Embodied Simulation in Multimodal Models Using Affordance Stimulus: A Probing Study

Kaiwen Bian, Alice Li, Sean Trott, & Cameron Jones

UC San Diego
Cognitive Science

Background

Embodied simulation, the theory that language comprehension in humans is rooted in physical experiences, **offers a framework for evaluating AI's potential to navigate and understand the world more intuitively**. Building on such theory and subsequent findings by that Large Language Models (LLMs) partially grasping object affordances without direct worldly experience, our study extends these inquiries to the capabilities of Multimodal Large Language Models (MLLMs) to discern object affordances. We present state-of-the-art MLLMs with 36 scenarios featuring Afforded, Non-Afforded, and Canonical objects to examine whether they assign higher probabilities to images which represent objects that are afforded or not in the context of their associated scenario.

Hypothesis

Multimodal Large Language Models (MLLMs) are capable of differentiating between Afforded and Non-Afforded objects, assigning higher probability scores to images which represent feasible actions in context as opposed to those that do not, despite having no lived experiences.

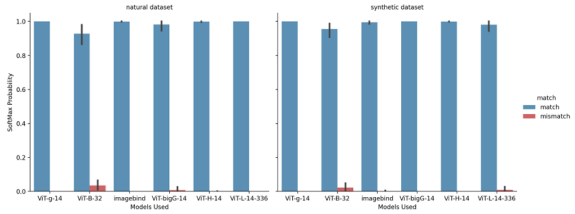
Data Collection & Normalization

Adapted scenarios from Jones et al. (2022):

- 18 scenarios with (Afforded, Non-Afforded, Canonical) objects.
- Created dataset with 3 images for each scenario with **synthetic** (DALL-E)& **natural** (manual internet collection) data set.
- **Normalized** with 6 MLLMs (ImageBind, CLIP ViT-B-32, CLIP ViT-L-14-336, ...) by presenting each data (image + text) and retrieve cosine distance to take Softmax operation and determine if they assign correct probability to each image.

Natural images:

Synthetic images with DALL-E:

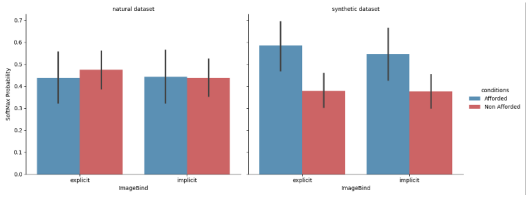


Softmax Result for 6 models' performance on distinguishing the natural & synthetic dataset created

Main Study 1

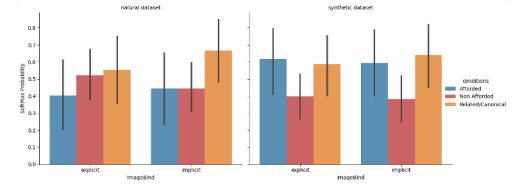
The main study focuses on Meta AI's ImageBind, which was one of the most powerful models at the time of this study and whose embedding space binds multiple sensory inputs together (visual + text). This is study is pre-registered on OSF.

- 2 different prompt types were tested:
  - **Explicit question** - "Brad was sitting in his office when an intruder threatened him with a knife. Which object did Brad use to defend himself?"
  - **Implicit 'this' statement** - "Brad was sitting in his office when an intruder threatened him with a knife. Brad used this to defend himself."
- **Afforded** and **Non-Afforded** images were presented to the model for the primary research question and Canonical image was used in the follow up manipulation check.
- Extracted **Cosine distance** (measures similarity where smaller distance is greater similarity) between the model's representation of the scenario description and each image.
- **Softmax function** applied to convert to probability distribution, allowing quantitative assessment of MLLM's ability to comprehend affordances.
- **ImageBind demonstrates some ability of understanding affordance relationships in the synthetic data,** but displays the reverse effect for natural data.



Afforded:
- Synthetic [t=2.14, p<0.05]
  - Reject H0
- Natural [t=-1.39. p>0.05]
  - Fail to reject H0

SoftMax result for ImageBind with 2 data set (natural, synthetic) and 2 relationships (Afforded & Non-afforded)

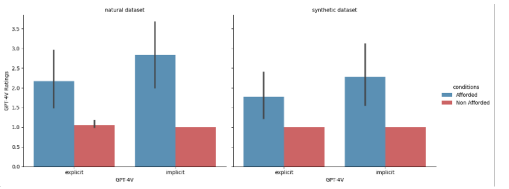


Canonical:
- Synthetic [t=2.38, p<0.05]
  - Reject H0
- Natural [t=2.12, p<0.05]
  - Reject H0

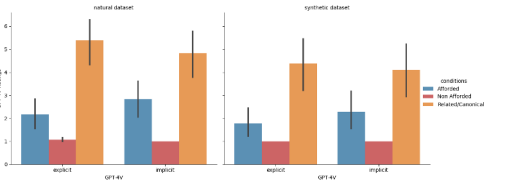
SoftMax result for ImageBind with 2 data set (natural, synthetic) and 3 relationships (Afforded, Non-afforded, & Canonical)

Follow-Up Study 2

Our follow-up study investigates GPT-4 Vision's (GPT-4V) ability to recognize object affordances in a multimodal context. Due to the lack of access to GPT-4V's internal embeddings, the study was conducted through GPT-4V's API, where each scenario was presented to the model to elicit a "sensibility ranking" from 1 (least sensible) to 7 (most sensible) for the use of each object.



Rating results using GPT-4V with 2 data set (Natural, Synthetic) and 2 relationships (Afforded & Non-Afforded)



Rating results using GPT-4V with 2 data set (Natural, Synthetic) and 3 relationships (Afforded, Non-Afforded, & Canonical)

The results were analyzed using a linear mixed-effects model, and we see that GPT-4V **significantly distinguished between Afforded and Non-Afforded objects**, as well as **between Canonical and Non-Afforded objects**, across both prompt types.

This highlights GPT-4V's potential in understanding the utility of objects in various contexts accurately.

Discussions

Results indicate that GPT-4 Vision can effectively differentiate between objects that are contextually appropriate for a given task and those that are not, **suggesting an emergent ability to understand the world's affordances**. Conversely, ImageBind displays a limited response to these distinctions, showing reduced sensitivity in recognizing affordances, particularly within the dataset reflecting real-world imagery and only marginal sensitivity in the dataset composed of artificially generated images. This suggests that even **without any embodied experiences, MLLMs can acquire implicit knowledge about the world.** However, this capability is not inherent to all models, underscoring that the mere integration of multimodal data does not universally afford models more advanced cognitive abilities.

- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
- Benjamin Bergen. 2015. Embodiment, simulation and meaning. In The Routledge handbook of semantics, pages 142–157. Routledge.
- Connell, L. (2007). Representing object colour in language comprehension. Cognition, 102(3), 476– 485. https://doi.org/10.1016/j.cognition.2006.02.009
- Feldman, J., & Narayanan, S. (2004). Embodied meaning in a neural theory of language. Brain and Language, 89(2), 385–392. https://doi.org/10.1016/s0093-934x(03)00355-9
- Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high- dimensional and embodied theories of meaning. Journal of Memory and Language, 43(3), 379–401. https://doi.org/10.1006/jmla.2000.2714
- 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
- Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. K. (2022). Distributional Semantics Still Can't Account for Affordances. Proceedings of the Annual Meeting of the Cognitive Science Society.
- Jones, C. R., & Trott, S. (2023). Multimodal Language Models Show Evidence of Embodied Simulation.
- Pecher, D., van Dantzig, S., Zwaan, R. A., & Zeelenberg, R. (2009). Short article: Language Comprehenders retain implied shape and orientation of objects. Quarterly Journal of Experimental Psychology, 62(6), 1108–1114. https://doi.org/10.1080/17470210802633255
- Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. Psychological Science, 12(2), 153–156. https://doi.org/10.1111/1467-9280.00326