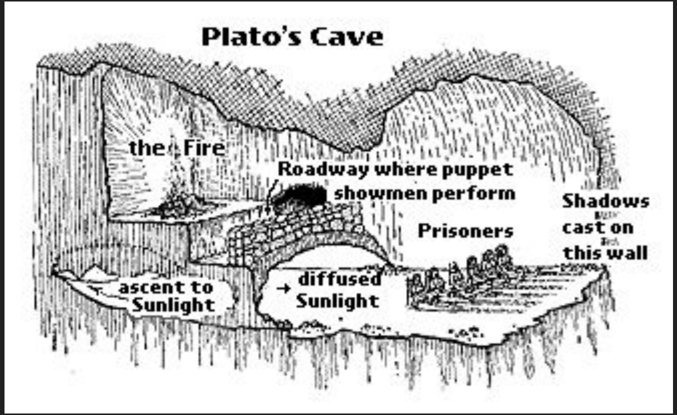
The Platonic Representation Hypothesis

Milton Lin, Krithika Ramesh

What is the Platonic Representation Hypothesis?



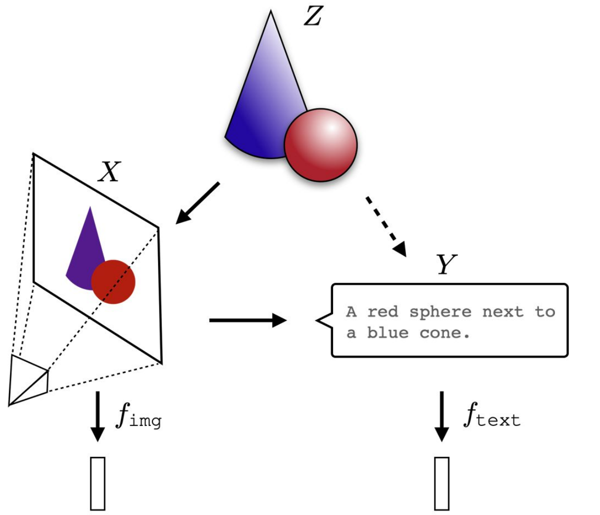


Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this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

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

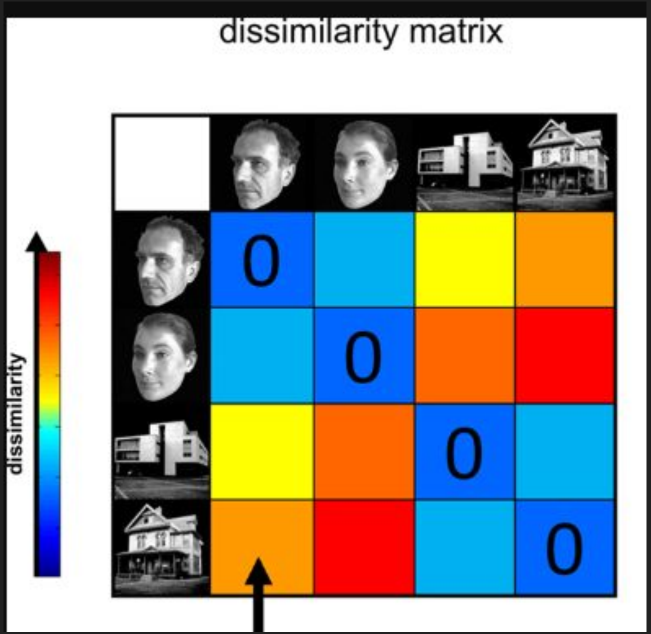
What do we mean by alignment of representations?

Terminology

- **Representation**: A function that maps every input to a feature vector f:X->V.
- **Kernel**: (of a representation) A measure of the distance/similarity between multiple input points
- **Kernel-alignment metric**: Measures the similarity between two kernels

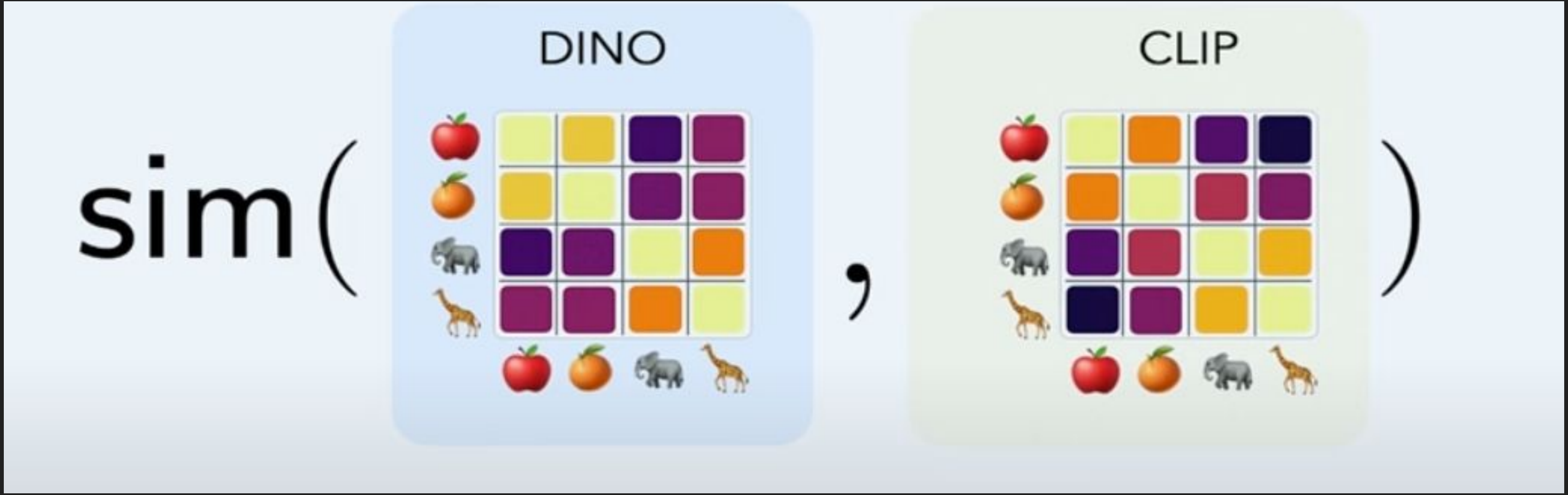
Example of kernels: Representation Dissimilarity Matrix

- Kriegeskorte et al. : Representation Dissimilarity Matrix in neuroscience.



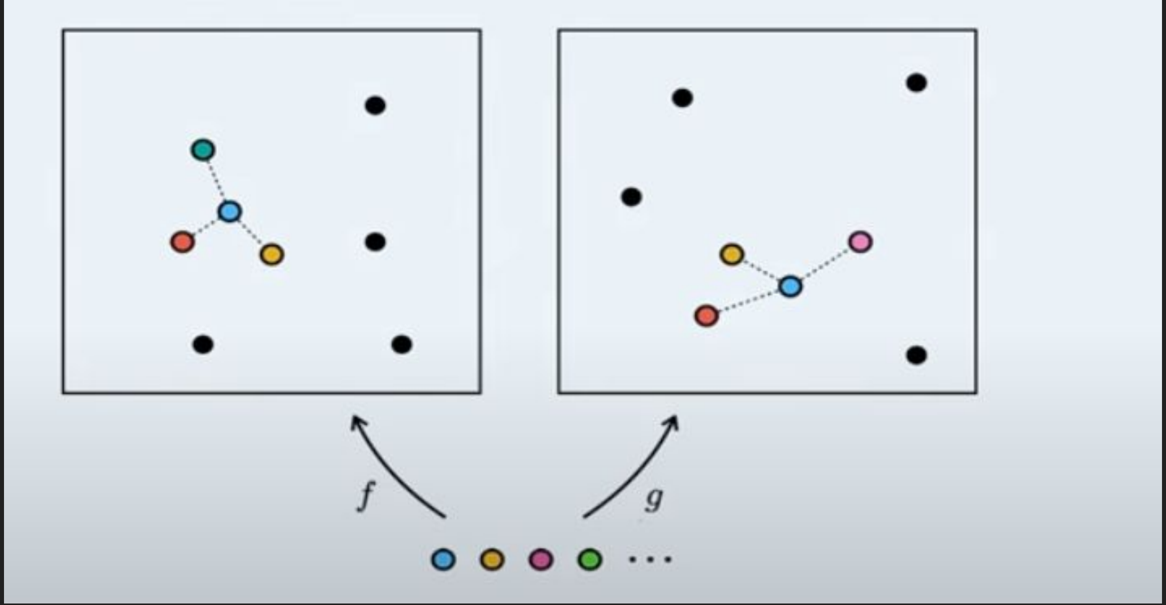
How to measure alignment of representations?

- Given two models f,g we compare their Kernels (previous slide) via **Mutual k-Nearest Neighbor Alignment Metric [appendix B]**



Nearest-neighbor kernel-alignment metric

a) Fix a node (blue). b) Find the k-nearest neighbors for each model f,g. c) Compare common neighbors.

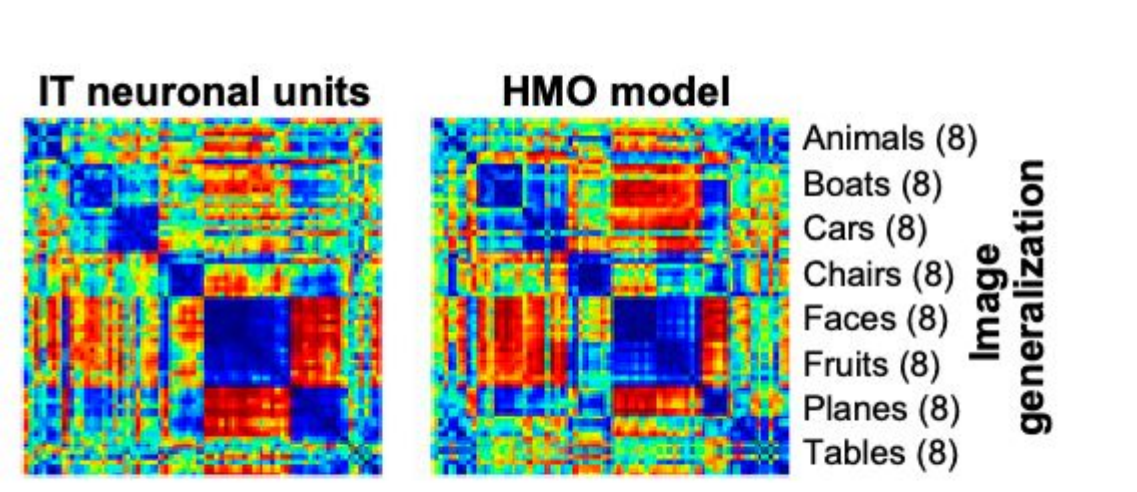


Models are increasingly aligning to brains

- Neural networks show alignment with biological brain representations, likely due to facing similar task and data constraints
- Studies show agreement between how humans and models perceive visual similarity even when models are trained on tasks that are seemingly unrelated to human perception.

Example: Neuroscience and Deep Neural Nets

- Yamins et al. Each entry in the matrix quantifies how dissimilar the neural population's response is to a pair of visual stimuli.



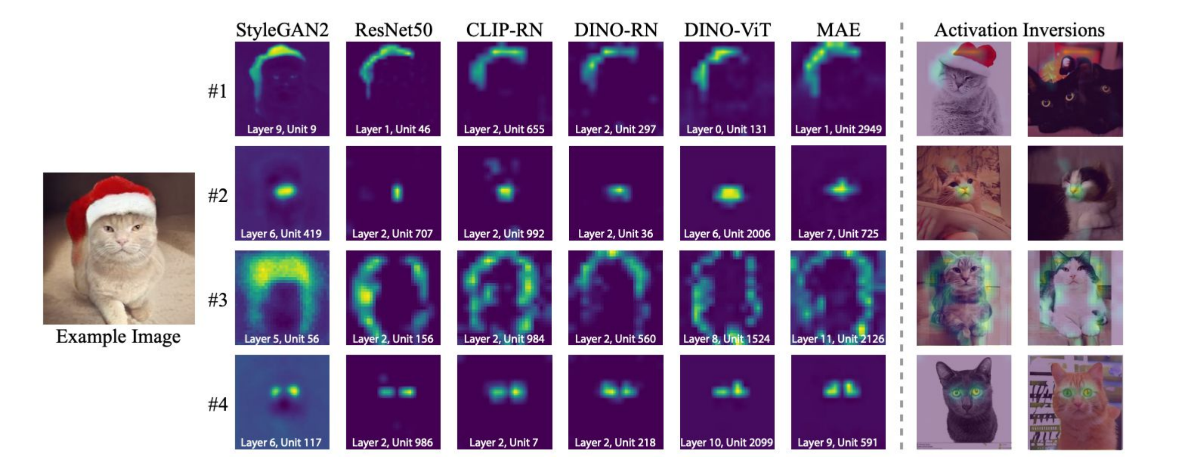
Background

Models with different architectures & objectives have aligned representations

- Lenc & Vedaldi (2015):
  - Vision model trained on ImageNet aligns with Places-365 model while maintaining performance.
  - Early layers of convolutional networks are more interchangeable than later layers.
- Bansal et al. (2021):
  - Model stitching shows close alignment between self-supervised and supervised models.
- Moschella et al. (2022):
  - Zero-shot model stitching is feasible without a stitching layer.
  - Different text models embed data similarly despite training on different modalities.

Example: Rosetta Neurons

- Dravid et al. (2023) - similar neurons that activate across different vision models:



Alignment of representations increases with scale and performance

Model alignment increases with model scale and dataset size.

Models with similar outputs tend to have similar internal activations and models with the same architecture often converge to the same weight basin.

\

This convergence occurs even with different initializations, allowing for permutations in weight space.

Merging separately trained models of the same architecture can combine their capabilities.

Alignment of representations increases with scale and performance

- They evaluate the performance of 78 vision models, trained with varying architecture, training objectives and datasets.

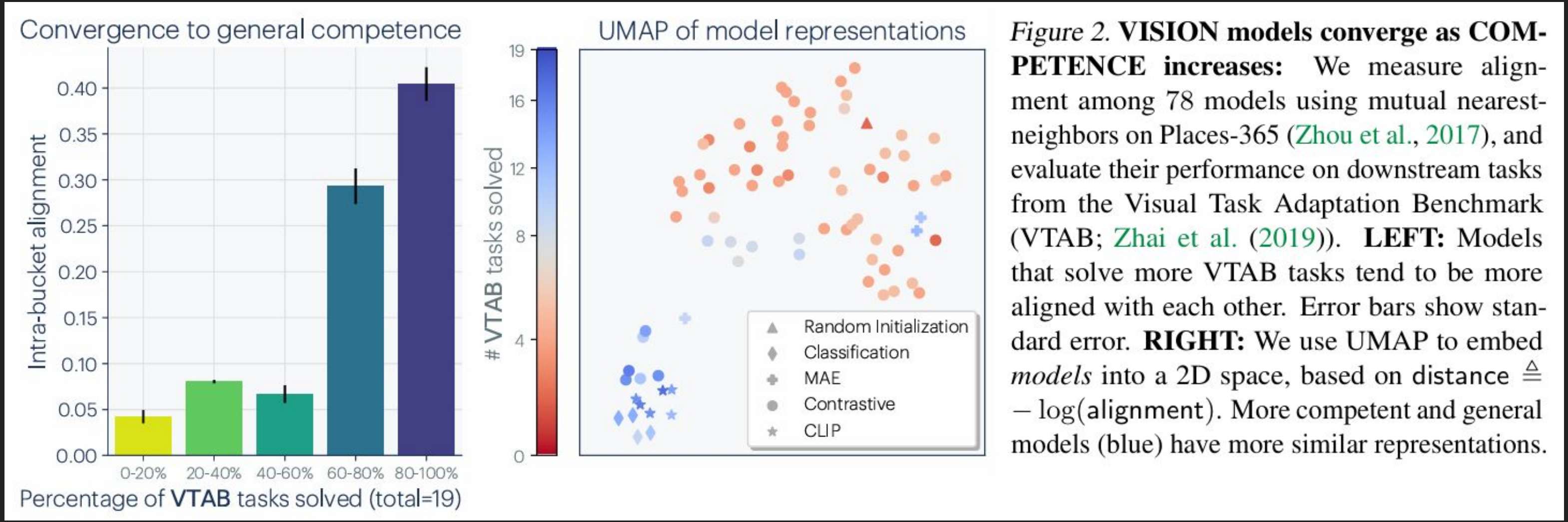


Figure 2. **VISION models converge as COMPETENCE increases:** We measure alignment among 78 models using mutual nearest-neighbors on Places-365 (Zhou et al., 2017), and evaluate their performance on downstream tasks from the Visual Task Adaptation Benchmark (VTAB; Zhai et al. (2019)). **LEFT:** Models 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 distance $\triangleq -\log(\text{alignment})$. More competent and general models (blue) have more similar representations.

Representations are converging across modalities

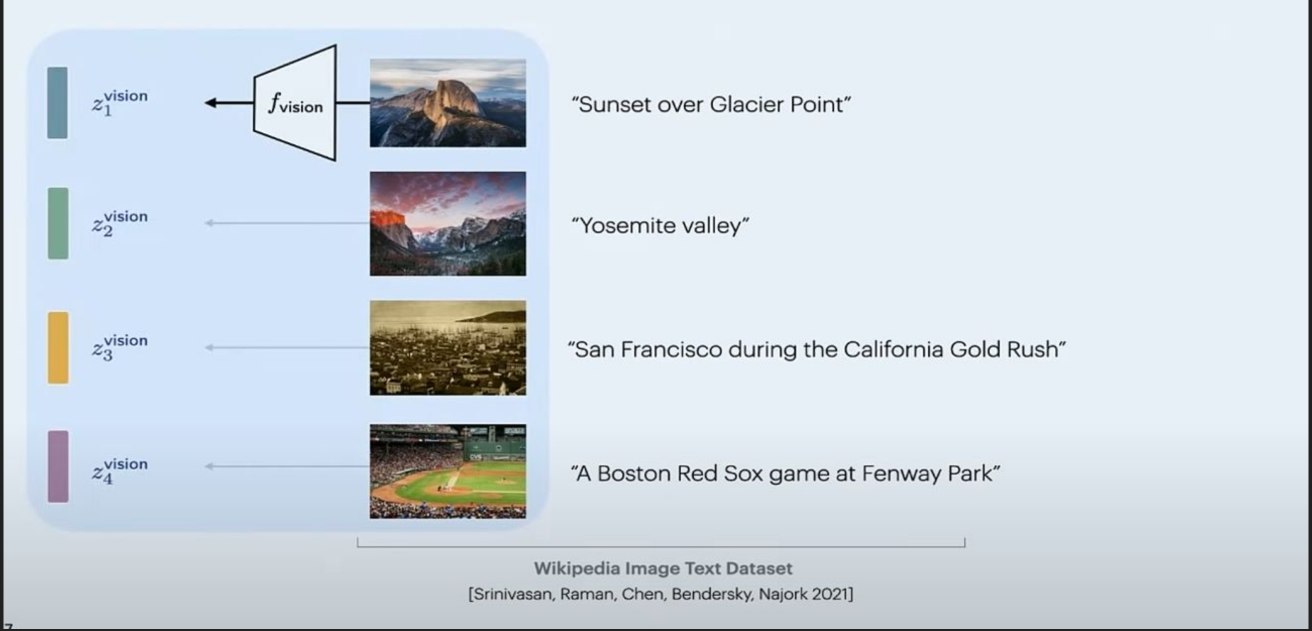
Do models trained on different data modalities also converge? ***Yes.***

- Merullo et al. (2022) showed that a single linear projection can effectively "stitch" a vision model to a language model for visual question answering and image captioning. This approach has also been found to work for aligning text inputs to visual outputs.
- LLaVA (Liu et al., 2023) achieve SoTA results by stitching pre-trained language and vision models.
- OpenAI discovered that jointly training language and vision models improves language task performance compared to training the language model alone.
- Prior work has also found that there is high semantic similarity between vision and language encoders across various training setups.

This work assesses alignment between vision and language models by using paired datasets to bridge the two modalities. They make use of Wikipedia images and the corresponding captions and then measure the alignment between the kernels corresponding to the language and vision models.

How to compare across model?

Use paired data: The Wikipedia caption dataset (WIT): paired image-text samples.



“Sunset over Glacier Point”

“Yosemite valley”

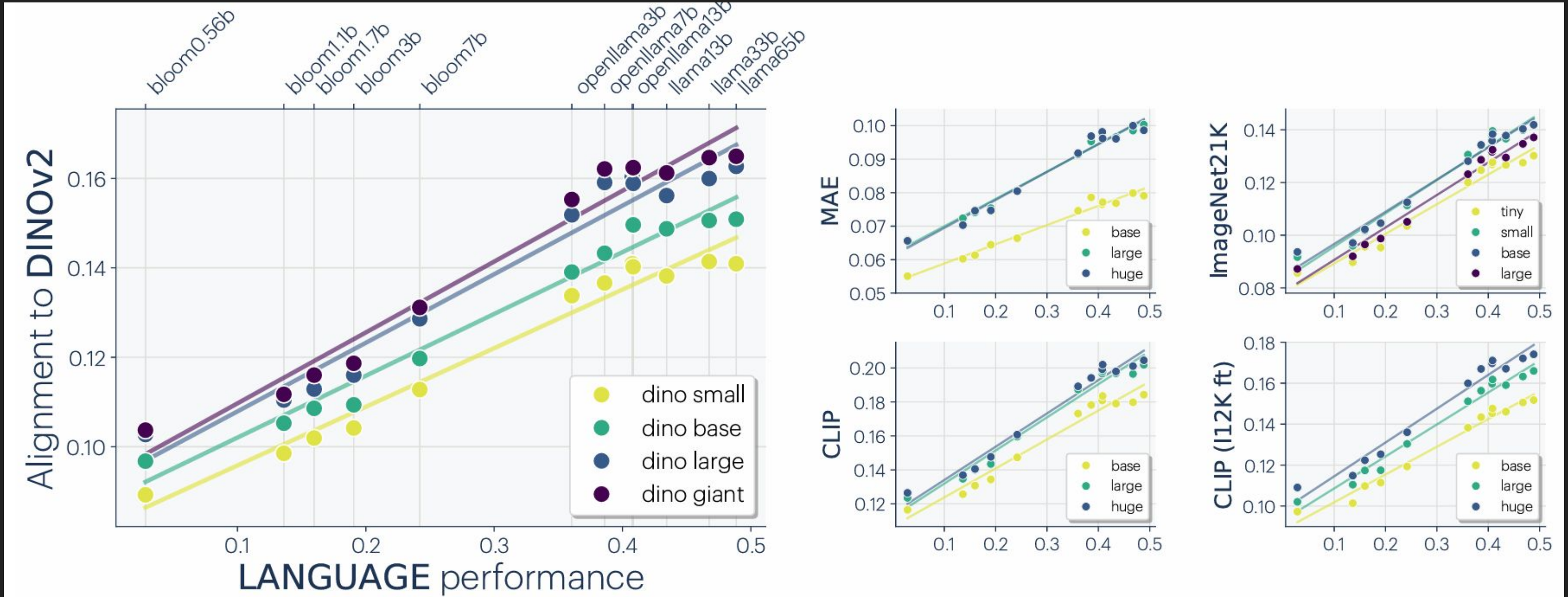
“San Francisco during the California Gold Rush”

“A Boston Red Sox game at Fenway Park”

Wikipedia Image Text Dataset

[Srinivasan, Raman, Chen, Bendersky, Najork 2021]

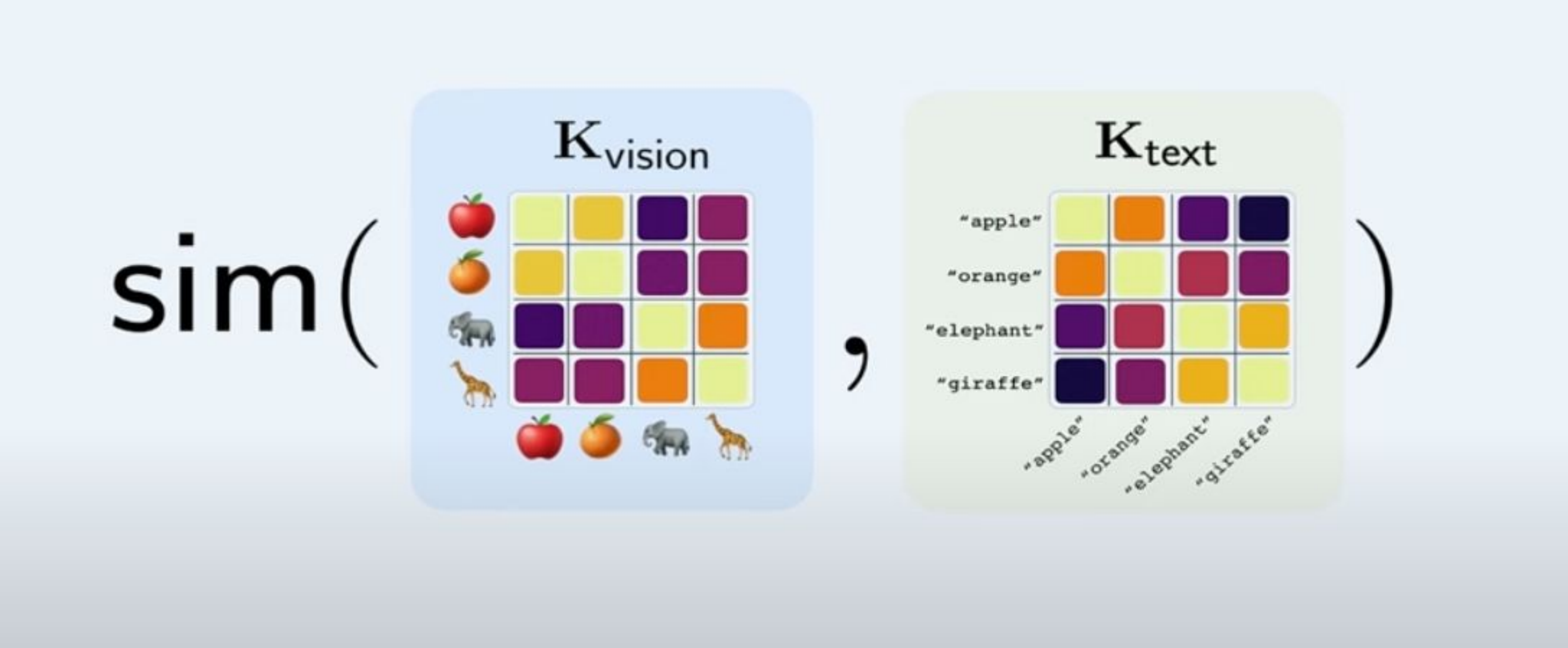
Representations are converging across modalities



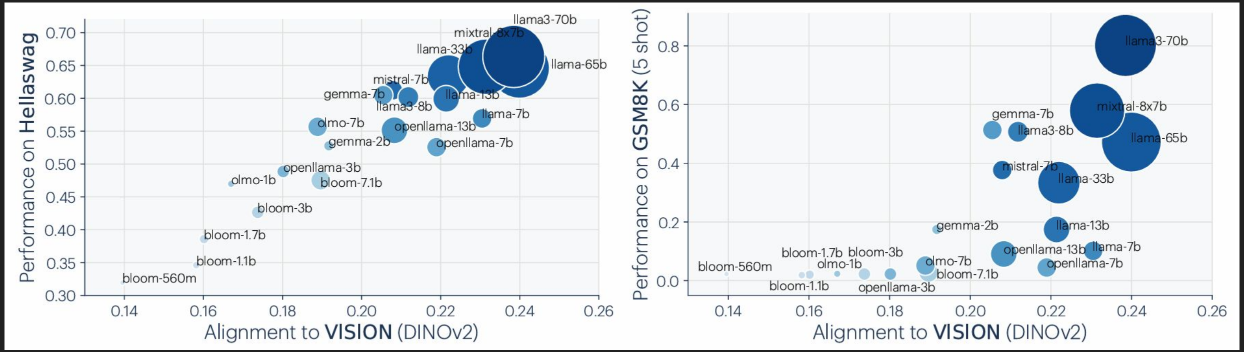
The better the LLM is at language modelling, the more it tends to align with vision models.

How to compare across model?

Use paired data: The Wikipedia caption dataset (WIT): paired image-text samples.



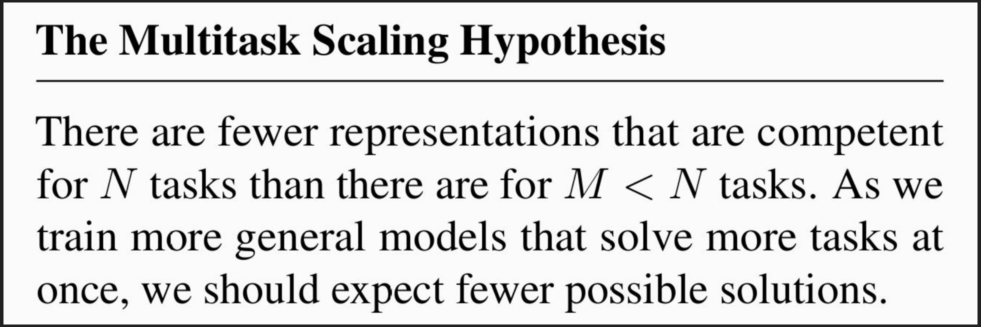
Does alignment predict downstream performance?



Why are representations converging?

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

Convergence via Task Generality



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.

Each data point places an additional constraint on the model.

As data scales, models that optimize the empirical risk also improve on the population risk.

Modern representation learning objectives optimize for multi-task solving anyway.

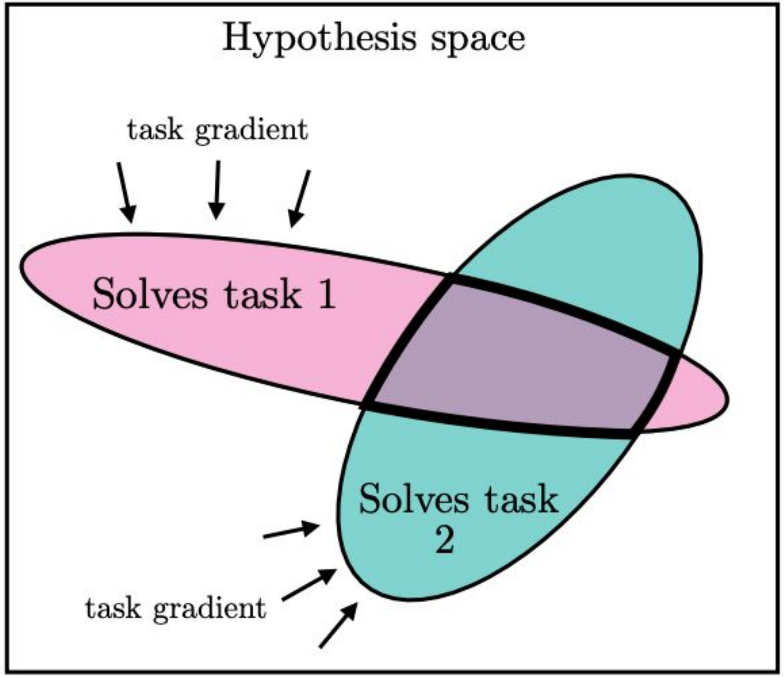
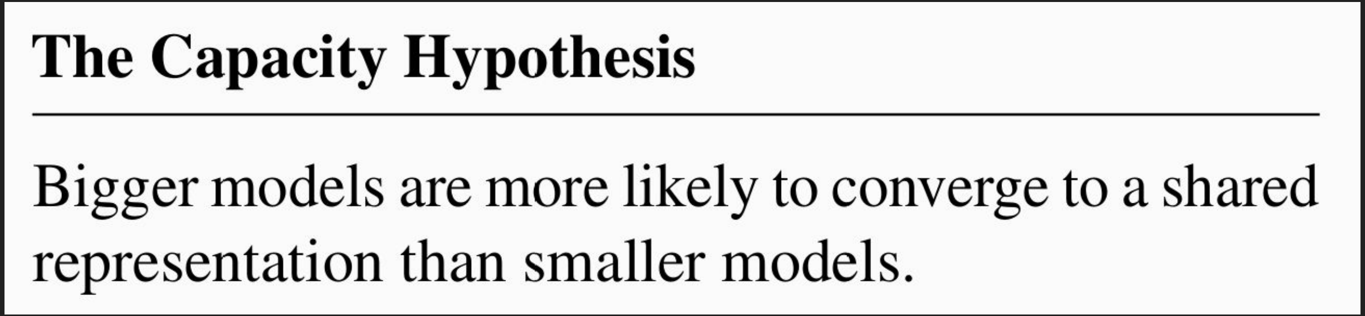


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

Convergence via Model Capacity



The Capacity Hypothesis

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

Scaling a model should be more effective at findings better approximations to the optimum.

With the same training objective, larger models, even with diff architectures will converge to the optimum.

When different training objectives share similar minimizers, larger models are better at findings these minimizers.

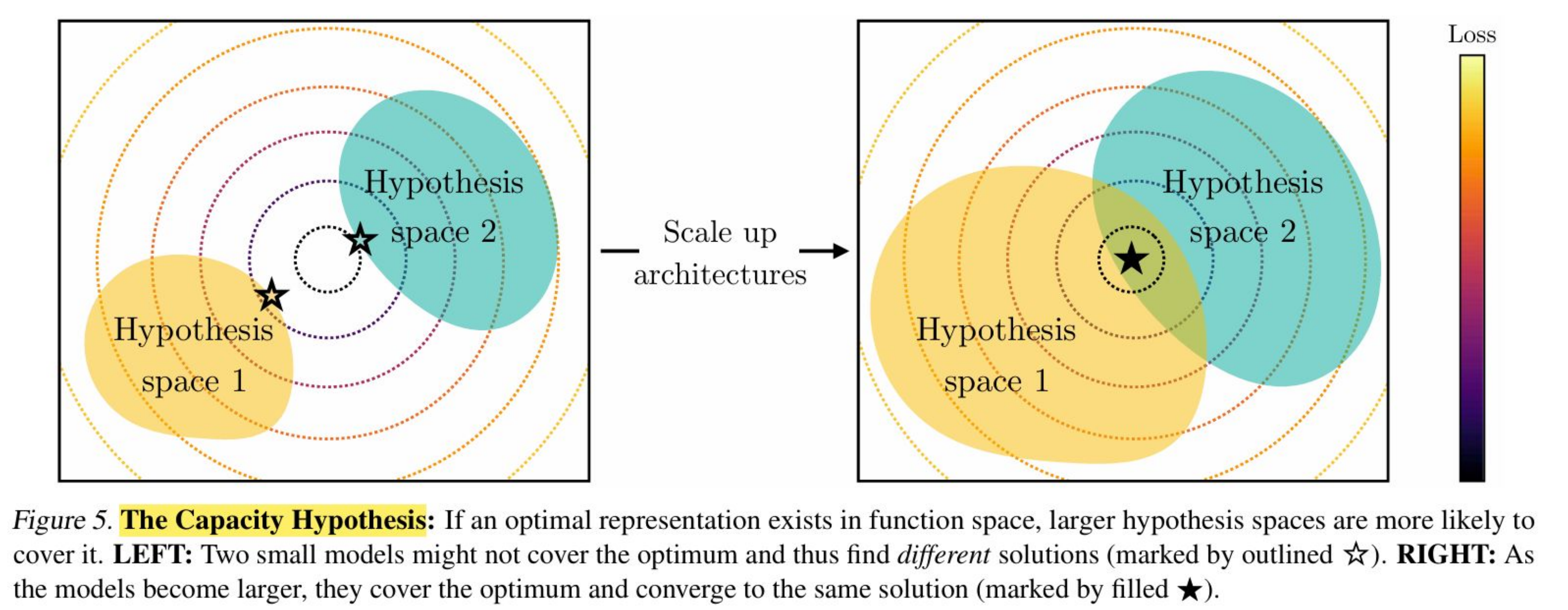
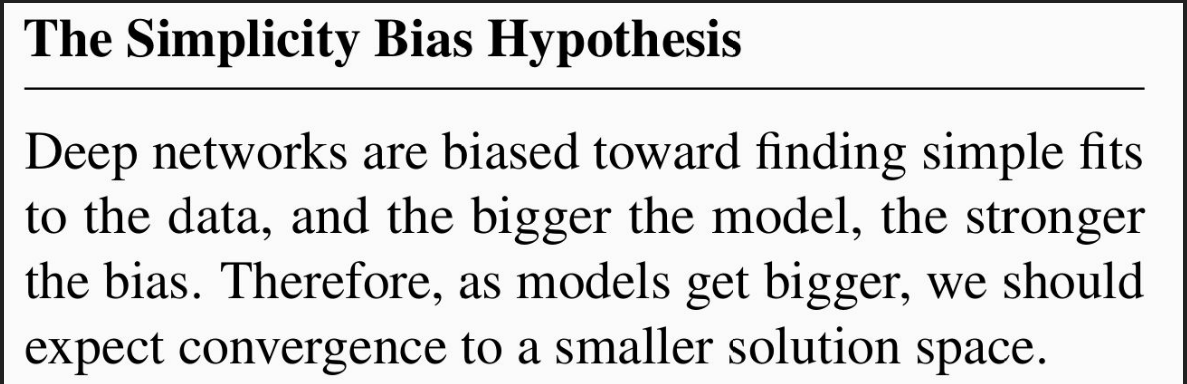


Figure 5. **The Capacity Hypothesis:** If an optimal representation exists in function space, larger hypothesis spaces are more likely to cover it. **LEFT:** Two small models might not cover the optimum and thus find *different* solutions (marked by outlined ☆). **RIGHT:** As the models become larger,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

The Capacity Hypothesis

Convergence via Simplicity Bias



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.

What prevents models from developing distinct internal representations?

Such simplicity bias could come from explicit regularization (e.g weight decay and dropout)

Even in the absence of regularization, deep networks implicitly favor simple solutions.

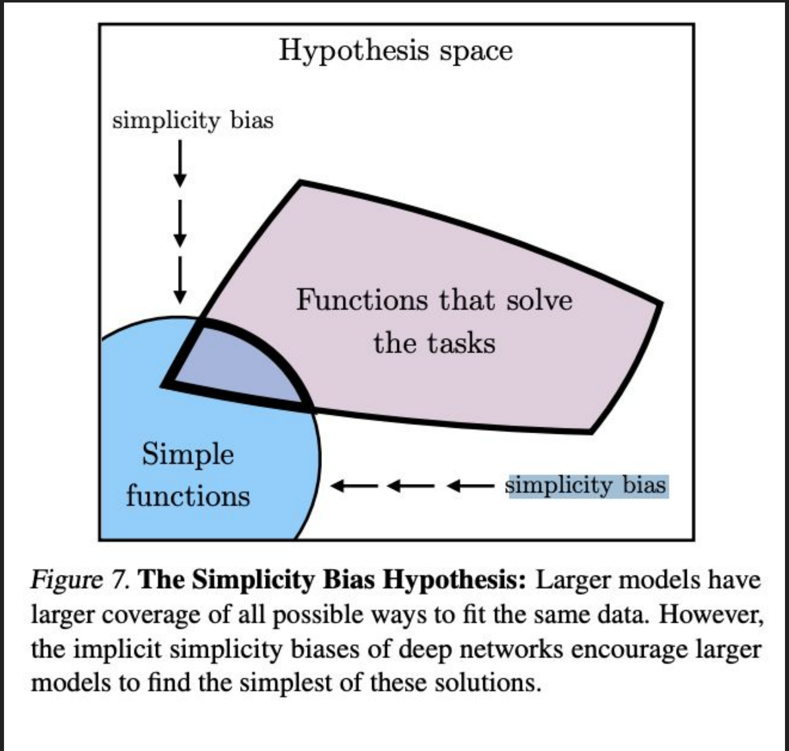
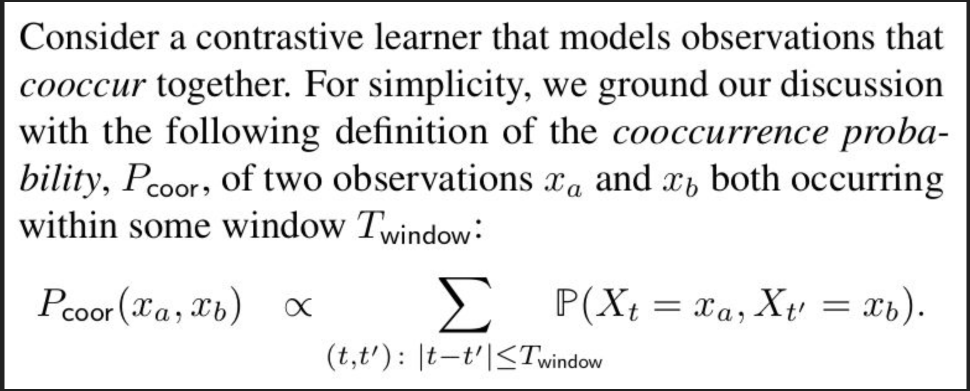


Figure 7. **The Simplicity Bias Hypothesis:** Larger models have larger coverage of all possible ways to fit the same data. However, the implicit simplicity biases of deep networks encourage larger models to find the simplest of these solutions.

Convergence via Simplicity Bias

A family of contrastive learners



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}:

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

A study of color

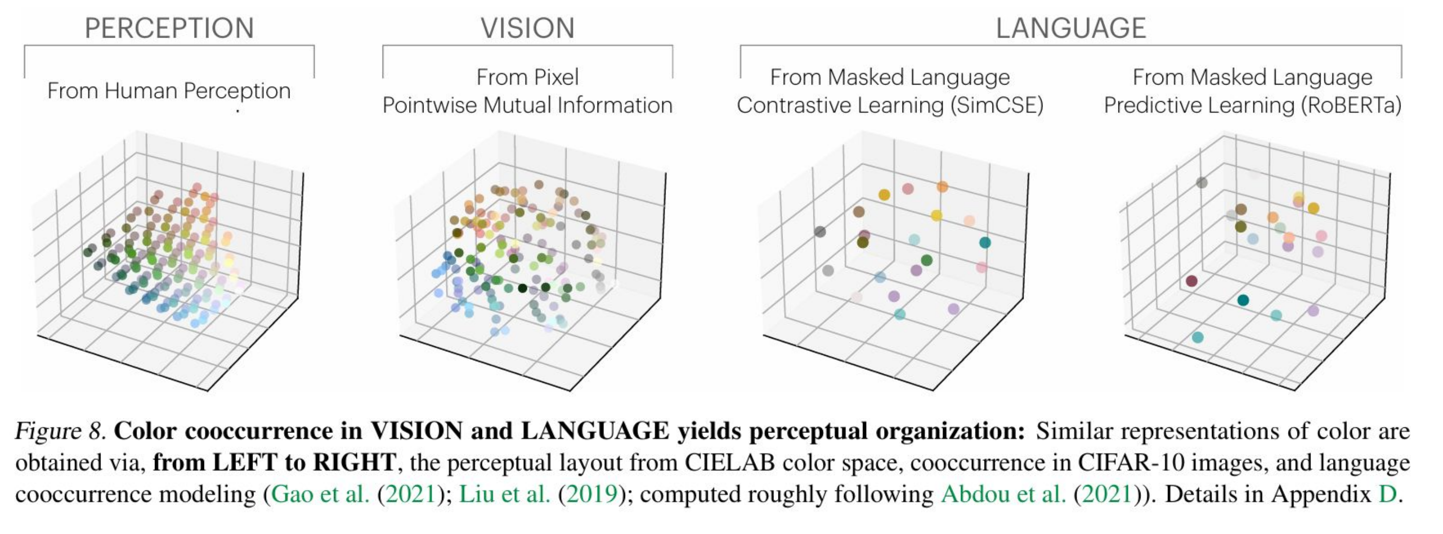


Figure 8. **Color cooccurrence in VISION and LANGUAGE yields perceptual organization:** Similar representations of color are obtained via, **from LEFT to RIGHT**, the perceptual layout from CIELAB color space, cooccurrence in CIFAR-10 images, and language cooccurrence modeling (Gao et al. (2021); Liu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

Implications of convergence

- Scaling is sufficient but not necessarily efficient
- Training data can be shared across modalities
- Ease of translation and adaptation across modalities
- Scaling may reduce hallucinations and bias

Limitations I

- Different modalities may contain different information
    - Does this apply to scenarios where these modalities are orthogonal to each other?
- The work presently focuses on vision and language modalities, and there is no telling that this hypothesis will translate across modalities
    - For example, the state space in robotics is presently not well-defined in that there is no standardized approach

Limitations II

- Collective preferences of researchers/developers within the community may filter down the pipeline and shape the trajectory of model development
    - Hardware lottery hypothesis challenges us to ask: Are there better alternatives to developing systems that mimic human reasoning?
- Other questions regarding measuring alignment, the convergence of special-purpose intelligence, etc.