1	Multimodal Human Perception of Object Dimensions: Evidence from Deep Neural
2	Networks And Large Language Models
3	
4	Florian Burger (F.Burger@westernsydney.edu.au)
5	The MARCS Institute for Brain, Behaviour and Development, Western Sydney University,
6	Sydney, Australia
7	
8	Manuel Varlet (M.Varlet@westernsydney.edu.au)
9	The MARCS Institute for Brain, Behaviour and Development, Western Sydney University,
10	Sydney, Australia
11	School of Psychology, Western Sydney University, Sydney, Australia
12	
13	Genevieve Quek (G.Quek@westernsydney.edu.au)
14	The MARCS Institute for Brain, Behaviour and Development, Western Sydney University,
15	Sydney, Australia
16	
17	Tijl Grootswagers (T.Grootswagers@westernsydney.edu.au)
18	The MARCS Institute for Brain, Behaviour and Development, Western Sydney University,
19	Sydney, Australia
20	School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney,
21	Australia
22	
23	
24	
25	
26	
37	
29	
30	
31	
32	
33	
34	
35	
30 27	
२८ २८	
30	
40	
41	
42	
43	
44	
43 44	

#### 45 Abstract

46 Human object recognition relies on both perceptual 47 and semantic dimensions. Here, we examined how 48 deep neural networks (DNNs) and large language 49 models (LLMs) capture and integrate human-derived 50 dimensions of object similarity. We extracted layer activations from CORnet-S and obtained BERT 51 52 embeddings for 1853 images from the THINGS 53 dataset. We used support vector regression (SVR) to 54 quantify explained variance in human-derived 55 dimensions. Results showed that multimodal 56 integration improved predictions in early visual 57 processing but offers limited additional benefits at 58 later stages, suggesting that deep perceptual 59 processing already encodes meaningful object 102 60 representations.

61

62 Keywords: Object Recognition, Multimodal 105
63 Integration, Deep Neural Networks, Large 106
64 Language Models, Human-Derived Dimensions 107

#### 65 Introduction

66 Human object recognition depends on both 67 perceptual dimensions, such as color or shape, and 68 semantic dimensions, such as category or conceptual 69 relationships. To investigate the relevance of 70 perceptual dimensions, previous research has often 71 used deep neural networks (DNNs) to identify layers 72 that correspond to human object recognition (Cichy et 73 al., 2016; Kriegeskorte, 2015) and found that early 74 layers in DNNs tend to capture on more simple, 75 perceptual features, while higher layers align more 76 closely with complex, semantic features (Guclu & Van 77 Gerven, 2015). Other studies have shown that large 78 language models (LLMs) capture semantic aspects of 124 79 dimensions underlying object recognition, with high 80 similarity to human judgments (Grand et al., 2022).

Combining LLMs with DNNs has been shown
to outperform each modality individually (Marjieh et
al., 2023), reinforcing the importance of multimodal
integration in object recognition (Martin, 2016). To
understand how perception and meaning interact in
object recognition, it is essential to determine at which
processing stages semantic information contributes.

88 In this study, we integrated LLM embeddings 89 with individual DNN layers and identified how 90 perceptual and conceptual representations align along the visual hierarchy. This allowed us to guantify 91 92 where semantic knowledge enhances predictions of 93 human-derived object dimensions, revealing the 94 dynamics of multimodal integration in object 95 recognition.

### 96 Methods

97 We used a publicly available dataset of 49 human-98 derived dimensions underlying human object recognition (Hebart et al., 2020) for 1853 images from 99 100 the THINGS database (Hebart et al., 2019). Individual 101 layer activations from CORnet-S (Kubilius et al., 2019) were extracted using a forward-pass per 103 image, separately for the V1, V2, V4, and IT layers. 104 To reduce dimensionality and equate feature space size between the DNN and LLM, probabilistic PCA 106 (Halko et al., 2010) was applied separately to each 107 DNN layer retaining the first 200 components. Next, 108 LLM embeddings were derived from the BERT model 109 (Devlin et al., 2019) by averaging hidden states for 110 prompts based on concept names from the THINGS 111 dataset (e.g., "fish"), followed by probabilistic PCA 112 retaining the first 200 components. Finally, a 113 combined predictor set was created by concatenating 114 the top 100 PCA components from the DNN with the 115 top 100 PCA components from the LLM embeddings 116 for each image. This resulted in three distinct 117 predictor sets, each containing 200 features per 118 image: (1) DNN predictors (first 200 PCA 119 DNN activations), (2) LLM components from 120 predictors (first 200 PCA components from BERT 121 embeddings), and (3) Combined predictors 122 (concatenated top 100 PCA components from both 123 the DNN and LLM).

Using these three different predictor sets, a 125 10-fold cross-validated support vector regression 126 (SVR) with a radial-basis function kernel was used to 127 predict the loading for each image on 49 dimensions 128 (Hebart et al., 2020). To evaluate model performance 129 and determine whether predictions exceeded 130 chance-level accuracy, we calculated the explained 131 variance (R<sup>2</sup>) for each individual dimension 132 separately. Statistical significance was assessed 133 using 1000 permutations to generate a null 134distribution of R² values. The same procedure was153135applied to other DNNs (AlexNet, VGG) and LLMs154136(RoBerta, BERT) to assess the generalizability of the155137results across different architectures.156

138 Results

139 We first assessed how perceptual features (DNN layers), semantic features (LLM embeddings), and 140 159 141 their combination (DNN+LLM) predict human-derived 160 142 dimensions. Explained variance (R<sup>2</sup>) increased 161 143 progressively along the DNN hierarchy from lower-162 144 level layers (V1, V2) to higher-level layers (V4, IT), 163 reflecting a clear hierarchical structure of visual 145 164 146 feature complexity (Fig. 1A). Semantic features from 165 147 LLM embeddings alone also explained substantial 166 148 variance (Fig. 1A). 167

However, across dimensions and layers, the combination of DNN and LLM features had some added benefit. For early DNN layers, the combination with LLM added to the performance while the benefit decreased as we moved up the hierarchy (Fig. 1C).

### 158 Conclusion

We investigated how individual DNN layers and LLM embeddings correspond to human-derived object dimensions, and whether combining both modalities enhances this correspondence. While multimodal integration benefits early visual processing, the strongest predictions occurred at later stages, where DNNs already capture high-level object representations effectively. Our results suggest that linguistic knowledge does not consistently enhance



**Figure 1**. Multimodal prediction of human object dimensions. **A**: R<sup>2</sup> scores for each dimension (dots) across DNN layers, LLM, and combined models; gray lines connect the same dimension across models. Colored lines show trajectory of top 3 dimensions averaged across all methods. **B**: Heatmap of R<sup>2</sup> scores per model/layer (rows) and dimension (columns); each square represents one prediction. **C**: Average R<sup>2</sup> by layer, showing improved performance for combined models. Shown trend was observed across a variety of methodological choices.

149Inspecting individual dimension predictions168150revealed notable variation, with some dimensions169151benefiting more from perceptual than semantic170152features (Fig. 1B).171

8 DNN-based representations in IT. This highlights that
9 deep perceptual processing in DNNs already
0 incorporates meaningful structure at higher levels of
1 the visual hierarchy.

172	Acknowledgments	20
173	This work was supported by the Australian	20
174	Research Council (DE230100380)	20
174		2
475		2
175	References	2
176	Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A.,	2
177	& Oliva, A. (2016). Comparison of deep	2
178	neural networks to spatio-temporal cortical	2
179	dynamics of human visual object recognition	2
180	reveals hierarchical correspondence.	2
181	Scientific Reports, 6(1), 27755.	2
182	https://doi.org/10.1038/srep27755	2
183	Devlin, J., Chang, MW., Lee, K., & Toutanova, K.	24
184	(2019). BERT: Pre-training of Deep	24
185	Bidirectional Transformers for Language	24
186	Understanding. Proceedings of NAACL-	24
187	HLT, 4171–4186.	24
188	Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E.	24
189	(2022). Semantic projection recovers rich	24
190	human knowledge of multiple object features	24
191	from word embeddings. Nature Human	24
192	Behaviour, 6(7), 975–987.	2
193	https://doi.org/10.1038/s41562-022-01316-8	2
194	Guclu, U., & Van Gerven, M. A. J. (2015). Deep	2
195	Neural Networks Reveal a Gradient in the	2
196	Complexity of Neural Representations	2
197	across the Ventral Stream. Journal of	2
198	Neuroscience, 35(27), 10005–10014.	2
199	https://doi.org/10.1523/JNEUROSCI.5023-	2
200	14.2015	2
201	Halko, N., Martinsson, PG., & Tropp, J. A. (2010).	2
202	Finding structure with randomness:	2
203	Probabilistic algorithms for constructing	2
204	approximate matrix decompositions (No.	2
205	arXiv:0909.4061). arXiv.	2
206	https://doi.org/10.48550/arXiv.0909.4061	۲.

207	Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W.
208	Y., Corriveau, A., Van Wicklin, C., & Baker,
209	C. I. (2019). THINGS: A database of 1,854
210	object concepts and more than 26,000
211	naturalistic object images. PLOS ONE,
212	14(10), e0223792.
213	https://doi.org/10.1371/journal.pone.022379
214	2
215	Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C.
216	I. (2020). Revealing the multidimensional
217	mental representations of natural objects
218	underlying human similarity judgements.
219	Nature Human Behaviour, 4(11), 1173–
220	1185. https://doi.org/10.1038/s41562-020-
221	00951-3
222	Kriegeskorte, N. (2015). Deep Neural Networks: A
223	New Framework for Modeling Biological
224	Vision and Brain Information Processing.
225	Annual Review of Vision Science, 1(1), 417–
226	446. https://doi.org/10.1146/annurev-vision-
227	082114-035447
228	Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R.,
229	Hong, H., Majaj, N., Issa, E., Bashivan, P.,
230	Prescott-Roy, J., Schmidt, K., Nayebi, A.,
231	Bear, D., Yamins, D. L., & DiCarlo, J. J.
232	(2019). Brain-Like Object Recognition with
233	High-Performing Shallow Recurrent ANNs.
234	Marjieh, R., Rijn, P. van, Sucholutsky, I., Sumers, T.
235	R., Lee, H., Griffiths, T. L., & Jacoby, N.
236	(2023). Words are all you need? Language
237	as an approximation for human similarity
238	judgments (No. arXiv:2206.04105). arXiv.
239	https://doi.org/10.48550/arXiv.2206.04105
240	Martin, A. (2016). GRAPES—Grounding
241	representations in action, perception, and
242	emotion systems: How object properties and
243	categories are represented in the human

- 244 brain. *Psychonomic Bulletin & Review*,
  245 23(4), 979–990.
  246 https://doi.org/10.3758/s13423-015-0842-3
- 247

# 248 Supplementary Material

## 249 A: Description of Human-Derived

## 250 Dimensions (Hebart et al., 2020)

251

1	'made of metal / artificial / hard'
2	'food-related / eating-related /
	kitchen-related'
3	'animal-related / organic'
4	'clothing-related / fabric / covering'
5	'furniture-related / household-
	related / artifact'
6	'plant-related / green'
7	'outdoors-related'
8	'transportation / motorized /
	dynamic'
9	'wood-related / brownish'
10	'body part-related'
11	'colorful'
12	'valuable / special occasion-
	related'
13	'electronic / technology'
14	'sport-related / recreational
	activity-related'
15	'disc-shaped / round'
16	'tool-related'
17	'many small things / course
	pattern'
18	'paper-related / thin / flat / text-
	related'
19	'fluid-related / drink-related'
20	'long / thin'
21	'water-related / blue'
22	'powdery / fine-scale pattern'
23	'red'

24	'feminine (stereotypically) /
	decorative'
25	'bathroom-related / sanitary'
26	'black / noble'
27	'weapon / danger-related /
	violence'
28	'musical instrument-related /
	noise-related'
29	'sky-related / flying-related /
	floating-related'
30	'spherical / ellipsoid / rounded /
	voluminous'
31	'repetitive'
32	'flat / patterned'
33	'white'
34	'thin / flat'
35	'disgusting / bugs'
36	'string-related'
37	'arms/legs/skin-related'
38	'shiny / transparent'
39	'construction-related / physical
	work-related'
40	'fire-related / heat-related'
41	'head-related / face-related'
42	'beams-related'
43	'seating-related / put things on
	top'
44	'container-related / hollow'
45	'child-related / toy-related'
46	'medicine-related'
47	'has grating'
48	'handicraft-related'
49	'cylindrical / conical'

252