arXiv:2405.07987v1 [cs.LG] 13 May 2024

The Platonic Representation Hypothesis

Minyoung Huh [*1] **Brian Cheung** [*1] **Tongzhou Wang** [*1] **Phillip Isola** [*1]

Abstract

We argue that representations in AI models, particularly deep networks, are converging. First, we survey many examples of convergence in the literature: over time and across multiple domains, the ways by which different neural networks represent data are becoming more aligned. Next, we demonstrate convergence across data modalities: as vision models and language models get larger, they measure distance between datapoints in a more and more alike way. We hypothesize that this convergence is driving toward a shared statistical model of reality, akin to Plato's concept of an ideal reality. We term such a representation the *platonic representation* and discuss several possible selective pressures toward it. Finally, we discuss the implications of these trends, their limitations, and counterexamples to our analysis.

Project Page: phillipi.github.io/prh

Code: github.com/minyoungg/platonic-rep

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

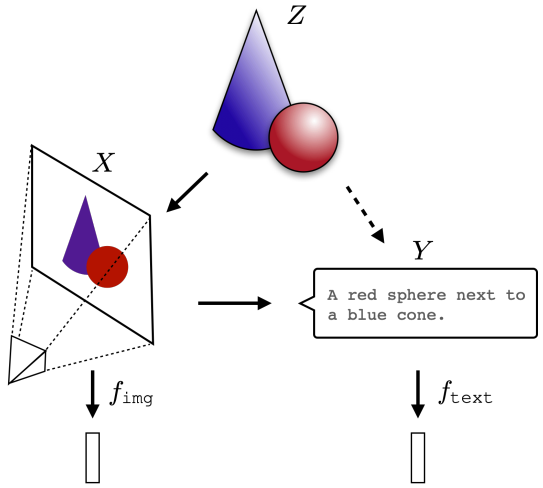


Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.

1. Introduction

AI systems are rapidly evolving into highly multifunctional entities. For example, whereas in the past we had special-purpose solutions for different language processing tasks (*e.g.*, sentiment analysis, parsing, dialogue), modern large language models (LLMs) are competent at all these tasks using a single set of weights (Srivastava et al., 2022). Unified systems are also being built across data modalities: instead of using a different architecture for processing images versus text, recent models, such as GPT4-V (Achiam et al., 2023), Gemini (Anil et al., 2023), and LLaVA (Liu et al., 2023), handle both modalities with a combined architecture. More and more systems are built off of general-purpose pretrained backbones, sometimes called foundation models (Bommasani et al., 2021), that support a large range of tasks, including robotics (Driess et al., 2023; Brohan et al., 2023), bioinformatics (Ma et al., 2024), and healthcare (Steinberg et al., 2021). In short, AI systems are becoming increasingly homogeneous in both their architectures and their capabilities.

This paper explores one aspect of this trend: representational convergence. We argue that there is a growing similarity in how datapoints are represented in different neural network models. This similarity spans across different model architectures, training objectives, and even data modalities.

What has led to this convergence? Will it continue? And ultimately, where does it end?

Our central hypothesis, stated above in Figure 1, is that there is indeed an endpoint to this convergence and a principle that drives it: different models are all trying to arrive at a *representation of reality*, meaning a representation of the

[*]Equal contribution [1]MIT. Correspondence to: Minyoung Huh <minhuh@mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

joint distribution over events in the world that generate the data we observe. Figure 1 conveys this hypothesis: there exists a real world (labeled Z), which we measure with various sensors, such as the camera shown to the left (X). Other *projections* of these measurements, such as the textual description shown, can be produced from the first set of measurements or mediated by some other set of measurements, *e.g.*, touch or other camera views (dotted arrow from X to Y)[1]. Representation learning algorithms find vector embeddings that statistically model the various measurements and projections. The resulting vector embeddings are all derived from the underlying reality in Z and thereby become aligned. As models are trained on more data and for more tasks, they require representations that capture more and more information about Z, and hence alignment toward Z increases toward a convergent point as a function of scale.

We call this converged hypothetical representation the "platonic representation" in reference to Plato's Allegory of the Cave (Plato, c. 375 BC), and his idea of an ideal reality that underlies our sensations. The training data for our algorithms are shadows on the cave wall, yet, we hypothesize, models are recovering ever better representations of the actual world outside the cave. This idea is not unique to Plato; our hypothesis is also related to the notion of "convergent realism" (Newton-Smith, 1981; Putnam, 1982; Doppelt, 2007; Hardin & Rosenberg, 1982) in the philosophy of science (*i.e.*, that science is converging on truth), and to many arguments that have been put forth in the representation learning literature (*e.g.*, Tian et al. (2020a); Zimmermann et al. (2021); Richens & Everitt (2024); Cao & Yamins (2024)).

Also closely related to our hypothesis is the "Anna Karenina scenario" described by Bansal et al. (2021), referring to the possibility that all well-performing neural nets represent the world in the same way. We discuss the evidence they give for this possibility in Section 2.[2] The platonic representation hypothesis refers to the situation where we are in an Anna Karenina scenario *and* the "happy representation" that is converged upon is one that reflects a statistical model of the underlying reality. We discuss the potential nature of this statistical model in more detail in Section 4.

2. Representations are converging

Preliminaries We restrict our attention to representations that are *vector embeddings*. We characterize such a representation by the similarity structure it induces, referred to as its kernel. Kernels are commonly used to assess representations (Kornblith et al., 2019; Klabunde et al., 2023); this can be justified by the fact that they capture the relative structures among data samples, which are also the learning signal for many machine learning algorithms (Aronszajn, 1950; Smola & Schölkopf, 1998). Following prior literature, we define *representational alignment* as a measure of the similarity of the similarity structures induced by two representations, *i.e.*, a similarity metric over kernels. We give the mathematical definition of these concepts below:

- A **representation** is a function $f\colon \mathcal{X} \to \mathbb{R}^n$ that assigns a feature vector to each input in some data domain $\mathcal{X}$.
- A **kernel**, $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, characterizes how a representation measures distance/similarity between datapoints. $K(x_i, x_j) = \langle f(x_i), f(x_j) \rangle$, where $\langle \cdot\,, \cdot \rangle$ denotes inner product, $x_i, x_j \in \mathcal{X}$ and $K \in \mathcal{K}$.
- A **kernel-alignment metric**, $m\colon \mathcal{K} \times \mathcal{K} \to \mathbb{R}$, measures the similarity between two kernels, *i.e.*, how similar is the distance measure induced by one representation to the distance measure induced by another. Examples include Centered Kernel Distance (CKA) (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), and nearest-neighbor metrics (Klabunde et al., 2023).

In our experiments, we use a *mutual nearest-neighbor metric* that measures the mean intersection of the k-nearest neighbor sets induced by two kernels, K_1 and K_2, normalized by k. This metric is a variant of those proposed in Klabunde et al. (2023) and Oron et al. (2017). See Appendix A for the exact definition and Appendix B for comparisons with alternative alignment metrics.

Next, we explore several ways in which representations are converging. First, we argue that different neural networks are converging to aligned representations. Then, we show that this continues to hold across modalities, where image embeddings in vision models align with text embeddings in language models.

2.1. Different models, with different architectures and objectives, can have aligned representations

One indication of representational convergence is the rising number of systems built on top of pre-trained foundation models. These models are becoming standard backbones across a growing spectrum of tasks. Their versatility across numerous applications implies a level of universality in the way they represent data.

While this trend implies convergence toward a relatively small set of foundation models, it does not imply that *different* foundation models will arrive at the same representation. Yet that is what has been observed by several recent papers.

Lenc & Vedaldi (2015) conducted one such study, in which

[1]Touch could convey the shapes in this example but not the colors. This is an important limitation to our hypothesis that we discuss at several points in the paper: different sensors and views might capture different information, which may limit their potential to converge to identical representations.

[2]Borrowed from Tolstoy (1877), similar analogies have been made in other domains, such as the "Anna Karenina principle" popularized by Diamond (1998) to explain animal domestication.

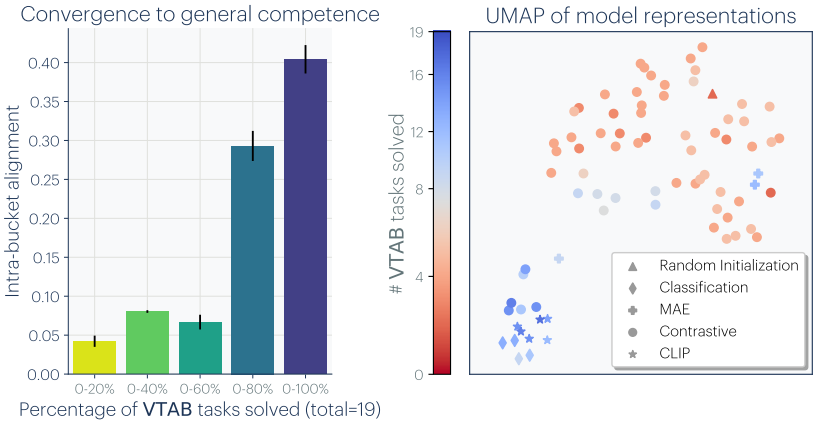


Figure 2. **VISION models converge as COMPETENCE increases:** We measure alignment among 78 models using mutual nearest-neighbors on Places-365 (Zhou et al., 2017), and evaluate their performance on downstream tasks from the Visual Task Adaptation Benchmark (VTAB; Zhai et al. (2019)). **LEFT:** Models that solve more VTAB tasks tend to be more aligned with each other. Error bars show standard error. **RIGHT:** We use UMAP to embed *models* into a 2D space, based on $\texttt{distance} \triangleq -\log(\texttt{alignment})$. More competent and general models (blue) have more similar representations.

they measured representational similarity through a technique called *model stitching*. Given two models, f and g, each composed of multiple layers ($f = f_1 \circ \cdots \circ f_n$, $g = g_1 \circ \cdots \circ g_m$), an intermediate representation from f is integrated into g via a learned affine stitching layer h, resulting in a new stitched model $F = f_1 \circ \cdots \circ f_k \circ h \circ g_{k+1} \circ \cdots \circ g_m$. If F has good performance, it indicates that f and g have compatible representations at layer k, up to the transform h.

In their study, Lenc & Vedaldi (2015) made two notable findings: (1) A vision model trained on ImageNet (Russakovsky et al., 2015) can be aligned with a model trained on the Places-365 dataset (Zhou et al., 2017) while maintaining good performance; (2) The early layers of these convolutional networks are more interchangeable than later layers. The first finding illustrates a level of data independence where distinct image datasets lead to similar representations. The second finding agrees with extensive research that oriented Gabor-like filters are common in both artificial and biological vision systems. This suggests a convergence to a similar initial layer of representation across various neural network architectures (Olshausen & Field, 1996; Krizhevsky et al., 2017). Bansal et al. (2021) expanded on the idea of model stitching, uncovering that models trained using self-supervised objectives align closely with their supervised counterparts.

Moschella et al. (2022) further demonstrated the feasibility of "zero-shot" model stitching without learning a stitching layer. Despite the fact that different text models were trained on different modalities, they found that the models often embed data in remarkably similar ways. In particular, they considered the kernel K defined by learned representations and showed that K serves as a bridge between models, allowing an encoder trained in one language, like English, to work effectively with a decoder in another, like French.

Dravid et al. (2023) extended this idea to individual neurons, and found "Rosetta Neurons" that are activated by the same pattern across a range of vision models. Such neurons form a common dictionary independently discovered by all models.

2.2. Alignment increases with scale and performance

Kornblith et al. (2019) observed model alignment not only exists but also increases with model scale. On CIFAR-10 classification (Krizhevsky et al., 2009), they found that larger models exhibit greater alignment with each other compared to smaller ones. Theoretically, Balestriero & Baraniuk (2018) showed that models with similar outputs (*e.g.*, as a result of having high performance) also have similar internal activations. With the continuing trend of models scaling up, this suggests model alignment will increase over time – we might expect that the next generation of bigger, better models will be even more aligned with each other.

We expand upon this observation by evaluating the transfer performance of 78 vision models. These models were trained with varying architectures, training objectives, and datasets (detailed in Appendix C.1). In Figure 2 (left), we bin these models based on their average transfer performance on the VTAB dataset (Zhai et al., 2019), and then measure the average kernel alignment of the models within each bin. The results indicate that models with high transfer performance form a tightly clustered set of representations, while models with weak performance have more variable representations. We further visualize this structure with UMAP (McInnes et al., 2018) over models representation in Figure 2 (right). This suggests that models that are competent all represent data in a similar way. Echoing Bansal et al. (2021) and Tolstoy (1877), we might say: all strong models are alike, each weak model is weak in its own way.

The discussion so far indicates that various models are aligning toward a unified representation. But does the convergence extend to model weights? While models with different architectures might not have compatible weight spaces, there exists ample evidence that models with the same architecture will often converge to the same basin of weights (Nagarajan & Kolter, 2019; Garipov et al., 2018; Lubana et al., 2023). This holds even for models with different initializations, up to permutations over weight space (Ainsworth et al., 2022). Because of this, it is possible to merge sepa-

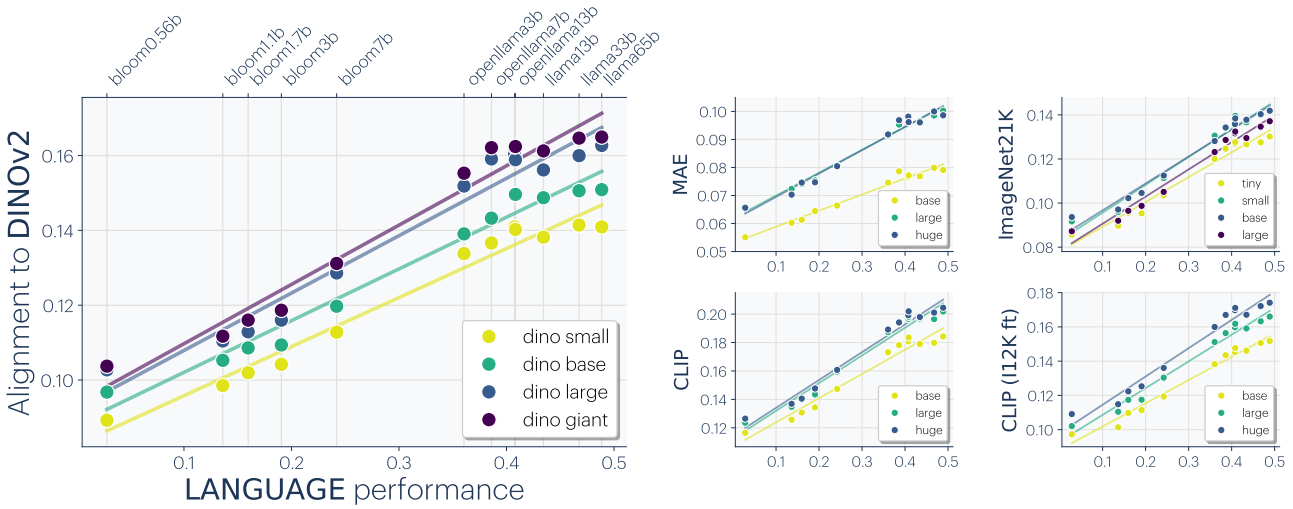


Figure 3. **LANGUAGE and VISION models align:** We measure alignment using mutual nearest-neighbor on the Wikipedia caption dataset (WIT) (Srinivasan et al., 2021). The x-axis is the language model performance measured over 4M tokens from the OpenWebText dataset (Gokaslan & Cohen, 2019) (see Appendix B for plots with model names). We measure performance using $1 - \texttt{bits-per-byte}$, where `bits-per-byte` normalizes the cross-entropy by the total bytes in the input text string. The results show a linear relationship between language-vision alignment and language modeling score, where a general trend is that more capable language models align better with more capable vision models. We find that CLIP models, which are trained with explicit language supervision, exhibit a higher level of alignment. However, this alignment decreases after being fine-tuned on ImageNet classification (labeled CLIP (I12K ft)).

rately trained models of the same architecture, and achieve some of the capabilities of all models in the mixture (Stoica et al., 2023; Jordan et al., 2022; Wortsman et al., 2022).

2.3. Representations are converging across modalities

Do models trained on different data modalities also converge? Several works indicate that the answer is *yes*.

Merullo et al. (2022) extended model stitching to the cross-modal setting, finding that a single linear projection is sufficient to stitch a vision model to an LLM and achieve good performance on visual question answering and image captioning. Koh et al. (2023) showed that linear stitching can also work in the opposite direction, aligning text inputs to visual outputs. In fact, many recent language-vision models stitch pre-trained language and vision models together. For example, LLaVA (Liu et al., 2023) demonstrated state-of-the-art results by projecting visual features into a language model with a 2-layer MLP.

Other works show further kinds of evidence of cross-modal synergy. Achiam et al. (2023) found that jointly training a language model with a vision model improves performance on language tasks, compared to training the language model on its own. Sharma et al. (2024) probed the visual knowledge of LLMs trained *only* on language data, by converting images into code that an LLM can process. They found that LLMs have rich knowledge of visual structures, to the extent that decent visual representations can be trained on images generated solely by querying an LLM to produce code and rendering the response. In visual generation, LLMs show abilities to augment captions with visual structures (*e.g.*, bounding boxes and locations) and improve generation quality (Betker et al., 2023; Lian et al., 2023a;b; Wu et al., 2023). In other modalities, Ngo & Kim (2024) showed auditory models are also roughly aligned with language models up to a linear transformation, and Ng et al. (2023) demonstrated the effectiveness of using pre-trained language models for facial motion prediction.

We set out to address these claims in a broader scope to determine whether models are indeed learning an increasingly modality-agnostic representation of the world. We sampled a variety of models trained either solely on vision or solely on language, and compared their representations as they became larger and more competent over many tasks.

In Figure 3, we assess alignment between a suite of language models and vision models. So far we have only defined alignment for two kernels defined over the same input space. To measure cross-modal alignment, we use paired datasets to bridge the two modalities. For vision and text, we use the Wikipedia captions dataset $\{(x_i, y_i)\}_i$ (Srinivasan et al., 2021), composed of images from Wikipedia (x_i) and their corresponding captions (y_i). We then measure alignment between a language model $f_{\texttt{text}}$ and a vision model $f_{\texttt{img}}$ as the alignment of the two following kernels:

$$K_{\texttt{img}}(i, j) = \langle f_{\texttt{img}}(x_i), f_{\texttt{img}}(x_j) \rangle \quad (1)$$

$$K_{\texttt{text}}(i, j) = \langle f_{\texttt{text}}(y_i), f_{\texttt{text}}(y_j) \rangle. \quad (2)$$

Using this analysis, we find that the better an LLM is at language modeling, the more it tends to aligns with vision

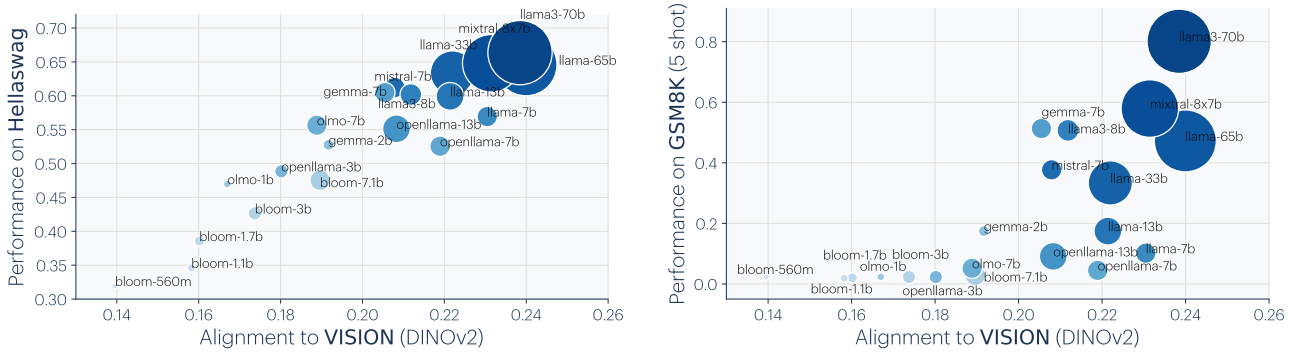


Figure 4. **Alignment predicts downstream performance:** We visualize correlation between LLM alignment score to DINOv2 (Oquab et al., 2023) and downstream task performance on Hellaswag (common-sense) (Zellers et al., 2019) and GSM8K (math) (Cobbe et al., 2021). LLMs are plotted with radii proportional to the size of the model, and color-coded by their rank order in language modeling scores ($1 - \texttt{bits-per-byte}$). We observe that models aligned more closely with vision also show better performance on downstream language tasks. For Hellaswag, there is a linear relationship with alignment score, while GSM8K exhibits an "emergence"-esque trend.

models, as shown in Figure 3. The converse effect also holds: the better a vision models is, the more it tends to align with LLMs. See Appendix C.2 for more details.

2.4. Models are increasingly aligning to brains

Neural networks also show substantial alignment with biological representations in the brain (Yamins et al., 2014). This commonality may be due to similarities in the task and data constraints both systems are confronted with. Even though the mediums may differ – silicon transistors versus biological neurons – the fundamental problem faced by brains and machines is the same: efficiently extracting and understanding the underlying structure in images, text, sounds, etc (Barlow et al., 1961; Olshausen & Field, 1997). The tasks that the human visual system has been honed to perform through evolution – like segmentation, detection, and whole-image classification – are also the ones that we train our neural nets to perform. Yamins et al. (2014) went as far as to title their work in the spirit that performance over such tasks implies brain alignment. Further, Conwell et al. (2022) showed that training data plays a large role in alignment. Psychophysical studies have also shown agreement between how humans perceive visual similarity and how models do, even when the models are trained on tasks, such as self-supervised prediction, that are seemingly unrelated to mimicking human perception (Zhang et al., 2018).

2.5. Does alignment predict downstream performance?

If models are converging towards a more accurate representation of reality, we expect that alignment should correspond to improved performance on downstream tasks. Figure 4 supports this hypothesis by demonstrating improved performance on commonsense reasoning (Hellaswag; Zellers et al. (2019)) and mathematical problem solving (GSM8K; Cobbe et al. (2021)) as alignment improves.

3. Why are representations converging?

Modern machine learning models are generally trained to minimize the empirical risk with possible implicit and/or explicit regularization:

$$\underbrace{f^*}_{\text{trained model}} = \arg\min_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x \sim \text{dataset}} [\underbrace{\mathcal{L}}_{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$

In the following sections, we lay out how each colored component in this optimization process potentially plays a role in facilitating representational convergence.

3.1. Convergence via Task Generality

Each training datapoint and objective (task) places an additional constraint on the model. As data and tasks scale, the volume of representations that satisfy these constraints must proportionately grow smaller, as visualized in Figure 6 and stated below:

The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

This has been previously termed as the Contravariance principle by Cao & Yamins (2024), which states that the set of solutions to an easy goal is large, while the set of solutions to a challenging goal is comparatively smaller. Moreover, we argue that this narrower solution set also generalizes better. As data scales, models that optimize the empirical risk $\mathbb{E}_{x \sim \text{dataset}}[\mathcal{L}(f, x)]$ also improve on the population risk $\mathbb{E}_{x \sim \text{reality}}[\mathcal{L}(f, x)]$, and become better at capturing statistical structures of the true data generating process (reality).

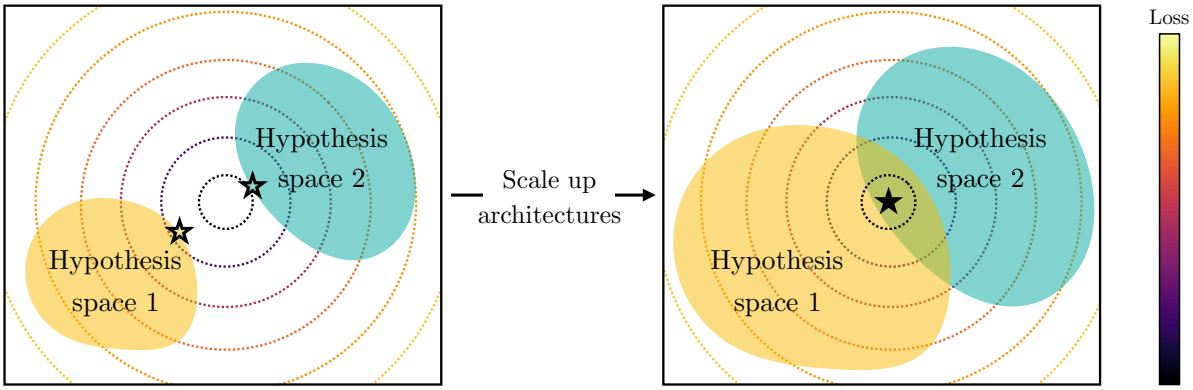


Figure 5. **The Capacity Hypothesis:** If an optimal representation exists in function space, larger hypothesis spaces are more likely to cover it. **LEFT:** Two small models might not cover the optimum and thus find *different* solutions (marked by outlined ☆). **RIGHT:** As the models become larger, they cover the optimum and converge to the same solution (marked by filled ★).

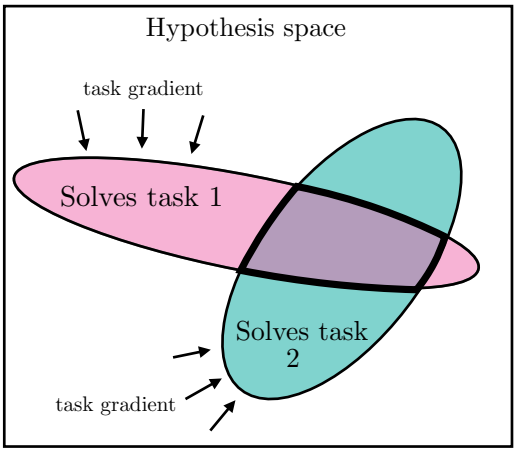


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

Recent work has demonstrated a power law relationship between data scale and model performance (Hestness et al., 2017). This implies that with enough data (*e.g.*, consisting of the entire internet and all offline scientific measurements) one ought to converge to a very small solution set with irreducible error – the inherent epistemic uncertainty of the world. As more models are trained on internet-scale data, the set of solutions that satisfies all data constraints must become relatively small.

In addition to data-scaling, many modern representation learning objectives $\mathcal{L}(f, x)$ directly optimize for multi-task solving. Contrastive learning finds a distance structure over data samples that optimizes many classification tasks (Arora et al., 2019b; Wang & Isola, 2020; Tian et al., 2020b). Masked Autoencoders (He et al., 2021) optimize randomly sampled reconstruction tasks. In fact, autoregressive language modeling can also be seen as optimizing a diverse set of tasks (Radford et al., 2019). Such multi-task objectives may be more effective than single-task ones (*e.g.*, ImageNet classification) due to the fact that they impose more task constraints on the representation, leading to a smaller and higher-quality solution space (Chen et al., 2020; He et al., 2020; Radford et al., 2017; 2019).

3.2. Convergence via Model Capacity

Suppose there is a globally optimal representation for standard learning objectives. Then, under sufficient data, *scaling* a model (*i.e.*, using larger function classes $\mathcal{F}$), as well as improved optimization, should be more effective at finding better approximations to this optimum, as illustrated in Figure 5. With the same training objective, larger models, even of different architectures, will thus tend to converge toward this optimum. When different training objectives share similar minimizers, larger models are better at finding these minimizers, and will train to similar solutions over the training tasks. We summarize this hypothesis as follows:

The Capacity Hypothesis

Bigger models are more likely to converge to a shared representation than smaller models.

3.3. Convergence via Simplicity Bias

Arriving at the same mapping on the *training data* does not prohibit the models from developing distinct internal representations. It is not unreasonable to posit that the representations used to detect a dog in a 1M parameter model could be quite different than that used by a 1B parameter model. What would stop a billion-parameter (and counting) model from learning an overly complicated and distinct representation? One key factor might be simplicity bias:

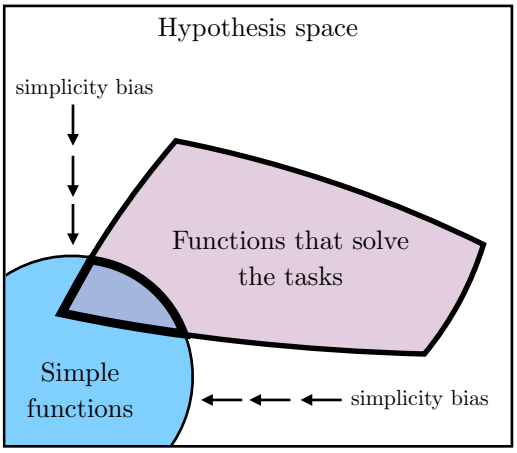


Figure 7. **The Simplicity Bias Hypothesis:** Larger models have larger coverage of all possible ways to fit the same data. However, the implicit simplicity biases of deep networks encourage larger models to find the simplest of these solutions.

The Simplicity Bias Hypothesis

Deep networks are biased toward finding simple fits to the data, and the bigger the model, the stronger the bias. Therefore, as models get bigger, we should expect convergence to a smaller solution space.

Such simplicity bias could be coming from explicit regularization $\mathcal{R}(f)$ commonly used in deep learning (*e.g.*, weight decay and dropout). However, even in the absence of external influences, deep networks naturally adhere to Occam's razor, implicitly favoring simple solutions that fit the data (Solomonoff, 1964; Gunasekar et al., 2018; Arora et al., 2019a; Valle-Perez et al., 2019; Huh et al., 2023; Dingle et al., 2018; Goldblum et al., 2023). Figure 7 visualizes how simplicity bias can drive convergence.

4. What representation are we converging to?

By now, we hope to have convinced the reader that task and data pressures, combined with increasing model capacity, can lead to convergence. We next turn our attention to *what* exactly is the endpoint of all this convergence.

Our central hypothesis, stated in Figure 1, is that the representation we are converging toward is a statistical model of the underlying reality that generates our observations. Consistent with the multitask scaling hypothesis, such a representation would naturally be useful toward many tasks (or at least toward any task grounded in reality). Additionally, this representation might be relatively simple, assuming that scientists are correct in suggesting that the fundamental laws of nature are indeed simple functions (Gell-Mann, 1995), in line with the simplicity bias hypothesis.

But what exactly do we mean by "a statistical model of the underlying reality." In this section, we formalize one definition with concrete mathematical statements. *Importantly*, this section should be read as just one concrete candidate for the form of the platonic representation; other candidates could be arrived at from other modeling assumptions.

4.1. An idealized world

We consider a world that works as follows, consistent with the cartoon in Figure 1. The world consists of a sequence of T discrete events, denoted as $\mathbf{Z} \triangleq [z_1, \ldots, z_T]$, sampled from some unknown distribution $\mathbb{P}(\mathbf{Z})$. Each event can be observed in various ways. An observation is a bijective, deterministic function $\mathsf{obs} : \mathcal{Z} \to \cdot$ that maps events to an arbitrary measurement space, such as pixels, sounds, mass, force, torque, words, etc. Later, in Section 6, we discuss limitations and potential extensions to continuous and unbounded worlds, and stochastic observations, that could yield a model that better reflects real learning scenarios.

One can think of an event as corresponding to the state of the world at some point in time[3], but it is also fine to simply consider an event as any variable that indexes observations, with no further physical meaning[4].

In this idealized world, knowing $\mathbb{P}(\mathbf{Z})$ would be useful for many kinds of predictions; this would constitute a world model over the events that cause our observations (Werbos, 1987; Ha & Schmidhuber, 2018; Richens & Everitt, 2024). We will next show that a particular representation of $\mathbb{P}(\mathbf{Z})$ is recovered by certain contrastive learners.

4.2. A family of contrastive learners converge to a representation of $\mathbb{P}(\mathbf{Z})$

Consider a contrastive learner that models observations that *cooccur* together. For simplicity, we ground our discussion with the following definition of the *cooccurrence probability*, P_{coor}, of two observations x_a and x_b both occurring within some window T_{window}:

$$P_{\mathsf{coor}}(x_a, x_b) \quad \propto \sum_{(t,t') \,:\, |t-t'| \leq T_{\mathsf{window}}} \mathbb{P}(X_t = x_a, X_{t'} = x_b).$$

Analogously, we can define P_{coor} for $\mathbf{Z}$ and other observation modalities. Note that P_{coor} is symmetric.

Consider *positive pairs* as two observations nearby in time (sampled from P_{coor}) and *negative pairs* as observations

[3]Here we only analyze temporal sequences, but note that the same could be done with respect to events laid out in space instead.

[4]This latter interpretation may be more consistent with Plato's intent. Scholars have argued that his allegory of the cave rejects any notion of a true world state (Nettleship, 1897). Instead, we could say that the joint distribution of observation indices is *itself* the platonic reality.

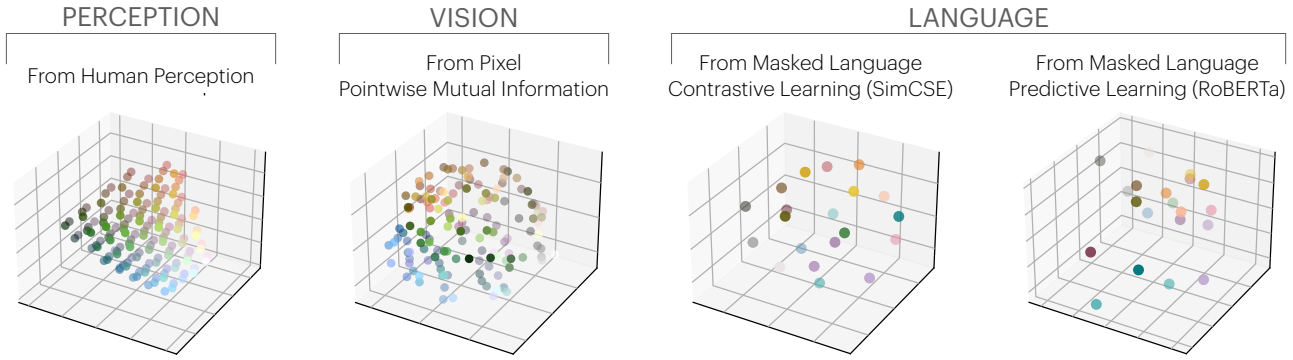


Figure 8. **Color cooccurrence in VISION and LANGUAGE yields perceptual organization:** Similar representations of color are obtained via, **from LEFT to RIGHT**, the perceptual layout from CIELAB color space, cooccurrence in CIFAR-10 images, and language cooccurrence modeling (Gao et al. (2021); Liu et al. (2019); computed roughly following Abdou et al. (2021)). Details in Appendix D.

drawn from any point in time (sampled independently from the marginal). Our contrastive learner tries to classify if a pair is positive or negative by learning a representation $f_X\colon X \to \mathbb{R}^d$ such that the dot-product kernel approximates the log odds ratio up to some offset:

$$\langle f_X(x_a), f_X(x_b)\rangle \approx \log \frac{\mathbb{P}(\mathsf{pos} \mid x_a, x_b)}{\mathbb{P}(\mathsf{neg} \mid x_a, x_b)} + \tilde{c}_X(x_a) \quad (3)$$

$$= \log \frac{P_{\mathsf{coor}}(x_a \mid x_b)}{P_{\mathsf{coor}}(x_a)} + c_X(x_a) \quad (4)$$

$$= K_{\mathsf{PMI}}(x_a, x_b) + c_X(x_a), \quad (5)$$

where K_{PMI} is the pointwise mutual information (PMI) kernel, and $c_X(x_a)$ is constant in x_b. We note that is a common setting for self-supervised contrastive learners with NCE objectives (Gutmann & Hyvärinen, 2010; Oord et al., 2018), such as SimCLR (Chen et al., 2020) and SimCSE (Gao et al., 2021). (See Oord et al. (2018) and Appendix F.1 for details.)

Under mild conditions that the world is smooth enough (see Appendix F.2), a choice of f_X can exactly represent K_{PMI}:

$$\langle f_X(x_a), f_X(x_b)\rangle = K_{\mathsf{PMI}}(x_a, x_b) + c_X, \quad (6)$$

where we observed that $c_X(x_a)$ from Equation (5) must be a constant since both sides are symmetric.

Therefore, the contrastive learners we consider are minimized by a representation f_X whose kernel is K_{PMI} (up to a constant offset). With sufficient data and optimization, we will observe convergence to this point.

Thus we have convergence to a representation of the statistics of X, but what about Z? Recall that our idealized world consists of *bijective* observation functions, which, over discrete random variables, preserve probabilities. So we have:

$$P_{\mathsf{coor}}(x_a, x_b) = P_{\mathsf{coor}}(z_a, z_b)$$
$$K_{\mathsf{PMI}}(x_a, x_b) = K_{\mathsf{PMI}}(z_a, z_b),$$

where we use P_{coor} and K_{PMI} in a modality-agnostic way to emphasize that different modalities share the same such quantities.

All these arguments hold not just for X but also for Y (or any other bijective, discrete modality), implying:

$$K_{\mathsf{PMI}}(z_a, z_b) = \langle f_X(x_a), f_X(x_b)\rangle - c_X \quad (7)$$
$$= \langle f_Y(y_a), f_Y(y_b)\rangle - c_Y. \quad (8)$$

Therefore, for any modality in our idealized world, we observe representational convergence to the same kernel, which represents certain pairwise statistics of $\mathbb{P}(\mathbf{Z})$.

This analysis suggests that certain representation learning algorithms may boil down to a simple rule: *find an embedding in which similarity equals PMI*. We note that this idea is consistent with prior works that have used PMI as a similarity measure for clustering in vision and language (*e.g.*, Isola et al. (2014); Isola (2015); Isola et al. (2016); Chambers & Jurafsky (2008)).

A study in color We conduct a case study to verify that convergence does happen on real data. Abdou et al. (2021) discovered that color distances in learned language representations, when trained to predict cooccurrences in *text* (Devlin et al., 2018), closely mirror human perception of these distances, which we reproduce in Figure 8 with both contrastive and predictive models. Interestingly, they noted an increasing similarity as models scale larger and become better at modeling *text* cooccurrences. In Figure 8, we also learn representations of color based on K_{PMI} from cooccurrences in *images*. Indeed, learning cooccurrence statistics in either domain recovers roughly the *same* perceptual representation. Details of this experiment are described in Appendix D.

We believe that our simple model encapsulates essential aspects of complex real-world systems, and offers a path toward understanding the representation that models are con-

verging to—a unified model that is proficient across various domains and modalities, grounded in the statistical properties of the underlying world. Section 6 further elaborates some limitations.

5. What are the implications of convergence?

Scaling is sufficient, but not necessarily efficient Our arguments are roughly in line with the claim that "scale is all you need" to reach high levels of intelligence. We have argued that as resources are scaled (# parameters, # datapoints, # flops), representations are converging, regardless of other modeling choices and even data modality. Does this mean that scale is all that matters? Not quite: different methods can scale with different levels of *efficiency* (Hestness et al., 2017; Kaplan et al., 2020), and successful methods must still satisfy some general requirements (*e.g.*, be a consistent estimator, model pairwise statistics of $\mathbb{P}(\mathbf{Z})$).

Training data can be shared across modalities Suppose you have access to N images and M sentences, and want to learn the best representation. If there is indeed a modality-agnostic platonic representation, then the image data should help find it, and so should the language data. The implication is that if you want to train the best vision model, you should train not just on N images but also on M sentences. This is already becoming common practice (Achiam et al., 2023; Radford et al., 2021). Many vision models are fine-tuned from pre-trained LLMs. The other direction is less common, but also is implied by our hypothesis: if you want to build the best LLM, *you should also train it on image data*. Indeed, Achiam et al. (2023) claim evidence that this is true, where training on images improved performance on text. In theory, there should be some conversion ratio: a pixel is worth a words for training LLMs, and a word is worth b pixels for training vision models.

Ease of translation and adaptation across modalities When two representations are aligned, transitioning from one to the other should be a simple function that's easily obtained. Our hypothesis could explain the phenomenon that conditional generation is easier than unconditional (Mirza & Osindero, 2014; Liu et al., 2020; Sauer et al., 2022), as the data we condition on may have the same platonic structure as the data we are generating. In line with this, recent work has found that representation-conditioning is even easier (Li et al., 2023). Similarly, representational convergence could act as a bridge that lets us find mappings between domains even without paired data; this may underlie the success of unpaired translation in vision (Zhu et al., 2017; Shi et al., 2024; Xie et al., 2022) and language (Tran et al., 2017; Lample et al., 2018). We emphasize that this doesn't mean that models trained on a single modality (*e.g.*, language) can immediately process raw data from another (*e.g.*, vision). What makes them adaptable to the new modalities is that they share a common modality-agnostic representation, and can readily process *representations* of new modalities. Furthermore, this implies that language models would achieve some notion of grounding in the visual domain even in the absence of cross-modal data.[5] The primary advantage of cross-modal data could then simply be sample efficiency.

Scaling may reduce hallucination and bias A prominent shortcoming of current LLMs is their propensity to hallucinate, or output false statements. If models are indeed converging toward an accurate model of reality, and scale powers this convergence, then we may expect hallucinations to decrease with scale. Of course, our hypothesis is conditioned on the training data for future models constituting a sufficiently lossless and diverse set of measurements. This may not come to pass, but it is an implication of our hypothesis worth pointing out. A similar argument can be made about certain kinds of bias. It has been shown that large models can exacerbate existing biases present in their training data (Hall et al., 2022). Our hypothesis implies that, while this may be true, we should expect *larger* models to amplify bias *less*. This does not mean bias will be removed, rather that the model's biases will more accurately reflect the data's biases, rather than exacerbating them.

6. Counterexamples and limitations

Different modalities may contain different information One immediate objection to our hypothesis is: what about the information that is unique to a given modality? Can language really describe the ineffable experience of watching a total solar eclipse? Or, how could an image convey the a concept like "I believe in the freedom of speech," which is easy to write in English? Two different models cannot converge to the same representation if they have access to fundamentally different information.

More precisely, our mathematical argument in Section 4 only strictly holds for bijective projections of $\mathbf{Z}$, so that the information in all the projections is equivalent to the information in the underlying world. This will not hold true for either lossy or stochastic observation functions. Nonetheless, similar arguments have been made theoretically and empirically that cooccurrence relations are learned by practical contrastive (Wang & Isola, 2020; Zimmermann et al., 2021) and predictive learners (Papyan et al., 2020). Lu

[5] In 1688, William Molyneux posed the question: could someone born blind, upon being given sight, be able to distinguish shapes by vision alone? (Locke, 1690) Our arguments suggest an answer: not immediately, but after a bit of visual experience (to form a visual representation) it should be easy (by mapping to prior touch-based representations). Empirical data shows that indeed congentially blind children given sight can quickly learn such abilities (Held et al., 2011).

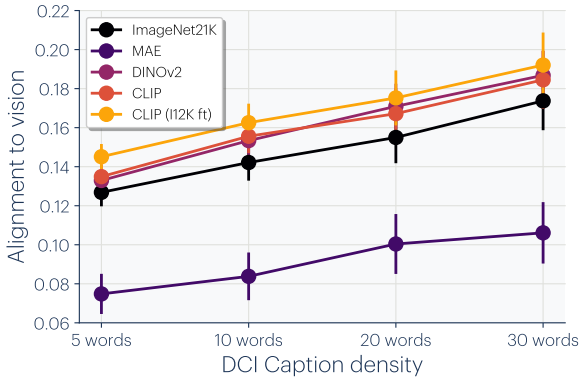


Figure 9. **Increasing caption density improves alignment:** We vary caption length using the Densely-Captioned-Images (DCI) dataset (Urbanek et al., 2023). Starting from a dense caption, we used LLaMA3-8B-Instruct (Meta, 2024) to summarize and generate coarse-grained captions. We compute the average alignment score across all vision and language models with standard deviation measured over the language models we evaluated. With denser captions, the mapping may become more bijective, leading to improved language-vision alignment scores.

et al. (2021) and Mirchandani et al. (2023) also showed that models trained to autoregressively generate text also capture statistical relations in many other modalities, including symbolic reasoning, vision, protein folding, and robotics.

A more nuanced version of our hypothesis will need to be developed to handle the case of non-bijective observations and abstract concepts. A starting point could be: different models will converge to the same representation *when the input signals are sufficiently high information and the models are sufficiently high capacity*; when they are not, the lower-information representation will only align with the higher-information one up to a level capped by the mutual information between the input signals and by the capacity of each model. This cap might or might not be practically important. Popular representations like CLIP are explicitly optimized to only capture the shared information between vision and language, yet are highly successful on many pure vision tasks. We perform a preliminary test of the effect of information level in Figure 9 (detailed in Appendix E), and find that the more descriptive (higher information) a caption is, the better its LLM representation aligns with the visual representation of the corresponding image.

Not all representations are presently converging Our argument has mainly focused on two modalities: vision and language. While we do expect other modalities will follow similar trends, we have yet to see the same level of convergence across all domains. For example, in robotics there is not yet a standardized approach to representing world states in the same way as there is for representing images and text. One limitation lies in the hardware used in robotics, which is often expensive and slow. This creates a bottleneck in the quantity and diversity of training data.

Sociological bias in producing AI models Researcher bias and collective preferences within the AI community have shaped the trajectory of model development. There is often an explicit or implicit goal of designing AI systems that mimic human reasoning and performance, and this could lead to convergence toward human-like representations even if other kinds of intelligence are in fact possible. Additionally, the "hardware lottery" (Hooker, 2021) suggests that the success of AI models can also depend on the compatibility of their design with available computational architectures, further contributing to convergent trends.

Special-purpose intelligences might not converge Different intelligent systems can be designed to accomplish different tasks. For instance: A bioinformatics systems might predict protein structure; an autonomous vehicle might follow lanes on highways. It's possible that not much is shared between these two narrow tasks. Our argument only holds for intelligences that are optimized to perform well on *many* tasks. We have argued that a representation of *reality* is a structure that is useful across many tasks, but for any special purpose there may be shortcuts, or even effective representations detached from reality. Such shortcuts may be more efficient and necessary for continued improvements in specific domains. This will become more relevant if continued scaling comes up against boundary conditions around resources like energy and compute.

How do we measure alignment? We focused on one particular alignment measure, mutual nearest-neighbor, in our experiments, and cited experiments using several others. However, there is active debate on the merits and deficiencies of all these ways of measuring alignment (Bansal et al., 2021; Sucholutsky et al., 2023). We discuss our choice and show results for other alignment metrics in Appendix A.

Lots left to explain We have shown results where different models arrive at *similar* but not the *same* representations. For example, in Figure 3, alignment clearly increases but only reaches a score of 0.16, according to our mutual nearest-neighbor metric. The maximum theoretical value for this metric is 1. Is a score of 0.16 indicative of strong alignment with the remaining gap being "noise" or does it signify poor alignment with major differences left to explain? We leave this as an open question.

Acknowledgements

We thank Lindsey & Brown for sharing their data for our experiments shown in Figure 8. We thank the anonymous reviewers for helpful feedback, and for providing the coun-

terexample on how to visually convey "I believe in the freedom of speech." Thanks for Yonglong Tian, Dilip Krishnan, Anna Decker, Yoon Kim, Jyo Pari, Ani Nrusimha, Dave Epstein, Victor Butoi, and Seungwook Han for helpful discussions and suggestions. This work was supported by a Packard Fellowship and a Sloan Research Fellowship to P.I., by the MIT-IBM Watson AI Lab, by ONR MURI grant N00014-22-1-2740, by the Center for Brains, Minds, and Machines, the MIT Quest for Intelligence, NSF STC award CCF-1231216, the DARPA Knowledge Management at Scale and Speed (KMASS) program, and the DARPA Machine Common Sense (MCS) program.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*, 2021.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019b.

Balestriero, R. and Baraniuk, R. G. A spline theory of deep learning. In *International Conference on Machine Learning*, pp. 374–383. PMLR, 2018.

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Baradad, M., Wulff, J., Wang, T., Isola, P., and Torralba, A. Learning to see by looking at noise. In *Advances in Neural Information Processing Systems*, 2021.

Baradad, M., Chen, R., Wulff, J., Wang, T., Feris, R., Torralba, A., and Isola, P. Procedural image programs for representation learning. *Advances in Neural Information Processing Systems*, 35:6450–6462, 2022.

Barlow, H. B. et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8, 2023.

BigScience, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Cao, R. and Yamins, D. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *Cognitive Systems Research*, 85, 2024.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chambers, N. and Jurafsky, D. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pp. 789–797, 2008.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve

math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pp. 2022–03, 2022.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Diamond, J. M. *Guns, germs and steel: a short history of everybody for the last 13,000 years*. Vintage London, 1998.

Dingle, K., Camargo, C. Q., and Louis, A. A. Input–output maps are strongly biased towards simple outputs. *Nature communications*, 9(1):761, 2018.

Doppelt, G. Reconstructing scientific realism to rebut the pessimistic meta-induction. *Philosophy of Science*, 74(1): 96–118, 2007.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dravid, A., Gandelsman, Y., Efros, A. A., and Shocher, A. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1934–1943, 2023.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Gao, T., Yao, X., and Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Gell-Mann, M. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Macmillan, 1995.

Geng, X. and Liu, H. OpenLLaMA: An open reproduction of LLaMA, May 2023. URL https://github.com/openlm-research/open_llama.

Gokaslan, A. and Cohen, V. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Goldblum, M., Finzi, M., Rowan, K., and Wilson, A. G. The no free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. *arXiv preprint arXiv:2304.05366*, 2023.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Hall, M., van der Maaten, L., Gustafson, L., Jones, M., and Adcock, A. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

Hardin, C. L. and Rosenberg, A. In defense of convergent realism. *Philosophy of Science*, 49(4):604–615, 1982.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2021.

Held, R., Ostrovsky, Y., de Gelder, B., Gandhi, T., Ganesh, S., Mathur, U., and Sinha, P. The newly sighted fail to match seen with felt. *Nature neuroscience*, 14(5):551–553, 2011.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hooker, S. The hardware lottery. *Communications of the ACM*, 64(12):58–65, 2021.

Huh, M., Mobahi, H., Zhang, R., Cheung, B., Agrawal, P., and Isola, P. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bCiNWDmlY2.

Isola, P. The discovery of perceptual structure from visual co-occurrences in space and time. In *MIT Ph.D. Thesis*, 2015.

Isola, P., Zoran, D., Krishnan, D., and Adelson, E. H. Crisp boundary detection using pointwise mutual information. In *ECCV*, 2014.

Isola, P., Zoran, D., Krishnan, D., and Adelson, E. H. Learning visual groups from co-occurrences in space and time. In *ICLR, Workshop paper*, 2016.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Jordan, K., Sedghi, H., Saukh, O., Entezari, R., and Neyshabur, B. Repair: Renormalizing permuted activations for interpolation repair. *arXiv preprint arXiv:2211.08403*, 2022.

Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 34(5):827–828, 1978.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Klabunde, M., Schumacher, T., Strohmaier, M., and Lemmerich, F. Similarity of neural network models: A survey of functional and representational measures. *arXiv preprint arXiv:2305.06329*, 2023.

Koh, J. Y., Salakhutdinov, R., and Fried, D. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pp. 17283–17300. PMLR, 2023.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. Phrase-based & neural unsupervised machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Lenc, K. and Vedaldi, A. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Li, T., Katabi, D., and He, K. Return of unconditional generation: A self-supervised representation generation method. *arXiv:2312.03701*, 2023.

Lian, L., Li, B., Yala, A., and Darrell, T. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023a.

Lian, L., Shi, B., Yala, A., Darrell, T., and Li, B. LLM-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023b.

Lindsey, D. T. and Brown, A. M. The color lexicon of american english. *Journal of vision*, 14(2):17–17, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Liu, S., Wang, T., Bau, D., Zhu, J.-Y., and Torralba, A. Diverse image generation via self-conditioned GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Locke, J. *An Essay Concerning Human Understanding*. 1690.

López-Cifuentes, A., Escudero-Vinolo, M., Bescós, J., and García-Martín, Á. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020.

Lu, K., Grover, A., Abbeel, P., and Mordatch, I. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 1, 2021.

Lubana, E. S., Bigelow, E. J., Dick, R. P., Krueger, D., and Tanaka, H. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pp. 22965–23004. PMLR, 2023.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Meta. Meta LLaMA 3, 2024. URL https://ai.meta.com/blog/meta-llama-3/.

Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., and Zeng, A. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., and Rodolà, E. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Nettleship, R. L. *Lectures on the 'Republic' of Plato*, volume 2. Macmillan, 1897.

Newton-Smith, W. *The Rationality of Science*. International Library of Philosophy, Psychology, and Scientific Method. Routledge & Kegan Paul, 1981. ISBN 9780710009135.

Ng, E., Subramanian, S., Klein, D., Kanazawa, A., Darrell, T., and Ginosar, S. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10083–10093, 2023.

Ngo, J. and Kim, Y. What do language models hear? probing for auditory representations in language models, 2024.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Oron, S., Dekel, T., Xue, T., Freeman, W. T., and Avidan, S. Best-buddies similarity—robust template matching using mutual nearest neighbors. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1799–1813, 2017.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Plato. Republic. c. 375 BC.

Putnam, H. Three kinds of scientific realism. *The Philosophical Quarterly (1950-)*, 32(128):195–200, 1982.

Radford, A., Jozefowicz, R., and Sutskever, I. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Richens, J. and Everitt, T. Robust agents learn causal world models. *ICLR*, 2024.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Sauer, A., Schwarz, K., and Geiger, A. StylegGAN-XL: Scaling StyleGAN to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Sharma, P., Rott Shaham, T., Baradad, M., Fu, S., Rodriguez-Munoz, A., Duggal, S., Isola, P., and Torralba, A. A vision check-up for language models. In *arXiv preprint*, 2024.

Shepard, R. N. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Smola, A. J. and Schölkopf, B. *Learning with kernels*, volume 4. Citeseer, 1998.

Solomonoff, R. J. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012.

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2443–2449, 2021.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Steinberg, E., Jung, K., Fries, J. A., Corbin, C. K., Pfohl, S. R., and Shah, N. H. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113: 103637, 2021.

Stoica, G., Bolya, D., Bjorner, J., Hearn, T., and Hoffman, J. Zipit! merging models from different tasks without training. *arXiv preprint arXiv:2305.03053*, 2023.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K. L., Oktar, K., Greff, K., Hebart, M. N., Jacoby, N., Zhang, Q., Marjieh, R., Geirhos, R., Chen, S., Kornblith, S., Rane, S., Konkle, T., O'Connell, T. P., Unterthiner, T., Lampinen, A. K., Müller, K.-R., Toneva, M., and Griffiths, T. L. Getting aligned on representational alignment, 2023.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020a.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 266–282. Springer, 2020b.

Tolstoy, L. *Anna Karenina*. The Russian Messenger, 1877.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tran, D., Burda, Y., and Sutskever, I. Feature-matching auto-encoders. 2017.

Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991.

Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., and Romero-Soriano, A. A picture is worth more than 77 text tokens: Evaluating CLIP-style models on dense captions, 2023.

Valle-Perez, G., Camargo, C. Q., and Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Werbos, P. J. Learning how the world works: Specifications for predictive networks in robots and brains. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics, NY*, 1987.

Wightman, R. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2021.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022.

Wu, T.-H., Lian, L., Gonzalez, J. E., Li, B., and Darrell, T. Self-correcting LLM-controlled diffusion models. *arXiv preprint arXiv:2311.16090*, 2023.

Xie, S., Ho, Q., and Zhang, K. Unsupervised image-to-image translation with density changing regularization. *Advances in Neural Information Processing Systems*, 35: 28545–28558, 2022.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. The visual task adaptation benchmark. 2019.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

A. Mutual k-Nearest Neighbor Alignment Metric

For two models with representations f, g the mutual k-nearest neighbor metric measures the average overlap of their respective nearest neighbor sets. In this section, we refer to this metric as m_{NN}, which we will formally define below.

For cross-modal domains, define $(x_i, y_i) \in \mathcal{X}$ as a sample from the data distribution $\mathcal{X}$ (*e.g.* image-caption dataset). For the single domain alignment measurements, the samples are equivalent $x_i = y_i$ (*e.g.*, images for vision, and text for language). Let $\{x_i, y_i\}_{i=1}^b$ be the corresponding mini-batch sampled from this data distribution. Then given two model representations f and g the corresponding features are: $\phi_i = f(x_i)$ and $\psi_i = f(y_i)$, where the collection of these features are denoted as $\Phi = \{\phi_1, \ldots, \phi_b\}$ and $\Psi = \{\psi_1, \ldots, \psi_b\}$. Then for each feature pair (ϕ_i, ψ_i), we compute the respective nearest neighbor sets $\mathcal{S}(\phi_i)$ and $\mathcal{S}(\psi_j)$.

$$d_{\mathsf{knn}}(\phi_i, \Phi \setminus \phi_i) = \mathcal{S}(\phi_i) \tag{9}$$

$$d_{\mathsf{knn}}(\psi_i, \Psi \setminus \psi_i) = \mathcal{S}(\psi_j) \tag{10}$$

where d_{knn} returns the set of indices of its k-nearest neighbors. Then we measure its average intersection via

$$m_{\mathsf{NN}}(\phi_i, \psi_i) = \frac{1}{k} |\mathcal{S}(\phi_i) \cap \mathcal{S}(\psi_j)| \tag{11}$$

where $|\cdot|$ is the size of the intersection.

The choice to use mutual nearest-neighbors Our initial efforts to measure alignment with CKA revealed a very weak trend of alignment between models, even when comparing models within their own modality. This has also been observed by (Bansal et al., 2021), which had relied on alternative metrics such as model-stitching as it "reveals aspects of representations that measures such as centered kernel alignment (CKA) cannot" (Bansal et al., 2021).

We chose to use nearest-neighbor as a metric, as methods like CKA has a very strict definition of alignment, which may not fit our current needs. For instance, understanding the precise similarity between unrelated items, such as an orange and Bill Gates, may not be critical.

Relationship between CKA and Mutual Nearest-Neighbors Let $\phi_i \in \mathbb{R}^n$ and $\psi_i \in \mathbb{R}^m$ be vectorized features of two models (*e.g.* language and vision models). Let $\mathbf{K}_{ij} = \kappa(\phi_i, \phi_j)$ and $\mathbf{L}_{ij} = \kappa(\psi_i, \psi_j)$ be the kernel matrices computed from a dataset using some kernel-function κ. Using an inner-product kernel, the ij-th entry of the centered counterpart of these Kernel matrices is:

$$\bar{\mathbf{K}}_{ij} = \langle \phi_i, \phi_j \rangle - \mathbb{E}_l[\langle \phi_i, \phi_l \rangle] \qquad \bar{\mathbf{L}}_{ij} = \langle \psi_i, \psi_j \rangle - \mathbb{E}_l[\langle \psi_i, \psi_l \rangle] \tag{12}$$

Then, the cross-covariance of $\mathbf{K}$ and $\mathbf{L}$ is given by:

$$\mathsf{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2} \mathsf{Trace}(\bar{\mathbf{K}}\bar{\mathbf{L}}) \tag{13}$$

which serves as an empirical estimator of the Hilbert-Schmidt Independence Criterion (Gretton et al., 2005). The Centered Kernel Alignment (CKA) (Kornblith et al., 2019) is then its normalized counterpart:

$$\mathsf{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\mathsf{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathsf{HSIC}(\mathbf{K}, \mathbf{K})\mathsf{HSIC}(\mathbf{L}, \mathbf{L})}} \tag{14}$$

CKA measures the congruence between two random variables, with a maximum alignment of 1 and a minimum of 0. It is invariant to isotropic scaling and offers a strict notion of alignment, measuring alignment across all samples. Hence, the CKA score reflects the global similarities of the models. This can be illustrated by expanding the trace term in HSIC:

$$\mathsf{Trace}(\bar{\mathbf{K}}\bar{\mathbf{L}}) = \sum_i \sum_j \left(\langle \phi_i, \phi_j \rangle - \mathbb{E}_l[\langle \phi_i, \phi_l \rangle]\right)\left(\langle \psi_i, \psi_j \rangle - \mathbb{E}_l[\langle \psi_i, \psi_l \rangle]\right) \tag{15}$$

One can modify the definition of alignment to restrict the cross-covariance measurement to samples considered to be nearest neighbors of the current sample i. This emphasizes similarity over dissimilarity, biasing the measure toward local alignment:

$$\mathsf{Align}_{\mathsf{knn}}(\mathbf{K}, \mathbf{L}) = \sum_i \sum_j \alpha(i,j) \cdot \left(\langle \phi_i, \phi_j \rangle - \mathbb{E}_l[\langle \phi_i, \phi_l \rangle]\right)\left(\langle \psi_i, \psi_j \rangle - \mathbb{E}_l[\langle \psi_i, \psi_l \rangle]\right) \tag{16}$$

$$\text{where} \qquad \alpha(i,j) = \mathbb{1}[\phi_j \in \mathsf{knn}(\phi_i) \wedge \psi_j \in \mathsf{knn}(\psi_i) \wedge i \neq j] \tag{17}$$

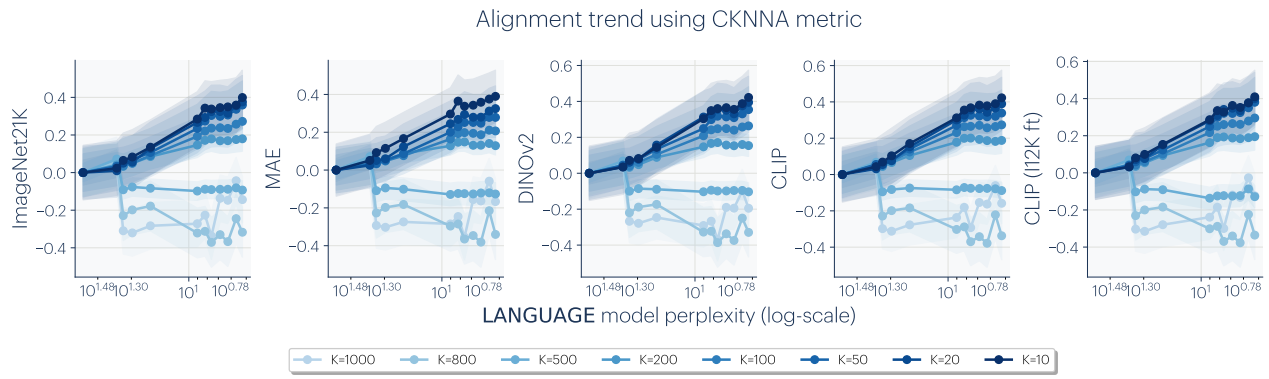


Figure 10. **Cross-modal alignment increases locally:** Alignment trend when varying the top-k nearest neighbors in the CKNNA metrics (Eqn. 18). We center alignment score to the smallest language model and divide the total trend by the standard deviation. When $k = 1024$, we recover the original CKA metric, and when $k < |\mathcal{X}|$ it closely resembles the mutual nearest-neighbor metric m_{NN}. Each line represents the average of all LLM models for a specific k. As we decrease k, the alignment becomes more pronounced. Certain visual tasks, such as CLIP, show global alignment, while methods like ImageNet1k classification only exhibit local alignment.

Where $\alpha(i, j)$ is a scalar weighting that assigns 1 if j is a mutual nearest neighbors to both ϕ_i and ψ_i, and 0 otherwise. We refer to this metric as the Centered Kernel Nearest-Neighbor Alignment (CKNNA) metric. As the number of nearest neighbors $k \to \dim(\mathbf{K})$, we recover the original CKA metric.

$$\mathsf{CKNNA}(\mathbf{K}, \mathbf{L}) = \frac{\mathsf{Align}_{\mathsf{knn}}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathsf{Align}_{\mathsf{knn}}(\mathbf{K}, \mathbf{K}), \mathsf{Align}_{\mathsf{knn}}(\mathbf{L}, \mathbf{L})}} \tag{18}$$

We can further relax the metric to treat the cross-covariance term identically across all nearest-neighbor samples. This is equivalent to the assumption that all nearby samples have the same distance. This simplification leads us back to the mutual nearest neighbor metric:

$$\sum_i \sum_j \alpha(i, j) \cdot 1 = n \cdot k \cdot m_{\mathsf{NN}}(\phi_i, \psi_i) \tag{19}$$

By equating these metrics, we analyze the changes in alignment between language and vision models as we vary the number of neighbors k in Eqn. 18. In Figure 10, we compute the average alignment score across all LLM models. For each k, we center the scores to the smallest vision model and divide by the standard deviation of the scores. We find that high values of k show less conclusive alignment across tasks while decreasing k shows a coherent trend across both models and tasks. We find that certain visual tasks, such as CLIP, exhibit global alignment, whereas methods like ImageNet21k classification show only local alignment. This observation suggests that cross-modal alignment occurs locally across most common visual tasks, and global alignment may require additional language grounding, as done in the CLIP objective.

B. Consistency across various metrics

We describe the metrics in Table 11 and their corresponding properties. The *symmetric* property implies that the metric is symmetric with respect to the data points $d(x, y) = d(y, x)$. The *global* property means all samples are used to compute the distance with respect to every sample. The *ordinal* property is when the ordering of the distance is taken into consideration. For example, mutual nearest neighbor is not ordinal since the nearest neighbors $\{a, b, c\}$ and $\{c, a, b\}$ are treated equally. The *batchable* property is a computational property that makes it feasible to compute in a reasonable time frame.

Vision-vision comparison In Figure 12, we evaluate Spearman's rank correlation among different metrics and hyperparameters over 78 vision models (details in Appendix C.1). We find most metrics highly correlated with each other.

Cross-modal comparison We measure vision-language alignment using a range of alternative metrics. We visualize the corresponding alignment results in Figure 13 and Figure 14. Our findings indicate that alignment sensitivity not only depends on the metric used to compute it but also varies according to the specific tasks on which the vision models are trained.

Metric	Property				Description
	symmetric	global	ordinal	batchable	
CKA	✓	✓	✓	✓	Centered Kernel Alignment (CKA; Kornblith et al. (2019)) measures the similarity of neural networks by comparing the alignment of their kernel induced by their feature spaces.
Unbiased CKA	✓	✓	✓	✓	Unbiased estimator of CKA that corrects for sample bias in HSIC (Song et al., 2012).
SVCCA	✓	✓	✓	✓	Singular Value Canonical Correlation Analysis (SVCCA; Raghu et al. (2017)) compares neural networks by decomposing their activities into singular vectors and measuring correlation.
Mutual k-NN	✓			✓	Measures the intersection over union (IoU) of nearest neighbors between two models.
CKNNA	✓	✓$*$	✓	✓	Modified CKA measure that computes the kernel alignment only for its nearest neighbors. See Appendix A.
Cycle k-NN				✓	Measures whether the nearest neighbor in one domain also considers the original sample as its nearest neighbor in the other domain.
Edit k-NN	✓	✓$*$	✓		Computes the edit distance required to match the nearest neighbors between two datasets. The score is normalized by the maximum edit distance.
LCS k-NN	✓	✓$*$	✓		Calculates the longest common subsequence of nearest neighbors and is normalized by the sequence length.

Figure 11. Comparative analysis of neural network similarity metrics. ✓$*$ indicates the metric is global and still meaningful when the nearest neighbor k is set to maximum batch-size $k = |\mathcal{X}|$.

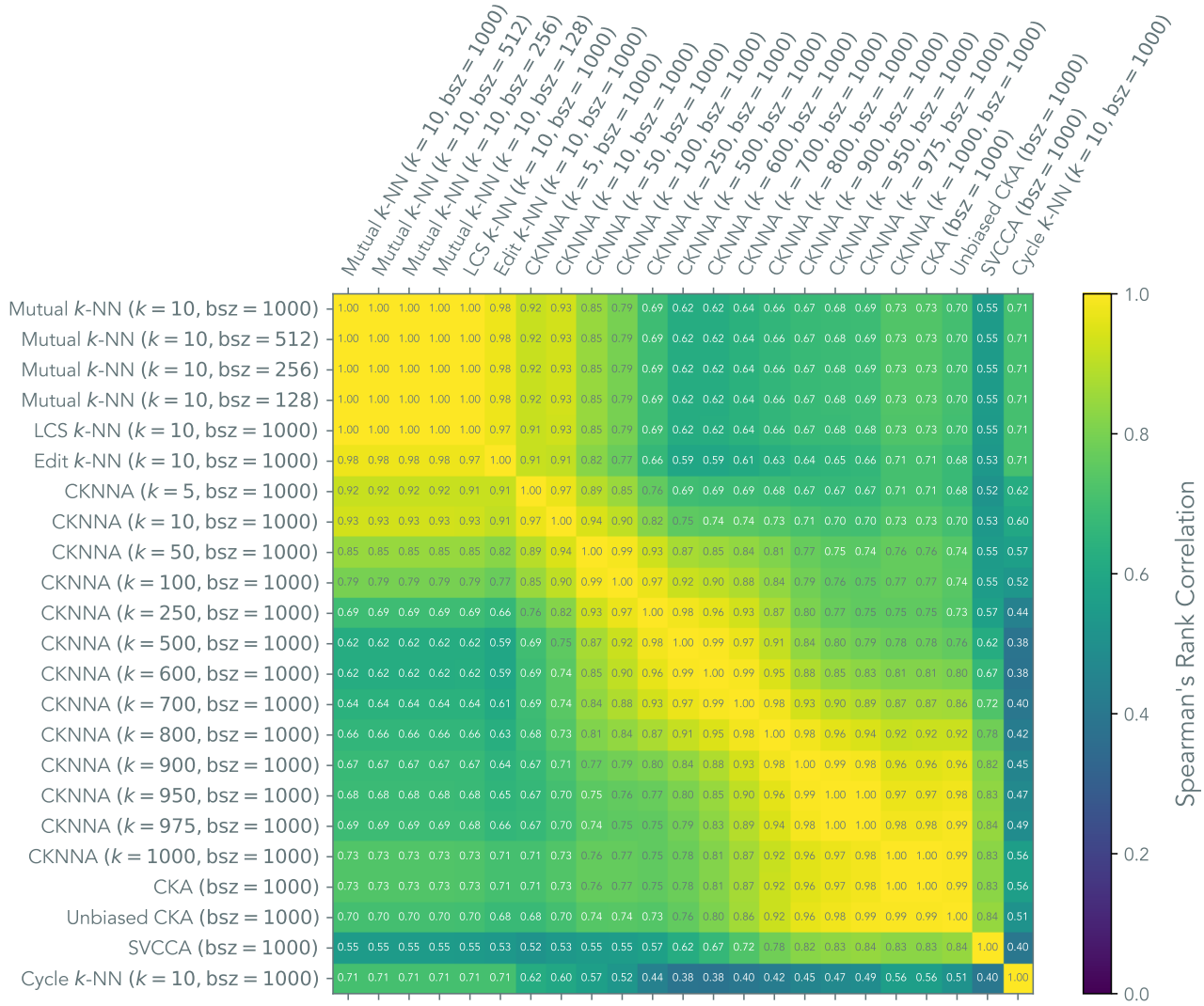
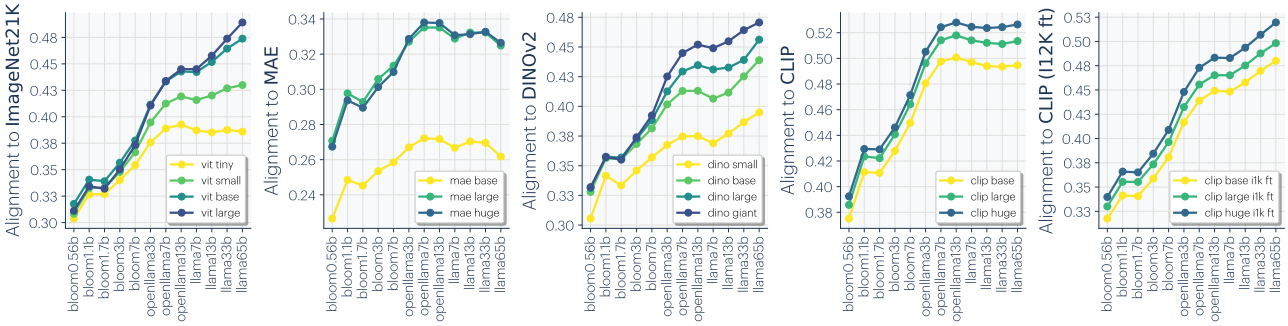
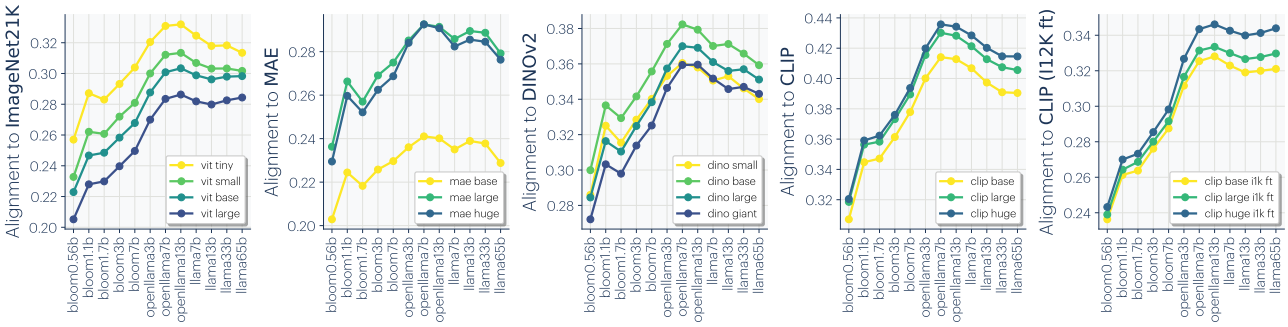


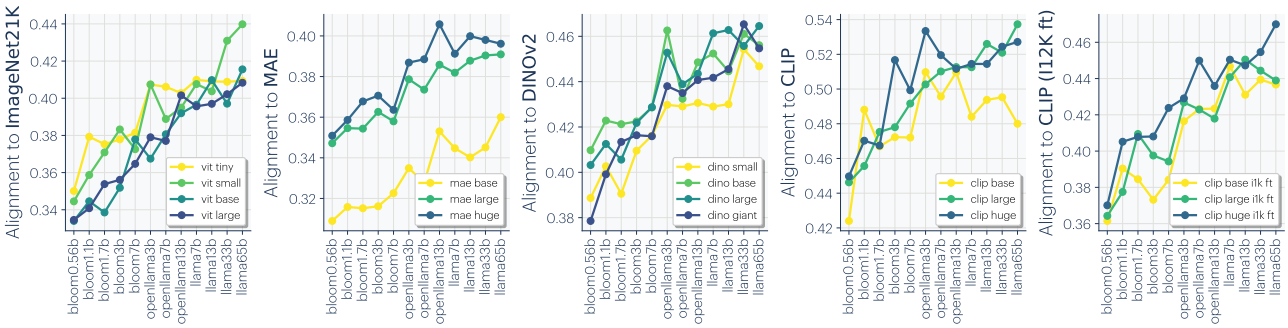
Figure 12. **Vision-vision alignment measured with various metrics.** Spearman's rank correlation among different metrics and batch sizes (bsz) when used to measure alignment among 78 vision models (see Appendix C.1 for details of these models). All p-values are below 2.24×10^{-105}. Our vision-vision analysis in Figure 2 is based on the first metric (Mutual k-NN with $k = 10$ and bsz $= 1000$).



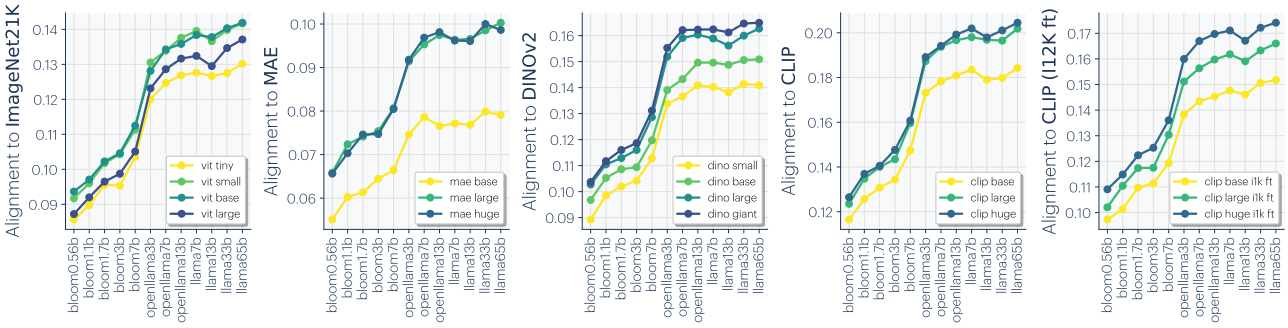
(a) CKA



(b) Unbiased CKA

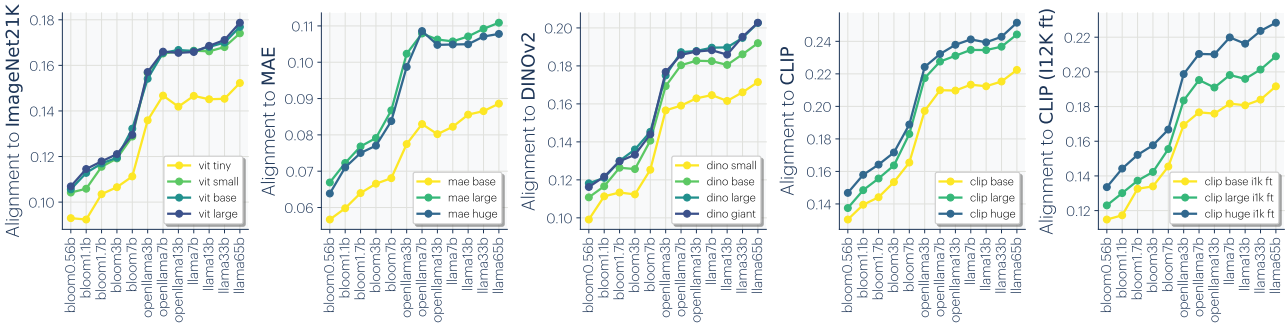


(c) SVCCA

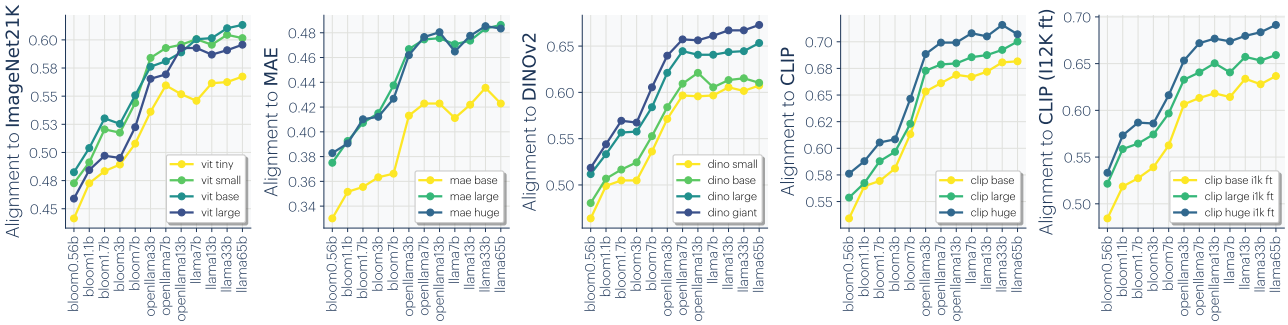


(d) Mutual k-NN ($k = 10$)

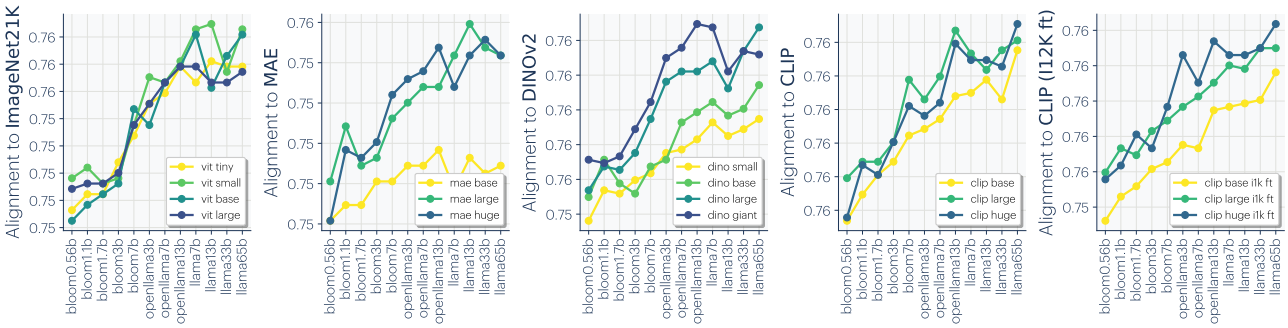
Figure 13. **Cross-modal alignment for various metrics**



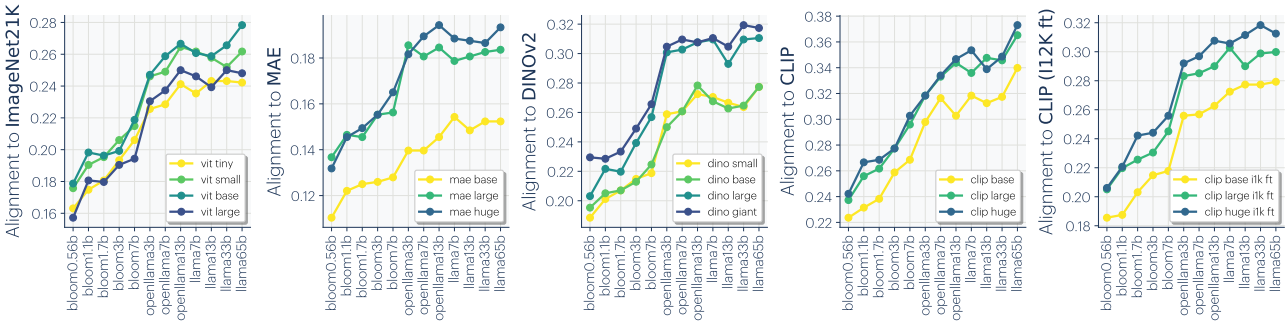
(a) CKNNA ($k = 10$)



(b) Cycle k-NN ($k = 10$)



(c) Edit-distance k-NN ($k = 10$)



(d) Longest-Common-Subsequence k-NN ($k = 10$)

Figure 14. **Cross-modal alignment measured with various metrics**

C. Experiments on Evaluating Alignment and Convergence

To demonstrate representational convergence, we take off-the-shelf models at multiple scales and multiple modalities and measure their representational alignment.

C.1. Vision-Vision Alignment and Representation Quality

We consider 78 vision models in total:

- 17 ViT models ranging from ViT-tiny to ViT-giant, trained on tasks including ImageNet-21k (Dosovitskiy et al., 2020) classification, Masked Autoencoders (He et al., 2021), DINO (Caron et al., 2021), and CLIP (Radford et al., 2021), including some finetuned on ImageNet-12k.
- 1 randomly initialized ResNet-50.
- 11 ResNet-50 models trained with contrastive learning on ImageNet-1k, Places-365 (Zhou et al., 2017; López-Cifuentes et al., 2020), and 9 synthetic image datasets used in Baradad et al. (2022).
- 49 ResNet-18 models trained with Alignment and Uniformity contrastive loss (Wang & Isola, 2020) on ImageNet-100, Places-365, and 47 realistic and synthetic image datasets from Baradad et al. (2021).

To test representation quality, we evaluate linear probing performance on all 19 VTAB classification tasks (Zhai et al., 2019), which is a standard multi-task transfer learning benchmark containing structured, specialized, and natural datasets covering diverse domains. To reduce compute requirements, we subsample training and validation datasets to have at most 10,000 samples. We consider a representation solves a task if its performance is $\geq 80\%$ of the best performance on that task across all 78 models.

To compute the alignment metric, we use $k = 10$ nearest neighbors over 1000 image representations computed on Places-365's validation dataset (Zhou et al., 2017). This dataset is disjoint from VTAB datasets, although both contain natural images.

C.2. Cross-Modal Alignment

We compare the representation of an image in a vision model to the representation of a caption describing that image in a language model. The language model families we consider are BLOOM (BigScience et al., 2022), OpenLlama (Geng & Liu, 2023), and LLama (Touvron et al., 2023). For Figure 4, we included more recent model families such as OLMo (Groeneveld et al., 2024), LLama3 (Meta, 2024), Gemma (Team et al., 2024), and Mistral/Mixtral (Jiang et al., 2023; 2024). These models were downloaded from Huggingface (Wolf et al., 2019).

For vision models, we consider ViT models (Dosovitskiy et al., 2020) of various sizes trained on various data and objectives. We mainly consider the popular vision models: classification on ImageNet21K (Russakovsky et al., 2015), MAE (He et al., 2021), DINOv2 (Oquab et al., 2023), CLIP (Radford et al., 2021), and CLIP finetuned on ImageNet12K. These models were downloaded from PyTorch Image Models (TIMM) (Wightman, 2021).

To compute the alignment metric, we use $k = 10$ nearest neighbors over 1024 samples from WIT (Wikipedia-based Image Text)(Srinivasan et al., 2021). For the vision model, we use class token of each layer, and for the language model, we average pool each layer to a single token. Since it is not trivial to determine where the alignment might occur, we draw inspiration from BrainScore(Schrimpf et al., 2018) and compute pairwise alignment scores, then take the maximum. One of these pairwise comparisons also includes concatenated features. We apply l_2 normalization to the features before measuring the distance. As transformer architectures have "emergent outliers" (Dettmers et al., 2022), we truncate the elements in the features that are above the 95-th percentile.

Simply taking the last token did not show any strong alignment signal. We also experimented with prompting the language model and taking the last token representation. The prompt we used was

```
An image with the caption '<caption>'. This is an image of a <fill>
```

Using prompting showed similar trends to average pooling but had slightly lower alignment scores.

D. Color Cooccurrence Experiment

Here we describe the details of how we created the four color representations visualized in Figure 8, from left to right.

Perceptual representation from CIELAB color space We embed pixels taken from the CIFAR-10 image dataset (Krizhevsky et al., 2009; Torralba et al., 2008) based on the CIELAB color space, which is designed as a *perceptually uniform* space that changes numerical values correspond to similar perceived changes in color.

Three representations from cooccurrence in VISION and LANGUAGE For these three representations, we first obtain a dissimilarity matrix over colors (in different ways detailed below), then use multidimensional scaling (Shepard, 1980) to find a 3-dimensional embedding in which Euclidean distance between the embeddings for A and B, z_A and z_B, best matches this dissimilarity matrix. We use 1,000 fits and take the best match. Afterward, we visually align it with the CIELAB space by finding the best rotation, translation, scaling, and flipping, by running the Kabsch-Umeyama algorithm (Kabsch, 1976; 1978; Umeyama, 1991) twice, once on $\mathbf{z}$ and once on $-\mathbf{z}$, to account for flipping. The dissimilarity matrix we used in each case is described as following:

- **VISION: Pixel cooccurrence.** We collect color cooccurrence statistics from the CIFAR-10 dataset, and estimate a joint distribution $p(A, B)$ over 300,000 randomly sampled pixel colors A and B that occur within a radius of at most 4 pixels of one another. Colors are quantized on a grid in RGB space and represented as discrete variables, and $p(A, B)$ is modeled as a table of normalized counts, from which we compute the empirical pointwise mutual information matrix $K_{\mathsf{PMI}}(A, B)$. Quantization ensures that there is no bias from how color distances are represented in RGB space. Dissimilarity matrix is defined as $-K_{\mathsf{PMI}}(A, B) + c$, where $c = \max_{A,B} K_{\mathsf{PMI}}(A, B)$ is an offset to ensure non-negativity (similar to the constant in Section 4.2 and Proposition F.1 that ensures neural networks can express K_{PMI}).

- **LANGUAGE.** We used an approach similar to Abdou et al. (2021).
  - We take 20 pairs of (color, word) appeared in the dataset collected by Lindsey & Brown (2014), where 51 participants were asked to free name each of the 330 colors from the Munsell Color Chart. We filtered words that appeared less than 100 times, and computed each word's associate color by taking the centroid in CIELAB space. Our filtering process followed Abdou et al. (2021) exactly, but resulted in 20 colors, a slightly different set than the 18 colors they claimed.
  - For each of the 20 color words `<col>`, we construct three sentences:

    ```
    The color <col>.
    This color is <col>.
    The color of this thing is <col>.
    ```

    and obtain the average sentence embedding from the language encoder, as the embedding for `<col>` (details below). We find this approach more effective than Abdou et al. (2021), which uses object names that potentially have color biases, even though the objects may appear in multiple colors.
  - Unlike Abdou et al. (2021), we did not perform linear regression from language embedding to CIELAB space, which distorts distances and easily overfits with only 20 samples. Instead, we used multidimensional scaling to best preserve distances, as described above.
  - **Masked language contrastive learning (SimCSE) embedding:** We used sentence embedding from the unsupervised SimCSE RoBERTa-L (Gao et al., 2021) to encode the above sentences into 1024-dimensional embeddings, and used the pairwise Euclidean distances among `<col>` embeddings as the dissimilarity matrix.
  - **Masked language predictive learning (RoBERTa) embedding:** We concatenated hidden states of the last four layers of RoBERTa-L (Liu et al., 2019), following (Devlin et al., 2018). We averaged across token dimensions, and obtained a 4096-dimensional embedding for each of the above sentences, and used the pairwise Euclidean distances among `<col>` embeddings as the dissimilarity matrix.

E. Caption Density Experiments

We use LLAMA3-8B-Instruct (Meta, 2024) to generate summary captions at various densities for images in the Densely Captioned Images dataset (Urbanek et al., 2023) from the train split. Following Urbanek et al. (2023), we prompt the language model with the following instructions to generate captions at differing granularity:

```
system: You are given a full-text description of an image. You should summarize it into about
<num_words> words, being sure to include as much salient visual information as possible given the
<num_words> word constraint, especially information from the start of the original description. The
new description should apply for the original image. Respond with only the summary, in one line.

user: <original_caption>
```

F. Analysis of Contrastive Learners

F.1. Contrastive objectives learn pointwise mutual information

There are two widely used forms of contrastive objectives. We now discuss each form in details and show how they both are minimized by the pointwise mutual information (PMI) as stated in Equation (5). To simplify notation, we consider learning the bivariate model $g(x_a, x_b) \in \mathbb{R}$. In Section 4, such g is optimized within the family of $\{g = \langle f_X, f_X \rangle \colon f_X \in \mathcal{F}_X\}$.

Recall that our positive pairs are sampled from $(x, x_+) \sim P_{\mathsf{coor}}$, and that the negative pairs are sampled independently from its marginals which we denote as $(x, x_-) \overset{\text{i.i.d.}}{\sim} P$ where $P(x) = \sum_{x_+} P_{\mathsf{coor}}(x, x_+)$.

1. **The binary NCE loss (Gutmann & Hyvärinen, 2010)** is defined with a certain prior over sampling positive vs. negative pairs. Let p_{pos} be the probability of sampling a positive pair. Then the loss is given by

$$\mathcal{L}_{\mathsf{binary\text{-}NCE}}(g) \triangleq p_{\mathsf{pos}} \cdot \mathbb{E}_{(x,x_+)\sim P_{\mathsf{coor}}} \left[-\log \sigma(g(x, x_+))\right] + (1 - p_{\mathsf{pos}}) \cdot \mathbb{E}_{(x,x_-)\overset{\text{i.i.d.}}{\sim} P} \left[-\log \sigma(-g(x, x_-))\right]. \tag{20}$$

The Bayes optimal solution is given by

$$\begin{aligned}
g(x_a, x_b) &= \log \frac{P(\mathsf{pos} \mid x_a, x_b)}{1 - P(\mathsf{pos} \mid x_a, x_b)} && (21)\\
&= \log \frac{P(\mathsf{pos}, x_a, x_b)}{P(\mathsf{neg}, x_a, x_b)} && (22)\\
&= \log \frac{p_{\mathsf{pos}} \cdot P_{\mathsf{coor}}(x_a, x_b)}{(1 - p_{\mathsf{pos}}) P(x_a) P(x_b)} && (23)\\
&= \log \frac{P_{\mathsf{coor}}(x_a, x_b)}{P(x_a) P(x_b)} + \log \frac{p_{\mathsf{pos}}}{1 - p_{\mathsf{pos}}} && (24)\\
&= K_{\mathsf{PMI}}(x_a, x_b) + c_X. && (25)
\end{aligned}$$

2. **The InfoNCE loss (Oord et al., 2018)** is defined with randomly sampling one positive pair along with K negative ones. With some hyperparameter $\tau > 0$, the loss is given by

$$\mathcal{L}_{\mathsf{InfoNCE}}(g) \triangleq \mathbb{E}_{\substack{(x,x_+)\sim P_{\mathsf{coor}} \\ (x_-^{(1)}, x_-^{(2)}, \ldots, x_-^{(K)}) \overset{\text{i.i.d.}}{\sim} P}} \left[-\log \frac{e^{g(x,x_+)/\tau}}{e^{g(x,x_+)/\tau} + \sum_{i=1}^{K} e^{g(x,x_-^{(i)})/\tau}}\right]. \tag{26}$$

The Bayes optimal solution is given by

$$\begin{aligned}
\frac{e^{g(x,x_+)/\tau}}{e^{g(x,x_+)/\tau} + \sum_{i=1}^{K} e^{g(x,x_-^{(i)})/\tau}} &= \frac{P_{\mathsf{coor}}(x_+ \mid x) \prod_j P(x_-^{(j)})}{P_{\mathsf{coor}}(x_+ \mid x) \prod_j P(x_-^{(j)}) + \sum_i P_{\mathsf{coor}}(x_-^{(i)} \mid x) P(x_+) \prod_{j \neq i} P(x_-^{(j)})} && (27)\\
&= \frac{P_{\mathsf{coor}}(x_+ \mid x)/P(x_+)}{P_{\mathsf{coor}}(x_+ \mid x)/P(x_+) + \sum_i P_{\mathsf{coor}}(x_-^{(i)} \mid x)/P(x_-^{(i)})}. && (28)
\end{aligned}$$

For $\tau = 1$, this optima corresponds to g choices where

$$g(x_a, x_b) = \log \frac{P_{\mathsf{coor}}(x_b \mid x_a)}{P(x_b)} + c_X(x_a) \tag{29}$$

$$= K_{\mathsf{PMI}}(x_a, x_b) + c_X(x_a). \tag{30}$$

For the general $\tau \neq 1$ case, we have g (and corresponding f_X) recovers K_{PMI} up to an offset and a scale. Our main argument in Section 4 that f_X recovers K_{PMI} still holds.

F.2. Contrastive learners can represent K_{PMI} exactly under smoothness conditions

We want to express $K_{\mathsf{PMI}} + C$ using some representation function $f_X \colon \mathcal{X} \to \mathbb{R}^n$ so that $< f_X(x_a), f_X(x_b) >= K_{\mathsf{PMI}}(x_a, x_b) + C$, for some C. For such f to exist, an equivalent criterion is that $K_{\mathsf{PMI}} + C$ is positive semi-definite (PSD), as can be seen from eigendecomposition.

Proposition F.1. *Suppose that the off-diagonal elements of K_{PMI} are bounded within $[\log \rho_{\mathsf{min}}, \log \rho_{\mathsf{min}} + \delta] \in (-\infty, 0]$. We have $K_{\mathsf{PMI}} + C$ is positive semi-definite (PSD) for some C if the joint distribution is sufficiently smooth:*

$$\frac{P_{\mathsf{coor}}(z_i \mid z_i)}{P_{\mathsf{coor}}(z_i)} \geq e^{N\delta} \rho_{\mathsf{min}}, \qquad \forall i. \tag{31}$$

Proof. Note that $K_{\mathsf{PMI}} + C$ still only has non-positive off-diagonal elements if

$$-C \geq \log \rho_{\mathsf{min}} + \delta. \tag{32}$$

For such C, it is diagonally dominant (and thus PSD) if,

$$\forall i, \qquad K_{\mathsf{PMI}}(z_i, z_i) + C \geq \sum_{j \neq i} |K_{\mathsf{PMI}}(z_i, z_j) + C| = -(N-1)C - \sum_{j \neq i} K_{\mathsf{PMI}}(z_i, z_j), \tag{33}$$

or equivalently,

$$\forall i, \qquad NC + \sum_j K_{\mathsf{PMI}}(z_i, z_j) \geq 0. \tag{34}$$

The following choice of C readily satisfies the above Equation (34):

$$C \triangleq -\min_i \frac{1}{N} \sum_k K_{\mathsf{PMI}}(z_i, z_j). \tag{35}$$

Therefore, it remains to show that Equation (32) is true. Note that

$$-C \triangleq \min_i \frac{1}{N} \sum_k K_{\mathsf{PMI}}(z_i, z_j) \geq \frac{N-1}{N} \log \rho_{\mathsf{min}} + \frac{1}{N} (\min_i K_{\mathsf{PMI}}(z_i, z_i)). \tag{36}$$

Therefore, it suffices to have

$$\log \rho_{\mathsf{min}} + \delta \leq \frac{N-1}{N} \log \rho_{\mathsf{min}} + \frac{1}{N} (\min_i K_{\mathsf{PMI}}(z_i, z_i)). \tag{37}$$

Rearranging terms gives the desired condition

$$\frac{P_{\mathsf{coor}}(z_i \mid z_i)}{P_{\mathsf{coor}}(z_i)} \geq e^{N\delta} \rho_{\mathsf{min}}, \qquad \forall i. \tag{38}$$

□

Remark F.2. Proposition F.1 is one example that a sufficiently smooth world or a sufficiently high sampling rate allows the PMI kernel K_{PMI} to be *exactly* represented as inner products of a learned feature space (up to a scale). The condition here can be satisfied, for example, if the off-diagonal terms decay linearly with respect to N and stay sufficiently close to each other. While the condition is somewhat strict, it captures the essence that smoothness and continuity allow easier learning. Nonetheless, we note that exact representation is not necessary for convergence, and thus this requirement can likely be relaxed. Please see Section 6 for discussions on practical settings.