Do Multimodal Large Language Models Show Evidence of Embodied Simulation?

Cameron Jones, Benjamin Bergen, and Sean Trott
Department of Cognitive Science, UC San Diego

Embodied simulation theories argue that human comprehenders ground language by simulating the sensorimotor experiences that it describes (Barsalou, 1999; Harnad, 1990). Empirical support for this theory shows that comprehenders activate relevant sensory representations when processing language (Hauk et al., 2004; Zwaan et al., 2002). However, there is debate over whether this activation is constitutive of grounded language understanding or merely epiphenomenal (Mahon & Caramazza, 2008). Large Language Models (LLMs) provide an opportunity to test embodied simulation theories. Lacking bodies or sensorimotor experience, LLMs ought not to display the same kinds of simulation effects as humans. Multilingual Large Language Models (MLLMs) have been proposed as a potential solution to the grounding problem by integrating language and other modalities (Li et al., 2019; Radford et al., 2021). But there is disagreement over whether MLLMs exhibit the necessary interaction between linguistic and sensorimotor inputs that underpins grounding in humans (Mollo & Millière, 2023; Pavlick, 2023).

We address both of these debates by adapting experiments used to provide evidence for grounding in humans. In these experiments, human participants were faster to verify that an image contained the same object as a previously presented sentence if the image matched visual features implied by the sentence. For instance, the sentence "He put the pen in the [cup/drawer]" facilitated recognition of a [horizontal/vertical] pen. Importantly, orientation is not explicitly mentioned in the sentence, suggesting that comprehenders automatically ground language by simulating features in other modalities. This paradigm has been extended to shape (Zwaan et al., 2002), color (Connell, 2007; Zwaan & Pecher, 2012), and size in vision, as well as volume in audition (Winter & Bergen, 2012).

We adapt all 5 of these stimulus sets for MLLMs to test whether they show the implied feature match effect that has been taken as evidence for sensorimotor grounding in humans. In Experiment 1, we use text-only LLMs as a baseline: asking whether p('horizontal') > p('vertical') following the sentence "He placed the pen in the cup. Now, the pen is…". An effect here would suggest that distributional linguistic information alone is sufficient to identify the implied sensorimotor features. In Experiment 2, we test whether MLLMs such as CLIP (Radford et al., 2021) and ImageBind (Girdhar et al., 2023) show an effect of implied feature match by measuring whether their representations of matching images and sentences are more similar than representations of non-matching pairs. A match effect would suggest that MLLMs reproduce an analogous process to embodied simulation in humans; linguistic stimuli that imply sensorimotor features cause models to activate relevant elements of their multimodal embedding space.

While these experiments could show that MLLMs are *sensitive* to implicit feature match across modalities, we are also interested in whether these models are capable of *explaining* the effects embodied simulation on human comprehenders. We test this in Experiment 3 by using (M)LLM predictions as a statistical baseline. We ask whether implicit feature match has an effect on human reaction times over and above the variance explained by model predictions. The extent to which LLMs and different MLLM architectures (e.g. fusion vs dual encoder) explain this effect provides evidence that the associations learned by the system are sufficient to account for embodied simulation in humans. The paradigm not only allows us to ask whether MLLMs show evidence of sensorimotor grounding in humans, but also to evaluate specific MLLM architectures as explicit candidate mechanisms for theories of sensorimotor grounding in humans (McClelland et al., 2020; Meteyard et al., 2012).

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., & Nisnevich, A. (2020). Experience Grounds Language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Connell, L. (2007). Representing object colour in language comprehension. *Cognition*, *102*(3), 476–485. https://doi.org/10.1016/j.cognition.2006.02.009

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). Imagebind: One embedding space to bind them all. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron*, *41*(2), 301–307.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). *VisualBERT: A Simple and Performant Baseline for Vision and Language* (arXiv:1908.03557). arXiv. http://arxiv.org/abs/1908.03557

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, *102*(1–3), 59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of Age: A Review of Embodiment and the Neuroscience of Semantics. *Cortex*, *48*(7), 788–804. https://doi.org/10.1016/j.cortex.2010.11.002

Mollo, D. C., & Millière, R. (2023). *The Vector Grounding Problem* (arXiv:2304.01481). arXiv. https://doi.org/10.48550/arXiv.2304.01481

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. http://arxiv.org/abs/2103.00020

Winter, B., & Bergen, B. (2012). Language comprehenders represent object distance both visually and auditorily. *Language and Cognition*, *4*(1), 1–16. https://doi.org/10.1515/langcog-2012-0001

Zwaan, R. A., & Pecher, D. (2012). Revisiting Mental Simulation in Language Comprehension: Six Replication Attempts. *PLOS ONE*, *7*(12), e51382. https://doi.org/10.1371/journal.pone.0051382

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168–171.