102(3), 476-485.

Girdhar, R. et al. (2023). Imagebind: One embedding space to bind them all. Proc IEEE/CVF Conf Comput Vis Pattern Recognit, 15180-15190.

Harnad, S. (1990). The symbol grounding problem. Physica D, 42(1-3), 335-346.

Ilharco, G. et al. (2021). Openclip.

Pecher, D. et al. (2009). Language comprehenders retain implied shape and orientation of objects. Q J Exp Psychol, 62(6), 1108-1114.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. Int Conf Mach Learn, 8748-8763.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. Psychol Sci, 12(2), 153-156.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. PLoS One, 7(12), e51382.