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

Kaiwen Bian
University of California, San Diego
kbian@ucsd.edu

Alice Li
University of California, San Diego
axl001@ucsd.edu

Abstract

Building on theories by Glenberg & Robertson (2000), which posited that language comprehension is partly rooted in embodied experience, and subsequent findings by Jones et al. (2022) that Large Language Models (LLMs) can partially grasp 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 (18 scenarios with 1 from each synthetic and natural dataset that we created) adapted from previous work featuring Afforded, Non-Afforded, and Canonical objects, and examine whether MLLMs assign higher probabilities to images which represent objects that are afforded or not in the context of their associated scenario. 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 physical 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.

1 Introduction

Advancements in computational technology and the diversity of available datasets have led to substantial improvements in Large Language Models (LLMs) and Computer Vision Models (CVMs), leading to the emergence of Multimodal Large Language Models (MLLMs) that integrate together both textual and visual data to improve understanding and interaction capabilities (Dosovitiskiy et al, 2021). Despite these advances, significant gaps remain in our understanding of how MLLMs synthesize and interpret this integrated data, particularly in relation to human-like language comprehension and real-world interaction. The "black box" nature of these models further complicates efforts to evaluate their reliability and interpretability. While MLLMs can process textual and visual data, the internal mechanisms by which they arrive at conclusions remain largely inaccessible. This echoes cognitive linguists' concerns about symbol grounding issues—how abstract computational entities relate to tangible, real-world entities and experiences (Harnad, 1990).

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. This theory suggests that human language comprehension is profoundly tied to physical experience(Bergen, 2015; Feldman and Narayanan,2003), and that reactivation of sensorimotor experiences is essential for resolving the symbol grounding issue. Empirical support for this theory is evident in phenomena such as the "match effect", where comprehension in humans is enhanced when sensory experiences align with linguistic inputs (Stanfield & Zwann, 2001; Pecher et al., 2009; Connell, 2007).

A potential implication of embodied simulation theories is that LLMs–which lack embodied experience–will be unable to model some aspects of human language comprehension. Some researchers have tested this empirically by asking whether LLMs are sensitive to distinctions that humans are thought to rely on simulation of embodied experience. Glenberg and Robertson (2000) found that while distributional models like Latent Semantic Analysis (LSA) are adept at capturing certain linguistic patterns, they are not sensitive to affordances — the set of actions an agent can take with objects in a given environment — in the same way humans are. (Gibson, 2014). Humans, drawing from their lived experiences, instinctively understand that a chair affords sitting for human-like bodies but not for elephants. This human sensitivity to affordances, which distributional models fail to replicate, raises a question: Is the inability of models like LSA to grasp such concepts due to the inherent limitations of using only distributional language statistics without connection to perceptual or actionable experiences?

Jones et al. (2022) revisit this question using contemporary LLMs like GPT-3 and replicating experiments that examine their sensitivity to the affordance of actions. Their findings show that while models do show sensitivity to afforded vs non-afforded items and are able to capture a third of the effect seen in human judgment, they still do not fully incorporate the nuanced understanding of physical interactions. As they still do not adequately account for affordances in comparison to humans, there remains the question of whether this gap in performance is due to the lack of physical interaction experience or the inherent nature of how these models are trained primarily on textual data (Bisk et al. 2020).

The present work explores whether models that integrate both textual and visual data can surpass the limitations of purely text-based systems by synthesizing information across modalities to understand object affordances and contextual interactions more deeply. By leveraging the strengths of both visual and textual data, these artificial systems might offer new pathways to address the symbol grounding problem, embodied cognition, and approach closer to human-like understanding and reasoning in real-world scenarios. Furthermore, by examining their performance in contextually rich, multimodal scenarios, we seek to enhance our understanding of AI's interpretability and reliability.

2 Study 1

Study 1 investigates the responsiveness of MLLMs to the concept of affordances, aiming to evaluate their linguistic comprehension and the extent to which they can apply knowledge beyond their training datasets.

The primary research question is whether MLLMs assign higher probability to images representing objects for which a linguistically described action is possible (Afforded), compared to actions that are not possible (Non-Afforded). To address this, we compare the probability assigned to Afforded images against Non-Afforded images across 18 different scenarios drawn from Glenberg & Robertson (2000), each associated with two image stimuli: Afforded and Non-Afforded. We also conduct a manipulation check, comparing the probability assigned to Canonical images against Non-Afforded images, to assess the MLLM's sensitivity to canonical affordances of objects.

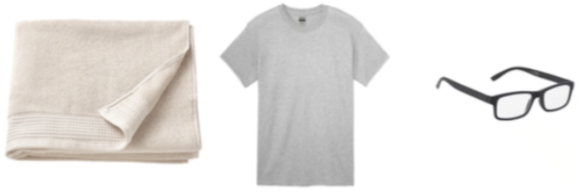

Figure 1: Example of natural datapoint.

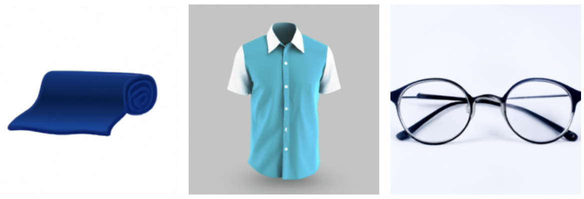

Figure 2: Example of synthetic datapoint.

3 Methods

3.1 Dataset

We adapted the scenarios from Jones et al. (2022), where they presented text to LLMs describing 18 scenarios each with three different types of objects for each scenario (Afforded, Non-Afforded, Canonical) and created a new dataset with 3 images for each scenario using both synthetic & natural data collection methods.

Here is an example data point, along with the set of possible images presented to the model: Scenario: "After wading barefoot in the lake, Erik needed something to get dry. What would he use?" Objects: Canonical Object: [towel], Afforded Object: [shirt], Non-Afforded Object: [glasses]

Natural images

Synthetic images with DALL-E

For our experiment with MLLMs, we developed custom datasets based on 18 distinct scenarios. Each scenario includes two images for every ob-

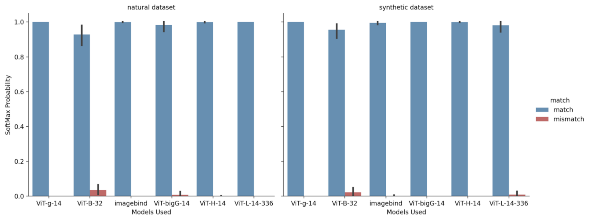

Figure 3: Softmax Result for 6 models' performance on distinguishing the natural & synthetic dataset created.

ject–one generated synthetically using OpenAI's diffusion model DALL-E (https://openai.com/dall-e) and the other sources manually from the internet. Synthetic images might present more simplified versions of objects, while real-world images provide more complex visual cues. By incorporating both a synthetic image set and natural image set we can evaluate the MLLM's ability to generalize across a broader spectrum of visual representations The DALL-E images were generated with the prompt "realistic object with white background" to ensure greater uniformity across the generated images, making up our synthetic dataset. Our natural dataset was handpicked using Google image search for <term>, and we chose the first image which matched closest with our specifications: the object image must have a white background, be positioned in its most natural angle, and have no odd discoloration.

To ensure that the images were representative of the objects they were intended to depict, we conducted a normalization study using 6 different MLLMs (CLIP ViT-B-32, CLIP ViT-L-14-336, CLIP ViT-L-14, CLIP ViT-H-14, CLIP ViT-G-14, CLIP ViT-bigG-14) (Radford et al., 2021). This process involved presenting the models with 3 sets of images, each representing a Canonical object, an Afforded object, or a Non-Afforded object (for example, an image of a towel, glasses, and a shirt), paired with their corresponding labels ('towel,' 'glasses,' 'shirt'). We calculated the cosine distance between each text label and its respective image to assess their similarity. Subsequently, we determined the probability for each image-text pair by computing the dot product of the vector space distances and applying a Softmax function. This analysis confirmed that the model's performance on matching word-image pairs met our expectations.

3.2 Model

Our model of interest for this study trained by Meta AI, ImageBind (Girdhar et al., 2023), is among the most powerful models at the time of this study. ImageBind is an MLLM which learns a joint embedding across six modalities from images, text, audio, depth, thermal to IMU data using Transformer architecture.

3.3 Procedure

For each of the 18 scenarios from each dataset, two different prompt types were tested: one with an explicit question (i.e. "Brad was sitting in his office when an intruder threatened him with a knife. Which object did Brad use to defend himself?") and another with a reference to an implicit 'this' statement (i.e. "Brad was sitting in his office when an intruder threatened him with a knife. Brad used this to defend himself."). By using both an explicit and an implicit prompt, we can evaluate how the framing of the question may potentially influence the MLLM's interpretation and response accuracy. Explicit prompts may guide the model more towards a targeted answer, while implicit prompts have the model infer and extract deeper levels of understanding on its own. Only the Afforded and Non-Afforded images were presented to the model for the primary research question. The Canonical image was used in the follow up manipulation check. After presentation of each prompt, we extracted the cosine distance between the model's representation of the scenario description and each image. This distance measures the similarity between the textual scenario and the visual representation of each object, where smaller distances indicate greater similarity. Probabilities were then calculated by soft-maxing the cosine distance between each image and the scenario description. This allows for quantitative assessment of MLLM's ability to comprehend affordances.

We have pre-registered our experiment on OSF https://osf.io/86aer.

3.4 Results

For the synthetic dataset Afforded condition, we reject the null hypothesis [t=2.14, p<0.05] where the probability assigned to Afforded images is significantly higher than that assigned to the Non-afforded image. But for the natural dataset, we fail to reject the null hypothesis [t=-1.39. p>0.05].

For both manipulation checks, however, the MLLM demonstrated sensitivity to the difference between Canonical and Non-Afforded images (syn-

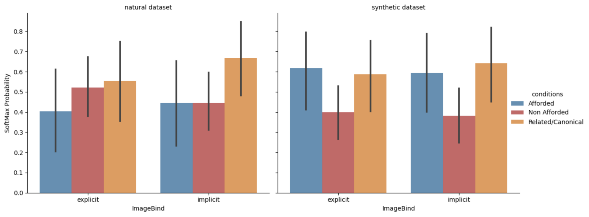

Figure 4: SoftMax result for ImageBind with 2 dataset (natural, synthetic) and 3 relationships (Afforded, Non-afforded, & Canonical)

thetic Canonical condition [t=2.38, p<0.05] and natural Canonical condition [t=2.12, p<0.05]).

3.5 Discussion

From our results, we conclude that ImageBind demonstrates some ability of understanding affordance relationships in the synthetic dataset but not in the artificial dataset if not a reverse effect. In the synthetic dataset, ImageBind appears to have a grasp of affordances, as indicated by the statistically significant difference in the model's ratings for Afforded versus Non-Afforded objects (t=2.14, p<0.05). This suggests that when presented with images that were created in accordance with an artificial system's understanding of the object, ImageBind can apply learned patterns to hypothesize about potential uses for the object. However, the same does not hold in the natural dataset, where the model was not able to consistently differentiate between Afforded and Non-Afforded images, failing to reject the null hypothesis here (t=-1.39, p>0.05). It's possible that differences in the datasets themselves contributed to the results. As models like ImageBind are usually exposed to more natural than synthetic-like images during training, the model may have greater sensitivity to subtle variations in natural scenarios that are not as pronounced in synthetic representations. Thus, the variance in model performance across datasets also reflects the varying exposure of the model to more commonly encountered image types during its training phase.

Furthermore, ImageBind demonstrates a clear sensitivity to canonical affordances across both datasets, reliably distinguishing between Canonical and Non-Afforded images (synthetic Canonical condition [t=2.38, p<0.05] and natural Canonical condition [t=2.12, p<0.05]). This consistency across both types of data highlights the model's ability to anchor its understanding in commonly accepted object functions and uses, regardless of whether the representation is synthetic or natural.

So while MLLMs like ImageBind can identify typical object uses, their capacity to understand less conventional affordances appears to be limited by the nature of their training data and the contexts they have been exposed to. This could reflect a gap in current MLLM training that could be bridged in the future by incorporating more diverse and context-rich experiences, more closely imitating "lived experiences" that humans have.

4 Study 2

Our second, follow-up study, delves into the ability of GPT-4 Vision (GPT-4V) to discern and prioritize object affordances in a multimodal context. Similar to ImageBind, the primary focus is to investigate whether GPT-4V can differentiate between afforded and non-afforded objects, based on scenarios accompanied by images.

4.1 Methods

4.1.1 Dataset

Study 2 uses the same dataset as the customized dataset we used in Study 1 (18 scenarios for the natural dataset and 18 scenarios for the synthetic dataset, with three images for each scenario).

4.1.2 Model

GPT-4 consistently outperforms existing LLMs on traditional ML benchmarks. It scores 40% higher than GPT-3.5 on internal evaluations, and GPT-4V augments GPT-4's capabilities by processing images alongside text https://openai.com/index/gpt-4v-system-card/. At the time when Study 2 was conducted, GPT-4V was the state-of-the-art MLLM.

4.1.3 Procedure

Since we do not have access to GPT-4V's internal embeddings at the time of this study, this test was conducted through interacting with GPT-4V's API. To obtain a "sensibility ranking" for the use of each object in the images within the context of the 18 scenarios, we issued prompts to the system that were preceded by a specific instruction designed to elicit this ranking:

"*In this task, you will read short passages and look at an image of an object. Please rate how sensible it would be to take the action described in the last sentence using the object in the image in the context of the whole passage. The scale goes from 1 (virtual nonsense) to 7 (completely sensible). Be sure to*

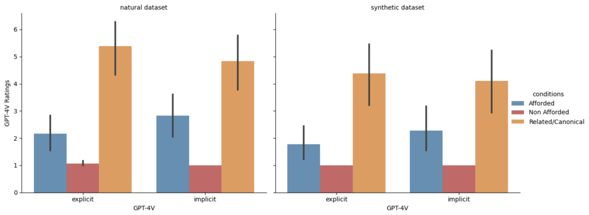

Figure 5: Rating results using GPT-4V with 2 dataset (Natural, Synthetic) and 3 relationships (Afforded, Non-Afforded, Canonical.)

read the sentences carefully. Please respond only with a number between 1 and 7."

We used the elicited ratings as a reflection of the model's judgment on how appropriate an object is for the described action. Additionally, to ensure deterministic answers, GPT-4V's system temperature is set to 0.

4.2 Results

For both our comparisons, we constructed a linear mixed-effects model. Between Afforded vs Non-Afforded Objects, our model analysis showed that Afforded objects were given higher sensibility rankings compared to Non-Afforded objects. From analysis, we see that (Afforded: M=2.10, SD=1.52, Non-Afforded: M=1.02, SD=0.19, $p < 0.001$). This suggests that GPT-4V is capable of distinguishing between objects that are appropriate for a given action and those that are not. Further analysis includes comparing sensibility rankings between Canonical objects and Non-Afforded objects, where Canonical objects expectedly received significantly higher rankings than Non-Afforded objects (Canonical: M=4.86, SD=2.28, Non-Afforded: M=1.02, SD=0.19, $p<0.001$), confirming that GPT-4V distinguishes between objects for their typical use and objects which are clearly used atypically in the given scenarios.

Regarding the prompt type, there appears to be no significant interaction between condition and prompt type ($p = 0.179$). We can assume from this that the difference between using an explicit as opposed to an implicit prompt does not impact the model's ability to discern affordances.

4.3 Discussion

Although the embeddings could not be extracted from GPT-4V for a more nuanced understanding of the model's capabilities when it comes to recognizing object affordances, using a heuristic of assigning ratings still underscores GPT-4V's ability to discern between afforded and non-afforded objects. The model's consistent performance regardless of the prompt type used implies that GPT-4V's ability to discern affordances is robust across different linguistic framing.

These findings, which illustrate an underlying grasp of sensibility across different affordance conditions, highlights the model's potential for nuanced understanding of context and utility, towards human-like processing. Even without direct sensory experience and interaction with the physical world, GPT-4V's training of textual and visual data has enabled the model to develop the capability to interpret and contextualize object affordances with a depth that approaches human-like intuitions about the world.

5 Conclusion

We found that GPT-4 Vision possesses the capability to effectively distinguish between contextually appropriate and inappropriate objects, while ImageBind exhibits limited sensitivity with regards to affordances. This variation highlights that the integration of multimodal data alone does not guarantee enhanced cognitive abilities across different MLLMs.

6 Limitations

A primary limitation of this study is the restricted dataset size on which the models were tested, as well as the type of data. The synthetic images, created using OpenAI's DALL-E, may not be fully representative of the complexities found in real-world objects. The natural images, sourced manually from the internet, may bias towards representations that are more commonly shared online. Additionally, with only 36 scenarios between the synthetic and natural datasets, the limited dataset size can impact the statistical power of the analysis. And while this study explores the capabilities of MLLMs, we lack a direct comparison with human data on the same tasks. Without comparison to human benchmarks, it's difficult to gauge whether the models' performance is honestly reflective of human-like understanding.

7 References

References

[1] 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. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. https://doi.org/10.48550/arXiv.2010.11929.

[2] Benjamin Bergen. *Embodiment, simulation and meaning*. In The Routledge Handbook of Semantics, pages 142–157. Routledge, 2015.

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

[4] J. Feldman, and S. Narayanan. *Embodied meaning in a neural theory of language*. Brain and Language, 89(2), 385–392, 2004. https://doi.org/10.1016/s0093-934x(03)00355-9.

[5] J. J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, 2014.

[6] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. *Imagebind: One Embedding Space to Bind Them All*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15180-15190, 2023.

[7] A. M. Glenberg, and D. A. Robertson. *Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning*. Journal of Memory and Language, 43(3), 379–401, 2000. https://doi.org/10.1006/jmla.2000.2714.

[8] S. Harnad. *The Symbol Grounding Problem*. Physica D: Nonlinear Phenomena, 42(1–3), 335–346, 1990. https://doi.org/10.1016/0167-2789(90)90087-6.

[9] C. R. Jones, T. A. Chang, S. Coulson, J. A. Michaelov, S. Trott, and B. K. Bergen. *Distributional Semantics Still Can't Account for Affordances*. Proceedings of the Annual Meeting of the Cognitive Science Society, 2022.

[10] C. R. Jones, and S. Trott. *Multimodal Language Models Show Evidence of Embodied Simulation*. 2023. https://doi.org/10.17605/osf.io/37pqv.

[11] D. Pecher, S. van Dantzig, R. A. Zwaan, and R. Zeelenberg. *Language Comprehenders retain implied shape and orientation of objects*. Quarterly Journal of Experimental Psychology, 62(6), 1108–1114, 2009. https://doi.org/10.1080/17470210802633255.

[12] R. A. Stanfield, and R. A. Zwaan. *The effect of implied orientation derived from verbal context on picture recognition*. Psychological Science, 12(2), 153–156, 2001. https://doi.org/10.1111/1467-9280.00326.

A Appendix

Further details of the models used for Study 1 and Study 2.

ViT-B/32: Trained on 400 million 224x224 pixel image-text pairs over 32 epochs, patch size of 32px and 120M parameters. It is the base model from Radford et al., 2021.

ViT-L/14: Trained on 400 million 224x224 pixel image-text pairs over 32 epochs, fine-tuned at 336px for an additional epoch, patch size of 14px and 430M parameters. Best-performing model from Radford et al., 2021.

ViT-H/14: Trained on LAION 2B dataset for 16 epochs, 1B parameters and based on the CLIP architecture.

ViT-G/14: Trained on the LAION 2B dataset for about a third of the epochs, 2B parameters and based on the CLIP architecture.

ImageBind: Consists of a Transformer architecture, in which the text and image encoders are based on the ViT-H/14 model. An MLLM which learns a joint embedding across six modalities from images, text, audio, depth, thermal to IMU data.