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

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

Abstract

Human learning and conceptual representation is grounded in sensorimotor experience, in contrast to state-of-the-art foundation models. In this paper, we investigate how well such large-scale models, trained on vast quantities of data, represent the *semantic feature norms* 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 norms 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 the 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]

1 Introduction

Multimodal models depend on vision encoders to provide information about the objects that are depicted, including their properties, 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 (Abdou et al., 2021; Merullo et al., 2023; Li et al., 2024; Huh et al., 2024). 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 concept? In other words, how similar, in terms of underlying attributes: is a CHAIR as seen by a vision encoder similar 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 (Huh et al., 2024)? Is a single modality, such as text, sufficient, or are multiple knowledge sources necessary? Early work on distributional representations, in text-only (Baroni and Lenci, 2008; Rubinstein et al., 2015; Lucy and Gauthier, 2017; Forbes et al., 2019; Misra et al., 2022, 2023) and multimodal (Bruni et al., 2014; Collell and Moens, 2016) models of static word embeddings, studied this question extensively. Recent advances in representation learning calls for revisiting this question to understand the relative representational power of each modality in modern models.

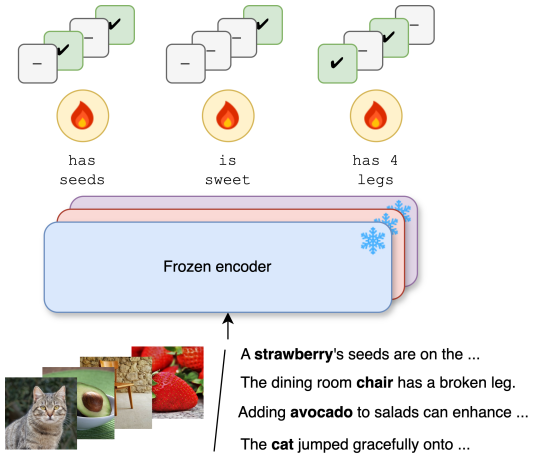

Figure 1: Given a dataset of concrete concepts, depicted using either visual or linguistic data, that are paired with semantic norms, we train linear probes on frozen modality-specific representations of to understand how well conceptual attributes can be extracted from models.

In this paper, we investigate how vision, language, and vision-and-language models represent concrete object concepts in terms of their associ-

[1]Code, datasets and results are available at: https://danoneata.github.io/seeing-what-tastes-good.

Findings of the Association for Computational Linguistics: ACL 2025, pages 24174–24191

ated 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 approach. The semantic norms cover many types of attributes, from visual-perceptual `is green`, to the functional `is eaten`, to the 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 (Waxman and Markow, 1995; Lupyan, 2012; Ivanova and Hofer, 2020; Benn et al., 2023). 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 as batch-wise nearest neighbours: 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 (McRae et al., 2005) to the concepts of the THINGS project (Hansen and Hebart, 2022), 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. (2016), which has been used to investigate language model representations (Utsumi, 2020; Turton et al., 2020; Chronis et al., 2023), but not, to our knowledge, visual or multimodal representations.

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.

The main contributions of this paper include:

- Improved understanding of the conceptual knowledge embedded in vision encoder models, ranging from self-supervised to class-supervised and language-supervised.
- **McRae×THINGS**: a new dataset of concepts densely annotated with semantic norms, using attributes from the McRae dataset and concepts from THINGS.
- Best practices for extracting representations for lexical semantic probing from LLMs.

2 Related Work

Understanding and evaluating the lexical semantics learned by language models via co-occurrence patterns 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 (e.g. using MEN (Bruni et al., 2014) and SimLex (Hill et al., 2015)). However, cosine similarity cannot uncover the underlying dimensions of meaning space, or how the space distinguishes between human-meaningful attributes. In contrast, testing for specific semantic attributes, by predicting semantic norms, can inform us about the underlying organisation of a model's representation space.

Baroni and Lenci (2008) were the first to use the McRae semantic norm dataset to evaluate the correspondence between early models of distributional semantics and cognitive concepts, using nearest-neighbors procedures. Using prediction models similar to linear probing, Rubinstein et al. (2015); Lucy and Gauthier (2017) find that static word embeddings encode taxonomic properties significantly more accurately than other types of properties, a finding we replicate. However, Sommerauer and Fokkens (2018) find that embeddings also reliably encode attributes that cut across taxonomic classes, such as `is dangerous`. Fagarasan et al. (2015) show that semantic norms can be predicted from word embeddings for unseen concepts. Contextual representation models outperform static word embeddings (Forbes et al., 2019; Bhatia and Richie, 2024). Misra et al. (2022, 2023) use semantic norms to explicitly probe for taxonomic generalization across concepts.

Conceptual attributes (either in the form of McRae norms directly or very similar data) have also been used to investigate the complementar-

ity of representations learned from language and vision. While Silberer et al. (2013); Derby et al. (2018); Derby (2022) show that multimodal representations can improve norm prediction, i.e. that two modalities are better than one, Bruni et al. (2014); Collell and Moens (2016) find only slight patterns of differences when they examine the differences between vision and language representations in predicting different attribute types.

This latter finding (which we confirm for more recent models) is in line with more recent work by Merullo et al. (2023); Li et al. (2024) which posits a linear relationship between vision and language encodings. These works also compare across different vision architectures with more or less supervision. Merullo et al. (2023) connect frozen visual encoders to frozen language models with a trained linear transform, and find that the performance on image captioning correlates with the amount of language supervision of the visual encoder: CLIP, trained with full captions, performs better than a model trained on category labels, and self-supervised BEiT, trained on image data alone, performs worst. Alternatively, Li et al. (2024) perform Procrustes analysis (a linear mapping) between image representations from ImageNet-trained vision models and language model representations for the same concepts, and find better alignment with larger language models, and with vision models that have been trained on supervised classification tasks, rather than self-supervised learning.

There is less work on the semantic alignment of vision model representations with human conceptual knowledge. In the computer vision literature, Muttenthaler et al. (2023); Mahner et al. (2024) has investigated the alignment between vision model representation spaces and human visual similarity judgements, using the THINGS dataset (Hebart et al., 2023). This work is directly analogous to evaluating pairwise similarities of language model representations against semantic similarity judgements, and as such, doesn't separate out individual concept attributes. Moreover, it assesses representations corresponding to instances (single images), rather than concepts (collections of instances). Mahner et al. (2024) compare sparse representations of human and model similarities, finding that while core dimensions overlap, humans use more semantic cues, and vision models rely more on visual cues, as well as many human-uninterpretable cues. In a study of several vision encoders, Muttenthaler et al. (2023) find that models trained on larger datasets and language supervision (CLIP) are more aligned with human similarity than smaller label- and self-supervised models. Finally, Suresh et al. (2024) show that image encoders trained to predict object attributes, rather than object classes, are more aligned with humans.

3 Concept Attributes: Datasets

Understanding concepts via a core set of distinctive attributes is a long-standing quest in cognitive science (Aristotle, 4th c. BC / 1928; Rosch and Mervis, 1975; Nosofsky et al., 2018; Gärdenfors, 2000). 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" (Rosch and Mervis, 1975) or "properties" (McRae et al., 2005) 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 not *complete*: less-salient, but nonetheless present, attributes of concepts are often missing (e.g. TIGER but not CAT `has teeth`). To remedy this first issue, we synthetically "complete" the attribute values from (McRae et al., 2005) across a large set of concepts. Secondly, norms are *biased* towards attributes that are easily lexicalised. We thus also explore a recent dataset of ratings across a fixed set of attributes related to sensory and neurological dimensions that are not based on elicited lexicalised norms (Binder et al., 2016).

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 (Hebart et al., 2019), which also includes a set of quality-controlled images for each concept.

McRae×THINGS norms. The original McRae semantic norms dataset (McRae et al., 2005) contains 541 concepts and 2 524 unique norms. The attributes are classified into different types, such as 'taxonomic', 'functional', 'visual-colour', corresponding to associated brain regions (Cree and McRae, 2003). We discard attributes appearing with fewer than five concepts; we also group highly similar attributes (e.g. `used by the military`, `used by soldiers`, `used by the army`) using se-

mantic similarity.[2] This results in a final set of 278 attributes. We then find the corresponding norms/attribute values for all 1 854 concepts in THINGS, resulting in a densely annotated dataset without missing norms.

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 in the original McRae set. We obtain a recall of 94–100% at responding correctly with respect to the human-authored attributes for a selection of ten attributes (one for each category), and, 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 (2022) 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. (2016) 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] and the final concept-attribute rating is the mean across participants.

4 Models

We primarily study the performance of image encoder models using Vision Transformers (ViT) backbones (Dosovitskiy et al., 2020), trained with different amounts of linguistic supervision. Table 1 presents a high-level overview. At one extreme, we use visual encoders trained *without any* label supervision. We also use encoders trained with object label classification supervision, e.g. trained on the ImageNet dataset. At the other end of the spectrum, we use visual encoders resulting from large-scale vision-language contrastive learning, and encoders derived from vision-language generative pretraining. The models were chosen so the encoders are approximately the same size, and operate over the same patch sizes. We also evaluate text-only embedding models, to compare the conceptual knowledge learned from the textual modality. Appendix B Table 4 shows the exact model names used in `timm` / HuggingFace Transformers.

4.1 Vision-only Models

ViT-MAE (He et al., 2022) is a self-supervised visual encoder pre-trained to reconstruct masked image patches at the pixel level using a deep Transformer encoder and decoder. **DINOv2** (Oquab et al., 2024) is also a self-supervised visual encoder pretrained using a combination of image-level objectives and patch-level objectives using a student and a teacher network (Moutakanni et al., 2024). This model is trained on a very large diverse dataset (142M images) without labels. **Swin-V2** (Liu et al., 2022) is a self-supervised visual encoder pretrained on ImageNet-21K to reconstruct masked image patches using a single linear layer (Xie et al., 2022). **Max ViT** (Tu et al., 2022) 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 (Radford et al., 2021) has separate visual and textual encoders that are jointly optimized to maximize the similarity of image–sentence pairs. **SigLIP** (Zhai et al., 2023) also has separate encoders that are trained to maximize a compute-efficient contrastive sigmoid loss. **PaliGemma** (Beyer et al., 2024) is a generative vision-language model initialized from the SigLIP-So400M visual encoder and the Gemma language model (Team et al., 2024). It is then further trained on a multimodal conditional language modelling task, and we use the visual encoder at the end of this multi-stage multimodal pretraining. **LLaVa-1.5** (Liu et al., 2024) is also a generative model that projects CLIP ViT/L embeddings into the Vicuna-

[2] We merge attributes with cosine similarity greater than 0.9, using the sentence embedding model all-MiniLM-L6-v2.

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

Model	Params.	Dataset	Size	Objective	Labels	IN-1K
FastText	–	CommonCrawl	840B	NLL	–	–
GLoVe	–	CommonCrawl	840B	NLL	–	–
Numberbatch	–	ConceptNet	N/A‡	PPMI	–	–
DeBERTa v3	86M	Wiki+Books	3.1B	RTD	–	–
Gemma	2B	Private	6T	NLL	–	–
ViT-MAE	304M	ImageNet-1K	1.3M	MSE	N/A	85.9
Max ViT (IN-1K)†	212M	ImageNet-1K	1.3M	Classification	Object classes	85.2
Max ViT (IN-21K)	212M	ImageNet-21K	14M	Classification	Object classes	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
LLaVa-1.5	324M	CC3M, OKVQA, etc.	1.2M	NLL	Sentences	N/A
Qwen2.5-VL	669M	Private	UNK	NLL	Sentences	N/A

Table 1: 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. ‡: ConceptNet is a knowledge graph of words and phrases with 8M nodes and 21M edges.

7B language model (Zheng et al., 2023) using an MLP projector. The model is multimodally trained on instruction data generated with GPT-4 on the CC3M dataset (Sharma et al., 2018), as well as on other scientific visual question answering datasets. **Qwen2.5-VL** (Bai et al., 2025) similarly integrates vision information through projection into an large language model, but in this model the image is input as a series of tokens, rather than as a single embedding. The model is trained in multiple stages on a wide variety of proprietary multimodal data.

4.3 Language-only Models

FastText (Mikolov et al., 2018) creates static word embeddings by combining character n-grams embeddings within a white space-delimited word. **GLoVe** (Pennington et al., 2014) 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). **Numberbatch** (Speer et al., 2017) embeddings (300D) are a combination of ConceptNet graph embeddings plus GLoVe and word2vec embeddings. **Gemma** (Team et al., 2024) 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** (Radford et al., 2021) also has a language encoder; we use the 151M parameter model that was trained with the visual encoder.

5 Methodology

We use trained linear probes (Alain and Bengio, 2017; Hupkes et al., 2018; Belinkov, 2022) 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 examples. Following standard methodology, the linear probes are trained on top of frozen representations, which allows us to estimate what is captured in the representations directly.

Each attribute is learned with a separate probe. For McRae×THINGS, we train a linear classifier for each attribute that maps a concept representation to a binary label, using a simple logistic regression.[4] For the Binder ratings dataset, we train a linear regression on each attribute to predict the mean rating

[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.

for each concept-attribute pair.[5] For both datasets, we generate 10 train–test splits for each attribute using 5-fold stratified cross-validation repeated twice, and report the average performance.

Visual concept representations. In the visual modality, a concept is represented by images from its THINGS concept class. The visual concept $\mathbf{e}_c$ 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, Numberbatch return an embedding directly, using only the surface form of the word.[6] Contextual language models (Gemma, DeBERTa v3) require a more careful methodology to extract meaningful vector representations. In these results, we always average over 10 sentences of the word in context (collected from the GPT4o API, see Appendix A), following (Vulić et al., 2020; Bommasani et al., 2020). We find that each model requires a different extraction technique in order to achieve reasonable performance; see Appendix E for failed attempts and suggestions for best practices. Briefly, the best representations are found from mean-pooling over multiple layers (Vulić et al., 2020). For Gemma, we obtain much better performance using only the last token of the target word, while for the masked language model (DeBERTa v3) we use the mean over all concept tokens.

5.1 Evaluation and Baselines

For McRae×THINGS, our main evaluation metric is F_1 score. Following (Hewitt and Liang, 2019), we calculate the *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. (A second random baseline is provided by the **SigLIP-Random** encoder, an untrained, randomly initialized, version of SigLIP.) For the linear regression results on Binder, we report root mean squared error (RMSE) as the main metric. (We also include F_1 accuracy results for logistic regression on a median-binarised version of Binder in Appendix C.)

[5]We use the `LinearRegression` implementation from `sklearn` with default settings: fit intercept, no regularisation.

[6]The static embeddings for multi-word concepts are averaged; homophones are not distinguished.

Model	McRae× THINGS F_1 sel ↑	Binder RMSE ↓
Vision models		
Random SigLIP	15.4	1.43
ViT-MAE	35.6	0.94
Max ViT (IN-1K)	29.0	1.37
Max ViT (IN-21K)	43.3	0.84
DINOv2	44.5	0.80
Swin-V2	**47.0**	**0.74**
Multimodal vision models		
LLaVA-1.5	45.0	0.83
Qwen2.5-VL	46.8	0.79
CLIP (image)	48.4	0.74
PaliGemma	49.9	0.73
SigLIP	**50.1**	**0.71**
Language models		
GloVe 840B	39.1	0.89
FastText	40.2	0.91
Numberbatch	44.1	0.83
CLIP (text)	43.0	0.81
DeBERTa v3	45.5	0.68
Gemma	**49.8**	**0.67**

Table 2: Performance of linear probes, averaged across attributes, for semantic norms on McRae×THINGS, and concept attribute ratings on Binder. We report F_1 selectivity on McRae×THINGS, which is corrected for random performance. On Binder we perform linear regression and report the root mean squared error (RMSE). More results can be found in Appendix Table 5.

6 Results

6.1 Main Results

The results for linear probe accuracy results are shown in Table 2; see also Appendix C, Table 5.

The impact of modality. Across the two datasets, the multimodal vision encoders are consistently amongst the highest performing models. However, the large text-only LLMs (Gemma-2B and DeBERTa v3) also rank highly. The self-supervised Swin-V2 model is the best vision model, and clearly outperforms (among others) the static word embedding models, despite having no access to lexical information.

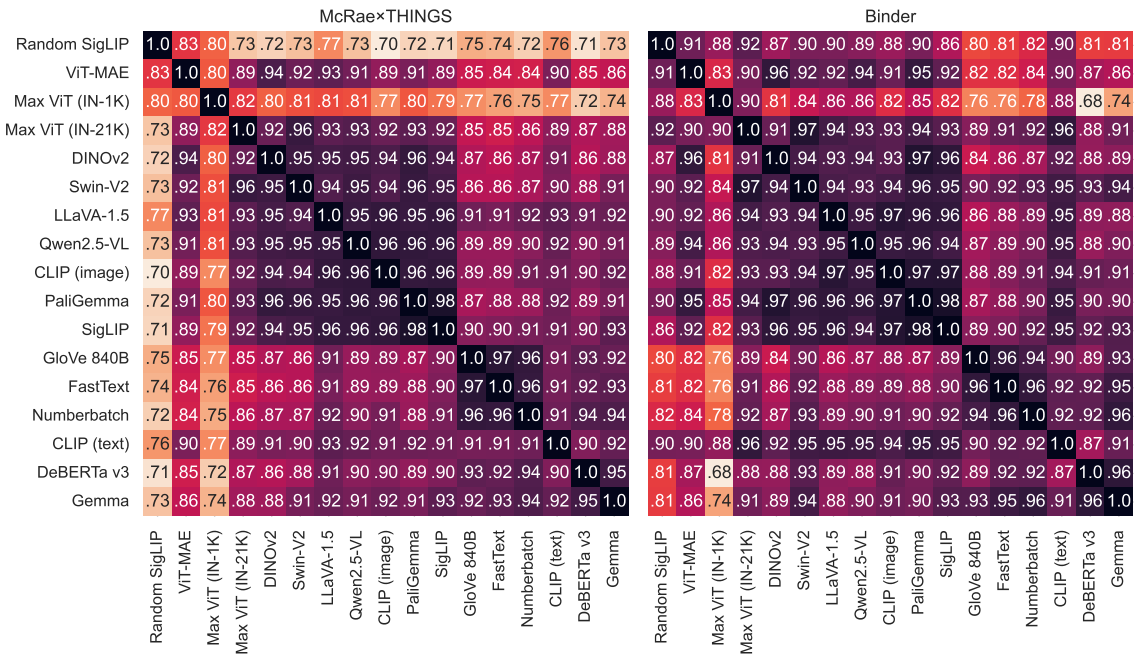

Figure 2: Per-attribute Pearson correlation between models on McRae×THINGS and Binder datasets.

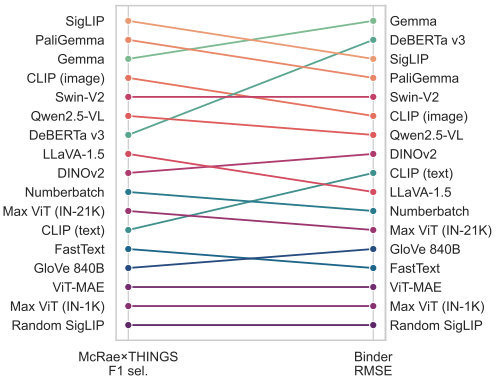

Figure 3: Relative rankings of models across the McRae×THINGS and Binder datasets (higher rank is better). The vision models are show in warm colours, language models in cool colours.

Dataset differences. Text-only models (especially Gemma and DeBERTa v3) perform relatively better on the Binder attribute dimensions, as seen in the rankings (Figure 3), while visually-informed models predict McRae×THINGS attributes better. Both dataset show large variation across different attributes.

Effect of training data amounts. Language models trained on larger amounts of data perform consistently better on McRae×THINGS and Binder. On the vision side, Swin-V2 learns better representations than DINOv2 for predicting semantic attributes, despite having seen one tenth as much data (14M vs 142M). Swin-V2 also outperforms the label-supervised ViT-MAE (IN-21K), having been trained on the same dataset, but with a less-informed objective. However, for Max ViT, the training data size has a substantial impact. For the multimodal vision models, the results on McRae×THINGS suggest that training data matters to some degree; for example, CLIP (image)[7] (400M) is outperformed by SigLIP and PaliGemma (5B). However, it is hard to disentangle the effect of dataset size from architecture and, in the case of language models, probing methodology (see Appendix E).

Correlation between model predictions. To understand the difference in model behaviour at the level of individual attributes, we calculate pairwise Pearson correlations between probe accuracy on different models (Figure 2). For the McRae×THINGS norms and Binder attributes, we see modality clusters, where vision encoders (with the exception of Max ViT IN-1K) are correlated with each other, and likewise the static word embedding models and the LLMS Gemma and DeBERTa v3. We also see some cross-modal correlations, with CLIP (text) correlating relatively higher

[7]We also evaluated the vision encoder of the performant open-weight (DFN2B-CLIP-ViT-L-14) trained on DataComp-1B for the same number of total training examples as OpenAI CLIP. It achieves an F_1 selectivity of 47.7 for McRae×THINGS and an RMSE of 0.79 on Binder.

with vision models in general (not only the CLIP image encoder), and Swin-V2 correlating more highly with language models on Binder. Overall, all (reasonable) model correlations are quite high, indicating that good encoders across modalities are rather similar. Inspecting the best and worst attribute for each model shows high consistency: For McRae×THINGS, the most accurate attribute across models is `is mammal`, while the worst is consistently `different sizes`.[8] For Binder the easiest attribute is `angry`, while the hardest is `sound` for most models. Figure 4 visualizes norm prediction performance of specific pairs of models (vision-only Swin-V2 vs text-only Gemma, CLIP image vs CLIP text), and qualitative examples can be found in Appendix D.

6.2 Attribute Type Results

Are vision encoders better at visual-perceptual features? Do language models encode more functional-encyclopedic features? To answer these questions we study performance aggregated by attribute type, as given by the datasets. Figure 5 presents the McRae×THINGS probing results per attribute type. Among the ten types, we see that taxonomic, visual-motion, and taste attributes are the easiest to predict. The vision models, especially the multimodal models, generally outperform the static word embeddings and to some extent the language models (Gemma and DeBERTa v3). This makes sense for visual attributes like colour, but, surprisingly, this is the case even for "encyclopedic" and "functional" attributes, which should be easier to learn from text than from visual inputs. Results by Binder attribute domain (Appendix Figs. 8 and 9) show similar patterns, with strong LLMs, multimodal vision encoders, and Swin-V2 performing similarly across attribute domains.

Possible confounds. Since linear probes are learned using attribute extensions (the set of positive examples of an attribute), we cannot be sure they actually learn the attribute characteristics, and not some closely correlated, but more visually or 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). 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 3) are highest for CLIP-image (0.594), FastText (0.578), and Numberbatch (0.573), a heterogenous set of models in terms of modality and their linear probing accuracy.

[8] Interestingly, this is an attribute that is clearly associated with the (variation shown by the) concept, instead of being associated with individual instances.

Model	Modality	Correlation
CLIP (image)	V(+L)	0.594
FastText	L	0.578
Numberbatch	L	0.573
LLaVA-1.5	V(+L)	0.565
GloVe 840B	L	0.564
SigLIP	V(+L)	0.561
PaliGemma	V(+L)	0.554
Qwen2.5-VL	V(+L)	0.553
DINOv2	V	0.552
CLIP (text)	L(+V)	0.550
Swin-V2	V	0.545
Gemma	L	0.543
Max ViT (IN-21K)	V	0.542
DeBERTa v3	L	0.536
ViT-MAE	V	0.495
Max ViT (IN-1K)	V	0.413
Random SigLIP	V	0.339

Table 3: McRae×THINGS dataset: Pearson correlation between per-norm probing performance, as measured by F_1 selectivity, and the proportion of the norm's extension belonging to a single supercategory (i.e. the extent to which predicting the supercategory would lead to high precision). Modality indicates which input each model operates on: vision (V) or language (L), with multimodality indicated in brackets.

7 Conclusion

This linear probing analysis on two datasets shows that multimodally-trained vision encoders represent conceptual attributes better than single-modality vision-only or text-only encoders. However, the single-modality encoders still perform well. In particular, the self-supervised Swin-V2, and to a lesser extent DINOv2 models, have learned a large amount of conceptual attribute knowledge, comparable to modern LLMs, and more than static word embeddings. This result is particularly surprising given that these vision models have

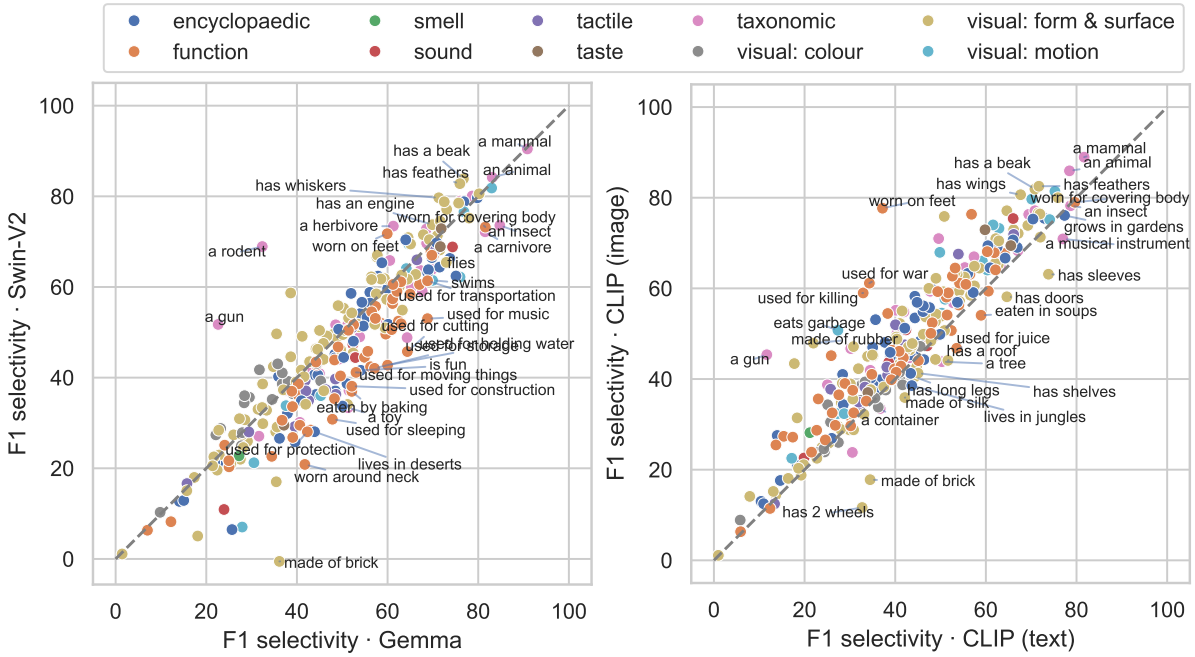

Figure 4: Per feature comparison between pairs of models in terms of the F1 selectivity score. Left: Swin-V2 vs Gemma. Right: CLIP (image) vs CLIP (text).

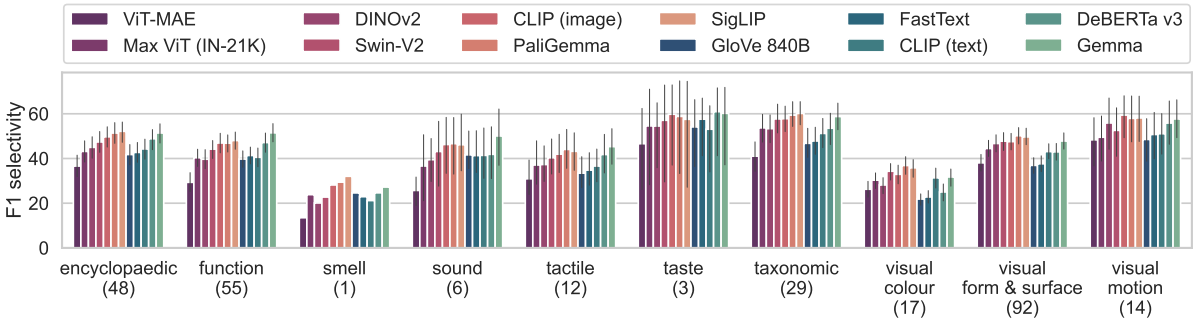

Figure 5: Results (F_1 selectivity) per attribute (norm) type on the McRae×THINGS data. The number below each type indicates the number of norms belonging to that type. The error bars denote 95% confidence intervals using bootstrapping. Vision models are in reddish colours, while language models are in greenish colours.

not been trained to distinguish between concepts, rather than instances, at all. Intriguingly, label-supervision of vision models seems to be harmful for learning human-aligned attributes, judging by the relatively worse performance of Max ViT, trained on ImageNet classification, compared to the self-supervised Swin-V2.

There is a long-held belief that we need multimodally-grounded representations to overcome the limitations of learning from only linguistic data. Our results suggest that Vision and Language encoders encode (somewhat) complementary views of concepts inasmuch same-modality models correlate stronger than different-modality models. However, overall correlations are high, indicating a level of convergence. Previous claims of modality convergence have used nearest-neighbours measures (Huh et al., 2024; Li et al., 2024); here we show similar convergence results using a very different linear probing methodology.

We expect models with conceptual knowledge organised in human-like ways, that are aware of the semantic attributes that underlie category memberships, would, in turn, achieve better downstream performance in language processing tasks. In future work, we will investigate the predictive power and utility of our probing tasks for multimodal training. This will also require going beyond simple object concepts to investigate more abstract, situational and configurational, concepts, in order to cover a larger proportion of the human conceptual repertoire.

Limitations

Linear probes Our linear probes assume that semantic attributes are encoded linearly in representation space. However, it is possible that semantic attributes are encoded as non-linear combinations: (Sommerauer and Fokkens, 2018) see increased probing accuracy with small MLPs compared to a logistic regression model such as we used. Our datasets are too small to learn MLPs without severe overfitting.

English-only Our experiments and analyses only concern evaluating the ability of models to predict the English semantic attributes of concepts expressed in English. This hinders our ability to make broader claims about the ability of models to perform this task in other languages, or for non-Western concrete concepts (Liu et al., 2021). In future work, we are interested in understanding the degree and quality of English-language influence on visual encoder representations.

Risks We forsee no risks associated with this research.

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

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? A case study in color. In *Proc. CoNLL*.

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

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

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

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

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

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

Sudeep Bhatia and Russell Richie. 2024. Transformer networks of human conceptual knowledge. *Psychological Review*, 131(1):271–306.

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

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

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

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. In *Proc. ACL*.

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

George S. Cree and Ken McRae. 2003. 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.

Steven Derby. 2022. *Interpretable Semantic Representations from Neural Language Models and Computer Vision*. Ph.D. thesis, Queen's University, Belfast.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020.

An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: Grounding semantic models in human perceptual data. In *Proc. IWCS*.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *Proc. CogSci*.

Peter Gärdenfors. 2000. *Conceptual Spaces: The Geometry of Thought*. The MIT Press.

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

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

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. 2023. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580.

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

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proc. EMNLP-IJCNLP*.

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

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The Platonic representation hypothesis. In *Proc. ICML*.

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

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

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. 2024. Do vision and language models share concepts? A vector space alignment study. *TACL*, 12:1232–1249.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proc. EMNLP*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proc. CVPR*.

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

Li Lucy and Jon Gauthier. 2017. 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*.

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

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

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

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

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

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proc. EACL*.

Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. 2022. A property induction framework for neural language models. In *Proc. CogSci*.

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

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

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

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

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

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

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

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. ACL*.

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

Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In *Proc. EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proc. AAAI*.

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

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

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

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating binder style word embeddings to new words. In *Proc. Workshop on Linguistic and Neurocognitive Resources*.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844.

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

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

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proc. NeurIPS*.

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 6 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.

ViT-MAE	facebook/vit-mae-large
Max ViT 1K	maxvit_large_tf_384.in1k
Max ViT 21K	maxvit_large_tf_224.in21k
DINOv2	facebook/dinov2-large
Swin-V2	swinv2_large_window12_192.ms_in22k
LLaVA-1.5	llava-hf/llava-1.5-7b-hf
Qwen2.5-VL	Qwen/Qwen2.5-VL-3B-Instruct
CLIP	openai/clip-vit-large-patch14
PaliGemma	google/paligemma-3b-mix-224
SigLIP	google/siglip-so400m-patch14-224
GLoVe	glove-840b-300d
DeBERTa v3	deberta-v3
Gemma	google/gemma-2b

Table 4: Precise names of the models used in this paper.

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 (`gpt4o-2024-08-06`) to collect the data. We also collected sentences with the addition constraint to avoid using any of the positively-labelled semantic norms for a given concept. (This was in order to reduce the chance that the resulting embedding literally included features about the expected norm.) Figure 7 shows the prompts used. The total cost of collecting the sentences was $26.24.

B Model Details

For reproducibility, Table 4 shows the exact model versions used in the experiments.

C Further Results

Detailed results. Table 5 presents the results in terms of precision, recall, raw F_1, and F_1 selectivity scores for the McRae×THINGS dataset and median-binarised Binder dataset. On the original Binder dataset, we report root mean squared and mean absolute errors.

Per-attribute results on Binder. Figure 8 presents the detailed results on each of the 67 attributes from the Binder dataset. Figure 9 shows the results aggregated per attribute type (7 types). We see that the auditory attributes (`audition`, `loud`, `sound`) are the most difficult. Distinguishing between positively and negatively associated concepts (`Benefit`, `Harm`, `Pleasant`, `Unpleasant`, `Happy`) is also surprisingly difficult. Interestingly, attributes to do with Time and negative Emotions (`sad`, `angry`, `disgusted`) are relatively easy for most models. Attributes that have directly to do with the human body (`Face`, `Body`, `Self`, `Human`) are also fairly easy.

D Qualitative Results

In Figure 10 we show results at the level of attributes and concepts. The results are four attributes (`has 4 legs`, `made of wood`, `is dangerous`, `tastes sweet`), and for each we show five random samples (concepts). For each sample we provide, the prediction using the same model selection as at the end of Section 6.1: that is, the best vision-only model (Swin-V2), the best language-only model (Gemma), and the language-and-vision models (CLIP image and CLIP text). Note that the models ingest the concept samples differently: the vision models average embeddings over multiple images, Gemma uses contextual sentences; so the images and concept word in Figure 10 are shown for illustrative purposes.

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.

E 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 language 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 inspec-

	McRae×THINGS				Binder (binarised)				Binder	
Model	P ↑	R ↑	F_1 ↑	F_1 sel ↑	P ↑	R ↑	F_1 ↑	F_1 sel ↑	RMSE ↓	MAE ↓
Vision models										
Random SigLIP	26.2	28.0	26.8	15.4	60.6	60.3	59.8	9.3	1.43	1.12
ViT-MAE	49.6	46.1	47.0	35.6	70.0	70.0	69.4	18.8	0.94	0.73
Max ViT (IN-1K)	38.7	44.1	40.4	29.0	62.2	61.0	61.0	10.4	1.37	1.07
Max ViT (IN-21K)	63.5	50.3	54.7	43.3	71.6	73.6	72.0	21.5	0.84	0.65
DINOv2	59.8	**54.0**	55.9	44.5	73.8	73.7	73.2	22.7	0.80	0.61
Swin-V2	**67.3**	53.8	**58.4**	**47.0**	**74.8**	**75.2**	**74.5**	**23.9**	**0.74**	**0.55**
Multimodal vision models										
LLaVA-1.5	59.1	55.5	56.4	45.0	74.6	74.0	73.8	23.2	0.83	0.64
Qwen2.5-VL	62.0	56.6	58.2	46.8	75.4	75.0	74.7	24.1	0.79	0.61
CLIP (image)	63.5	58.1	59.8	48.4	**77.0**	**76.2**	**76.1**	**25.5**	0.74	0.56
PaliGemma	67.3	58.2	61.3	49.9	76.0	76.1	75.5	25.0	0.73	0.55
SigLIP	**67.5**	**58.4**	**61.5**	**50.1**	76.8	76.0	75.8	25.2	**0.71**	**0.53**
Language models										
GloVe 840B	51.9	51.1	50.5	39.1	74.6	74.1	73.9	23.3	0.89	0.69
FastText	55.1	50.7	51.6	40.2	74.0	74.1	73.5	22.9	0.91	0.71
Numberbatch	59.6	54.0	55.5	44.1	75.0	75.0	74.5	23.9	0.83	0.65
CLIP (text)	60.2	51.7	54.4	43.0	73.2	72.7	72.5	21.9	0.81	0.63
DeBERTa v3	64.2	53.2	56.9	45.5	76.9	76.1	75.9	25.3	0.68	0.52
Gemma	**68.7**	**57.2**	**61.2**	**49.8**	**77.1**	**76.5**	**76.3**	**25.7**	**0.67**	**0.51**

Table 5: Detailed results, in terms of precision (P), recall (R), F_1 score (F_1) and F_1 selectivity score (F_1 sel), of concept norm linear probes on the McRae×THINGS and binarised Binder datasets. On the original Binder dataset we report root mean squared error (RMSE) and mean absolute error (MAE).

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 6: 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 7: The prompts used to collect sentence contexts for each concept in the THINGS dataset. Top: Unconstrained prompt; Bottom: Constrained prompt. The constraint tries to prevent GPT4o from mentioning the attributes already associated with a concept.

					McRae×THINGS			
	Model	Input	Seq.	Layer	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.

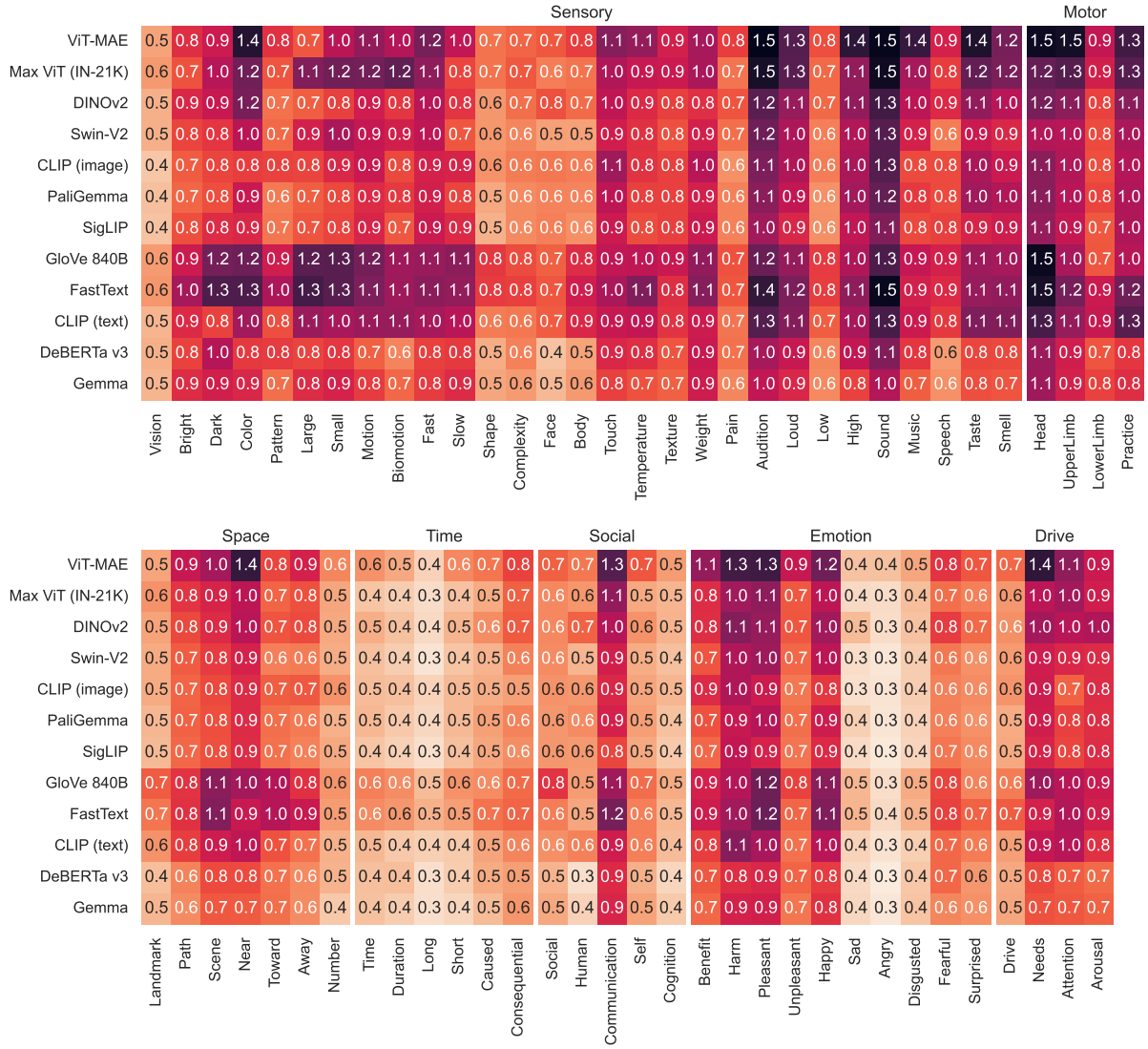

Figure 8: Per-attribute RMSE on Binder attribute ratings, across models. Lower is better.

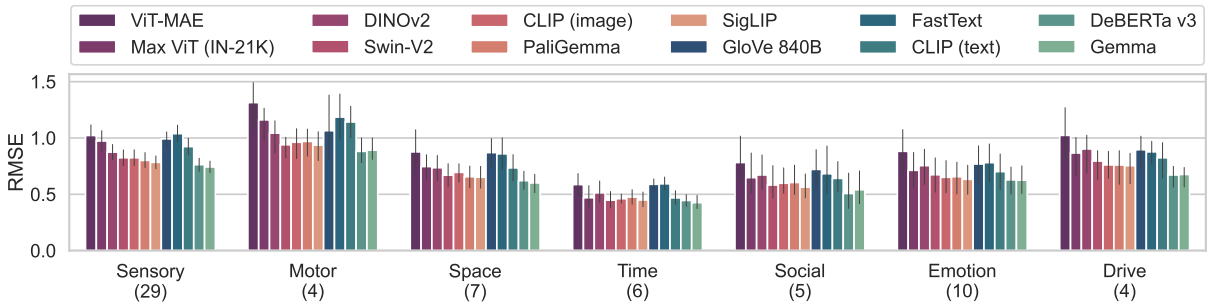

Figure 9: Results (RMSE) aggregated over attribute domain on the Binder data (note: lower is better). The number below each domain indicates the number of attributes belonging to that domain. The error bars denote 95% confidence intervals using bootstrapping. Vision models are in reddish colours, while language models are in greenish colours.

Model	F1 sel.	Five random samples per attribute 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 10: 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 colour-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.

tion 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**). Nevertheless, this approach was still substantially below the performance that we expected. Following Bommasani et al. (2020), 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 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.