Do Multimodal Large Language Models and Humans Ground Language Similarly?

Cameron R. Jones*
University of California, San Diego
Department of Cognitive Science
cameron@ucsd.edu

Benjamin Bergen
University of California, San Diego
Department of Cognitive Science
bkbergen@ucsd.edu

Sean Trott
University of California, San Diego
Department of Cognitive Science
sttrott@ucsd.edu

Large Language Models (LLMs) have been criticized for failing to connect linguistic meaning to the world—for failing to solve the "symbol grounding problem." Multimodal Large Language Models (MLLMs) offer a potential solution to this challenge by combining linguistic representations and processing with other modalities. However, much is still unknown about exactly how and to what degree MLLMs integrate their distinct modalities—and whether the way they do so mirrors the mechanisms believed to underpin grounding in humans. In humans, it has been hypothesized that linguistic meaning is grounded through "embodied simulation," the activation of sensorimotor and affective representations reflecting described experiences. Across four preregistered studies, we adapt experimental techniques originally developed to investigate embodied simulation in human comprehenders to ask whether MLLMs are sensitive to sensorimotor features that are implied but not explicit in descriptions of an event. In Experiment 1, we find sensitivity to some features (color and shape) but not others (size, orientation, and volume). In Experiment 2, we identify likely bottlenecks to explain an MLLM's lack of sensitivity. In Experiment 3, we find that despite sensitivity to implicit sensorimotor features, MLLMs cannot fully account for human behavior on the same task. Finally, in Experiment 4, we compare the psychometric predictive power of different MLLM architectures and find that ViLT, a single-stream architecture, is more predictive of human responses to one sensorimotor feature (shape)

* Corresponding author.

Action Editors: Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. Submission received: 18 December 2023; revised version received: 30 April 2024; accepted for publication: 19 June 2024.

https://doi.org/10.1162/coli_a_00531

than CLIP, a dual-encoder architecture—despite being trained on orders of magnitude less data. These results reveal strengths and limitations in the ability of current MLLMs to integrate language with other modalities, and also shed light on the likely mechanisms underlying human language comprehension.

1. Introduction

Advances in Large Language Models (LLMs) have led to impressive performance on a range of linguistic tasks (Hu et al. 2022; Trott et al. 2023; Dillion et al. 2023; Chang and Bergen 2024). Yet despite these improvements, a common criticism of contemporary LLMs is that they are trained on linguistic input alone (Bender and Koller 2020; Bisk et al. 2020). Lacking bodies or sensorimotor experience, they have no way to "ground" the symbols they are trained on, which some (Harnad 1990) have argued is necessary for true language understanding. A natural solution to this problem could be found in Multimodal Large Language Models (MLLMs) (Driess et al. 2023; Girdhar et al. 2023; Huang et al. 2023; Radford et al. 2021), which learn to associate linguistic representations with information from other *modalities*, such as vision or sound. However, there is still considerable disagreement over whether MLLMs exhibit the necessary interaction between linguistic and sensorimotor inputs that appears to underpin grounding in humans (Mollo and Millière 2023; Tong et al. 2024; Shanahan 2023; Bisk et al. 2020). How tightly do MLLMs integrate representations of information from distinct inputs (e.g., vision and language), and how *humanlike* is the manner in which they do this?

We address this gap directly by turning to the evidentiary basis for grounding in humans (Bergen 2015). A range of experimental evidence suggests that humans ground language—in part—through *embodied simulation* of the sensorimotor experiences that language describes. By applying techniques originally developed to probe the representations and mechanisms underlying grounding in human language comprehension, we can ask to what extent MLLMs use analogous representations and mechanisms. Our approach builds on past work investigating whether and to what extent equipping neural networks with grounded (e.g., visual) information produces more humanlike representations (Bruni, Tran, and Baroni 2014; Chrupała, Kádár, and Alishahi 2015; Kiros, Chan, and Hinton 2018; Kádár, Chrupała, and Alishahi 2017; Peng and Harwath 2022; Harwath and Glass 2015); indeed, there is evidence that representations from grounded models are more useful for predicting explicit human judgments, for example, similarity ratings about concrete words (Kiros, Chan, and Hinton 2018). This approach offers a unique opportunity for cross-disciplinary symbiosis: MLLMs (and LLMs) with different architectures can act as implementations of existing theories of language comprehension, and can thus help refine and resolve outstanding debates about the functional role of grounding in human comprehenders.

1.1 Evidence for Embodied Simulation in Humans

The theory of **embodied simulation** claims that human comprehenders ground language by simulating the sensorimotor experiences that it describes (Barsalou 1999; Harnad 1990). For example, understanding a sentence like "She tossed the ball" would involve activating the same (or a subset of the) neural tissue that is involved in either perceiving or participating in that event (Bergen 2012).

This theory enjoys empirical support in the form of both behavioral (Zwaan, Stanfield, and Yaxley 2002; Pecher et al. 2009; Winter and Bergen 2012; Stanfield

and Zwaan 2001) and neuroimaging (Hauk, Johnsrude, and Pulvermüller 2004; Pulvermüller 2013) evidence. One particularly prominent experimental paradigm is the *sentence-picture verification task* (Stanfield and Zwaan 2001; Winter and Bergen 2012; Zwaan, Stanfield, and Yaxley 2002; Pecher et al. 2009). In this task, participants read a sentence (e.g., "He hammered the nail into the wall"), then see a picture of an object (e.g., a nail) and must indicate whether that object was mentioned in the preceding sentence. On critical trials, the depiction of the object is manipulated to either match *implicit features* (e.g., color, shape, orientation) from the sentence or mismatch them. For example, the sentence "He hammered the nail into the wall" implies that the nail is horizontal, while "He hammered the nail into the floor" implies that the nail is vertical. Crucially, these features are not mentioned explicitly in the sentence. Thus, if human participants respond faster or more accurately to pictures that match implied perceivable features of the event described in the sentence, this suggests that they have spontaneously *inferred* this sensory information.

The sentence-picture verification task has been used to demonstrate evidence for embodied simulation across multiple visual features, including orientation (Stanfield and Zwaan 2001), shape (Pecher et al. 2009), distance (Winter and Bergen 2012), and color (Connell 2007; Zwaan and Pecher 2012). It has also been adapted for other modalities, such as sound (e.g., implied volume) (Winter and Bergen 2012). In each case, a facilitatory effect of the experimental manipulation is generally interpreted as reflecting the activation of *implicit sensorimotor features* from linguistic input; participants' responses to real sensorimotor stimuli are influenced by whether the stimulus matches simulated features.

1.2 Debates Over the Interpretation of Evidence

Despite widespread evidence for some degree of sensorimotor activation, there remains considerable debate over which *mechanisms* are most likely to give rise to this effect.

One question revolves around the functional role played by sensorimotor activation. Much of the current evidence cannot adjudicate whether simulation plays an epiphenomenal or necessary role in the process of understanding language (Mahon and Caramazza 2008; Ostarek and Bottini 2021). On functional accounts (Barsalou 1999), embodied simulation is causally important for inferring implicit features. Comprehenders use sensorimotor representations and the simulation process itself to infer that the nail is likely to be *horizontal* if it is being hammered into a *wall*. As Mahon and Caramazza (2008) point out, however, sensorimotor simulation could also occur as a byproduct of spreading activation during language comprehension. On this account, processing of the sentence might generate amodal or linguistic representations of the implied features—for example, "horizontal nail." These amodal representations, in turn, could activate relevant sensorimotor representations without the sensorimotor representations playing any causal role in comprehension. Under this account, the match effect observed on the target picture trial would not require direct activation of *visual* features but could be explained by the activation of correlated *linguistic features*.

Another question is *architectural* in nature: If semantic representations are multimodal, when and how is information from different modalities integrated? Here, the possibilities range from "full integration" (i.e., semantic representations are fully multimodal) to "grounding by interaction" (i.e., semantic representations are partially "symbolic," but can be grounded on the fly) (Mahon and Caramazza 2008; Meteyard et al. 2012).

In each case, answering these questions has proven extremely challenging for the field. It is difficult to specify verbal theories in sufficient detail that they make divergent predictions that could be used to test them. One path forward is to identify suitable *computational operationalizations* of these verbal theories and the effects they predict—such as LLMs and MLLMs.

1.3 Adapting Psycholinguistic Techniques for (M)LLMs

LLMs are neural networks with billions of parameters trained on billions or even trillions of words to predict missing tokens from a sequence. LLMs are trained on linguistic input alone—which is often cited as a limitation with respect to sensorimotor grounding. MLLMs provide a potential solution to this problem by linking linguistic input to another *modality*, typically (though not always) vision (Driess et al. 2023; Girdhar et al. 2023; Huang et al. 2023). For example, CLIP (Contrastive Language-Image Pretraining) models are trained on image-caption pairs (Radford et al. 2021), and thus learn to map flexibly between linguistic and visual representations.

Much remains unknown about exactly how MLLMs' representations differ from those of unimodal LLMs. Additionally, there is considerable variance within MLLMs in terms of their *architecture*, for example, whether linguistic and visual representations are integrated during encoding ("fusion architectures") or encoded separately, then integrated later on ("dual-encoder architectures"). It is unclear how this variation affects the nature of cross-modal representations formed.

Careful application of methodologies developed for humans, such as the sentence-picture verification task (Stanfield and Zwaan 2001; Zwaan, Stanfield, and Yaxley 2002), can address both of these questions. In doing so, it also informs debates around embodied simulation in humans (see Section 1.2).

Here we attempt to address both sets of questions by administering adapted versions of sentence-picture verification tasks to LLMs and MLLMs. First, we test whether MLLMs show a stronger association between matching sentence-picture pairs vs. non-matching pairs. This allows us to ask: *To what extent do MLLM representations encode implicit sensorimotor features, and for which features (e.g., orientation vs. shape) or modalities (e.g., vision vs. sound) are these activations strongest?*

Second, we probe both MLLMs and LLMs to ask when the relevant information (e.g., a nail's implied orientation) becomes accessible. This is important for identifying the mechanisms by which implicit features are activated and diagnosing the reasons for insensitivity where models fail. Specifically, we ask: *Can MLLMs' text-encoders extract implied sensorimotor features from text-only descriptions* and *can MLLMs map explicit descriptions of a feature to matching images or sounds?*

Third, as a further test of mechanism, we use representations elicited from both MLLMs and LLMs to predict human behavior on this task. That is: *Does either model provide a plausible explanatory account of the human match/mismatch effect?*

Finally, we compare a suite of MLLMs, ranging from dual-encoder models to single-stream fusion models, and ask: *Are architectures with more integration between modalities more sensitive to implicit sensorimotor features and better at explaining human data?*

2. Experiment 1

In Experiment 1, we test whether ImageBind (Girdhar et al. 2023), a state-of-the-art MLLM, is sensitive to whether or not sensorimotor features implied by sentences are explicitly present in images and sounds.

2.1 Methods

2.1.1 Materials. We draw experimental stimuli from existing sentence-picture or sentence-sound verification experiments designed to test for effects of sensorimotor simulation in humans. Items for each task are organized as quadruplets, consisting of a pair of sentences and a pair of media stimuli (images or sounds). Sentence pairs differ by implying that an object has a given sensorimotor property (e.g., color or volume). Each of the media stimuli in a pair match one of the sentences by explicitly displaying the implied feature (and therefore mismatch the other sentence; see Figure 1).

We draw stimuli from five different experiments, each of which manipulates a different sensorimotor feature:

1. SHAPE: Pecher et al. (2009) collected a set of 60 quadruplets that varied the implied shape of the object (see Figure 1, top left). A sentence such as "There was an eagle in the [nest/sky]." implies that the eagle's wings are either folded or out-stretched. A pair of black-and-white images of eagles each match one of these sentences by displaying the relevant property.

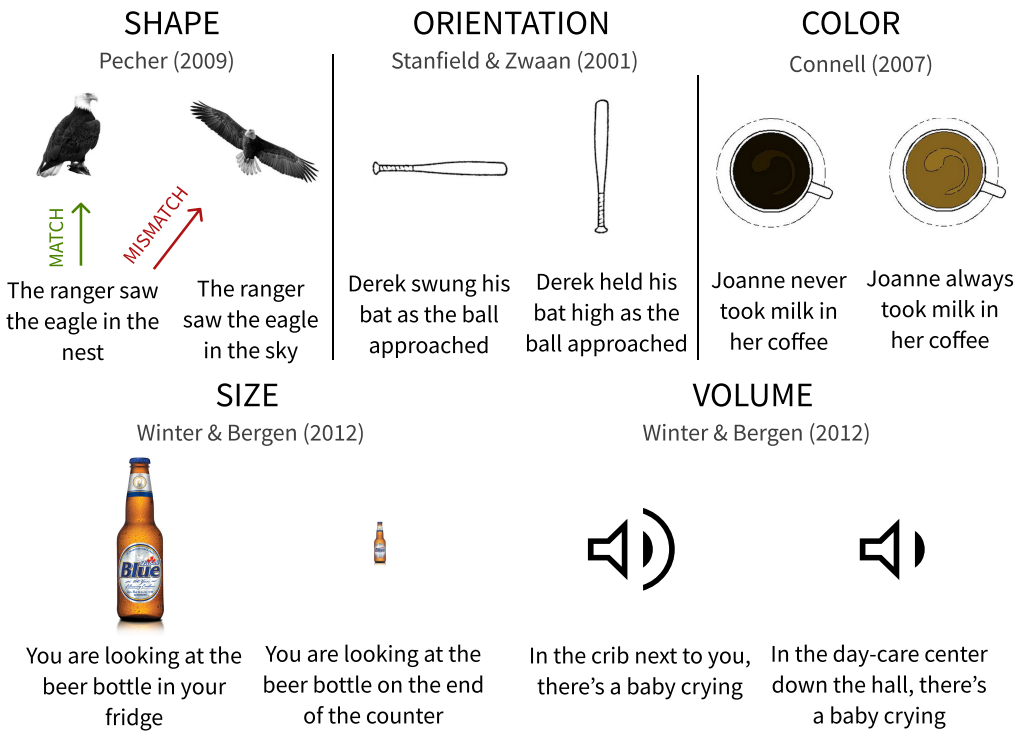

Figure 1
The dataset consisted of pairs of sentences and images or sounds, forming quadruplets. Each sentence in a pair implied that an object had a certain visual or auditory property (e.g., brown color). Each implied sensorimotor property was matched by one of the pair of media (images or sounds). The implied visual properties included SHAPE (**Top Left**, Pecher et al. 2009), COLOR (**Top Center**, Connell 2007), ORIENTATION (**Top Right**, Stanfield and Zwaan 2001), SIZE (**Bottom Left**, Winter and Bergen 2012), and VOLUME (**Bottom Right**, Winter and Bergen 2012).

2. ORIENTATION: Stanfield and Zwaan (2001) collected 24 quadruplets of sentences implying different orientations of an item, and line-drawings that were rotated to match the implied orientation (Figure 1, top center). For instance, "Derek swung his bat as the ball approached" suggests a horizontal bat, while "Derek held his bat high as the ball approached" suggests a vertical bat.
3. COLOR: 12 quadruplets from Connell (2007) vary the implied color of an object. "Joanne [never/always] took milk in her coffee" implies black/brown coffee. The only difference between matching images was their color (Figure 1, top right).
4. SIZE: Winter and Bergen (2012) Experiment 1 manipulates the implied apparent size of objects from the viewer's perspective by varying the viewer's distance from the object, e.g., "You are looking at the beer bottle [in your fridge / on the end of the counter]". Corresponding images display the same object at different scales (Figure 1, bottom left).
5. VOLUME: Winter and Bergen (2012) Experiment 2 manipulates the implied volume of sounds by varying the viewer's distance from the sound, e.g., "In the [crib next to you / day-care center down the hall], there's a baby crying". Matching audio stimuli vary the volume of the sound described in the sentence (Figure 1, bottom right).

2.2 Models

For our preregistered analyses, we selected ImageBind (Girdhar et al. 2023) as our primary model to evaluate due to its strong performance on a variety of tasks and its ability to process inputs from multiple modalities (allowing us to test the vision and audio tasks with the same model). ImageBind is an MLLM that learns a joint embedding across six modalities, including images, text, audio, depth, thermal, and inertial measurement unit data.

Internally, ImageBind uses a Transformer architecture for each modality (Vaswani et al. 2017). For the image encoder, ImageBind uses a Vision Transformer (ViT) architecture that adapts the transformer to handle visual data (Dosovitskiy et al. 2021). The ViT divides an image into fixed-size non-overlapping patches that are then linearly embedded into input vectors. A classification head is attached to the output to produce the final prediction. Despite their simplicity and lack of inductive biases (e.g., convolutional layers), ViTs have achieved competitive performance on various visual tasks, especially when pre-trained on large datasets (Dosovitskiy et al. 2021; Schuhmann et al. 2022). The text encoder, following Radford et al. (2021), is based on the GPT-2 architecture (Radford et al. 2019). Text inputs are appended with a special [EOS] token, and the activations of the highest layer of the transformer at the [EOS] token are treated as the feature representation of the text. Following Gong, Chung, and Glass (2021), ImageBind encodes audio into a 2D spectrogram and uses a ViT to process spectrograms as images.

More specifically, ImageBind uses a frozen CLIP ViT-H/14 (Ilharco et al. 2021) for its vision (630M parameters) and text (302M parameters) encoders. ViT-H/14 is a dual encoder vision language model trained using CLIP. CLIP uses contrastive learning to associate images with text descriptions (Radford et al. 2021). The model jointly trains a ViT image encoder and a text encoder to predict the correct pairings of (image, text)

pairs. This allows CLIP to learn a shared semantic space between images and text. ImageBind is additionally trained using the contrastive loss objective between images and each of the other modalities, to learn to project features from each modality to a shared embedding space.

In addition to ImageBind, to contextualize the model's performance and to test the generalizability of our results, we evaluated 3 additional CLIP models on the vision datasets, and a single CLAP model on the audio dataset:

ViT-B/32: The base model from Radford et al. (2021). ViT-B/32 uses a patch size of 32px and has 120M parameters. It was trained on 400 million 224×224 pixel image-text pairs over 32 epochs.

ViT-L/14: The best-performing model from Radford et al. (2021), described in the paper as ViT-L/14@336px. ViT-L/14 uses a patch size of 14px and has 430M parameters. It was pre-trained in the same manner as ViT-B/32 and then fine-tuned at 336px for one additional epoch.

ViT-H/14: A larger model based on the CLIP architecture (Ilharco et al. 2021). ViT-H/14 has 1B parameters and was trained on the LAION 2B dataset for 16 epochs (Schuhmann et al. 2022).

CLAP: CLAP (Contrastive Language-Audio Pretraining) uses a CNN14 model with 81M parameters as the audio encoder, and a BERT base model with 110M parameters as the text encoder. The audio and text embeddings are projected into a shared 1024-dimensional multimodal space. CLAP was trained on 128,010 audio-text pairs from 4 datasets: FSD50K (36,796 pairs), ClothoV2 (29,646 pairs), AudioCaps (44,292 pairs), and MACS (17,276 pairs).

We access CLIP models using the OpenCLIP Python package (version 2.23.0; ML Foundations 2023), and CLAP through the Python transformers package (version 4.35.2; The HuggingFace Team and Contributors 2023).

2.3 Model Evaluation

To evaluate MLLMs, we implemented a computational analogue of the sentence-picture verification task. Our primary question was whether a model's representation of a given linguistic input (e.g., "He hammered the nail into the wall") was more similar to its representation of an image or sound that matched an implied sensorimotor feature (e.g., horizontal orientation) compared to an image or sound that did not (e.g., a vertical nail). For each sentence-media pair, we found the cosine distance between the ImageBind embedding of the sentence and the media stimulus. This value quantifies the similarity between the linguistic and modal representations within the model

$$similarity_{ij} = cosine(S_i, I_j)$$

where S_i is the embedding for sentence i, I_j is the embedding for image j. To statistically evaluate the model's performance, we constructed a linear mixed-effects model predicting $similarity_{ij}$ on the basis of Match condition, with random intercepts by quadruplet id. We were interested in two different kinds of question, for which we performed separate analyses. First we asked whether there was an effect of match overall, across all datasets. A significant result, where the matching probabilities are greater than mismatching ones, would indicate that the MLLM's representations are sensitive to the sensorimotor properties implied by the linguistic input. In this model we included an additional random intercept by dataset. Second, we asked whether ImageBind

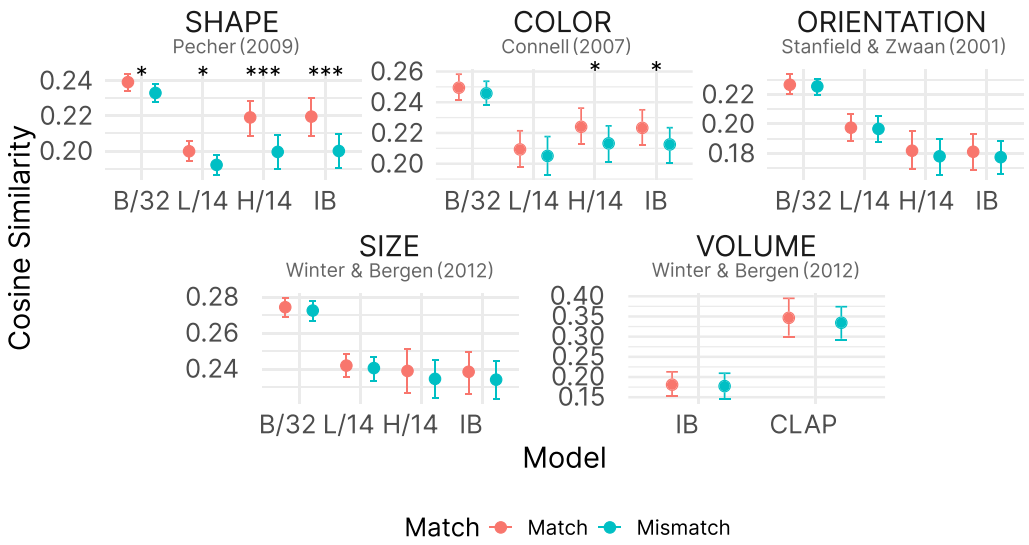

Figure 2
Comparison of mean cosine distances for each model between representations of media that either matched (blue bars) or did not match (red bars) implied sensorimotor features of a sentence. Error bars denote 95% bootstrapped confidence intervals. ImageBind showed significantly higher similarity for SHAPE and COLOR, but not for any other modality. All other vision models showed the SHAPE effect, but only ViT-H/14 showed the COLOR effect. CLAP showed no significant effect on the VOLUME dataset.

showed effects of match within each dataset individually. All stimuli, hypotheses, and analysis were pre-registered on the Open Science Foundation (`https://osf.io/37pqv`).

2.4 Results

We used linear mixed-effects models to analyze the effects of feature match on the cosine distance between ImageBind's representations of text-media pairs. These models account for the hierarchical structure of the data by including random effects for factors such as dataset and item. The *t*-statistics and *p*-values reported indicate the significance of the fixed effects (e.g., match condition) in predicting the dependent variable (cosine distance).

Overall, there was a significant positive effect of match on the cosine distance between ImageBind representation of text-media pairs [$t(469.1) = 3.06), p = 0.002$], meaning that the model's representations of matching text-media pairs were more similar than those of mismatching pairs. However, this effect was only detected in two of the individual datasets: SHAPE [$t(179) = 3.72), p < 0.001$] and COLOR [$t(35) = 2.164), p = 0.037$; all other ps > 0.4]; see Figure 2. The effect of SHAPE appeared to be robust, occurring in all of the other 3 CLIP models (all $p < 0.05$). The color effect, by contrast, occurred only in ViT-H/14 [$t(35) = 2.160), p = 0.038$]. None of the other models showed any sensitivity to any of the other modalities (all $p > 0.13$).

2.5 Discussion

The results suggest that ImageBind is sensitive to whether the images match an object's implied shape or color, but not other sensorimotor modalities including orientation,

size, and volume. The effect of SHAPE was observed in all of the MLLMs we tested, suggesting that this result generalizes to even very small CLIP-based models. Only the largest CLIP model (ViT-H/14, the visual encoder on which ImageBind is based) showed the color effect, suggesting that this result may not be as generalizable.

The effect of shape and color match is analogous in some ways to effects observed in humans; as a result, these effects suggest that exposure to language-image pairs is sufficient to generate sensitivity to implicit relationships between language and the world. In humans, this effect is interpreted as reflecting embodied simulation during language comprehension. Thus, the effect of shape and color match in MLLMs could be interpreted as deflationary, suggesting that simulation is not required to produce the sensitivity that underlies these effects. Alternatively, MLLMs could be interpreted as a mechanistic model of sensorimotor simulation in humans. In this sense, the integration of text and sensorimotor information that occurs in MLLMs would be viewed as a mechanistic explanation of how sensorimotor grounding of language works in human language comprehenders. In either case, the results suggest that learning to project modality-specific representations to a shared representational space through contrastive learning is a viable mechanism for generating sensitivity to the shape and color of objects implied by sentences.

Many of the effects in this study are numerically small ($\sim$0.01). This suggests that, even though the effect is systematic and significant, it has a very small influence on models' representations. In some psychological research, large effect sizes are crucial to assessing the importance of effects (Funder and Ozer 2019). However, in other settings, including the embodied cognition debate, the validity of hypotheses can be sensitive to small effects. Reaction time effects in the original studies are also numerically small, but evidence that these responses systematically vary at all significantly in response to particular, pre-specified aspects of a stimulus are used to adjudicate among theories of the mechanisms that produce behavior (Zwaan, Stanfield, and Yaxley 2002; Connell 2007).

In contrast to SHAPE and COLOR, there was no effect of implied feature match in the ORIENTATION, SIZE, and VOLUME datasets, for ImageBind or any of the other models evaluated. This suggests that these models' mechanisms are *not* capable of generating sensitivity to implicit sensorimotor features in general, undermining claims that MLLMs ground language in sensorimotor representations.

Why do we see sensitivity for some feature types and not others? There are a variety of possible reasons, including differences between the stimulus sets, inherent differences in how features are represented, and differences in the distribution of features across training sets. Shape and color are both relatively predictable on an observer-independent basis. By contrast, orientation, distance, and amplitude can vary greatly depending on the perceiver's perspective. Models may have learned to ignore these features in their representations (indeed, rotation invariance is sometimes seen as a desirable feature of computer vision models [Kalra et al. 2021]). Experiment 2 addresses the question of why and where models succeed or fail more directly.

3. Experiment 2

MLLMs might fail to activate implicit sensorimotor features for at least two reasons. First, they might fail to infer the feature from the linguistic description of the sentence, for example, a model's representation of the text "He hammered the nail into the floor" might not contain information about the orientation of the nail. Alternatively, the model

might successfully infer relevant features from text (e.g., orientation) but fail to exhibit sensitivity to this feature in images.

In order to tease apart these possibilities, we break the task used in Experiment 1 into two distinct sub-tasks. We first construct a set of "explicit" sentences that describe the relevant feature directly (e.g., "the nail is horizontal."). We then test whether models show a higher degree of association between these explicit feature labels and each of the types of stimuli used in Experiment 1.

The first sub-task, **[Implicit Text → Explicit Text]**, tests whether sentences that imply a sensorimotor feature are more closely associated with explicit descriptions of the feature. This allows us to ask whether the implied feature is being encoded in the model's representation of the sentence

In the second sub-task, **[Media → Explicit Text]**, we test whether these explicit text feature labels are more closely associated with the images and sounds depicting the features. This provides a more direct test of whether the model's representations of the media encode sensorimotor features such as orientation and size.

In addition, to investigate the role of learning from multimodal data in generating sensitivity to sensorimotor features in text, we evaluate the representations of a text-only transformer language model (GPT-2 large) for sensitivity on the **[Implicit Text → Explicit Text]** task.[1] GPT-2 large (Radford et al. 2019) is based on a similar architecture to the ImageBind text-encoder and has a similar number of parameters (762M) to the vision and text encoder used by ImageBind (934M). If GPT-2's representations are sensitive to the match between implicit and explicit descriptions of sensorimotor features, it would suggest that exposure to language alone is sufficient to develop this sensitivity. To the extent that contrastive learning from multimodal data increases sensitivity, it would suggest that the training signal from multimodal inputs is important to learning how to transform linguistic representations in a way that emphasizes relevant sensorimotor features.

3.1 Methods

3.1.1 Materials. The implicit sentences and media (images and sounds) were identical to those used in Experiment 1. We generated explicit descriptions of the manipulated sensorimotor feature for each item. These were designed to be concise, neutral, and to reflect the relevant feature as it was contrasted in both the sentence and media pairs.

1. SHAPE: We created short sentences describing an attribute of the object (e.g., "The eagle's wings were [folded/outstretched]").
2. ORIENTATION: All sentences were of the form "the [object] was [orientation]" (e.g., "The nail was [horizontal/vertical]").
3. COLOR: All sentences were of the form "the [object] was [color]" (e.g., "The steak was [red/brown]").
4. SIZE: In order to ensure that the feature made sense with respect to the implicit sentences (that varied distance from the object), we described the object's apparent size from the viewer's perspective. All sentences were of the form "the [object] looks relatively [size] from your perspective"

1 We originally pre-registered this analysis using the ViT-H/14 text-encoder due to a misunderstanding about when the model was frozen during pretraining. We amended our pre-registration before conducting the analysis with GPT-2 large.

(e.g., "The fire hydrant looks relatively [large/small] from your perspective").

5. VOLUME: As with size, these sentences described the sound's apparent volume from the viewer's perspective. All sentences were of the form "the sound of the [object] is relatively [volume] from your perspective" (e.g., "The sound of the handgun is relatively [loud/quiet] from your perspective").

3.1.2 Model Evaluation. The procedure was similar to the one used in Experiment 1, except that rather than comparing the models' representations of images and implicit sentences, we compared representations of (i) explicit sentences to implicit sentences, then (ii) explicit sentences to images/sounds. In each case, we found the cosine similarity between representations of the relevant stimulus, and tested whether similarities show an effect of feature match. As in Experiment 1, we pre-registered our analyses for ImageBind, but report results for the same additional models to provide context and test generalizability. We also presented the explicit-implicit text pairs to GPT-2 to test whether language exposure alone is sufficient to develop sensitivity to implied sensorimotor orientation. We operationalize GPT-2 representations as the mean of the activations in the last layer of the model.

3.2 Results

Cosine distances between GPT-2 representations of the **[Implicit Text → Explicit Text]** sentences were not sensitive to the match vs mismatch effect overall [$t(455) = 0.045), p = 0.964$] or for any of the individual datasets (all $p > 0.819$; see Figure 3). The text-based similarity analysis using ImageBind representations showed a match effect overall [$t(106) = 2.56), p = 0.010$], and in the SHAPE [$t(417) = 2.99), p = 0.003$] but no other datasets (all $p > 0.37$). There was no significant interaction between match and model (GPT-2 vs ImageBind) on cosine similarity [$t(106) = 1.28), p = 0.203$]. Among the other models evaluated, the SHAPE dataset also produced a significant match effect in B/32 [$t(179) = 2.09), p = 0.038$] and H/14 [$t(179) = 3.03), p = 0.003$] representations, and CLAP showed a very marginally significant effect on the VOLUME dataset [$t(71) = 2.00), p = 0.049$]. Several of the effects for the COLOR dataset also approached significance. Given that the dataset here was relatively small and that significance thresholds are necessarily arbitrary (Gelman and Stern 2006), these may well represent real effects that this study was underpowered to detect.

In the **[Media → Explicit Text]** task, ImageBind showed an overall match effect [$t(455) = 6.41), p < 0.001$]. More granular analysis showed effects for three features: SHAPE [$t(179) = 5.11), p < 0.001$], ORIENTATION [$t(71) = 5.95), p < 0.001$], and COLOR [$t(35) = 8.79), p < 0.001$]. The model did not show a significant match effect for either SIZE [$t(95) = 0.682), p = 0.497$] or VOLUME [$t(71) = 0.17), p = 0.863$]. The results of other models patterned similarly, with significant match effects on SHAPE and COLOR for all three CLIP models, and effects on ORIENTATION for B/32 and H/14 (all $p < 0.02$). No other effects approached significance ($p > 0.2$).

3.3 Discussion

This diagnostic analysis allows us to identify, for each feature, the mechanism by which the model achieves sensitivity to sensorimotor features, or where this mechanism fails.

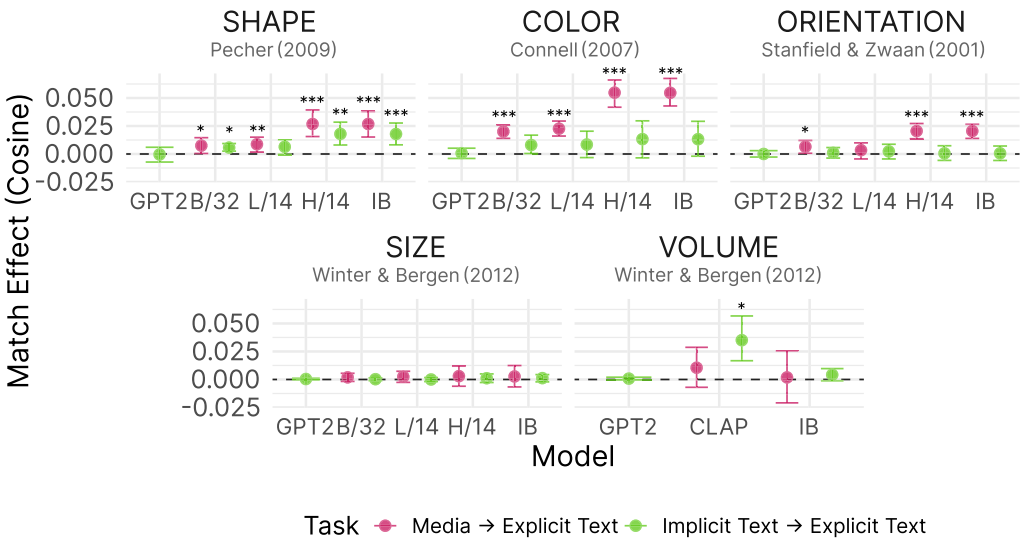

Figure 3
Match effects (cosine similarity for matching – mismatching pairs) for two diagnostic tasks: comparing media and sentences to explicit text feature labels. In the **[Implicit Text → Explicit Text]** task, ImageBind showed a significant effect for SHAPE. B/32 and H/14 also show a SHAPE effect, and CLAP showed an effect on the VOLUME dataset. GPT-2 showed no effect of match on any of the datasets, and no other effects reached significance. In the **[Media → Explicit Text]** task (pink), ImageBind showed a match effect for three features (SHAPE, COLOR, and ORIENTATION). All MLLMs showed the SHAPE and COLOR effects, but only B/32 and H/14 show the ORIENTATION EFFECT. None of the models showed effects of SIZE and VOLUME.

In the simplest case, SHAPE, the results of Experiment 1 suggest that ImageBind is sensitive to matches between sensorimotor features implicit in sentences and explicit in media. The present analysis shows that the model can also match features of explicit text descriptions to implicit text descriptions and to images.

In the case of ORIENTATION, the results suggest that ImageBind is capable of discriminating orientation in images, but does not associate sentences that *imply* a given orientation with explicit descriptions of that orientation. This implies that extracting the feature from the implicit description is a bottleneck for the multimodal model overall, and may account for the model's insensitivity to orientation in Experiment 1 (Kamath, Hessel, and Chang 2023).

For SIZE and VOLUME, the model shows no sensitivity in either matching explicit to implicit text descriptions, or matching explicit features to media. This suggests that the model is not sensitive to these features at all. It could be that models never learn to be sensitive to these features because they are not normally relevant for image-text matching tasks that the model is trained on.

The case of COLOR is harder to interpret. ImageBind showed a match effect for color between implicit sentences and images in Experiment 1, and between explicit descriptions and images in Experiment 2, but not between explicit and implicit text descriptions. This could suggest that the model's representations of text descriptions of color can be used to identify images that match the color, but not sentences that imply the object's color. One possible explanation for this is that the explicit color descriptions (e.g., "The steak is [brown/red]."") are poorly designed. However, the model is able to use these descriptions to identify images matching these colors. It is possible that

other features of the implicit sentences contribute to the model's representations in ways that align with visual representations but not explicit textual descriptions. Importantly, however, the match effect on the text version of the color task was reasonably large (0.13). Given the small number of items in this dataset, and the fact that we see effects that approach significance across several models, it seems likely that MLLM text encoders show a real effect of this manipulation, but one that this study was underpowered to detect.

Finally, GPT-2 showed no sensitivity to the match/mismatch distinction on any of the datasets. This is consistent with the hypothesis that multimodal feedback (produced through contrasting learning on text-image pairs) is crucial to developing sensitivity to implicit descriptions of sensorimotor features. However, although GPT-2 has a similar architecture and number of parameters to ViT H/14 (on which the ImageBind text and image encoders are based), it is trained on a next-word prediction objective which might not provide incentives for generating representations that encode relevant features. To test this possibility we conducted a follow-up study with BERT (Devlin et al. 2019). Part of BERT's pretraining involves classifying sentences using the activations on a special [CLS] token, similar to the [EOS] token used in CLIP-based text encoders. We performed the same analysis using activations in the final layer of BERT-large on the [CLS] token, and found a similar pattern of results. The largest difference in cosine similarity was observed for the SHAPE dataset [$t(179) = 1.05, p = 0.29$] (all other ps $>$ 0.73). These results are consistent with the proposal that feedback from multimodal stimuli helps to shape the representations of MLLM text encoders so that they are more sensitive to implicit sensorimotor features. In turn this provides support for embodied theories of human cognition, which suggest that sensorimotor experience plays a crucial role in grounding our understanding of language (Barsalou 1999; Bisk et al. 2020).

4. Experiment 3

An important test of a model's cognitive plausibility is its ability to predict human behavior on relevant psycholinguistic tasks (Kuribayashi, Oseki, and Baldwin 2023). Thus, in Experiment 3, we asked whether and to what extent the match/mismatch effects exhibited by MLLMs (and LLMs) tested in Experiments 1–2 successfully predict variance in human reaction time on the same task.

This analysis is helpful for establishing the cognitive plausibility of a particular model. Further, if a given model *explains away* the main effect of match/mismatch in human data, it provides a viable mechanism for what human comprehenders might be doing on that same task.

4.1 Methods

4.1.1 Materials. We used model predictions from Experiments 1–3 and human data from pre-existing studies:[2]

1. SHAPE: 8,410 trials from 348 participants from Zwaan and Pecher (2012) Experiment 1.

2 We were not able to extract item-level condition information for each trial for the COLOR dataset in Experiment 3 of Zwaan and Pecher (2012) and so this dataset was excluded from the human baseline analysis.

2. ORIENTATION: 7,480 trials from 336 participants from Zwaan and Pecher (2012) Experiment 2.
3. SIZE: 349 trials from 22 participants from Winter and Bergen (2012) Experiment 1.
4. VOLUME: 749 trials from 32 participants from Winter and Bergen (2012) Experiment 2.

We implemented participant-level exclusions using the same criteria as the original studies. We performed trial-level filtering in a consistent way across all studies. We removed trials where reaction times were < 300ms (indicating guessing) or > 3000ms (indicating inattention).

4.1.2 Analysis. In order to test whether (M)LLM representations could account for the effects of implicit feature match observed in human experiments, we reanalyzed human experimental data using model responses as a control predictor.

We constructed linear mixed effects models predicting human reaction times to each item, with random intercepts by participant id and quadruplet id. We constructed three distinct models for comparison. The first, with just match condition as a predictor, functions as a reproduction of the original experiment, to test whether the match effect is detectable using our statistical analysis. The second model predicts reaction time on the basis of the cosine similarity between the MLLM representations of the relevant stimuli. The model measures the extent to which MLLMs are predictive of human behavior. Finally, we constructed a model with both match condition and cosine similarity as predictors. We used hierarchical model comparison to test whether match condition has a residual effect on reaction time, over and above the variance predicted by the model.

4.2 Results

We first established that the main effect of match/mismatch observed in human data could be reproduced when data were analyzed using a linear mixed effects model. A full model including a fixed effect of Match (and random intercepts for participant and quadruplet id) explained significantly more variance than a model omitting only that variable [$\chi^2 = 63.93, p < .001$]; as expected, reaction time was slower in the mismatch condition [$\beta = 47.06, SE = 5.88, p < .001$]. The chi-square values (χ^2) and associated p-values indicate the significance of the improvement in model fit when adding the Match variable to the base model. The beta coefficients (β) represent the estimated effect of each predictor on the dependent variable. For example, the positive beta coefficient for Match ($\beta = 47.06$) indicates that reaction times are estimated to be 47.06ms slower in the mismatch condition compared to the match condition, holding other predictors constant.

We then added a fixed effect of *text-based similarity*: the cosine similarity scores between GPT-2 representations on the **[Implicit Text → Explicit Text]** task. GPT-2 similarity and human reaction times were not significantly correlated [$\beta = -43.7, SE = 137.9, p = 0.751$]. Critically, Match continued to explain variance above and beyond the base model using GPT-2 similarity as a predictor [$\chi^2 = 63.98, p < .001$], indicating that distributional information could not explain away the human effect (see Figure 4).

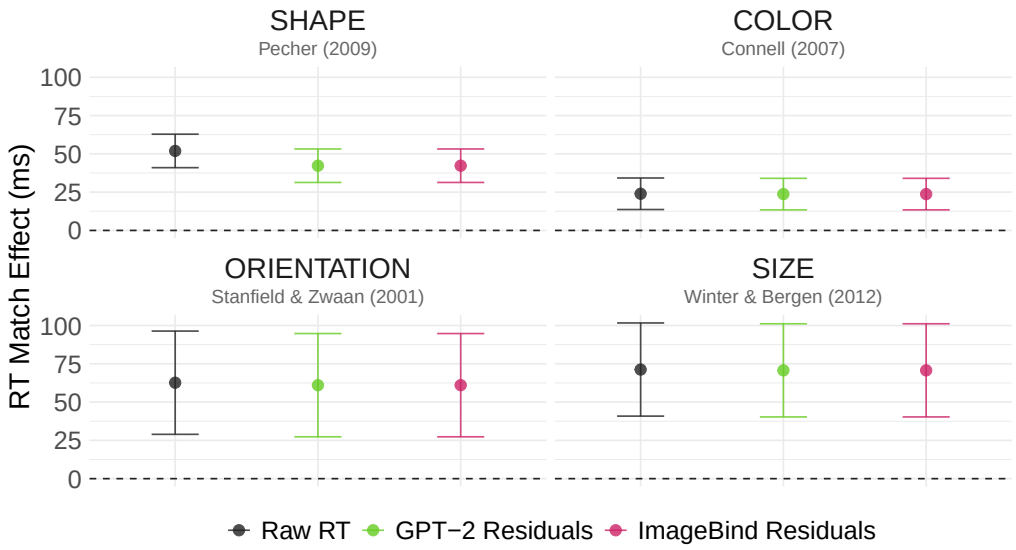

Figure 4
Match effects (reaction time for mismatch - match conditions) for human participants in raw RT data (black), and after controlling for linear model predictions on the basis of GPT-2 cosine similarity (green) and the ImageBind model's cosine similarities (pink). Match effects on human RTs were significant when controlling for both models on each of the four datasets, indicating that MLLMs cannot account for the effect of implicit feature match on human comprehenders.

We carried out an identical analysis, but using *MLLM similarity* (i.e., the ImageBind cosine similarities extracted in Experiment 1) instead of text-based similarity. The coefficient for MLLM similarity was significantly negative [$\beta = -823, SE = 92.2, p < .001$]. The negative beta coefficient for MLLM similarity ($\beta = -823$) suggests that a difference between a cosine similarity of 0 and 1 in MLLM similarity is associated with an estimated decrease of 823*ms* in reaction time, controlling for other predictors. However, as with the GPT-2 predictor, Match continued to explain variance above and beyond the base model including only MLLM similarity as a predictor [$\chi^2 = 38.41$, $p < .001$].

We carried out each analysis within each dataset (SHAPE, SIZE, etc.). ImageBind cosine distance was predictive of reaction time in the SHAPE dataset [$\beta = -1309, p < .001$], but no other dataset (all $p > 0.13$). In each case, there was a significant effect of match after controlling for cosine similarity (all $p < 0.02$). The results of the additional models patterned very similarly to ImageBind. All three of the CLIP models were predictive of SHAPE reaction time (all $p < 0.001$). ViT-B/32 and ViT-L/14 were also predictive of ORIENTATION RT, but the effects were marginal ($p = 0.032$, and $p = 0.049$, respectively). There were significant effects of match on RT after controlling for each of the other models (all $p < 0.025$).

4.3 Discussion

MLLM representations appear to be *correlated* with human representations to some extent, as indicated by the fact that measures of similarity derived from the models

were significantly predictive of human reaction time. Crucially, however, these representations were *insufficient* to account for the main effect of the experimental manipulation (i.e., match/mismatch)—even in cases where the models showed a robust effect of match/mismatch as well (e.g., as in SHAPE). These results suggest that implicit sensorimotor features play a stronger role in human semantic processing than they do in the MLLMs tested here: either because these features are automatically activated or because they are strategically engaged in a task-dependent manner.

A candidate explanation for the difference between humans and the models has to do with *architecture*: ImageBind (like the other CLIP models) maintains distinct representational spaces for distinct modalities (i.e., language, vision, and sound), which are projected into a shared embedding space through a contrastive learning process (Girdhar et al. 2023). This means that the benefits of multimodality can only be observed *after* this projection: For the most part, representations of language or vision are still unimodal. If human semantic representations are more tightly integrated, as in some proposals (Meteyard et al. 2012; Binder and Desai 2011), then this could account for the stronger effect of match/mismatch in human data. In Experiment 4, we ask whether different MLLM architectures exhibit different psychometric predictive power (Kuribayashi, Oseki, and Baldwin 2023).

5. Experiment 4

There is more than one way to be "multimodal" (Bugliarello et al. 2021). *Dual-encoder models* like CLIP (Radford et al. 2021) maintain distinct representations of each modality for the majority of the network, and only integrate these representations via a thin projection to a shared embedding space. In contrast, *fusion* models process text and modal inputs in concert, allowing the representation of one to influence the other. In *dual-stream* fusion models, like BridgeTower (Xu et al. 2023), modalities are encoded separately but influence one another through mechanisms such as cross-attention. In *single-stream* fusion models, such as ViLT (Kim, Son, and Kim 2021), modalities are encoded jointly and interact via conventional self-attention mechanisms (Wu et al. 2023).

It remains unclear exactly how MLLM architecture affects the degree of "cross-talk" between modalities and the multimodal representations each MLLM forms. It is also unknown which architecture most closely mirrors the mechanism by which humans ground language: Indeed, there is considerable debate about whether human semantic representations are fully multimodal (i.e., "single-stream"), or whether grounding occurs primarily through selective "interaction" between distinct, unimodal representations (Mahon and Caramazza 2008; Meteyard et al. 2012).

We address both questions by asking whether different architectures vary in the match/mismatch effect; we then quantify the **psychometric predictive power (PPP)** of each architecture (Kuribayashi, Oseki, and Baldwin 2023). Because the MLLMs tested vary in more than just their architecture (e.g., amount of training data), this analysis is an imperfect test of the hypothesis that variance in architecture contributes to variance in PPP. However, it is a relatively strong test of the more specific hypothesis that single-stream models will exhibit higher PPP than dual-stream models: This is because the dual-stream model tested (CLIP) was trained on orders of magnitude more data ($\sim$ 400M image/caption pairs; Radford et al. 2021) than ViLT, the single-stream model ($\sim$ 4M image/caption pairs; Kim, Son, and Kim 2021).

5.1 Methods

5.1.1 Multimodal Models. We compared three models:

CLIP ViT-B/32 (Radford et al. 2021). A dual-encoder model composed of a 63M-parameter 12-layer 512-wide text-encoder with 8 attention heads, and a Vision Transformer (Dosovitskiy et al. 2021) with a 32px patch size trained on 224×224px images. The model is trained by contrastive learning on a dataset of 400M image-text pairs.

BridgeTower (Xu et al. 2023). A dual-stream fusion model composed of a 12-layer RoBERTa-based text-encoder and a 12-layer CLIP-ViT-B/16 vision transformer with 6 cross-modal layers that perform *co-attention* between text and image representations. The model is trained on 4M image-text pairs on both Masked Language Modeling (MLM) and Image-Text Matching (ITM) objectives.

ViLT (Kim, Son, and Kim 2021). A single-stream fusion model composed of a single BERT-based 12-layer transformer that implements self-attention over concatenated word and image embeddings. The model uses a 32px patch size for processing images and is trained on 4 million image-text pairs using MLM and ITM.

We run ViLT and BridgeTower through the Python transformers package (version 4.35.2; The HuggingFace Team and Contributors 2023). Because more integrated models do not produce independent representations of each modality, we cannot use cosine similarity to compare model representations as we did in Experiments 1–3. Instead, we use the logits that each model produces for each sentence-image pair, representing the association strength between the sentence and the image. To quantify the similarity between linguistic and visual representations within the model, we apply a softmax function to these logits, converting them into probabilities. This process effectively normalizes the logits across all sentence-image pairs, allowing us to interpret them as the model's confidence in associating a specific image with a given sentence:

$$p_{ij} = \frac{\exp(\text{logit}_{ij})}{\sum_{k=1}^{2} \exp(\text{logit}_{ik})} \tag{1}$$

where logit_{ij} is the logit score assigned by the model for the pairing of sentence i with image j, and p_{ij} is the softmax probability of this pairing.

5.1.2 Procedure. We conducted two analyses. First, in a series of pairwise comparisons, we asked whether MLLMs differed in the size of their match/mismatch effect: specifically, for each pair of models, we constructed a linear regression with an interaction between Model Type (e.g., ViLT vs. CLIP) and Match/Mismatch, predicting Similarity.

Second, for each MLLM under consideration, we constructed a linear mixed effects model predicting human reaction time, with MLLM similarity as a fixed effect, and random intercepts for subjects and items. We operationalized the PPP of each model as Akaike Information Criterion (AIC). The AIC is a metric used for comparing statistical models, where a lower AIC indicates a better balance between model fit and simplicity, suggesting the model is a more parsimonious explanation of the data.

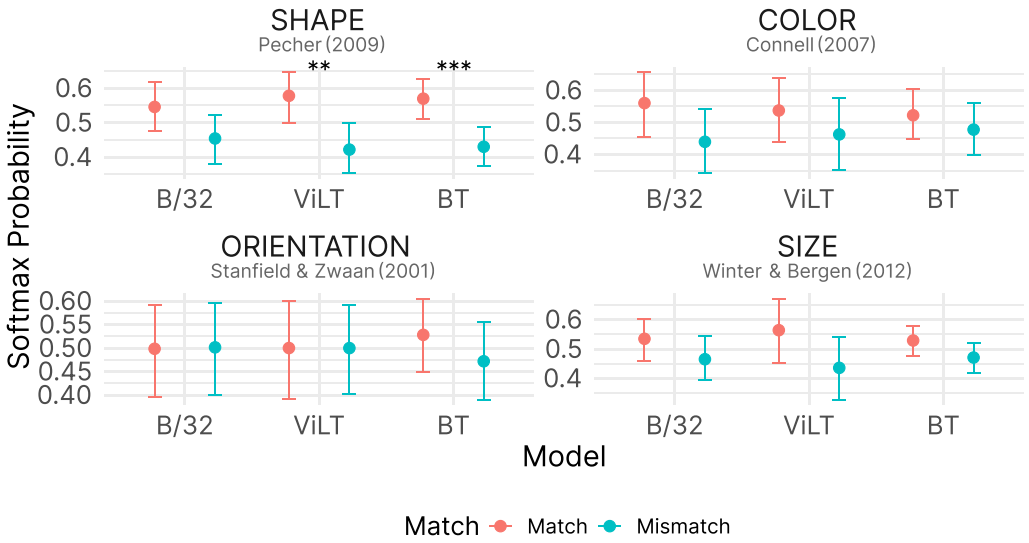

Figure 5
Main effects of match on the softmax probability assigned to text labels for each image from CLIP B/32, BridgeTower, and ViLT models. On the SHAPE dataset, there was a main effect on BridgeTower [$t(510) = 3.50, p < 0.001$] and ViLT [$t(510) = 2.94, p = 0.004$] probabilities. There was also an effect for CLIP ViT-B/32 that approached significance [$t(510) = 1.80, p = 0.072$]. No other dataset-level effects showed significant effects.

5.2 Results

There was a main effect of match overall in each of the three models: CLIP ViT-B/32 [$t(510) = 2.28, p = 0.023$], BridgeTower [$t(510) = 3.88, p < 0.001$], and ViLT [$t(510) = 3.22, p = 0.001$]. However, when examining each dataset individually, there was only a main effect of condition on similarity in the SHAPE dataset for BridgeTower [$t(510) = 3.50, p < 0.001$] and ViLT [$t(510) = 2.94, p = 0.004$]. There was also an effect for CLIP ViT-B/32 that approached significance [$t(510) = 1.80, p = 0.072$]. No other dataset-level effects were significant (all $p > 0.09$, see Figure 5). There was no difference in the size of the match/mismatch effect across MLLMs, as confirmed by the lack of a significant interaction between Model Type and Match (all $p > .1$).

For the SHAPE dataset, the models varied considerably in how successfully they predicted human reaction time, as operationalized by AIC (see Table 1). ViLT exhibited the lowest (best) AIC, followed by CLIP B/32 and then BridgeTower. These differences were substantial (> 30 between each pair of models).[3] That is, variation in model architecture was related to variation in PPP. Variation in the ORIENTATION and SIZE datasets was much smaller, perhaps reflecting the fact that *none* of the models tested were sensitive to those implicit features (see Experiment 1).

3 Differences in AIC are generally not interpreted using traditional significance testing. However, differences larger than 4 are generally interpreted as reflective meaningful differences in model fit (Burnham and Anderson 2004).

Table 1
Fit of linear models predicting human reaction times on the basis of measures from different MLLMs. Estimate shows the expected change in RT for a softmax probability of 1 vs 0. *p*-values are for the effect of modal probability on RT. ΔAIC is the difference in AIC from each model to the best-performing model in each dataset. **Bold** *p*-values indicate significant effects while underlined ΔAIC values are the best for each dataset. On the SHAPE dataset, **ViLT** produces a much better fit than ViT **B/32** ($\Delta AIC = 74$)—a dual encoder model)—and **B**ridge**t**ower ($\Delta AIC = 143$): a dual-stream fusion model. There are no meaningful AIC differences (all < 3) for the other two datasets.

Dataset	Model	Estimate	*p*-value	ΔAIC
Orientation	BT	−5.82	0.750	2.12
Orientation	B/32	−25.91	0.121	<u>0.00</u>
Orientation	ViLT	−15.80	0.308	1.51
Shape	BT	−80.01	**< 0.001**	143.07
Shape	B/32	−129.85	**< 0.001**	74.20
Shape	ViLT	−169.34	**< 0.001**	<u>0.00</u>
Size	BT	−38.94	0.514	1.70
Size	B/32	76.94	0.139	<u>0.00</u>
Size	ViLT	−16.81	0.646	2.89

5.3 Discussion

These results were partially consistent with the hypothesis that MLLMs with more integration between their modalities would exhibit higher PPP: Specifically, ViLT, a single-stream model, was more predictive of human behavior than CLIP, a dual-stream model. Notably, this is despite the fact that CLIP was trained on orders of magnitude more training data than ViLT.

One interpretation of this result could be that ViLT's architecture is more reflective of human semantic representations, and thus it needs fewer training exemplars to predict human behavior (i.e., it exhibits better *data efficiency*). On the other hand, this is a single empirical result; this explanation also does not account for why BridgeTower performed worse than CLIP. The question of model architecture and how it relates to PPP and data efficiency is explored in more detail in the General Discussion.

6. General Discussion

We evaluated MLLMs on five embodied simulation experiments in order to ask whether MLLMs exhibit the same kind of integration between modalities that is thought to constitute grounding in humans. Our results present mixed evidence on this question. In Experiment 1, ImageBind representations were sensitive to whether the implied shape or color of an object in a sentence was matched in an image. Importantly, these visual features were not explicitly mentioned in the sentences. The model's sensitivity to implied SHAPE and COLOR suggests that it is activating representations of an object that are specific to (and implicit in) the event described in the sentence (e.g., the implication that an eagle in the sky would have its wings outstretched, while an eagle in its nest would have its wings folded). In humans, an analogous effect is taken as evidence of embodied simulation (Stanfield and Zwaan 2001; Bergen 2015). The findings

here suggest that such an effect can be produced via exposure to large-scale statistical associations between patterns in images and patterns in text.

In contrast, we found no match effect for ORIENTATION, SIZE, or VOLUME. Our diagnostic analysis in Experiment 2 suggests different reasons for these failures. ImageBind was sensitive to whether explicit labels for object orientations matched images, suggesting that the model is capable of accurately representing this feature. However, there was no match effect between these explicit orientation labels and sentences that indirectly imply orientation. This suggests that the text-encoder is the bottleneck for MLLM sensitivity to orientation overall. Kamath, Hessel, and Chang (2023) find a more general version of this effect by testing whether input captions are encoded with enough specificity to be recovered using a text-only probe. They find that a small proportion of inputs are recoverable, especially for more compositional captions, suggesting that MLLMs would be incapable of discriminating matching images even if their image representations were sufficiently granular. They find that some text-encoders perform significantly better than others. Future work could investigate whether this is also true for the sentence-picture verification experiments used here.

For the SIZE and VOLUME stimuli, we found that ImageBind was not able to discriminate matches between either explicit and implicit text labels, or explicit labels and media (images and sounds). This suggests that the model may not be capable of representing these features *in general*. One possible reason for this insensitivity could be the MLLMs' training data and objectives: MLLMs may not have had sufficient exposure to scenarios where these features are crucial for understanding the context or the content of the image-text pairs (Lin et al. 2014). This lack of emphasis in the training data could lead to underdeveloped representation of these features in the model's architecture. Alternatively, the nature of the features themselves could present an obstacle: The perception of size and volume is inherently relational and can vary greatly depending on the context. Capturing and encoding such context-dependent features might require a more sophisticated approach to multimodal integration than what current MLLMs employ (Liu, Emerson, and Collier 2023). Future research could focus on enhancing the training paradigms and exploring architectures that can better handle compositional and context-dependent features.

Even in the case of SHAPE and COLOR, ImageBind was not sufficiently sensitive to implicit feature match to explain its effects on human comprehenders (see Experiment 3). While the model's behavior was correlated with human behavior, it did not eliminate the main effect of the experimental manipulation. This mirrors other recent work (Jones et al. 2022; Trott and Bergen 2023; Trott et al. 2023) suggesting that even LLMs that are *sensitive* to certain semantic features do not display equivalent sensitivity as human comprehenders. One possible explanation for this discrepancy is that sensorimotor information plays a more prominent role in human semantic representations, that is, human representations of meaning exhibit more multimodal integration than dual-stream models like ImageBind or CLIP. More generally, humans have much richer sources of interaction with the world beyond labeled image/caption pairs, and this may be reflected in their representations of meaning.

This explanation is also consistent with the results of Experiment 4, which reveal the effects of MLLM architecture on psychometric predictive power. Specifically, ViLT's single-stream approach was best at predicting human behavior. Given that ViLT was also trained on many fewer exemplars than dual-stream models like CLIP, this suggests that the benefits of architectural integration—at least when it comes to predicting human behavior—might compensate for a relative dearth of training data. It is still worth noting, however, that even ViLT failed to fully account for the effect of match/mismatch

in humans. Future work could investigate whether more training data would close this gap in ViLT; if it does *not*, it would suggest that the rich, interactional nature of grounded human experience may play an important role in their semantic representations.

6.1 Are MLLM Representations of Color and Shape Grounded?

The notion of "grounding" is highly polysemous (Bisk et al. 2020; Mollo and Millière 2023). One interpretation of grounding is sensorimotor grounding, which Mollo and Millière (2023) is defined as a *link* (or integration) between a conceptual representation and sensorimotor representations. This is the interpretation primarily explored in this article, and the one often used in past work on multimodal grounding of language models (Bruni, Tran, and Baroni 2014; Chrupała, Kádár, and Alishahi 2015; Kiros, Chan, and Hinton 2018; De Sa and Ballard 1998; Kádár, Chrupała, and Alishahi 2017; Peng and Harwath 2022; Harwath and Glass 2015). Do MLLMs exhibit sensorimotor grounding? In one sense, the answer is clearly yes: MLLMs have the *opportunity* to ground concepts using the specific modalities they are trained on (e.g., vision). In another sense, however, "grounding" could be taken to mean actually *deriving* and *deploying* specific semantic features that could in principle be gleaned from those sensorimotor modalities, such as the implied shape of an object. Under this narrower interpretation of grounding, the MLLMs tested were grounded with respect to shape and color, but not to size, orientation, or volume.

Further, as Mollo and Millière (2023) note, "grounding" could be interpreted in other ways as well. It might refer to a system's ability to relate a given concept to other concepts, as in a semantic network, namely, *relational grounding*; or it might involve a system's ability to "anchor a representation" in the world itself (Mollo and Millière 2023), that is, *referential grounding*. Mollo and Millière (2023) argue that referential grounding may be a particularly important source of linguistic meaning, and that this depends on understanding the causal history of how specific labels or expressions are used to refer to specific meanings—that is, how language is anchored to the world.

While the current work cannot address the more general question of what constitutes "grounding," it does offer empirical evidence that can inform theories of whether MLLMs are, in fact, grounded. Namely, by probing MLLMs with experiments designed to test grounding in humans, we can identify whether and how specifically sensorimotor representations are activated (or fail to be activated) when MLLMs are exposed to linguistic descriptions of events.

6.2 MLLMs as Explicit Computational Models of Grounding in Humans

As well as evaluating grounding in MLLMs per se, we were interested in investigating MLLMs as explicit computational models of grounding mechanisms in humans. Explicit computational models provide a helpful tool for adjudicating between theories of grounding in humans. Theories must be specified in great detail in order to be implemented in code. Once implemented, these models allow us to test whether specific mechanisms are capable of producing the behavior that they have been proposed to explain.

MLLMs in general represent models of a certain class of grounding theories. They learn passively from exposure to mixed-modality inputs to associate distributional patterns from one modality with another. Importantly, they lack the ability to interact with the world in order to seek out information or test theories about how modalities relate—a capacity that has been implicated in grounding for humans (Varela, Thompson, and

Rosch 2017). In this sense, they have been described as "learning language from the television," hardly better, perhaps, than from the radio (Bisk et al. 2020). Nevertheless, we find some evidence that MLLMs develop sensitivity to implicit sensorimotor features (SHAPE and COLOR). These results provide some limited support to the theory that MLLMs' associative mechanisms are capable of generating schematic representations of textual descriptions with sufficient granularity that implicit sensorimotor features of described events can be related to real modal inputs.

In contrast, however, ImageBind did not show sensitivity to three other features (ORIENTATION, SIZE, and VOLUME). We therefore fail to find evidence that the mechanism tested here is capable of generating sensitivity to implicit sensorimotor features in general. In addition, even in the cases of SHAPE and COLOR, the model explained a very small proportion of the effect of Match on human reaction times (see Figure 4). This result is consistent with theories that human comprehenders make use of other resources or mechanisms that go beyond contrastive learning from independent encodings of images and text.

At a more granular level, specific MLLM architectures provide loose analogies for proposed mechanistic models of embodied simulation in humans (Meteyard et al. 2012; Binder and Desai 2011). Dual Encoder models can be roughly aligned with Secondary Embodiment theories (Mahon and Caramazza 2008; Patterson, Nestor, and Rogers 2007), where modality-specific inputs are processed independently and used to inform a higher-level non-modality-specific representation. Dual-Stream fusion models loosely operationalize weak embodiment theories (Barsalou 1999), where processing input from one modality is partly dependent on one's representations of another. Single-stream fusion models have the potential to implement the strongest kind of multimodal interaction (Strong embodiment; Gallese and Lakoff 2005), where linguistic inputs are processed using exactly the same neural resources as sensorimotor inputs. However, it is an empirical question whether models actually learn to do this.

We tested the plausibility of MLLMs as operationalizations of the mechanism of embodied simulation in humans by asking what proportion of the variance in human behavioral data they can account for. The fact that the single-stream fusion model (ViLT) provided significantly better predictions of human behavior than CLIP provides tentative support for stronger embodiment theories that hypothesize close integration between conceptual and perceptual representations. However, there were a variety of differences between models in our analysis which limit the strength of the inferences that can be drawn. Ideally, future work will hold training data and parameter counts constant while varying architectural features systematically. Such careful work, with close reference to the theoretical embodiment literature, could help to shed light on what kinds of mechanisms are necessary to realize embodied grounding in humans.

7. Conclusion

MLLMs have been proposed as solutions to the so-called symbol grounding problem (Harnad 1990). However, it is unclear whether MLLM representations are *grounded* in similar ways and to similar degrees as many believe human representations are (Bergen 2015). A large body of experimental evidence has emerged suggesting that humans understand language in part by activating relevant sensorimotor features, for example, the implied shape or orientation of an object (Zwaan, Stanfield, and Yaxley 2002; Pecher et al. 2009). By adapting techniques originally designed to probe grounding in humans, we found that MLLMs are sensitive to some implicit features and not others—and that MLLMs fail to fully account for the effect of grounding in humans.

This work has implications for debates about whether MLLMs are grounded (Mollo and Millière 2023); it also suggests a potentially fruitful direction for cross-pollination between psycholinguistics and computational linguistics.

Ethics Statement

This study did not involve new data collection from human participants, focusing instead on the analysis of pre-existing datasets with Multilingual Large Language Models (MLLMs). The study also did not involve training models, and the computational demands of our research, including energy usage and potential environmental impacts, were negligible.

Acknowledgments

We would like to thank Diane Pecher and Rolf Zwaan for making their stimuli and data for Zwaan and Pecher (2012) available via OSF and assistance in implementing our analysis; and Bodo Winter for providing stimuli and human subjects data for Winter and Bergen (2012).

References

Barsalou, Lawrence W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660. https://doi.org/10.1017/S0140525X99002149, PubMed: 11301525

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

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

Bergen, Benjamin K. 2012. *Louder than Words: The New Science of How the Mind Makes Meaning*. Basic Books.

Binder, Jeffrey R. and Rutvik H. Desai. 2011. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11):527–536. https://doi.org/10.1016/j.tics.2011.10.001, PubMed: 22001867

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47. https://doi.org/10.1613/jair.4135

Bugliarello, Emanuele, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994. https://doi.org/10.1162/tacl_a_00408

Burnham, Kenneth P. and David R. Anderson. 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304. https://doi.org/10.1007/b97636

Chang, Tyler and Benjamin Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350. https://doi.org/10.1162/coli_a_00492

Chrupała, Grzegorz, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 112–118. https://doi.org/10.3115/v1/P15-2019

Connell, Louise. 2007. Representing object colour in language comprehension. *Cognition*, 102(3):476–485. https://doi.org/10.1016/j.cognition.2006.02.009, PubMed: 16616075

De Sa, Virginia R. and Dana H. Ballard. 1998. Category learning through multimodality sensing. *Neural Computation*, 10(5):1097–1117. https://doi.org/10

.1162/089976698300017368, PubMed: 9654768

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600. https://doi.org/10.1016/j.tics.2023.04.008, PubMed: 37173156

Dosovitskiy, Alexey, 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. *arXiv preprint arXiv:2010.11929*.

Driess, Danny, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Funder, David C. and Daniel J. Ozer. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156–168. https://doi.org/10.1177/2515245919847202

Gallese, Vittorio and George Lakoff. 2005. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22(3–4):455–479. https://doi.org/10.1080/02643290442000310, PubMed: 21038261

Gelman, Andrew and Hal Stern. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331. https://doi.org/10.1198/000313006X152649

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457

Gong, Yuan, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. *arXiv preprint arXiv:2104.01778*. https://doi.org/10.21437/Interspeech.2021-698

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

Harwath, David and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. https://doi.org/10.1109/ASRU.2015.7404800

Hauk, Olaf, Ingrid Johnsrude, and Friedemann Pulvermüller. 2004. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307. https://doi.org/10.1016/S0896-6273(03)00838-9, PubMed: 14741110

Hu, Jennifer, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*. https://doi.org/10.18653/v1/2023.acl-long.230

Huang, Shaohan, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Ilharco, Gabriel, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP (0.1). Zenodo. https://doi.org/10.5281/zenodo.5143773

Jones, Cameron R., Tyler A. Chang, Seana Coulson, James A. Michaelov, Sean Trott, and Benjamin Bergen. 2022. Distributional semantics still can't account for affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44:482–489.

Kádár, Akos, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*,

43(4):761–780. https://doi.org/10.1162/COLI_a_00300

Kalra, Agastya, Guy Stoppi, Bradley Brown, Rishav Agarwal, and Achuta Kadambi. 2021. Towards rotation invariance in object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3530–3540. https://doi.org/10.1109/ICCV48922.2021.00351

Kamath, Amita, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944. https://doi.org/10.18653/v1/2023.emnlp-main.301

Kim, Wonjae, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594, PMLR.

Kiros, Jamie, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933. https://doi.org/10.18653/v1/P18-1085

Kuribayashi, Tatsuki, Yohei Oseki, and Timothy Baldwin. 2023. Psychometric predictive power of large language models. *arXiv preprint arXiv:2311.07484*. https://doi.org/10.18653/v1/2024.findings-naacl.129

Lin, Tsung Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, Fangyu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651. https://doi.org/10.1162/tacl_a_00566

Mahon, Bradford Z. and Alfonso Caramazza. 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, PubMed: 18448316

Meteyard, Lotte, Sara Rodriguez Cuadrado, Bahador Bahrami, and Gabriella Vigliocco. 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, PubMed: 21163473

ML Foundations. 2023. OpenCLIP. https://github.com/mlfoundations/open_clip. Python package version 2.23.0.

Mollo, Dimitri Coelho and Raphaël Millière. 2023. The vector grounding problem. *arXiv preprint arXiv:2304.01481*.

Ostarek, Markus and Roberto Bottini. 2021. Towards strong inference in research on embodiment–possibilities and limitations of causal paradigms. *Journal of Cognition*, 4(1):5. https://doi.org/10.5334/joc.139, PubMed: 33506171

Patterson, Karalyn, Peter J. Nestor, and Timothy T. Rogers. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12):976–987. https://doi.org/10.1038/nrn2277, PubMed: 18026167

Pecher, Diane, Saskia van Dantzig, Rolf A. Zwaan, and René Zeelenberg. 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, PubMed: 19142833

Peng, Puyuan and David Harwath. 2022. Word discovery in visually grounded, self-supervised speech models. *arXiv preprint arXiv:2203.15081*. https://doi.org/10.21437/Interspeech.2022-10652

Pulvermüller, Friedemann. 2013. How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9):458–470. https://doi.org/10.1016/j.tics.2013.06.004, PubMed: 23932069

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, and Mitchell Wortsman. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Shanahan, Murray. 2023. Talking about large language models. *Communications of the ACM*, 67(2):68–79. https://doi.org/10.1145/3624724

Stanfield, Robert A. and Rolf A. Zwaan. 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, PubMed: 11340925

The HuggingFace Team and Contributors. 2023. Transformers: State-of-the-art machine learning for JAX, PyTorch and TensorFlow. https://github.com/huggingface/transformers. Python package version 4.35.2.

Tong, Shengbang, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Trott, Sean and Benjamin Bergen. 2023. Word meaning is both categorical and continuous. *Psychological Review*, 30(5):1239–1261. https://doi.org/10.1037/rev0000420, PubMed: 36892900

Trott, Sean, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309. https://doi.org/10.1111/cogs.13309, PubMed: 37401923

Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 2017. *The Embodied Mind, revised edition: Cognitive Science and Human Experience*. MIT Press. https://doi.org/10.7551/mitpress/9780262529365.001.0001

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 30:5998–6008.

Winter, Bodo and Benjamin Bergen. 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

Wu, Jiayang, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data*, pages 2247-2256. https://doi.org/10.1109/BigData59044.2023.10386743

Xu, Xiao, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. BridgeTower: Building bridges between encoders in vision-language representation learning. https://doi.org/10.1609/aaai.v37i9.26263

Zwaan, Rolf A. and Diane Pecher. 2012. Revisiting mental simulation in language comprehension: Six replication attempts. *PloS ONE*, 7(12):e51382. https://doi.org/10.1371/journal.pone.0051382, PubMed: 23300547

Zwaan, Rolf A., Robert A. Stanfield, and Richard H. Yaxley. 2002. Language comprehenders mentally represent the shapes of objects. *Psychological Science*, 13(2):168–171. https://doi.org/10.1111/1467-9280.00430, PubMed: 11934002