Back into Plato's Cave: Examining Cross-modal Representational Convergence at Scale

A. Sophia Koepke[1,2,3], Daniil Zverev[2], Shiry Ginosar[4], and Alexei A. Efros[1]

[1]UC Berkeley [2]Technical University Munich, MCML
[3]University of Tübingen, Tübingen AI Center [4]Toyota Technical Institute at Chicago

Abstract. The Platonic Representation Hypothesis [40] suggests that neural networks trained on different modalities (e.g., text and images) align and eventually converge toward the same representation of reality. If true, this has significant implications for whether modality choice matters at all. We show that the experimental evidence for this hypothesis is fragile and depends critically on the evaluation regime. Alignment is measured using mutual nearest neighbors on small datasets (≈1K samples) and degrades substantially as the dataset size is scaled up to millions of samples. The alignment that remains between model representations primarily reflects coarse semantic overlap rather than consistent fine-grained structure. Moreover, the evaluations in Huh et al. [40] are done in a one-to-one image-caption setting, a constraint that breaks down in realistic many-to-many settings and further reduces alignment. We also find that the reported trend of stronger language models increasingly aligning with vision does not hold for newer models. Overall, our findings suggest that the current evidence for cross-modal representational convergence is considerably weaker than subsequent works have taken it to be. Models trained on different modalities may learn equally rich representations of the world, just not the same one.

Project page: https://akoepke.github.io/cave_umwelten

1 Introduction

The success of Large Language Models (LLMs) is causing much hand-wringing in the computer vision community: do we even need pixels to build machines that understand our world, or is language "all you need"?

Several works have demonstrated that models trained only on text data have made progress in solving what were thought to be fundamentally visual problems, such as visual question answering (VQA) [30, 38], visual reasoning [2, 10, 37, 91], or embodied robotics applications [1, 54]. This resonates with the suggestion that text data may make other modalities redundant [83], on the premise that the part of the world that is relevant to humans is manifest in language. On the other hand, it is argued that linguistic data alone cannot yield genuine understanding [8] or allow actual embodiment. After all, there is a reason we visit art museums rather than just read descriptions of paintings in a catalogue. This raises a central question: how do models trained on different modalities represent reality?

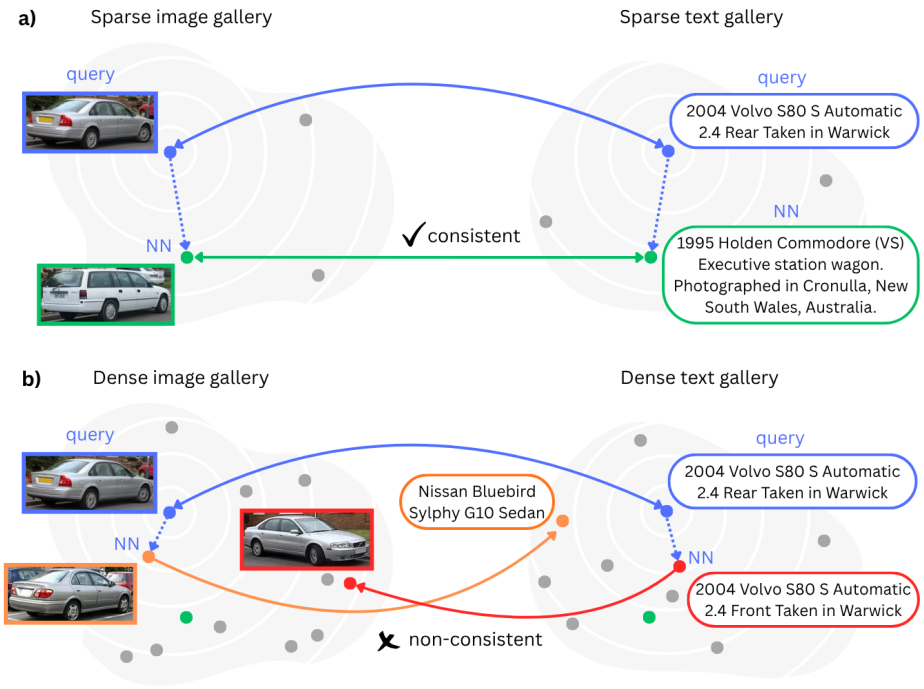


Fig. 1: Illustration of the mutual nearest neighbor metric used by Huh et al. [40] to measure cross-modal alignment. (a) Sparse regime: given a query image and caption (blue), nearest neighbors (NN) are retrieved independently in image and text embedding spaces. Mutual NN alignment measures whether the NNs are consistent across modalities. (b) Dense regime: as dataset size increases, NNs within each modality get better. The vision model retrieves a car in the same pose, and the language model retrieves a caption of the same car model regardless of pose. At scale, improved within-modality organization does not translate into cross-modal agreement.

The Platonic Representation Hypothesis [40] offers a compelling answer: as neural networks grow larger and consume more data, their learned representations will become more and more aligned, no matter which data modality (text, vision, audio, touch, etc.) was used for training. Proponents of language-only learning have interpreted this as validation of their approach: since the choice of modality does not matter as they all lead to the same shared representation, one might as well use language as the most convenient source of data.[1] However, the strength of a hypothesis depends on the strength of its evidence, and the experimental protocol underpinning the claim rests on specific methodological choices that have largely gone unexamined in subsequent work.

In this paper, we take a closer look at the experimental evidence for the hypothesis and find it to be fragile and to depend critically on the evaluation regime. Huh et al. [40] conducted their analysis on small, sparse datasets with one-to-one correspondences between modalities. However, real-world multi-modal

[1] But analogously, the same argument could be made for vision-only learning [41].

data is large, dense, and inherently many-to-many: one image has many valid descriptions, and a single caption can correspond to many plausible images. These differences fundamentally change what it means for two representations to "align".

In a small dataset, weakly related samples may become nearest neighbors simply because no better alternatives exist (Fig. 1a). Here, two models can agree despite organizing their representations differently. As the dataset grows (i.e. the gallery used for retrieving nearest neighbors gets denser), both models find closer neighbors and cross-modal consistency requires more fine-grained structural alignment (Fig. 1b). A vision model may retrieve an image of a car taken from a similar angle as the query, while the language model retrieves a caption describing the same car model as the query but in a different pose. Both are valid, but inconsistent between modalities, producing a mismatch that gets penalized under the mutual nearest-neighbor metric. This illustrates how mutual nearest-neighbor agreement becomes an increasingly strict measure for alignment in many-to-many regimes. Using a mutual kNN metric with $k > 1$ is less strict, but does not fundamentally change the conclusion.

In this paper, we examine how cross-modal alignment changes in evaluation settings with large, dense, and non-bijective datasets, and observe the following:

Alignment degrades with scale: Increasing the gallery from 1024 to millions of samples causes a sharp drop in cross-modal mutual nearest-neighbor agreement.

Coarse agreement persists but fine-grained agreement does not: In controlled settings (e.g., ImageNet), vision and language models reliably retrieve correct-class neighbors but rarely agree on the same instance.

Many-to-many correspondence reduces alignment: Allowing multiple valid correspondences per sample leads to drops in agreement, even when retrieved neighbors are semantically sensible.

Previously reported trends may not hold for newer models: The claim that stronger language models align better with vision seems to weaken for more recent models.

These findings paint a more mixed picture than the small-gallery results from Huh et al. suggest. Models trained on different modalities can learn rich and semantically meaningful structure, yet still organize that structure differently. Low agreement does not imply poor representations, it reflects differences in how information is arranged. Nearly a century ago, von Uexküll [89] argued that every organism inhabits its own perceptual world, or *Umwelt*, shaped by its senses rather than by an observer-independent reality. The same, we believe, might hold for our models: each constructs its own representational structure, determined by its modality and training data, rather than converging toward a shared model of reality. Though it is still early days, we suspect future evidence will favor von Uexküll over Plato.

2 Related Work

One Platonic Ideal vs many *Umwelten*. In his "Theory of Forms", Plato argued that every physical object we perceive is a flawed imitation (a shadow) of some eternal, abstract "ideal" form [71], and only by escaping from the tyranny of our physical senses (leaving the cave of shadows), we can achieve true understanding. But in the 20th century, this argument for a single, unified Platonic Ideal representation has been repeatedly undercut by biologists, psychologists, and philosophers. Biologist von Uexküll argued that every organism inhabits its own perceptual environment, or *Umwelt* [89]: a tick lives in a world of thermal gradients, a bat in a world of echoes. The different Umwelten might have only little overlap with each other[2]. Gibson's ecological psychology [25] pushed this further, proposing that perception is shaped by what an organism can *do* in its environment, not by an observer-independent reality. Philosopher Wittgenstein, thinking about language, arrived at a strikingly similar conclusion. He famously argued: "If a lion could speak, we could not understand him" [94], meaning that the lion's world (goals, instincts, perceived reality) is so utterly different from our own, that even if it spoke English, we would not comprehend the meaning[3]. Building on Wittgenstein, psychologist Rosch developed her Prototype Theory of Categorization [75], arguing forcefully against a single platonic ideal as a representation of object categories, proposing a data-driven clustering-based model instead.

Representational alignment. The question of representational similarity has been studied extensively in the neurosciences [19,33,48]. In machine learning, the parallel question of whether independently trained networks learn similar internal structure has received growing attention. Lenc and Vedaldi [52] investigated the equivalence of representations from different trained models and found that early convolutional layers are more interchangeable than later ones. This task, also referred to as "model stitching", was later revisited by Bansal et al. [6]. Related to this, Li et al. [53] proposed methods to align neurons across independently trained networks. More recently, Dravid et al. [18] introduced "Rosetta Neurons," showing that different vision models share common units corresponding to similar visual concepts across architectures, tasks, and training data.

Furthermore, alignment has been linked with shared model capabilities measured by task performances [5,6,40,45,64]. To directly quantify representational similarities, several metrics have been used to measure correlations between features [36,64]. Kornblith et al. [47] introduced Central Kernel Alignment (CKA) as a robust measure invariant to orthogonal transformations and isotropic scaling. Huh et al. [40] found the CKA metric to reveal only a "very weak trend of alignment between models" and therefore proposed the use of the mutual kNN metric that measures the overlap of two sets of neighborhoods of size k.

[2] For a tour of von Uexküll's ideas, see Koenderink's delightful book [46].

[3] For a great treatment of Wittgenstein's argument in popular culture, see the episode *Darmok* of the American TV series *Star Trek: The Next Generation*.

Multi-modal alignment. Early efforts to connect images and text utilized human annotations [90]. The curation of large-scale paired image-caption datasets, such as MS-COCO [55] and Visual Genome [49], facilitated the systematic study of cross-modal correspondence and models. The CLIP model [73] by Radford et al. formed a turning point by demonstrating that contrastive learning on web-scale image-text pairs could produce shared embedding spaces.

Since, a growing body of work has investigated whether such alignment arises even without explicit joint training. Merullo et al. [62] showed that a simple learned linear transformation could map between frozen vision encoders and LLMs. Moschella et al. [65] use similarities to an anchor set. Maniparambil et al. [59] demonstrated that even unaligned unimodal encoders possess high semantic similarity. Fully unsupervised approaches include blind vision-language matching [77] and unpaired embedding translation via cycle-consistency [42, 98]. Finally, Gupta et al. [31] show that an orthogonal map can map between independently trained multi-modal contrastive models. These results are often seen as evidence for representational convergence. However, they are obtained in restricted settings (e.g. [77] experiments on CIFAR-100 and ImageNet-100) and do not scale to real-world multi-modal data. Our work examines whether alignment survives beyond these constraints, showing that it decreases at scale and reflects coarse categorical agreement rather than shared fine-grained structure.

Limits and measurement of emergent cross-modal structure. Several analyses show that alignment between independently trained unimodal encoders depends strongly on data, architecture, and evaluation protocol. Tjandrasuwita et al. [86] find that alignment varies with modality similarity and the balance of shared versus unique information, while Hadgi et al. [32] report weaker alignment for "pure" 3D encoders without careful subspace selection. Zhu et al. [99] further show that video–text alignment depends on temporal richness and text availability.

Gröger et al. [29] show that global similarity measures such as CKA are sensitive to network scale and can be altered via null calibration, largely removing evidence of global convergence while leaving local neighborhood similarity (e.g., mutual kNN) more stable, though still evaluated under small-scale and bijective regimes. Beyond similarity metrics, Smith et al. [81] and Kumar et al. [50] show that functional agreement and output behavior can persist even when internal representations are misaligned or entangled, suggesting that behavioral compatibility does not imply shared structure. These caveats echo grounding arguments that text-only learning may be insufficient to recover perceptual structure [7, 51], and motivate multimodal foundation models that integrate perception and language at scale [4, 35, 39, 63].

3 Experimental setup

Mutual kNN metric. To measure alignment between representations from different models, we use the mutual k-nearest-neighbor metric (illustrated in Fig. 1), following Huh et al. [40]. Given a shared gallery set of n datapoints (referred to as mini-batch sampled from the data distribution in [40]) encoded

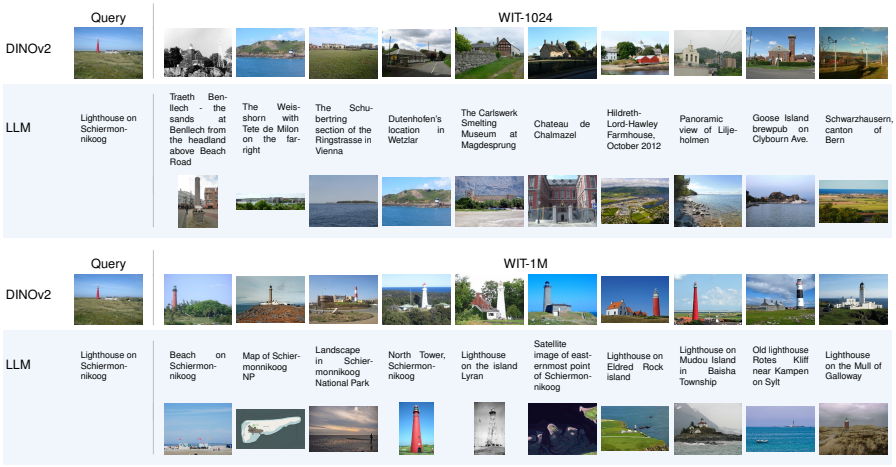


Fig. 2: Nearest-neighbor quality depends on data density. We show 10 within-modality nearest neighbors for image (DINOv2) and text (LLM) embeddings on a sparse WIT-1024 gallery (top) and a denser WIT-1M gallery (bottom). For text queries, retrieved captions and their corresponding reference images are shown. At smaller scale, nearest neighbors are less semantically precise. Nearest-neighbor structure becomes more semantically refined as gallery density increases.

into feature vectors $\mathbf{a_i} \in \mathbb{R}^{d_1}$ and $\mathbf{b_i} \in \mathbb{R}^{d_2}$ by two models with $i \in \{1, \cdots, n\}$, we first L2-normalize each representation. We then retrieve the k nearest neighbors of every query point independently for each model (e.g. image and text query for vision and language encoders):

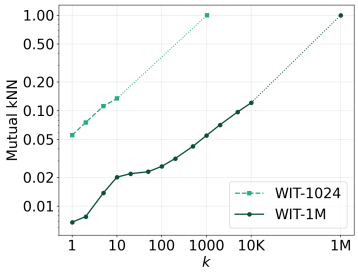
$$\mathcal{N}_k^{\mathbf{a}}(i) = \text{argtopk}_{j \neq i}\, \mathbf{a}_i^\top \mathbf{a}_j, \qquad \mathcal{N}_k^{\mathbf{b}}(i) = \text{argtopk}_{j \neq i}\, \mathbf{b}_i^\top \mathbf{b}_j.$$

The per-sample score is the number of overlapping samples normalized by k:

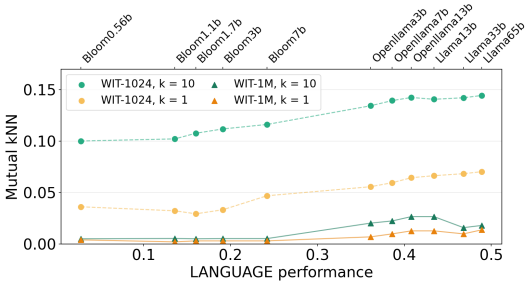
$$s_i = \frac{|\mathcal{N}_k^{\mathbf{a}}(i) \cap \mathcal{N}_k^{\mathbf{b}}(i)|}{k},$$

and the overall mutual-kNN score is the mean over all samples. A score of 1 means that every point's k nearest neighbors are identical in both spaces, and a score of 0 means that the k nearest neighbors do not overlap. In the sparse gallery in Fig. 1, the query (blue) retrieves the same neighbor in both image and text spaces, giving a mutual kNN score of 1 for k=1. A score of $\frac{k}{n}$ suggests chance-level expected overlap for independent random retrieval, which decreases for growing n. Note that throughout we report raw mutual kNN (as in [40]).

Implementation details. For most experiments, we use DINOv2-base [69] as the vision encoder. We refer to this model as DINOv2 in the following. Our primary language model is OpenLlama3b [24,87] (abbreviated as OpenLlama). Additional models are considered in the supplementary material (Sec. A.2). For each image and text sample, we extract the representations from all layers of their



(a) Effect of neighborhood sizes k in mutual kNN (both axes log-scaled). Trivially, mutual kNN converges to 1.0 as k approaches the full gallery size. [40] utilized mutual kNN for $k = 10$.



(b) Alignment between DINOv2 and different LLMs, measured on WIT-1024 and WIT-1M. As observed in [40], alignment (mutual kNN) increases with language performance (measured as $1 - \text{bitsperbyte}$ from [40]) on the 1024-sample set, but this trend breaks with larger gallery size.

Fig. 3: Mutual kNN text-image feature alignment when scaling from WIT-1024 to WIT-1M. (a) shows the dependence on neighborhood size k, while (b) examines alignment for different LLMs. *The observation from [40], that more capable language models align better with vision largely vanishes at WIT-1M scale.*

respective encoders and follow the experimental protocol from [40]. Details about additional models used in Sec. 4 are provided in Sec. E.3 in the supplementary material. We use Faiss [17] for nearest neighbor computation at scale. Specifically, we use their exact nearest neighbor implementation with `IndexFlatL2` which is equivalent to using cosine similarity on normalized vectors.

4 How much do representations align?

In this section, we take a close look at the experimental evidence underpinning the Platonic Representation Hypothesis [40]. The experiments in [40] rest on two foundations that warrant scrutiny: the use of mutual kNN alignment on a small evaluation set of only 1024 samples from the Wikipedia Image-Text (WIT) dataset [82] (WIT-1024), and the use of data with bijective (one-to-one) image-text correspondences. Typically, these choices are not acknowledged when the hypothesis is cited [9, 13, 57, 60, 76]. The claim is usually invoked in its broad, appealing form rather than in the narrow terms under which experimental support was provided.

Here, we analyze how alignment behaves for a finer-grained metric (k=1 instead of k=10), and a denser gallery (million(s of) instead of 1024 samples). We then decompose what mutual kNN alignment actually measures in a controlled setup on ImageNet. This reveals that models individually retrieve correct-class neighbors but rarely agree on which one, suggesting that information is organized differently in each unimodal model. We then turn to the bijective assumption, and examine what happens when it is relaxed. Finally, we perform a trend check to ask whether the predictions from [40] have held up as models have improved.

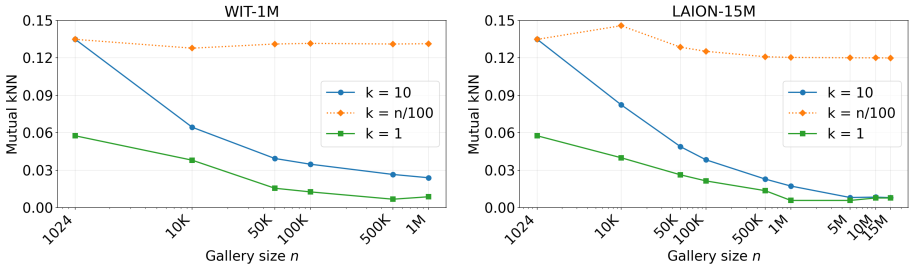


Fig. 4: Scaling the gallery size to 1M (WIT) and 15M (LAION) shows a large drop in mutual kNN alignment for k=1 and k=10 for DINOv2 and OpenLlama features.

Sensitivity to k in mutual kNN. Huh et al. [40] reported mutual kNN alignment for k=10. We additionally evaluate at k=1, which requires the two representation spaces to agree on the single nearest neighbor. As shown in Fig. 3a, the metric trivially converges to 1 as k approaches the full gallery size n, since both neighbor sets then contain all samples. Even moderate values of k can inflate scores by capturing broadly similar rather than precisely matching neighbors. In our analyses at larger gallery scales, we perform deduplication to prevent near-duplicate samples from trivially inflating neighborhood overlap (see Sec. E.1 in the supplementary material for details).

4.1 Alignment across dataset scales

Nearest neighbors in sparse gallery. We now turn to the data, and ask whether the 1024-sample gallery used in [40] is too sparse to capture more than coarse structural agreement. As shown in Tab. 1, the mean cosine similarity between queries and nearest neighbors in terms of both image (DINOv2) and text features (OpenLlama) is significantly lower for the WIT-1024 gallery compared to WIT-1M (e.g. 0.799 compared to 0.906 for DINOv2 at k=1). Note that in both cases, we use WIT-1024 as the query set.

We visualize nearest neighbors for k=10 for image and text features on WIT-1024 and WIT-1M in Fig. 2. At low density, semantically unrelated samples may end up as nearest neighbors as there is nothing closer available, meaning that measured mutual kNN agreement mainly can reflect the shared lack of alternatives. To get more meaningful insights, we scale the density of the retrieval gallery in the following section.

Table 1: Nearest-neighbor quality across gallery sizes. As the gallery grows, nearest neighbors get closer to the query set in both DINOv2 and OpenLlama embedding spaces, facilitating the more fine-grained analysis of cross-modal alignment.

Gallery	Type	k=1	k=10
WIT-1024	DINOv2	0.799	0.717
WIT-1024	OpenLlama	0.502	0.400
WIT-1M	DINOv2	0.906	0.888
WIT-1M	OpenLlama	0.757	0.701

Densification by scaling the gallery size. Having established that the WIT-1024 gallery captures mainly coarse structure, we densify the gallery and test whether alignment persists. We evaluate on up to 1M and 15M gallery samples

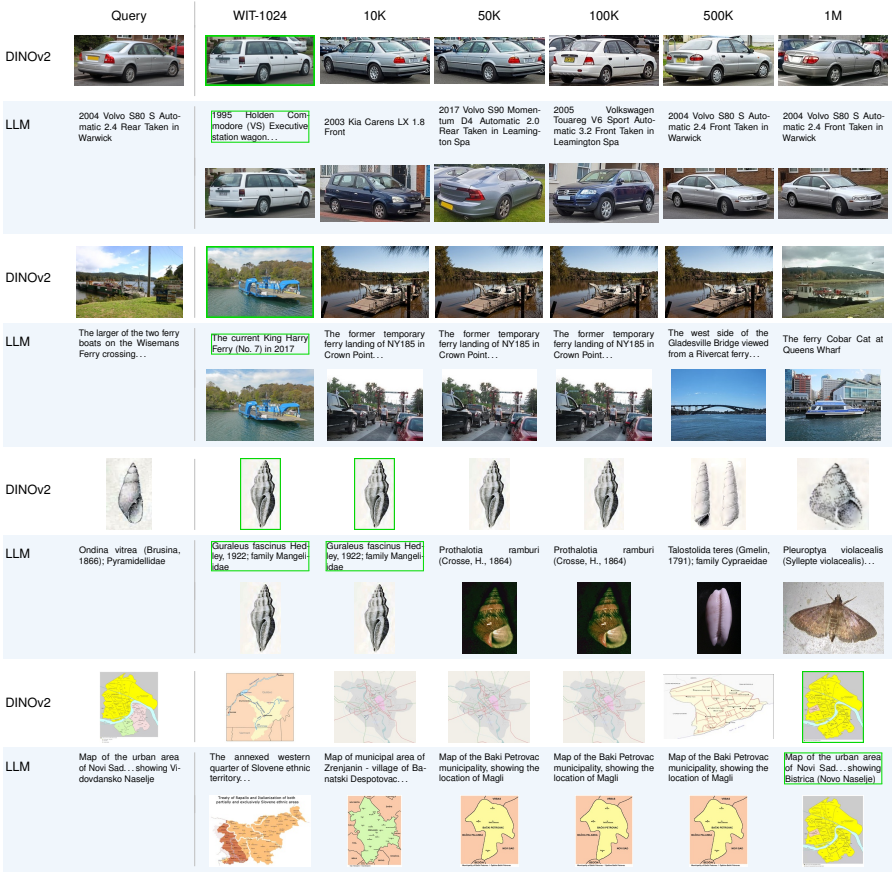


Fig. 5: Nearest-neighbor (k=1) examples with DINOv2 and OpenLlama across gallery scales on WIT-1M. Captions are shown with corresponding images. Mutual kNN matches across modalities are framed green. *While the bottom example shows a match at 1M scale, at larger scales each model finds closer but different matches (top three).*

from the English-text WIT [82] and LAION400M [78] respectively. The best layer pair was determined on the 1024-sample subset of WIT, following [40]. As shown in Fig. 4, alignment scores decrease as gallery size grows for fixed k and query set (WIT-1024). The mutual kNN alignment scores drop from 0.135 and 0.058 on the 1024-sample gallery to 0.008 and 0.001 on LAION-15M for $k = 10$ and $k = 1$ respectively. This confirms that the agreement observed at small scale declines with the transition to finer-grained evaluation at large scale. Nearest neighbors become closer and more semantically similar to the query, placing greater demand on the two representation spaces to agree on subtle distinctions.

Interestingly, alignment at k=$n/100$ remains relatively stable across scales, suggesting that models share some degree of coarse structural agreement. We hypothesize that this amounts to precisely the kind of broad categorical corre-

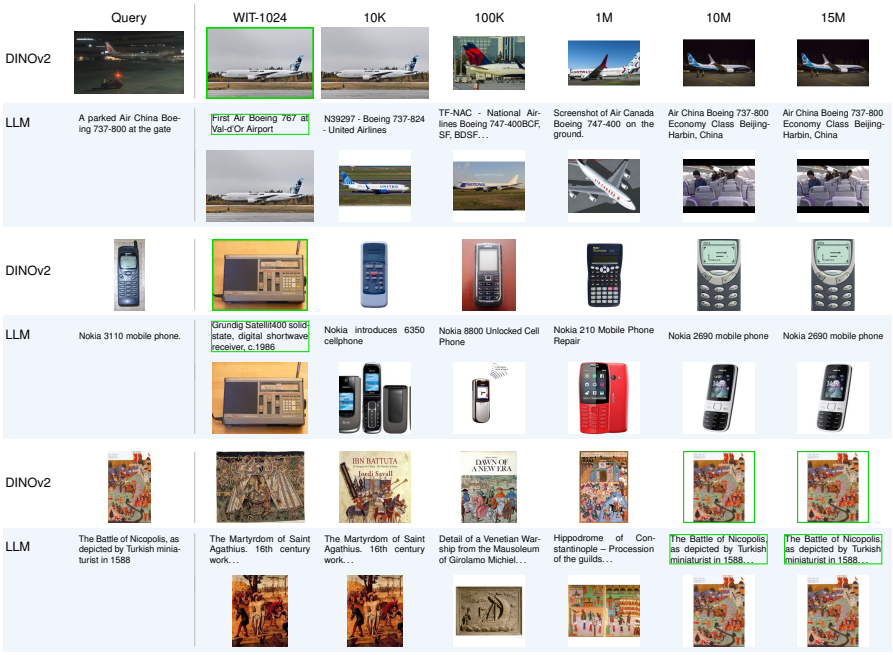


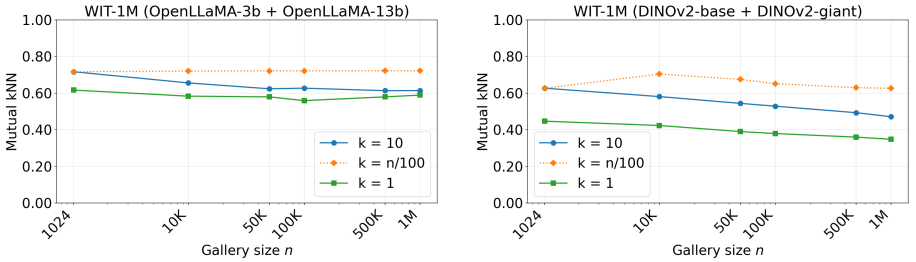
Fig. 6: Nearest-neighbor (k=1) examples with DINOv2 and OpenLlama across gallery scales on LAION-15M. *As the gallery densifies, each model finds closer but different matches (top two). The match at 15M (bottom right) is a near-duplicate that survived our deduplication pipeline.*

spondence one would expect from models trained on overlapping internet data, and lacks signal about whether representations are organized in the same way.

We also analyze how mutual kNN alignment for various LLMs and DINOv2 behaves at the WIT-1M scale. Reproducing the setting of [40], Fig. 3b shows a clear trend on WIT-1024: stronger language models exhibit higher alignment with visual features. This is a central finding of [40] and one of their most compelling pieces of evidence. However, when we scale the gallery to 1M samples, this trend largely vanishes. The gap between LLMs narrows considerably, and the relationship between model capability and alignment weakens. This suggests that the observation in [40] may be a result of the sparse evaluation setting.

The nearest-neighbor examples in Figs. 5 and 6 further illustrate this effect. Matches that appear semantically meaningful at small gallery sizes often break down as more candidates are introduced. In vision space, we find better neighbors that deviate from the best text neighbors at larger scale. There are only very few matches found at 1M and 15M data scale, most of which are near-duplicates that our deduplication pipeline did not catch (e.g. a crop shifted by a few pixels). We show additional visualizations in Figs. 21 to 24 in the supplementary material.

We additionally test whether the alignment drop with increasing gallery size is merely an artifact of the mutual kNN metric being harder at scale. Specifically, we



(a) OpenLlama-3b and OpenLlama-13b **(b)** DINOv2-base and DINOv2-giant

Fig. 7: Unimodal mutual kNN alignment as a function of gallery size on WIT-1M. *In contrast to cross-modal alignment (Fig. 4), unimodal alignment remains significantly more stable across scales.*

measure within-modality alignment for two pairs of models: two language models of different scale (OpenLlama-3b and OpenLlama-13b), and, separately, two vision models (DINOv2-base and DINOv2-giant). If mutual kNN alignment collapses for dense galleries regardless of the models being compared, the cross-modal drop observed would be uninformative. If within-modality alignment remains stable, the cross-modal drop is meaningful.

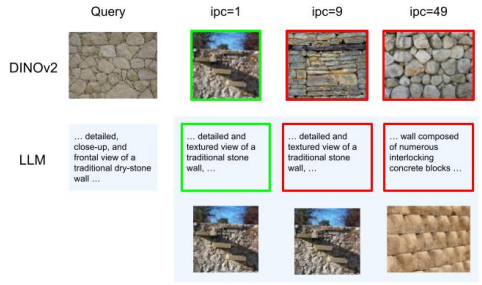
As shown in Fig. 7, unimodal alignment remains much more stable across gallery sizes than the cross-modal alignment reported in the main paper. For the OpenLlama pair, mutual kNN at k=1 stays between $[0.59, 0.62]$, and for the DINOv2 pair between $[0.35, 0.45]$, across all gallery scales. This confirms that mutual kNN does not inherently collapse at scale.

Observation 1: Mutual kNN alignment scores decrease for denser galleries for fixed small k, suggesting that mutual kNN is sensitive to gallery sparsity. Furthermore, the reported trend that stronger language models align better with vision weakens substantially at WIT-1M scale, with all models scoring near zero.

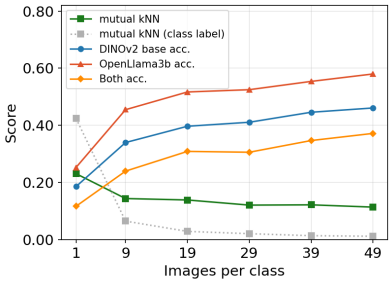
4.2 What is captured by cross-modal mutual kNN alignment?

Low mutual kNN alignment at fixed small k could mean two things: the models individually retrieve poor neighbors, or they each retrieve good neighbors but different ones. The stability at $k = \frac{n}{100}$ hints at the models agreeing at a coarse level but diverging on fine-grained structure. To test this directly, we use the ImageNet [15] validation set, where class labels let us evaluate each model's retrieval independently.

We decompose each query into: (i) whether each model individually retrieves a correct-class neighbor, (ii) whether both do, and (iii) whether they agree on the exact same gallery item (mutual kNN with k=1). Our query set consists of one image per class (1000 images), and we vary the number of images per class (ipc) in the gallery from 1 to 49. We use detailed image captions (981 words on average)



(a) The query image (left) is matched with galleries of increasing density. As the gallery becomes more dense, DINOv2 and OpenLlama retrieve from the same class, but different instances, illustrating how within-class structure is organized differently across modalities.



(b) Per-modality retrieval accuracy and cross-modal mutual kNN alignment (k=1) as images / captions per class in gallery increase. Modalities individually improve with gallery density, but alignment does not.

Fig. 8: Decomposing cross-modal alignment on ImageNet val. (a) shows a qualitative retrieval example where both models find plausible neighbors but disagree on the specific instance. (b) quantifies this: individual class-level retrieval accuracy improves with gallery density, yet strict alignment remains flat, illustrating that **models organize within-class structure differently**.

generated by gemini-3-flash-preview [70, 85], making this a favorable setting for alignment (details are provided in Sec. E.2 in the supplementary material).

Fig. 8b reveals that as the gallery densifies, in line with Cover and Hart [12], both models individually improve at retrieving correct-class neighbors. This indicates some degree of shared coarse structure. At larger scale, both models retrieve reasonable neighbors but different ones. We see similar trends when looking at coarser evaluation with k=10 (see Sec. B.3 in the supplementary material). At 49 images per class in the gallery, DINOv2 succeeds 46.1% of the time and OpenLlama 58.0%. Yet strict alignment on the exact same gallery item remains flat around 11%, even with detailed captions. For reference, alignment with class-name-only captions drops from 0.42 to near zero as ipc increases. The models are individually capable but organize within-class structure differently (Fig. 8a). At ipc=1, strict alignment (23.1%) actually exceeds the rate at which both models retrieve a correct-class neighbor (11.7%), meaning the models often agree on semantically plausible but technically incorrect neighbors (Fig. 9).

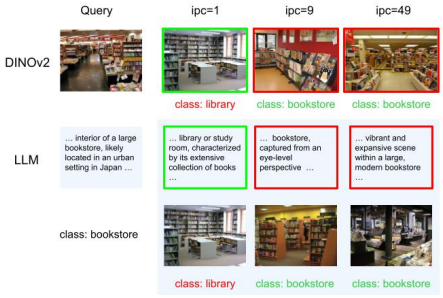


Fig. 9: Shared mistake at ipc=1. The query image (bookstore) is matched by both DINOv2 and OpenLlama to a library image. The models agree, but on the wrong answer.

This reveals what mutual kNN actually captures. It does not measure unimodal representation quality, but agreement on fine-grained structure. Our experiments provide direct evidence that low cross-modal alignment in terms of mutual kNN is not due to poor representations but rather due to fun-

damentally different representational organization within modalities. Both models learn structured, high-quality representations, they simply do not structure them the same way.

Observation 2: As the data gets denser, both models retrieve correct-class neighbors at increasing rates, yet strict cross-modal mutual kNN alignment is flat at 11%. This is not a failure of unimodal representation quality, but of unaligned representational organization across modalities.

4.3 What happens when the data is not bijective?

In practice, the relationship between modalities, such as image and text, is inherently many-to-many: a single image can be described by countless text descriptions, and a single text caption can correspond to a large set of visually distinct images. More fundamentally, modalities often differ in information content. Specifically, images encode spatial, textural, and perceptual structure that text captures only to a limited extent. On the other hand, text encodes abstraction, negation, and compositional semantics that images do not. One could, in principle, bridge this gap trivially. For instance, one could encode pixel values as text or render captions as images and establish a bijection between those. Those preserve the information, but the inductive structure (the modality-specific properties that make each modality useful) of each modality is lost.

To test what happens when bijectivity is relaxed, we use the CycleReward dataset [3] which pairs each real sample with multiple synthetic candidates. The I2T subset contains 11 generated captions per real image, and T2I consists of 12 synthetic images for each text prompt. This directly breaks the bijection, i.e. one-to-one matching, that our earlier analysis assumes.

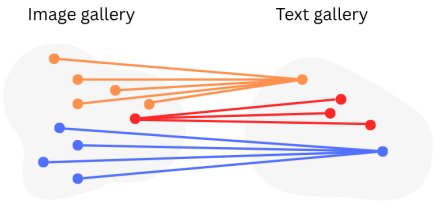


Fig. 10: Illustration of non-bijective (many-to-many) correspondence between image and captions. The nearest neighbor of a text caption for one image (blue) is a caption for a different image (red). However, the nearest image neighbor for a given image may be another image with the same caption.

We evaluate mutual kNN by densifying one modality at a time: for T2I we keep the text fixed and increase the number of generated images per prompt, and for I2T we keep the image fixed and add generated captions. For illustration, let us consider the T2I experiments where the closest neighbor in the densified modality is more likely to be a similar image that is associated with the same caption, while in the sparse text space the NN will be a caption for a different image. This creates a scenario where mutual kNN fails (see Fig. 10).

We adapt the mutual kNN metric where a match is counted when the retrieved item corresponds to the same source sample, even if it is not the exact same caption or image. In Fig. 11, we see that the mutual kNN scores decrease as

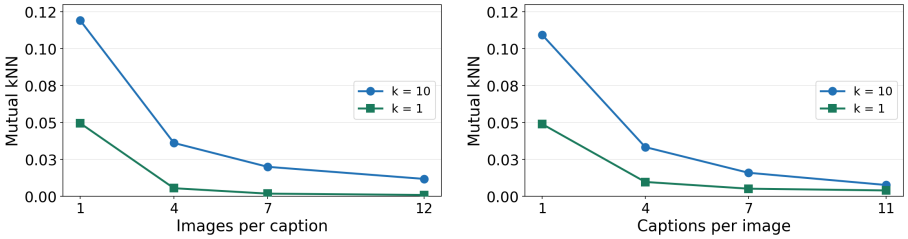


Fig. 11: Effect of relaxing the bijective assumption on text-image alignment, using the CycleReward dataset [3]. We densify one modality by adding more images per caption (left) or more captions per image (right) while keeping the other fixed. Mutual kNN alignment decreases consistently for both k=1 and k=10.

the one-to-one assumption is relaxed. Whether this is due to reduced alignment or just a limitation of the metric is an open question. Regardless, the original evidence for convergence depends on an assumption that real-world multi-modal data rarely satisfies.

Observation 3: A single image can be described in countless ways, and a single caption can match many visually distinct images. When we progressively relax the one-to-one assumption, mutual kNN alignment drops consistently. The mutual kNN metric cannot distinguish between genuine misalignment and many-to-many correspondence.

4.4 Trend check: Are the predictions from [40] holding up so far?

Huh et al. [40] predict that as LLMs become stronger, their representations align more with vision representations. This claim is evaluated using three proxies for language performance: HellaSwag [97], GSM8K [11], and (1 − bitsperbyte).

In this section, we revisit this trend analysis with an extended set of models and benchmarks. We evaluate 55 LLMs, spanning from BLOOMZ [66] to recently released models. In addition to the evaluations in [40], we use the ARC Challenge [10], MMLU [34], and LogiQA2 [58] benchmarks. These probe arithmetic reasoning, general knowledge, and logical reasoning. We present results for models that surpass Llama-3-70B [27] (the strongest model in [40]) on at least one benchmark, testing whether the alignment-performance trend continues. The full set of 55 models and results on the largely saturated benchmarks are included in Sec. D in the supplementary material.

In Fig. 12, we observe that recent models do not continue on the scaling lines extrapolated from the model set from [40]. Instead, their points do not follow the predicted trend and hint at saturation with respect to DINOv2 features. Furthermore, the R^2 averaged across all regression lines ranges from -8.6 to -3.7, confirming that the extrapolated trend is not continued by recent models.

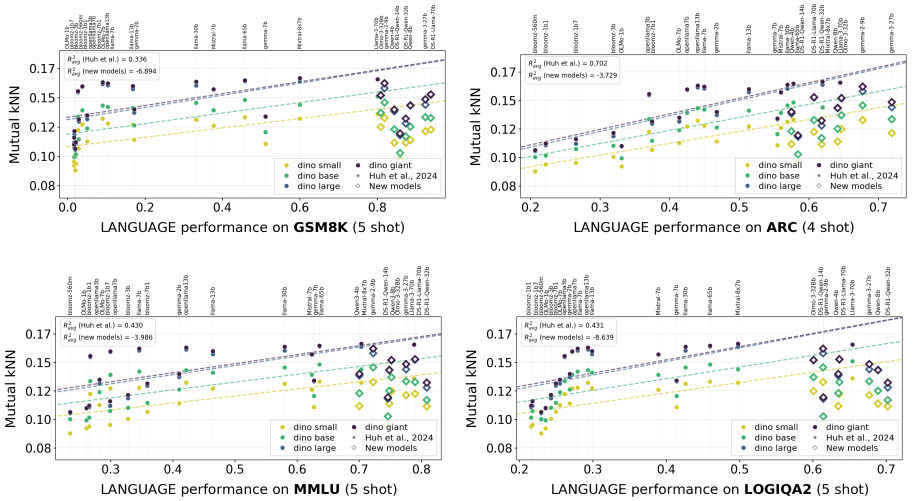


Fig. 12: Testing whether the alignment-LLM performance trend from [40], tested on WIT-1024, holds for recent LLMs on ARC, GSM8K, MMLU, and LogiQA2. Dashed lines show the trend for models from [40] (circles). Recent LLMs (diamonds) do not follow the trend: stronger language models do not seem to be more aligned with DINOv2.

5 Discussion and future work

The Platonic Representation Hypothesis suggests that models trained on different modalities converge toward a shared representation of reality as they scale. Our results indicate a more conditional interpretation of this claim. Mutual kNN agreement is highly sensitive to the evaluation regime: it drops sharply when moving from small galleries to million-scale datasets and degrades further under many-to-many cross-modal correspondences. Moreover, the previously reported trend that stronger language models yield higher alignment does not consistently hold for recent models.

These findings do not rule out shared structure across modalities. Overall, our results indicate that small-scale mutual nearest-neighbor evaluations may overstate the degree of convergence by relying on restrictive gallery sizes and one-to-one pairing. Low mutual kNN agreement should not be conflated with weak representations. As our analysis on ImageNet shows, it may instead reflect differences in how fine-grained structure is organized. Rather than converging on a single Platonic representation of reality, modalities appear to inhabit in their own *Umwelten*, a distinct but coherent representational "cave" where alignment between them is local and partial.

Future work: in search of bijection. Prior work on the Platonic Representation Hypothesis [40] evaluates alignment under a one-to-one correspondence assumption between modalities. We have shown that mutual kNN agreement is not reliable once this assumption is relaxed. Likewise, interpretations of the Platonic Representation Hypothesis as evidence that "language is all you need" [92]

often rely on evaluation regimes that effectively assume near-bijective structure between images and text.

However, real-world image–text data is fundamentally many-to-many, and the extent to which any approximate bijection exists at the level of representations remains unclear. A key direction for future work is to directly test this assumption, for example by studying whether language can serve as a lossless bottleneck for image reconstruction (i.e. an image-text-image autoencoder). If, as we suspect, this proves illusive for realistic settings (e.g. text bottlenecks of under a thousand words), it would be very interesting to identify and model the part of the joint text-image space forming a bijection (the intersection of the Venn diagram), and disentangle it from the parts that do not.

Acknowledgements: This work was in part supported by the BMFTR (FKZ: 16IS24060), the DFG (SFB 1233, project number: 276693517), NSF IIS-2403305, and ONR MURI. This research utilized compute resources at the Tübingen Machine Learning Cloud. The authors thank all Efros group members for valuable discussions that shaped this work, and particularly Tyler Bonnen and Amil Dravid for proofreading the draft. Lastly, we thank Phillip Isola for feedback and for sparking this conversation by inviting us out of the cave, which ended up motivating us to go back in to examine the shadows in a new light.

References

1. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 (2022) 1
2. Akyürek, E., Damani, M., Zweiger, A., Qiu, L., Guo, H., Pari, J., Kim, Y., Andreas, J.: The surprising effectiveness of test-time training for few-shot learning. arXiv preprint arXiv:2411.07279 (2024) 1
3. Bahng, H., Chan, C., Durand, F., Isola, P.: Cycle consistency as reward: Learning image-text alignment without human preferences. arXiv preprint arXiv:2506.02095 (2025) 13, 14, 25
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 5
5. Balestriero, R., et al.: A spline theory of deep learning. In: ICML (2018) 4
6. Bansal, Y., Nakkiran, P., Barak, B.: Revisiting model stitching to compare neural representations. In: NeurIPS (2021) 4
7. Bender, E.M., Koller, A.: Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020) 5
8. Browning, J., LeCun, Y.: Ai and the limits of language. Noema Magazine (2022) 1
9. Chai, W., Song, E., Du, Y., Meng, C., Madhavan, V., Bar-Tal, O., Hwang, J.N., Xie, S., Manning, C.D.: Auroracap: Efficient, performant video detailed captioning and a new benchmark. In: ICLR (2025) 7
10. Chollet, F.: On the measure of intelligence. arXiv preprint arXiv:1911.01547 (2019) 1, 14, 27

11. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021) 14, 27
12. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory (1967) 12
13. Dar, G.: mini-vec2vec: Scaling universal geometry alignment with linear transformations. arXiv preprint arXiv:2510.02348 (2025) 7
14. DeepSeek-AI, Guo, D., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 34
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 11, 33
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 34
17. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024) 7
18. Dravid, A., Gandelsman, Y., Efros, A.A., Shocher, A.: Rosetta neurons: Mining the common units in a model zoo. In: ICCV (2023) 4
19. Edelman, S.: Representation is representation of similarities. Behavioral and brain sciences (1998) 4
20. Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., Zou, A.: The language model evaluation harness. Zenodo (07 2024). https://doi.org/10.5281/zenodo.12608602, https://zenodo.org/records/12608602 27
21. Gemma Team, G.D.: Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118 (2024) 34
22. Gemma Team, G.D.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024) 34
23. Gemma Team, G.D.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025) 34
24. Geng, X., Liu, H.: Openllama: An open reproduction of llama (2023), https://github.com/openlm-research/open_llama 6, 26, 34
25. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979) 4
26. Gokaslan, A., Cohen, V.: Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus (2019) 27
27. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 14, 34
28. Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A.H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K.R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M.E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N.A., Hajishirzi, H.: Olmo: Accelerating the science of language models. In: ACL (2024) 34

29. Gröger, F., Wen, S., Brbić, M.: Revisiting the platonic representation hypothesis: An aristotelian view. arXiv preprint arXiv:2602.14486 (2026) 5
30. Gu, S., Clark, C., Kembhavi, A.: I can't believe there's no images! learning visual tasks using only language supervision. In: ICCV (2023) 1
31. Gupta, S., Kansal, S., Jegelka, S., Isola, P., Garg, V.: Canonicalizing multimodal contrastive representation learning. In: ICLR (2026) 5
32. Hadgi, S., Moschella, L., Santilli, A., Gomez, D., Huang, Q., Rodolà, E., Melzi, S., Ovsjanikov, M.: Escaping plato's cave: Towards the alignment of 3d and text latent spaces. In: CVPR (2025) 5
33. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science (2001) 4
34. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: ICLR (2021) 14, 27
35. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025) 5
36. Hotelling, H.: Relations between two sets of variates. In: Breakthroughs in statistics: methodology and distribution (1992) 4
37. Hu, X., Storks, S., Lewis, R.L., Chai, J.: In-context analogical reasoning with pre-trained language models. In: ACL (2023) 1
38. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N.A., Luo, J.: Promptcap: Prompt-guided image captioning for vqa with gpt-3. In: ICCV (2023) 1
39. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., Wei, F.: Language is not all you need: Aligning perception with language models. In: NeurIPS (2023) 5
40. Huh, M., Cheung, B., Wang, T., Isola, P.: The platonic representation hypothesis. In: ICML (2024) 1, 2, 4, 5, 6, 7, 8, 9, 10, 14, 15, 27, 28, 29, 30, 31, 34
41. Isola, P.: Personal communication (2025) 2
42. Jha, R., Zhang, C., Shmatikov, V., Morris, J.X.: Harnessing the universal geometry of embeddings. In: NeurIPS (2025) 5
43. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., El Sayed, W.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023) 34
44. 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) 34
45. Jiang, J., Zhou, J., Zhu, Z.: Tracing representation progression: Analyzing and enhancing layer-wise similarity. arXiv preprint arXiv:2406.14479 (2024) 4
46. Koenderink, J.J.: Sentience. De Clootcrans Press, Trajectum, Netherlands (2019) 4
47. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: ICML (2019) 4
48. Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience (2008) 4
49. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017) 5

50. Kumar, A., Clune, J., Lehman, J., Stanley, K.O.: Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. arXiv preprint arXiv:2505.11581 (2025) 5
51. LeCun, Y., et al.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Openreview (2022) 5
52. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. In: CVPR (2015) 4
53. Li, Y., Yosinski, J., Clune, J., Lipson, H., Hopcroft, J.: Convergent learning: Do different neural networks learn the same representations? In: ICLR (2016) 4
54. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. In: ICRA (2023) 1
55. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 5
56. Liu, A.H., Subramanian, S., Jouault, V., Sadé, A., et al.: Ministral 3. arXiv preprint arXiv:2601.08584 (2026) 34
57. Liu, D., Zhao, S., Zhuo, L., Lin, W., Xin, Y., Li, X., Qin, Q., Qiao, Y., Li, H., Gao, P.: Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657 (2025) 7
58. Liu, H., Liu, J., Cui, L., Teng, Z., Duan, N., Zhou, M., Zhang, Y.: Logiqa2.0: The logicqa dataset for logical reasoning. IEEE Transactions on Audio, Speech, and Language Processing (2023) 14, 27
59. Maniparambil, M., Akshulakov, R., Djilali, Y.A.D., Narayan, S., Seddik, M.E.A., Mangalam, K., O'Connor, N.E.: Do vision and language encoders represent the world similarly? In: CVPR (2024) 5
60. Marcos-Manchón, P., Fuentemilla, L.: Shared representations in brains and models reveal a two-route cortical organization during scene perception. arXiv preprint arXiv:2507.13941 (2026) 7
61. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843 (2016) 27
62. Merullo, J., Castricato, L., Eickhoff, C., Pavlick, E.: Linearly mapping from image to text space. In: ICLR (2023) 5
63. Meta AI: The llama 4 herd: The beginning of a new era of natively multimodal ai innovation (2025), https://ai.meta.com/blog/llama-4-multimodal-intelligence/ 5
64. Morcos, A.S., Raghu, M., Bengio, S.: Insights on representational similarity in neural networks with canonical correlation. In: NeurIPS (2018) 4
65. Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., Rodolà, E.: Relative representations enable zero-shot latent space communication. In: ICLR (2023) 5
66. Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., Bari, M.S., Shen, S., Yong, Z.X., Schoelkopf, H., Tang, X., Radev, D., Aji, A.F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., Raffel, C.: Crosslingual generalization through multitask finetuning. In: ACL (2023) 14, 34
67. OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Ettinger, A., Guerquin, M., Heineman, D., Ivison, H., Koh, P.W., Liu, J., Malik, S., Merrill, W., Miranda, L.J.V., Morrison, J., Murray, T., Nam, C., Poznanski, J., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer,

L., Farhadi, A., Smith, N.A., Hajishirzi, H.: 2 olmo 2 furious. arXiv preprint arXiv:2501.00656 (2025) 34
68. OpenAI: Introducing gpt-oss (2025), https://openai.com/index/introducing-gpt-oss/ 34
69. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. TMLR (2024) 6, 26, 34
70. Pichai, S., Hassabis, D., Kavukcuoglu, K.: A new era of intelligence with Gemini 3. Google Blog (The Keyword) (Nov 2025), https://blog.google/products-and-platforms/products/gemini/gemini-3/, accessed: 2026-01-01 12, 33
71. Plato: Republic (c 375 BC) 4
72. Qwen Team, A.C.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 34
73. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 5
74. Research, I.: Granite 3.3 8b base (2025), https://huggingface.co/ibm-granite/granite-3.3-8b-base 34
75. Rosch, E.: Principles of categorization. In: Rosch, E., Lloyd, B.B. (eds.) Cognition and Categorization, pp. 27–48. Lawrence Elbaum Associates (1978) 4
76. Ruan, J., Abudula, A., Liu, X., Li, B., Li, Y., Wang, C., Fan, Y., Ge, Y., Xiao, T., Zhu, J.: Ndp: Next distribution prediction as a more broad target. arXiv preprint arXiv:2408.17377 (2024) 7
77. Schnaus, D., Araslanov, N., Cremers, D.: It's a (blind) match! towards vision-language correspondence without parallel data. In: CVPR (2025) 5
78. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021) 9, 29, 33
79. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Public multimodal dataset (PMD), https://huggingface.co/datasets/facebook/pmd 32
80. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: CVPR (2022) 32
81. Smith, D., Mannering, H., Marcu, A.: Functional alignment can mislead: Examining model stitching. In: ICML (2025) 5
82. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: ACM SIGIR conference on research and development in information retrieval (2021) 7, 9, 25, 29, 32
83. Sutskever, I.: "the mastermind behind gpt-4 and the future of ai" — eye on a.i. (podcast, season 2 episode 118). https://podcasts.apple.com/us/podcast/ilya-sutskever-the-mastermind-behind-gpt-4-and/id1438378439?i=1000604382855 (mar 2023), accessed: 2026-02-28 1
84. Team, F.L.: The falcon 3 family of open models (2024), https://huggingface.co/blog/falcon3 34
85. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 12, 33

86. Tjandrasuwita, M., Ekbote, C., Ziyin, L., Liang, P.P.: Understanding the emergence of multimodal representation alignment. In: ICML (2025) 5
87. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 6, 34
88. 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) 34
89. Uexküll, J.B., Kriszat, G.: Streifzuge durch die Umwelten von Tieren und Menschen Ein Bilderbuch unsichtbarer Welten. Springer (1934) 3, 4
90. Von Ahn, L., Ginosar, S., Kedia, M., Blum, M.: Improving image search with phetch. In: ICASSP (2007) 5
91. Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., Goodman, N.D.: Hypothesis search: Inductive reasoning with language models. arXiv preprint arXiv:2309.05660 (2023) 1
92. Wang, S.L., Isola, P., Cheung, B.: Words that make language models perceive. arXiv preprint arXiv:2510.02425 (2025) 15
93. Wightman, R.: Pytorch image models (2019), https://github.com/huggingface/pytorch-image-models 34
94. Wittgenstein, L.: Philosophical Investigations. Wiley-Blackwell (1953) 4
95. Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Wang, G., Li, H., Zhu, J., Chen, J., et al.: Yi: Open foundation models by 01.ai. arXiv preprint arXiv:2403.04652 (2024) 34
96. Zauner, C.: Implementation and benchmarking of perceptual image hash functions (2010) 29
97. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: Hellaswag: Can a machine really finish your sentence? In: ACL (2019) 14, 27
98. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017) 5
99. Zhu, T., Han, T., Guibas, L., Pătrăucean, V., Ovsjanikov, M.: Dynamic reflections: Probing video representations with text alignment. In: ICLR (2026) 5

Supplementary Material
Back into Plato's Cave: Examining Cross-modal Representational Convergence at Scale

A Is the drop in mutual *k*NN alignment at scale caused by the caption quality, or by specific model choices?

In this section, we provide additional experiments to verify that the main findings in the paper are not due to a confounding variable.

A.1 WIT-1M-recap: Is the alignment drop caused by poor captions?

One might hypothesize that the alignment drop at scale is driven by low quality of the WIT captions rather than by a fundamental cross-modal difference. To test this, we recaption WIT-1M using gemini-3-flash-preview as described in Sec. E.2. The resulting WIT-1M-recap dataset contains visually detailed descriptions of around 500 words per image. As shown in Fig. 13, mutual kNN alignment still drops with gallery size. More detailed captions give overall higher mutual kNN scores, but do not prevent the decline in scores. **This suggests that caption quality is not the primary driver of the mutual kNN alignment drop**.

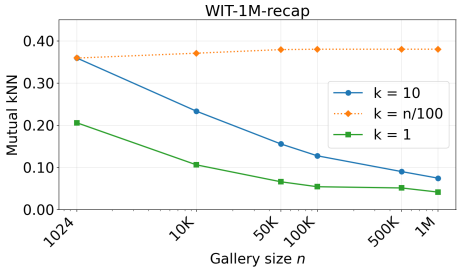
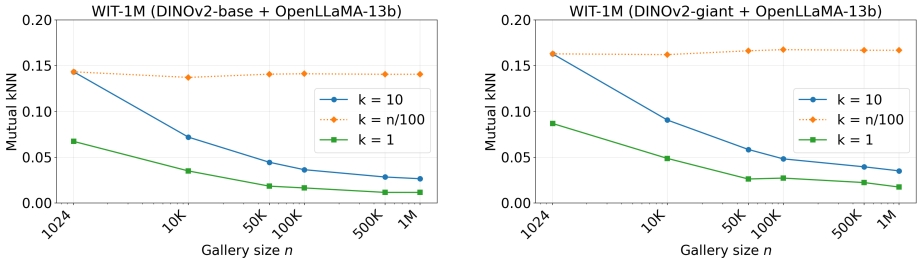


Fig. 13: Cross-modal mutual kNN alignment on images recaptioned using gemini-3-flash-preview (WIT-1M-recap) as the gallery grows to 1M samples. *Detailed captions result in overall higher mutual kNN scores, but do not prevent the drop in scores.*

A.2 Mutual cross-modal *k*NN alignment drops at scale across model pairs

The cross-modal alignment drop reported in Fig. 4 in the paper uses DINOv2-base and OpenLlama-3b. Here, we examine whether similar patterns hold for stronger models. In Fig. 14, we repeat the scaling experiment for two additional model pairs: DINOv2-base with OpenLlama-13b, and DINOv2-giant with OpenLlama-13b.



(a) DINOv2-base and OpenLlama-13b **(b)** DINOv2-giant and OpenLlama-13b

Fig. 14: Cross-modal mutual kNN alignment as gallery grows from WIT-1024 to WIT-1M for additional, stronger model pairs. Replacing DINOv2-base with the stronger DINOv2-giant and OpenLlama-3b (Fig. 4 in the paper) with OpenLlama-13b does not prevent the drop. *This suggests that the degradation in mutual kNN alignment was not a result of the limitation of any individual model.*

Replacing DINOv2-base with the stronger DINOv2-giant and OpenLlama-3b with OpenLlama-13b does not change the pattern. We observe that mutual kNN still drops at scale. This is consistent with Fig. 3b in the paper, which already shows low alignment scores across different LLMs at WIT-1M scale. **This confirms that the degradation reported in the paper is not specific to that particular choice of models**.

B Additional ImageNet experiments

The controlled experimental setting on the ImageNet validation set in Sec. 4.2 of the paper provides one of our key findings: models individually retrieve correct-class neighbors at increasing rates as the gallery densifies, yet cross-modal agreement remains flat. Here, we verify that the ImageNet validation set serves as a suitable test bed. Furthermore, we confirm that our observations are not limited to our choice of models or metric settings.

B.1 The ImageNet validation set is denser than WIT-1024

A natural question is how the gallery density in our ImageNet experiments compares to the WIT data. As shown in Tab. 2, nearest-neighbor cosine similarities on the ImageNet validation set are substantially higher than on WIT-1024 and comparable to WIT-1M. Even for only one image per class in the gallery (ipc=1), the ImageNet validation set provides a denser retrieval setting than WIT-1024.

This confirms that the ImageNet experiments operate in a denser retrieval regime comparable to WIT-1M, making this a meaningful test bed.

Table 2: Nearest-neighbor distances across gallery sizes. ImageNet, even with only one image per class in the gallery (ipc=1), has neighbor distances comparable to WIT-1M, confirming that it operates in a similarly dense retrieval regime.

Gallery	Model	Dim	k=1	k=10
WIT-1024	DINOv2-base	768	0.799	0.717
WIT-1024	OpenLlama3b	3200	0.502	0.400
WIT-1M	DINOv2-base	768	0.906	0.888
WIT-1M	OpenLlama3b	3200	0.757	0.701
ImageNet ipc=1	DINOv2-base	768	0.823	0.763
ImageNet ipc=1	DINOv2-giant	1536	0.609	0.496
ImageNet ipc=1	OpenLlama3b	3200	0.928	0.904
ImageNet ipc=1	LLaMA-65B	8192	0.890	0.858
ImageNet ipc=49	DINOv2-base	768	0.887	0.861
ImageNet ipc=49	DINOv2-giant	1536	0.749	0.690
ImageNet ipc=49	OpenLlama3b	3200	0.954	0.944
ImageNet ipc=49	LLaMA-65B	8192	0.926	0.912

B.2 Stronger models do not close the gap for ImageNet

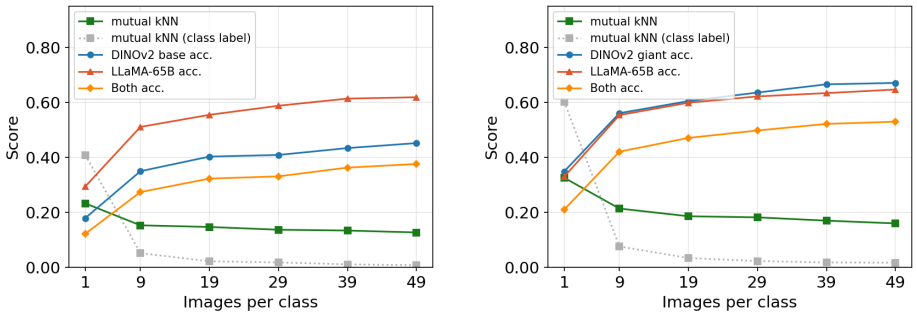
The ImageNet decomposition experiments in Sec. 4.2 in the main paper use DINOv2-base and OpenLlama-3b as its vision and language model respectively. Here, we probe whether stronger models would show better cross-modal agreement that closes the gap to unimodal retrieval accuracy.

In Fig. 15, we repeat the experiment with DINOv2-base paired with OpenLlama-65b, and DINOv2-giant paired with OpenLlama-65b. The pattern is unchanged: both models individually improve at retrieving correct-class neighbors as the gallery densifies, but strict cross-modal alignment remains flat. **Using substantially stronger models on both sides does not close the gap between individual retrieval accuracy and cross-modal agreement**.

B.3 ImageNet ablation shows a similar pattern for $k = 10$

The main paper reports the ImageNet decomposition experiments with mutual kNN scores for k=1. Here, we verify that the finding is not an artifact of this strict setting. We additionally present how mutual kNN with k=10 evolves when the gallery grows in Fig. 16 for two different model pairs.

We again observe that individual retrieval accuracy improves with gallery density while cross-modal alignment, here in terms of mutual kNN with k=10, does not.



(a) DINOv2-base and Llama-65b **(b)** DINOv2-giant and Llama-65b

Fig. 15: ImageNet per-modality retrieval accuracy and cross-modal mutual kNN alignment (k=1) as images / captions per class in gallery increase for different model pairs. *Even with substantially stronger models (OpenLlama-65b, DINOv2-giant), individual retrieval improves with gallery density while cross-modal alignment remains flat.*

C What happens with non-synthetic data that is not bijective?

In the main paper (Sec. 4.3), we use the CycleReward dataset [3] to test alignment when the bijective (one-to-one) assumption is relaxed with *synthetic* multi-modal correspondences. Here, we complement this analysis using non-synthetic many-to-many correspondences from the WIT dataset [82].

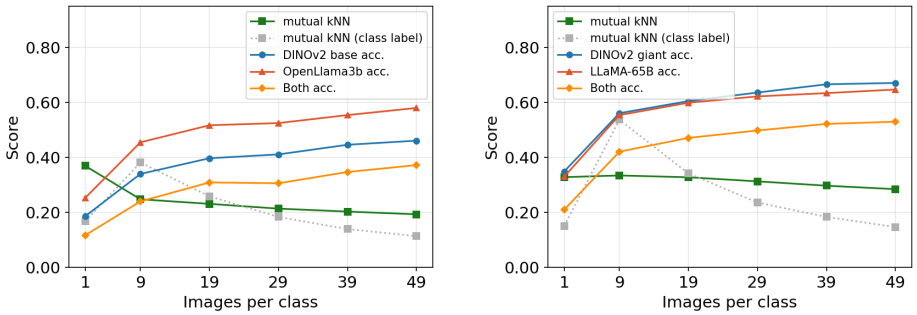
C.1 Non-synthetic dataset with many-to-many correspondences

Natural duplicates in WIT. The WIT dataset naturally contains many-to-many correspondences between images and captions: the same caption can describe many visually distinct images, and the same image is reused across Wikipedia articles with different corresponding text. Specifically, 7.1% of the captions are associated with more than one image, and 24.6% of the images have more than one caption before deduplication (see Sec. E.1). These naturally occurring one-to-many and many-to-one correspondences provide a complementary test bed for relaxing the bijective (one-to-one) setting without relying on generated images or captions.

Within-group deduplication. Grouping by caption text (for T2I) or by image (for I2T) can include *within-group* duplicates, i.e. a caption group may contain duplicate image, and an image group may contain repeated captions. After within-group deduplication, the number of qualifying one-to-many samples decreases from 7,844 to 4,975 for T2I and from 38,254 to 24,853 for I2T (Tab. 3).

Table 3: Natural duplicates in the WIT dataset after within-group deduplication.

	T2I	**I2T**
Unique elements	3.2 M	2.5 M
Appearing >1 time	7.1%	24.6%
Groups with ≥5 corresp.		
Before dedup	7,844	38,254
After dedup	4,975	24,853



(a) DINOv2-base and OpenLlama-3b **(b)** DINOv2-giant and OpenLlama-65b

Fig. 16: Per-modality retrieval accuracy and cross-modal mutual kNN alignment (k=10) as images / captions per class in gallery increase for two different model pairs. *Again, modalities individually improve with gallery density, but mutual kNN alignment, here with k=10, does not.*

We construct two complementary one-to-many datasets to mirror the experimental setup in Sec. 4.3 of the paper:

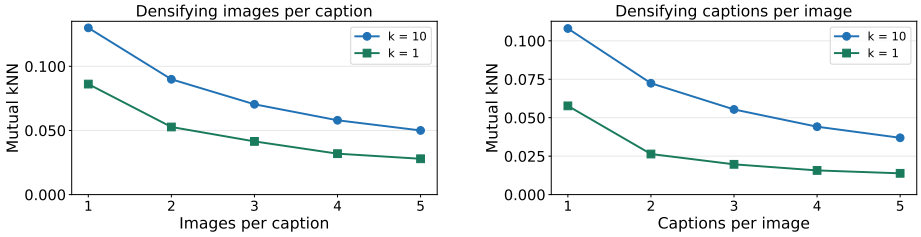
T2I (text-to-images): We select all 4,975 captions that are associated with at least 5 unique images. For each caption, we take 5 images, yielding a flat dataset of 24,875 image-text pairs.

I2T (image-to-texts): We identify 24,853 unique images that are paired with at least 5 distinct captions (by exact string matching). To match the T2I dataset size, we randomly subsample 4,975 images. For each image, we take 5 captions, again yielding 24,875 samples.

C.2 Mutual kNN also decreases on non-synthetic data when the bijective assumption is relaxed

We evaluate alignment between DINOv2-base [69] and OpenLlama-3b [24] on the WIT-based T2I and I2T datasets. Fig. 17 shows mutual kNN alignment for k=1 and k=10 as the number of images per caption and vice versa increases from 1 to 5. In both directions, alignment decreases as bijectivity is relaxed.

This is consistent with the results on the CycleReward dataset in the main paper (Fig. 10) and reinforces the conclusion that mutual kNN alignment is sensitive to the bijective assumption. When multiple valid correspondences exist for a query, the two modalities are less likely to agree on the same nearest neighbor, even if each individually retrieves a good match. **This confirms that the observed drop in alignment for non-bijective setting is not an artifact of *synthetic data*.**



(a) T2I: increasing the number of unique images per caption from 1 (bijective) to 5.

(b) I2T: increasing the number of unique captions per image from 1 (bijective) to 5.

Fig. 17: Effect of relaxing the bijective assumption on mutual kNN alignment using naturally occurring many-to-many correspondences in the WIT data (I2T and T2I subset). *Mutual kNN alignment on non-synthetic drops consistently as we increase the number of images per caption and vice versa. This confirms that the pattern observed on CycleReward is not an artifact of synthetic data.*

D Does the alignment vs performance trend predicted by Huh *et al.* [40] continue with recent LLMs?

To assess whether the alignment vs performance trend predicted by Huh *et al.* [40] continues with recent language models, we evaluate 55 LLMs (see Sec. E.3 for full list) on six standard benchmarks using the LM Evaluation Harness framework [20]. [40] originally used three benchmarks to measure language capability: HellaSwag [97], GSM8K [11], and $(1 - \texttt{bitsperbyte})$ on OpenWebText [26]. We replace OpenWebText with Wikitext [61] and extend this analysis to three additional benchmarks that probe different aspects of language understanding: ARC Challenge [10], MMLU [34], and LogiQA2 [58].

Benchmarks and metrics. Tab. 4 summarizes the evaluation configuration for each benchmark used in Sec. 4.4 of the paper (we used the default configurations from [20]).

Table 4: Overview of language model benchmarks used in Sec. 4.4 of the paper.

Benchmark	Capability	Few-shot	Metric
HellaSwag [97]	Commonsense reasoning	0	Accuracy
Wikitext [61]	Language modeling	0	$1 -$ bits per byte
ARC Challenge [10]	Science QA	4	Accuracy
GSM8K [11]	Math reasoning	5	Exact match
MMLU [34]	General knowledge	5	Accuracy
LogiQA2 [58]	Logical reasoning	5	Accuracy

Table 5: Average R^2 of the linear regression (fitted on the 19 *base models* from Huh *et al.* [40]) evaluated on the *base models* themselves and on the 36 recent models, across all four DINOv2 variants. Positive R^2_{avg}(new) indicates that the trend from the *base models* is a good predictor for the new models. Negative values indicate that the regression line is a worse predictor than the mean.

Benchmark	R^2_{avg} **(Huh *et al.*)**	R^2_{avg} **(new)**
HellaSwag	0.752	0.297
Wikitext	0.729	0.489
ARC	0.702	−0.575
GSM8K	0.336	−1.753
MMLU	0.430	−0.662
LogiQA2	0.431	−1.414

Does the alignment vs performance trend hold? For each benchmark and each DINOv2 variant, we fit a linear regression on the *base models* used in [40], predicting mutual kNN alignment a from benchmark performance p. We then evaluate how well this trend describes two populations:

R^2 **(Huh *et al.*):** The standard coefficient of determination on the data is used to fit the regression, *i.e.* R^2(Huh *et al.*) $= r^2$, where r is the Pearson correlation between mutual kNN alignment and language modelling benchmark score across the 19 *base models* from Huh *et al.* [40].

R^2 **(new models):** We apply the line fitted on the *base models* to the 36 recent models and compute the generalized R^2:

$$R^2(\text{new}) = 1 - \frac{\sum_{i \in \mathcal{M}_{\text{new}}} (a_i - \hat{a}_i)^2}{\sum_{i \in \mathcal{M}_{\text{new}}} (a_i - \bar{a}_{\text{new}})^2},$$

where $\hat{a}_i$ are the linear regression alignment predictions based on language performance p_i, a_i is the alignment score for the i-th model and $\bar{a}_{\text{new}}$ is the mean alignment of the new models. When $R^2(\text{new}) > 0$, the relation between alignment and language performance predicted in Huh *et al.* extrapolates; when $R^2(\text{new}) < 0$, the regression line is a worse predictor than simply predicting the average $\bar{a}_{\text{new}}$. R^2_{avg} values are reported in Tab. 5 and Figs. 18 and 19.

The results reveal a split across language modelling benchmarks. For HellaSwag and Wikitext, the relation between alignment and language performance observed by Huh *et al.* partially extends to recent models: the R^2_{avg} on new models remains positive (0.297 and 0.489, respectively), indicating that stronger language models according to these benchmarks have higher mutual kNN alignment with DINOv2. Both benchmarks primarily measure next-token prediction quality and commonsense language understanding, which are closely related to the pretraining objective of autoregressive LLMs.

In contrast, for the four benchmarks that probe more specialized reasoning abilities: ARC (science QA), GSM8K (arithmetic), MMLU (general knowledge),

and LogiQA2 (logical reasoning), the relation between alignment and language performance predicted from the *base models* Huh *et al*. does not appear to hold for this set of recent models.

Specifically, the R^2_{avg} on new models is consistently negative, ranging from -0.575 (ARC) to -1.753 (GSM8K). This means that the linear fit from the *base models* from Huh *et al*. [40] is a worse predictor of alignment for recent models than simply predicting the mean. In Figs. 18 and 19, we observe that recent models that are stronger than the best *base model* (Meta-Llama-3-70B) do not show higher mutual kNN alignment with DINOv2 features. Instead, their alignment scores seem to saturate or decrease.

We note that the 36 added (new) models are heterogeneous. They include new models trained on next-token prediction (pre-training), instruction-tuned models, and reasoning-distilled models (e.g. DeepSeek-R1-Distill). We treat them as a single population to test whether the trend extrapolates to recent LLMs.

The above results support and extend the finding from Sec. 4.4 of the main paper. **The relationship between alignment and language performance from [40] holds for core language modelling benchmarks, but does not seem to generalize to reasoning benchmarks.**

E Experimental setup

In Sec. E.1, we provide additional details about the deduplication pipeline for the WIT-1M and LAION-15M datasets. We then describe the captioning pipeline used for WIT-1M-recap and for the ImageNet validation set in Sec. E.2.

E.1 WIT-1M and LAION-15M datasets

Image deduplication. We deduplicate the gallery pools from WIT [82] and LAION400M [78] at the image level using perceptual hashing [96]. For each image, a 64-bit hash $\mathbf{h}_i \in \{0,1\}^{64}$ is computed. For this, we first convert the image to grayscale, resize it to 32×32, apply a 2D Discrete Cosine Transform, and threshold the top-left 8×8 low-frequency coefficients against their median. This produces a binary fingerprint that is robust to minor changes, e.g. due to recompression. We consider images i and j duplicates if their Hamming distance satisfies

$$d\text{H}(\mathbf{h}_i, \mathbf{h}_j) \leq 2,$$

measuring the number of bit positions at which two hashes differ:

$$d_{\text{H}}(\mathbf{h}_i, \mathbf{h}_j) = \sum_{b=1}^{64} \mathbf{1}[h_{i,b} \neq h_{j,b}].$$

We perform deduplication of the gallery against the WIT-1024 images, and within the gallery by keeping the first occurrence in the case of image duplicates.

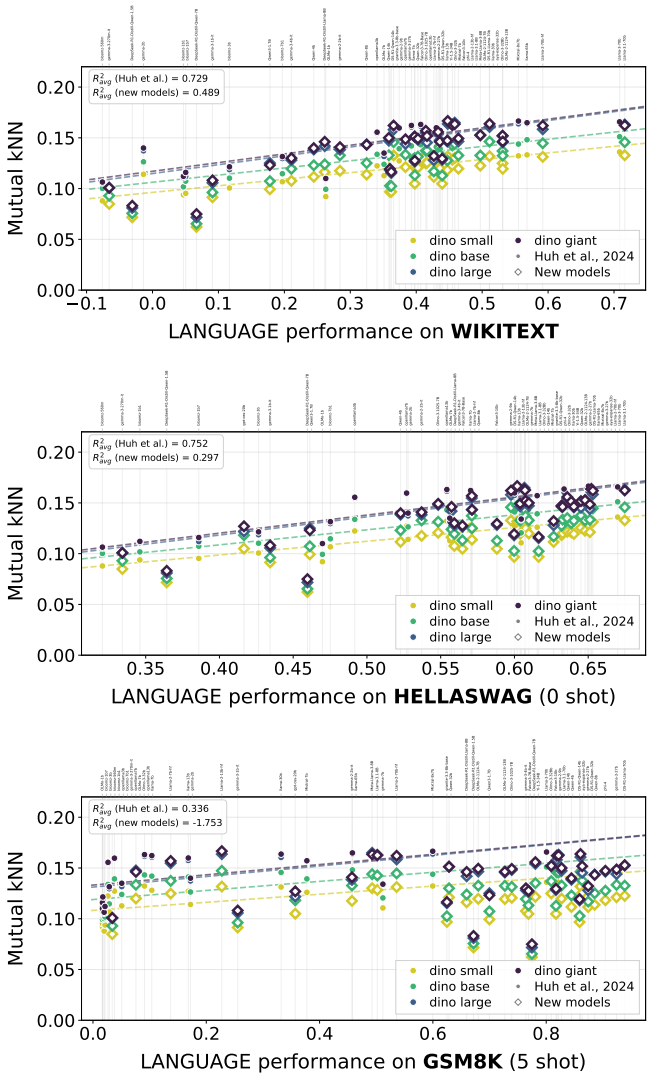


Fig. 18: Mutual kNN alignment vs. language benchmark performance for 55 LLMs across four DINOv2 variants, on WikiText, HellaSwag, and GSM8K. Dashed lines show the linear trend fit to the 19 *base models* from [40]. For WikiText and HellaSwag (top two plots), recent models roughly follow the trend. For GSM8K (bottom plot), the trend is not followed.

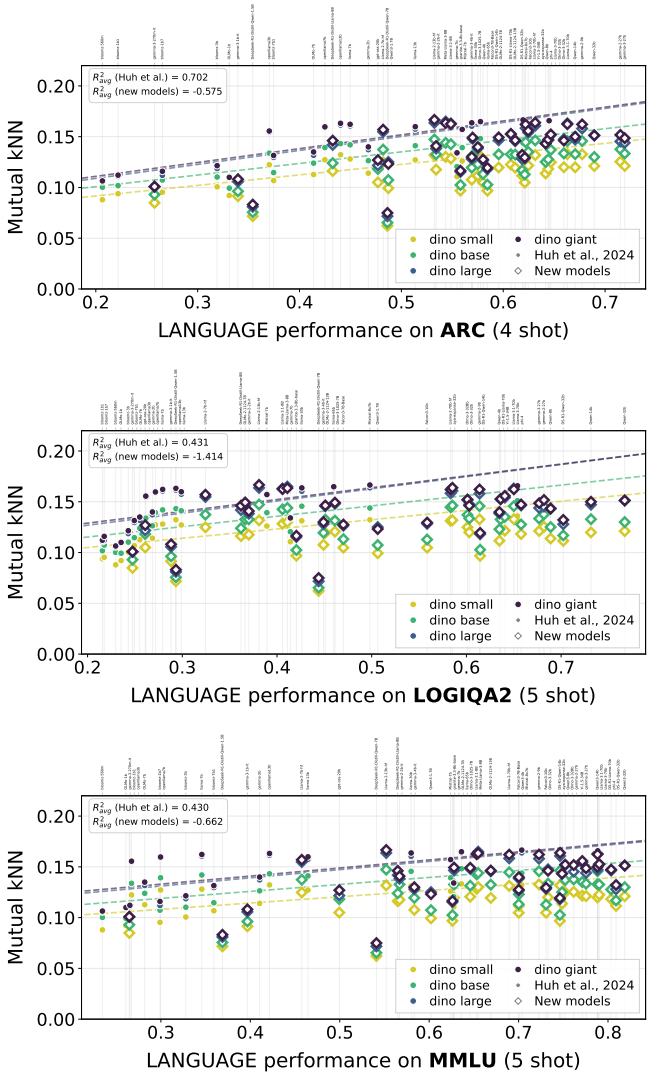


Fig. 19: Mutual kNN alignment vs. language benchmark performance for 55 LLMs across four DINOv2 variants, on ARC, LogiQA2, and MMLU. As with GSM8K, the alignment-performance trend from [40] does not extrapolate to recent models on any of these reasoning benchmarks. Stronger models do not appear to show higher mutual kNN alignment with DINOv2 features.

Table 6: Deduplication statistics for the WIT-1M and LAION-15M gallery pools. Image duplicates are detected via perceptual hashing (pHash) with Hamming distance ≤ 2. Caption duplicates are detected by exact string matching. WIT-1M and LAION-15M are 1M and 15M image-caption pairs randomly sampled from the remaining final pool.

	WIT-1M	**LAION-15M**
Raw pool	3,582,610	20,000,000
Image deduplication		
Duplicates (with WIT-1024)	2,847	53
Duplicates (within gallery)	2,164,343	3,371,128
Pool after image deduplication	2,486,852	17,941,016
Caption deduplication		
Duplicates (with WIT-1024)	52	14
Duplicates (within gallery)	97,654	642,895
Final pool size	**2,389,146**	**17,298,107**

Table 7: Distribution of caption duplicates in the WIT and LAION galleries after image deduplication. Note that the "unique captions" include some captions that are removed as query matches for the final pool.

	WIT		**LAION**	
Copies per caption	Unique captions	Total samples	Unique captions	Total samples
1	2,357,353	2,357,353	16,897,221	16,897,221
2	22,799	45,598	316,109	632,218
3	3,807	11,421	48,864	146,592
4	1,529	6,116	15,422	61,688
5	787	3,935	6,932	34,660
6–10	1,520	11,431	9,596	69,614
11–100	1,283	29,765	3,838	76,200
101–1,000	118	27,164	103	12,376
>1,000	2	5,077	4	14,588
Total	2,389,198	2,486,852	17,298,111	17,941,016

Caption deduplication. In addition to image deduplication, we do a text deduplication pass to remove gallery samples with captions identical to another gallery sample or to a WIT-1024 query caption. Duplicate captions are undesirable because they allow trivial text-based query-gallery matching, inflating retrieval scores regardless of visual representations.

We use exact string matching and remove any gallery samples that match WIT-1024 captions. Among the gallery samples, we discard duplicates and keep only the first occurrence of a duplicate caption-sample.

WIT-1M. We obtain a raw pool of 3,582,610 samples from the English-text WIT dataset [82]. To construct the English-only subset of the dataset, we used a subset of [79,80]. Since [79] only contains the image URLs, we retrieved the corresponding images from [82]. The raw pool undergoes our deduplication pipeline, resulting in 2,389,146 samples. We randomly sampled 1 million samples for the WIT-1M dataset. Deduplication statistics are provided in Tab. 6 and Tab. 7. In WIT,

31,845 captions appear repeatedly, accounting for 129,499 samples (5.21%); the most frequent captions are `coat of arms` (2716×) and `Town hall` (2361×).

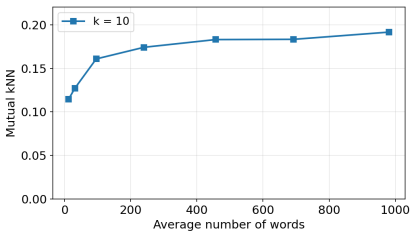
LAION-15M. We randomly sample 20M samples from the LAION-400M dataset [78] which consists of English image-text pairs. We randomly sample 20M samples as our raw pool which undergoes our deduplication pipeline. Finally, we randomly sample 15M from the final pool after deduplication, resulting in our LAION-15M data pool. Deduplication statistics are provided in Tab. 6 and Tab. 7. In LAION, 400,890 unique captions appear repeatedly, accounting for 1,043,795 samples (5.82%); the most frequent captions are `Patent Drawing` (10027×) and `Throw Pillow` (3246×).

E.2 Captioning pipeline for the ImageNet validation set and WIT-1M-recap

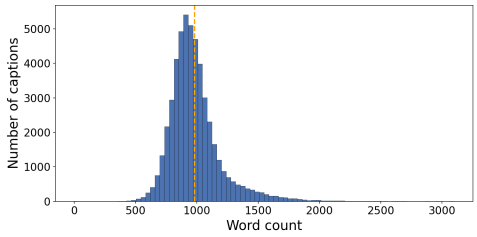
We used gemini-3-flash-preview [70,85] for captioning the images in the ImageNet validation [15] set and in WIT-1M. Specifically, we used the following text prompt for each image.

```
You are a precise image description system. Describe the image in the
    following JSON format.
Return ONLY a valid JSON object with exactly these 7 keys. No text before or
    after the JSON.
{
  "one_sentence": "<exactly one sentence, strictly fewer than 15 words>",
  "short":        "<2-3 sentences, 20-40 words total>",
  "100w":         "<a paragraph, approximately 100 words>",
  "250w":         "<several paragraphs, approximately 250 words>",
  "500w":         "<detailed description, approximately 500 words>",
  "750w":         "<thorough description covering all visual details,
      approximately 750 words>",
  "extreme_long": "<maximally detailed description covering every visible
      element, texture, color, spatial relationship, lighting, and context.
      YOU MUST WRITE AT LEAST 1000 WORDS. If your draft is under 1000 words,
      keep adding more detail about textures, materials, lighting, spatial
      layout, colors, and any other visible elements until you reach at least
       1000 words.
  Target 1000-1500 words.>"
}
Be factual and visual. Describe what you actually see: objects, people,
    animals, colors, textures, spatial relationships, background, lighting,
    and mood. Do not invent information not visible in the image.
```

For the ImageNet validation set, we perform experiments with captions of the `extreme_long` type. As shown in Fig. 20a, alignment scores increase with caption length. Captions of approximately 500 words achieve scores close to the maximum, while the longest captions yield the best mutual kNN alignment scores between DINOv2-base and OpenLlama-3b on the ImageNet validation set. A distribution over the number of words for captions of the `extreme_long` caption type is shown in Fig. 20b. Despite prompting the model to produce at least 1,000 words per caption, 63.1% of captions fall below this target. The average caption length is 981 words.



(a) Mutual kNN alignment for DINOv2-base and OpenLlama-3b increases with longer captions on the ImageNet validation set.



(b) Distribution of word counts for Gemini-generated `extreme_long` captions across 49,984 ImageNet validation images.

Fig. 20: Generated image captions for the ImageNet validation set. a) shows the mutual kNN alignment using captions of different lengths between DINOv2-base and OpenLlama3b on the ImageNet validation set. As also shown in [40], longer detailed captions yield higher alignment scores. b) shows the distribution over caption length (word count). We use captions of on average 981 words for our ImageNet experiments.

For WIT-1M-recap, we caption the 1 million images in the WIT-1M dataset using the `500w` variant for computational efficiency. This resulted in captions for 999,971 (29 images did not get processed by gemini-3-flash-preview) images of on average 478 words.

E.3 Models and feature extraction pipeline

Our feature extraction pipeline is based on the experimental protocol from Huh *et al*. [40], which we extend to include additional LLMs.

Vision models. On the vision side, we use four DINOv2 [69] variants: ViT-S/14 (384-d), ViT-B/14 (768-d), ViT-L/14 (1024-d), and ViT-G/14 (1536-d) [16], loaded via the `timm` [93] library. For each image, we extract the CLS token representation from every transformer layer, yielding a per-sample feature tensor of shape $L \times d$, where L is the number of layers and d the feature dimensionality.

Language models. We evaluate 55 large language models spanning 13 model families. The first group comprises the 19 *base models* used by Huh *et al*. [40]: BLOOMZ (560M–7.1B) [66], OpenLlama (3B–13B) [24], LLaMA (7B–65B) [87], OLMo (1B, 7B) [28], Gemma (2B, 7B) [22], Mistral-7B [43], Mixtral-8×7B [44], and Meta-Llama-3-70B [27].
For the trend analysis in Sec. D, we extend this set with 36 recent models: LLaMA-2 (7B–70B) [88], Llama-3/3.1 [27], OLMo-2/3 [67], Ministral-3 (3B–14B) [56], Gemma-2 (2B–27B) [21], Gemma-3 (270M–27B) [23], DeepSeek-R1-Distill (1.5B–70B) [14], Qwen3 (1.7B–32B) [72], Falcon3 (7B, 10B) [84], 01.AI Yi-1.5 (34B) [95], IBM Granite (8b) [74], and OpenAI GPT-OSS (20b) [68].
For each model, we extract hidden-state representations from all layers. Following [40], we apply average pooling over non-padding tokens to obtain a single vector per layer.

F Additional qualitative results

We present additional qualitative results for nearest-neighbor retrieval at different gallery scales on WIT-1M and LAION-15M in Figs. 21 and 22 and Figs. 23 and 24 respectively. In addition to the near-duplicate matches in Figs. 5 and 6 of the main paper, we here show further examples for cross-modal agreement (green-bordered matches) at scale when the modalities happen to select the same neighbor. Others show agreement at WIT-1024 that breaks down as the gallery densifies. In those cases, each modality individually finds a better match at scale, but they no longer agree on the same one.

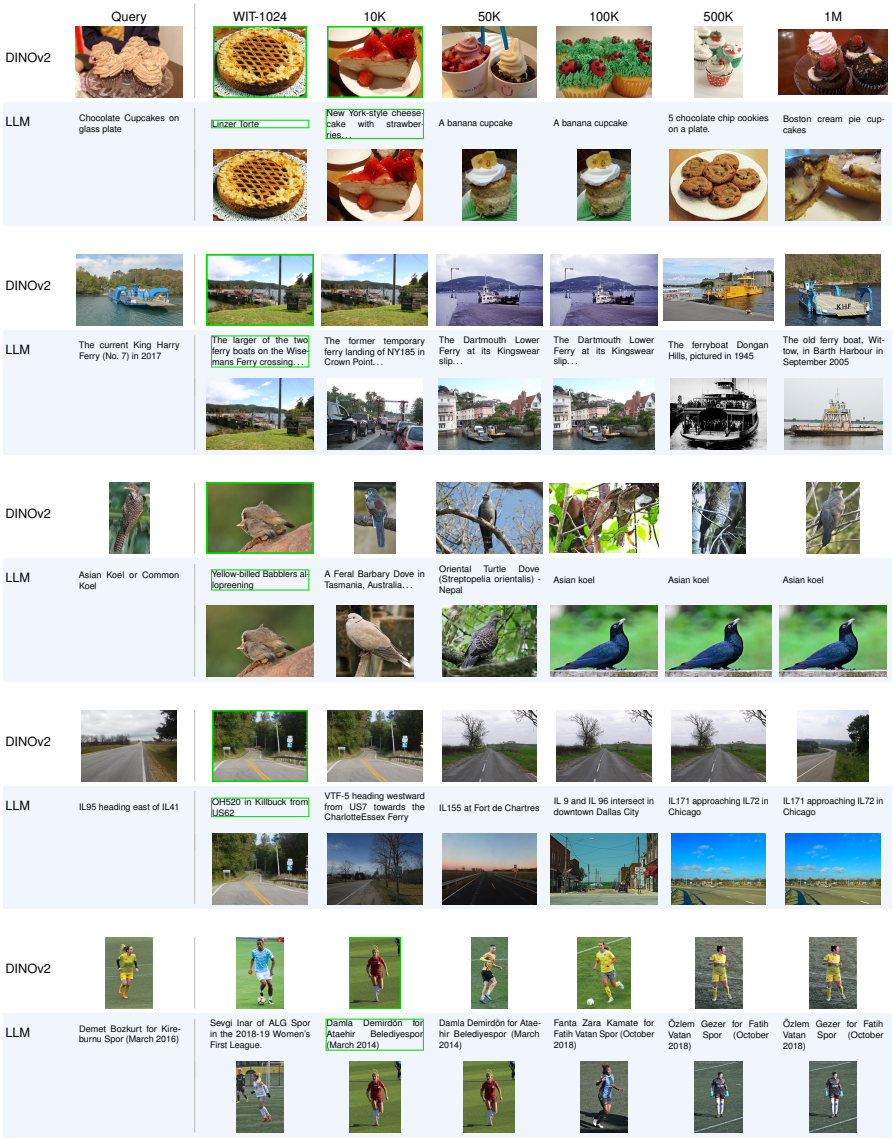


Fig. 21: Additional nearest-neighbor examples with DINOv2 and OpenLlama-3b for k=1 across gallery scales on the WIT-1M dataset. For OpenLlama-3b, we show the (partial) retrieved captions along with the corresponding reference image (LLM-ref) for visualisation. Green-bordered captions and images indicate a mutual kNN match across modalities.

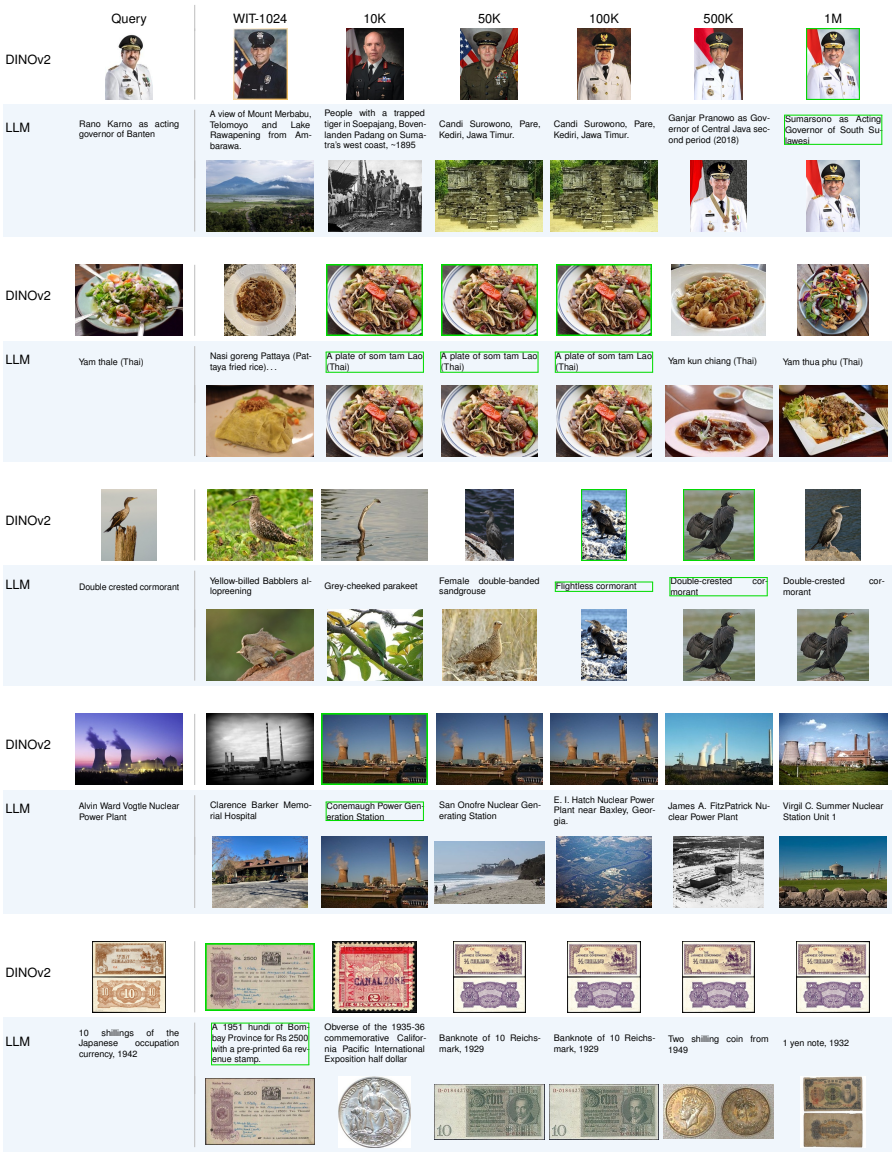


Fig. 22: Additional nearest-neighbor examples with DINOv2 and OpenLlama-3b for k=1 across gallery scales on the WIT-1M dataset. Green-bordered captions and images indicate a mutual kNN match across modalities.

Fig. 23: Additional nearest-neighbor examples with DINOv2 and OpenLlama-3b for k=1 across gallery scales on the LAION-15M dataset. Green-bordered captions and images indicate a mutual kNN match across modalities.

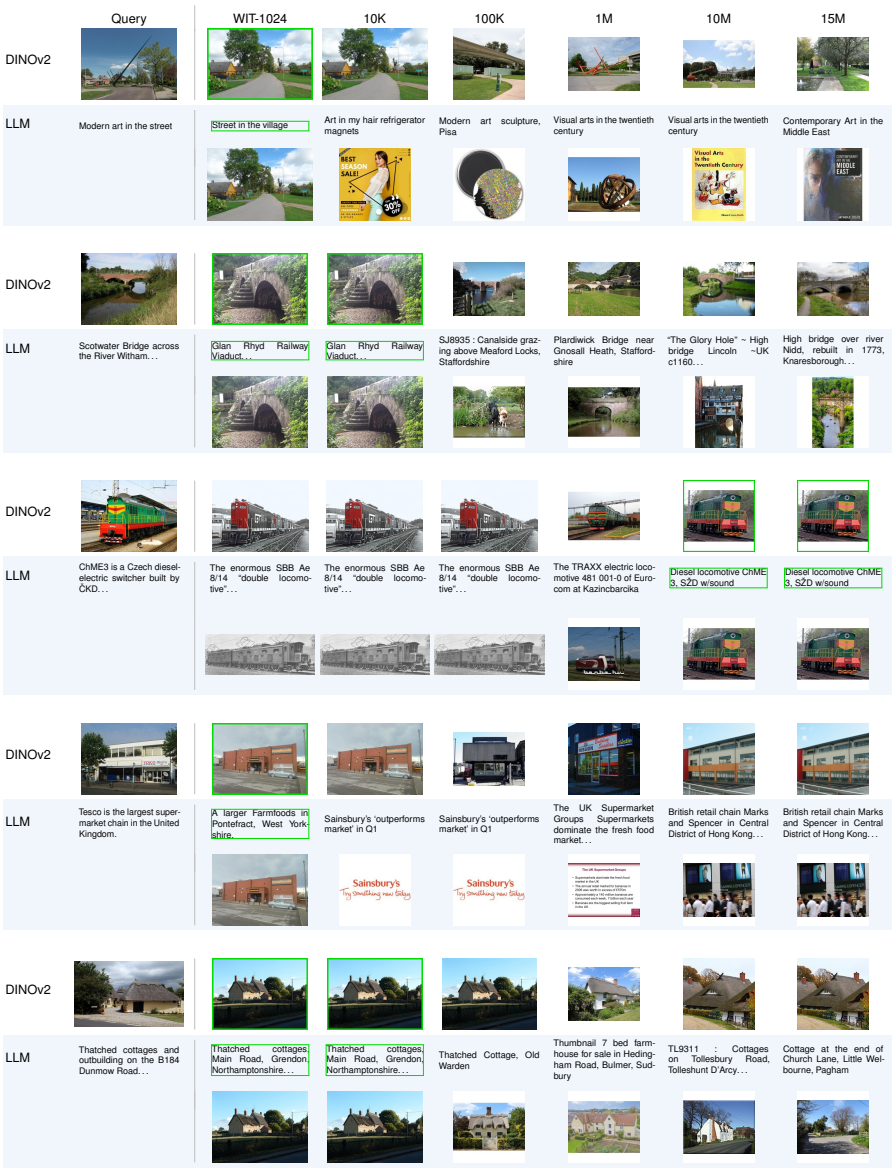


Fig. 24: Additional nearest-neighbor examples with DINOv2 and OpenLlama-3b for k=1 across gallery scales on the LAION-15M dataset. Green-bordered captions and images indicate a mutual kNN match across modalities.