Do Vision and Language Models Share Concepts? A Vector Space Alignment Study

Jiaang Li[†] **Yova Kementchedjhieva**[‡] **Constanza Fierro**[†] **Anders Søgaard**[†]
[†]University of Copenhagen, Denmark
[‡]Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
{jili,c.fierro,soegaard}@di.ku.dk, yova.kementchedjhieva@mbzuai.ac.ae

Abstract

Large-scale pretrained language models (LMs) are said to ''lack the ability to connect utterances to the world'' (Bender and Koller, 2020), because they do not have ''mental models of the world'' (Mitchell and Krakauer, 2023). If so, one would expect LM representations to be unrelated to representations induced by vision models. We present an empirical evaluation across four families of LMs (BERT, GPT-2, OPT, and LLaMA-2) and three vision model architectures (ResNet, SegFormer, and MAE). Our experiments show that LMs partially converge towards representations isomorphic to those of vision models, subject to dispersion, polysemy, and frequency. This has important implications for *both* multi-modal processing and the LM understanding debate (Mitchell and Krakauer, 2023).[1]

1 Introduction

The debate around whether LMs can be said to understand is often portrayed as a back-and-forth between two opposing sides (Mitchell and Krakauer, 2023), but in reality, there are many positions. Some researchers have argued that LMs are ''all syntax, no semantics'', i.e., that they learn form, but not meaning (Searle, 1980; Bender and Koller, 2020; Marcus et al., 2023).[2] Others have argued that LMs have inferential semantics, but not referential semantics (Rapaport, 2002; Sahlgren and Carlsson, 2021; Piantadosi and Hill, 2022),[3] whereas some have posited that a form of externalist referential semantics is possible, at least for chatbots engaged in direct conversation (Cappelen and Dever, 2021; Butlin, 2021; Mollo and Millière, 2023; Mandelkern and Linzen, 2023). Most researchers agree, however, that LMs ''lack the ability to connect utterances to the world'' (Bender and Koller, 2020), because they do not have ''mental models of the world'' (Mitchell and Krakauer, 2023).

This study provides evidence to the contrary: Language models and computer vision models (VMs) are trained on independent data sources (at least for unsupervised computer vision models). The only common source of bias is the world. If LMs and VMs exhibit similarities, it must be because they both model the world. We examine the representations learned by different LMs and VMs by measuring how similar their geometries are. We consistently find that the better the LMs are, the more they induce representations similar to those induced by computer vision models. The similarity between the two spaces is such that from a very small set of parallel examples we are able to linearly project VMs representations to the language space and retrieve highly accurate captions, as shown by the examples in Figure 1.

Contributions. We present a series of evaluations of the vector spaces induced by three families of VMs and four families of LMs, i.e., a total of fourteen VMs and fourteen LMs. We show that

[1]Code and dataset: https://github.com/jiaangli/VLCA.

[2]The idea that computers are ''all syntax, no semantics'' can be traced back to German 17th century philosopher Leibniz's Mill Argument (Lodge and Bobro, 1998). The Mill Argument states that mental states cannot be reduced to physical states, so if the capacity to understand language requires mental states, this capacity cannot be instantiated, merely imitated, by machines. In 1980, Searle introduced an even more popular argument against the possibility of LM understanding, in the form of the so-called Chinese Room thought experiment (Searle, 1980). The Chinese Room presents an interlocutor with no prior knowledge of a foreign language, who receives text messages in this language and follows a rule book to reply to the messages. The interlocutor is Searle's caricature of artificial intelligence, and is obviously, Searle claims, not endowed with meaning or understanding, but merely symbol manipulation.

[3]See Marconi (1997) for this distinction.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1232–1249, 2024. https://doi.org/10.1162/tacl_a_00698
Action Editor: Deyi Xiong. Submission batch: 1/2024; Revision batch: 7/2024; Published 9/2024.

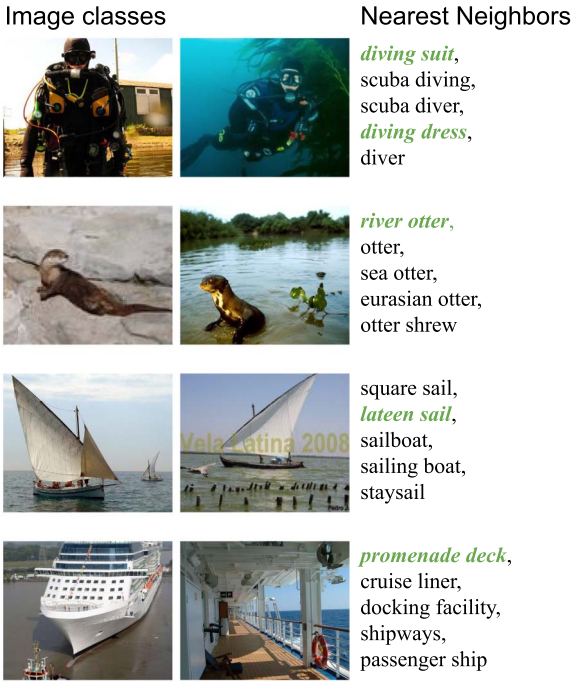

Figure 1: Mapping from MAE_{Huge} (images) to $\text{OPT}_{30\text{B}}$ (text). Gold labels are in green.

within each family, the larger the LMs, the more their vector spaces become structurally similar to those of computer vision models. This enables retrieval of language representations of images (referential semantics) with minimal supervision. Retrieval precision depends on dispersion of image and language, polysemy, and frequency, but consistently improves with language model size. We discuss the implications of the finding that language and computer vision models learn representations with similar geometries.

2 Related Work

Inspiration from Cognitive Science. Computational modeling is a cornerstone of cognitive science in the pursuit for a better understanding of how representations in the brain come about. As such, the field has shown a growing interest in computational representations induced with self-supervised learning (Orhan et al., 2020; Halvagal and Zenke, 2022). Cognitive scientists have also noted how the objectives of supervised language and vision models bear resemblances to predictive processing (Schrimpf et al., 2018; Goldstein et al., 2021; Caucheteux et al., 2022; Li et al., 2023) (but see Antonello and Huth (2022) for a critical discussion of such work).

Studies have looked at the alignability of neural language representations and human brain activations, with more promising results as language models grow better at modeling language (Sassenhagen and Fiebach, 2020; Schrimpf et al., 2021). In these studies, the partial alignability of brain and model representations is interpreted as evidence that brain and models might process language in the same way (Caucheteux and King, 2022).

Cross-modal Alignment. The idea of cross-modal retrieval is not new (Lazaridou et al., 2014), but previously it has mostly been studied with practical considerations in mind. Recently, Merullo et al. (2023) showed that language representations in LMs are *functionally* similar to image representations in VMs, in that a linear transformation applied to an image representation can be used to prompt a language model into producing a relevant caption. We dial back from function and study whether the concept representations converge toward structural similarity (isomorphism). The key question we address is whether despite the lack of explicit grounding, the representations learned by large pretrained language models structurally resemble properties of the physical world as captured by vision models. More related to our work, Huh et al. (2024) proposes a similar hypothesis, although studying it from a different perspective, and our findings corroborate theirs.

3 Methodology

Our primary objective is to compare the representations derived from VMs and LMs and assess their alignability, i.e., the extent to which LMs converge toward VMs' geometries. In the following sections, we introduce the procedures for obtaining the representations and aligning them, with an illustration of our methodology provided in Figure 2.

Vision Models. We include fourteen VMs in our experiments, representing three model families: SegFormer (Xie et al., 2021), MAE (He et al., 2022), and ResNet (He et al., 2016). For all three types of VMs, we only employ the encoder component as a visual feature extractor.[4]

[4]We ran experiments with CLIP (Radford et al., 2021), but report on these separately, since CLIP does not meet the

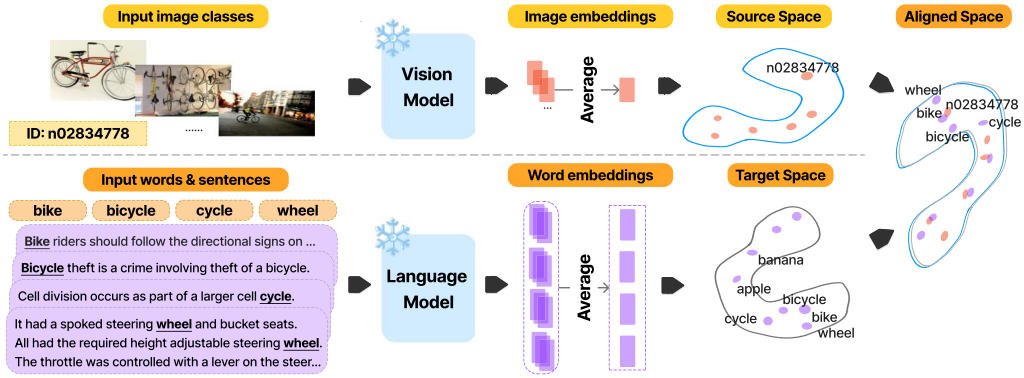

Figure 2: Experiments stages: During our experiments, words, sentences, and images are selected from the aliases list (wordlist and ImageNet-21K aliases), Wikipedia and ImageNet-21K, respectively. The source and target spaces are constructed utilizing image and word embeddings which are extracted by specialized vision and language models.

SegFormer models consist of a Transformer-based encoder and a light-weight feed-forward decoder. They are pretrained on object classification data and finetuned on scene parsing data for scene segmentation and object classification. We hypothesize that the reasoning necessary to perform segmentation in context promotes representations that are more similar to those of LMs, which also operate in a discrete space (a vocabulary). The SegFormer models we use are pretrained with ImageNet-1K (Russakovsky et al., 2015) and finetuned with ADE20K (Zhou et al., 2017).

MAE models relies on a Transformer-based encoder-decoder architecture, with the Vision-Transformer (ViT) (Dosovitskiy et al., 2021) as the encoder backbone. MAE models are trained to reconstruct masked patches in images, i.e., a fully unsupervised training objective, similar to masked language modeling. The encoder takes as input the unmasked image patches, while a lightweight decoder reconstructs the original image from the latent representation of unmasked patches interleaved with mask tokens. The MAE models we use are pretrained on ImageNet-1K.

ResNet models for object classification consist of a bottleneck convolutional neural network with residual blocks as an encoder, with a classification head. They are pretrained on the ImageNet-1K.

Language Models. We include fourteen Transformer-based LMs in our experiments, representing four model families: BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), OPT (Zhang et al., 2022), and LLaMA-2 (Touvron et al., 2023). We use six different sizes of BERT (all uncased): $\text{BERT}_{\text{Base}}$ and $\text{BERT}_{\text{Large}}$, which are pretrained on the BooksCorpus (Zhu et al., 2015) and English Wikipedia (Foundation), and four smaller BERT sizes, distilled from $\text{BERT}_{\text{Large}}$ (Turc et al., 2019). GPT-2, an auto-regressive decoder-only LM, comes in three sizes, pretrained on the WebText dataset (Radford et al., 2019). OPT also comes in three sizes, pretrained on the union of five datasets (Zhang et al., 2022). LLaMA-2 was pretrained on two trillion tokens.

Vision Representations. The visual representation of a concept is obtained by embedding the images available for the concept with a given VM encoder and then averaging these representations. When applying SegFormer, we average the patches' representations from the last hidden state as the basis for every image, whereas we use the penultimate hidden state for MAE models.[5] ResNet models generate a single vector per input image from the average pooling layer.

Language Representations. The LMs included here were trained on text segments, so applying them to words in isolation could result in unpredictable behavior. We therefore represent words by embedding English Wikipedia sentences, using the token representations that form the concept,

criteria of our study, being trained on a mixture of text and images. CLIP results are presented in Appendix C.

[5]We also experimented with utilizing the representations from the last hidden state; however, the results were not as promising as those obtained from the penultimate hidden state. Caron et al. (2021) demonstrate the penultimate-layer features in ViTs trained with DINO exhibit strong correlations with saliency information in the visual input, such as object boundaries and so on.

decontextualizing these representations by averaging across different sentences (Abdou et al., 2021). In the case of masked language models, we employ an averaging approach on the token representations forming the concept; otherwise, we choose for the last token within the concept (Zou et al., 2023).

Linear Projection. Since we are interested in the extent to which vision and language representations are isomorphic, we focus on linear projections.[6] Following Conneau et al. (2018), we use Procrustes analysis (Schönemann, 1966) to align the representations of VMs to those of LMs, given a bimodal dictionary (§ 4.1). Given the VM matrix A (i.e., the visual representations of concepts) and the LM matrix B (i.e., the language representation of the concepts) we use Procrustes analysis to find the orthogonal matrix Ω that most closely maps source space A onto the target space B. Given the constrain of orthogonality the optimization $\Omega = \min_R \|RA - B\|_F$, s.t. $R^T R = I$ has the closed form solution $\Omega = UV^T, U\Sigma V = \text{SVD}(BA^T)$, where SVD stands for singular value decomposition. We induce the alignment from a small set of dictionary pairs, evaluating it on held-out data (§ 4.2). Given the necessity for both the source and target space to have the same dimensionality, we employ principal component analysis (PCA) to reduce the dimensionality of the larger space in cases of a mismatch.[7]

4 Experimental Setup

In this section, we discuss details around bimodal dictionary compilation (§ 4.1), evaluation metrics, as well as our baselines (§ 4.2).

4.1 Bimodal Dictionary Compilation

We build bimodal dictionaries of image-text pairs based on the ImageNet-**21K** dataset (Russakovsky et al., 2015) and the CLDI (cross-lingual dictionary induction) dataset (Hartmann and Søgaard, 2018). In ImageNet, a concept class has a unique ID and is represented by multiple images and one or more names (which we refer to as *aliases*), many of which are multi-word expressions. We filter the data from ImageNet-21K: keeping classes with over 100 images available, aliases that appear at least five times in Wikipedia, and classes with at least one alias. As a result, 11,338 classes and 13,460 aliases meet the criteria. We further filter aliases that are shared by two different class IDs, and aliases for which their hyponyms are already in the aliases set.[8] To avoid *any* form of bias, given that the VMs we experiment with have been pretrained on ImageNet-1K, we report results on ImageNet-21K excluding the concepts in ImageNet-1K (Exclude-1K).

Set	Num. of classes	Num. of aliases	Num. of pairs
Only-1K	491	655	655
Exclude-1K	5,942	7,194	7,194
EN-CLDI	1,690	1,690	1,690

Table 1: Statistics of the bimodal dictionaries.

One important limitation of the Exclude-1K bimodal dictionary is that all concepts are nouns. Therefore, to investigate how our results generalize to other parts of speech (POS), we also use the English subset of CLDI dataset (EN-CLDI), which contains images paired with verbs and adjectives. Each word within this set is unique and paired with at least 22 images. Final statistics of the processed datasets are reported in Table 1.

The pairs in these bimodal dictionaries are split 70-30 for training and testing based on the class IDs to avoid train-test leakage.[9] We compute five such splits at random and report averaged results. See § 6 for the impact of training set size variations.

4.2 Evaluation

We induce a linear mapping Ω based on training image-text pairs sampled from A and B, respectively. We then evaluate how close $A\Omega$ is to B by computing retrieval precision on held-out image-text pairs. To make the retrieval task as challenging as possible, the target space B is expanded with 65,599 words from an English wordlist in addition to 13,460 aliases, resulting in a total of 79,059 aliases in the final target space.

[6] For work on non-linear projection between representation spaces, see Nakashole (2018), Zhao and Gilman (2020), and Glavaš and Vulić (2020).

[7] The variance is retained for most models after dimensionality reduction, except for a few cases where there is some loss of information. The cumulative of explained variance ratios for different models are presented in Table 8.

[8] We obtain the aliases hypernyms and hyponyms from the Princeton WordNet (Fellbaum, 2010).

[9] In the EN-CLDI set, we simply use words to mitigate the risk of train-test leakage.

Metrics. We evaluate alignment in terms of precision-at-k (P@k), a well-established metric employed in the evaluation of multilingual word embeddings (Conneau et al., 2018), with $k \in \{1, 10, 100\}$.[10] Note that this performance metric is much more conservative than other metrics used for similar problems, including pairwise matching accuracy, percentile rank, and Pearson correlation (Minnema and Herbelot, 2019). Pairwise matching accuracy and percentile rank have random baseline scores of 0.5, and they converge in the limit. If a has a percentile rank of p in a list $\mathcal{A}$, it will be higher than a random member of $\mathcal{A}$ p percent of the time. Pearson correlation is monotonically increasing with pairwise matching accuracy, but P@k scores are more conservative than any of them for reasonably small values of k. In our case, our target space is 79,059 words, so it is possible to have P@100 values of 0.0 and yet still have near-perfect pairwise matching accuracy, percentile rank, and Pearson correlation scores. P@k scores also have the advantage that they are intuitive and practically relevant, e.g., for decoding.

Random Retrieval Baseline. Our target space of 79,059 words makes the random retrieval baseline:

$$\text{P@1} = \frac{1}{N}\sum_{i=1}^{N}\frac{n_i}{U} \quad (1)$$

where N represents the total number of image classes; i iterates over each image class; n_i denotes the number of labels for image class i; U refers to the total number of unique aliases. From Equation 1, we get P@1$\approx 0.0015\%$.

Length-frequency Alignment Baseline. The random retrieval baseline tells us how well we can align representations across the two modalities in the absence of any signal (by chance). However, the fact that we can do better than a random baseline, does not, strictly speaking, prove that our models partially converge toward any sophisticated form of modeling the world. Maybe they simply pick up on shallow characteristics shared across the two spaces. One example is frequency: frequent words may refer to frequently depicted objects. Learning what is rare *is* learning about the world, but more is at stake in the debate around whether LMs understand. Or consider length: word length may correlate with the structural complexity of objects (in some way), and maybe this is what drives our alignment precision? To control for such effects, we run a second baseline aligning representations from computer vision models to two-dimensional word representations, representing words by their length and frequency. We collected frequency data based on English Wikipedia using NLTK (Bird et al., 2009) for all aliases within our target space. We use PCA and Procrustes Analysis or ridge regression (Toneva and Wehbe, 2019) to map into the length-frequency space and report the best of those as a second, stronger baseline.

Baseline	P@1	P@10	P@100
Random retrieval	0.0015	0.0153	0.1531
Length-frequency alignment	0.0032	0.0127	0.6053
Non-isomorphic alignment	0.0000	0.0121	0.1105

Table 2: Alignment results for our baselines. All the Precision@k scores are reported in percentage.

Non-isomorphic Alignment Baseline. The former two baselines examine the possibility of aligning representations across two modalities based on chance or shallow signals. While informative, neither strictly demonstrates that a linear projection cannot effectively establish a connection between two non-isomorphic representation spaces, potentially outperforming the random or length-frequency baselines. To rigorously explore this, we disrupt the relationship between words and their corresponding representations by shuffling them. This permutation ensures that the source and target spaces become non-isomorphic. Specifically, we shuffled OPT_{30B} three times at random and report the alignment results between those and original OPT_{30B}, we use the same Procrustes analysis for computing the alignment. Table 2 presents a comparison of the three different baselines. All baselines have P@100 well below 1%. Our mappings between VMs and LMs score much higher (up to 64%), showing the strength of the correlation between the geometries induced by these models with respect to a conservative performance metric.

[10]For example, we could use the mapping of the image of an apple into the word 'apple', and the mapping of the image of a banana into the word 'banana', as training pairs to induce a mapping Ω. If Ω then maps the image of a lemon onto the word 'lemon' as its nearest neighbor, we say that the precision-at-one for this mapping is 100%. If two targ et aliases were listed in the bimodal dictionary for the source image, mapping the image onto either of them would result in P@1 $= 100\%$.

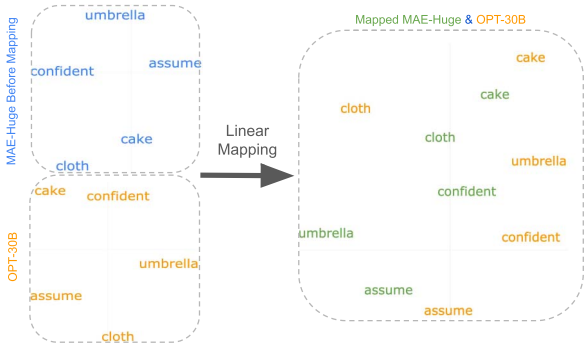

Figure 3: t-SNE plot of 5 words mapped from MAE_{Huge} (blue) to $\text{OPT}_{30\text{B}}$ (orange) using Procrustes analysis. The green represent the mapped MAE_{Huge} embeddings.

5 Results

Similarities between visual and textual representations and how they are recovered through Procrustes analysis are visualized through t-SNE in Figure 3. Our main results for nine VMs and all LMs are presented in Figure 4. The best P@100 scores are around 64%, with baseline scores lower than 1% (Table 2). In general, even the smallest language models outperform the baselines by orders of magnitude. We focus mainly on P@10 and P@100 scores because P@1 only allows one surface form to express a visual concept, but in reality, an artifact such as a vehicle may be denoted by many lexemes (car, automobile, SUV, etc.), each of which may have multiple inflections and derivations (car, cars, car's, etc.). Figure 5 shows examples where the top predictions seem ''as good'' as the gold standard. We find that a region of 10 neighbours corresponds roughly to grammatical forms or synonyms, and a neighbourhood of 100 word forms corresponds roughly to coarse-grained semantic classes. Results of P@10 in Figure 4, show that up to one in five of all visual concepts were mapped to the correct region of the language space, with only a slight deviation from the specific surface form. Considering P@100, we see that more than two thirds of the visual concepts find a semantic match in the language space when using ResNet152 and OPT or LLaMA-2, for example. We see that ResNet models score highest overall, followed by SegFormers, while MAE models rank third. We presume that this ranking is the result, in part, of the model's training objectives: Object classification may induce a weak category-informed bias into the ResNet encoders. In this sense, the performance of MAE models, which are fully unsupervised, presents the strongest evidence for the alignability of vision and language spaces—the signal seemingly could not come from anywhere else but the intrinsic properties of the visual world encoded by the models.

In Figures 4 and 6, we see a clear trend: As model size increases, structural similarity to VMs goes up. The correlation between VM size and similarity is slightly weaker. Specifically, ResNet152 and $\text{OPT}_{30\text{B}}$ obtain the best results, with a P@100 of 64.1%, i.e., 6/10 visual concepts are mapped onto a small neighborhood of 100 words—out of total set of 79,059 candidate words. Around 1/3 images are correctly mapped onto neighborhoods of 10 words, and about 1/20, onto exactly the right word. The scaling effect seems log-linear, with no observed saturation.[11]

6 Analysis

Here, we test whether our findings extend to different parts of speech, as well as how alignment precision is influenced by factors such as dispersion, polysemy, and frequency, or by the size of the seed used for Procrustes analysis. For all experiments in § 6, we use the largest model per model family for both VMs and LMs and the Exclude-1K bimodal dictionary, unless stated otherwise.

Part of Speech. ImageNet mostly contains nouns. To measure the generalization of the learned mapping to other parts of speech, we train and/or test it on adjectives, verbs, and nouns from EN-CLDI (Hartmann and Søgaard, 2018). The target space (language) in this dataset consists of 1690 concepts filtered from 79,059 concepts, which leads to a P@100 baseline below 8%. We consider two settings: (1) to evaluate whether our approach is robust across parts of speech we use concepts of all POS (nouns, adjectives, and verbs) as training and evaluation data; and (2) to evaluate whether the mapping learned by nouns generalizes to other POS, we use *only nouns* as training data and evaluate on adjectives and verbs. The results are in Table 3, along with bimodal pairs counts. In the first experiment the results are strong, for instance ResNet152 and $\text{OPT}_{30\text{B}}$ lead to a P@100

[11] We also investigate the effects of incorporating text signals during vision pretraining by comparing pure vision models against selected CLIP vision encoders. The findings are unsurprising—more details are presented in Appendix C.

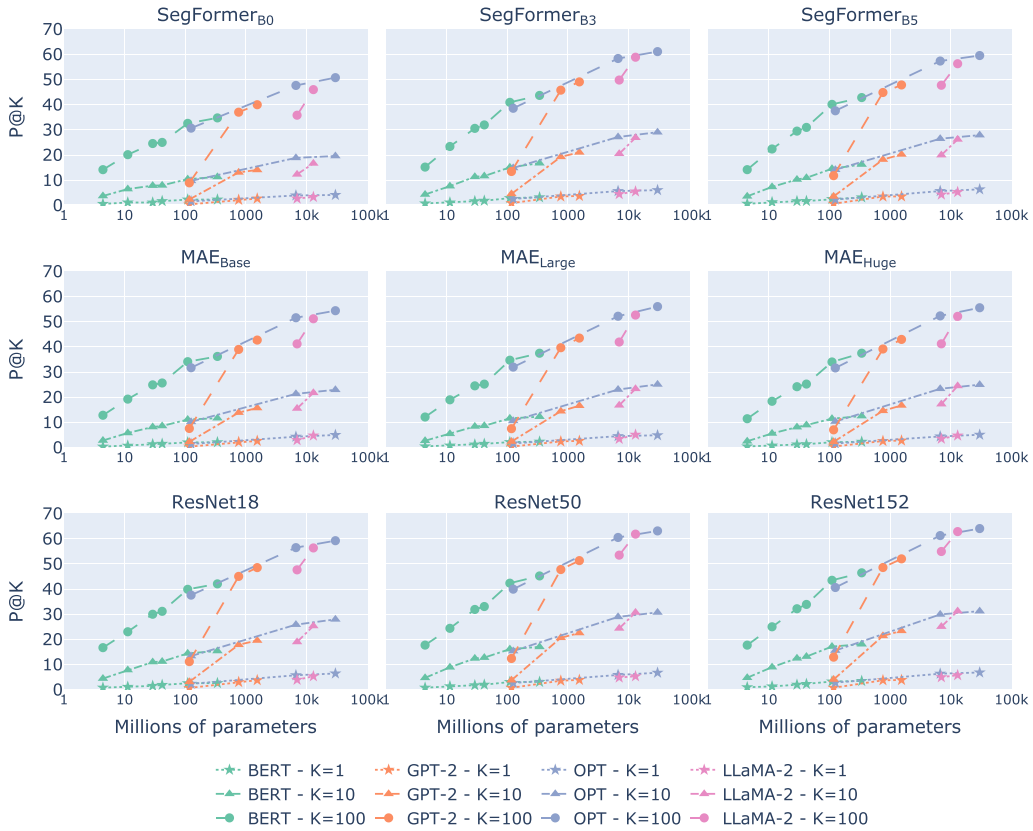

Figure 4: LMs converge toward the geometry of visual models as they grow larger on Exclude-1K set.

Figure 5: Examples featuring the 100 nearest neighbors in the mapping of image classes into the language representation space (from MAE$_{Huge}$ to OPT$_{30B}$). The golden labels are highlighted in green.

of 81.9%. Regarding the second setting, the numbers are lower but still well above our baseline, suggesting that shared concepts between VMs and LMs extend well beyond nouns.

Image Dispersion. Image dispersion is calculated by averaging the pair-wise cosine distance between all images associated with a concept (Kiela et al., 2015). We partition all concepts within the bimodal dictionary into three equally-sized distributed bins (low, medium, high) based on their dispersion. Subsequently, we classify the held-out concepts into these bins and present the results in Table 4. Results

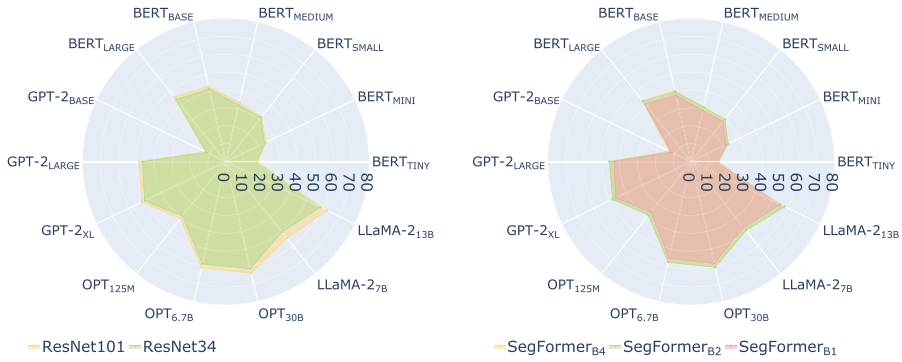

Figure 6: Illustrating the impact of scaling VMs up on Exclude-1K set. The incremental growth in P@100 for scaled-up VMs is marginal, contrasting with the more substantial increase observed when scaling up LMs in the same family.

Models	Train	Test	P@1	P@10	P@100
MAE_{Huge}	Noun 1337	Adj. 157	2.5	15.9	50.9
SF-B5			**3.2**	19.1	53.5
ResNet152			0.6	**22.9**	**57.3**
MAE_{Huge}	Noun 1337	Verb 196	0.5	8.7	49.5
SF-B5			**1.0**	10.7	56.1
ResNet152			0.0	**13.8**	**61.2**
MAE_{Huge}	Mix 1337	Mix 353	6.7	45.6	77.8
SF-B5			**6.9**	48.8	80.1
ResNet152			4.8	**51.7**	**81.9**

Table 3: Evaluation of POS impact on OPT_{30B} and largest VMs across various families, by showing the influence of different POS on EN-CLDI set. ‘‘Mix’’ denotes a combination of all POS categories.

Models	Disp.	Pairs	$BERT_{Large}$	$GPT\text{-}2_{XL}$	$LLaMA\text{-}2_{13B}$	OPT_{30B}
SF-B5	low	703.8	43.2	47.8	**58.7**	**60.6**
	med.	744.4	42.2	**48.7**	56.9	60.4
	high	714.8	**43.3**	47.1	53.3	57.6
MAE_{Huge}	low	847.6	**40.9**	**44.7**	**54.9**	**56.7**
	med.	683.8	37.7	43.6	53.8	55.3
	high	631.6	32.7	40.1	46.7	54.3
ResNet152	low	683.0	**50.6**	**57.6**	**71.3**	**70.5**
	med.	739.8	45.8	51.8	60.8	63.4
	high	740.2	43.1	46.8	56.7	58.4

Table 4: Effect of image dispersion on mapping performance of various LMs and VMs across different levels of image dispersion in terms of P@100 scores on the Exclude-1K set. SF = SegFormer.

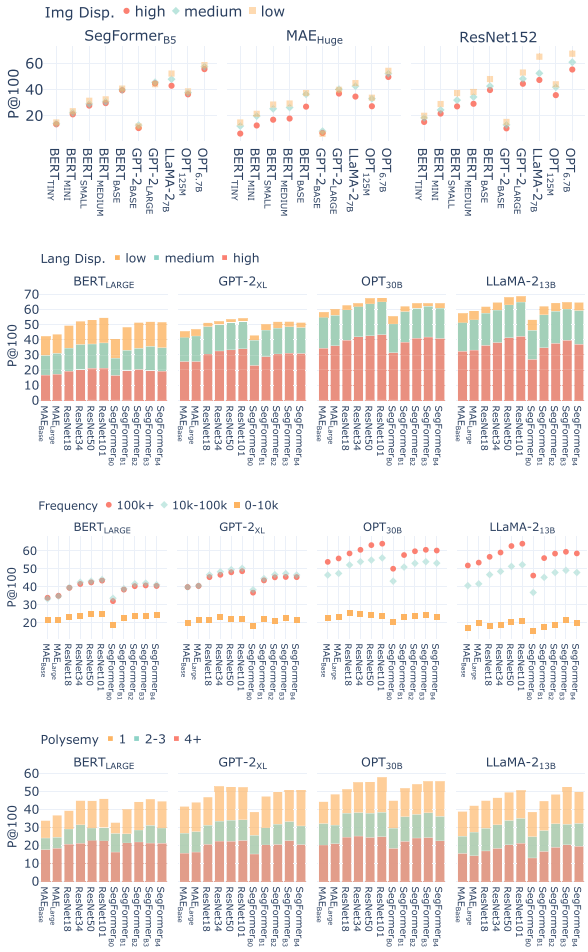

Figure 7: Performance of LMs and VMs across varied levels of (from top to bottom) image dispersion, language dispersion, frequency, and polysemy. Results are presented in terms of P@100 on the Exclude-1K dataset. Img = Image; Lang = Langauge; Disp = Dispersion.

for ResNet152 and MAE_{Huge} show concepts of lower dispersion are easier to align, while results for SegFormer-B5 are mixed. See Figure 7 for the same consistent results observed across the remaining LMs.

Polysemy. Words with multiple meanings may have averaged-out LM representations, and as such we would expect higher polysemy to cause a drop in precision. We obtain polysemy counts from BabelNet (Navigli and Ponzetto, 2012) for the aliases in our bimodal dictionary and measure precision over non-polysemous words, words with two or three meanings, and words with four and

Models	Polysemy	Pairs	MAE_{Huge}	ResNet152	SF-B5	Frequency Rank	Pairs	MAE_{Huge}	ResNet152	SF-B5
$BERT_{Large}$	1	87.0	**38.3**	**47.3**	**42.5**	0–10k	96.6	23.0	25.0	23.7
	2–3	89.0	24.4	29.7	28.1	10k–100k	700.0	34.3	**44.8**	40.1
	4+	106.4	17.4	23.6	19.2	100k+	1366.4	**35.3**	43.3	**40.6**
$GPT\text{-}2_{XL}$	1	87.0	**42.1**	**52.4**	**48.9**	0–10k	96.6	21.0	22.1	22.0
	2–3	89.0	28.0	34.9	32.7	10k–100k	700.0	39.7	**50.4**	**45.4**
	4+	106.4	15.7	22.8	21.0	100k+	1366.4	**40.2**	48.7	44.7
$LLaMA\text{-}2_{13B}$	1	87.0	**40.6**	**51.8**	**48.9**	0–10k	96.6	18.7	21.6	18.1
	2–3	89.0	26.1	35.5	31.3	10k–100k	700.0	40.8	52.3	46.9
	4+	106.4	14.8	21.3	18.1	100k+	1366.4	**53.1**	**64.0**	**57.0**
OPT_{30B}	1	87.0	**47.4**	**57.2**	**55.9**	0–10k	96.6	24.5	24.7	25.4
	2–3	89.0	31.8	39.6	34.4	10k–100k	700.0	47.2	56.1	52.3
	4+	106.4	19.9	25.4	25.2	100k+	1366.4	**55.2**	**64.0**	**59.0**

Table 5: Comparison of different levels of alias polysemy and frequency on mapping performance in terms of P@100 scores on the Exclude-1K set. SF = SegFormer.

more meanings. We ignore aliases not in BabelNet. Precision scores for these three bins are presented in Table 5, alongside counts of the image-text pairs located in each bin. Unlike other results reported in the paper, precision is computed separately for each alias, i.e., if a visual concept is associated with 4 aliases and only three of those appear among the nearest 100 neighbors, the P@100 for this concept will be reported as 75%. This is necessary since different aliases for the same concept may land in different bins. The trend, as expected, is for non-polysemous aliases to yield higher precision scores, regardless of VM and LM. For ResNet152 and the largest LM in our experiments, OPT_{30B}, P@100 is 57.2% for the aliases with a single meaning, but only 25.4% for aliases with four or more meanings. See Figure 7 for the same consistent results observed across the remaining LMs.

Language Dispersion. Since many of the aliases in our bimodal dictionary are *not* in BabelNet, e.g., multi-word expressions, we also consider 'language dispersion' in Wikipedia. This is a proxy for the influence of polysemy and is measured in the same way as image dispersion. The definition and corresponding equation are in Appendix B.2. As seen in Table 6 and Figure 7, we observe a consistent trend across all four LM families, with lower language dispersion correlating with higher alignment precision.

Frequency. We proceed to investigate the influence of word frequency in our study. To gauge this influence, we collect and rank word frequency data from the English Wikipedia for all the unigrams and bigrams, and split them into three distinct frequency bins: the top 10,000 aliases, aliases falling within the word frequency range of 10,000 to 100,000, and aliases ranking beyond 100,000 in terms of word frequency. Then we assess precision for the aliases within our bimodal dictionary across these bins. The precision scores for these three bins are detailed in Table 5, along with the corresponding counts of image-text pairs found within each bin. Our findings reveal a discernible trend where lower-frequency aliases consistently yield higher precision scores, for all VM and LM combinations. For instance, when utilizing ResNet152 and the most substantial LM in our experimentation, OPT_{30B}, P@100 reaches 64.0% for aliases positioned beyond the 100,000 frequency mark, but decreases to 24.7% for aliases within the top 10,000 frequency bin. The results of other frequency experiments for the remaining

Models	Dispersion	Pairs	MAE_{Huge}	ResNet152	SF-B5
$BERT_{Large}$	low	1100.6	**44.1**	**54.5**	**51.2**
	medium	571.8	30.7	38.4	34.1
	high	490.6	16.9	21.8	19.1
$GPT\text{-}2_{XL}$	low	815.2	**46.3**	**53.9**	**50.3**
	medium	768.8	42.0	52.3	47.0
	high	579.0	25.2	34.3	30.7
$LLaMA\text{-}2_{13B}$	low	702.8	**58.7**	**68.5**	**62.3**
	medium	707.8	52.9	65.1	57.3
	high	752.4	32.3	42.1	36.8
OPT_{30B}	low	779.8	**60.4**	**67.6**	**63.2**
	medium	721.2	55.1	65.1	59.7
	high	662.0	35.6	43.7	40.4

Table 6: Effect of language dispersion on mapping performance of various LMs and VMs across different levels of language dispersion in terms of P@100 scores on the Exclude-1K set. SF = SegFormer.

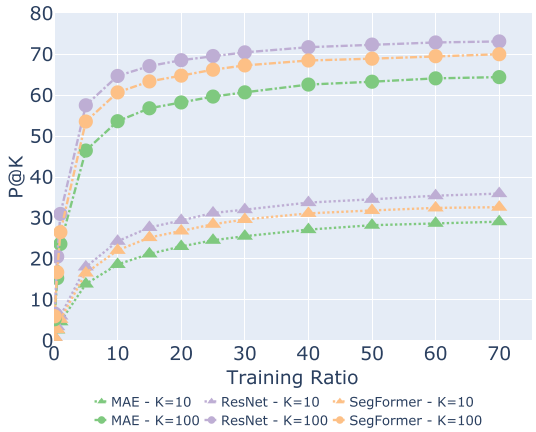

Figure 8: Effect of training data size variation on OPT_{30B}, for the largest vision models from three VM families.

VMs and the largest LM combinations are shown in Figure 7.

Dictionary size. The training data in the bimodal dictionary—used to induce linear projections—consist of thousands of items. Here, we compare the impact of varying the size of the training data, evaluating the different mappings on the same 5,942 concept representations within the Exclude-1K set. The P@100 baseline for these results is below 2%. According to Figure 8, mappings well above this baseline are induced with as few as 297 training pairs (5% of the Exclude-1K dictionary).

In sum, we have found that alignability of LMs and VMs is sensitive to image and language dispersion, polysemy, and frequency. Our alignment precision is nevertheless very high, highlighting the persistence of the structural similarities between LMs and VMs.

7 Discussion

Having established that language and vision models converge towards a similar geometry, we discuss the implications of this finding from various practical and theoretical perspectives within AI-related fields and beyond.

Implications for the LM Understanding Debate. Bender and Koller (2020) have been cited for their thought experiment about an octopus listening in on a two-way dialogue between two humans stuck on deserted islands, but Bender and Koller (2020) also present the following *more constrained* version of their thought experiment:

> . . . imagine training an LM (again, of any type) on English text, again with no associated independent indications of speaker intent. The system is also given access to a very large collection of unlabeled photos, but without any connection between the text and the photos. For the text data, the training task is purely one of predicting form. For the image data, the training task could be anything, so long as it only involves the images. At test time, we present the model with inputs consisting of an utterance and a photograph, like *How many dogs in the picture are jumping?* or *Kim saw this picture and said ''What a cute dog'' What is cute?*

This second thought experiment highlights the importance of relating word representations to representations of what they refer to. It also shows what their argument hinges on: If unsupervised or very weakly supervised alignment of LMs and VMs is possible, their argument fails. The question then is whether such alignment is possible? Bender and Koller deem LMs unable to 'connect their utterances to the world' (or images thereof), because they assume that their representations are unrelated to representations in computer vision models. If the two representations were structurally similar, however, it would take just a simple linear mapping to make proxy inferences about the world and to establish reference. Thought experiments are, in general, fallible intuition pumps (Brendel, 2004), and we believe our results strongly suggest that the second thought experiment of Bender and Koller (2020) is misleading.

Implications for the Study of Emergent Properties. The literature on large-scale, pretrained models has reported seemingly emergent properties (Søgaard et al., 2018; Manning et al., 2020; Garneau et al., 2021; Teehan et al., 2022; Wei et al., 2022), many of which relate to induction of world knowledge. Some have attributed this to memorization, e.g.:

> It is also reasonable to assume that more parameters and more training enable better memorization that could be helpful for tasks requiring world knowledge. (Wei et al., 2022)

while others have speculated if this is an effect of compression dynamics (Søgaard et al., 2018; Garneau et al., 2021). The alignability of different modalities can prove to be a very suitable test bed in the study of emergent properties relating to world knowledge. Our experiments above provide initial data points for the study of such properties.

Implications for Philosophy. Our results have direct implications for two long-standing debates in philosophy: the debate around *strong artificial intelligence* and the so-called *representation wars*. Searle's original Chinese Room argument (Searle, 1980) was an attempt to refute artificial general intelligence, including that it was possible for a machine to *understand* language. Instead, Searle claims, the interlocutor in his experiment is not endowed with meaning or understanding, but mere symbol manipulation. Here we have showed that an interlocutor endowed only with text converges on inducing the same representational geometry as computer vision models with access to visual impressions of the world. In effect, our experiments show that some level of referential semantics (in virtue of internal models of the world) emerges from training on text alone.

The ''representation wars'' (Williams, 2018) refers to a controversy around the role of mental representations in cognitive processes. Cognitive scientists and philosophers aligned with cognitive science often postulate discrete, mental representations to explain observed behavior. Neo-behaviorists, physicalist reductionists, and proponents of embodied cognition have argued to the contrary. To some extent, the positions have softened a bit in recent years. Many now use the term *representation* to mean a state of a cognitive system, i.e., neural network, that responds selectively to certain bodily and environmental conditions (Shea, 2018). This is similar to how the term is used in machine learning and is certainly compatible with neo-behaviorism and physical reductionism. This leaves us with the problem of externalism. Proponents of embodied cognition claim that mental representations give us at best a partial story about how meaning is fixed, because meaning depends on external factors. The meaning of a proper name, for example, depends on causal factors relating the name to an initial baptism. This has also been proposed as a possible account for referential semantics in language models (see Butlin, 2021; Cappelen and Dever, 2021; Mollo and Millière, 2023; Mandelkern and Linzen, 2023). We believe that our experiments suggest a way to reconcile this dispute by accounting for how external factors influence our mental representations, bringing us a bit closer to a solution to this long-standing debate.

Limitations of Our Findings. Our results show that visual concepts can be mapped onto language concepts with high precision when the parameters of the mapping are learned in a supervised fashion. While this experimental setup is sufficient to uncover the structural similarities between visual and language spaces, the argument would be even stronger if we could show that the mapping can be induced in an unsupervised fashion as well, as has been done for cross-lingual embedding spaces (Conneau et al., 2018). We experimented with algorithms for unsupervised embedding alignment (Conneau et al., 2018; Artetxe et al., 2018; Hoshen and Wolf, 2018) and found that all suffer from the degenerate solution problem described in Hartmann et al. (2018), i.e., that no algorithm is currently available that can effectively map between modalities without supervision. We also experimented with initializing the linear transformation in unsupervised algorithms from the one learned with supervision with various amounts of noise added. We found that the unsupervised algorithms were able to recover from the offset up to a certain point, suggesting that with a sufficiently large number of random restarts, unsupervised mapping would be possible. Our experiments demonstrate linear projections can align vision and language representations for concrete noun concepts, as well as, to a lesser extent, for verbs and adjectives. Future models may enable even better linear mappings. Whether some concept subspaces are inherently (linearly) unalignable, remains an open question.

8 Conclusion

In this work, we have studied the question of whether language and computer vision models learn similar representations of the world, despite being trained on independent data from independent modalities. We evaluated the structural similarity of the representations learned by these models for different sizes and architectures. We found that the geometries of these spaces are surprisingly similar, and that similarity increases with

model size. These results seem to challenge the second thought experiment in Bender and Koller (2020). In our experiments, our baseline never goes beyond 1% at P@100, but our linear maps exhibit P@100 scores of up to 64%—an inferential ability which, in our view, strongly suggests the induction of internal world models, something which many previously have deemed impossible. We have discussed various implications, but not all: In the past, researchers have speculated if image representations could act as an interlingua for cross-lingual knowledge transfer (Bergsma and Van Durme, 2011; Kiela and Bottou, 2014; Vulić et al., 2016; Hartmann and Søgaard, 2018). Our results suggest this is viable, and that the quality of such transfer should increase log-linearly with model size.

Acknowledgments

We thank the reviewers and action editors for their invaluable feedback. Special thanks to Serge Belongie and Vésteinn Snæbjarnarson for helpful discussions. Jiaang Li is supported by Carlsberg Research Foundation (grant: CF221432) and the Pioneer Centre 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 *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.conll-1.9

Richard Antonello and Alexander Huth. 2022. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, pages 1–16. https://doi.org/10.1162/nol_a_00087, PubMed: 38645616

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*. https://doi.org/10.18653/v1/P18-1073

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 1764–1769. AAAI Press.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Elke Brendel. 2004. Intuition pumps and the proper use of thought experiments. *Dialectica*, 58(1):89–108. https://doi.org/10.1111/j.1746-8361.2004.tb00293.x

Patrick Butlin. 2021. Sharing our concepts with machines. *Erkenntnis*, pages 1–17. https://doi.org/10.1007/s10670-021-00491-w

Herman Cappelen and Josh Dever. 2021. *Making AI Intelligible: Philosophical Foundations*, New York, USA: Oxford University Press. https://doi.org/10.1093/oso/9780192894724.001.0001

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660. https://doi.org/10.1109/ICCV48922.2021.00951

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022. Long-range and hierarchical language predictions in brains and algorithms. *Nature Human Behaviour*, abs/2111.14232. https://doi.org/10.48550/arXiv.2111.14232

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*. https://doi.org/10

.1038/s42003-022-03036-1, PubMed: 35173264

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR 2018*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net.

Christiane Fellbaum. 2010. Wordnet. *Theory and Applications of Ontology: Computer Applications*. Springer, pages 231–243. https://doi.org/10.1007/978-90-481-8847-5_10

Wikimedia Foundation. Wikimedia downloads.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2021. Analogy training multilingual encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12884–12892. https://doi.org/10.1609/aaai.v35i14.17524

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.675

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Se Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2021. Thinking ahead: Spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv*. https://doi.org/10.1101/2020.12.02.403477

Manu Srinath Halvagal and Friedemann Zenke. 2022. The combination of hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *bioRxiv*. https://doi.org/10.1101/2022.03.17.484712

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2018. Why is unsupervised alignment of English embeddings from different algorithms so hard? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 582–586, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1056

Mareike Hartmann and Anders Søgaard. 2018. Limitations of cross-lingual learning from image search. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 159–163, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3021

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. https://doi.org/10.1109/CVPR52688.2022.01553

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE. https://doi.org/10.1109/CVPR.2016.90

Yedid Hoshen and Lior Wolf. 2018. An iterative closest point method for unsupervised word translation. In *CoRR*, page 1801.06126. https://doi.org/10.18653/v1/D18-1043

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1005

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1015

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1132

Jiaang Li, Antonia Karamolegkou, Yova Kementchedjhieva, Mostafa Abdou, Sune Lehmann, and Anders Søgaard. 2023. Structural similarities between language models and neural response measurements. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*.

Paul Lodge and Marc Bobro. 1998. Stepping back inside Leibniz's mill. *The Monist*, 81(4):553–572. https://doi.org/10.5840/monist199881427

Matthew Mandelkern and Tal Linzen. 2023. Do language models refer? https://doi.org/10.1162/coli_a_00522

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054. https://doi.org/10.1073/pnas.1907367117, PubMed: 32493748

D. Marconi. 1997. *Lexical Competence*. MIT Press, Boston, MA.

Gary Marcus, Evelina Leivada, and Elliot Murphy. 2023. A sentence is worth a thousand pictures: Can large language models understand human language?

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*.

Gosse Minnema and Aurélie Herbelot. 2019. From brain space to distributional space: The perilous journeys of fMRI decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2021

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120. https://doi.org/10.1073/pnas.2215907120, PubMed: 36943882

Dimitri Coelho Mollo and Raphaël Millière. 2023. The vector grounding problem.

Ndapa Nakashole. 2018. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 512–522, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1047

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.

Artificial Intelligence, 193:217–250. https://doi.org/10.1016/j.artint.2012.07.001

Emin Orhan, Vaibhav Gupta, and Brenden M. Lake. 2020. Self-supervised learning through the eyes of a child. In *Advances in Neural Information Processing Systems*, volume 33, pages 9960–9971. Curran Associates, Inc.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Steven Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

William J. Rapaport. 2002. Holism, conceptual-role semantics, and syntactic semantics. *Minds and Machines*, 12(1):3–59. https://doi.org/10.1023/a:1013765011735

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y

Magnus Sahlgren and Fredrik Carlsson. 2021. The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4(682578). https://doi.org/10.3389/frai.2021.682578, PubMed: 34557662

Jona Sassenhagen and Christian J. Fiebach. 2020. Traces of meaning itself: Encoding distributional word vectors in brain activity. *Neurobiology of Language*, 1(1):54–76. https://doi.org/10.1162/nol_a_00003, PubMed: 36794005

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10. https://doi.org/10.1007/BF02289451

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *bioRxiv*. https://doi.org/10.1073/pnas.2105646118, PubMed: 34737231

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*. https://doi.org/10.1101/407007

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424. https://doi.org/10.1017/S0140525X00005756

Nicholas Shea. 2018. *Representation in Cognitive Science*. Oxford University Press. https://doi.org/10.1093/oso/9780198812883.001.0001

Anders Søgaard, Sebastian Ruder, and Ivan Vulíc. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL 2018*. https://doi.org/10.18653/v1/P18-1072

Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 2022. Emergent structures and training dynamics in large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating*

Large Language Models, pages 146–159, virtual+Dublin. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.11

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 188–194, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2031

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Daniel Williams. 2018. Predictive processing and the representation wars. *Minds and Machines*, 28(1):141–172. https://doi.org/10.1007/s11023-017-9441-6, PubMed: 31258246

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090. Curran Associates, Inc.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Jiawei Zhao and Andrew Gilman. 2020. Non-linearity in mapping based cross-lingual word embeddings. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3583–3589, Marseille, France. European Language Resources Association.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. *Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2017.544

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)* pages 19–27. https://doi.org/10.1109/ICCV.2015.11

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Models	Links
$\text{BERT}_{\text{TINY}}$	https://huggingface.co/google/bert_uncased_L-2_H-128_A-2
$\text{BERT}_{\text{MINI}}$	https://huggingface.co/google/bert_uncased_L-4_H-256_A-4
$\text{BERT}_{\text{SMALL}}$	https://huggingface.co/google/bert_uncased_L-4_H-512_A-8
$\text{BERT}_{\text{MEDIUM}}$	https://huggingface.co/google/bert_uncased_L-8_H-512_A-8
$\text{BERT}_{\text{BASE}}$	https://huggingface.co/bert-base-uncased
$\text{BERT}_{\text{LARGE}}$	https://huggingface.co/bert-large-uncased
$\text{GPT-2}_{\text{BASE}}$	https://huggingface.co/openai-community/gpt2
$\text{GPT-2}_{\text{LARGE}}$	https://huggingface.co/openai-community/gpt2-large
GPT-2_{XL}	https://huggingface.co/openai-community/gpt2-xl
OPT_{125M}	https://huggingface.co/facebook/opt-125m
$\text{OPT}_{\text{6.7B}}$	https://huggingface.co/facebook/opt-6.7b
OPT_{30B}	https://huggingface.co/facebook/opt-30b
$\text{LLaMA-2}_{\text{7B}}$	https://huggingface.co/meta-llama/Llama-2-7b
$\text{LLaMA-2}_{\text{13B}}$	https://huggingface.co/meta-llama/Llama-2-13b
SegFormer-B0	https://huggingface.co/nvidia/segformer-b0-finetuned-ade-512-512
SegFormer-B1	https://huggingface.co/nvidia/segformer-b1-finetuned-ade-512-512
SegFormer-B2	https://huggingface.co/nvidia/segformer-b2-finetuned-ade-512-512
SegFormer-B3	https://huggingface.co/nvidia/segformer-b3-finetuned-ade-512-512
SegFormer-B4	https://huggingface.co/nvidia/segformer-b4-finetuned-ade-512-512
SegFormer-B5	https://huggingface.co/nvidia/segformer-b5-finetuned-ade-640-640
MAE_{BASE}	https://huggingface.co/facebook/vit-mae-base
$\text{MAE}_{\text{LARGE}}$	https://huggingface.co/facebook/vit-mae-large
MAE_{HUGE}	https://huggingface.co/facebook/vit-mae-huge
ResNet18	
ResNet34	
ResNet50	https://pypi.org/project/img2vec-pytorch/
ResNet101	
ResNet152	
CLIP-RN50	
CLIP-RN101	
CLIP-RN50*64	https://github.com/openai/CLIP
CLIP-VIT-B-32	
CLIP-VIT-L-14	

Table 7: Sources of models in our experiments.

A Detailed Experimental Settings

A.1 Computational Environment

Our primary software toolkits included HuggingFace Transformers 4.36.2 (Wolf et al., 2020), PyTorch 2.1.2 (Paszke et al., 2019). We ran our experiments on 2 NVIDIA A100s 40G.

A.2 Model Details

Except for the ResNet and CLIP models, all other models used in this study are from the HuggingFace Transformers (Table 7).

B Dispersion Details

B.1 Image Dispersion

The image dispersion d of a concept alias a is defined as the average pairwise cosine distance between all the image representations $i_1, i_2 ... i_n$ in the set of n images for a given alias (Kiela et al., 2015):

$$d(a) = \frac{2}{n(n-1)} \sum_{k<j\leq n} 1 - \frac{i_j \cdot i_k}{|i_j||i_k|}$$

B.2 Language Dispersion

The language dispersion d of a concept alias a is defined as the average pairwise cosine distance between all the corresponding word representations $w_1, w_2 ... w_n$ in the set of n sentences for a given alias:

$$d(a) = \frac{2}{n(n-1)} \sum_{k<j\leq n} 1 - \frac{w_j \cdot w_k}{|w_j||w_k|}$$

Model	Explained Variance Ratio (Sum)						
	256	512	768	1024	1280	2048	Max
MAE_{Huge}	0.9735	0.9922	0.9975	0.9994	1.0000	–	1.0000
ResNet152	0.9795	0.9942	0.9974	0.9987	0.9993	1.0000	1.0000
SegFormer-B5	0.9685	1.0000	–	–	–	–	1.0000
$\text{LLaMA-2}_{\text{13B}}$	0.5708	0.6662	0.7277	0.7725	0.8077	0.8814	1.0000
OPT_{30B}	0.4926	0.6002	0.6664	0.7164	0.7554	0.8360	1.0000

Table 8: The cumulative of explained variance ratios for different models and sizes.

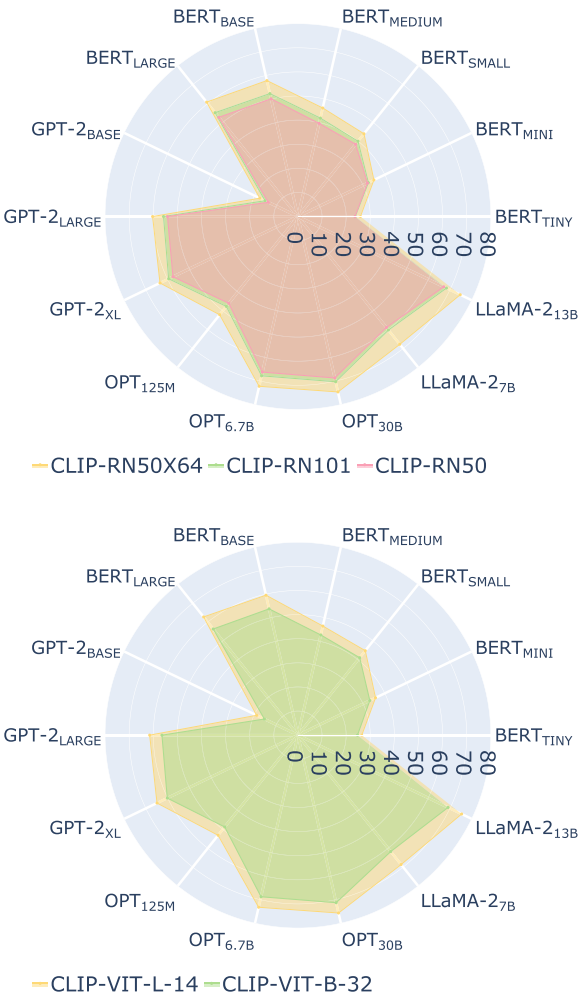

Figure 9: Illustrating the impact of scaling CLIP models up on Exclude-1K set. The incremental growth in P@100 for scaled-up CLIP models is marginal, contrasting with the more substantial increase observed when scaling up LMs in the same family.

C More Results

Cumulative Percentage of Variance Explained. In Table 8, we present the cumulative percentage of variance explained by each selected component after PCA.

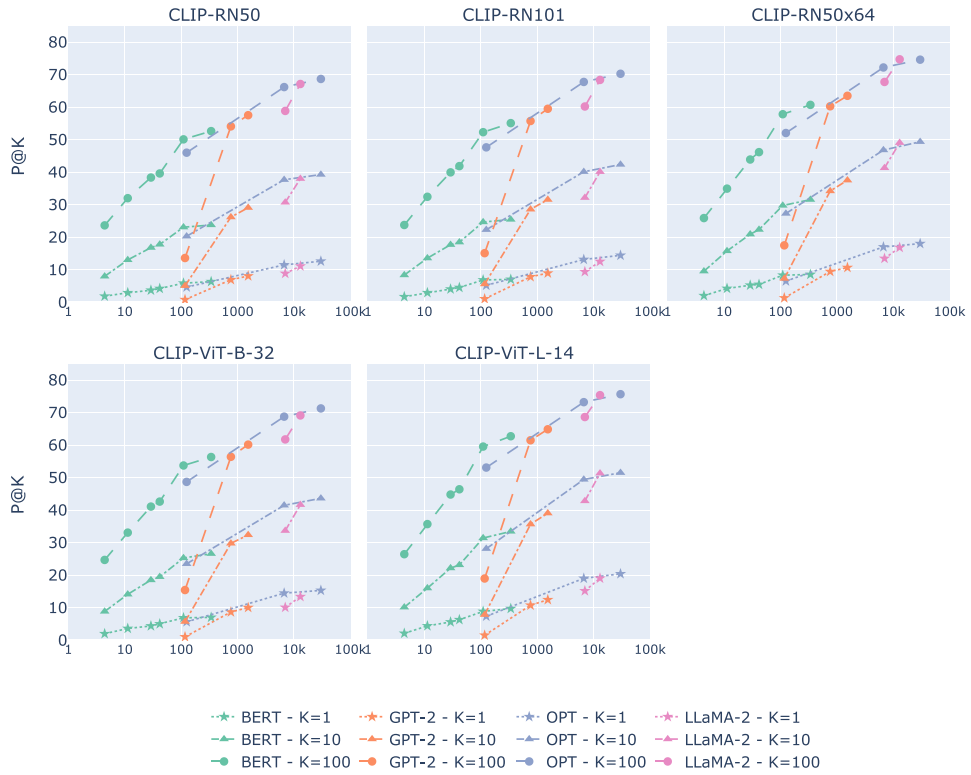

Figure 10: LMs converge toward the geometry of CLIP models as they grow larger on Exclude-1K set.

CLIP Results. We also investigate the effects of incorporating text signals during vision pre-training by comparing pure vision models against selected CLIP (Radford et al., 2021) vision encoders (ResNet50, ResNet101, ResNet50×60, ViT-Base-Patch32, and ViT-Large-Patch14). The results align with our expectations, indicating that the CLIP vision encoders exhibit better alignment with LMs. The findings also support our previous observation that larger LMs tend to demonstrate better alignment. However, it would be unfair to directly compare the results from CLIP with pure vision models, as the pretraining datasets they utilize differ significantly in scale and scope. Detailed results are presented in Figure 9 and Figure 10.

Models	Train	Test	P@1	P@10	P@100
CLIP-ViT-L	Noun	Adj.	**12.7**	**52.2**	**85.4**
CLIP-RN50×64	1337	157	7.0	45.2	84.7
CLIP-ViT-L	Noun	Verb.	**12.2**	**55.1**	**93.9**
CLIP-RN50×64	1337	196	9.2	46.4	89.3
CLIP-ViT-L	Mix	Mix.	**39.1**	**81.0**	**94.1**
CLIP-RN50×64	1337	353	33.7	79.9	93.8

Table 9: Evaluation of POS impact on OPT_{30B} and different CLIP models using EN-CLDI set. ‘‘Mix’’ denotes a combination of all POS categories.

POS Impact on CLIP and OPT. In Table 9, we report the POS impact on OPT_{30B} and two best CLIP vision encoders in our experiments.