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

Anonymous ACL submission

Abstract

There is widespread agreement about the grounded nature of human learning and representation, and the belief that computational models of meaning *need* to be multimodal. In this paper, we ask to what degree does this belief hold in the era of models trained on billions of examples? We investigate the ability of pre-trained vision models to 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 set 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 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 (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 particular 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)? Early work on distributional representations, in text-only (Baroni and Lenci, 2008; Rubinstein et al., 2015; Lucy and Gauthier, 2017) and multimodal (Bruni et al., 2014; Collell and Moens, 2016) 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.

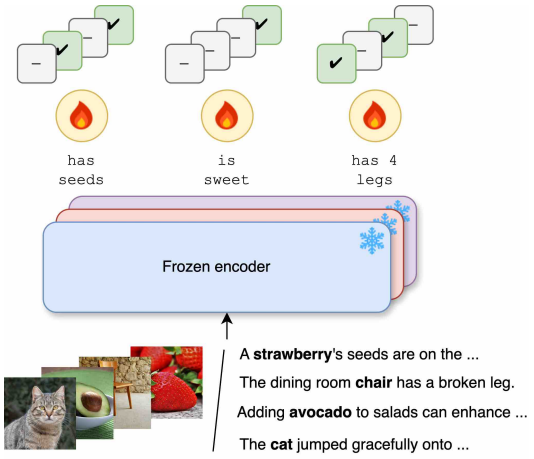

Figure 1: Given a dataset of concrete concepts, depicted using either visual or linguistic data, that are paired with semantic norms, we train linear probes on frozen modality-specific representations of to understand how well conceptual attributes can be extracted from models.

In this paper, we investigate how vision, language, and vision-and-language models represent concrete object concepts in terms of their associated attributes (semantic norms). We use a linear probing methodology to test whether model representations make distinctions corresponding to at-

tributes associated with concepts, depicted visually or in text. Figure 1 presents an overview of our 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 (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 is an extension of the classic McRae (McRae et al., 2005) semantic norms with cleaned synthetic norms from the THINGS project (Hansen and Hebart, 2022), to create the new McRae++ dataset. The second is a dataset of neuro-cognitive attribute ratings from Binder et al. (2016).

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 sufficent data of either modality.

The main contributions[1] of this paper include:

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

[1]The datasets and code are available at: [ANONYMIZED].

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) test static word embeddings and find that they 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, while (Derby et al., 2019; Bhatia and Richie, 2024) add semantic norms as additional input to models of distributional semantics.

Conceptual attributes (either in the form of McRae norms directly or very similar data) have also been used to investigate the complementarity 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 Li et al. (2024) and Merullo et al. (2023) 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. Here we use an extended version of the McRae norms (McRae et al., 2005).

While commonly used, semantic norm data have two important weaknesses. Firstly, they are biased towards attributes that are easily lexicalized. Secondly, they are not complete: less-salient, but present, attributes are often missing (e.g. TIGER but not CAT `has teeth`). We thus also use a recent dataset of *ratings* across a fixed set of attributes related to sensory and neurological dimensions (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++ norms. The classic McRae semantic norms dataset (McRae et al., 2005) contains 541 concepts and 2 524 unique norms. The norms are classified into different types, such as 'taxonomic', 'functional', 'visual-color', corresponding to associated brain regions (Cree and McRae, 2003).

Hansen and Hebart (2022) extend the McRae dataset using GPT-3 to the 1 854 concepts in THINGS (Hebart et al., 2019). The GPT-3 generated norms are sufficiently similar to human-generated norms, in terms of their distribution across feature types, predicted super-category structure, and concept similarity, but have greater coverage in terms of concepts and features. However, the norms are often noisy (`m sorry`, `is vitamins`), redundant (`is gray`–`is gray in color`), or too generic (`comes in different sizes and colors`).

We create a new cleaner set of semantic norms for the THINGS concept set, with the aim is to increase the set of associated concepts for each norm. To do this, we map the GPT-3-generated Hansen

Name	Feature norm	Concepts
McRae	has seeds	{APPLE, AVOCADO, CANTALOUPE, CUCUMBER, DANDELION, GRAPE, HONEYDEW, LEMON, LIME, MANDARIN, ...} (21 concepts)
Hansen	has seeds (sim: 1.0) has many seeds (sim: 0.9)	{APPLE, BANANA, BELL PEPPER, ...} (45 concepts) {POMEGRANATE} (one concept)
McRae++	has seeds	{APPLE, AVOCADO, CANTALOUPE, ... BANANA, BELL PEPPER, BERRY, ..., POMEGRANATE} (49 concepts)

Table 1: McRae++ extends the McRae norms to the THINGS concepts through the Hansen norms. In this example, the "has seeds" McRae norm is mapped to the "has seeds" and "has many seeds" Hansen norms and inherits all their associated concepts (in orange). We also include the original concepts in McRae that overlap with THINGS (dark blue) and discard the others (light blue). We obtain a more complete set then either McRae or Hansen.

Name	Norms	N $\geq$ 10	Concepts	Types
Original dataset				
McRae	2 524	122	541	10
Hansen	84 561	1 548	1 854	N/A
McRae++	2 202	533	1 854	10
Binder	65	65	534	(14)/7
Filtered variants				
McRae++	533	533	1 854	10
Binder	65	65	155	7

Table 2: Summary of norm datasets: McRae and Hansen datasets contain many concepts and norms but few norms that are mentioned across sufficient ($\geq$ 10) concepts. McRae++ more than quadruples the number of well-represented norms. Types marks the annotation of norms with higher-level attribute types. We use filtered McRae++ and Binder. Filtered McRae++ contains only norms with $\geq$ 10 THINGS concepts; filtered Binder only the concrete object concepts in THINGS.

norms to McRae norms, using cosine similarity calculated by a sentence embedding model[2]. If a Hansen norm matches to a McRae norm (similarity > 0.9), then we add all the concepts with that Hansen norm to the matched McRae norm (Figure 1). Finally we filter out any McRae concepts that are not in THINGS. This procedure results in a set of 2 202 feature norms, corresponding to the feature norms in McRae with concepts in THINGS. We then filter to only keep the feature norms with more than 10 concepts, resulting in 533 norms.

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], which we binarize using the median value for each attribute.

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 3 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. Table 6 shows the precise models 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

[2] 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
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 visual encoder models studied in this paper. The number of parameters in the visual encoder, the type and size of the pretraining data, the pretraining objective, and, for context, the reported **I**mage**N**et**1K** classification accuracy at 224px × 224px, except where noted otherwise. †: 384px × 384px

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 a small dataset (ImageNet-1K).

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.

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). **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. The probes aim to separate the concepts that are positive for a given attribute (norm, rating) from the concepts that don't mention the norm (McRae), or are rated too low (Binder). Concretely, for each feature norm f, we train a linear classifier[4] that maps a representation $\mathbf{e}_c$ of a concept c to a binary label y. For each feature norm, we generate 10 train–test splits using 5-fold stratified cross-validation repeated twice, and report the average performance.

[4]We use a simple logistic regression function (`sklearn`'s default implementation without regularization, and increasing 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.

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) return an embedding directly, using only the surface form of the word[5]. Contextual language models (Gemma, DeBERTa v3, and CLIP) 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 C), 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 B 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

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). Selectivity-F_1 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.

6 Results

6.1 Main Results

The results for linear probe accuracy results are shown in Table 4.

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

Model	McRae++	Binder
Vision models		
Random SigLIP	7.2	9.3
ViT-MAE	23.3	18.8
Max ViT (IN-1K)	24.9	10.4
Max ViT (IN-21K)	30.2	21.5
Swin-V2	31.4	**23.9**
DINOv2	**33.2**	22.7
Multimodal vision models		
SigLIP	37.0	25.2
PaliGemma	37.1	25.0
CLIP (image)	**37.2**	**25.5**
Language models		
GloVe 840B	29.1	23.3
FastText	30.2	22.9
DeBERTa v3	30.9	**25.5**
CLIP (text)	30.9	21.9
Gemma	**35.9**	**25.5**

Table 4: Average F_1-*selectivity* performance of linear probes for semantic norms on the McRae++ data and concept attribute ratings on the Binder data, corrected (per-probe) for random performance. Note that performance for Binder seems relatively low, because the baseline for F_1-selectivity for evenly matched data is 50.0. The full results can be found on Appendix 8.

The impact of modality. Across the two datasets, the multimodal vision encoders achieve the highest probing performance. However, the text-only Gemma-2B LLM is competitive, especially on the Binder dataset. Amongst the vision-only models, the self-supervised models DINOv2 and Swin-V2 are best, but worse than multimodal vision and the better language encoders.

Effect of training data. CLIP learns representations that are equal quality to SigLIP and PaliGemma for predicting semantic attributes, despite having seen one tenth as much data (400M vs 4–5B). Likewise, for self-supervised vision models, Swin-V2 (14M) and DINOv2 (142M) perform similarly. Interestingly, Swin-V2 outperforms label-supervised ViT-MAE (IN-21K), having been trained on the same dataset, but with a less-informed objective. For the language models, training size also matters to some degree, particularly on McRae++, but it is harder to disentangle the effect of architecture and probing methodology (see Appendix B).

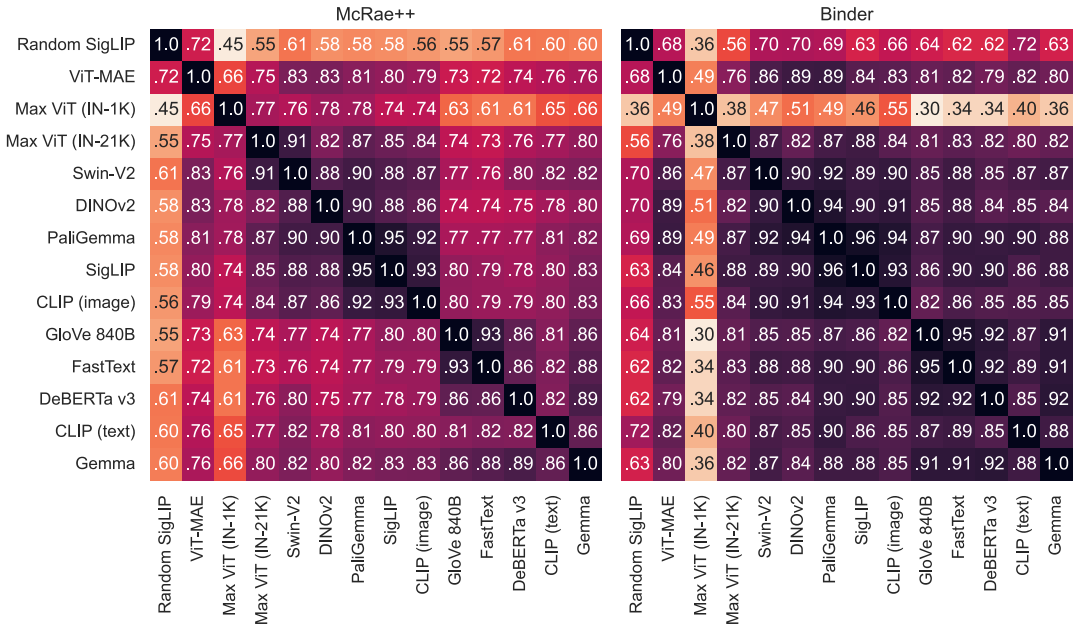

Figure 2: Pearson correlation between models based on attribute performance on the McRae++ and Binder datasets.

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 (Figure 2). For the McRae++ norms, and to a lesser extent on 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 language encoders. Correlations are quite high overall, however, indicating that all encoders across modalities are rather similar. Figure 3 visualizes norm prediction performance of specific pairs of models (vision-only DINOv2 vs text-only Gemma, CLIP image vs CLIP text), and qualitative examples can be found in Appendix A.

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 4 presents the McRae++ probing results per attribute type. Among the ten types, we see that taxonomic attributes are the easiest to predict, followed by visual-motion, for both vision and language models. The vision models, especially the multimodal models, generally outperform the language models, except Gemma-2B. This makes sense for visual attributes like color, but, surprisingly, this is the case even for "encyclopedic" and "functional" attributes, which should be easier to learn from text than from visual inputs. The "taste" and "sound" attributes seem easier for language models, but there are very few norms in these categories. Results by Binder attribute domain (App. Figs. 6 and 7) differ less across model modalities.

Possible confounds. 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 (App. Table 7) are highest for ViT-MAE (0.455), DINOv2 (0.417), and CLIP-image (0.416), a heterogenous set of models in terms of their linear probing behaviour.

7 Conclusion

This linear probing analysis on two datasets shows that multimodally-trained vision encoders rep-

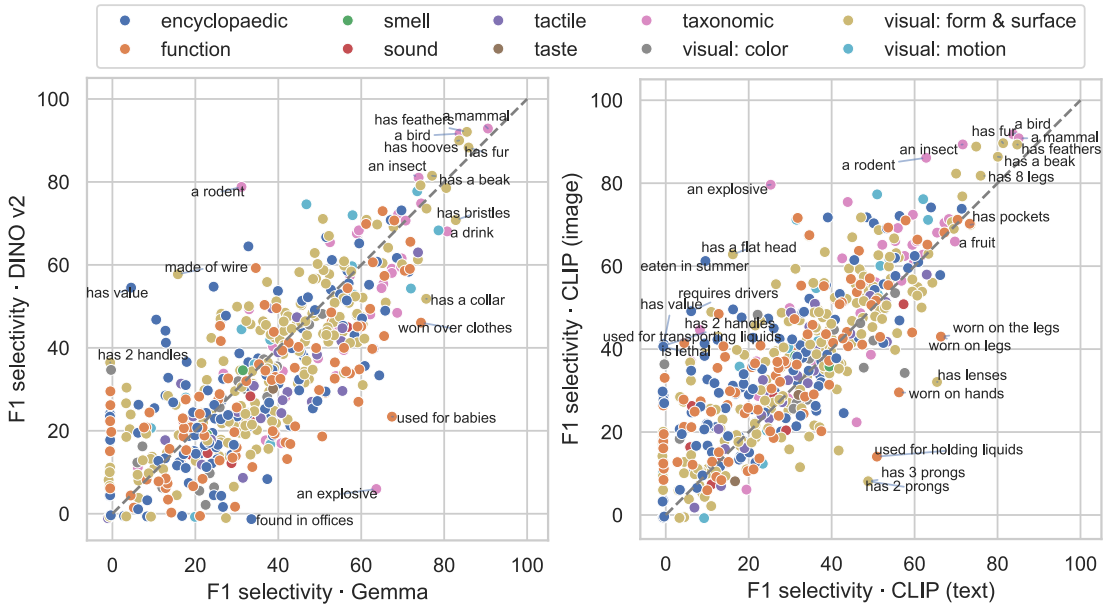

Figure 3: Per feature comparison between pairs of models in terms of the F1 selectivity score. Left: DINO v2 vs Gemma. Right: CLIP (image) vs CLIP (text).

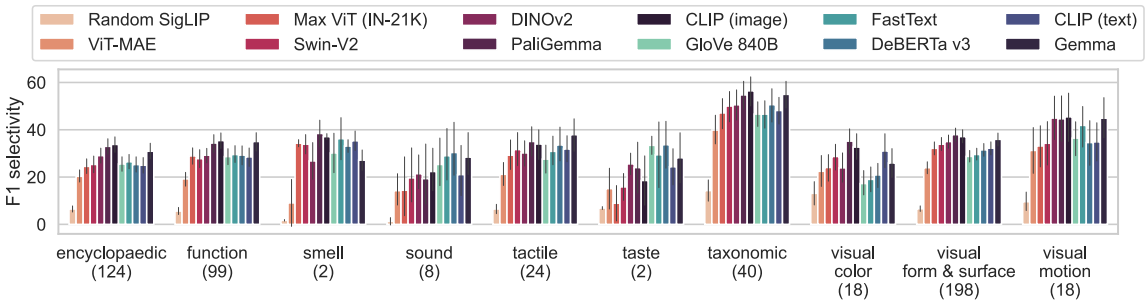

Figure 4: Results (F_1-sel) per attribute (norm) type on the McRae++ data. The number below each type indicates the number of norms belonging to that type. The error bars denote 95% confidence intervals using bootstrapping. Vision models are in reddish colors, while language models are in greenish colors.

resent conceptual attributes better than single-modality encoders. However, the single-modality encoders still perform well. In particular, the self-supervised DINOv2 and Swin-V2 models have learned a comparable amount of conceptual attribute knowledge, which is surprising given that they have not been trained to distinguish between concepts (rather than instances). Intriguingly, label-supervision seems to be harmful for learning human-aligned attributes, judging by the relatively worse performance of Max ViT compared to Swin-V2. For the language models, we find that static embeddings perform well for this kind of concept-level task, particularly given the efforts required to get concept representations from Gemma-2B.

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 better than different-modality models. However, overall correlations are high, indicating a level of convergence. Previous claims of modality convergence have used nearest-neighbors measures (Huh et al., 2024; Li et al., 2024); we show convergent 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 language processing tasks. In future work, we will investigate the predictive power and utility of our probing tasks for multimodal training.

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.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes.

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

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

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

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 *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, and Barry Devereux. 2019. Feature2Vec: Distributional semantic modelling of human property knowledge. *Preprint*, arXiv:1908.11439.

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 *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 260–270, Brussels, Belgium. Association for Computational Linguistics.

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 *International Conference on Learning Representations*.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: Grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*.

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 *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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 *Forty-First International Conference on Machine Learning*.

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 Aleksandrovna Ivanova and Matthias Hofer. 2020. Linguistic Overhypotheses in Category Learning: Explaining the Label Advantage Effect.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. 2024. Do Vision and Language Models Share Concepts? A Vector Space Alignment Study. *Transactions of the Association for Computational Linguistics*, 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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019.

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*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

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. *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 *ICLR*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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 *NeurIPS*.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. 2023. Human alignment of neural network representations. In *International Conference on Learning Representations (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 *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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 *International conference on machine learning*, pages 8748–8763. PMLR.

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 *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 726–730, Beijing, China. Association for Computational Linguistics.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Pia Sommerauer and Antske Fokkens. 2018. Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Brussels, Belgium. Association for Computational Linguistics.

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?

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 *European conference on computer vision*, pages 459–479. Springer.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

A Qualitative Results

In Figure 5 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 (DINO 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 5 are shown for illustrative purposes.

For the `has 4 legs` attribute we observe that the vision models (DINO v2 and CLIP) predict as positive for STOOL or MOLE, but are incorrectly penalised due to the missing positive concepts. For the `is dangerous`, the language models identify predict razor as positive, and are again penalised by an arguably missing annotation.

B 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 5 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

Model	F1 sel.	Five random samples per feature norm and their predictions									
		has 4 legs (visual: form & surface)									
		DOG	+	GOAT	+	STOOL	−	ANTEATER	−	MOLE	−
DINO v2	69.7		✓		✓		✓		✓		•
Gemma	63.8		✓		✓		•		•		•
CLIP (image)	66.7		✓		✓		✓		•		✓
CLIP (text)	55.1		✓		•		•		•		•
		made of wood (visual: form & surface)									
		AXE	+	SKI	+	DYNAMITE	−	LOVESEAT	−	BOW	−
DINO v2	50.9		•		✓		✓		•		•
Gemma	53.6		✓		•		•		•		•
CLIP (image)	51.6		✓		✓		•		•		•
CLIP (text)	47.7		✓		✓		•		✓		•
		is dangerous (encyclopaedic)									
		DYNAMITE	+	AXE	+	RAZOR	−	TUMBLEWEED	−	MOLE	−
DINO v2	25.4		•		•		•		✓		•
Gemma	48.3		✓		✓		✓		•		•
CLIP (image)	37.0		✓		✓		•		•		•
CLIP (text)	37.5		✓		•		✓		•		•
		tastes sweet (taste)									
		PLUM	+	RAISIN	+	WATERMELON	+	PINEAPPLE	+	LAVENDER	−
DINO v2	30.1		✓		✓		•		•		•
Gemma	38.7		✓		•		•		✓		•
CLIP (image)	29.0		✓		✓		✓		•		•
CLIP (text)	32.0		✓		✓		•		•		✓

Figure 5: Five random predictions of linear probes trained on four feature norms. Positive concepts are indicated by +, negative concepts by −. The linear probes are trained on embeddings from one of the four models: DINO v2, Gemma, CLIP image and text encoders. If a model predicts a concept as having the feature norm, 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 feature norms.

					McRae++		
	Model	Input	Seq.	Layer	P	R	F_1
A	Gemma	word	mean	0 (emb)	21.0	10.0	12.8
B	Gemma	word (space)	mean	0 (emb)	31.3	15.6	19.8
C	Gemma	sentences (10)	mean	1	36.8	19.7	24.3
D	Gemma	sentences (10)	mean	18 (last)	41.2	28.2	31.9
E	Gemma	sentences (10)	last	18 (last)	44.5	32.9	36.0
F	Gemma	sentences (10)	mean	0–6	38.5	23.1	27.3
G	Gemma	sentences (10)	mean	0–9	40.0	24.9	29.1
H	Gemma	sentences (10)	mean	9–18	45.8	30.1	34.6
I	Gemma	sentences (10)	last	9–18	48.9	33.7	37.9
J	Gemma	sentences (50)	mean	18 (last)	40.2	28.1	31.5
K	Gemma	sentences (50, constr.)	mean	18 (last)	39.6	27.5	30.9
L	DeBERTa v3	sentences (10)	mean	12 (last)	24.4	21.8	21.9
M	DeBERTa v3	sentences (10)	mean	0–4	41.5	26.3	30.6
N	DeBERTa v3	sentences (10)	mean	0–6	44.1	28.5	32.9
O	GPT2	sentences (10)	mean	12 (last)	27.5	21.5	22.8
P	BERT base uncased	sentences (10)	mean	0–4	29.5	19.4	22.2
Q	BERT base uncased	sentences (10)	mean	0–6	31.4	20.8	23.8

Table 5: 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.

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 6: Precise names of the models used in this paper.

Model	Correlation
Max ViT (IN-21K)	0.268
DeBERTa v3	0.268
Swin-V2	0.306
CLIP (text)	0.316
Max ViT (IN-1K)	0.326
Gemma	0.336
Random SigLIP	0.346
PaliGemma	0.371
FastText	0.374
SigLIP	0.384
GloVe 840B	0.385
CLIP (image)	0.416
DINOv2	0.417
ViT-MAE	0.455

Table 7: McRae++ dataset: 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).

Gemma-2B tokenizer tokenizes single word inputs, <bos>aardvark→{aard, vark}, instead of →{_aard, vark}. In order to extract a (more useful) static embedding, we needed to include a space before the concept token in order to achieve correct tokenization (**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 C 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.

C Collecting Textual Contexts of Concepts using the GPT-4o API

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.) Figures 8 and 9 show the prompts used. The total cost of collecting the sentences was $26.24 and the data will be made publicly available for future research.

D Model Names

For reproducibility, Table 6 shows the precise names of the models used in this paper.

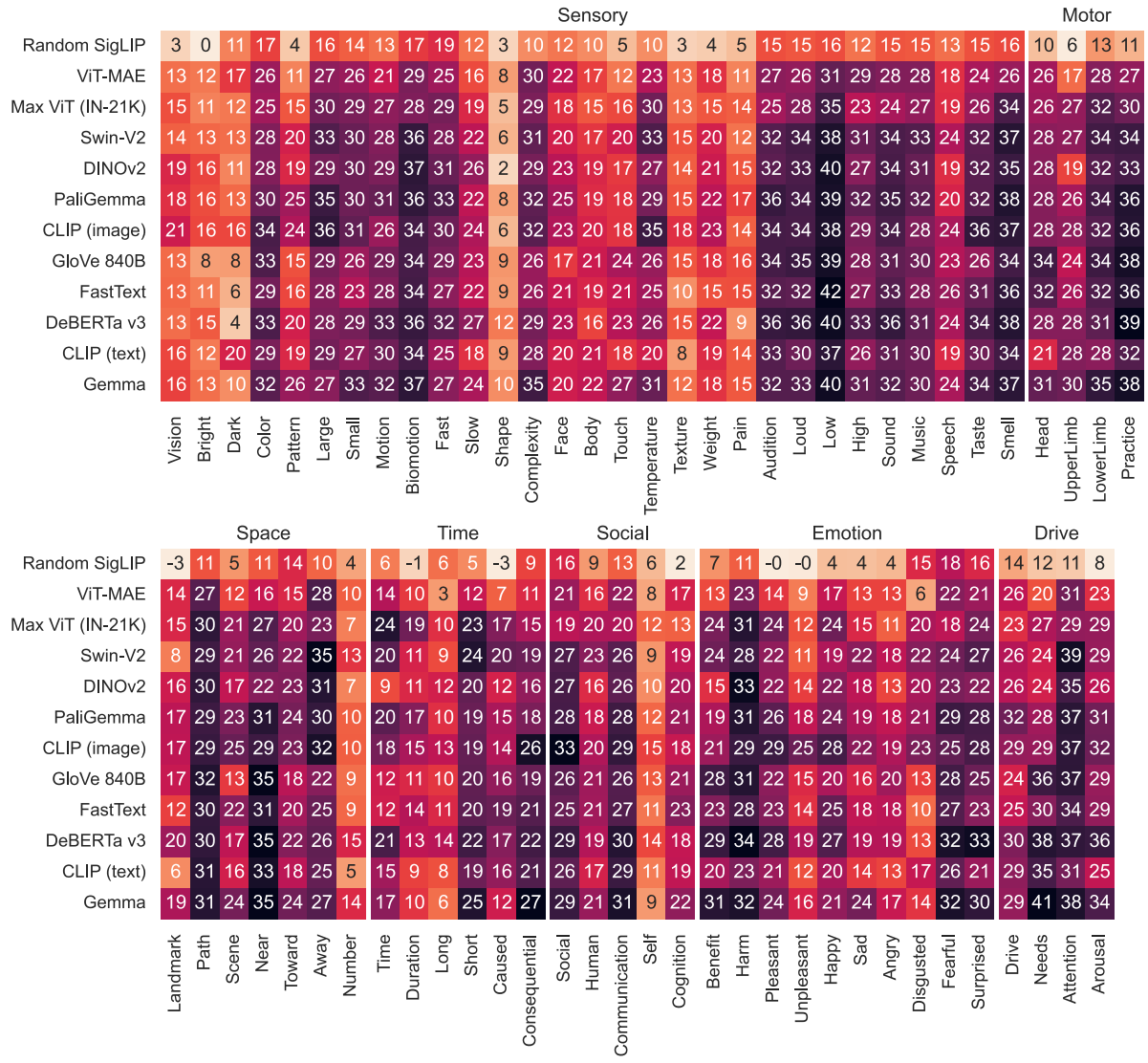

Figure 6: F_1 selectivity for the Binder attribute ratings. Note that raw F_1 score is much higher: the random baseline (against which F_1 selectivity is calculated) is 50% for evenly-distributed data.

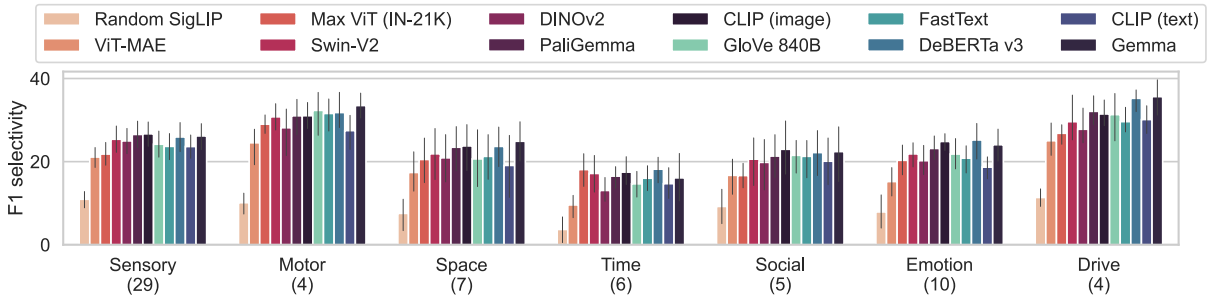

Figure 7: Results (F_1-sel) per attribute domain on the Binder data. The number below each domain indicates the number of attributes belonging to that domain. The error bars denote 95% confidence intervals using bootstrapping. Vision models are in reddish colors, while language models are in greenish colors.

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."

Figure 8: The prompt used to collect textual contexts for each concept in the 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. Try to avoid using the following phrases in any of the sentences: {positive_norms}"

Figure 9: The prompt used to collect constrained textual contexts for each concept in the THINGS dataset. The constraint tried to prevent GPT4o from using the positive norms associated with a concept.

	McRae++				Binder			
Model	P	R	F_1	F_1-sel	P	R	F_1	F_1-sel
Vision models								
Random SigLIP	9.2	10.0	9.2	7.2	60.6	60.3	59.8	9.3
ViT-MAE	28.8	24.7	25.3	23.3	70.0	70.0	69.4	18.8
Max ViT (IN-1K)	28.2	29.2	26.9	24.9	62.2	61.0	61.0	10.4
Max ViT (IN-21K)	42.8	28.0	32.2	30.2	71.6	73.6	72.0	21.5
Swin-V2	**44.5**	29.0	33.4	31.4	**74.8**	**75.2**	**74.5**	**23.9**
DINO v2	41.5	**33.2**	**35.2**	**33.2**	73.8	73.7	73.2	22.7
Multimodal vision models								
SigLIP	48.2	35.6	39.0	37.0	76.8	76.0	75.8	25.2
PaliGemma	**48.5**	35.6	39.1	37.1	76.0	76.1	75.5	25.0
CLIP (image)	46.2	**37.0**	**39.2**	**37.2**	**77.0**	**76.2**	**76.1**	**25.5**
Language models								
GloVe 840B	36.5	29.7	31.1	29.1	74.6	74.1	73.9	23.3
FastText	40.0	29.5	32.3	30.2	74.0	74.1	73.5	22.9
DeBERTa v3	44.1	28.5	32.9	30.9	**77.0**	**76.3**	**76.1**	**25.5**
CLIP (text)	42.8	29.0	32.9	30.9	73.2	72.7	72.5	21.9
Gemma	**48.9**	**33.7**	**37.9**	**35.9**	**77.0**	76.2	**76.1**	**25.5**

Table 8: Detailed results, in terms of precision (P), recall (R), F_1 score and F_1-selectivity score, of concept norm linear probes on the McRae++ and Binder data.