Seeing What Tastes Good: Revisiting Multimodal Distributional Semantics in the Billion Parameter Era

Dan Oneata* Desmond Elliott†,‡ Stella Frank‡,†
*POLITEHNICA Bucharest ‡Pioneer Centre for AI
†Department of Computer Science, University of Copenhagen
dan.oneata@gmail.com stfr@di.ku.dk

Abstract

Accurate understanding of a concept includes representing the common attributes and affordances of that concept across multiple modalities. We investigate the ability of pretrained vision models to represent the semantic attributes of concrete object concepts, e.g. a ROSE `is red`*,* `smells sweet`*, and* `is a flower`*. More specifically, we use probing tasks to test which properties of objects these models are aware of. We evaluate image encoders trained on image data alone, as well as multimodally-trained image encoders and language-only models, on predicting an extended denser version of the classic McRae semantic attribute norms, widely used in NLP, and the newer Binder dataset of attribute ratings. We find that multimodal image encoders slightly outperform language-only approaches, and that image-only encoders perform comparably to language models, even on non-visual attributes that are classified as "encyclopedic" or "function". These results offer new insights into what can be learned from pure unimodal learning, and the complementarity of the modalities.*[1]

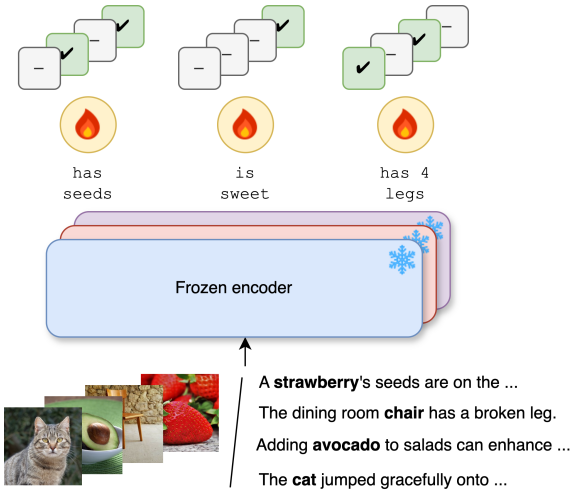

Figure 1. Given a dataset of concrete object concepts, depicted using either visual or linguistic data, we train linear probes on frozen modality-specific representations of to understand how well conceptual attributes can be extracted from models.

1. Introduction

Multimodal models depend on vision encoders to provide information about the objects that are depicted and their properties, their spatial configuration, lighting, and scene information. Recent work has highlighted a degree of linear alignment between neural network representations of the vision and language modalities [20, 23, 29]. This implies that the respective representation spaces have similar configurations, in terms of the local organisation (nearest neighbours) of concepts. However, there remains an open the question of *how* the different modalities "understand" or represent the concepts: which attributes are salient for a particular concept? In other words, how similar, in terms of underlying attributes, is a CHAIR as seen by a vision encoder to a CHAIR as encoded by a language model? This question concerns the complementarity of vision and language: are different modalities distinct, or in fact convergent [20]? Early work on distributional representations, in text-only [3, 25, 38] and multimodal [9, 10] models of static word embeddings, studied this question extensively. Recent advances in representation learning necessitates that we revisit this question to understand the relative representational power of each modality in modern models.

In this paper, we investigate how vision, language, and vision-and-language models represent concrete object concepts in terms of their associated attributes (semantic norms). We use a linear probing methodology to test whether model representations make distinctions corresponding to attributes associated with concepts, depicted visually or in text. Figure 1 presents an overview of our

[1] Results, code, and data are available on the project webpage: https://danoneata.github.io/seeing-what-tastes-good/

approach. The semantic norms cover many different types of attributes, from visual-perceptual `is green` to functional `is eaten` to encyclopedic `grows on trees`. Our first question is whether different encoders, from different modalities, capture particular attribute types more or less well.

Secondly, the models we evaluate correspond to a set of hypotheses about the role of language and labelling in conceptualization and category learning, a hotly debated topic in cognitive and neuroscience [5, 22, 26, 44]. At one extreme are pure vision encoders (ViT-MAE, DINOv2) trained without any language or category label supervision. At the other, models like CLIP and SigLIP learn to represent the visual input by aligning it to text: a form of language-steered world learning. We also evaluate text-only models that get categories for free (via word labels) but have to infer perceptual and other attributes from distributional semantics. Inasmuch language "carves up the world", visual encoders with more language input should be better aligned with semantic norms for English concepts.

We test these hypotheses using two concept attribute datasets. The first dataset links the semantic norms from the McRae dataset [28] to the concepts of the THINGS project [15], with an additional expansion step, to create the new McRae×THINGS dataset. The second is a dataset of neuro-cognitive attribute ratings from Binder et al. [7]. Our results demonstrate strong conceptual awareness in multimodal visual encoders across all attribute types. Moreover, while single-modality models behave most similarly (i.e. vision models and language models correlate most strongly within-modality), all performant models are highly correlated, indicating a degree of convergence, given exposure to sufficient data of either modality.

2. Related Work

Understanding the lexical semantics learned by language models is a long-standing concern in distributional semantics. A popular method for evaluating vector representations of lexemes is the correlation between the cosine similarity of two words in model space compared to human ratings of word similarity [9, 19]. Analogous work in the computer vision literature recently investigated the alignment between vision model representation spaces and human visual similarity judgements [27, 32, 40].

However, similarities alone cannot explain the dimensions of meaning space, or how the space distinguishes between human-meaningful attributes. In contrast, conceptual attributes (e.g., in the form of McRae norms [28]) can inform us about the organisation of a model's representation space. This type of data has been used to assess the complementarity of representations learned from language and vision [9, 10, 12, 13, 39, 47]. While some of this prior work shows that multimodal representations can improve attribute prediction [12, 13, 39], others observe only slight patterns of differences [9, 10, 47].

This latter finding (which we confirm for more recent models) is in line with more recent work [20, 23, 29], which posits a linear relationship between vision and language encodings. These works also compare across multiple vision models trained with different levels of supervision, and find that language supervision induces better downstream performance [29] and alignment with language model representations [20, 23] than no supervision.

3. Concept Attributes: Datasets

Understanding concepts via a core set of distinctive attributes is a long-standing quest in cognitive science [2, 14, 33, 37]. One method of discovering which attributes are important for human categorisation is *semantic norm elicitation*: participants are asked to write down the "characteristics and attributes" [37] or "properties" [28] they associate with a particular concept. Pooled over many participants, semantic norms thus represent a concept as a set of frequently mentioned salient attributes. While commonly used, semantic norm data have two important weaknesses. Firstly, they are biased towards attributes that are easily lexicalised. Secondly, they are not complete: less-salient, but present, attributes are often missing (e.g. TIGER but not CAT `has teeth`). To remedy this second issue, we synthetically "complete" the attribute values from McRae [28] across a large set of concepts. In addition, we also explore a recent dataset of *ratings* across a fixed set of attributes related to sensory and neurological dimensions not based on elicited lexicalised norms [7]. Since we are exploring visual and linguistic representations, the concepts we consider are concrete objects, corresponding to English nouns. We use the set of object concepts from THINGS [17], which also includes a set of quality-controlled images for each concept.

McRae×THINGS norms. The classic McRae semantic norms dataset [28] contains 541 concepts and 2 524 unique attributes. The attributes are classified into different types, such as "taxonomic", "functional", "visual-color", corresponding to associated brain regions [11]. We use only the attributes appearing with more than 5 concepts; we also group redundant attributes (*e.g.*, `used by the military`, `used by soldiers`, `used by the army`) using semantic similarity,[2] resulting in a final set of 278 attributes. We then find the corresponding norm/attribute values for all 1 854 concepts in THINGS. To obtain a complete mapping between concepts and attributes, we ask GPT-4o to annotate whether or not each attribute is a common trait of each concept (see Appendix A); each concept is briefly disambiguated and described using a definition extracted from the THINGS metadata. As a sanity check we verify that the norms (concept–attribute pairs) produced by our method include the norms

[2] Attributes whose cosine similarity is greater than 0.9 are merged, using the sentence embedding model all-MiniLM-L6-v2

in the original McRae set. As desired, the number of concepts positively associated with a given attribute increases. For example, the number of positive concepts for `tastes good` increases from 28 to 335; for `lays eggs` from 39 to 83; for `is dangerous` from 121 to 299. We note that Hansen and Hebart [15] also used an LLM-based process to collect norms for THINGS, but their process was designed to elicit more (potentially unique) norms for these concepts, whereas ours has the goal of comprehensive attribute annotation to avoid false negatives (missing positive values).

Binder ratings. Binder et al. [7] collected dense ratings for 65 "experiential attributes" of 534 concepts, of which we use the 155 concepts also found in THINGS. The experiential attributes correspond to lower-level conceptual dimensions such as visual `brightness`, somatic `pain`, or motor `movements in the upper/lower body`, and are organized into 14 different fine-grained domains (vision, somatic, etc.), collapsed to 7 coarser domains (sensory, motor, etc.). Participants used a 7-level rating scale,[3] which we binarize using the median value for each attribute.

4. Models

We study the performance of image-only, language-only, and multimodally-trained models on concept-attribute prediction. See Appendix C for further model details.

4.1. Vision-only Models

ViT-MAE [16] is a self-supervised visual encoder pretrained to reconstruct masked image patches at the pixel level using a deep Transformer encoder and decoder. **DINOv2** [34] is also a self-supervised visual encoder pretrained using a combination of image-level and patch-level objectives using a student and a teacher network [31]. This model is trained on a very large diverse dataset (142M images) without labels. **Swin-V2** [24] is a self-supervised visual encoder pretrained on ImageNet-21K to reconstruct masked image patches using a single linear layer [45]. **Max ViT** [42] is a Vision Transformer with Transformer blocks that combine convolution, block attention, and grid-based attention. This model is directly trained with a multi-class classification objective on ImageNet (IN-1K or IN-21K).

4.2. Multimodal Models

CLIP [36] has separate visual and textual encoders that are jointly optimized to maximize the similarity of image–sentence pairs. **SigLIP** [46] also has separate encoders that are trained to maximize a compute-efficient contrastive sigmoid loss. **PaliGemma** [6] is a generative vision-language model initialized from the SigLIP visual encoder and the Gemma language model [41], and is then further trained on a multimodal conditional language modelling task. We evaluate the visual encoder at the end of this multi-stage multimodal pretraining.

4.3. Language-only Models

FastText [30] creates static word embeddings by combining character n-grams embeddings within a white space-delimited word. **GLoVe** [35] also creates static embeddings based on aggregated global word-word co-occurrence statistics. For both FastText and GLoVe we use 300D embeddings trained on Common Crawl (840B tokens). **Gemma** [41] is a 2B parameter causal language model trained on 3T tokens. **DeBERTa-V3** is an language encoder trained on Wikipedia and the Books Corpus (3.1B words) to detect replaced tokens in sentences. **CLIP** [36] also has a language encoder; we use the 151M parameter model that was trained with the visual encoder.

5. Methodology

Each concept is represented as the average-pool of instance representations extracted from a frozen encoder (see Appendix B for details). We use trained linear probes [1, 4, 21] to measure the extent to which conceptual attributes (McRae feature norms or Binder attribute ratings) are evident in image and text representations. This evaluation requires generalizing attributes to unseen concepts, based on a small set of positive example concepts.

Each attribute is learned with a separate probe. The probes aim to separate the concepts that are positive for a given attribute (norm, rating) from the negative (McRae$\times$THINGS), or low-rated (Binder), concepts. For each attribute, we train a linear classifier that maps a concept representation to a binary label, using a simple logistic regression.[4] We generate 10 train–test splits for each attribute using 5-fold stratified cross-validation repeated twice, and report the average performance.

Our main evaluation metric is F_1 score. Following [18], we calculate the F_1 *selectivity* of each probe as the difference between the F_1 score on the correct labelling minus the expected random performance (i.e. the expected performance of a probe that learned a frequency bias). F_1 selectivity results are thus already with regard to a random baseline.

6. Results

The main linear probe accuracy results are shown in Table 1. Further results, including qualitative examples and the breakdown by attribute-type for the Binder dataset, can be found in Appendices D and E.

[3] Participants answered the question "To what degree do you think of CONCEPT as having/being associated with ATTRIBUTE?"

[4] We use `sklearn`'s default implementation without regularization and increase the maximum number of iterations to 1 000). We cannot train more elaborate (MLP) probes since our training datasets are very small, with few positive examples.

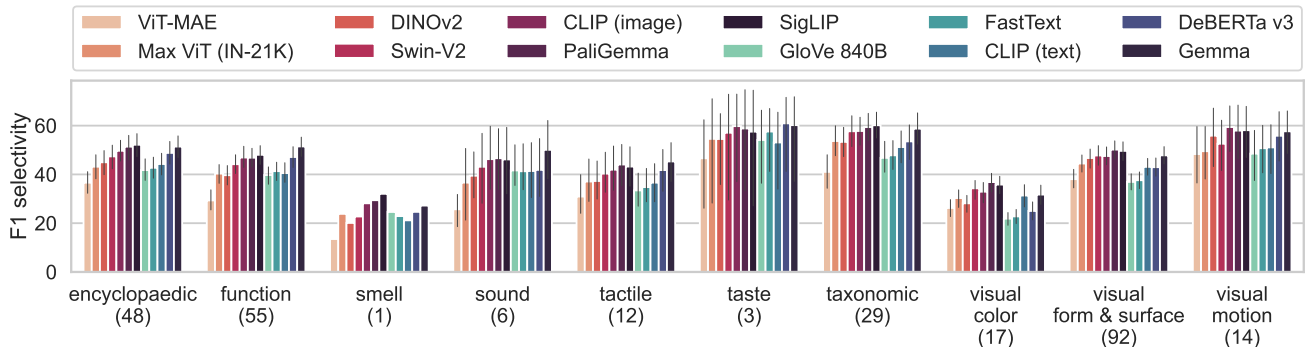

Figure 2. Performance per attribute type on the McRae×THINGS data with 95% confidence intervals using bootstrapping. Numbers below each type are the number of attributes per type. Vision models are in reddish colors; language models are in greenish colors.

Model	McRae×THINGS	Binder
Vision models		
Random SigLIP	15.4	9.3
ViT-MAE	35.6	18.8
Max ViT (IN-1K)	29.0	10.4
Max ViT (IN-21K)	43.3	21.5
DINOv2	44.5	22.7
Swin-V2	**47.0**	**23.9**
Multimodal vision models		
CLIP (image)	48.4	**25.5**
PaliGemma	49.9	25.0
SigLIP	**50.1**	25.2
Language models		
GloVe 840B	39.1	23.3
FastText	40.2	22.9
CLIP (text)	43.0	21.9
DeBERTa v3	45.5	25.3
Gemma	**49.8**	**25.7**

Table 1. Average F_1 *selectivity* of linear probes for semantic norms on the McRae×THINGS data and concept attribute ratings on the Binder data, corrected (per-probe) for random performance.

The impact of modality. Across the two datasets, the multimodal vision encoders are consistently amongst the highest performing models. However, the text-only Gemma-2B LLM also ranks highly. The self-supervised models Swin-V2 is the best vision model, but is outperformed by the multimodal vision encoders and Gemma.

Attribute type results. Are vision encoders better at visual-perceptual features? Do language models encode more functional-encyclopedic features? Figure 2 shows the results aggregated per the attribute types given in the McRae×THINGS dataset. We see that taxonomic, visual-motion, and taste attributes are the easiest to predict. The vision models, especially the multimodal models, generally outperform the language models, apart from Gemma. This makes sense for visual attributes like colour, but, surprisingly, this is the case even for encyclopedic and functional attributes, which could be easier to learn from textual inputs. Appendix F explores the effect of confounds. Overall, we see large variation between individual attributes within the same type. The Binder dataset shows similar patterns of variance and slightly less differentiation across model modalities (see Appendix Figures 5 and 6).

Effect of naming. Intriguingly, simple label-supervision in vision encoders seems to be harmful for learning human-aligned attributes, considering the relatively worse performance of Max ViT compared to Swin-V2. Richer text inputs, as seen by CLIP, SigLIP and PaliGemma, are necessary for multimodal complementarity.

7. Conclusion

Our linear probing analysis on two datasets shows that multimodally-trained vision encoders represent conceptual attributes better than single-modality encoders. However, single-modality encoders still perform well. Self-supervised models such as Swin-V2 have learned a sizeable amount of conceptual attribute knowledge, despite not having been trained to distinguish between concepts (rather than instances). For the language models, we find that simple static embeddings perform well at this concept-level task, especially compared with the effort required to get concept representations from Gemma-2B (Appendix G).

There is a long-held belief that multimodally-grounded representations are needed to overcome the limitations of learning from only linguistic or visual data. Our results suggest that Vision and Language encoders encode somewhat complementary views of concepts, inasmuch same-modality models correlate better than different-modality models (Appendix Figure 7). However, overall correlations are high, indicating a level of convergence. Previous claims of modality convergence have used nearest-neighbours measures [20, 23]; we show similar convergence results using a different linear probing methodology.

Acknowledgments

Dan Oneata is supported by the EU Horizon project AI4TRUST (No. 101070190) and by CNCS-UEFISCDI (PN-IV-P8-8.1-PRE-HE-ORG-2023-0078). Desmond Elliott is supported by a research grant (VIL53122) from VILLUM FONDEN. Stella Frank is supported by the Pioneer Center for AI, DNRF grant number P1.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *Proc. ICLR Workshop Track*, 2017. 3

[2] Aristotle. *Categories (Translated by E. M. Edghill)*. 4th c. BC / 1928. 2

[3] Marco Baroni and Alessandro Lenci. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1): 55–88, 2008. 1

[4] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. 3

[5] Yael Benn, Anna A Ivanova, Oliver Clark, Zachary Mineroff, Chloe Seikus, Jack Santos Silva, Rosemary Varley, and Evelina Fedorenko. The language network is not engaged in object categorization. *Cerebral Cortex*, 33(19):10380–10400, 2023. 2

[6] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 3

[7] Jeffrey R. Binder, Lisa L. Conant, Colin J. Humphries, Leonardo Fernandino, Stephen B. Simons, Mario Aguilar, and Rutvik H. Desai. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4):130–174, 2016. 2, 3

[8] Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proc. ACL*, 2020. 7, 9

[9] E. Bruni, N. K. Tran, and M. Baroni. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014. 1, 2

[10] Guillem Collell and Marie-Francine Moens. Is an image worth more than a thousand words? On the fine-grain semantic differences between visual and linguistic representations. In *Proc. COLING*, 2016. 1, 2

[11] George S. Cree and Ken McRae. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2):163–201, 2003. 2

[12] Steven Derby. *Interpretable Semantic Representations from Neural Language Models and Computer Vision*. PhD thesis, Queen's University, Belfast, 2022. 2

[13] Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. In *Proc. CoNLL*, 2018. 2

[14] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, 2000. 2

[15] Hannes Hansen and Martin N. Hebart. Semantic features of object concepts generated with GPT-3. In *Proc. CogSci*, 2022. 2, 3

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, 2022. 3

[17] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14 (10):e0223792, 2019. 2

[18] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proc. EMNLP-IJCNLP*, 2019. 3

[19] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015. 2

[20] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic representation hypothesis. In *Proc. ICML*, 2024. 1, 2, 4

[21] Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926, 2018. 3

[22] Anna A. Ivanova and Matthias Hofer. Linguistic overhypotheses in category learning: Explaining the label advantage effect. In *Proc. CogSci*, 2020. 2

[23] Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? A vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024. 1, 2, 4

[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proc. CVPR*, 2022. 3

[25] Li Lucy and Jon Gauthier. Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, 2017. 1

[26] Gary Lupyan. Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 3, 2012. 2

[27] Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *arXiv preprint arXiv:2406.19087*, 2024. 2

[28] Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris Mcnorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005. 2

[29] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *Proc. ICLR*, 2023. 1, 2

[30] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proc. LREC*, 2018. 3

[31] Théo Moutakanni, Maxime Oquab, Marc Szafraniec, Maria Vakalopoulou, and Piotr Bojanowski. You don't need domain-specific data augmentations when scaling self-supervised learning. In *Proc. NeurIPS*, 2024. 3

[32] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *Proc. ICLR*, 2023. 2

[33] Robert M. Nosofsky, Craig A. Sanders, Brian J. Meagher, and Bruce J. Douglas. Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, 50(2):530–556, 2018. 2

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, 2014. 3

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021. 3

[37] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605, 1975. 2

[38] Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. How well do distributional models capture different types of semantic knowledge? In *Proc. ACL-IJCNLP*, 2015. 1

[39] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Models of semantic representation with visual attributes. In *Proc. ACL*, 2013. 2

[40] Siddharth Suresh, Wei-Chun Huang, Kushin Mukherjee, and Timothy T Rogers. Categories vs semantic features: What shapes the similarities people discern in photographs of objects? In *Proc. ICLR Workshop on Representational Alignment*, 2024. 2

[41] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 3

[42] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxVIT: Multi-axis vision transformer. In *Proc. ECCV*, 2022. 3

[43] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics. In *Proc. EMNLP*, 2020. 7

[44] Sandra R. Waxman and Dana B. Markow. Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3):257–302, 1995. 2

[45] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Proc. CVPR*, 2022. 3

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. ICCV*, 2023. 3

[47] Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. Visual grounding helps learn word meanings in low-data regimes. In *Proc. ACL*, 2024. 2

A. Data Collection

Concept–attribute norm annotations. To obtain a complete representation of the THINGS concepts in terms of the (most frequent) attributes appearing in the McRae norms, we asked GPT-4o (`gpt-4o-2024-08-06`) whether each norm is a valid trait of each concept; Figure 3 shows the exact prompts. Given 1854 concepts and 278 attributes, this yields over 515k queries. We used the OpenAI Batch API for a the total cost of $127.64.

Annotation validation. When extracting the annotations from the GPT-4o output, we observed that the format was not always consistent: *e.g.*, the `valid` field was usually either `true` or `false`, but sometimes also `True`, `TRUE`, `yes`, `Yes`, `sometimes`, `False`, `no`, `No` (sometimes rendered as a string, sometimes as a literal); sometimes the `valid` field also included explanations for the chosen answer or the concept definition; sometimes the produced JSON used single quotes, sometimes double quotes. In retrospect, many of these exceptions may have been prevented by a more precise prompting, but they were not apparent when testing at smaller scale. To account for all these exceptions, we defined a custom parser that managed to extract a boolean value for each of the outputs. The resulting data is available on the project's webpage.

Textual contexts. The best performance for contextualized language models depends on having a collection of sentences in which the concepts appear. In the absence of a large and naturally occurring dataset of such sentences, we prompted the GPT-4o API (`gpt-4o-2024-08-06`) to collect the data. We also collected sentences with the addition constraint to avoid mentioning any of the positively-labelled attributes for a given concept. (This was in order to reduce the chance that the resulting embedding literally included features for the attributes.) Figure 4 shows the prompts used. The total cost of collecting the sentences was $26.24.

SYSTEM: "You are asked to decide whether an attribute is a common trait of a concept (to follow). Please answer the request in JSON format with the following structure: {'concept': CONCEPT, 'attribute': ATTRIBUTE, 'valid': ANSWER}"

USER: "Is {attribute} a common trait of {concept}, in the sense of {concept_definition}?"

Figure 3. The prompt used to collect the McRae×THINGS dataset.

SYSTEM: "You are asked to write {num} short sentences about a word (to follow). Answer the request by returning a list of numbered sentences, 1–{num}."

USER: "Write {num} short sentences about {concept}. You must use {concept} as a noun in each sentence."

SYSTEM: "You are asked to write {num} short sentences about a word (to follow). Answer the request by returning a list of numbered sentences, 1–{num}."

USER: "Write {num} short sentences about {concept}. You must use {concept} as a noun in each sentence. Try to avoid using the following phrases in any of the sentences: {positive_attributes}"

Figure 4. The prompts used to collect sentence contexts for each concept in the THINGS dataset. Top: Unconstrained prompt; Bottom: Constrained prompt. The constraint tried to prevent GPT4o from mentioning the attributes already associated with a concept.

B. Feature Extraction Details

Visual concept representations. In the visual modality, a concept is represented by all images from its THINGS concept class. The visual concept is computed by averaging the embeddings extracted from the last layer of a given vision encoder. Since many of the vision models produce a dense grid of embeddings, we obtain a single vector by average pooling the embeddings spatially.

Textual concept embeddings. In the language modality, a concept is represented by the English noun label given by McRae. Static word embedding models (GloVe, FastText) return an embedding directly, using only the surface form of the word. The static embeddings for multi-word concepts are averaged; homophones are not distinguished. Contextual language models (Gemma, DeBERTa v3, and CLIP) require words to be embedded in context to extract meaningful vector representations. In our experiments, a word's conceptual representation is the average over 10 sentences of the word in context (collected from the GPT4o API, see Appendix A), following [8, 43]. We find that each model requires a different extraction technique in order to achieve reasonable performance (see Appendix G). The best representations are found by mean-pooling over multiple layers [43]. For Gemma, we obtain much better performance using only the last token of the target concept word, while for the masked language model (DeBERTa v3) we use the mean over all concept tokens.

ViT-MAE	`facebook/vit-mae-large`
DINOv2	`facebook/dinov2-large`
Swin-V2	`swinv2_large_window12_192.ms_in22k`
Max ViT-1K	`maxvit_large_tf_384.in1k`
CLIP	`openai/clip-vit-large-patch14`
SigLIP	`google/siglip-so400m-patch14-224`
PaliGemma	`google/paligemma-3b-mix-224`
GLoVe	`glove-840b-300d`
Gemma-2B	`google/gemma-2b`
DeBERTa-v3	`deberta-v3`

Table 2. Models used in this paper.

C. Model Details

Table 2 shows the exact model versions used in the experiments and Table 3 provides an overview of the models in terms of their training data, objective function, type of supervision, and performance on the ImageNet-1K dataset.

D. Further Results

Detailed results. Table 4 presents the results in terms of precision, recall, raw F_1, and F_1 selectivity scores for both the McRae×THINGS and Binder datasets. The model rankings are similar across all metrics.

Per-attribute results on Binder. Figure 5 presents the detailed results on each of the 67 attributes from the Binder

Model	Params.	Dataset	Size	Objective	Labels	IN-1K
FastText	–	CommonCrawl	840B	NLL	–	–
GLoVe	–	CommonCrawl	840B	NLL	–	–
DeBERTa-V3	86M	Wiki+Books	3.1B	RTD	–	–
Gemma-2B	2B	Private	6T	NLL	–	–
ViT-MAE	304M	ImageNet-1K	1.3M	MSE	N/A	85.9
Max ViT†	212M	ImageNet-1K/-21K	1.3M/14M	Classification	Object classes	85.2/88.3
Swin-V2†	197M	ImageNet-21K	14M	SimMIM	N/A	87.7
DINOv2	304M	LVD	142M	DINO + iBOT	N/A	86.3
CLIP	304M	Private	400M	Contrastive	Sentences	83.9
SigLIP	400M	Private	4B	Sigmoid Contr.	Sentences	83.2
PaliGemma	400M	Private	1B	NLL	Sentences	N/A

Table 3. Overview of the models studied in this paper. The number of parameters in the encoder, the type and size of the pretraining data, the pretraining objective, and, where applicable, the reported **I**mage**N**et**1K** classification accuracy at 224px × 224px, except where noted otherwise. †: 384px × 384px

dataset. Figure 6 shows the results aggregated per attribute type (7 types). Results by attribute are somewhat unintuitive: visual attributes like "vision", "shape", "texture' have low scores, compared to some "social", "emotion", or "drive" attributes. This can be explained by observing that high values of these attributes are for concepts that (for example) "are associated with *having* a texture", rather than a specific texture: this may be something that vision models, that are trained to distinguish specific *kinds of* texture, have trouble with. We see that all models pattern quite similarly, though the stronger language models (Gemma, DeBERTa v3) are slightly better for "drive", "emotion", and "motor" domains.

Correlation between model predictions. Figure 7 shows the Pearson correlation of probe accuracy for all pairs of models. For the McRae×THINGS dataset, and to a lesser extent on Binder, we see modality clusters, where vision encoders (with the exception of Max ViT IN-1K) are correlated with each other, and likewise the language encoders. Correlations are quite high overall, however, indicating that all encoders across modalities are rather similar. Figure 8 visualizes norm prediction performance of specific pairs of models: vision-only Swin-V2 vs text-only Gemma, and CLIP image vs CLIP text.

E. Qualitative Results

In Figure 9 we show examples of predictions for four attributes (`has 4 legs`, `made of wood`, `is dangerous`, `tastes sweet`); for each we show five randomly sampled concepts. We show the best vision-only model (Swin-V2), the best language-only model (Gemma), and the language-and-vision models (CLIP image and CLIP text). For the attribute `has 4 legs` we see that the vision-based models (Swin-V2 and CLIP-image) label TABLECLOTH as positive, likely due to visual co-occurrence with TABLE. All models struggle with the difficult cases of KANGAROO, predicted as `having 4 legs`, and SKI, predicted as not `made of wood`. Some concept–attribute pairs are arguably ambiguous—is a CORKSCREW `dangerous`? is a TOMATO SAUCE `sweet`?—resulting in disagreements between models.

F. Possible Confounds between Attributes

Since linear probes are learned using attribute extensions (the set of positive examples of an attribute), we can't be sure they actually learn the attribute characteristics, and not some closely correlated, but more visually/textually available, attribute. For example, the two taste attributes (`tastes good` and `tastes sweet`) have extensions that are subsets of the food supercategory, which is learnable from visual features alone (e.g. as demonstrated by high performance on the taxonomic `is food` norm for all models, including DINOv2). Likewise, many of the motion attributes capture subsets of animals (`eats grass`). As a initial analysis, we check whether models are better at learning attributes that coincide with taxonomic supercategories, as provided by the THINGS dataset. The resulting correlations (Table 5) are highest for CLIP-image (0.594), FastText (0.578), and GloVe 840B (0.564), a heterogenous set of models in terms of modality and their linear probing accuracy.

G. Failures in Extracting Contextualized Textual Representations

Concept representations can, in principle, be extracted from any language model using just the surface-form of the concept label token(s). Here, we report a collection of negative results for this seemingly simple task using contextual lan-

Model	McRae×THINGS				Binder			
	P	R	F_1	F_1 sel	P	R	F_1	F_1 sel
Vision models								
Random SigLIP	26.2	28.0	26.8	15.4	60.6	60.3	59.8	9.3
ViT-MAE	49.6	46.1	47.0	35.6	70.0	70.0	69.4	18.8
Max ViT (IN-1K)	38.7	44.1	40.4	29.0	62.2	61.0	61.0	10.4
Max ViT (IN-21K)	63.5	50.3	54.7	43.3	71.6	73.6	72.0	21.5
DINOv2	59.8	**54.0**	55.9	44.5	73.8	73.7	73.2	22.7
Swin-V2	**67.3**	53.8	**58.4**	**47.0**	**74.8**	**75.2**	**74.5**	**23.9**
Multimodal vision models								
CLIP (image)	63.5	58.1	59.8	48.4	**77.0**	**76.2**	**76.1**	**25.5**
PaliGemma	67.3	58.2	61.3	49.9	76.0	76.1	75.5	25.0
SigLIP	**67.5**	**58.4**	**61.5**	**50.1**	76.8	76.0	75.8	25.2
Language models								
GloVe 840B	51.9	51.1	50.5	39.1	74.6	74.1	73.9	23.3
FastText	55.1	50.7	51.6	40.2	74.0	74.1	73.5	22.9
CLIP (text)	60.2	51.7	54.4	43.0	73.2	72.7	72.5	21.9
DeBERTa v3	64.2	53.2	56.9	45.5	76.9	76.1	75.9	25.3
Gemma	**68.7**	**57.2**	**61.2**	**49.8**	**77.1**	**76.5**	**76.3**	**25.7**

Table 4. Detailed results, in terms of precision (P), recall (R), F_1 score (F_1) and F_1 selectivity (F_1 sel) score, of concept attribute linear probes on the McRae×THINGS and Binder data.

Model	Correlation
Random SigLIP	0.339
Max ViT (IN-1K)	0.413
ViT-MAE	0.495
DeBERTa v3	0.536
Max ViT (IN-21K)	0.542
Gemma	0.543
Swin-V2	0.545
CLIP (text)	0.550
DINOv2	0.552
PaliGemma	0.554
SigLIP	0.561
GloVe 840B	0.564
FastText	0.578
CLIP (image)	0.594

Table 5. McRae×THINGS dataset: Correlation between per-attribute probing performance, as measured by F_1-selectivity, and the proportion of the attribute's extension belonging to a single super-category (i.e. the extent to which predicting the supercategory would lead to high precision).

guage models. Table 6 presents the complete results of our endeavours. Initial experiments with the Gemma-2B language model focused on using only the static embedding layer, which resulted in complete failure to train meaningful probes (**A**). Closer inspection revealed that the Gemma-2B tokenizer tokenizes single word inputs differently from words appearing in a sentence (i.e., words preceded by a space): `<bos>aardvark`→{`aard`, `vark`} instead of {`_aard`, `vark`}. Using the within-sentence (space-prepended) tokenization, performance improved but was still lower than expected (**B**). Following Bommasani et al. [8], we decided to collect contextualized sentence representations over a set of textual contexts for each concept. We collected 50 sentences from the GPT-4o API for each context (see Appendix A for details). These per sentence embeddings are averaged over multiple sentences, analogous to averaging the embeddings over multiple image instances. This greatly improved performance compared to using the embedding layer (**C**), and extracting the representation from the last later further improved performance (**D**). Another improvement was obtained by extracting the representation from the final subword token of a concept, i.e. `vark` in the tokenization of `aardvark` (**E**), and the final improvement involved extracting the representation as an average over multiple Transformer layers (**I**). The representations obtained from 50 sentences did not improve performance (**J**). Performance was slightly reduced using the contexts generated with the semantic norm constraints (**K**), indicating the model could use information from context sentences for this task. With this methodology fixed, we quickly found better representations for the DeBERTa v3 language encoder (**N**), and confirmed that this would also result in marginal improvements

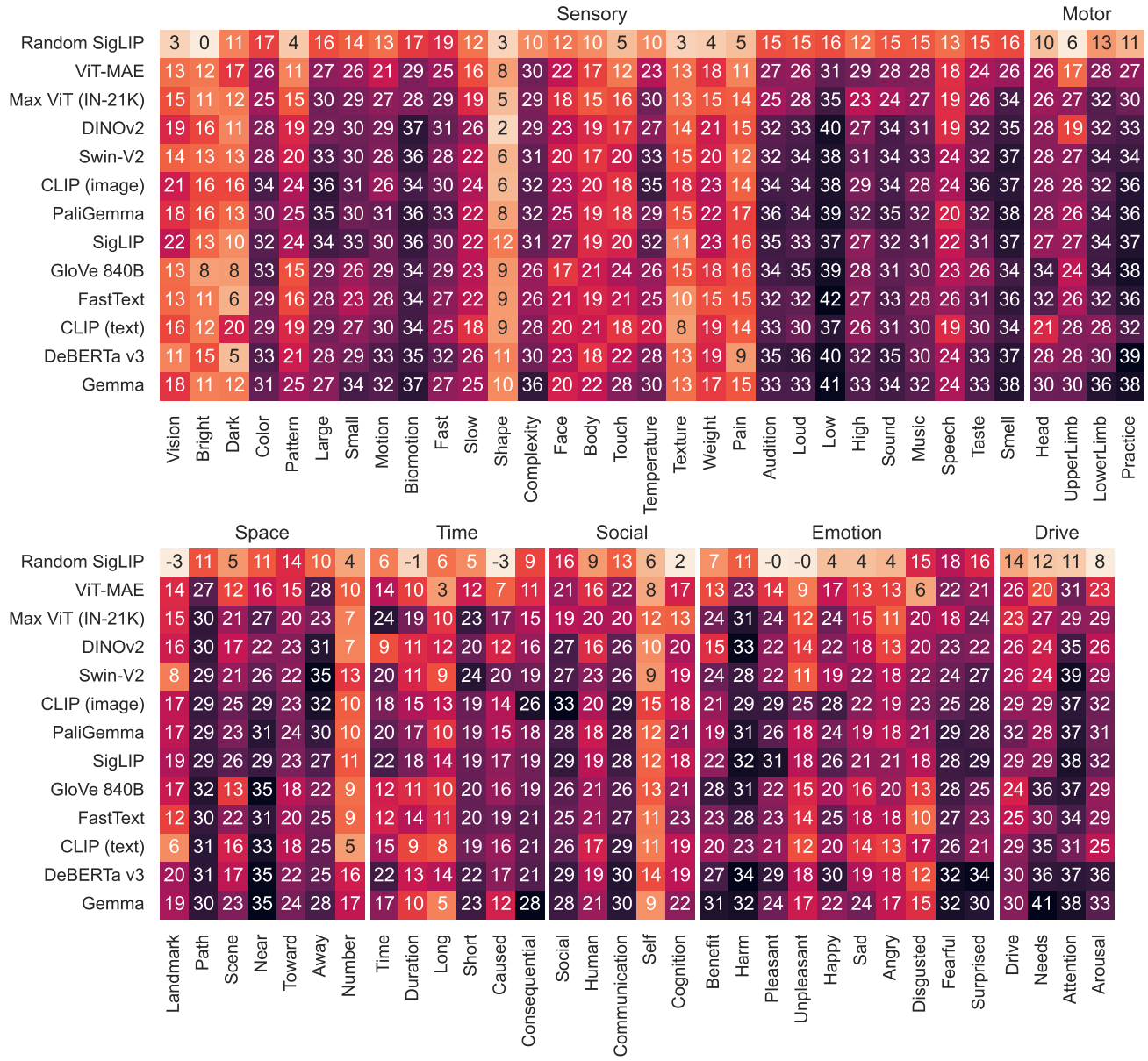

Figure 5. F_1 selectivity for the Binder attribute ratings. Note that raw F_1 score is much higher: the random baseline (against which F_1 selectivity is calculated) is 50% for evenly-distributed data.

for BERT (**Q**). We also report results for BERT base (uncased) and GPT-2 for completeness. We find that BERT base (uncased) performs much worse than DeBERTa v3 in similar conditions (**N** vs **Q**), and that GPT-2 also performs much worse than Gemma (**O** vs **D**). Given these findings, we do not include BERT or GPT-2 in our main results.

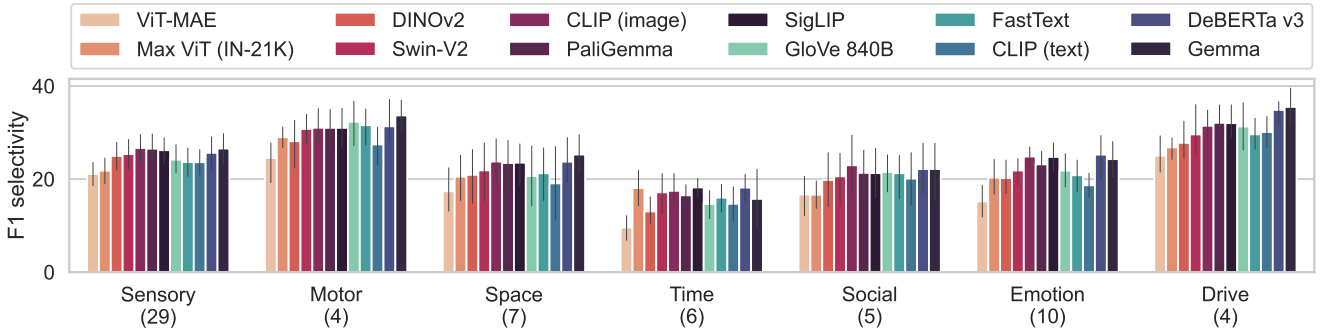

Figure 6. Results (F_1-sel) per attribute domain on the Binder data. The number below each domain indicates the number of attributes belonging to that domain. The error bars denote 95% confidence intervals using bootstrapping. Vision models are in reddish colors, while language models are in greenish colors.

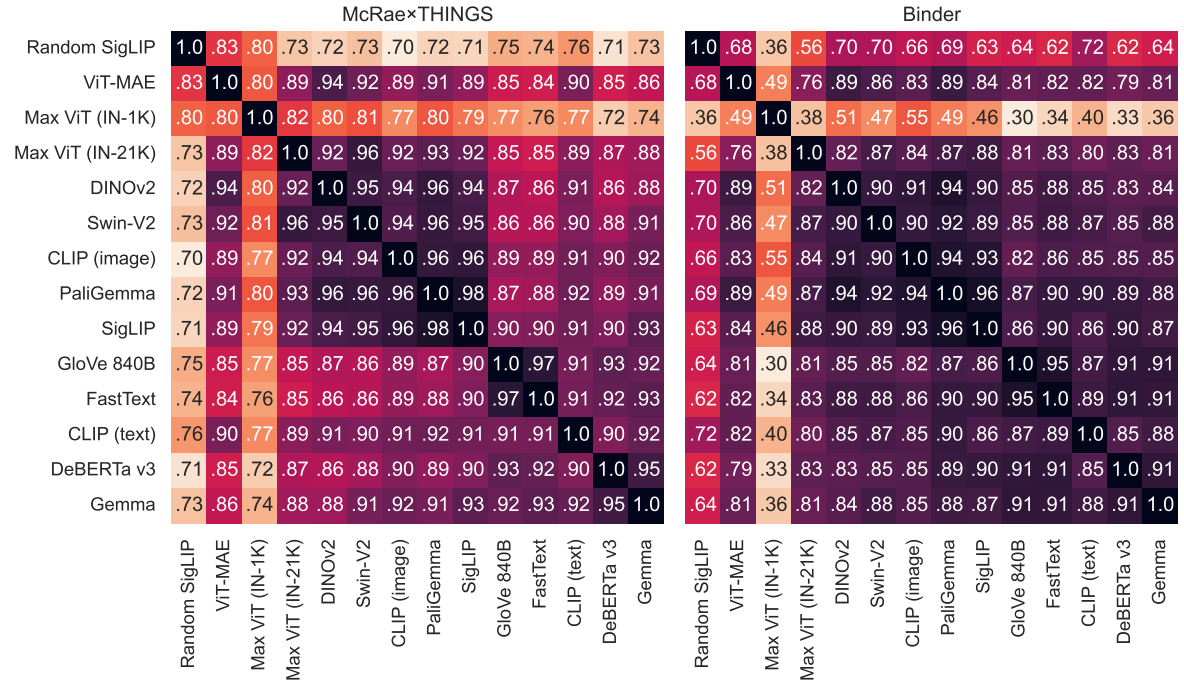

Figure 7. Pearson correlation between models based on attribute performance on the McRae×THINGS and Binder datasets.

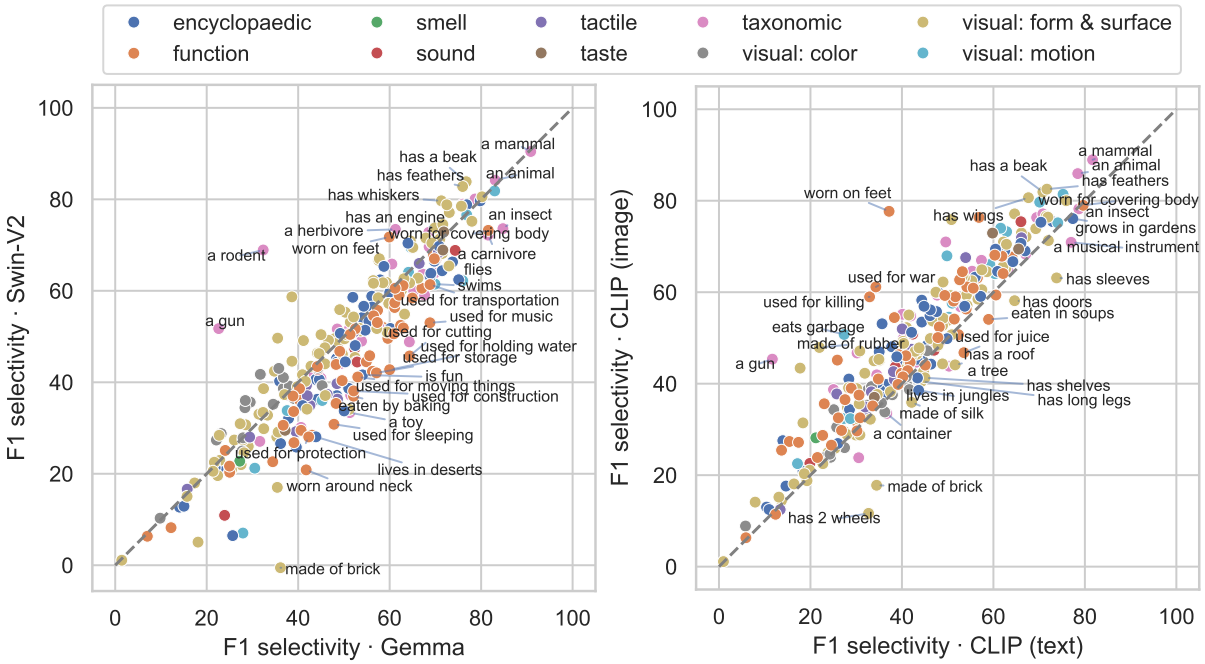

Figure 8. Per attribute comparison between pairs of models in terms of the F1 selectivity score. Left: Swin-v2 vs Gemma. Right: CLIP (image) vs CLIP (text).

Model	F1 sel.	Five random samples per feature norm and their predictions									
		has 4 legs (visual: form & surface)									
		DOG	+	STOOL	+	TABLECLOTH	−	ALTAR	−	KANGAROO	−
Swin-V2	78.5		✓		✓		✓		●		✓
Gemma	75.7		✓		●		●		✓		✓
CLIP (image)	76.6		✓		✓		✓		✓		✓
CLIP (text)	71.6		✓		✓		✓		✓		✓
		made of wood (visual: form & surface)									
		AXE	+	SKI	+	BOW3	−	PUPPET	−	CARDBOARD	−
Swin-V2	46.1		✓		●		●		●		✓
Gemma	49.7		✓		●		●		✓		✓
CLIP (image)	47.8		✓		●		●		●		●
CLIP (text)	43.8		✓		●		✓		✓		✓
		is dangerous (encyclopaedic)									
		DYNAMITE	+	BISON	+	RAZOR	+	CORKSCREW	−	TATTOO	−
Swin-V2	38.7		✓		✓		●		✓		●
Gemma	51.0		✓		●		✓		●		✓
CLIP (image)	44.8		✓		✓		●		●		●
CLIP (text)	38.9		✓		✓		✓		●		●
		tastes sweet (taste)									
		PLUM	+	RAISIN	+	CAKE MIX	+	TOMATO SAUCE	−	CRYSTAL1	−
Swin-V2	72.9		✓		✓		●		●		✓
Gemma	71.8		✓		●		✓		✓		●
CLIP (image)	72.9		✓		✓		●		●		●
CLIP (text)	59.8		✓		✓		✓		●		●

Figure 9. Five random predictions of linear probes trained on four attributes. Positive concepts are indicated by **+**, negative concepts by **–**. The linear probes are trained on embeddings from one of the four models: Swin-V2, Gemma, CLIP image and text encoders. If a model predicts a concept as having the attribute, we indicate this by ✓; otherwise we use ●. The correctness of the prediction is color-coded: green for a correct prediction, red for an incorrect one. In the second column, we show the F1 selectivity (%) for the each of the models and attributes.

	Model	Input	Seq.	Layer	McRae×THINGS P	R	F_1	F_1 sel
A	Gemma	word	mean	0 (emb)	43.2	25.3	30.3	18.8
B	Gemma	word (space)	mean	0 (emb)	58.3	37.9	44.2	32.8
C	Gemma	sentences (10)	mean	1	61.2	41.8	47.9	36.5
D	Gemma	sentences (10)	mean	18 (last)	63.8	52.4	56.3	44.9
E	Gemma	sentences (10)	last	18 (last)	66.5	56.8	60.2	48.8
F	Gemma	sentences (10)	mean	0–6	62.2	46.3	51.5	40.1
G	Gemma	sentences (10)	mean	0–9	62.3	48.7	53.2	41.8
H	Gemma	sentences (10)	mean	9–18	65.9	53.9	58.0	46.6
I	Gemma	sentences (10)	last	9–18	68.7	57.2	61.2	49.8
J	Gemma	sentences (50)	mean	18 (last)	62.7	52.1	55.8	44.4
K	Gemma	sentences (50, constr.)	mean	18 (last)	62.1	51.6	55.2	43.8
L	DeBERTa v3	sentences (10)	mean	12 (last)	43.9	42.9	42.8	31.4
M	DeBERTa v3	sentences (10)	mean	0–4	62.9	51.6	55.3	43.9
N	DeBERTa v3	sentences (10)	mean	0–6	64.2	53.2	56.9	45.5
O	GPT2	sentences (10)	mean	12 (last)	45.4	41.1	42.4	31.0
P	BERT base uncased	sentences (10)	mean	0–4	48.9	41.1	43.5	32.0
Q	BERT base uncased	sentences (10)	mean	0–6	50.9	42.7	45.2	33.8

Table 6. The effects of input (isolated concept word or contextual sentences), sequence pooling (mean or last token), and layer (individual layer or averaged over a range of layers) for the contextualised language models.