MODULATING CROSS-MODAL CONVERGENCE WITH SINGLE-STIMULUS, INTRA-MODAL DISPERSION

Eghbal A. Hosseini[1]*, Brian Cheung[2], Evelina Fedorenko[1] & Alex H. Williams[3,4]
[1]Department of Brain and Cognitive Sciences, MIT
[2]CSAIL, MIT
[3]Center for Neural Science, NYU
[4]Center for Computational Neuroscience, Flatiron Institute
{ehoseini, evelina9}@mit.edu, cheungb@mit.edu, aw4614@nyu.edu

ABSTRACT

Neural networks exhibit a remarkable degree of representational convergence across diverse architectures, training objectives, and even data modalities. This convergence is predictive of alignment with brain representation. A recent hypothesis suggests this arises from learning the underlying structure in the environment in similar ways. However, it is unclear how individual stimuli elicit convergent representations across networks. An image can be perceived in multiple ways and expressed differently using words. Here, we introduce a methodology based on the Generalized Procrustes Algorithm to measure intra-modal representational convergence at the single-stimulus level. We applied this to vision models with distinct training objectives, selecting stimuli based on their degree of alignment (intra-modal dispersion). Crucially, we found that this intra-modal dispersion strongly modulates alignment between vision and language models (cross-modal convergence). Specifically, stimuli with low intra-modal dispersion (high agreement among vision models) elicited significantly higher cross-modal alignment than those with high dispersion, by up to a factor of two (e.g., in pairings of DINOv2 with language models). This effect was robust to stimulus selection criteria and generalized across different pairings of vision and language models. Measuring convergence at the single-stimulus level provides a path toward understanding the sources of convergence and divergence across modalities, and between neural networks and human neural representations.

1 INTRODUCTION

Artificial neural networks show a remarkable ability to learn representations that generalize across tasks, and predict neural representations in humans and other animals. Even though details of implementation — including training, architecture, and input modality — vary between networks, they appear to have minimal influence on the final representation, and models converge onto similar representations. The Platonic Representation Hypothesis (Huh et al., 2024) suggests this convergence arises from learning shared environmental priors. Hosseini et al. (2024) found similar evidence in convergence across neural networks and their predictivity of brain representation. Both lines of work however established convergences across groups of stimuli, and thus globally.

Individual observations are critical for probing and understanding representations and can potentially drive models, and humans, toward either convergence or divergence. Humans often find themselves interpreting differently the same work of art, a painting for example. Existing methods are often ill-suited for probing convergence at single stimulus level, as they either measure local effects (Feather et al., 2023), average similarity over a large set of stimuli datasets (Hosseini et al., 2024), or lack a proper metric space for rigorous comparison (Sucholutsky et al., 2024; Harvey et al., 2023). This highlights the need for a robust, stimulus-specific measure of alignment.

*Now at Google DeepMind

Building on Generalized Procrustes Analysis (Gower, 1975), we introduce a stimulus-specific measure of representational convergence across vision models (Williams et al., 2021; Haxby et al., 2020). We then demonstrate that our measure can be used to select stimuli that effectively modulate representational alignment across modalities, from vision to language.

2 METHODS

2.1 GENERALIZED PROCRUSTES METHOD

Generalized Procrustes Analysis (GPA) is a method for aligning multiple datasets, or embedding spaces, into a common reference frame. The core idea is to find a set of optimal transformations, restricted to orthogonal transformations, that minimize the discrepancy between a shared consensus configuration and each individually transformed dataset. Stimulus-level residuals are a classical by-product of GPA (Gower, 1975); our contribution is their application to comparing neural network representations and to modulating cross-modal alignment.

Formally, consider a set of M representation matrices $\{N_1, N_2, \ldots, N_M\}$, where each representation $N_i \in \mathbb{R}^{m \times n_i}$, with $i \in \{1, \ldots, M\}$ indexing models, represents m corresponding samples (e.g., stimuli) in an n_i-dimensional space. The dimensionality n_i can differ across the representations. GPA seeks to find an optimal consensus representation, $N_{\text{joint}} \in \mathbb{R}^{m \times k}$, and a set of corresponding orthogonal transformation matrices, $\{T_1, T_2, \ldots, T_M\}$, where each $T_i \in \mathbb{R}^{n_i \times k}$, by minimizing the sum of squared Frobenius distances. The optimization problem is defined as:

$$\min_{N_{\text{joint}}, \{T_i\}_{i=1}^{M}} \sum_{i=1}^{M} \|N_i T_i - N_{\text{joint}}\|_F^2 \tag{1}$$

subject to each transformation matrix being orthogonal, $T_i^T T_i = I_k$ for all $i \in \{1, \ldots, M\}$. Here, $\|\cdot\|_F$ denotes the Frobenius norm, and the constraint $T_i^T T_i = I_k$ ensures that the transformations do not distort the internal geometry of each representation N_i. This process effectively rotates each embedding space to achieve maximal alignment with the emergent consensus space N_{joint}. Prior to performing GPA, we first center each embedding along the second dimension n_i, zero-pad dimensions across all models to a common dimensionality, and normalize each representation $\|N_i\|_F = 1$.

Generalized Procrustes Analysis can also be interpreted as computing a **barycenter** of neural representations in Procrustes shape space. In classical shape analysis, each centered, scale-normalized configuration N_i corresponds to a point on a quotient manifold obtained by modding out rotations,[1] and Generalized Procrustes iteratively aligns all configurations to the barycenter (also called the Fréchet mean or Karcher mean), represented by N_{joint}. This perspective clarifies that GPA is not merely an arbitrary linear alignment, but a principled procedure for finding the central tendency of a population of neural representations with respect to a scale and rotation-invariant metric space (Williams et al., 2021).

We restrict each T_i to orthogonal transformations, rather than a general linear map, because they preserve the internal geometry of each representation: pairwise distances and angles between stimuli are unchanged under T_i. A general linear map can absorb scale and reshape covariance structure, which are not desirable for defining metrics on representation (Williams et al., 2021). Scale is instead handled explicitly by the Frobenius normalization $\|N_i\|_F = 1$.

2.2 QUANTIFYING SINGLE-STIMULUS DISPERSION

To quantify representational convergence at the single-stimulus level, we measure the dispersion between each model's transformed representation and the shared joint representation, N_{joint}. Specifically, after computing the joint space via GPA, we first project each model's representation into this common space using its corresponding transformation matrix, T_i. For each stimulus $j \in \{1, \ldots, m\}$, we then compute the Euclidean distance between its projected representation from

[1] Here we include reflections as well as rotations in the equivalence class. This is common in the analysis of neural representations since even a permutation of the neural unit labels can result in a reflection.

model i and its joint representation. This yields the **Procrustes residual**, r_{ij}, defined as:

$$r_{ij} = \|(N_i T_i)_j - (N_{\text{joint}})_j\|_2 \tag{2}$$

This process results in a residual matrix $R \in \mathbb{R}^{m \times M}$, where each entry r_{ij} quantifies the dispersion for a given stimulus j and model i (see Fig 1C). A low residual value signifies high alignment (low dispersion), while a high value indicates low alignment (high dispersion).

We employ two strategies to summarize the overall dispersion for each stimulus across all M models:

1. **Mean Dispersion:** Our first approach is to average the residuals across all models for each stimulus. We rank the stimuli from least to most dispersed. While straightforward, this metric can be sensitive to outlier models (Fig 1D)
2. **Principal Component of Dispersion:** To capture the primary axis of disagreement in a more robust manner, our second approach utilizes Principal Component Analysis (PCA) on the residual matrix R. We use the score along the first principal component (PC1) as a more nuanced measure of dispersion for each stimulus, as it reflects the most significant shared variance in model disagreement (Fig 1E).

2.3 Vision Models

We inspected representations across ViT models trained on diverse objectives (Huh et al., 2024). These objectives include Masked Autoencoders (MAE) (He et al., 2022), DINO (Caron et al., 2021), CLIP (Radford et al., 2021), and CLIP with additional finetuning on ImageNet-12K. To compute the joint Procrustes representation, we selected a large architecture in each class: MAE: `vit_huge_patch14_224.mae`; DINO: `vit_giant_patch14_dinov2.lvd142m`; CLIP: `vit_huge_patch14_clip_224.laion2b`; CLIP + finetuning on ImageNet-12K: `vit_huge_patch14_clip_224.laion2b_ft_in12k`. All ViTs have a dimension of 1280, except for the DINO model, which had a dimension of 1536. We thus zero-padded all other model representations to match the DINO dimensions. We extracted the CLS-token representation from the penultimate transformer block of each ViT and used it for Procrustes alignment, and identified stimuli with varying degrees of dispersion.

2.4 Language Models

Similar to Huh et al. (2024) we compared representations of ViT vision model to LLMs across model families. We focused on BLOOM (Workshop et al., 2023) , OpenLLaMA (Geng & Liu, 2023), and LLaMA model class (Touvron et al., 2023). For each LLM, we extracted token embeddings from each transformer block and mean-pooled across tokens to obtain a single representation per caption, we then used the block that showed best alignment with the vision models.

2.5 Dataset

Following Huh et al. (2024), we used Wikipedia caption dataset (Srinivasan et al., 2021) to measure convergence between modalities. This dataset included a set of 1024 image/caption pairs from Wikipedia articles. Images and captions are used as released in the Wikipedia caption dataset (WIT); no additional filtering was applied.

2.6 Alignment measure

To measure the local alignment between vision and language modalities, we employed Centered Kernel k-Nearest Neighbors Alignment (CKNNA) (Huh et al., 2024; Kornblith et al., 2019). This method adapts the standard Centered Kernel Alignment (CKA) to focus on local representational structure. Intuitively, CKNNA computes the cross-covariance only between a sample and its nearest neighbors, thereby assessing the local, rather than global, similarity between two representation spaces. We considered 10 nearest neighbors for each sample when reporting alignment (Fig. 2).

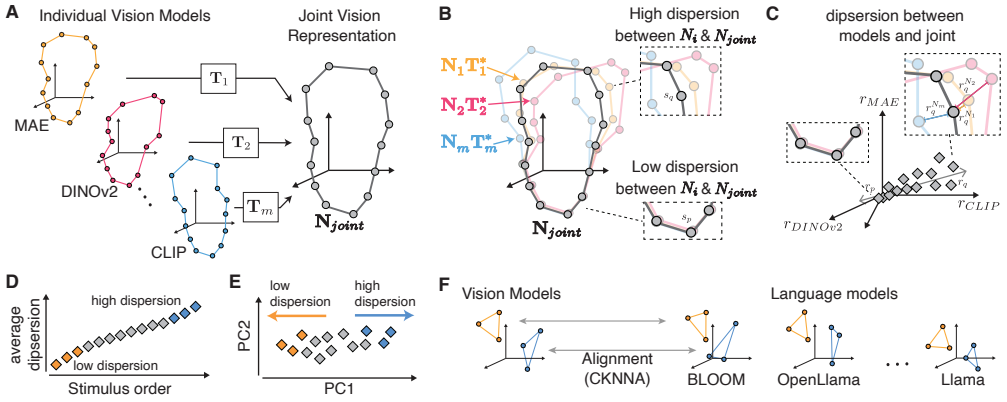


Figure 1: **(A)** Overview of the Generalized Procrustes Analysis (GPA) problem. Given representations from diverse vision models, our goal is to learn a set of model-specific transformations to construct a single joint representation. **(B)** When individual model representations are projected onto this joint space, stimuli can exhibit either low dispersion between joint and individual models (bottom) or instead high dispersion (top). **(C)** This dispersion is quantified in a residual space, where each stimulus is represented as a point whose coordinates reflect its distance between each model's representation and the joint space. **(D, E)** We use two approaches to identify stimuli with varying degrees of dispersion: (D) ranking stimuli based on their average dispersion across all models and (E) using the score along the first principal component (PC1) of the residual space as a more robust measure. **(F)** After identifying stimuli with high and low dispersion, we measure the alignment between the vision models and language models using a local similarity metric (CKNNA) to test how stimulus selection modulates cross-modal convergence.

3 Results

We investigated the dependence of convergence between visual and linguistic representations on stimuli using the Procrustes-based methodology described above. As a stricter test of the Platonic Representation Hypothesis, we considered representations learned across ViT transformer architectures with different objectives, an effect not explored in the original work. After computing a joint representation across these vision models, we extracted stimuli with varying degrees of dispersion to test their impact on cross-modal alignment.

Stimuli with low intra-modal dispersion yielded a substantially higher degree of cross-modal convergence between visual and linguistic representations. We first identified stimuli with the least and most dispersion using a rank-ordering of their mean residuals (Fig. 2A). We selected three sets: (1) **low-dispersion**, (2) **high-dispersion**, and (3) **random baseline**. We then compared vision model alignment with three classes of LLMs. The low-dispersion set showed significantly higher convergence, up to a twofold increase in some cases (DinoV2, CLIP+ft on ImN12K), over the random and high-dispersion stimuli (Fig. 2B-E).

The correspondence between intra-modal dispersion and cross-modal convergence generalizes to different sampling strategies. Instead of rank ordering, we performed PCA on the residual space and selected stimuli along the first principal component (PC1), as shown in (Fig. 2F). When we repeated the alignment experiment, we found a consistent pattern where the degree of intra-modal dispersion strongly correlated with cross-modal convergence (Fig. 2G-J). This confirms that the significant gap in alignment between low- and high-dispersion stimuli is a robust phenomenon across all tested models.

4 Discussion

We applied Generalized Procrustes Analysis to identify stimuli with varying representational dispersion across vision models at the single-stimulus level. We showed that selecting for low-dispersion stimuli can boost cross-modal alignment with language models close to twofold, and is robust across different selection criteria.

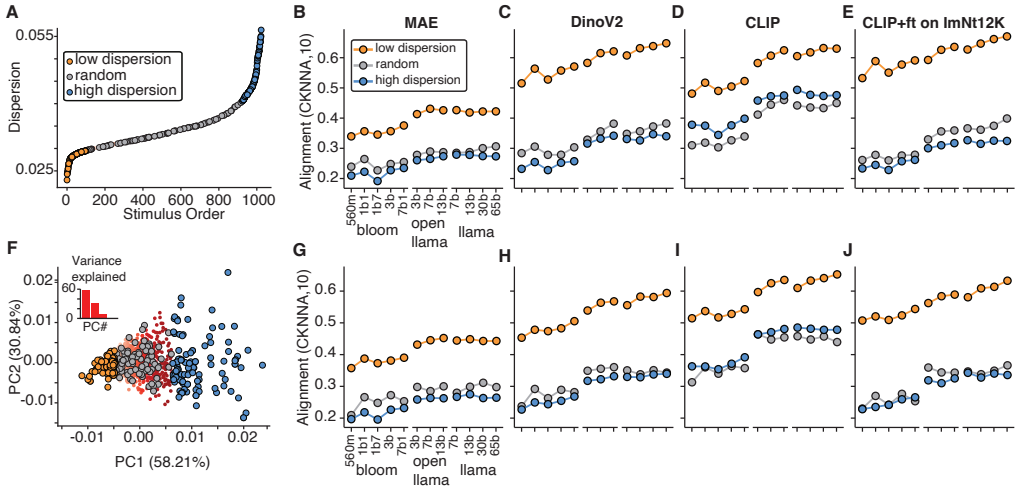


Figure 2: **Stimulus-specific dispersion modulates vision-language alignment.** **(A)** rank based stimulus selection:stimuli are sorted by mean dispersion to create low-, high-, and random-dispersion sets. **(B-E)** Vision-language alignment (CKNNA) is then measured for each set across four vision models (ViT-MAE, DINOv2, CLIP, and CLIP+FT). **(F)** PCA based stimulus selection: stimuli are selected based on their score along the first principal component (PC1) of the dispersion space. **(G-J)** The alignment experiment is repeated using these sets. For both selection methods, low-dispersion stimuli consistently yield the highest vision-language alignment across all tested models.

Our approach helps uncover what features drive representational convergence and could potentially offer a more stringent test for comparing artificial and biological neural networks. Future work could extend these findings to more datasets and, critically, identify the common features within low- and high-dispersion stimuli that are responsible for modulating alignment.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, and Josh H McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nat. Neurosci.*, 26 (11):2017–2034, November 2023.

Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023. URL `https://github.com/openlm-research/open_llama`.

John C. Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Sarah E. Harvey, Brett W. Larsen, and Alex H. Williams. Duality of bures and shape distances with implications for comparing neural representations, 2023. URL `https://arxiv.org/abs/2311.11436`.

James V Haxby, J Swaroop Guntupalli, Samuel A Nastase, and Ma Feilong. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, 9:e56601, jun 2020. ISSN 2050-084X. doi: 10.7554/eLife.56601. URL `https://doi.org/10.7554/eLife.56601`.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Eghbal Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024. doi: 10.1101/2024.12.26.629294. URL https://www.biorxiv.org/content/early/2024/12/26/2024.12.26.629294.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 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*, SIGIR '21, pp. 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463257. URL https://doi.org/10.1145/3404835.3463257.

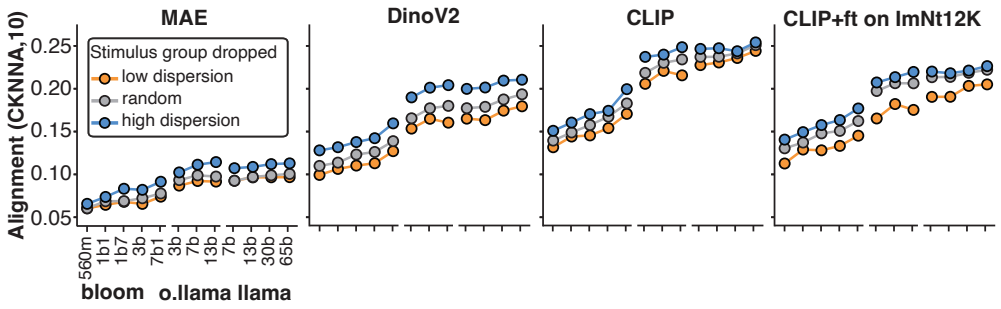
Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, 2024. URL https://arxiv.org/abs/2310.13018.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

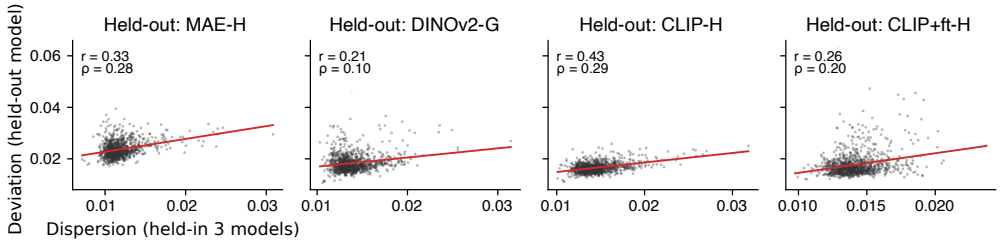
Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, and François Yvon et. al. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL https://arxiv.org/abs/2211.05100.

SUPPLEMENTARY MATERIAL



Supplementary Figure 1: **Dropping the highest-dispersion stimuli systematically increases vision–language alignment, whereas dropping the lowest-dispersion stimuli decreases it**. We removed the 25% of stimuli with the highest mean dispersion (*drop high-dispersion*, blue) or the 25% with the lowest (*drop low-dispersion*, orange), and recomputed the vision-language local alignment (CKNNA, $k = 10$) on the remaining 768 stimuli from WIT. The full-set baseline (*full*, gray) is shown for reference. Each of the four panels corresponds to one vision model (MAE, DINOv2, CLIP, CLIP+FT on ImageNet-12K), with alignment reported against 12 language models from three families: BLOOM (`560m`, `1b1`, `1b7`, `3b`, `7b1`), OpenLLaMA (`3b`, `7b`, `13b`), and LLaMA (`7b`, `13b`, `30b`, `65b`). Across all vision–LLM pairings, dropping high-dispersion stimuli raises CKNNA above the full-set baseline, while dropping low-dispersion stimuli lowers it.



Supplementary Figure 2: **Dispersion computed from a subset of the vision models predicts the held-out model's deviation from the consensus** For each of the four vision models in turn, GPA is re-run on the other three, producing a partial-dispersion score per stimulus (x-axis). The held-out model is then projected into the remaining consensus and its per-stimulus deviation from that consensus is measured (y-axis). Each point is one WIT stimulus (n = 1024); the red line is the linear fit. Pearson r is significantly positive for all four held-out models: MAE-H $r = 0.33$, $p < 0.001$; DINOv2-G $r = 0.21$, $p < 0.001$; CLIP-H $r = 0.43$, $p < 0.001$; CLIP+ft-H $r = 0.26$, $p < 0.001$ (ρ values show Spearman correlations). This suggests that stimulus-level dispersion captures a property shared across model families.