

Human-like object concept representations emerge naturally in multimodal large language models

Changde Du^{1,2}, Kaicheng Fu^{1,2}, Bincheng Wen³, Yi Sun^{1,2}, Jie Peng^{1,2}, Wei Wei¹, Ying Gao¹, Shengpei Wang¹, Chuncheng Zhang¹, Jinpeng Li⁴, Shuang Qiu¹, Le Chang³, and Huiguang He^{1,2,5,*}

¹State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

³Institute of Neuroscience, State Key Laboratory of Brain Cognition and Brain-Inspired Intelligence Technology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

⁴School of Automation Science and Engineering, South China University of Technology, Guangzhou, China

⁵Zhongguancun Academy, Beijing, China

*corresponding author: Huiguang He (huiguang.he@ia.ac.cn)

ABSTRACT

Understanding how humans conceptualize and categorize natural objects offers critical insights into perception and cognition. With the advent of Large Language Models (LLMs), a key question arises: can these models develop human-like object representations from linguistic and multimodal data? In this study, we combined behavioral and neuroimaging analyses to explore the relationship between object concept representations in LLMs and human cognition. We collected 4.7 million triplet judgments from LLMs and Multimodal LLMs (MLLMs) to derive low-dimensional embeddings that capture the similarity structure of 1,854 natural objects. The resulting 66-dimensional embeddings were stable, predictive, and exhibited semantic clustering similar to human mental representations. Remarkably, the dimensions underlying these embeddings were interpretable, suggesting that LLMs and MLLMs develop human-like conceptual representations of objects. Further analysis showed strong alignment between model embeddings and neural activity patterns in brain regions such as EBA, PPA, RSC, and FFA. This provides compelling evidence that the object representations in LLMs, while not identical to human ones, share fundamental similarities that reflect key aspects of human conceptual knowledge. Our findings advance the understanding of machine intelligence and inform the development of more human-like artificial cognitive systems.

Introduction

The ability to categorize and conceptualize objects forms the bedrock of human cognition, influencing everything from perception to decision-making. When confronted with diverse objects, humans can often differentiate their categories and concepts by making structured comparisons between them. This process is an essential part of human cognition in tasks ranging from everyday communication to problem-solving. In this cognitive process, our mental representations serve as a substrate, aiding in the recognition of objects^{1,2}, formation of categories^{3–5}, organization of conceptual knowledge^{6,7}, and the prediction of behaviors based on experiences. Therefore, understanding the structure of these representations is a fundamental pursuit in cognitive neuroscience and psychology^{8–11}, underpinning significant research advancements in the field. For instance, various studies have identified potential dimensions that organize these representations, such as animals versus non-animals^{12–15}, natural versus human-made^{16,17}, and large versus small^{18,19}.

The cognitive plausibility of deep learning systems has sparked significant debate^{20,21}, with recent works often focusing on diverse neural networks pretrained on limited datasets for specific computer vision tasks like image classification^{22–27}. While these endeavors have led to notable advancements^{27–30}, including some evidence of human-like representations emerging from self-supervised learning^{31–34}, a critical question remains: to what extent can complex, task-general psychological representations emerge without explicit task-specific training, and how do these compare to human cognitive processes across a broad range of tasks and domains? LLMs, such as OpenAI’s ChatGPT and Google’s Gemini, have emerged as potent tools in text and image understanding, generation, and reasoning. These models exhibit impressive capabilities in tasks like object identification, information categorization, concept communication, and inference. Unlike task-specific small-scale neural network models, LLMs utilize generic neural network architectures with billions of parameters, trained through next token

prediction on massive text corpora (and images for MLLMs) comprising trillions of tokens. Despite ongoing debates about their capacities³⁵⁻³⁷, one potential strength lies in their adeptness at problem-solving with minimal task-specific training, often requiring only straightforward task instructions without parameter updates. These features raised the question of whether LLMs have developed human-like conceptual representations about natural objects.

In this study, we used a data-driven approach to explore the core dimensions of mental representations in LLM (ChatGPT-3.5) and MLLM (Gemini Pro Vision 1.0). Inspired by previous work conducted on human similarity judgments using visual object images, we adopted a similar methodology to both the LLM and MLLM. Unlike presenting visual stimuli to human participants and MLLMs, we presented corresponding textual descriptions of visual images to the LLMs. Harnessing the models' ability to perform a triplet odd-one-out task, a well-established paradigm in cognitive psychology^{10,16,17,38}, we collected extensive datasets comprising 4.7 million triplet similarity judgments for both the LLM and MLLM. Each dataset is rich in triple similarity judgment entries, drawn from a pool of 1,854 unique objects. This diverse collection enables the examination and capture of visual and conceptual mental representations spanning a wide array of natural objects.

Using a representation learning method previously designed for human participants^{16,39}, we identified 66 sparse, non-negative dimensions underlying LLMs' similarity judgments that lead to excellent predictions of both single-trial behavior and similarity scores between pairs of objects. We demonstrated that these dimensions are interpretable, exhibited spontaneous semantic clustering, and characterized the large-scale structure of LLMs' mental representations of natural objects. Furthermore, by comparing the identified dimensions with the core dimensions observed in human cognition, we found close alignment between model and human embeddings. Finally, we found strong correspondence between the model embeddings and neural activity patterns in category-selective brain Region of Interests (ROIs, e.g., EBA, PPA, RSC, FFA), underscoring the generalization of these learned mental representations and offering a compelling evidence that the object representations in LLMs, while not identical to those in the human, share fundamental commonalities that reflect key schemas of human conceptual knowledge. These results enrich the growing body of work characterizing the emergent characteristics of LLMs⁴⁰⁻⁴⁹, showcasing their potential to capture and reflect human-like conceptualizations of real-world objects.

Results

We initiated our study by selecting a diverse set of objects from the THINGS database⁵⁰, encompassing 1,854 common objects (Fig. 1a). To compare LLMs' mental representations with humans, we adopted the triplet odd-one-out task, effective for modeling human mental dimensions^{10,16,17,38,51} (Figs. 1b-d). Given the impracticality of conducting 1.06 billion triplet judgments, we approximated the similarity matrix using approximately 0.44% of the total judgments, following established methods^{16,17}. Human similarity judgments were collected from 4.7 million trials via Amazon Mechanical Turk¹⁷, and LLMs' behavioral data mirrored these trials. Fig. 1e displays examples of prompts and responses from GPT-3.5-Turbo and Gemini Pro Vision, detailing choice derivation. We utilized the Sparse Positive Similarity Embedding (SPoSE) method^{16,39} (Fig. 1f) to infer LLMs' low-dimensional representations, optimizing object weights to predict behavioral judgments. We validated the generalization of LLM embeddings on the Natural Scenes Dataset (NSD)⁵² and applied Representational Similarity Analysis (RSA)⁵³ to assess correlations with neural activity (Figs. 1g-h).

Low-dimensional embeddings identified from LLMs are stable and predictive

Given the stochastic nature of SPoSE modeling (see Methods), we conducted multiple reruns with different random initializations, yielding slightly varied embeddings. Dimensions were sorted by their total object weights, and redundant dimensions (correlation > 0.4) were pruned, retaining only one. This reduced redundancy, as most dimensions appeared consistently across runs. To evaluate retained dimensions, we gathered triplet judgments for 48 typical objects (these triplet judgments are not included in the SPoSE model's training data), comparing choice probabilities with predictions from the SPoSE embedding. Fig. 2a shows that predictive performance stabilizes as dimensions increase, saturating at 60 dimensions for LLM, MLLM, and human. We chose the top 66 dimensions for LLM and MLLM to align with the 66 core dimensions from human similarity judgments¹⁷, as dimensions beyond the 66th contribute minimally to object similarity prediction.

Figs. 2b-d illustrate strong correlations between the model-predicted and behaviorally-measured Representational Similarity Matrices (RSMs) for LLM (0.71), MLLM (0.85), and human (0.9), validating the close reflection of behavioral similarity space. This result shows that, despite the complex object pool, a low-dimensional embedding can capture a large portion of the representational structure derived from similarity judgments.

Next, we calculated reproducibility scores for each retained dimension (see Methods). In Fig. 2e, all LLM embedding dimensions scored above 0.51, with 37 dimensions exceeding 0.90. Fig. 2f shows that MLLM dimensions had reproducibility scores above 0.36, except one at 0.22, with 31 dimensions exceeding 0.80. Human dimensions in Fig. 2g showed comparable reproducibility. These findings confirm that the embeddings are stable across reruns.

We also evaluated the ability of these embeddings to predict choices in the odd-one-out task using model's own held-out behavioral choice test set. As shown in Fig. 2h, accuracies were 56.7% ($\pm 0.22\%$), 63.4% ($\pm 0.25\%$), and 64.1% ($\pm 0.18\%$)

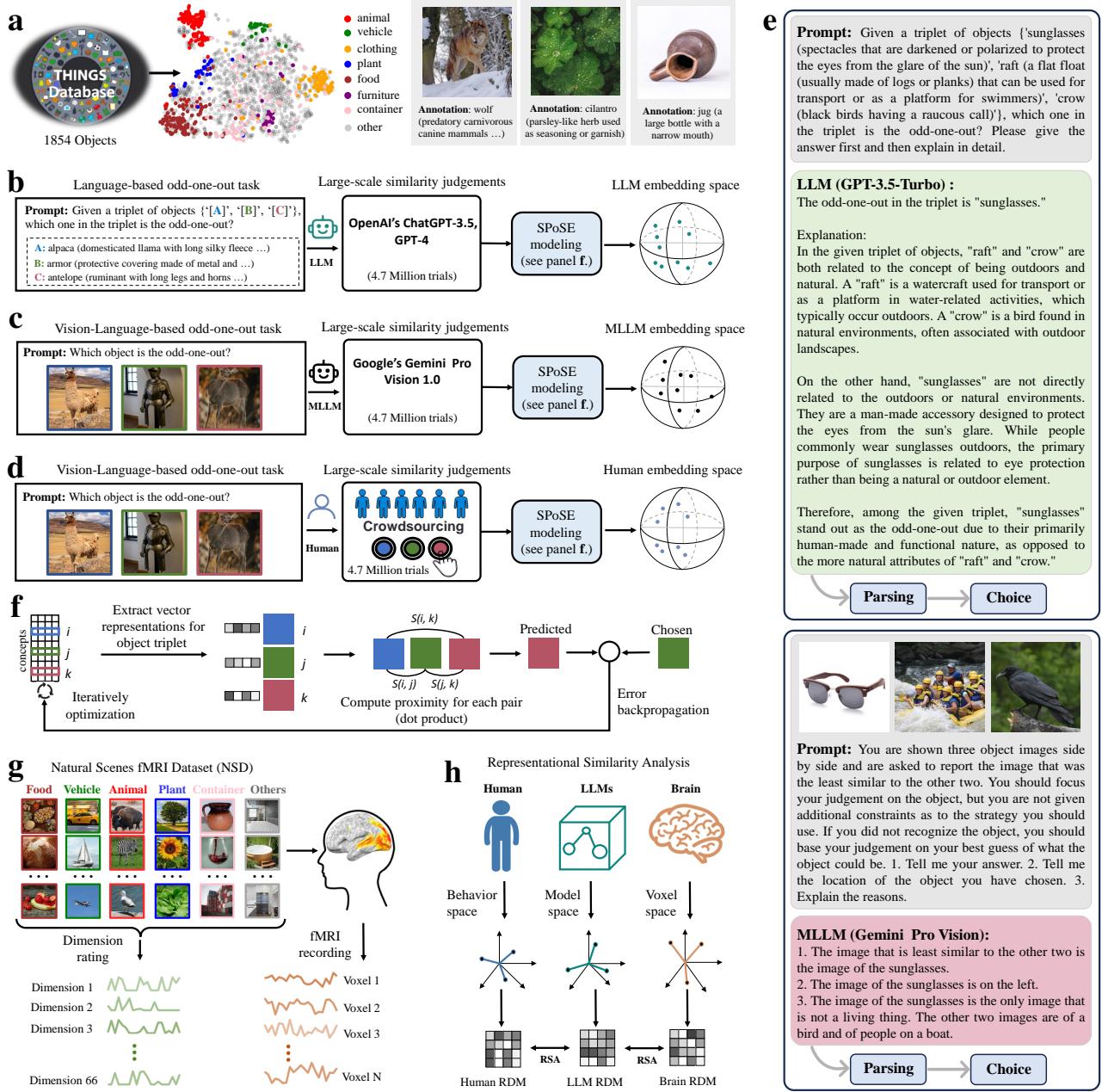


Fig. 1. Schematic diagrams of the experiment and analysis methods. **a**, THINGS database and examples of object image with their language descriptions at the bottom. **b-d**, Pipelines of mental embedding learning under the triplet odd-one-out paradigm for LLM, MLLM, and humans, respectively. Odd-one-out judgments were collected for approximately 4.7 million triplets, and modeled using the SPoSE approach to derive the corresponding low-dimensional embedding. **e**, Examples of prompts and responses for LLM and MLLM. **f**, Illustration of the SPoSE modeling approach. **g**, Illustration of the NSD dataset with dimension ratings for stimulus images. The schematic structure incorporates elements adapted from Figure 1A of Horikawa et al. (2020)⁵⁴ (<https://doi.org/10.1016/j.jisci.2020.101060>), published under a CC BY 4.0 license. **h**, Overview of the comparisons between space of LLMs, human behavior and brain activity. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.

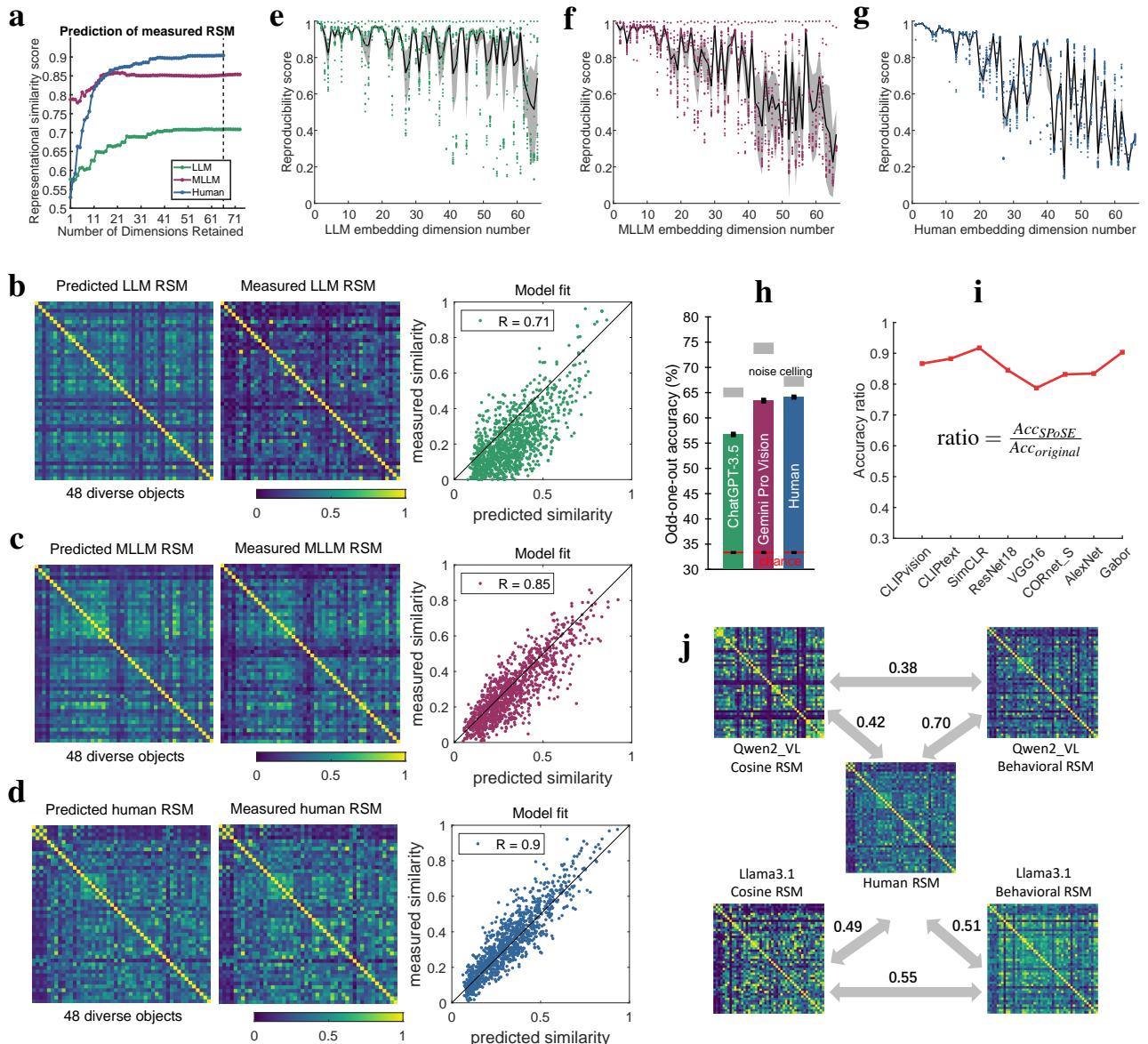


Fig. 2. Validation of the embeddings derived from similarity judgments over 4.7 million trials. **a**, Prediction performance of the measured similarity matrix with varying dimensions of the SPoSE embedding. **b-d**, RSMs for a subset consisting of 48 objects, created by estimating similarity based on the model embedding (left) and by fully sampling all possible triplets in a validation behavioral experiment (middle). Here, the similarity between two objects is operationalized as the proportion of times they are judged to be similar, across all trials. Correlation between the predicted and measured similarity on all object pairs were shown in right. **e-g**, Reproducibility of dimensions in the chosen 66-dimensional embedding. The dimensions were sorted in descending order by the sum of their weights across objects. The scores are presented as mean \pm 95% confidence intervals (CIs), and shaded areas reflect the 95% CIs ($n=20$ runs, and each dot represents the highest correlation of each selected dimension with all dimensions of a single run). **h**, Odd-one-out prediction performance on the model's own held-out behavioral choice test set. Results and chance-levels are presented as mean \pm 95% CIs, and the error bars reflect 95% CIs ($n=1000$ bootstraps). The noise ceilings were estimated from the additional behavioral datasets for each model separately, and were presented as mean \pm 95% CIs (shaded bands). **i**, How closely SPoSE embeddings mimic model's original features in odd-one-out predictions. The vertical axis represents the ratio of the SPoSE embedding accuracy to the original feature accuracy on the held-out test set constructed using cosine distances. **j**, How correlated are the model probing methods based on behavioral choices with those based on cosine distance. The numbers on the gray arrows represent the Pearson correlation between different RSMs (of the 48 objects).

for LLM, MLLM, and human, respectively (chance = 33.3%, 95% CI = [33.19%, 33.47%], 1,000 permutation tests). Noise ceilings for fitting individual-trial behavior were 65.1% ($\pm 0.96\%$), 73.8% ($\pm 1.12\%$), and 67.2% ($\pm 1.04\%$), indicating that the low-dimensional embeddings achieve up to 87.1%, 85.9%, and 95.4% of the optimal predictive accuracy for LLM, MLLM, and human, respectively.

Furthermore, we compared SPoSE embedding's predictive performance to that of the original model features using open-source models. As shown in Fig. 2*i*, the accuracy ratios demonstrate that SPoSE embeddings closely approximate the original features (with ratios around 90%), highlighting their effectiveness as compressed representations (see Extended Data Fig. 1*a* for the number of retained dimensions for these models and their predictive performance curves). Additionally, in Fig. 2*j*, we compared two model probing methods: the behavioral judgment method and the cosine distance method. For the pure language model Llama3.1, the correlation between the two methods was relatively strong ($r = 0.55$), while for the vision-language model Qwen2_VL⁵⁵ (7B version), it was lower ($r = 0.38$). Importantly, the behavioral judgment method aligned better with human-derived RSM than the cosine distance method (0.70 vs. 0.42 for Qwen2_VL, and 0.51 vs. 0.49 for Llama3.1). These results suggest the feasibility of using SPoSE embeddings derived from behavioral judgments to probe the closed-source LLMs/MLLMs where direct feature extraction is infeasible.

Overall, SPoSE modeling generated a low-dimensional, stable, and predictive mental embedding, excelling in predicting triplet similarity judgments and reconstructing their representational space. This indicates that LLM (particularly MLLM) judgments of natural objects are structured and principled. In the following sections, we explore key schemas in this embedding and their connections to human mental representations.

Emergent object category information

Natural object categories emerge from mental embeddings derived from human similarity judgments^{16,38}. To assess whether embeddings from LLM and MLLM also show emergent category structures, we used 18 high-level categories from the THINGS database⁵⁰ and applied a cross-validated nearest-centroid classifier to predict the category membership for each of the 1,112 objects of these categories (see Methods).

As seen in Fig. 3*a*, LLM embeddings achieved 83.4% top-1 accuracy (chance = 9.8%, 95% CI = [8.2%, 11.4%]), while MLLM reached 78.3% (chance = 9.9%, 95% CI = [8.2%, 11.5%]). Human embeddings performed best with 87.1% top-1 accuracy (chance = 10.3%, 95% CI = [8.6%, 12.0%]). Fig. 3*b* shows similar categorization performance between SPoSE embeddings and original features across models, confirming SPoSE's effectiveness in capturing object categories if the model itself is powerful in object representation²⁴. Figs. 3*c-d* visualize the global structure of embeddings via a t-SNE plot (dual perplexity: 5 and 30; 1,000 iterations) initialized with multidimensional scaling (MDS). Objects with similar values cluster together, showing that items from the same category group across LLM, MLLM, and human data. Thus, LLMs inherently capture object category structures without explicit representational constraints. Compared to traditional supervised models (like VGG16⁵⁶) or self-supervised models (like SimCLR⁵⁷), LLMs and humans exhibit superior object category information. Overall, LLM and MLLM results support known distinctions between animate/inanimate and man-made/natural objects, consistent with previous human studies¹⁶.

The embedding dimensions of the LLMs are interpretable and informative

While past research has explored multidimensional mental representations in humans^{16,17}, this study is the first to examine LLMs. We focused on analyzing these dimensions to identify properties prioritized by LLM and MLLM when assessing object similarity. Figs. 4*a-d* visually represent selected dimensions in LLM and MLLM by showing object images weighted most heavily in those dimensions. These dimensions are interpretable, reflecting conceptual and perceptual traits. We assigned intuitive labels (e.g., "animal-related" and "food-related"; see Methods) to dimensions from LLM and MLLM. Some dimensions appear to represent semantic categories (e.g., food, animals, vehicles) (Fig. 4*a*), while others capture perceptual features like hardness, value, temperature, or texture (Fig. 4*b*). Certain MLLM dimensions seem to reflect global spatial properties (e.g., crowded) (Fig. 4*c*), while some convey shape (flatness, elongation) and color (Fig. 4*d*). Dimensions also distinguish user specificity (children vs. adults, everyday consumers vs. experts) (Extended Data Fig. 1*b*), physical composition (wood, ceramic, metal) (Extended Data Fig. 1*c*), and environment-related traits (land vs. sea, indoor vs. outdoor) (Extended Data Fig. 1*d*). See Extended Data Figs. 2-6 for a visual display of all 66 dimensions. Each dimension in LLM or MLLM embodies multiple attributes, but we offer a single interpretation per dimension to showcase the concepts they represent.

We categorized the dimensions into three groups: shared across all three (LLM, MLLM, human), unique to human, and missing from human but present in LLM/MLLM. Shared dimensions include "animal-related" (2, 3), "food-related" (2, 3, 6, 18, 41, 58), "electronics/technology" (5, 11), "transportation/movement" (8, 19, 52, 58), and more. Unique human dimensions include "white" (22), "red" (24), "black" (27), "tubular" (31), "grid/grating-related" (33), "spherical/voluminous" (36), "elliptical/curved" (41), and more. Dimensions missing in humans but present in LLM/MLLM include "vegetable-related" (13, 28), "frozen treats/drink" (22), "presentation/display-related" (23), "headwear-related" (25), "livestock-related" (26), and

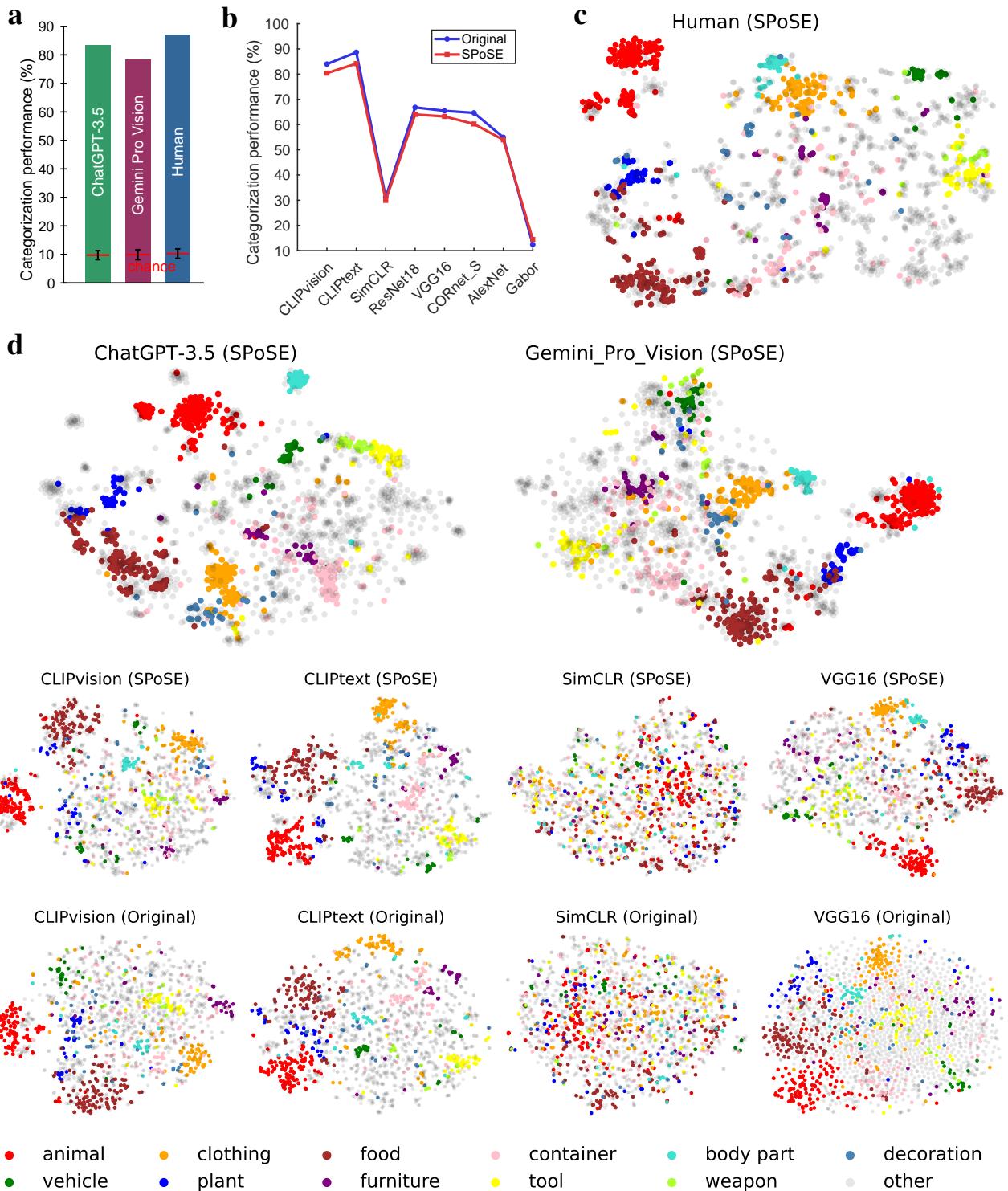


Fig. 3. Emerging object category information in the derived embeddings. **a**, Categorization performance of different embeddings, tested on 18 categories in the THINGS database. Chance-levels are presented as mean \pm 95% CIs, and the error bars reflect 95% CIs ($n=1000$ bootstraps). **b**, Categorization performance comparisons between the SPoSE embedding and original model feature. **c-d**, t-SNE visualization of 1,854 objects, showing emergent category clusters in the learned embedding space of human and models. Dots correspond to objects, and were colored according to their labels.

more. In general, categories such as animals, food, and technology are universally recognized across humans, LLMs, and MLLMs, indicating a common conceptual basis. Humans excel at distinguishing object differences through perceptual features like color, shape, and texture, which are less pronounced in LLM and MLLM. Moreover, LLM and MLLM tend to form more specific categories (e.g., fruits, vegetables, headwear) than humans' broader categorizations. The absence of certain dimensions in human representations does not imply an inability to perceive them; rather, these dimensions may emerge at a higher level, such as humans consolidating "vegetable-related" and "nut-related" dimensions under a "food-related" dimension.

The dimensions derived from LLM and MLLM appear to exhibit a degree of interpretability, as evidenced by the ability to assign intuitive labels to them. These labels were listed in Extended Data Table 1. We also annotated these dimensions using MLLM, comparing human-generated vs. MLLM-generated labels in Extended Data Table 2. In addition, we divided all dimensions into visual, semantic, and mixed visual-semantic groups (based on examination by human experts) and calculated the proportion for each group (Fig. 4e). LLM and MLLM have more semantic dimensions, while humans are better at using visual information. In contrast, the purely vision model SimCLR (a self-supervised learning model) shows minimal ability to learn semantic dimensions (Extended Data Fig. 7), whereas the dimensions derived from random representations lack any interpretability (Supplementary Fig. 1). We also categorized dimensions by ease of interpretation (based on whether they can be clearly explained by a single label), finding that most dimensions are easy to interpret (Fig. 4f). Specifically, 60/66 dimensions for LLM, 57/66 for MLLM, and 62/66 for humans are easy to interpret, with humans having the fewest hard-to-interpret dimensions.

We examined the composition of dimensions for specific objects. Fig. 4g uses circular bar plots to represent objects, where petal angle and color denote dimensions, and length indicates the dimension's importance. For example, "almond" is primarily food-related, while "satellite" is associated with electronics and flying. These plots also demonstrate that objects are indeed characterized by a rather small number of dimensions, indicating that not all 66 dimensions are necessary for particular similarity judgment. To quantify this, we progressively eliminated less significant dimensions for each object and assessed model performance. We found that retaining 3 to 8 dimensions for LLM, 2 to 10 for MLLM, and 7 to 13 for humans suffices to achieve 95-99% of the full model's performance in explaining behavioral judgments within the odd-one-out context (Fig. 4h). LLM exhibits lower dimensionality than humans, likely due to its lack of visual input. Although MLLM can access visual data, its multimodal integration remains inferior to human capabilities, limiting dimensions related to shape or color, inherently tied to human visual experience.

Comparison between models and humans

We employed two approaches to assess model-human alignment: one measuring consistency in similarity judgments⁵⁸ and the other analyzing core dimension relationships.

Using comprehensive triplet sampling on 48 objects, we estimated similarity via choice probabilities and correlated model and human similarity matrices with Pearson correlation. Fig. 5a compares various models, including visual-only, visual-language, LLMs, MLLMs, and a Gabor baseline, revealing higher human-consistency for LLM and MLLM. A preliminary comparison between ChatGPT-3.5 and GPT-4 in Fig. 5b, directly based on their choice consistency with human on 2,171 triplets, shows that notable differences remain between LLMs and human. To delve deeper into the reasons behind these differences, we show in Fig. 5c the most relevant dimensions that humans and models rely on to make choices (see Methods). We see that human and models make different choices because of the differently key dimensions they rely on. For example, human can make choice based on color (like "red"), while LLM only makes choice based on semantics (like "protective"). More examples are in Extended Data Fig. 1f.

Next, we explored the relationship between the core dimensions of LLMs and humans, as shown in Fig. 5d. The matrices are generally sparse, indicating that a dimension in one system strongly correlates with only a few dimensions in the other. Many dimensions even show a strong one-to-one mapping. Quantitatively, 31 out of the 66 LLM dimensions and 42 out of the 66 MLLM dimensions strongly correlate with human dimensions ($r > 0.4$), indicating substantial alignment. In MLLM, several human dimensions are subdivided (e.g., human dim. 18 "fluid-related" splitting into MLLM dims. 18 "container" and 22 "fluid-related") or amalgamated (e.g., human dims. 3 "animal-related" and 40 "disgusting" merging into MLLM dim. 34 "insect-related"). Similarly, LLM shows adaptations, particularly in semantics, though it lacks sensory dimensions like color or shape. For example, LLM distinguishes between dim. 22 "frozen treats" and dim. 57 "hot drinks" (or dim. 2 "wild animals" vs. dim. 26 "livestock," dim. 13 "vegetables" vs. dim. 18 "fruits," etc.). While MLLM still lacks specific color-related dimensions (e.g., "red," "black"), it aligns more closely with humans, especially in dimensions like shape (e.g., dim. 35 "grainy," dim. 64 "round/curvature") and spatial features (e.g., dim. 8 "serried/stacked," dim. 44 "dense/many small things"). This shows that MLLM, like humans, can perceive a large amount of visual information. Quantitatively, Fig. 5e shows the number of shared and unique dimensions ($r > 0.2$) between models and humans, where 38 of 66 dimensions being shared across the three systems.

Relationship to the cerebral representational geometries

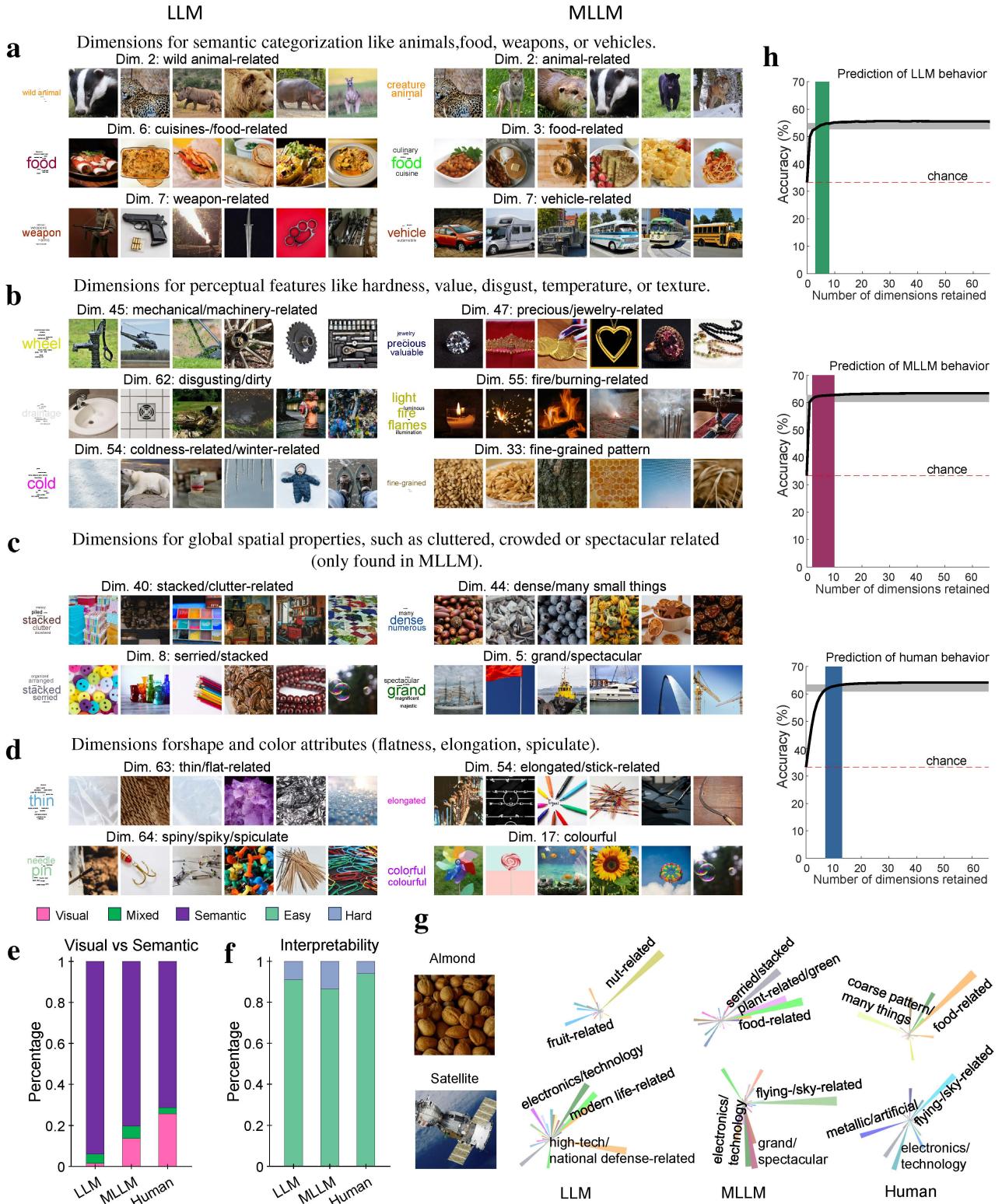


Fig. 4. Object dimensions illustrating their interpretability. **a-d**, For each dimension, visualization includes the top 6 images carrying the greatest weights, accompanied by a word cloud reflecting human's annotations for what is captured by the dimension. For LLM, we replaced linguistic descriptions with images of the related objects to aid visualization. **e**, Proportions of visual, semantic, and mixed visual-semantic dimensions. **f**, Proportions of easy and hard to interpret. **g**, Illustration of example objects with their dominant dimensions. **h**, To explain 95 to 99% of the predictive performance in behavior, how many dimensions are required. For subfigures **a-d**, **g**, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.

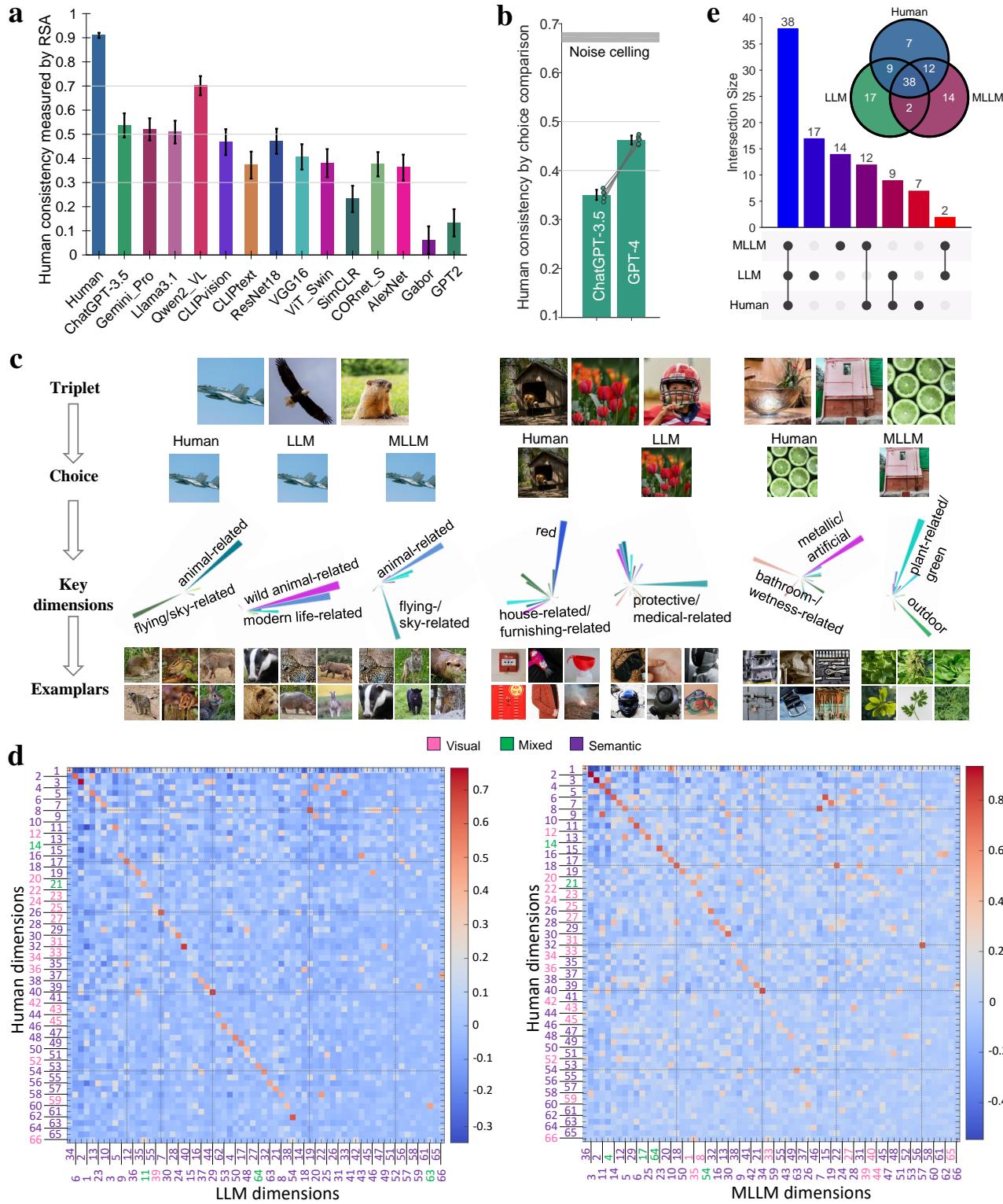


Fig. 5. Comparison between models and humans. **a**, Human-model consistency (Pearson's r) between human and model object similarity matrices. Left blue bar shows baseline between-human consistency. Data are presented as mean \pm 95% CIs, and the error bars reflect 95% CIs ($n=1000$ bootstraps). **b**, Preliminary comparison between ChatGPT-3.5 and GPT-4. The error bars reflect standard deviation (SD), and data are presented as mean \pm SD ($n=5$ samplings, and dots represent the result of each time). **c**, Key dimensions that underpin specific behavioral choices made by human and models. **d**, Cross-correlation matrix between each pair of model systems (human-LLM, human-MLLM, and LLM-MLLM (in Extended Data Fig. 1e)). **e**, Quantification of shared ($r > 0.2$) and non-shared dimensions between different systems. For subfigure **c**, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.

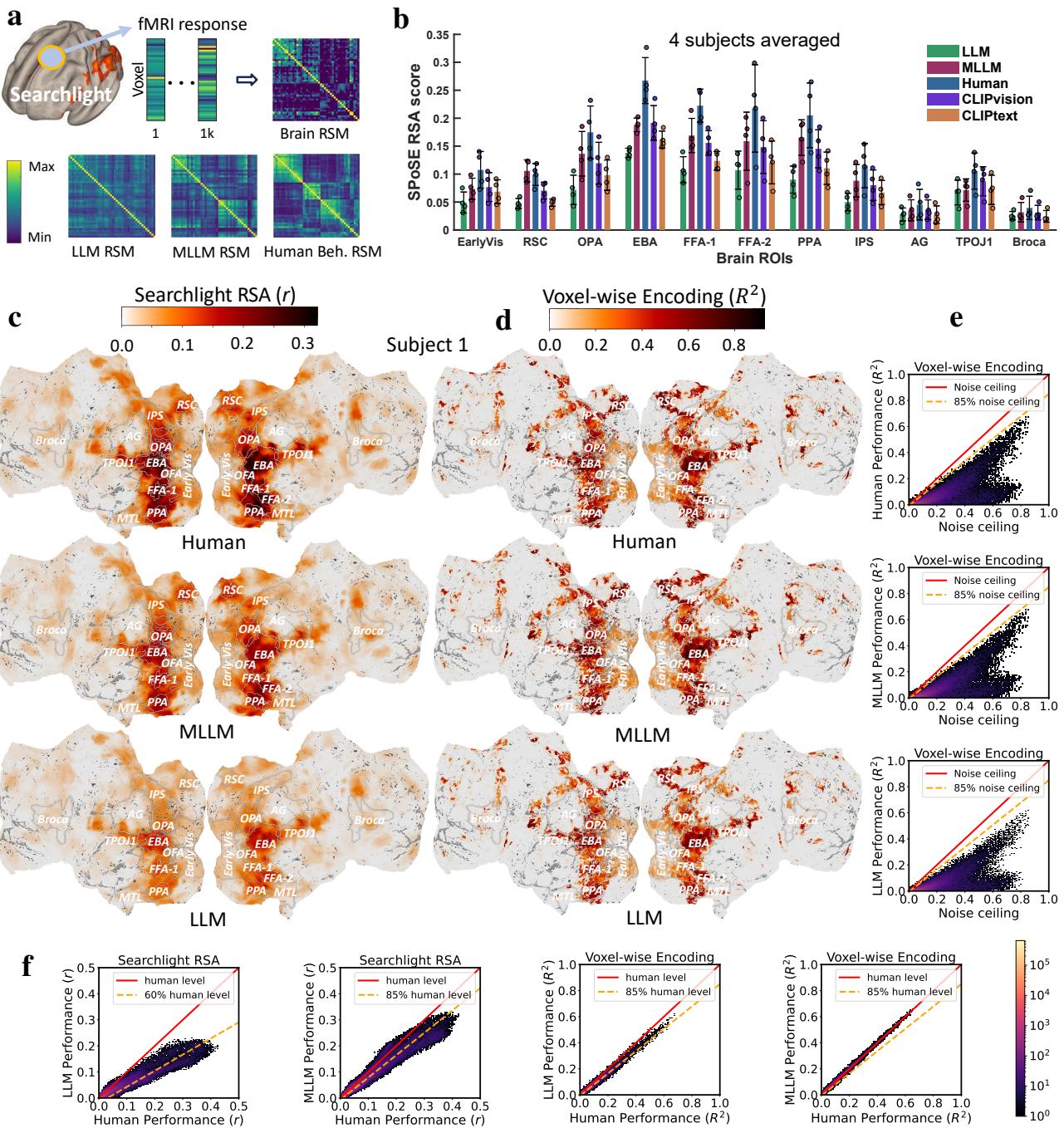


Fig. 6. Relationship to the cerebral representational geometries. **a**, Searchlight brain RSM and the varied model RSMs on the NSD shared_1k dataset. **b**, RSA between model RSM and brain ROI RSM constructed from the SPoSE embedding of that brain ROI (see Methods). The error bars reflect SD, and data are presented as mean \pm SD ($n=4$ subjects, and dots represent the scores of different individuals). **c-d**, Cortical maps of searchlight RSA and voxel-wise encoding (evaluated by using R^2 with noise ceiling normalization). For visualization purpose, we only conducted noise ceiling normalization for voxels that have the predicted $R^2 > 0.2$. **e**, 2-D histograms of human, LLM and MLLM performance in R^2 against noise ceiling across all voxels in the whole brain. **f**, 2-D histograms of LLM, MLLM against human performance.

To link LLMs' embeddings with brain responses, we applied searchlight RSA⁵³ (see Fig. 6a) using fMRI data from the NSD dataset⁵². Independent dimension rating models were fitted for each dimension, and these models predicted multi-dimensional embeddings for objects, creating a representational geometry. We then compared this predicted RSM to SPoSE embedding RSMs of brain ROIs and searchlight RSMs of brain sectors to gauge how well the LLM's embedding aligns with brain regions.

The representational similarity scores for each model and brain ROI are depicted in Fig. 6b. It should be noted that we adopted the SPoSE method to infer low-dimensional embeddings for CLIP⁵⁹ (here used as a strong baseline⁶⁰) and brain ROIs, using cosine distance as a metric to construct the desired odd-one-out records. Human and MLLM embeddings outperform LLM and CLIP, particularly in functionally defined, category-selective ROIs (e.g., EBA, PPA, RSC, FFA). However, ROI-based analysis may miss fine-grained spatial patterns, as similar scores can conceal spatial differences.

Figs. 6c&d display fine-grained cortical maps of human, LLM, and MLLM embeddings using searchlight RSA and voxel-wise encoding (see Methods) for subject S1, highlighting only significant voxels ($P < 0.05$, FDR-corrected). Additional models and subjects are shown in Extended Data Fig. 8a. Visual inspection shows MLLM and human embeddings align more closely with most of the brain regions than LLM and CLIP, and the contrast of local details can also be clearly viewed. This performance difference is most obvious under searchlight RSA, and relatively moderate in voxel-wise encoding. Beyond overall performance metric, peaks in the cortical maps align with scene-selective⁶¹ (PPA, RSC, OPA), body-selective⁶² (EBA) and face-selective^{63,64} (FFA, OFA) ROIs, suggesting MLLM captures semantic relationships similar to human cognition. Furthermore, both the overall performance levels and the pattern consistency remain stable across multiple subjects (Extended Data Fig. 8a). Voxel-wise encoding results based on the original CLIP embedding and its low-dimensional SPoSE embedding (Extended Data Fig. 8b) also provide strong evidence that SPoSE is an effective intrinsic dimension learning method. Fig. 6e presents 2-D histograms of human, LLM and MLLM performance in R^2 against noise ceiling across all voxels. For human and MLLM, most voxels in the category-selective ROIs (e.g., EBA, PPA, RSC, FFA) are predicted close to their 85% noise ceiling, while LLM is slightly worse. Fig. 6f presents 2D histograms comparing LLM and MLLM to human performance across whole brain voxels. LLM and MLLM achieve about 60% and 85% of human performance under searchlight RSA, respectively. In voxel-wise encoding, LLM reaches 90% of human performance, while MLLM nearly matches human levels.

Discussion

The present study comprehensively investigates object concept representations in LLMs and MLLMs, and their relationship to human cognition and brain representations. We collected 4.7 million behavioral judgments to derive 66 stable dimensions predicting object similarity, uncovering semantic clustering in both LLM and MLLM embeddings, resembling human mental structures. Despite differing architectures, these models developed conceptual representations similar to humans, supported by interpretable dimensions reflecting core aspects of object understanding. MLLM, which integrates visual and linguistic data, predicted individual choices at 85.9% of the noise ceiling, consistent with findings that multimodal learning enhances representation robustness and generalizability^{65–67}. Moreover, the strong alignment between MLLM embeddings and neural activity in regions like EBA, PPA, RSC, and FFA suggests that MLLM representations share similarities with human conceptual knowledge⁶⁸.

Broad applications of the derived embeddings

The low-dimensional mental embeddings identified in this study can be used in human-machine representation alignment and fusion, potentially enhancing human-machine interfaces and collaborative systems by revealing shared object representation schemas. Practically, these interpretable dimensions could inform the development of more human-like artificial cognitive systems, improving their natural interaction with humans⁶⁹. To better align LLM and MLLM with human reasoning in the odd-one-out task, we can explore the method of guiding model attention to human-preferred dimensions. By tailoring prompts to emphasize specific attributes (e.g., "red" or "artificial"), we believe that models could make choices more consistent with human judgments (i.e., explicit guidance can help bridge the gap between model and human reasoning; Supplementary Figs. 2–4). Moreover, the collected extensive machine behavioral datasets offer a valuable benchmark for evaluating AI model representations.

Relationship to the other related studies

Both the human brain and large-scale AI models are complex systems, typically analyzed through dimensionality reduction. Recent hypotheses like the "low-rank"⁷⁰ and "distributed information bottleneck"⁷¹ propose solutions to identifying optimal latent dimensions. Our findings align with these concepts, demonstrating that LLMs can develop human-like object representations using fundamental dimensions, akin to the brain's capacity to derive rich conceptual knowledge from simple neural mechanisms. Exploring these low-dimensional structures could deepen our understanding of cognition in both biological and artificial systems.

The similarity between LLMs and human representations, despite differing input modalities, suggests a convergence beyond data covariance. This is consistent with findings on innate semantic transformations in the visual system⁷², and is further supported by the interpretability of LLMs' embeddings, reflecting fundamental semantic structures. Prior studies^{73–75} demonstrate that artificial models can predict visual brain activity, which aligns with our results showing model-neural correlations in higher cortical regions. These findings suggest LLMs develop representations that capture key aspects of human conceptual knowledge^{76,77}, further highlighting the natural alignment between language and vision^{78,79}. Previous fMRI studies have revealed diverse organizational principles in the brain for processing external stimuli. The primary visual cortex exhibits retinotopy through eccentricity and angle selectivity^{80,81}. These principles of dimensional organization extend to higher-order information^{82–88}. Our study expands this research to the conceptual representations of natural objects.

Traditionally, neural network representations are analyzed by examining neuron activation patterns^{89–92}. However, as AI systems grow in complexity, neuron-level approaches become less effective. Instead, inspired by cognitive psychology, behavioral methods can infer AI system representations through actions. Decades of research have developed techniques to elucidate mental representations from human behavior^{16,93}. Our study adopts this behavioral approach for LLMs, complementing existing neuron-level methods. Probing LLMs from a cognitive perspective has gained attention^{35,94–98}, revealing insights into areas like color processing⁹⁹, emotion analysis^{100,101}, memory^{102,103}, morality¹⁰⁴, and decision-making^{40,105,106}. Understanding the parallels between human cognition and LLMs offers exciting opportunities to explore the intersections of AI and cognitive science^{37,69}.

Limitations and future directions

One potential limitation of this study is its focus on ChatGPT-3.5 and Gemini Pro Vision (v1.0), which may not encompass the full spectrum of models. However, the methodology is extendable to other state-of-the-art LLMs such as GPT-4V¹⁰⁷. This extension could reveal the generalization of identified dimensions and highlight the unique aspects of different AI architectures. Another potential limitation is that the impact of varying language prompts on LLMs' responses. In this study, the language prompts we used were carefully designed to ensure that the LLMs understand the task instructions correctly. We think that these considerations have a negligible impact on the study's overall conclusions. Moreover, we only employed object-level annotations in the language prompts of LLM. Object-level annotations focus on abstract categories, while image-level annotations (generated by a vision-language model or human annotators) can capture more image-specific visual attributes like color and texture (Supplementary Fig. 5). Using the image-level annotations will make LLM more consistent with human judgments (this can be confirmed in the MLLM probing experiments, which is equivalent to using image-level annotation in essence), highlighting the importance of visual information in similarity judgments (Supplementary Figs. 6–8).

Future work could leverage instruction fine-tuning for LLM/MLLM on large-scale triplet odd-one-out question-answer pairs, where answers include both human choices and the underlying reasoning dimensions, to improve model-human alignment.

Methods

Stimuli and triplet odd-one-out task. In selecting stimulus objects, our preference was for the THINGS database⁵⁰, a resource designed to encompass 1,854 living and non-living objects based on their practical usage in daily life. During the triplet odd-one-out task, participants (humans or LLMs) encountered three objects drawn from the THINGS database, either through images or textual descriptions. Their objective was to identify the object with the highest dissimilarity among the three. This task evaluates the relationship between two objects considering the context set by a third object. Featuring a diverse range of objects, this method provides a systematic means to assess perceived similarity unaffected by context, thus minimizing response bias. Moreover, it enables the measurement of context-dependent similarity, such as by restricting similarity evaluations to specific higher-level categories like animals or vehicles.

Behavioral responses from humans. The human behavioral dataset utilized in our research originated from a recent study¹⁷, where 5,517,400 human similarity judgments were collected via Amazon Mechanical Turk. After quality control—which excluded 818,240 trials (14.83%) based on overly fast responses (>25% trials <800ms and >50% <1,100ms), repetitive patterns (outside central 95% distribution in ≥200 trials), and inconsistent demographic reporting (>3 ages provided)—the final dataset comprised 4,699,160 valid trials from 12,340 participants. Participants (6,619 female; 4,400 male; 56 other/unspecified; mean age = 36.71 years, SD = 11.87; 41.9% unreported age) were right-handed with normal/corrected vision, compensated at \$0.10 per 20 trials. The protocol, approved by the NIH Institutional Review Board (93-M-0170) and NIH Office of Human Research Subject Protection, obtained informed consent. While self-selection bias (tech-savvy English-speakers) and handedness exclusion may limit generalizability, the focus on relative similarity judgments—demonstrated robust across demographics¹⁶—reduces population-specific effects.

Collecting behavioral responses from LLM. For our study, we gathered all human-used similarity judgments, totaling 4.7 million trials. To solicit responses from ChatGPT-3.5 (gpt-3.5-turbo), Llama3.1 (Meta-Llama-3.1-8B-Instruct), and GPT-4 (gpt-4-0314), we employed a prompt where each image was represented by its object name and descriptions, as image input processing was not supported by these models. These text descriptions are sourced from definitions of object names in WordNet, Google, or Wikipedia, and have been compiled and made publicly available at <https://osf.io/jum2f/>. For model comparison, Llama3.1 was used to collect the full sampling of triplets (91,568 trials) of the 48 typical objects. Due to cost constraints, GPT-4 only amassed a total of 2,171 trials, primarily for initial comparisons with ChatGPT-3.5.

The prompt structure used was standardized: "*Given a triplet of objects {[Object_A], [Object_B], [Object_C]}, which one in the triplet is the odd-one-out? Please give the answer first and then explain in detail.*" In practice, [Object_A], [Object_B], and [Object_C] were replaced with the respective object descriptions for each trial. The temperature parameter, dictating response randomness in LLMs, was set to 0.01. Because of the well-structured nature of the model's responses, we parsed the model choice from the first sentence of their response using string matching. To assess the upper limit of predictability under dataset randomness (the noise ceiling), we randomly selected 1,000 triplets and conducted a minimum of 14 trials and a maximum of 25 trials for each using the same prompt, evaluating consistency in choices across trials.

Collecting behavioral responses from MLLM. Regarding collecting behavioral responses from Gemini Pro Vision (v1.0), we adopted a similar strategy. The prompt we used is as follows: "*You are shown three object images side by side and are asked to report the image that was the least similar to the other two. You should focus your judgment on the object, but you are not given additional constraints as to the strategy you should use. If you did not recognize the object, you should base your judgment on your best guess of what the object could be. 1. Tell me your answer. 2. Tell me the location of the object you have chosen. 3. Explain the reasons.*" In some trials, the Gemini Pro Vision model refused to respond because it believed that the given images contained some unknown sensitive information. In this case, we applied a method akin to image replacement to address the issue.

The temperature parameter for determining response randomness in Gemini Pro Vision was also configured to 0.01, with images displayed at 512 x 512 pixels. Since the model's responses are well structured, we extracted the keyword about the position of the object in its answers (e.g., "left," "middle," or "right") to determine the model's choice. Similarly, to gauge the noise ceiling and potential predictability, we additionally sampled 1,000 randomly chosen triplets and ran a minimum of 14 trials and a maximum of 25 trials for each of them using the same prompt for each trial and estimated the consistency of choices for each triplet across trials.

As for the model of Qwen2_VL-7B, we used a similar strategy to collect the full sampling of triplets for the 48 typical objects.

Constructing behavioral responses for the other models. For models do not have visual or language-based question-answer capabilities (such as CLIP, SimCLR, VGG16, etc.), we first used the pre-trained model to extract the features of the object images (or their language descriptions), and then constructed the required odd-one-out data based on the cosine distance of the features.

Feature extractors. For the pre-trained models originally used for classification tasks (such as VGG16, ResNet18, etc.), we extracted the penultimate layer features, rather than the head. For CLIP, we extract features in the final embedding layer. For GPT2 and Llama3.1, we extracted features by averaging the last hidden state activations across all tokens to obtain sentence embeddings. For Qwen2_VL, we extracted image features from the last layer of its visual branch, which is based on a 600M-parameter ViT. Some of the pretrained models sourced from the following repositories: the Torchvision model zoo, the Pytorch-Image-Models (timm) library, the VISSL (self-supervised) model zoo, the OpenAI CLIP collection, and the Transformer python library. In particular, the Gabor model feature extractor consists of a single fixed set of convolutions: 12 Gabor wavelets with spatial frequency log-spaced between 3 and 72 cyc/stimulus at 6 evenly-spaced orientations between 0 and π , following previous work¹⁰⁸.

Natural Scene Dataset (NSD). NSD⁵², recognized as the largest neuroimaging dataset linking brain insights with artificial intelligence, involves richly sampled fMRI data from 8 subjects. Across 30-40 MRI sessions, each subject observed between 9,000-10,000 distinct natural scenes using whole-brain gradient-echo EPI at 1.8 mm isotropic resolution and 1.6 s TR during 7T scanning. Image stimuli were drawn from the COCO dataset¹⁰⁹, with corresponding captions retrievable using COCO ID. To assess the generalization ability of the low-dimensional embeddings learned from humans and LLMs across datasets, the shared_1k subset from the NSD were chosen as the test set (because the stimuli in this subset were shared by all 8 subjects). Additionally, fMRI responses linked to the shared_1k stimuli across subjects S1, S2, S5, and S7 were earmarked for subsequent analysis (because subjects S3, S4, S6, and S8 did not complete the full fMRI data acquisition).

Sparse Positive Similarity Embedding (SPoSE). Utilizing the SPoSE approach^{16,39}, we derived embedding representations for 1,854 objects based on similarity judgment data from LLM and MLLM, respectively. The PyTorch implementation for this process can be accessed at <https://github.com/ViCCo-Group/SPoSE>. Initially, an embedding matrix \mathbf{X} was created with random weights in the range of 0 to 1 across 100 latent dimensions for each object, resulting in a 1854-by-100 matrix. Stochastic gradient descent was subsequently applied to fine-tune this embedding matrix using odd-one-out responses. The optimization objective function aimed to minimize a combination of cross-entropy loss concerning triplet choice probabilities for all options and an L1-norm on the weights to promote sparsity:

$$\min \mathcal{L}(\mathbf{x}) = \sum_{i=1}^n \log \left(\frac{\exp(\mathbf{x}_i \mathbf{x}_j)}{\exp(\mathbf{x}_i \mathbf{x}_j) + \exp(\mathbf{x}_i \mathbf{x}_k) + \exp(\mathbf{x}_j \mathbf{x}_k)} \right) + \lambda \sum_{i=1}^m \|\mathbf{x}\|_1, \quad (1)$$

where \mathbf{x} corresponds to an object vector; i , j and k to the indices of the current triplet; n to the number of triplets; and m to the number of objects. The regularization parameter λ , which controls the trade-off between sparsity and model performance, was determined using cross-validation on the training set ($\lambda = 0.004$ for LLM, 0.0035 for MLLM, 0.00385 for humans, and 0.007 for the other models and brain ROIs). In addition to sparsity, the optimization was constrained by strictly enforcing weights in the embedding \mathbf{X} to be positive. The minimization of this objective was carried out using stochastic gradient descent with an Adam optimizer¹¹⁰ (with default parameters) and a batch size of 100 on triplet odd-one-out judgments. After the optimization was complete, dimensions with weights below 0.1 for all objects were eliminated. Finally, the dimensions underwent sorting based on the sum of their weights across objects in descending order.

This model operates under two key theoretical assumptions. Firstly, it postulates sparsity within the embedding space dimensions, indicating that each object primarily influences certain dimensions rather than all. Secondly, it assumes positivity in these dimensions. Consequently, an object's weight on a specific dimension signifies the extent of the related property within the object. These assumptions diverge from typical dimensionality reduction approaches like Principal Component Analysis (PCA), which assume dense dimensions across the real number spectrum. Furthermore, SPoSE facilitates cross-correlations among dimensions while PCA assumes independence. Consequently, SPoSE often uncovers a greater number of dimensions, reflecting finer details or attributes, which are more easily interpretable compared to PCA dimensions. Notably, the weight an object holds on a dimension directly corresponds to the presence of the associated property within the object.

We opted for the behavioral odd-one-out task and the SPoSE method to learn the low-dimensional embeddings of LLMs rather than attempting to directly access their internal features, primarily due to the challenges associated with extracting features from modern, large-scale LLMs that are often proprietary or too vast to navigate directly. This approach allows us to circumvent the limitations imposed by the closed nature or sheer scale of contemporary LLMs, providing us with a more feasible avenue to explore their mental representations.

Reproducibility of embedding dimensions. Considering the stochastic nature of the optimization process, the SPoSE method yields varying sets of dimensions upon each reiteration. To assess the stability of the 66-dimensional embedding, we conducted 20 model runs with distinct random initializations. Evaluating each original dimension against all dimensions in the 20 reference embeddings, we identified the best-matching dimension based on the highest correlation. Consistent with previous research¹⁶, a Fisher z-transform was applied to these correlations, averaged across the 20 reference embeddings, and then reversed to obtain a mean reliability value for each dimension across all 20 embeddings.

Category prediction. Evaluating the representational embeddings' categorization performance involved testing them across 18 out of the 27 THINGS database categories. Objects falling into multiple categories were excluded from the analysis, resulting in the removal of 9 categories. Among these excluded categories, 7 were subcategories or had less than ten unique objects post-filtering. The remaining 18 categories included clothing, toy, vehicle, container, electronic device, animal, furniture, body part, food, musical instrument, plant, home decor, sports equipment, office supply, part of car, medical equipment, tool, and weapon, totaling 1,112 objects. Classification was conducted through leave-one-object-out cross-validation. Training involved computing category centroids by averaging the 66-dimensional vectors of all objects within each category, excluding the left-out object. The category membership of the excluded object was predicted based on the smallest Euclidean distance to the respective centroid. This process was iterated for all 1,112 objects, with prediction accuracy averaged across the dataset. The chance level is determined by 1000 permutation tests.

Evaluating consistency between humans and models by comparing behaviors. With the exception of GPT-4, all other models (and human) have completed behavioral data acquisition on the full sample triples of the 48 typical objects described above. For each model, we constructed its RSM for the 48 objects by calculating the choice probability of each object pair. To estimate human consistency, following previous work⁵⁸, we computed the Pearson correlation on the behavioral RSMs from the model (m) and the human (h) and we then divide that raw Pearson correlation by the geometric mean of the split-half internal

reliability measured for each system as follows:

$$\tilde{\rho}(m, h) = \frac{\rho(RSM_m, RSM_h)}{\sqrt{\rho(RSM_m^{half_1}, RSM_m^{half_2})\rho(RSM_h^{half_1}, RSM_h^{half_2})}}, \quad (2)$$

where $RSM_m^{half_1}$ and $RSM_m^{half_2}$ were computed by using the split-half behavioral data of triples of the 48 typical objects, and similar for $RSM_h^{half_1}$ and $RSM_h^{half_2}$.

Comparison between ChatGPT-3.5 and GPT-4 was conducted directly based on their choice consistency with human on a specific set of 2,171 triplets. We conducted a total of 5 comparisons, each based on randomly selecting 1,000 samples from these 2,171 samples, and finally reported the average result.

Dimensional relevance score for odd-one-out choice. For a given triplet, we compute the original predicted softmax probabilities based on the entire low-dimensional embeddings of each image within the triplet. Then, we iteratively remove a certain dimension from the low-dimensional embeddings, calculate the softmax probabilities predicted by the pruned embeddings, and then compute the difference between the softmax probabilities obtained before and after pruning. This difference is taken as the relevance score for that dimension. This approach has been used in a previous study²⁶.

Dimension naming. In defining the human mental embedding, the dimension names from a previous investigation were employed as references¹⁷. However, for LLM and MLLM, each of the 66 dimensions within the embedding was associated with common-sense labels through a straightforward naming procedure. Specifically, we analyze a set of 1-by-12 images of objects and identify shared properties described in the images. Each array consisted of images selected from the top of one dimension from the embedding. Ten of the authors provided concise labels, limited to 1–2 words, describing the arrayed images. Subsequently, word clouds were generated to visualize dimension names, showcasing the distribution of labels based on frequency, utilizing the wordcloud function in MATLAB (Mathworks) with default settings. Finally, the lead authors of this study gave intuitive labels for each dimension. Dimension labels were also summed up by the MLLM (here gemini-pro-1.5-exp) with the prompt as follows: *"There are 9 subfigures in the picture. Please use 1-2 English words or phrases to describe the common theme represented by these 9 subfigures."*

Dimension rating for NSD images. We predicted the 66 object dimensions for each image within the NSD dataset. Specifically, we leveraged the OpenAI-trained CLIP model⁵⁹ (with "ViT-L/14" as the backbone), which is a multimodal model trained on image-text pairs and which was recently demonstrated to yield excellent prediction of human similarity judgments^{111,112}. For each of the 1,854 object images in the THINGS dataset, we extracted the image and text features from the final layer of the CLIP image and text encoders, respectively. Subsequently, for each of the 66 dimensions of LLM (or MLLM, or Human), we fitted a ridge regression model to predict dimension values, using a concatenation of the extracted image and text features from CLIP as input. The optimal regularization hyperparameters were determined by using 5-fold cross-validation across the training set (100 candidate parameters spaced evenly on a log scale from 10^{-3} to 10^3 , that is np.logspace(-3, 3, 100)). These trained regression models were then applied to the extracted features across all images in the NSD dataset.

Searchlight RSA. For fMRI, local cerebral RSMs were computed in subject space within a grey-matter spherical region (6 mm diameter) centered at each voxel location. RSA analyses assessed the Pearson correlation r between the local cerebral RSM and each kind of the model RSMs.

SPoSE RSA. For each brain ROI, we extracted the fMRI signal in that region on the shared_1k dataset and constructed a large number of odd-one-out data based on the cosine distance. After that, SPoSE learning was used to obtain the corresponding low-dimensional embeddings of each brain ROI, and the RSMs of each ROI were calculated using the learned low-dimensional embeddings. Finally, Pearson correlations between the brain ROI RSM and the model RSM were calculated.

Voxel-wise encoding. For each subject in the NSD, we built a ridge regression model to predict the fMRI response to each test image per voxel. The images of the training set are subject-specific, but the images of the test set are shared (that is, shared_1k). For all training and testing images, we first used the dimension rating model to predict the low-dimensional embeddings, and then conducted voxel-wise fitting based on the predicted embeddings. The regularization parameter for each voxel was selected autonomously through a 5-fold cross-validation process on the training dataset. We explored 100 evenly spaced regularization parameters on a logarithmic scale ranging from 10^{-3} to 10^3 , which corresponds to the np.logspace(-3, 3, 100) function in Python. The model's accuracy was assessed on the test dataset utilizing both Pearson's correlation coefficient (r) and the noise

ceiling normalized coefficient of determination (R^2). Following the NSD work⁵², the noise ceiling was calculated by:

$$NC = 100 \times \frac{ncsnr^2}{ncsnr^2 + \frac{1}{n}}, \quad (3)$$

where n indicates the number of trials that are averaged together ($n = 3$ for subjects S1, S2, S5, and S7), and $ncsnr$ indicates the noise ceiling signal-to-noise ratio which has been provided in NSD. To ascertain the statistical significance of our predictions, we conducted a bootstrapping procedure, resampling the test dataset with replacement 2,000 times, and subsequently calculated the False Discovery Rate (FDR) adjusted P -values.

Abbreviation of Brain ROIs. EarlyVis: early visual cortex; Scene, PPA: parahippocampal place area, OPA: occipital place area, RSC: retrosplenial cortex; Body, EBA: extrastriate body area; Face, FFA-1: fusiform face area 1, FFA-2: fusiform face area 2; Mind and Language, TPOJ-1: temporoparietal junction 1, AG: angular gyrus, Broca, MTL: medial temporal lobe.

Visualization of cerebral cortex. To visualize the analytical outcomes across the entire cortical region, we employed flattened cortical surfaces derived from individual subjects' anatomical images. FreeSurfer¹¹³ facilitated the generation of cortical surface meshes from T1-weighted anatomical images. This process involved applying five relaxation cuts on each hemisphere's surface and excluding the corpus callosum. Subsequently, functional images were registered to the anatomical images and mapped onto the surfaces for visualization purposes using Pycortex¹¹⁴.

Data availability

The THINGS database is accessible at <https://osf.io/jum2f/>. The behavioral triplet odd-one-out datasets for Human, ChatGPT-3.5, and Gemini Pro Vision 1.0 can be found at <https://osf.io/f5rn6/>, <https://osf.io/qn5uv/>, and <https://osf.io/qn5uv/>, respectively. Those interested in the preprocessed NSD fMRI dataset supporting this research can obtain it from <http://naturalscenesdataset.org/>. Language descriptions for the 1,854 THINGS objects, the learned mental embeddings of LLM and MLLM, as well as the human and MLLM annotated dimension names are shared in <https://osf.io/qn5uv/>.

Code availability

The code used for data collection, embedding learning, dimension rating, result analysis, and visualization in this study is publicly available on GitHub (https://github.com/ChangdeDu/LLMs_core_dimensions¹¹⁵).

Acknowledgements

This work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB1010202); in part by the National Natural Science Foundation of China under Grant 62020106015 and Grant 62206284; in part by Beijing Natural Science Foundation under Grant L243016, and in part by the Beijing Nova Program under Grant 20230484460. We would like to thank Martin N. Hebart for sharing the THINGS database and 4.7 million human behavioral responses. We also thank Emily J. Allen and Kendrick Kay for sharing the NSD fMRI data. All illustrative images in this article were sourced from Pixabay and Pexels due to copyright restrictions.

Author contributions

C.D. and H.H. designed the research. C.D. conducted the experiments. C.D., Y.S., K.F., and J.P. collected the data. C.D. wrote the paper. C.D., B.W., W.W., Y.G., S.W., C.Z., J.L., S.Q., L.C. and H.H. analyzed the results. All authors read and approved the paper.

Competing interests

The authors declare no competing interests.

References

1. Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychol. review* **94**, 115 (1987).
2. Edelman, S. Representation is representation of similarities. *Behav. brain sciences* **21**, 449–467 (1998).
3. Nosofsky, R. M. Attention, similarity, and the identification–categorization relationship. *J. experimental psychology: Gen.* **115**, 39 (1986).
4. Goldstone, R. L. The role of similarity in categorization: Providing a groundwork. *Cognition* **52**, 125–157 (1994).
5. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. *Cogn. psychology* **8**, 382–439 (1976).
6. Mahon, B. Z. & Caramazza, A. Concepts and categories: A cognitive neuropsychological perspective. *Annu. review psychology* **60**, 27–51 (2009).
7. Rogers, T. T. & McClelland, J. L. *Semantic cognition: A parallel distributed processing approach* (MIT press, 2004).
8. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).
9. Battleday, R. M., Peterson, J. C. & Griffiths, T. L. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nat. communications* **11**, 5418 (2020).
10. Jagadeesh, A. V. & Gardner, J. L. Texture-like representation of objects in human visual cortex. *Proc. Natl. Acad. Sci.* **119**, e2115302119 (2022).
11. Grand, G., Blank, I. A., Pereira, F. & Fedorenko, E. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. human behaviour* **6**, 975–987 (2022).
12. Connolly, A. C. *et al.* The representation of biological classes in the human brain. *J. Neurosci.* **32**, 2608–2618 (2012).
13. Downing, P. E., Chan, A.-Y., Peelen, M. V., Dodds, C. & Kanwisher, N. Domain specificity in visual cortex. *Cereb. cortex* **16**, 1453–1461 (2006).
14. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
15. Caramazza, A. & Shelton, J. R. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *J. cognitive neuroscience* **10**, 1–34 (1998).
16. Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. human behaviour* **4**, 1173–1185 (2020).
17. Hebart, M. N. *et al.* THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife* **12**, e82580 (2023).
18. Konkle, T. & Oliva, A. A real-world size organization of object responses in occipitotemporal cortex. *Neuron* **74**, 1114–1124 (2012).
19. Konkle, T. & Oliva, A. Canonical visual size for real-world objects. *J. Exp. Psychol. human perception performance* **37**, 23 (2011).
20. Bowers, J. S. *et al.* Deep problems with neural network models of human vision. *Behav. Brain Sci.* **46**, e385 (2023).
21. Hermann, K., Nayebi, A., van Steenkiste, S. & Jones, M. For human-like models, train on human-like tasks. *Behav. Brain Sci.* **46**, e394 (2023).
22. Jha, A., Peterson, J. C. & Griffiths, T. L. Extracting low-dimensional psychological representations from convolutional neural networks. *Cogn. science* **47**, e13226 (2023).
23. Nadler, E. O. *et al.* Divergences in color perception between deep neural networks and humans. *Cognition* **241**, 105621 (2023).
24. Cohen, U., Chung, S., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. communications* **11**, 746 (2020).
25. Dobs, K., Martinez, J., Kell, A. J. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. advances* **8**, eabl8913 (2022).
26. Mahner, F. P., Muttenthaler, L., Güçlü, U. & Hebart, M. N. Dimensions underlying the representational alignment of deep neural networks with humans. *arXiv preprint arXiv:2406.19087* (2024).

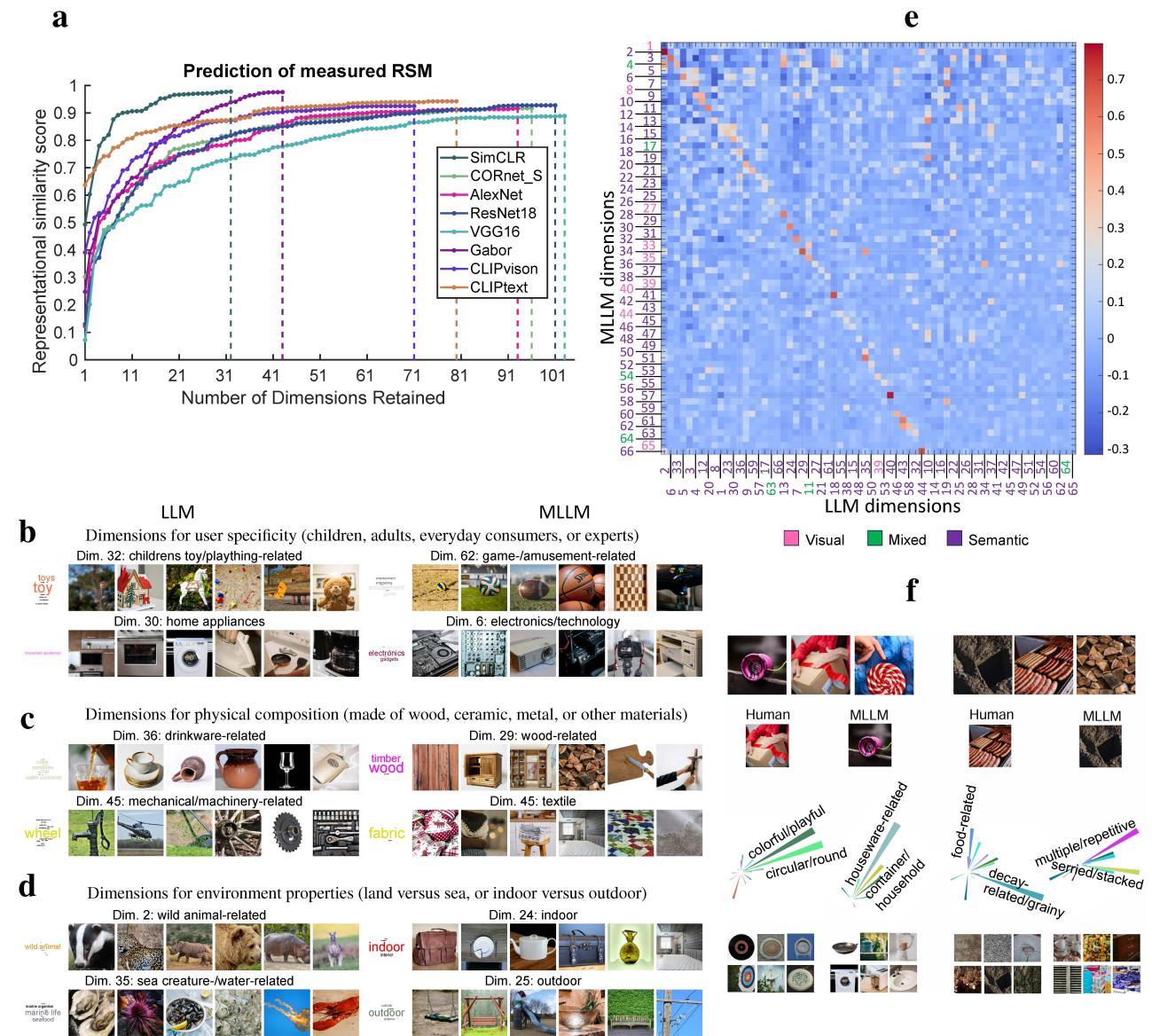
27. Jacob, G., Pramod, R., Katti, H. & Arun, S. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat. communications* **12**, 1872 (2021).
28. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nat. neuroscience* **25**, 369–380 (2022).
29. Muttenthaler, L. & Hebart, M. N. Interpretable object dimensions in deep neural networks and their similarities to human representations. *J. Vis.* **22**, 4516–4516 (2022).
30. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2021).
31. Prince, J. S., Alvarez, G. A. & Konkle, T. Contrastive learning explains the emergence and function of visual category-selective regions. *Sci. Adv.* **10**, eadl1776 (2024).
32. Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. *Nat. communications* **13**, 491 (2022).
33. Zhuang, C. *et al.* Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci.* **118**, e2014196118 (2021).
34. Feather, J., Leclerc, G., Mądry, A. & McDermott, J. H. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nat. Neurosci.* **26**, 2017–2034 (2023).
35. Demszky, D. *et al.* Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701 (2023).
36. Dillon, D., Tandon, N., Gu, Y. & Gray, K. Can AI language models replace human participants? *Trends Cogn. Sci.* (2023).
37. Messeri, L. & Crockett, M. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
38. Josephs, E. L., Hebart, M. N. & Konkle, T. Dimensions underlying human understanding of the reachable world. *Cognition* **234**, 105368 (2023).
39. Zheng, C. Y., Pereira, F., Baker, C. I. & Hebart, M. N. Revealing interpretable object representations from human behavior. In *International Conference on Learning Representations* (2019).
40. Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).
41. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
42. Wei, J. *et al.* Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).
43. Schaeffer, R., Miranda, B. & Koyejo, S. Are emergent abilities of large language models a mirage? *Adv. Neural Inf. Process. Syst.* **36** (2024).
44. Hagendorff, T. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988* (2023).
45. Hagendorff, T., Fabi, S. & Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nat. Comput. Sci.* **3**, 833–838 (2023).
46. Strachan, J. W. *et al.* Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 1–11 (2024).
47. Kumar, S. *et al.* Shared functional specialization in transformer-based language models and the human brain. *Nat. communications* **15**, 5523 (2024).
48. Chen, Y., Liu, T. X., Shan, Y. & Zhong, S. The emergence of economic rationality of gpt. *Proc. Natl. Acad. Sci.* **120**, e2316205120 (2023).
49. Zhang, R. *et al.* Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? (2024). [2403.14624](#).
50. Hebart, M. N. *et al.* Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS one* **14**, e0223792 (2019).
51. Wei, C., Zou, J., Heinke, D. & Liu, Q. CoCoG: Controllable visual stimuli generation based on human concept representations. In *the 33rd International Joint Conference on Artificial Intelligence* (2024).
52. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. neuroscience* **25**, 116–126 (2022).

53. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. systems neuroscience* **2**, 249 (2008).
54. Horikawa, T., Cowen, A. S., Keltner, D., and Kamitani, Y. (2020). The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. *iScience*, 23(5):101060.
55. Wang, P. *et al.* Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
56. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR* (2015).
57. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (2020).
58. Rajalingham, R. *et al.* Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
59. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
60. Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J. & Wehbe, L. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nat. Mach. Intell.* **5**, 1415–1426 (2023).
61. Epstein, R. A. & Baker, C. I. Scene perception in the human brain. *Annu. review vision science* **5**, 373–397 (2019).
62. Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* **293**, 2470–2473 (2001).
63. Sergent, J., Ohta, S. & Macdonald, B. Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain* **115**, 15–36 (1992).
64. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
65. Chang, Y. *et al.* A survey on evaluation of large language models. *ACM Transactions on Intell. Syst. Technol.* **15**, 1–45 (2024).
66. Minaee, S. *et al.* Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).
67. Yin, S. *et al.* A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).
68. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv* 2022–03 (2022).
69. Zador, A. *et al.* Catalyzing next-generation artificial intelligence through neuroAI. *Nat. communications* **14**, 1597 (2023).
70. Thibeault, V., Allard, A. & Desrosiers, P. The low-rank hypothesis of complex systems. *Nat. Phys.* 1–9 (2024).
71. Murphy, K. A. & Bassett, D. S. Information decomposition in complex systems via machine learning. *Proc. Natl. Acad. Sci.* **121**, e2312988121 (2024).
72. Doerig, A. *et al.* Semantic scene descriptions as an objective of human vision (arxiv: 2209.11737). arxiv (2022).
73. Conwell, C., Prince, J., Alvarez, G. & Konkle, T. The unreasonable effectiveness of word models in predicting high-level visual cortex responses to natural images. In *Conference on Computational Cognitive Neuroscience 2023*.
74. McMahon, E., Conwell, C., Garcia, K., Bonner, M. F. & Isik, L. Language model prediction of visual cortex responses to dynamic social scenes. *J. Vis.* **24**, 904–904 (2024).
75. Conwell, C. *et al.* Monkey see, model knew: Large language models accurately predict human and macaque visual brain activity. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models 2024*.
76. Tuckute, G., Kanwisher, N. & Fedorenko, E. Language in brains, minds, and machines. *Annu. Rev. Neurosci.* **47** (2024).
77. Tuckute, G. *et al.* Driving and suppressing the human language network using large language models. *Nat. Hum. Behav.* **8**, 544–561 (2024).
78. Popham, S. F. *et al.* Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat. neuroscience* **24**, 1628–1636 (2021).
79. Roads, B. D. & Love, B. C. Learning as the unsupervised alignment of conceptual systems. *Nat. Mach. Intell.* **2**, 76–82 (2020).

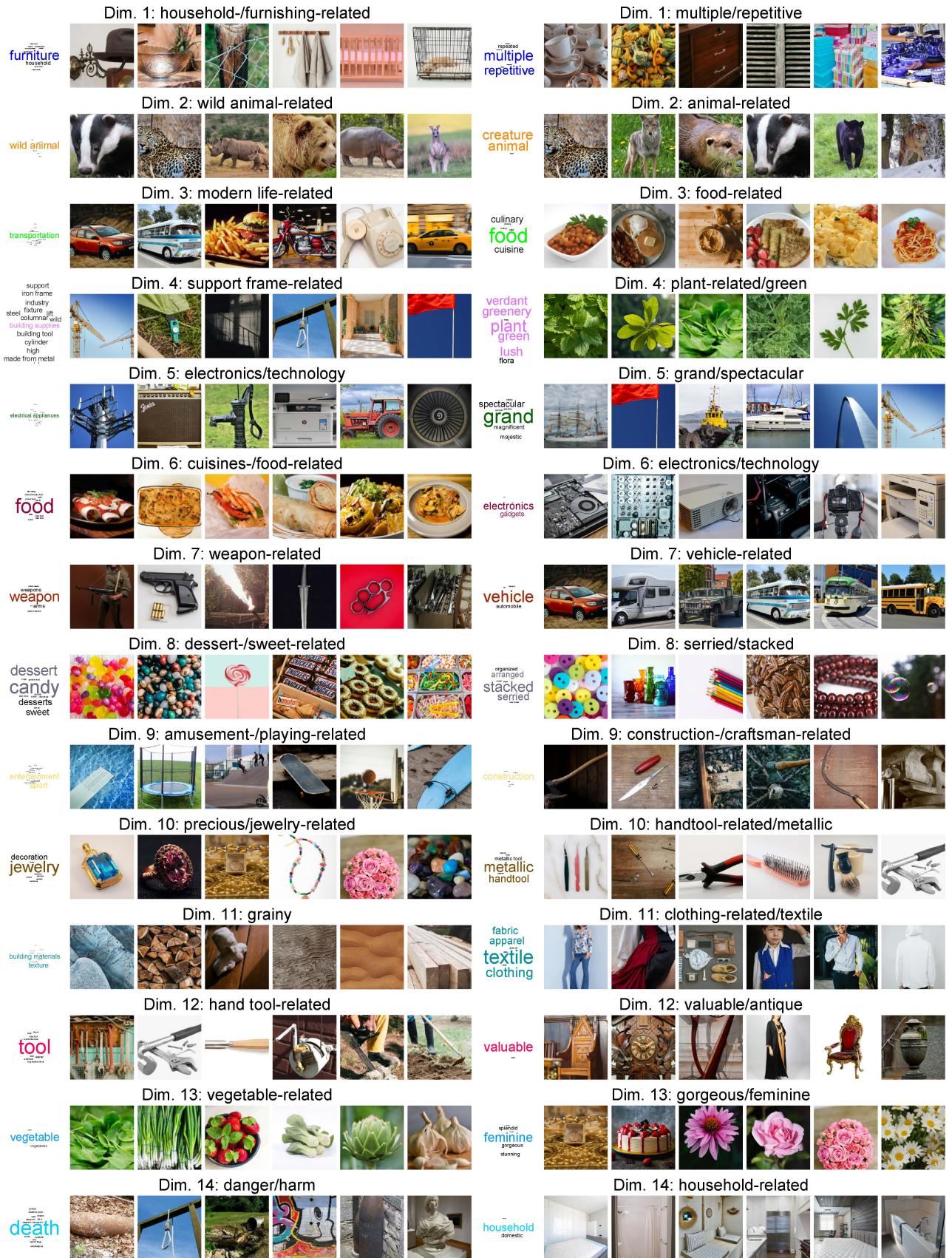
- 80.** Sereno, M. I. *et al.* Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* **268**, 889–893 (1995).
- 81.** Engel, S. A., Glover, G. H. & Wandell, B. A. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. cortex (New York, NY: 1991)* **7**, 181–192 (1997).
- 82.** Hansen, K. A., Kay, K. N. & Gallant, J. L. Topographic organization in and near human visual area V4. *J. Neurosci.* **27**, 11896–11911 (2007).
- 83.** Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
- 84.** Harvey, B. M., Klein, B. P., Petridou, N. & Dumoulin, S. O. Topographic representation of numerosity in the human parietal cortex. *Science* **341**, 1123–1126 (2013).
- 85.** Sha, L. *et al.* The animacy continuum in the human ventral vision pathway. *J. cognitive neuroscience* **27**, 665–678 (2015).
- 86.** Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
- 87.** Margulies, D. S. *et al.* Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci.* **113**, 12574–12579 (2016).
- 88.** Huntenburg, J. M., Bazin, P.-L. & Margulies, D. S. Large-scale gradients in human cortical organization. *Trends cognitive sciences* **22**, 21–31 (2018).
- 89.** Bau, D. *et al.* Understanding the role of individual units in a deep neural network. *Proc. Natl. Acad. Sci.* **117**, 30071–30078 (2020).
- 90.** McGrath, T. *et al.* Acquisition of chess knowledge in alphazero. *Proc. Natl. Acad. Sci.* **119**, e2206625119 (2022).
- 91.** Achitbat, R. *et al.* From attribution maps to human-understandable explanations through concept relevance propagation. *Nat. Mach. Intell.* **5**, 1006–1019 (2023).
- 92.** Bills, S. *et al.* Language models can explain neurons in language models. URL <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>. (Date accessed: 14.05. 2023) (2023).
- 93.** Sanborn, A. N., Griffiths, T. L. & Shiffrin, R. M. Uncovering mental representations with markov chain monte carlo. *Cogn. psychology* **60**, 63–106 (2010).
- 94.** Mahowald, K. *et al.* Dissociating language and thought in large language models. *Trends Cogn. Sci.* (2024).
- 95.** Qu, Y. *et al.* Integration of cognitive tasks into artificial general intelligence test for large models. *Iscience* **27** (2024).
- 96.** Meng, J. AI emerges as the frontier in behavioral science. *Proc. Natl. Acad. Sci.* **121**, e2401336121 (2024).
- 97.** Marjeh, R., Sucholutsky, I., van Rijn, P., Jacoby, N. & Griffiths, T. What language reveals about perception: Distilling psychophysical knowledge from large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45 (2023).
- 98.** Campbell, D., Kumar, S., Giallanza, T., Griffiths, T. L. & Cohen, J. D. Human-like geometric abstraction in large pre-trained neural networks. *arXiv preprint arXiv:2402.04203* (2024).
- 99.** Kawakita, G., Zelezniak-Johnston, A., Tsuchiya, N. & Oizumi, M. Comparing color similarity structures between humans and llms via unsupervised alignment. *arXiv preprint arXiv:2308.04381* (2023).
- 100.** Li, C. *et al.* Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760* (2023).
- 101.** Sabour, S. *et al.* EmoBench: Evaluating the emotional intelligence of large language models. In *the 62nd Annual Meeting of the Association for Computational Linguistics* (2024).
- 102.** Janik, R. A. Aspects of human memory and large language models. *arXiv preprint arXiv:2311.03839* (2023).
- 103.** Huff, M. & Ulakçı, E. Towards a psychology of machines: Large language models predict human memory. *arXiv preprint arXiv:2403.05152* (2024).
- 104.** Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).
- 105.** Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D. & Griffiths, T. L. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).

106. Alsagheer, D. *et al.* Comparing rationality between large language models and humans: Insights and open questions. *arXiv preprint arXiv:2403.09798* (2024).
107. Achiam, J. *et al.* GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
108. St-Yves, G., Allen, E. J., Wu, Y., Kay, K. & Naselaris, T. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nat. communications* **14**, 3329 (2023).
109. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *13th European Conference on Computer Vision*, 740–755 (Springer, 2014).
110. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
111. Hebart, M. N., Kaniuth, P. & Perkühn, J. Efficiently-generated object similarity scores predicted from human feature ratings and deep neural network activations. *J. Vis.* **22**, 4057–4057 (2022).
112. Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A. & Kornblith, S. Human alignment of neural network representations. In *Proc. of the 11th International Conference on Learning Representations* (2022).
113. Fischl, B. Freesurfer. *Neuroimage* **62**, 774–781 (2012).
114. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. neuroinformatics* **23** (2015).
115. Du, C. & CDDU. ChangdeDu/LLMs_core_dimensions. *Zenodo*, <https://zenodo.org/record/15090332> (2025).

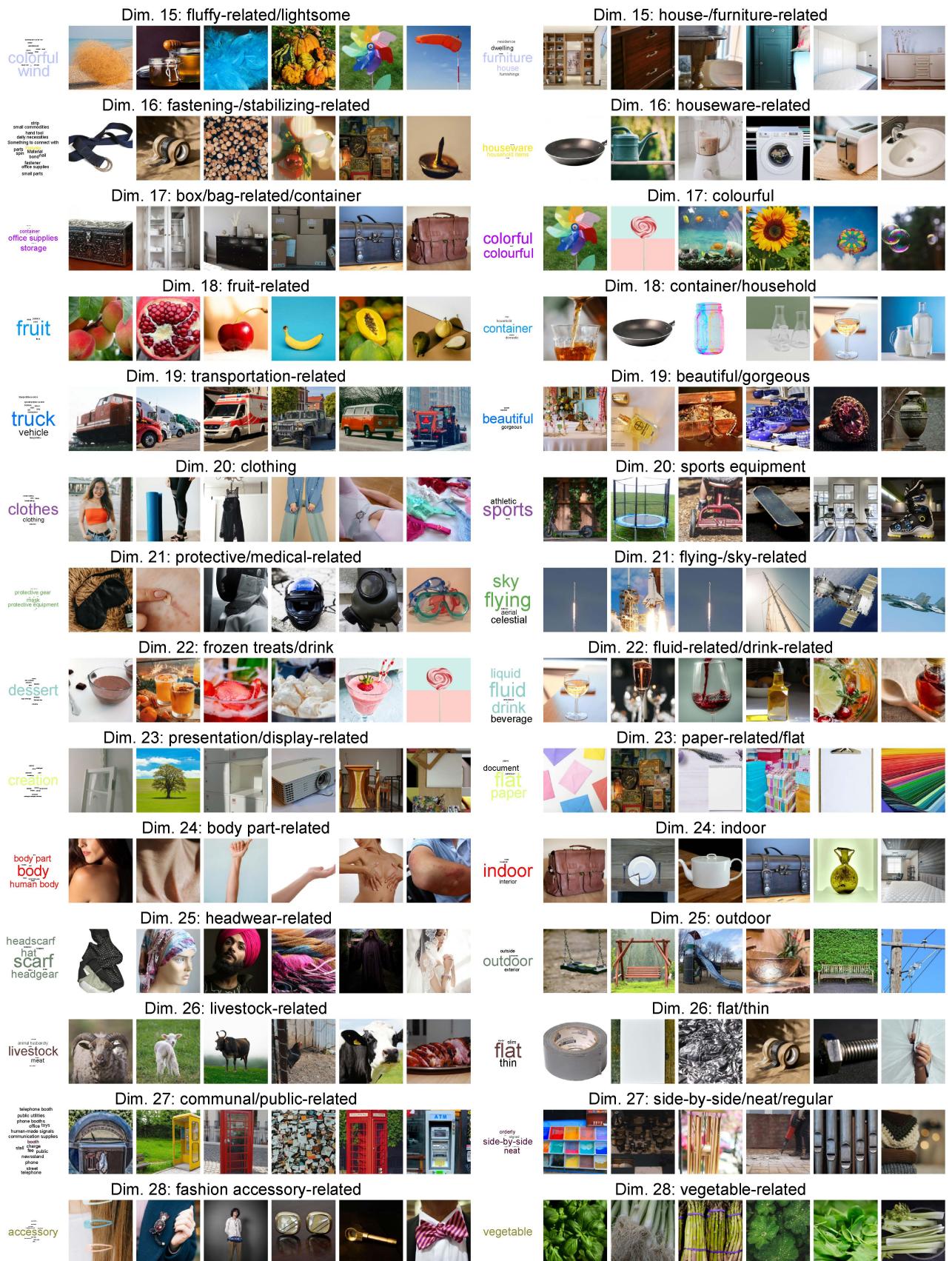
Extended data



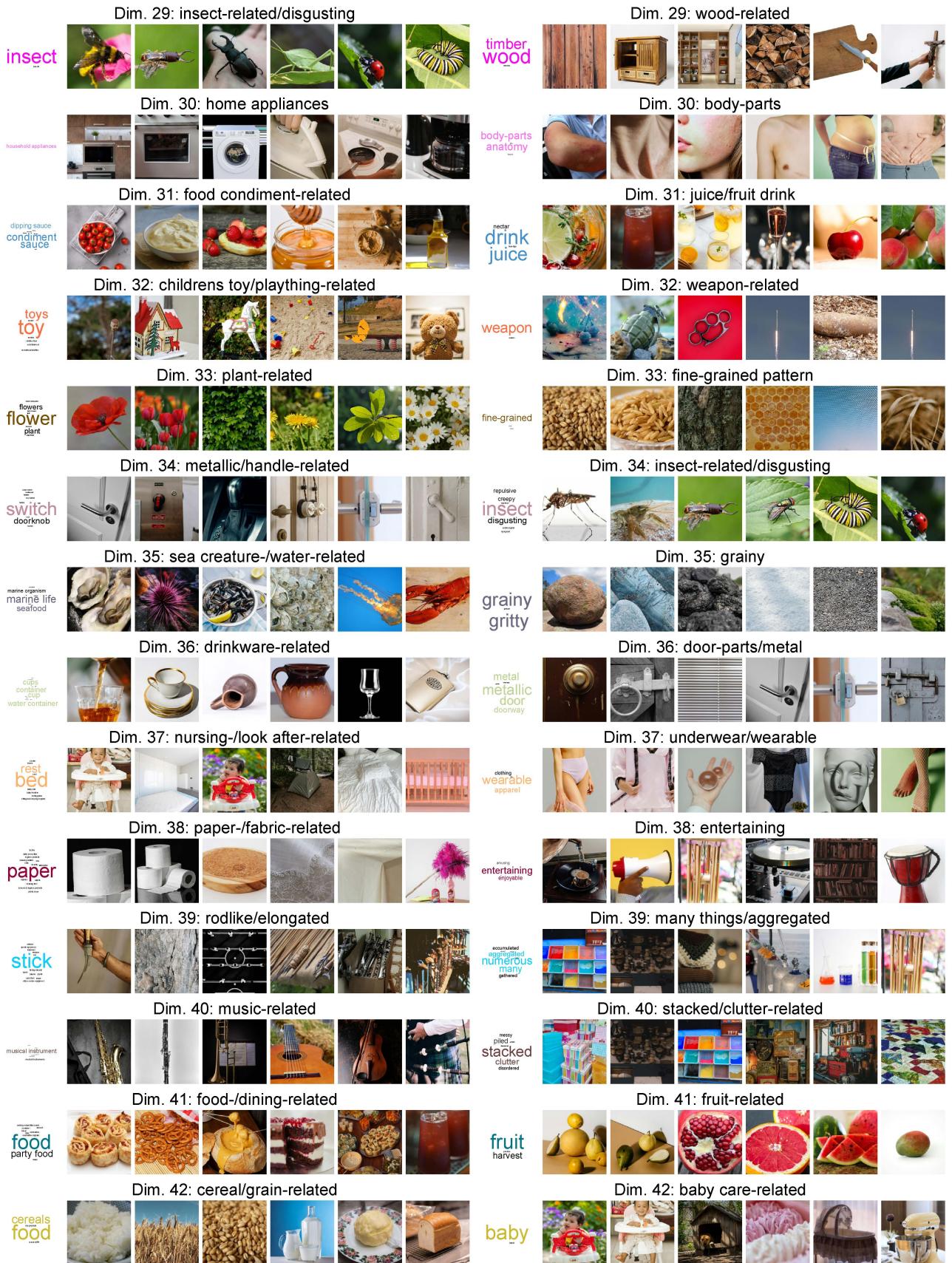
Extended Data Fig. 1. Object dimensions learned by different models and their interpretations (related to Figs. 2, 4 and 5). **a.**, Dimensions retained by different models and the ability to predict their behavioral RSMs. **b-d.**, Object dimensions illustrating their interpretability for LLM and MLLM. **e.**, Cross-correlation matrix between LLM and MLLM. **f.**, Key dimensions that underpin the different choices that humans and models made.



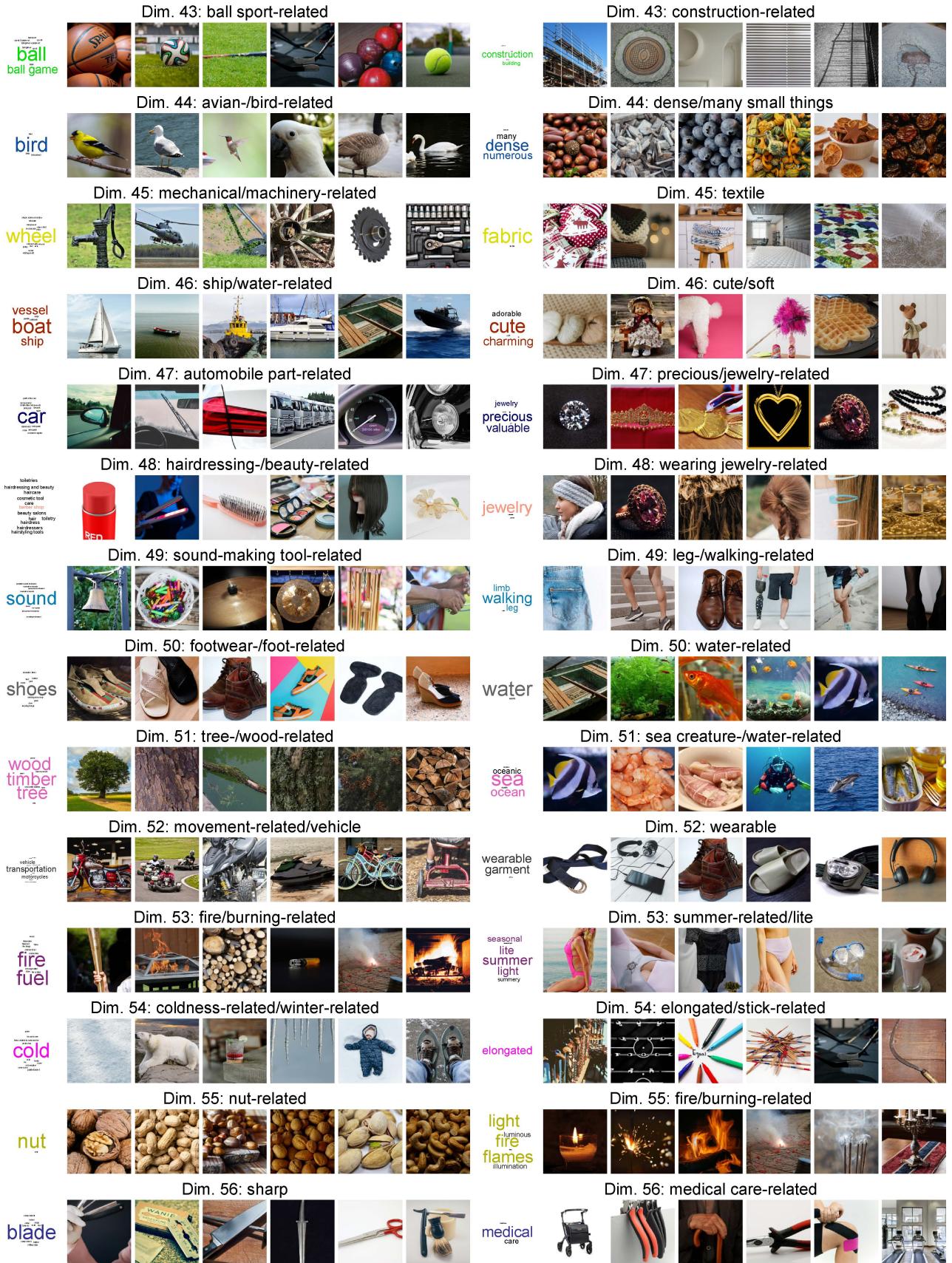
Extended Data Fig. 2. Object dimensions (1-14) illustrating their interpretability for LLM (left) and MLLM (right)(related to Fig. 4). Each dimension is illustrated with the top 6 images with the highest weights along this dimension.



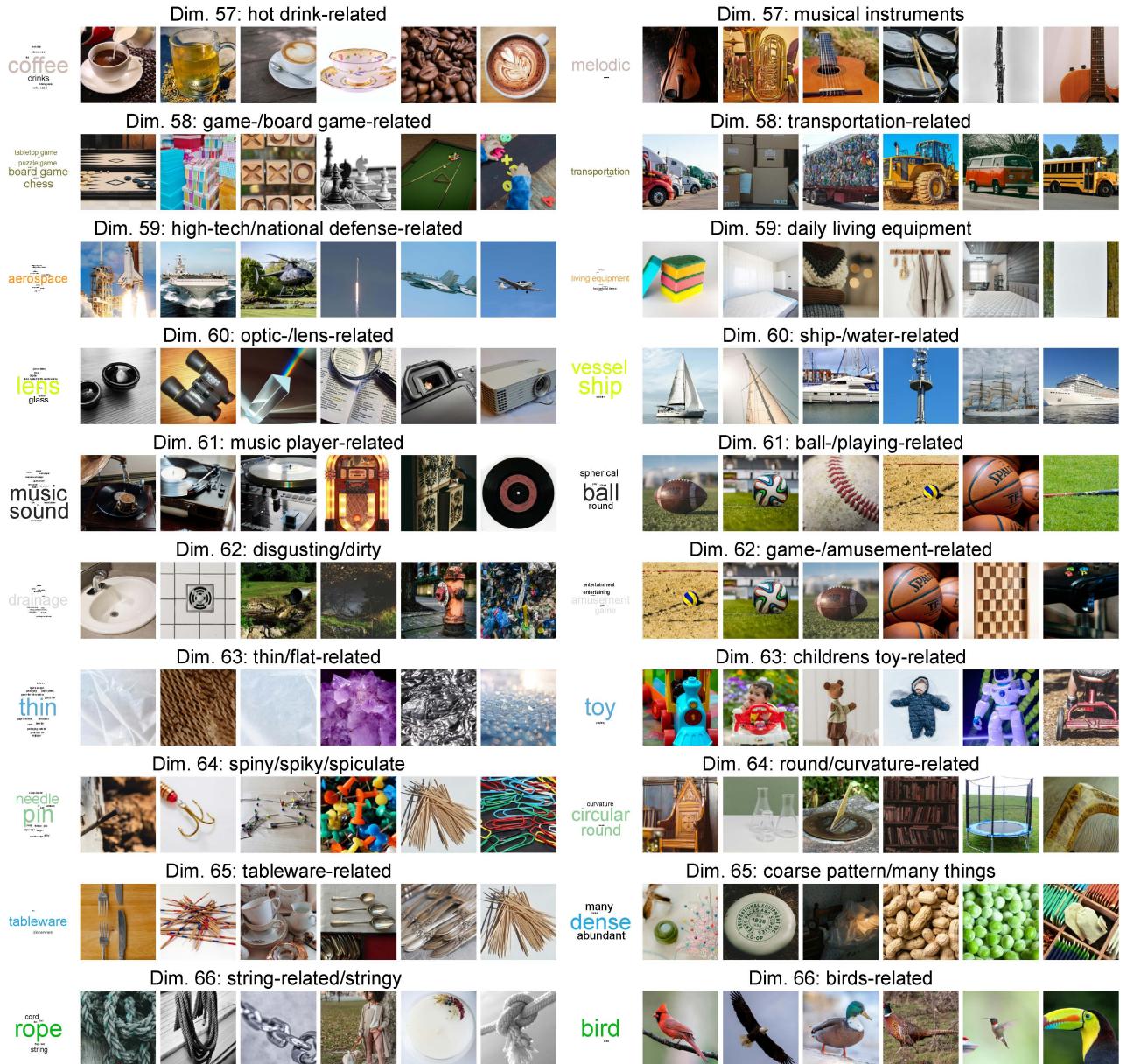
Extended Data Fig. 3. Object dimensions (15-28) illustrating their interpretability for LLM (left) and MLLM (right)(related to Fig. 4). Each dimension is illustrated with the top 6 images with the highest weights along this dimension.



Extended Data Fig. 4. Object dimensions (29-42) illustrating their interpretability for LLM (left) and MLLM (right)(related to Fig. 4). Each dimension is illustrated with the top 6 images with the highest weights along this dimension.



Extended Data Fig. 5. Object dimensions (43–56) illustrating their interpretability for LLM (left) and MLLM (right)(related to Fig. 4). Each dimension is illustrated with the top 6 images with the highest weights along this dimension.



Extended Data Fig. 6. Object dimensions (57-66) illustrating their interpretability for LLM (left) and MLLM (right)(related to Fig. 4). Each dimension is illustrated with the top 6 images with the highest weights along this dimension.

Extended Data Table 1. List of all dimensions and their intuitive labels summed up by the human experts (related to Fig. 4).

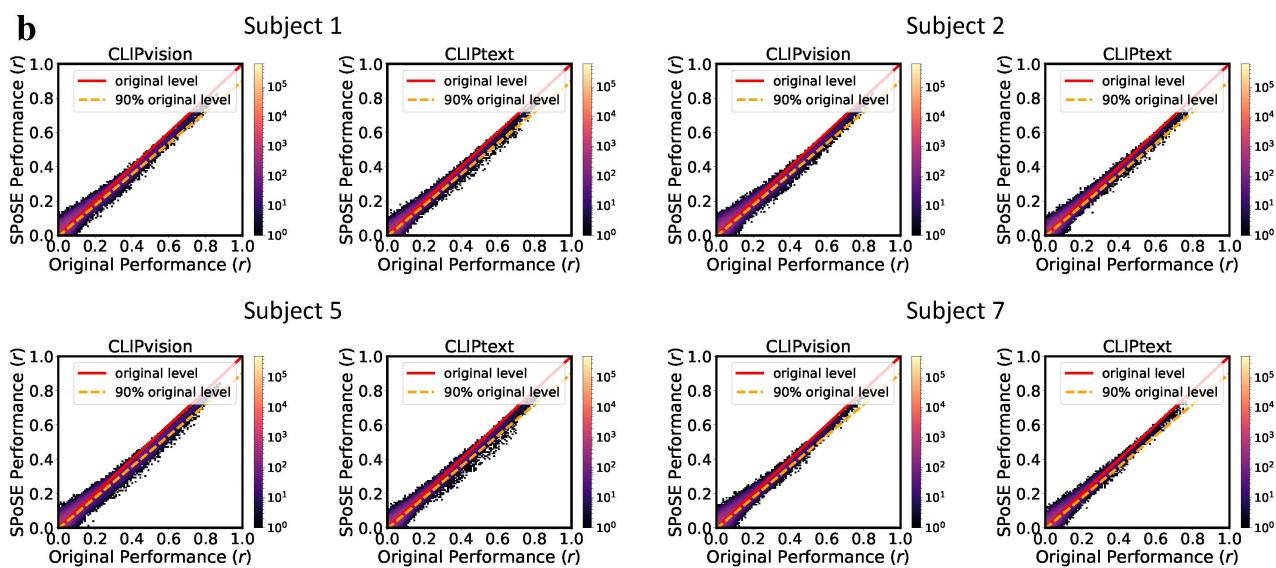
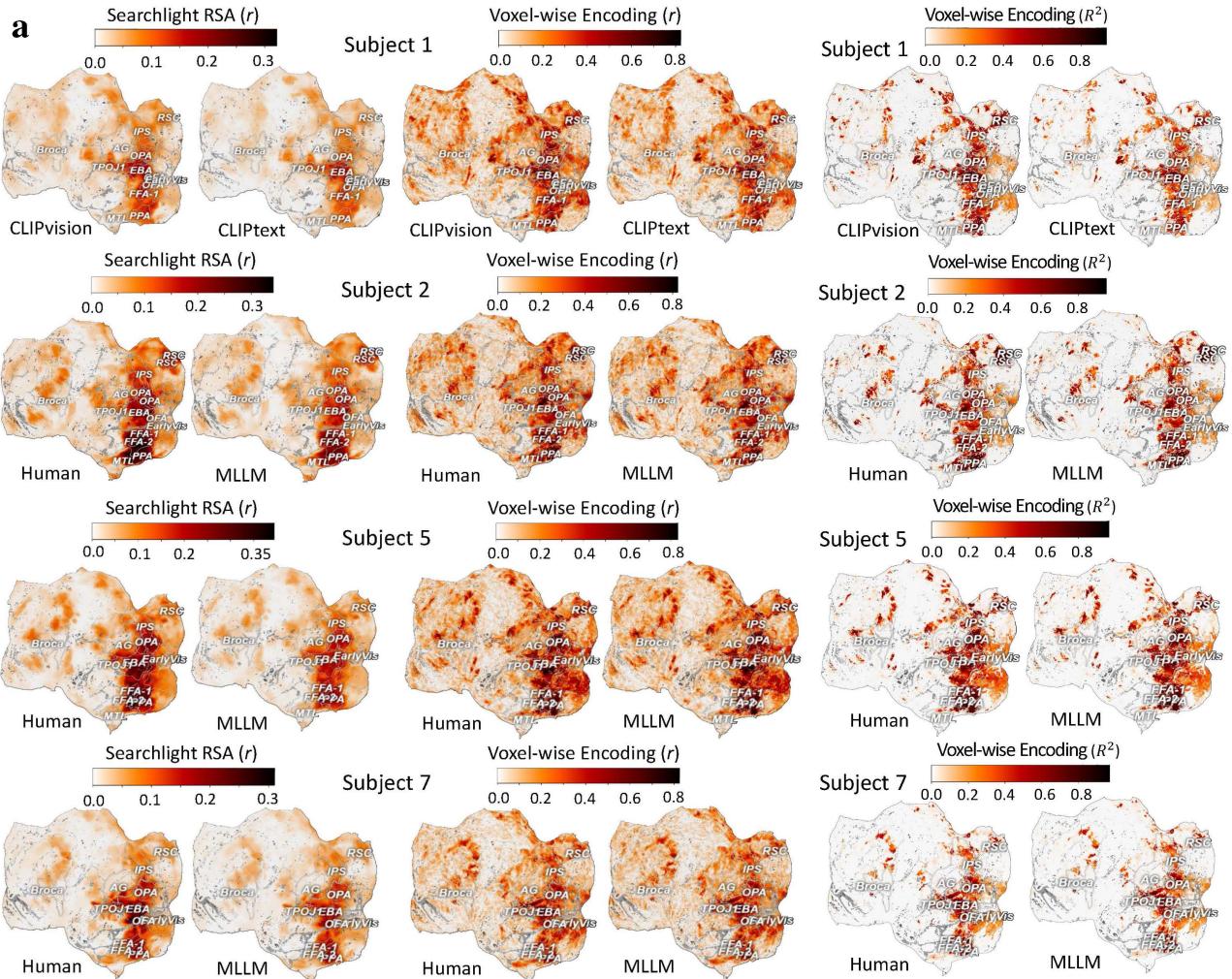
Dim. No.	LLM (GPT3.5-Turbo)	MLLM (Gemini Pro Vision 1.0)	Humans
1	household-/furnishing-related	multiple/repetitive	metallic/artificial
2	wild animal-related	animal-related	food-related
3	modern life-related	food-related	animal-related
4	support frame-related	plant-related/green	textile
5	electronics/technology	grand/spectacular	plant-related
6	cuisines-/food-related	electronics/technology	house-related/furnishing-related
7	weapon-related	vehicle-related	valuable/precious
8	dessert/sweet-related	serried/stacked	transportation-/movement-related
9	amusement-/playing-related	construction-/craftsman-related	body-/people-related
10	precious/jewelry-related	handtool-related/metallic	wood-related/brown
11	grainy	clothing-related/textile	electronics/technology
12	hand tool-related	valuable/antique	colorful/playful
13	vegetable-related	gorgeous/feminine	outdoors
14	danger/harm	household-related	circular/round
15	fluffy-related/lightsome	house-/furniture-related	paper-related/flat
16	fastening-/stabilizing-related	houseware-related	sports-/playing-related
17	box/bag-related/container	colourful	tools/elongated
18	fruit-related	container/household	fluid-related/drink-related
19	transportation-related	beautiful/gorgeous	water-related
20	clothing	sports equipment	oriented/many things
21	protective/medical-related	flying-/sky-related	decay-related/grainy
22	frozen treats/drink	fluid-related/drink-related	white
23	presentation/display-related	paper-related/flat	coarse pattern/many things
24	body part-related	indoor	red
25	headwear-related	outdoor	long/thin
26	livestock-related	flat/thin	weapon-/danger-related
27	communal/public-related	side-by-side/neat/regular	black
28	fashion accessory-related	vegetable-related	household
29	insect-related/disgusting	wood-related	feminine (stereotypical)
30	home appliances	body-parts	body part-related
31	food condiment-related	juice/fruit drink	tubular
32	childrens toy/plaything-related	weapon-related	music-/hearing-/hobby-related
33	plant-related	fine-grained pattern	grid-/grating-related
34	metallic/handle-related	insect-related/disgusting	repetitive/spiky
35	sea creature-/water-related	grainy	construction-/craftsman-related
36	drinkware-related	door-parts/metal	spherical/voluminous
37	nursing-/look after-related	underwear/wearable	string-related/stringy
38	paper-/fabric-related	entertaining	seating-/standing-/lying-related
39	rodlike/elongated	many things/aggregated	flying-/sky-related
40	music-related	stacked/clutter-related	disgusting/slimy
41	food-/dining-related	fruit-related	elliptical/curved
42	cereal/grain-related	baby care-related	sand-colored
43	ball sport-related	construction-related	green
44	avian-/bird-related	dense/many small things	bathroom-/wetness-related
45	mechanical/machinery-related	textile	yellow
46	ship/water-related	cute/soft	heat-/light-related
47	automobile part-related	precious/jewelry-related	beams-/mesh-related
48	hairdressing-/beauty-related	wearing jewelry-related	foot-/walking-related
49	sound-making tool-related	leg-/walking-related	box-related/container
50	footwear/foot-related	water-related	stick-shaped/cylindrical
51	tree-/wood-related	sea creature-/water-related	head-related
52	movement-related/vehicle	wearable	upright/elongated/volumous
53	fire/burning-related	summer-related/lite	pointed/spiky
54	coldness-related/winter-related	elongated/stick-related	child-related/cute
55	nut-related	fire/burning-related	farm-related/historical
56	sharp	medical care-related	seeing-related/small/round
57	hot drink-related	musical instruments	medicine-related
58	game-/board game-related	transportation-related	dessert-related
59	high-tech/national defense-related	daily living equipment	orange
60	optic-/lens-related	ship-/water-related	thin/flat
61	music player-related	ball-/playing-related	cylindrical/conical/cushioning
62	disgusting/dirty	game-/amusement-related	coldness-related/winter-related
63	thin/flat-related	childrens toy-related	measurement-related/numbers-related
64	spiny/spiky/spiculate	round/curvature-related	fluffy/soft
65	tableware-related	coarse pattern/many things	masculine (stereotypical)
66	string-related/stringy	birds-related	fine-grained pattern

Extended Data Table 2. Dimension labels summed up by the human experts and the MLLM (here, gemini-pro-1.5-exp, related to Fig. 4). MLLM matches human annotation highly consistently marked with **✓✓**, consistent with **✓**, and inconsistent with **✗**. While MLLM excels at concrete comparative tasks (like triplet odd-one-out selection), it shows limitations in dimension naming tasks that require abstracting and generalizing across diverse visual and semantic features.

Dimension labels of LLM (GPT3.5-Turbo)		Dimension labels of MLLM (Gemini Pro Vision 1.0)		
Dim.	Annotated by human experts	Annotated by MLLM	Annotated by human experts	Annotated by MLLM
1	household-/furnishing-related	household items/home furnishings ✓✓	multiple/repetitive	secondhand goods/flea market ✗
2	wild animal-related	wild animals/animals in the wild ✓✓	animal-related	wild animals ✓
3	modern life-related	modes of transportation ✗	food-related	breakfast foods/brunch dishes ✓
4	support frame-related	simple machines/mechanical advantage ✗	plant-related/green	green plants ✓✓
5	electronics/technology	old technology/obsolete technology ✓	grand/spectacular	different watercrafts/vessels ✗
6	cuisines-/food-related	dishes/food ✓✓	electronics/technology	electronic devices/obsolete technology ✓✓
7	weapon-related	weapons/weaponry ✓✓	vehicle-related	modes of transportation/vehicles ✓✓
8	dessert-/sweet-related	sweets/candy ✓✓	serried/stacked	round objects/circular shapes ✗
9	amusement-/playing-related	recreational activities/outdoor fun ✓	construction-/craftsman-related	hand tools/tools ✓
10	precious/jewelry-related	jewelry & gems ✓✓	handtool-related/metallic	household tools ✓
11	grainy	raw materials ✗	clothing-related/textile	clothing, apparel ✓✓
12	hand tool-related	tools/hand tools ✓✓	valuable/antique	antique/vintage ✓
13	vegetable-related	vegetables/produce ✓✓	gorgeous/feminine	gifts/presents ✗
14	danger/harm	death/suffering ✗	household-related	furniture/home furnishings ✓
15	fluffy-related/lightsome	fall/autumn ✗	house-/furniture-related	home furniture ✓✓
16	fastening-/stabilizing-related	craft supplies/crafting materials ✗	houseware-related	household appliances/items ✓✓
17	box/bag-related/container	storage/containers ✓✓	colourful	bright colors ✓
18	fruit-related	fruits/fruit varieties ✓✓	container/household	glass containers/containers for liquids ✓
19	transportation-related	modes of transportation/vehicles ✓✓	beautiful/gorgeous	luxury/wealth ✗
20	clothing	women's clothing ✓	sports equipment	children's toys/recreational equipment ✗
21	protective/medical-related	safety equipment/protective gear ✓✓	flying-/sky-related	air & space/flight & aerospace ✓
22	frozen treats/drink	desserts/sweet treats ✓	fluid-related/drink-related	beverages/drinks ✓✓
23	presentation/display-related	home improvement ✗	paper-related/flat	office supplies / stationery ✓
24	body part-related	body parts ✓✓	indoor	household items/home goods ✗
25	headwear-related	head coverings/headwear ✓✓	outdoor	outdoor scenes ✓✓
26	livestock-related	farm animals/livestock ✓✓	flat/thin	common materials/everyday items ✗
27	communal/public-related	street furniture/public amenities ✓	side-by-side/neat/regular	storage/supplies/inventory/stock ✗
28	fashion accessory-related	fashion accessories ✓✓	vegetable-related	green vegetables/leafy greens ✓✓
29	insect-related/disgusting	insects/bugs ✓✓	wood-related	wood products/wooden objects ✓✓
30	home appliances	home appliances ✓✓	body-parts	human body parts ✓✓
31	food condiment-related	food spreads/condiments ✓✓	juice/fruit drink	food and drinks ✓
32	childrens toy/plaything-related	children's toys/playthings ✓✓	weapon-related	weapons/explosives ✓✓
33	plant-related	flowers/plants ✓✓	fine-grained pattern	raw materials/natural resources ✗
34	metallic/handle-related	door hardware/door parts ✓	insect-related/disgusting	insects and invertebrates ✓✓
35	sea creature-/water-related	seafood/sea creatures ✓	grainy	natural materials/raw materials ✗
36	drinkware-related	household items/containers. ✓	door-parts/metal	door hardware/door parts ✓✓
37	nursing-/look after-related	furniture/household items ✗	underwear/wearable	women's apparel/lingerie ✓
38	paper-/fabric-related	household items ✗	entertaining	analog technology/vintage items ✗
39	rodlike/elongated	long/thin/rod-like objects ✓✓	many things/aggregated	industrial supplies ✗
40	music-related	musical instruments ✓✓	stacked/clutter-related	garage sale/flea market ✗
41	food-/dining-related	party food/snacks ✓	fruit-related	fruits/fruit variety ✓✓
42	cereal/grain-related	wheat products/foods made from wheat ✓✓	baby care-related	baby items/baby products ✓✓
43	ball sport-related	sports equipment ✓✓	construction-related	urban infrastructure/city elements ✗
44	avian-/bird-related	birds/bird species ✓✓	dense/many small things	natural textures/organic materials ✗
45	mechanical/machinery-related	simple machines/mechanical parts ✓✓	textile	household linens/fabric goods ✓
46	ship/water-related	boats/watercraft ✓✓	cute/soft	handmade/crafts ✗
47	automobile part-related	car parts ✓✓	precious/jewelry-related	valuable possessions/precious items ✓
48	hairdressing-/beauty-related	hair styling/hair care ✓✓	wearing jewelry-related	hair accessories ✗
49	sound-making tool-related	musical instruments/sound makers ✓✓	leg-/walking-related	human legs ✓✓
50	footwear-/foot-related	footwear/shoes ✓✓	water-related	water activities/aquatic life ✓
51	tree-/wood-related	trees/wood ✓✓	sea creature-/water-related	fishing/seafood ✗
52	movement-related/vehicle	recreational vehicles/personal vehicles ✓	wearable	personal accessories/personal items ✓
53	fire/burning-related	fire/sources of fire ✓✓	summer-related/lite	beach vacation/swimming gear ✗
54	coldness-related/winter-related	winter/cold weather ✓✓	elongated/stick-related	handmade crafts/diy projects ✗
55	nut-related	nuts/edible nuts ✓✓	fire/burning-related	fire/combustion ✓✓
56	sharp	sharp objects/blades ✓✓	medical care-related	assistive devices/mobility aids ✓
57	hot drink-related	coffee & tea ✓✓	musical instruments	musical instruments ✓✓
58	game-/board game-related	board games/indoor games ✓✓	transportation-related	vehicles/motor vehicles ✓✓
59	high-tech/national defense-related	vehicles/transportation ✗	daily living equipment	hotel linens/hotel supplies ✗
60	optic-/lens-related	optical lenses/optics ✓✓	ship-/water-related	sea vessels/watercraft ✓✓
61	music player-related	music players/audio devices ✓✓	ball-/playing-related	sports equipment ✓
62	disgusting/dirty	water infrastructure/urban utilities ✗	game-/amusement-related	sports/games ✓
63	thin/flat-related	shiny materials/reflective surfaces ✗	childrens toy-related	children's toys ✓✓
64	spiny/spiky/spiculate	sharp objects/pointy things ✓✓	round/curvature-related	obsolete technology ✗
65	tableware-related	kitchen utensils ✓✓	coarse pattern/many things	arts and crafts ✗
66	string-related/stringy	knots and cords ✓	birds-related	birds/bird species ✓✓

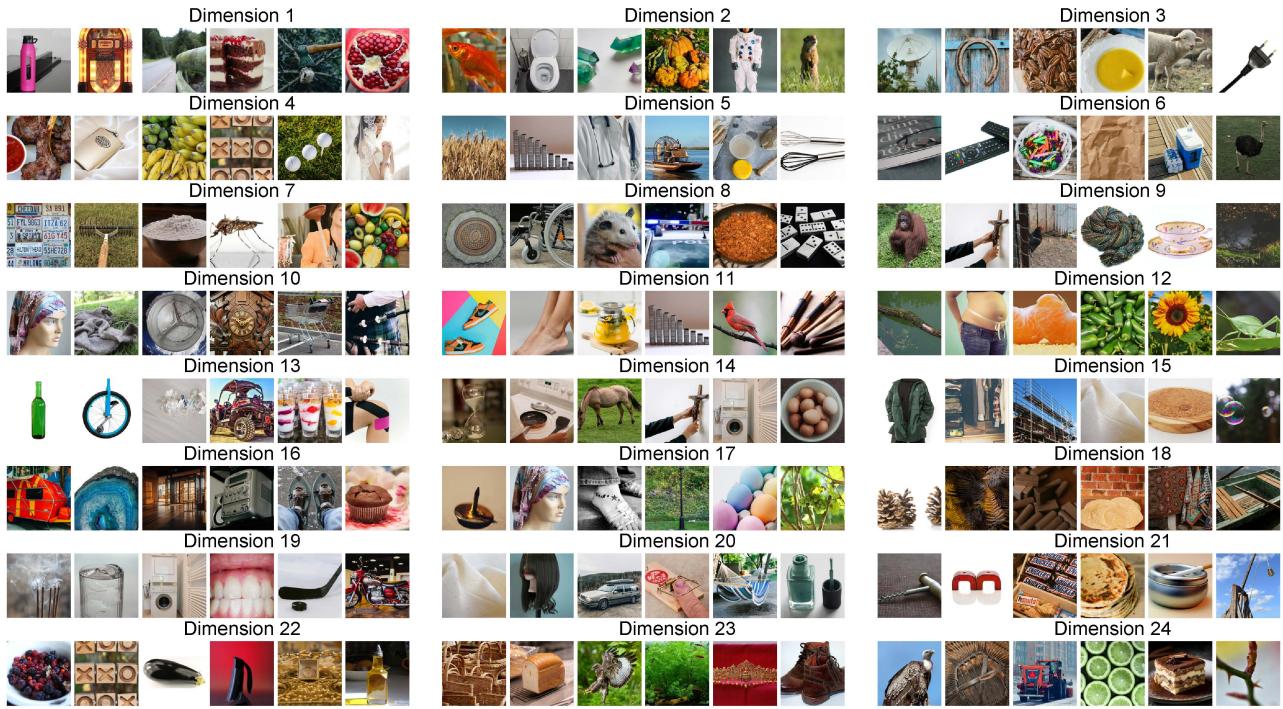


Extended Data Fig. 7. Object dimensions (1-32) illustrating their interpretability for self-supervised learning model SimCLR (related to Fig. 4). **a**, Each dimension is illustrated with the top 6 images with the highest weights along this dimension. **b**, Dimensions retained by SimCLR and the ability to predict its behavioral RSMs. **c**, Attribution of the 32 dimensions of the SimCLR model, where the visual dimensions occupy the vast majority, and only a few semantic dimensions.

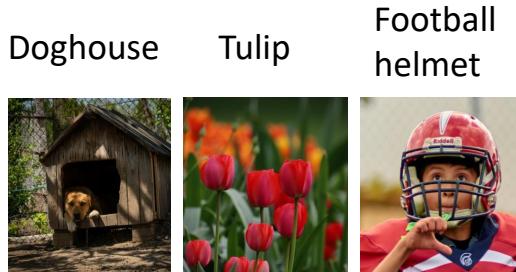


Extended Data Fig. 8. More results on the relationship between model and brain representations (related to Fig. 6). **a**, Flattened cortical maps for more models and subjects. Performance was evaluated by using both Pearson’s correlation (r) and the noise-normalized R^2 . **b**, Voxel-wise encoding performance using the original high-dimensional model features and the low-dimensional SPOSE embeddings of CLIP model.

Supplementary information



Supplementary Fig. 1. Top 24 dimensions for "random representation" model (related to Fig. 4). We constructed representations of the 1,854 object concepts using 1,000-dimensional random vectors, generated 4.7 million odd-one-out data points based on cosine distances, and then applied the SPoSE method to learn low-dimensional embeddings. Each dimension was illustrated with the top 6 images with the highest weights along this dimension. These dimensions exhibit no interpretability whatsoever. This strongly suggests that the interpretability of the dimensions obtained from LLM/MLLM is primarily attributable to the models' representations rather than the SPoSE method itself. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.



Prompt: Given a triplet of objects {'doghouse (outbuilding that serves as a shelter for a dog)', 'tulip (any of numerous perennial bulbous herbs having linear or broadly lanceolate leaves and usually a single showy flower)', 'football helmet (a padded helmet with a face mask to protect the head of football players)'}, which one in the triplet is the odd-one-out? Please give the answer first and then explain in detail.

LLM (GPT-3.5-Turbo) :
'Tulip' is the odd-one-out in the triplet.

Explanation:
In this triplet of objects:

1. Doghouse: Represents a structure that serves as a shelter for a dog.
2. Tulip: Represents a type of flower, a perennial bulbous herb.
3. Football helmet: Represents protective headgear worn by football players.

The odd-one-out in this triplet is 'Tulip' because it is different from the other two objects in terms of its category and function:

- Doghouse and Football helmet are both related to protection and serve functional purposes. They are items designed for specific uses: sheltering dogs and protecting football players' heads.
- Tulip, on the other hand, is a flower and does not serve a functional or protective purpose like the other two objects. It is a natural object associated with aesthetics and decoration rather than utility or protection.

Therefore, 'Tulip' stands out as the odd-one-out in this triplet of objects due to its nature as a decorative flower compared to the functional and protective roles of 'Doghouse' and 'Football helmet.'

Guiding model attention to the dimension of "red"



Prompt: Given a triplet of objects {'doghouse (outbuilding that serves as a shelter for a dog)', 'tulip (any of numerous perennial bulbous herbs having linear or broadly lanceolate leaves and usually a single showy flower)', 'football helmet (a padded helmet with a face mask to protect the head of football players)'} and consider the aspect of "red" or "color" as the main focus, which one in the triplet is the odd-one-out? Please give the answer first and then explain in detail.

LLM (GPT-3.5-Turbo) :
The odd-one-out is the doghouse.



Explanation:
When considering the aspect of "red" or "color," the distinction lies in the typical association of these objects with color:

1. Tulip: Tulips are inherently associated with a wide range of colors, including red. They are commonly admired and categorized based on their colors, with red being a very prominent and well-known variation.
2. Football helmet: Football helmets often come in various colors, including red, as they are typically designed to match the team's colors or branding. Red is a common and deliberate design choice in this context.
3. Doghouse: While doghouses can technically be painted any color, they are not inherently associated with color in the way tulips or football helmets are. Doghouses are usually functional structures, and their color is not as significant or defining compared to the other two objects. Red, or any other color, is not a characteristic typically emphasized in relation to doghouses.

Thus, the doghouse stands out as the odd-one-out in terms of its weaker connection to the concept of "red" or "color."

Supplementary Fig. 2. Guiding LLM's attention to the target dimension by using tailored prompts (related to Fig. 5).
We added the phrase "consider the aspect of "red" or "color" as the main focus" to the prompt of LLM. As can be seen, when the prompt included guidance on the dimensions prioritized by humans ("red"), the LLM was able to make choice consistent with human judgment. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.



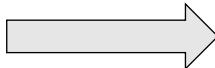
Prompt: You are shown three object images side by side and are asked to report the image that was the least similar to the other two. You should focus your judgement on the object, but you are not given additional constraints as to the strategy you should use. If you did not recognize the object, you should base your judgement on your best guess of what the object could be. 1. Tell me your answer. 2. Tell me the location of the object you have chosen. 3. Explain the reasons.

MLLM (Gemini Pro Vision 1.0):

1. The image that is least similar to the other two is the downspout.
2. The downspout image is in the center of the three images.
3. The first image shows a wooden trough filled with water, exhibiting various shades of brown, green, and dark gray. The third image displays a pile of lime wedges, predominantly bright green. Both images feature natural, organic colors. In contrast, the middle image shows a light pinkish-beige wall with a light beige downspout. These colors are significantly different from the natural, cool tones of the other two images.



Guiding model
attention to the
dimension of
"artificial"



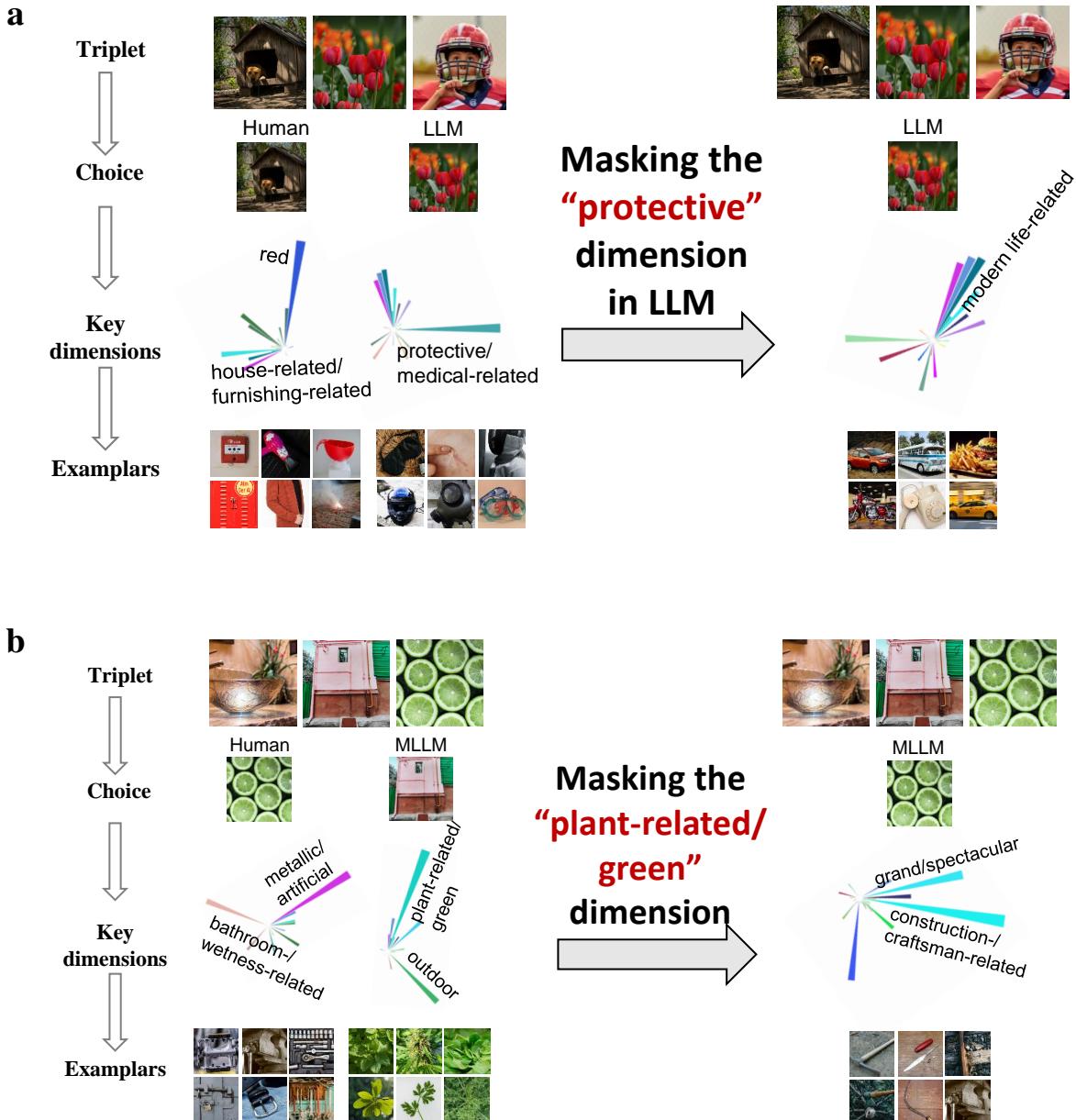
Prompt: You are shown three object images side by side and are asked to report the image that was the least similar to the other two. You should focus your judgement on the object and consider the aspect of "human-made" and "artificial" as your judging criteria. If you did not recognize the object, you should base your judgement on your best guess of what the object could be. 1. Tell me your answer. 2. Tell me the location of the object you have chosen. 3. Explain the reasons.

MLLM (Gemini Pro Vision 1.0):

1. The object that is least similar to the other two is the limes.
2. The limes are located in the third position (right).
3. The first object (a wooden water trough) and the second object (a downspout) are both human-made structures that serve practical purposes in water management. They are designed and constructed by humans to fulfill specific functions related to water use and drainage. In contrast, the third object (limes) is a natural fruit, not human-made or artificial. While limes can be processed or used in cooking, they do not fit the criteria of being a constructed object like the other two. Thus, the limes stand out as the least similar object in terms of being human-made and artificial.



Supplementary Fig. 3. Guiding MLLM's attention to the target dimension by using tailored prompts (related to Fig. 5). We added the phrase "consider the aspect of "human-made" and "artificial" as your judging criteria" to the prompt of MLLM. As can be seen, when the prompt included guidance on the dimensions prioritized by humans ("artificial"), the MLLM was able to make choice consistent with human judgment. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.



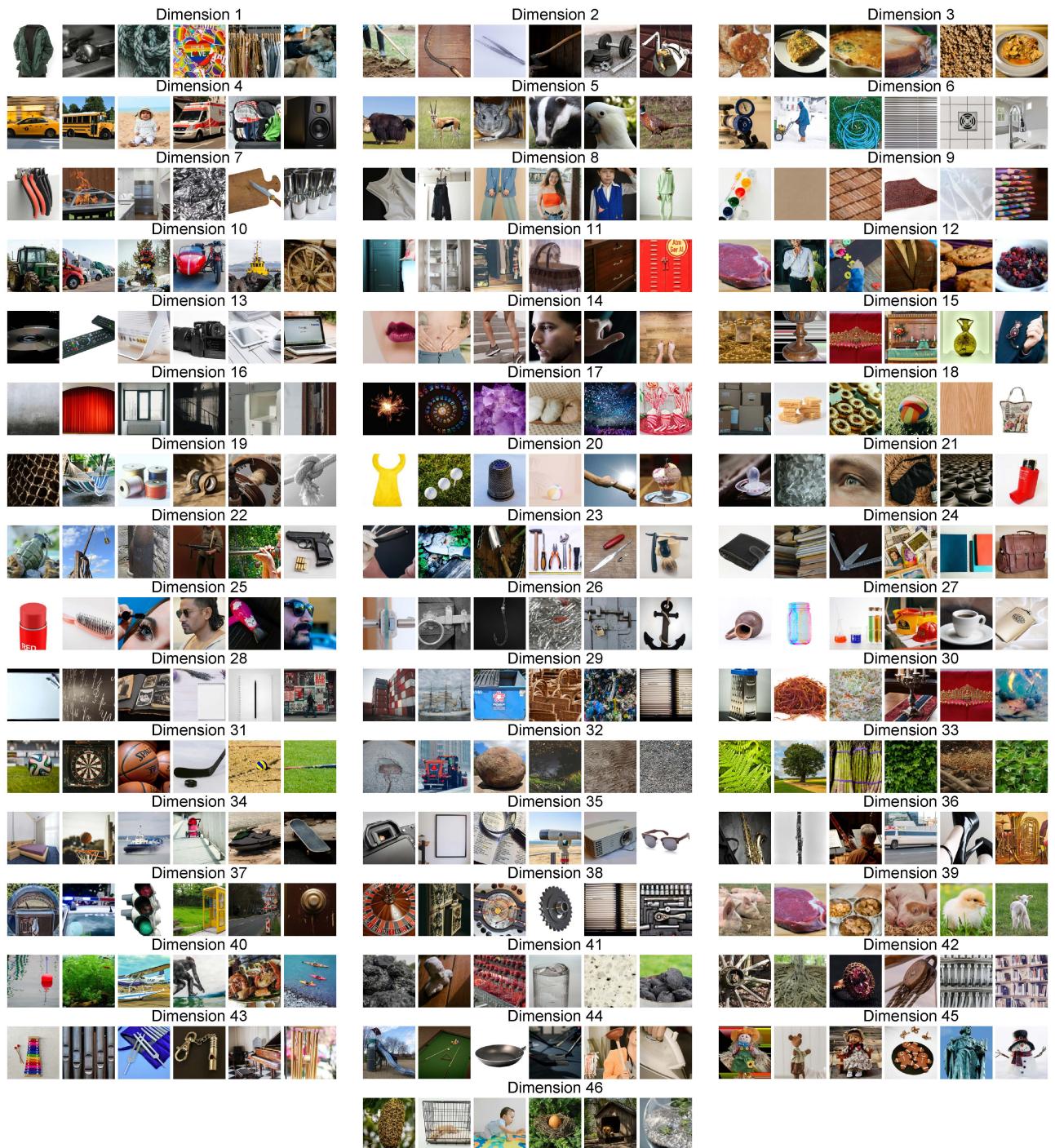
Supplementary Fig. 4. Masking the most critical dimension currently prioritized by the model but deviating from human preferences (related to Fig. 5). **a**, After masking the “protective” dimension, the LLM’s odd-one-out choice using the remaining 65 dimensions remained unchanged, but the key dimension it relied on shifted to “modern life-related.” **b**, After masking the “plant-related/green” dimension, the MLLM’s choice changed from “downspout” to “limes,” and the key dimension it relied on shifted to “construction-/craftsman-related.” From these two examples, it can be seen that directly masking certain key dimensions of the LLM/MLLM may or may not change the model’s behavioral choices. This intervention method has poor controllability over the model’s behavioral choices and the key dimensions it relies on, making it difficult to ensure that the model’s choices and the dimensions it relies on will become more aligned with human judgments. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.

Image	Name + definition (object-level)	Caption generated from MLLM (image-level)
	bagel ((Yiddish) glazed yeast-raised doughnut-shaped roll with hard crust)	a brown wicker basket placed on a wooden dining table, containing three large, tasty-looking bagels
	bear (massive plantigrade carnivorous or omnivorous mammals with long shaggy coats and strong claws)	a large brown bear walking across a grassy field with a stream nearby
	car (a motor vehicle with four wheels; usually propelled by an internal combustion engine)	a bright orange electric car parked in a parking lot
	flashlight (a small portable battery-powered electric lamp)	a black flashlight with a metal clip, sitting on a table
	sunflower (any plant of the genus <i>Helianthus</i> having large flower heads with dark disk florets and showy yellow rays)	a beautiful sunflower field with numerous sunflowers swaying in the breeze

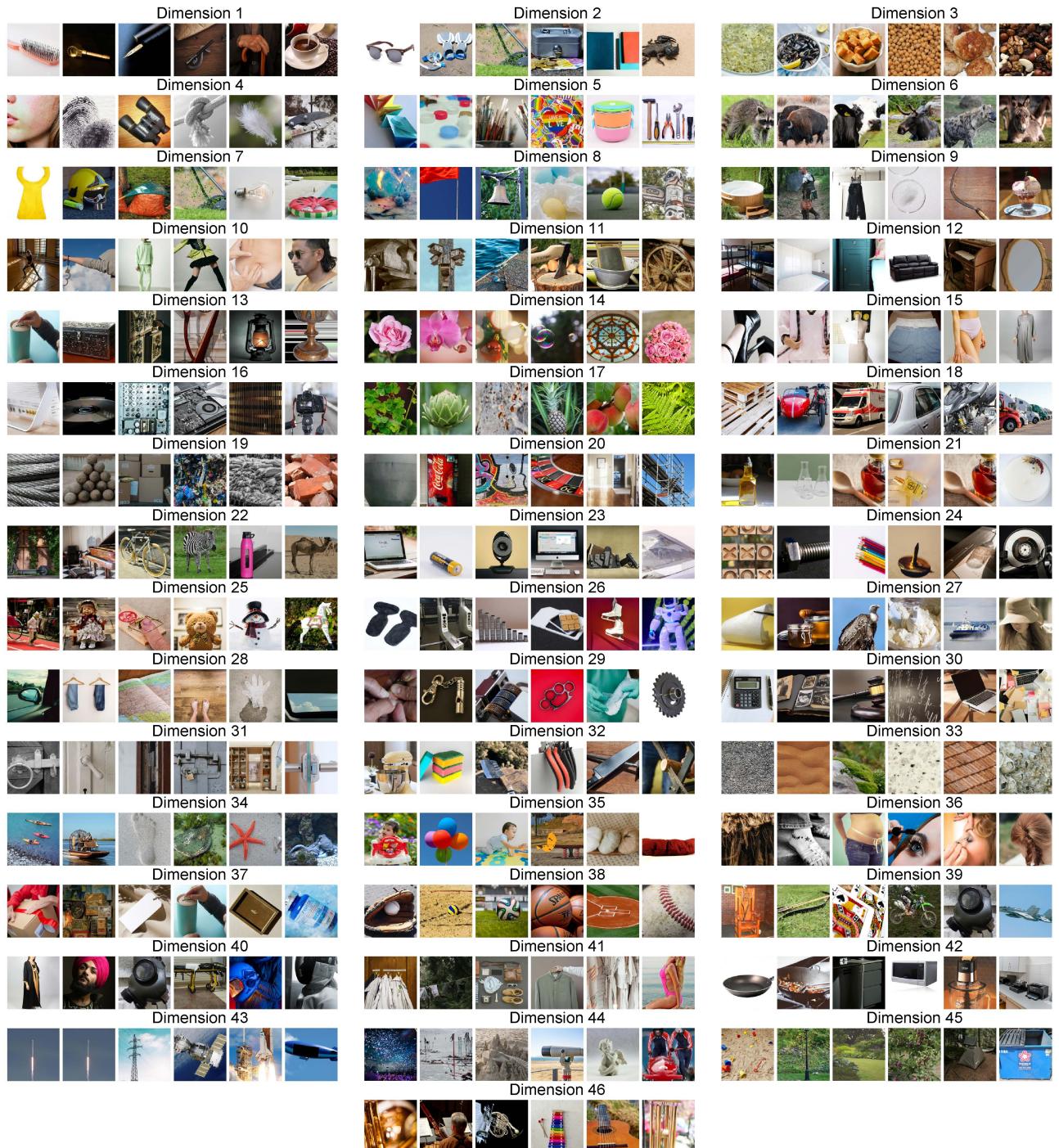
Supplementary Fig. 5. Two kinds of textual descriptions for example images (related to Fig. 1).

Object-level annotations: These annotations focus on the abstract, categorical representation of objects, typically using object names and definitions. They are well-suited for probing high-level conceptual understanding and are less sensitive to visual variations within a category. In our study, the LLM experiments using category-based annotations can be viewed as an "object-level" analysis, as they primarily assess the model's ability to distinguish between objects based on their conceptual categories.

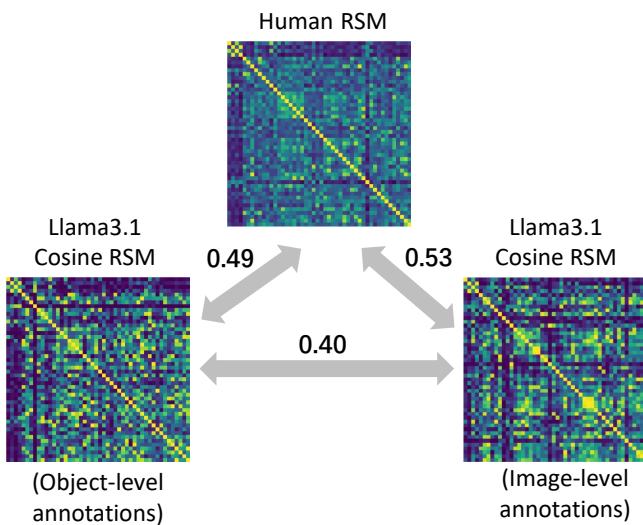
Image-level annotations: Here, the MLLM used for image caption generation was LLaVA-13B-v1-1 with the prompt as "Generate a detailed textual description of the image." These annotations capture detailed visual attributes of individual images, such as color, texture, and spatial relationships. They are more appropriate for tasks that require fine-grained visual discrimination or analysis of within-category variations. In our study, the MLLM experiments, which directly process the visual content of images, can be viewed as an "image-level" analysis, as they assess the model's ability to distinguish objects based on their visual features. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.



Supplementary Fig. 6. Object dimensions (1-46) illustrating their interpretability for LLaMA3.1 with object-level annotations (related to Fig. 4). We extracted representations from the object-level descriptions and efficiently constructed 4.7 million odd-one-out triplets based on their cosine distance. We then applied the SPoSE method to learn low-dimensional embeddings from these data, and each dimension was illustrated with the top 6 images with the highest weights along this dimension. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.



Supplementary Fig. 7. Object dimensions (1-46) illustrating their interpretability for LLama3.1 with image-level annotations (related to Fig. 4). We extracted representations from the image-level descriptions and efficiently constructed 4.7 million odd-one-out triplets based on their cosine distance. We then applied the SPoSE method to learn low-dimensional embeddings from these data, and each dimension was illustrated with the top 6 images with the highest weights along this dimension. In contrast to object-level approach, image-level approach resulted in the emergence of dimensions related to spatial (e.g., Dims. 3, 5, 19), textual (e.g., Dim. 33) and color (e.g., Dim. 14) attributes. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay and Pexels.



Supplementary Fig. 8. Comparison of the RSMs on the 48 typical objects measured by using different image annotation approaches (object-level vs. image-level) (related to Fig. 4). Cosine RSM was calculated from the model's cosine distance-based odd-one-out data. The numbers on the gray arrows represent the Pearson correlation between different RSM pairs. As can be seen, the RSM corresponding to the image-level annotation method aligns more closely with human judgments (0.53 vs. 0.49), primarily due to the fact that this annotation method leverages a vision-language model to generate image descriptions (effectively providing it with "eyes").