When Does Perceptual Alignment Benefit Vision Representations?

Shobhita Sundaram[1*]   **Stephanie Fu**[2*]   **Lukas Muttenthaler**[3,4†]

Netanel Y. Tamir[5]   **Lucy Chai**[1]   **Simon Kornblith**[6]   **Trevor Darrell**[2]   **Phillip Isola**[1]

[1]MIT   [2]U.C. Berkeley   [3]TU Berlin   [4]BIFOLD   [5]Weizmann Institute of Science   [6]Anthropic
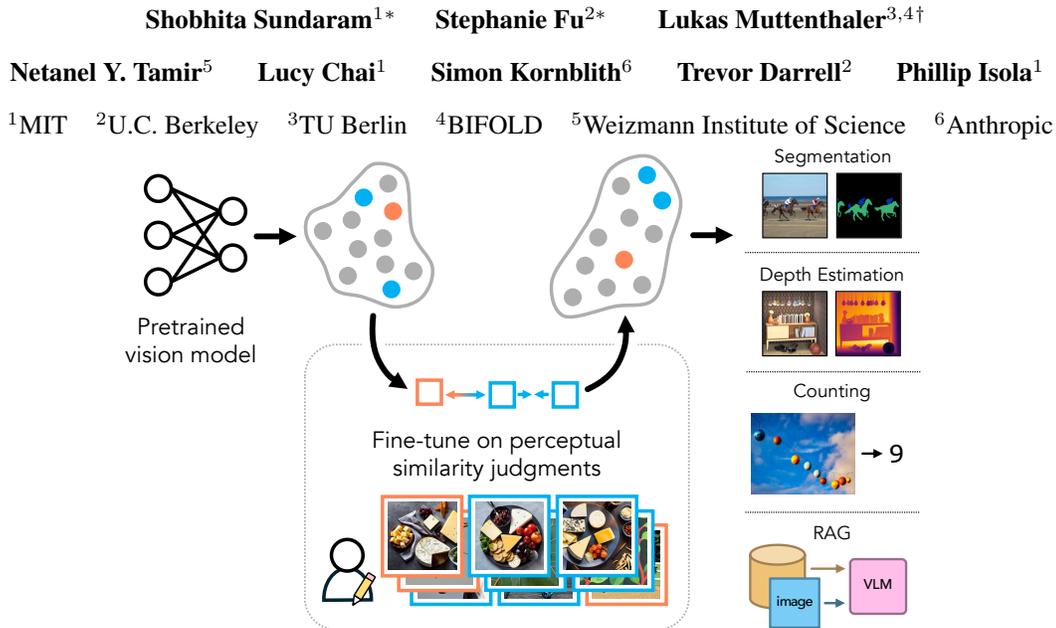
Figure 1: **Does human perceptual alignment improve vision representations?** Vision models have been shown to learn useful image representations through large-scale pretraining (e.g., CLIP, DINO). We find that additionally aligning these models to human perceptual judgments yields representations that *improve* upon the original backbones across many downstream tasks, including counting, segmentation, depth estimation, instance retrieval, and retrieval-augmented generation, while degrading performance in natural classification tasks. Our blog post and code are available at `percep-align.github.io`.

Abstract

Humans judge perceptual similarity according to diverse visual attributes, including scene layout, subject location, and camera pose. Existing vision models understand a wide range of semantic abstractions but improperly weigh these attributes and thus make inferences misaligned with human perception. While vision representations have previously benefited from alignment in contexts like image generation, the utility of perceptually aligned representations in general-purpose settings remains unclear. Here, we investigate how aligning vision representations to human perceptual judgments impacts their usability across diverse vision tasks. We finetune state-of-the-art models on human similarity judgments for image triplets and evaluate them across standard benchmarks. We find that perceptual alignment yields representations that *improve* upon the original backbones across many tasks, including counting, segmentation, depth estimation, instance retrieval, and retrieval-augmented generation, while deteriorating performance on natural classification. Performance is widely preserved on other tasks, including specialized out-of-distribution domains such as in medical imaging and 3D environment frames. Our results suggest that injecting an inductive bias about human perceptual knowledge into vision models can contribute to better representations.

---

[*]Equal contribution.   [†]Work partly done while a Student Researcher at Google DeepMind.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Our sense of similarity is crucial to how we perceive and act in the world. Consider the range of factors that might influence your judgment of visual similarity: layout, color, perspective, semantics, and more. These characteristics shape our local inferences about object relationships, enabling us to build a comprehensive understanding of the visual world.

Given the importance of similarity judgments in our visual perception, it follows that aligning vision models to these judgments could help them develop more human-like visual capabilities. In the language domain, we have seen the power of aligning LLMs to human feedback [RLHF; 9], resulting in safer and more useful models [3]. Similar trends are emerging in vision, where there is growing interest in alignment with human perceptual judgments to improve vision representations [e.g., 61, 17, 42, 18, 43]. Human feedback has also been used to improve the aesthetic quality of diffusion-generated images [58, 16, 30], increase retrieval capabilities [18], and improve downstream task performance of image/text models [42]. While acknowledging this wide space of human preference data used for alignment in vision and language, we focus our study on image similarity judgments, which give us a more stable and shared measure of human visual perception [18, 41, 54, 43].

Although there is consensus that alignment to perceptual judgements can enable specific goals (e.g., similarity [18, 61, 41, 42]), their utility in creating *general-purpose* representations is less clear. Does human alignment improve model performance by leveraging human-provided labels as an inductive bias, or compromise the model's original representational power by diverting it towards a separate task? Recent work suggests that the conclusions are nuanced: naively incorporating human percep­tual knowledge into model finetuning can distort representations, requiring strong regularization to maintain downstream performance while increasing the representations' interpretability and align­ment [42]. Furthermore, objective function and training data appear to matter more than architecture or model size for aligning to perceptual judgments [41]. The nature of human preference labels also plays a crucial role, as different types of labels produce distinct learning signals. For instance, fine-tuning vision backbones with mid-level perceptual similarity judgments yields representations well-suited for image retrieval tasks [18], while low-level similarity labels are more effective for image reconstruction losses [61].

It is clear from this body of work that careful perceptual alignment helps with specific tasks, however the effect of these adjustments on the models' representation spaces is less well understood. That is, are models aligned to perceptual judgments only better at specific tasks – such as predicting image similarity – or do they, like humans, actually have better general-purpose representations? Here, we evaluate the usefulness of human-aligned representations not just in predicting perceptual judgments but also on standard vision benchmarks requiring diverse notions of visual understanding. We finetune several state-of-the-art models – including CLIP, DINO, DINOv2, and SynCLR – on NIGHTS, a dataset of human similarity judgments over synthetic image triplets [18], and evaluate the finetuned models on standard tasks. Our experiments suggest that these human-aligned representations demonstrate strong gains compared to the original model backbones, even on tasks requiring skills beyond those needed during training or perceptual alignment. We also identify limitations of human perceptual alignment, and find that finetuning can sacrifice performance on some natural data tasks in which models had strong prior performance. In summary, our contributions are the following:

- We investigate the effects of aligning pretrained vision models to human perceptual judg­ments on various downstream visual recognition tasks.
- We find that propagating global image-level human similarity annotations to ViT patch tokens benefits downstream dense prediction tasks such as depth prediction and segmentation.
- We show that human-aligned representations are also beneficial in retrieval-based tasks requiring global image understanding, including retrieval-augmented generation for recent vision-language models, counting-based retrieval, and instance-based retrieval.
- We ablate the effect of various human-annotated visual similarity datasets, and find that mid-level image similarity, as opposed to low-level pixel-level variations or high-level semantic associations, offers the largest improvements in generalization capabilities.

2 Related Works

Vision backbones as learned feature extractors. Using the intermediate activations of deep net­works has been a long-standing strategy, originally used as a way of harnessing data-driven priors

2

from models trained on fewer large datasets and repurposing them for downstream tasks where such expensive supervision is less readily available [12, 50]. For instance, ImageNet pretrained models have proven useful for tasks such as texture synthesis [19], style-transfer [20], and super-resolution [27]. But given the challenges of collecting large-scale labelled datasets using human annotators, other approaches such as self-supervised learning [6, 7, 59, 4, 44] and contrastive learning on image/alt text pairs [46, 8, 26] have since superseded supervised learning as the standard approach for training vision models. Such models have been shown to yield rich multipurpose representations that can generalize to a variety of tasks involving visual perception, scene understanding, and image generation [51, 38, 52, 31, 39, 57].

View selection for self-supervised learning. The self-supervised, contrastive-learning objective aims to maximize the feature similarity between similar views of the data and minimize the similarity between different (negative) views. While originally the negative views are selected randomly from a pool of images [6, 22], this often results in requiring large batch sizes to learn useful representations. The process of selecting these positive and negative learning examples remains an active research area. Recent approaches have suggested that alternative strategies including hard-negative mining [47], nearest neighbors for positive pairs [14], and supervised labels when available [28] can provide useful learning signals. Other avenues include training the contrastive framework on synthetic data from generative models [25, 55], which provides an infinite data source of image variations rather than the traditional data augmentations typically used to generate alternative views.

Learning with human alignment. Learning directly from human feedback can provide models with more targeted supervision using fewer examples [9], and, thus, has been beneficial for fine-tuning large models towards specific human preferences [53, 62]. Ding et al. [11] show that image similarity metrics, trained on human feedback, can be subsequently used for evaluating the diversity of a text-to-image model. Several datasets aim to annotate human visual preferences in image similarity, including low-level [61] and high-level [23] image variations. In particular, Zhang et al. [61] and Fu et al. [18] use these datasets to learn a more human-aligned perceptual metric that improves the image retrieval abilities of vision models, while Muttenthaler et al. [42] used THINGS for learning a linear transform on top of vision representations to increase downstream task performance and improve alignment with human similarity judgments. These annotated human preference datasets serve as useful signals for contrastive objectives, providing direct supervision for positive and negative pairings that are aligned with human decisions; here, we investigate how fine-tuning self-supervised models using these human annotations impacts model performance on various downstream tasks.

3 Learning from perceptual judgments

We propose to use the method described below as a "second pretraining stage", which aligns the feature representations from large vision models with human perceptual judgments before applying them to downstream tasks. We note that prior work on this dataset aimed to develop a model for measuring image similarity based on human judgments. Here, we investigate if pretraining on this dataset leads to a better general-purpose representation, as measured by performance on different downstream tasks.

3.1 Human Similarity Annotations

We use the NIGHTS dataset to produce human-aligned variations of several large vision models [18]. The NIGHTS dataset consists of 20k synthetically generated image triplets, annotated with two alternative forced-choice human similarity judgments. These triplets are collected so that each has 6-10 unanimous human ratings, thus eliminating ambiguous cases where humans are likely to disagree.

NIGHTS consists of image triplets varying in *mid-level* information. Images in a triplet roughly share the same semantic content; however, they vary in pose, layout, shape, color, and the number of objects (see Fig. 15 in the Appendix for examples). Thus, the perceptual judgments indicate the shared visual appearance properties, as opposed to requiring higher-level semantic knowledge about the image content.

3.2 Image-level objective

Given a pre-trained backbone f_θ, we fine-tune its parameters θ on a dataset of triplets $\mathcal{D} = \{(x, \tilde{x_0}, \tilde{x_1}), y\}$, where x denotes a reference image, and $\tilde{x_0}$ and $\tilde{x_1}$ denote two variation images. The judgement $y \in \{0, 1\}$ indicates which of $\tilde{x_0}$ and $\tilde{x_1}$ is more similar to x. We measure distance

3

(dissimilarity) between two images $(x, \tilde{x_0})$ using the cosine distance between their respective image features $(f_\theta(x), f_\theta(\tilde{x_0}))$, which is defined as:

$$d(x, \tilde{x_0}) = 1 - \frac{f_\theta(x) \cdot f_\theta(\tilde{x_0})}{|f_\theta(x)||f_\theta(\tilde{x_0})|}. \tag{1}$$

We use an alignment loss to encourage the model to match human preferences:

$$\mathcal{L}_{\text{alignment}}(\theta) = \max(0, m - \Delta d \cdot \bar{y}), \tag{2}$$

where $\Delta d = d(x, \tilde{x_0}) - d(x, \tilde{x_1})$, $\bar{y}$ maps $y \in \{0, 1\} \to \{-1, 1\}$, and m is the margin, which we set to 0.05 following [18]. Note that this loss is equivalent to the triplet loss [5], and thus minimizes the cosine distance between the representations of the more similar pair and maximizes the distance between the representations of the other pair.

3.3 Patch-level objective
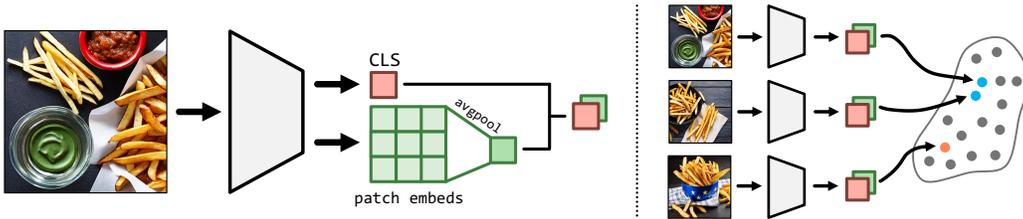
Figure 2: **Diagram of our feature extraction method when training with a patch-level objective.** Left: We extract the CLS and patch embeddings from DINO and DINOv2, perform a spatial average-pool on the patch embeddings, and concatenate [CLS, patch] vectors. Right: We train these concatenated features with a hinge loss, identical to the image-level objective.

The NIGHTS dataset contains global annotations of similarity at the image level, but the holistic label is the result of several local attributes, such as perspective, layout, foreground appearance etc. In addition to the global CLS token from the Vision Transformer model backbones, each model also contains a set of spatial patch embeddings. Propagating this global human annotation to individual patch tokens allows for spatial representations that are aligned with human similarity preferences. Thus, we formulate a patch alignment objective to optimize these local patch features jointly with the global image label.

The local objective only differs from the global objective in how the features are extracted. Instead of computing $\mathcal{L}(\text{CLS}_A, \text{CLS}_B)$, we compute

$$\mathcal{L}\left(\text{cat}[\text{CLS}_A, \text{pool}(\text{PATCH}_A)], \text{cat}[\text{CLS}_B, \text{pool}(\text{PATCH}_B)]\right).$$

CLS is of dimension $(1, d)$ and PATCH is (s, s, d) where s is the number of patches along each spatial dimension. We spatially average the patch tokens to get dimension $(1, d)$. We then concatenate the CLS and pooled patch tokens to get dimension $(1, 2d)$. We fine-tune the same alignment loss (see Eq. 2) on concatenated CLS and averaged-pooled patch tokens, and train heads for semantic segmentation and depth estimation on the resulting patch embeddings. Note that only experiments reported in Sec. 4.1 use this objective, as they require local features. For all other evaluations, we exclusively use CLS tokens (see Eq. 2) if not mentioned otherwise.

3.4 Implementation details

Vision Model Backbones. We fine-tune several state-of-the-art Vision Transformer [VIT; 13] backbones including DINO [4], DINOv2 [44], CLIP [46], OpenCLIP [8], and SynCLR [55]. For DINO, DINOv2, MAE, and SynCLR, we use the CLS token of the final layer. We extract the CLS token before layer norm for MAE and after it for the other backbones. For CLIP and OpenCLIP, we use the representations of the image encoder. We also experiment with concatenated features from DINO, CLIP, and OpenCLIP, used to train the DreamSim Ensemble in [18] (referred to as "ensemble" in our results). For each model, we use its base size, ViT-B.

We implicitly ablate the usage of synthetic triplets in NIGHTS by including SynCLR in our experiments; as that backbone was contrastively trained on generated images, additional performance changes can be attributed to human perceptual alignment.

4

Finetuning with human preference labels. We finetune each backbone using Low-Rank Adaptation (LoRA), which was found to achieve better alignment performance and efficiency than full fine-tuning in [18]. For more training and technical details see Sections C.1 and C.6 in the Appendix.

4 Experiments

In this section, we evaluate perceptually-aligned backbones against base models on common vision tasks. We study global representations through instance retrieval, object-counting, and retrieval-augmented generation experiments. Additionally, we find that local patch-level representations can be improved by tuning on image-level perceptual judgments, and show performance increases on semantic segmentation and depth estimation.

4.1 Dense Prediction

Semantic segmentation. Following the procedure detailed in Section 3.3 and Fig. 2, we LoRA-tune new backbones with perceptually-aligned CLS and patch tokens. To evaluate segmentation performance, we freeze these backbones and train a single linear layer transforming patch tokens to a segmentation map. We evaluate DINO and DINOv2 on standard segmentation benchmarks in Table 1 and show that human-aligned models boost performance in 16 out of 20 cases. Across all datasets and metrics, human-aligned DINO (denoted as DINO-HA) outperforms the base model and often achieves the highest mIoU and Pixel Accuracy (P.A.) overall. DINOv2-HA also outperforms its nonaligned counterpart on COCO and DAVIS2017.

We flag datasets already seen in the DINOv2 retrieval pretraining [44], such as Pascal VOC, ADE20k, and Cityscapes. If a dataset is already in-distribution for a backbone, fine-tuning on different data may be more likely to change the feature space such that that dataset is more out-of-distribution. Thus this is a potential confounding factor.

	Pascal VOC*		ADE20k*		Cityscapes*		COCO		DAVIS2017	
	mIoU	P.A.	mIoU	P.A.	mIoU	P.A.	mIoU	P.A.	mIoU	P.A.
DINO	0.729	0.800	0.342	0.548	0.562	0.749	0.810	0.901	0.809	0.886
DINO-HA	**0.745**[†]	**0.840**[†]	**0.391**	**0.602**	**0.573**[†]	**0.769**[†]	**0.816**[†]	**0.914**[†]	**0.818**[†]	**0.913**[†]
DINOv2	**0.686**	**0.803**	**0.441**[†]	0.645	**0.535**	0.737	0.759	0.877	0.794	0.906
DINOv2-HA	0.635	0.771	0.418	**0.700**[†]	0.528	**0.739**	**0.761**	**0.891**	**0.810**	**0.908**

Table 1: **Base and human-aligned model performance on semantic segmentation.** Aligned models largely outperform baselines, with DINO-HA achieving the highest performance across models for 4 out of 5 datasets. Note that Pascal VOC, ADE20k, and Cityscapes were included in DINOv2's retrieval pretraining. † indicates best score in the column.

Depth estimation. We follow the evaluation protocol of [44, 37] and train a single linear layer on frozen patch tokens to output a depth map with values mapped into 256 uniformly-distributed bins. This head is trained with a scale-invariant log loss introduced in [15] and a scale-invariant gradient-matching term as described in [36]. In Table 2, we report performance on monocular depth estimation and show that human-aligned models outperform base models in 27 out of 36 cases. Consistent with segmentation performance, human-aligned DINO outperforms the base model on all metrics across all datasets, and is often the highest-performing model overall (denoted by †). We also evaluate out-of-distribution generalization by training a depth head on NYUv2 and evaluating on the 4D Light Field dataset; combined with the performance boost even on datasets that DINOv2 was trained on (NYUv2 and SUN-RGBD), these results demonstrate that human-aligned models have strong generalization capabilities prior to any training for downstream tasks.

4.2 Retrieval-augmented generation

First introduced for text generation with non-parametric memory [34], retrieval-augmented generation (RAG) has become a popular method for selecting relevant few-shot examples when prompting large vision-language models (VLMs) [1, 2, 32, 33]. RAG evaluations go beyond conventional retrieval benchmarks on top-k recall, offering a more informative indicator of downstream large-model performance and utility in the multimodal domain. We evaluate OpenFlamingo [2]'s few-shot classification accuracy by using a vision backbone to retrieve a query image's 3 nearest neighbors and prepending the query image with those examples along with their class labels. Following OpenFlamingo's image classification evaluation framework [2], we extract the model's logits per

5

	NYUv2[*]					
	RMSE ($\downarrow$)	AbsRel ($\downarrow$)	log10 ($\downarrow$)	$\delta > 1.25$ ($\uparrow$)	$\delta > 1.25^2$ ($\uparrow$)	$\delta > 1.25^3$ ($\uparrow$)
DINO	1.034	**3.517**	0.173	0.415	0.746	0.895
DINO-HA	**1.032**	3.759	**0.169**[†]	**0.445**[†]	**0.761**	**0.900**
DINOv2	**1.003**[†]	3.188	**0.178**	**0.445**[†]	**0.785**[†]	**0.907**[†]
DINOv2-HA	1.062	**3.167**[†]	0.185	0.419	0.749	0.891
	NYUv2 $\rightarrow$ 4D LF					
	RMSE ($\downarrow$)	AbsRel ($\downarrow$)	log10 ($\downarrow$)	$\delta > 1.25$ ($\uparrow$)	$\delta > 1.25^2$ ($\uparrow$)	$\delta > 1.25^3$ ($\uparrow$)
DINO	3.817	0.543	0.380	0.160	0.335	0.454
DINO-HA	**3.757**	**0.538**	**0.362**	**0.187**	**0.346**	**0.463**
DINOv2	**3.460**[†]	0.497	**0.337**[†]	0.194	0.337	**0.464**[†]
DINOv2-HA	3.720	**0.487**[†]	0.356	**0.203**[†]	**0.356**[†]	0.455
	SUN-RGBD[*]					
	RMSE ($\downarrow$)	AbsRel ($\downarrow$)	log10 ($\downarrow$)	$\delta > 1.25$ ($\uparrow$)	$\delta > 1.25^2$ ($\uparrow$)	$\delta > 1.25^3$ ($\uparrow$)
DINO	4.900	2.350	0.533	0.205	0.385	0.524
DINO-HA	**4.788**[†]	**2.300**[†]	**0.526**[†]	**0.230**	**0.418**[†]	**0.531**[†]
DINOv2	5.200	3.023	0.615	0.173	0.309	0.444
DINOv2-HA	**5.082**	**2.904**	**0.599**	**0.237**[†]	**0.409**	**0.487**

Table 2: **Human-aligned DINO and DINOv2 performance on monocular depth estimation benchmarks.** Note that NYUv2 and SUN-RGBD were included in DINOv2's retrieval pretraining set, yet human-aligned DINOv2 still outperforms the base model on SUN-RGBD. Along with the results on an unseen test data domain (train on NYUv2 $\rightarrow$ test on 4D Light Field), these results demonstrate strong generalization performance of models aligned to human perceptual judgments. † indicates best score in the column.
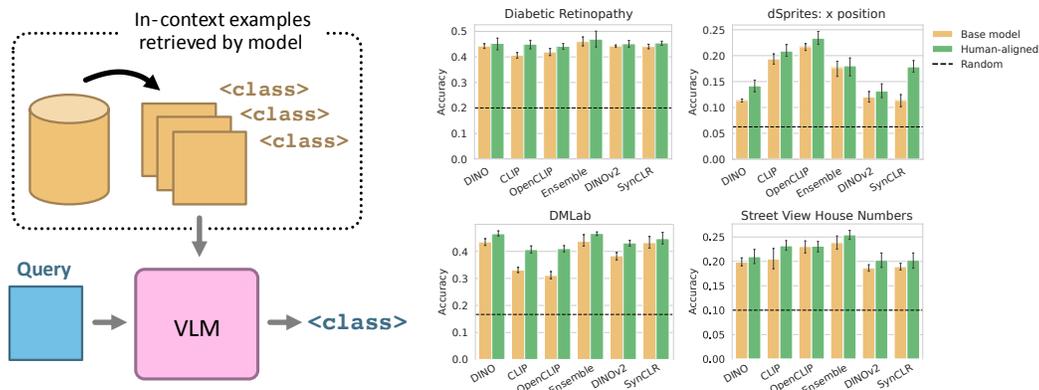
Figure 3: **Left: Diagram of evaluation setup for retrieval-augmented generation.** We retrieve the top-3 nearest image-prompt examples for each datasets and prompt OpenFlamingo with them before inputting the query image. **Right: Classification accuracy on VTAB [35] from wide-varying domains.** Error bars indicate 95% confidence interval over 5 random seeds.

object class and determine the model's decision by selecting the class with the largest log probability. See the Appendix for full details on the RAG experimental setup.

As illustrated in Fig. 3, classification accuracy on a wide variety of data domains improves with prompts retrieved by human-aligned models, compared to the original model backbones. Even in out-of-distribution domains such as medical imagery and 2D renders of game scenes, human-aligned models retrieve in-context examples that boost OpenFlamingo classification accuracy. These results suggest that human-aligned models can select more informative examples for in-context learning, thereby boosting the few-shot generalization abilities of a downstream multimodal VLM.

6

4.3 Counting

A well-documented limitation of large vision backbones is their performance on compositional tasks: in particular, on object counting [45]. We investigate how aligning to perceptual judgments affects performance on counting tasks via the FSC147, CARPK, and Clevr-Count (adapted from the original Clevr dataset by [60]) benchmarks by computer k-Nearest Neighbors accuracy on frozen vision representations. We report results in Table 3 and retrieval visualizations for few ($n = 3$) and many ($n = 8, 10$) objects in Fig. 4 and find that, across 6 different models, the human-aligned versions outperform their counterparts in 35 out of 36 cases. See the Appendix for full details on the counting experimental setup. Given that the perceptual similarity dataset we use for finetuning
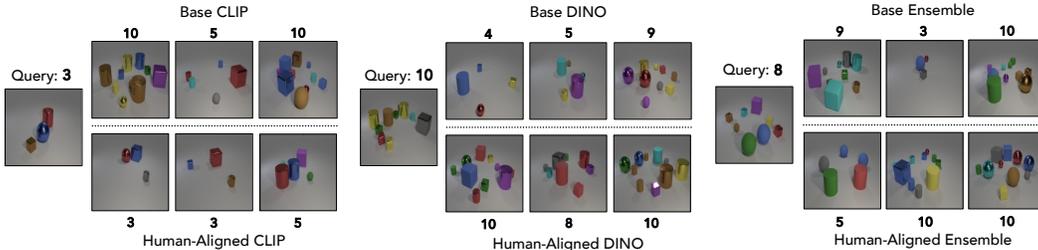
Figure 4: **Visualizations of nearest-neighbor examples retrieved by CLIP, DINO, and Ensemble models as well as their human-aligned versions.** Overall, we see retrieved images with more accurate object counts in CLIP-HA, DINO-HA, and Ensemble-HA across multiple nearest neighbors.

contains image-level similarity judgments, the consistent improvements that we observe on counting tasks, which requires local object awareness, is somewhat surprising. We hypothesize that the sensitivity of our human-aligned models to object counts may be a byproduct of NIGHTS examples themselves, many of which include image triplets with varying numbers of objects (see Fig. 15 in the Appendix). In terms of human perception, we note that humans consider object count when evaluating image similarity as soon as they develop counting profiency [40]. Thus, it is possible that given the prevalence of triplets with object-count variations in NIGHTS, the human annotations naturally capture this counting aware effect in the global image-level labels and propagate this information to the human-aligned models.
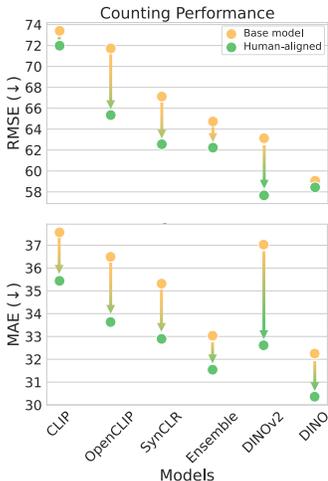
Figure 5: Performance improvements on Clevr-Count visualized by backbone for RMSE (top) and MAE (bottom), averaged across all datasets. Lower is better.

	FSC147		CARPK		Clever-Count	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
DINO	44.1	**118.7**	51.4	56.8	1.25	1.70
DINO-HA	**41.3**†	119.3	**48.7**	**54.5**†	**1.08**	**1.50**
DINOv2	57.5	128.3	52.4	59.5	1.18	1.60
DINOv2-HA	**44.9**	**113.6**†	**52.1**	**58.2**	**0.84**	**1.20**
CLIP	59.1	158.4	52.5	60.3	1.09	1.52
CLIP-HA	**53.2**	**156.0**	**52.3**	**58.8**	**0.81**	**1.15**†
OpenCLIP	54.4	153.5	54.1	60.3	0.97	1.36
OpenCLIP-HA	**50.2**	**139.1**	**49.9**	**55.8**	**0.80**†	**1.15**†
SynCLR	50.6	139.6	54.3	60.3	1.05	1.45
SynCLR-HA	**46.4**	**128.1**	**51.3**	**58.2**	**0.98**	**1.37**
Ensemble	48.4	132.4	49.6	60.3	1.10	1.51
Ensemble-HA	**45.4**	**130.3**	**48.4**†	**55.2**	**0.83**	**1.19**

Table 3: Error comparisons for base and human-aligned models on standard counting benchmarks. Though FSC147 and CARPK have examples with extreme object counts (tens and hundreds) unseen in the NIGHTS data, human-aligned models still achieve higher performance in each pair. † indicates best score in the column, lower is better.

4.4 Instance retrieval

7

In this section, we evaluate how aligning models to perceptual judgments affects their performance on the instance retrieval task. This task aims to retrieve images from a gallery containing a shared subject or object with the query image. At test time, retrieval is performed by computing the cosine similarity between the extracted features of the query and gallery images. Succeeding at this task requires that a representation be robust to recognizing instance identities under different lighting, backgrounds, poses, and other such shifts.

We evaluate all base models and their human-aligned counterparts on the Consumer-to-Shop benchmark of the DeepFashion2 dataset [21]. The benchmark consists of 10990 consumer "in the wild images" as the query set, and 21438 gallery images with matching clothing items to consumer images. Following the evaluation protocol from [21], we report Top-1, 3, and 5 accuracy in Table 4. Human aligned models outperform base

	DeepFashion2		
	Top-1	Top-3	Top-5
DINO	8.02	12.15	14.44
DINO-HA	**11.69**	**17.55**	**20.84**
DINOv2	5.95	8.47	9.98
DINOv2-HA	**9.03**	**13.57**	**16.30**
CLIP	4.59	6.89	8.23
CLIP-HA	**8.62**	**13.02**	**15.60**
OpenCLIP	14.88	22.74	27.36
OpenCLIP-HA	**16.85**	**24.58**	**28.64**
SynCLR	4.88	7.34	9.02
SynCLR-HA	**8.35**	**12.31**	**14.86**
Ensemble	13.54	20.01	23.54
Ensemble-HA	**23.39**[†]	**32.47**[†]	**37.20**[†]

Table 4: Top-1, -3, and -5 recall scores for instance retrieval on DeepFashion 2. † indicates best score in the column, higher is better.

models by a significant margin across all metrics and backbones (visualized in Fig. 6. In Fig. 7 we provide qualitative retrieval results. These results agree with prior work showing that training on NIGHTS improves performance in retrieving similar images to queries [18].
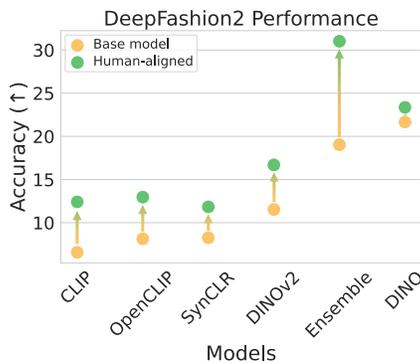
Figure 6: Performance improvements on the DeepFashion2 instance retrieval, task visualized by backbone and averaged across all k for top-k recall. Higher is better.
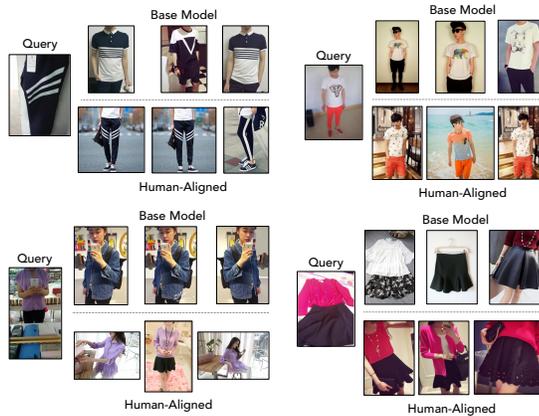
Figure 7: Examples of top-3 retrievals for a given query image on DeepFashion2. Overall, the human-aligned models return matching clothing items more frequently.

4.5 What type of human similarity annotation is most beneficial?

As evidenced by the current widespread use of large vision models [4, 44, 46], trained on massive datasets, model performance largely correlates with data scale. This raises the question: are our performance gains purely due to training the base models on additional data, or as a result of the perceptual qualities embedded in NIGHTS? To investigate this, we ablate the training dataset - rather than tune on 13,900 NIGHTS triplets, we train on three other image triplet datasets of the same size:

1. BAPPS [61]: Originally the training set for the LPIPS perceptual similarity metric [61], BAPPS consists of image patch triplets with various low-level distortions applied (e.g. color jitter, gaussian blur, JPEG compression artifacts).

2. THINGS [24]: This dataset contains image triplets with each image encoding a different concept (e.g. a triplet of images categorized as {airplane, elephant, football}), labeled by humans tasked to determine which concept is the odd-one-out.

3. ImageNet [48]: To ablate whether perceptual judgments at any level are needed, we construct an image triplet dataset by randomly selecting two images from one category and one image from another, labeling the first image pair as more similar to each other.

8

For all three datasets, we apply the same training settings as with the original LoRA-tuning on NIGHTS. See Fig. 8 for dataset ablations on object counting and instance retrieval.

Tuning on NIGHTS indeed provides the largest improvements across these tasks, with THINGS worsening performance overall and BAPPS/ImageNet having minimal effect. These trends may appear because BAPPS' photometric distortions are too low-level to impart any perceptual signal onto the backbones, THINGS encodes a higher-level conceptual similarity irrelevant to these mid-level vision tasks, and pre-trained vision models already perform quite well at discriminating ImageNet categories. Indeed, previous work [18] has found that similarity judgments by perceptual metrics trained on BAPPS correlate better to low-level metrics such as color than to semantic attributes.

In Section A.2 of the Appendix we find that the performance boost from NIGHTS over other datasets is also consistent across semantic segmentation and depth estimation. Conversely, tuning on NIGHTS fails to improve over other datasets on classification datasets; we further discuss this finding in Sections 5 and A.1.
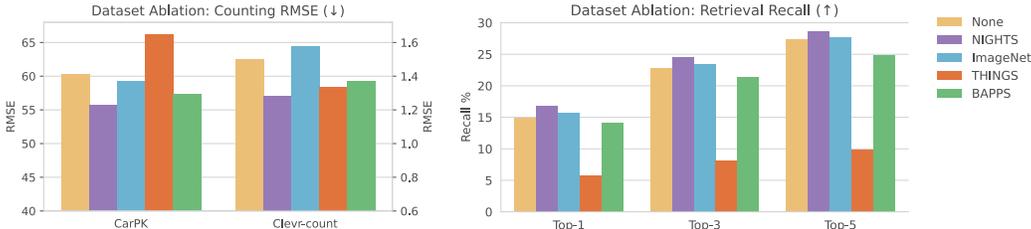
Figure 8: Evaluations comparing dataset utility on counting tasks (lower RMSE is better) and DeepFashion2 instance retrieval (higher recall is better). Across each task, tuning on NIGHTS yields the largest improvements while THINGS worsens performance and BAPPS/ImageNet makes minimal changes.

5 Discussion

Recently, the vision research community has converged on the idea that using human perception to improve machine perception can bolster the transfer of vision representations to downstream tasks [61, 17, 41, 18, 42, 43]. However, it remained unclear which tasks benefit most from alignment with human perceptual judgments; different tasks may demand representations that encode different levels of granularity and semantics. Here, we fine-tune modern backbones, pretrained on different tasks, on human perceptual judgments [18] and subsequently investigate which downstream tasks benefit. We develop a better understanding of how perceptual alignment affects performance on important tasks – e.g., segmentation, depth estimation, RAG – which may in turn inform future model training and data curation decisions across different applications. By evaluating competency at downstream tasks, we quantify what these representations capture and make decodable, enabling a better understanding of the human-aligned feature space. While we probe representations in terms of competency, understanding them in terms of their mechanism is a rich direction for future work.

We find widespread benefits from perceptual alignment across both image- and patch-level tasks. At the global level, performance consistently improves for retrieval-augmented generation, counting-based and instance-based retrieval. Moreover, propagating the supervision from image-level similarity judgments to ViT patch tokens improves performance on dense prediction tasks (semantic segmentation and depth prediction). Since model performance is closely linked to data scale, we also ablate the choice of perceptual dataset. While fine-tuning on NIGHTS — a dataset of mid-level perceptual judgments — leads to improvements across various tasks, fine-tuning on triplets from ImageNet [10], BAPPS [61], and THINGS [23] preserves or deteriorates transfer performance.

Why does fine-tuning on NIGHTS in particular lead to improvements? We hypothesize that the variations found in BAPPS and THINGS are solely high- or low-level, whereas the mid-level distortions in NIGHTS cover salient features that humans use when making inferences about what they see; these characteristics include style, pose, color, and count (see Fig.15), and largely correlate with the characteristics a model must successfully extract for many computer vision tasks. Previous work [18] found that models fine-tuned on NIGHTS seem to attend to both low-level and semantic attributes. Aligning a feature space to these concepts may be useful for visual tasks requiring both visual and semantic knowledge, such as retrieval, counting, segmentation, etc. This hypothesis may

9

also explain why tuning on NIGHTS hurts performance on fine-grained tasks, in which perceptually similar images may belong to different categories.

Limitations. Perceptual alignment does not appear to improve performance for standard image classification tasks such as natural image datasets in the VTAB benchmark [60] (see Tables 5a-b in the Appendix). This is surprising in light of recent findings that demonstrate downstream task improvements in image classification tasks for human-aligned representations [42]. Although it is hard to pinpoint the exact cause, we hypothesize two reasons: First, perceptual judgments at different levels of abstraction may be helpful for different downstream tasks. While the mid-level perceptual judgments in NIGHTS boost performance for retrieval-based and dense prediction tasks, they may not impart a useful inductive bias for standard image classification tasks; high-level semantic associations could simply be better suited for these kinds of tasks. Alternatively, the visual features that humans use to judge similarity may not be appropriately captured in classification accuracy metrics. Some VTAB datasets have numerical ground truth in which the distance from the correct answer is meaningful; thus, it may be ill-suited for classification, and better suited to continuous evaluations, which we report in sections of the paper.

Additionally, a key insight from our dataset ablations is that not all human preferences improve performance. Finetuning on perceptual datasets with solely high-level (THINGS) or low-level (BAPPS) variations hurts performance for many downstream tasks (Section 4.5 and A.2). Similarly, Muttenthaler et al. [43] recently discovered that finetuning vision models on THINGS can hurt downstream task transfer. Due to our evaluation of the synthetically-pretrained SynCLR [55], we attribute the drop in performance to perceptual alignment. Furthermore, by ablating the number of fine-tuning steps in Section A.2, we show that overfitting to perceptual judgments may also harm downstream performance.

Societal impacts. Beyond our findings, there are other possibilities for harm outside the scope of the type of perceptual annotations we have studied: Human preferences may reflect unwanted biases. A long-standing problem in both language and vision is that biases (e.g. gender, racial) reflected in Internet language/images are inherited by large models, and reflected in their embeddings. One can imagine similar phenomena to happen in visual preferences [29]. Humans may disagree on a preference label, or even disagree with themselves if asked at different points in time. This may lead to noisy data if not filtered carefully, thus harming representations [54, 43]. Without sufficient demographic diversity in the annotator group, emerging biases may be reflected in the model. For example, some RLHF-trained language models have been shown to develop a bias towards the opinions of high-income, liberal individuals over their non-RLHF counterparts [49].

Acknowledgements

We thank Tali Dekel and Richard Zhang for helpful discussions. This work was supported by the Sagol Weizmann-MIT Bridge Program and by a Packard Fellowship and a Sloan Research Fellowship to P.I. and an NSF GRFP Fellowship to S.S.

10

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[4] 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.

[5] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22243–22255. Curran Associates, Inc., 2020.

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

[9] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[11] Li Ding, Jenny Zhang, Jeff Clune, Lee Spector, and Joel Lehman. Quality diversity through human feedback. *arXiv preprint arXiv:2310.12103*, 2023.

[12] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pp. 647–655. PMLR, 2014.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

11

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014.

[16] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models, 2023.

[17] Thomas FEL, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9432–9446. Curran Associates, Inc., 2022.

[18] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768. Curran Associates, Inc., 2023.

[19] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

[20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[21] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5332–5340, 2019.

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[23] Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multi-dimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, October 2020. doi: 10.1038/s41562-020-00951-3.

[24] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, feb 2023. ISSN 2050-084X. doi: 10.7554/eLife.82580.

[25] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021.

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text 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. 4904–4916. PMLR, 18–24 Jul 2021.

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

[28] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

12

[29] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf.

[30] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023.

[31] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022.

[32] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[33] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

[34] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[35] Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, and Liwei Wang. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7212–7220, 2019.

[36] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation, 2022.

[38] Yiqing Liang, Eliot Laidlaw, Alexander Meyerowitz, Srinath Sridhar, and James Tompkin. Semantic attention flow fields for monocular dynamic scene decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21797–21806, 2023.

[39] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.

[40] Kelly S. Mix. Similarity and numerical equivalence: Appearances count. *Cognitive Development*, 14(2):269–297, 1999. ISSN 0885-2014. doi: https://doi.org/10.1016/S0885-2014(99)00005-2.

[41] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *11th International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, Mai 01-05, 2023*. OpenReview.net, 2023.

[42] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50978–51007. Curran Associates, Inc., 2023.

[43] Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024.

13

[44] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[45] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3147–3157, 2023.

[46] 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.

[47] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[49] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

[50] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

[51] Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. Materialistic: Selecting similar materials in images. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.

[52] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.

[53] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

[54] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

[55] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *ArXiv*, abs/2312.17742, 2023.

[56] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, June 2024.

[57] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. *arXiv preprint arXiv:2403.14548*, 2024.

14

[58] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2096–2105. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00200.

[59] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. 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. 12310–12320. PMLR, 18–24 Jul 2021.

[60] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[61] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[62] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023.

15

Appendix

In the Appendix, Section A contains additional experiments and ablations, Section B contains qualitative results and examples from different perceptual datasets, and Section C contains methodological/implementation details.

A Experiments

A.1 Classification with the VTAB Benchmark

We evaluate on the Visual Task Adaptation Benchmark (VTAB) [60], a comprehensive set of nineteen classification datasets that is commonly used to study the transferability of representations. VTAB covers a broad set of domains beyond classic natural image datasets, including medical images, scene understanding, and fine-grained classification. Several of the datasets are drawn from domains that are not included in the NIGHTS dataset.

In Tables 5a-b, we show classification results on all VTAB datasets for each backbone, and its human-aligned version (denoted with -HA). The VTAB datasets are divided into three categories: Natural, Specialized, and Structured. Natural datasets include standard vision benchmarks; specialized datasets include benchmarks captured through specialized imaging equipment, such as satellites and microscopes; structured datasets include benchmarks that focus on structure and layout, with both naturally-captured and simulated images.

We find that human-aligned models largely lead to worse performance on standard natural tasks. Results on specialized tasks are mixed; for most datasets, human alignment helps or hurts by a marginal amount. Results on structured datasets (Table 6) are also mixed, however on particular datasets such as sNORB, dSprites-Orientation, and dSprites-Location, alignment improves performance across backbones. We remark, however, that these structured datasets may have limited evaluation utility. The Clevr-Count task, for instance, has a numerical ground truth (i.e. the number of objects in an image) in which the distance from the correct answer is meaningful; thus it may be ill-suited to classification, and better suited to continuous evaluations such as RMSE or MAE, which we report in Section 4.3.

| | **Natural** | | | | | | | **Specialized** | | | |
Model	Caltech101	CIFAR-100	DTD	Flowers102	Pets	Sun397	SVHN	Camelyon	EuroSAT	Resisc45	Retinopathy
DINO	**95.3**	**83.0**	**78.7**	**96.7**	**93.1**	**74.8**	70.8	**85.8**	97.2	**93.8**	**77.9**
DINO-HA	93.6	80.8	72.6	93.2	91.0	69.7	**73.2**	83.7	**97.3**	93.2	76.9
DINOv2	95.5	**88.9**†	**82.8**	**99.7**†	**95.9**†	**81.9**	61.2	**86.7**†	96.4	92.7	**78.6**
DINOv2-HA	**96.2**	87.7	80.2	97.9	89.7	77.0	**63.3**	84.6	**97.0**	**93.5**	77.9
CLIP	95.4	**76.5**	**76.6**	**95.5**	**87.6**	**80.6**	65.4	83.7	95.3	92.9	75.8
CLIP-HA	**95.7**	76.4	76.0	93.4	85.5	78.1	**69.2**	**84.0**	**95.9**	**93.0**	**76.1**
OpenCLIP	**97.3**	**82.4**	**81.6**	**97.0**	**90.4**	**82.3**	**73.1**	82.3	96.4	83.3	75.9
OpenCLIP-HA	96.0	75.8	77.3	93.5	83.7	76.6	70.9	**82.9**	**96.5**	**92.6**	**76.2**
SynCLR	**95.5**	**74.3**	**80.5**	**98.8**	**92.9**	**81.4**	56.2	**86.0**	96.6	**94.0**	**78.8**†
SynCLR-HA	95.0	69.7	77.8	95.9	90.5	75.7	**70.9**	85.1	**96.8**	93.4	78.0
Ensemble	**97.4**†	**86.8**	**84.3**†	**98.8**	**94.4**	**83.9**†	80.1	**85.7**	**98.0**†	**96.1**†	**78.7**
Ensemble-HA	97.3	85.0	81.4	96.4	91.9	82.5	**82.2**†	83.5	97.9	**96.1**†	78.1

Table 5: **Performance on VTAB natural subset (left) and specialized subset (right).**

A.2 Additional dataset ablations

In Section 4.5 we show that tuning on NIGHTS leads to larger performance improvements on object counting and instance retrieval than training on other triplet datasets. Here, we show that this finding is consistent across dense prediction tasks as well. We evaluate models trained on NIGHTS, BAPPS, THINGS, and ImageNet – as described in section 4.5 – on semantic segmentation and depth estimation. As shown in Fig. 9- 10, training on NIGHTS outperforms training on all other datasets.

16

	Structured							
Model	Clevr-Count	Clevr-Dist	DMLab	dSpr-Loc	dSpr-Orti	KITTI-Dist	sNORB-Azim	sNORB-Elev
DINO	**82.2**	**58.1**	**53.6**	63.8	57.1	**72.6**	61.8	**56.5**
DINO-HA	79.2	57.2	51.1	**81.5**	**66.6**	66.2	**65.8**	55.3
DINOv2	**87.1**	54.6	**57.7**	59.7	68.2	62.3	55.3	52.8
DINOv2-HA	83.6	**59.0**	53.7	**79.2**	**74.1**	**69.1**	**64.7**	**55.9**
CLIP	70.6	55.0	50.2	67.6	55.6	**65.3**	37.5	47.0
CLIP-HA	**72.9**	**57.8**	**51.4**	**80.2**	**60.3**	64.6	**42.6**	**47.2**
OpenCLIP	**78.5**	53.8	**52.0**	63.3	55.0	**67.9**	38.0	**50.4**
OpenCLIP-HA	76.7	**57.6**	51.5	**84.9**	**64.1**	63.3	**46.7**	50.1
SynCLR	**90.9**†	**63.8**†	**57.3**	68.1	55.2	**76.9**†	52.1	**60.5**
SynCLR-HA	81.7	61.5	52.9	**86.1**	**64.3**	63.6	**66.6**	58.5
Ensemble	**89.2**	63.5	**58.4**†	72.5	67.3	**76.5**	76.9	**65.5**†
Ensemble-HA	85.5	**63.8**†	57.1	**91.8**†	**76.4**†	70.0	**81.0**†	62.2

Table 6: **Performance on VTAB structured subset.**

We additionally run this ablation on a subset of VTAB (see Fig. 11a). On these classification datasets, base models perform best; the exception is sNORB (pose prediction) for which NIGHTS is best. Amongst perceptual datasets, NIGHTS is sometimes outperformed by BAPPS/ImageNet. This result is consistent with our findings in Section A.1, that tuning on NIGHTS often fails to improve classification performance.

Finally, we ablate the *strength* of alignment – i.e. the training loss when fine-tuning on different perceptual datasets. In Fig. 11b we show the downstream performance on DeepFashion2 when fine-tuning for increasing numbers of steps on different datasets. Tuning on NIGHTS outperforms other datasets over the full training trajectory. Moreover, while performance rises significantly with a small amount of alignment to NIGHTS, it trends down after >1000 steps. This indicates that a small amount of alignment is helpful, however overfitting may harm performance.
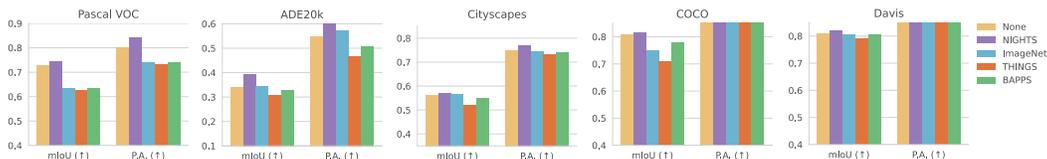
Figure 9: **Dataset ablations on semantic segmentation.** Following the same procedure as in section 4.5 and the segmentation training in section 4.1 of the main paper, we ablate low/mid/high-level similarity. We find that tuning on mid-level similarity with NIGHTS provides the most improvement, with many other cases degrading the base DINO backbone.
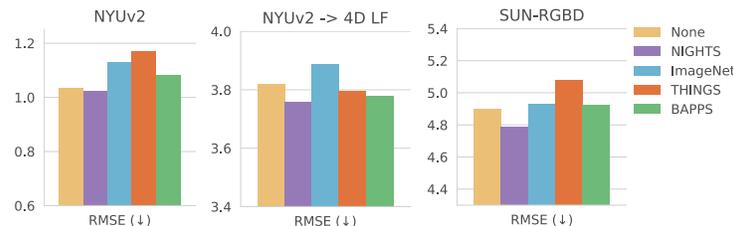
Figure 10: **Dataset ablations on depth estimation.** Following the ablation and depth training setups described in the main paper, we evaluate low/mid/high-level similarity. We find that tuning on mid-level similarity with NIGHTS results in the lowest error (RMSE), and that the other cases often result in worse performance than base DINO.

17
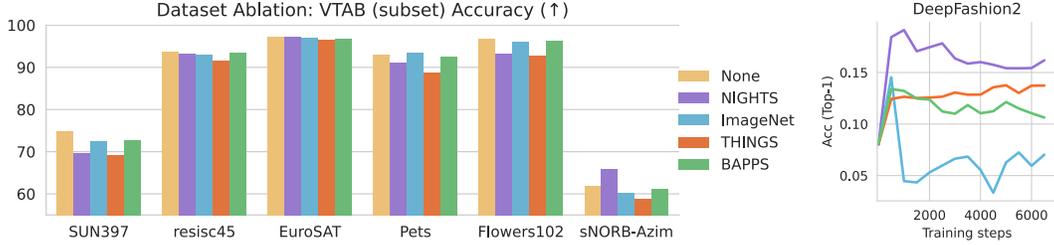
Figure 11: **(left) Dataset ablations on a subset of VTAB.** Following the same procedure as in section 4.5 and the segmentation training in section 4.1 of the main paper, we ablate low/mid/high-level similarity. We find that all perceptual datasets degrade the DINO backbone, with the exception of sNORB, for which NIGHTS is best. We also note that training on BAPPS or ImageNet often better preserves base model performance than NIGHTS. **(right) Training-step ablations on DeepFashion2.** We ablate low/mid/high-level similarity and test models finetuned for different numbers of steps on instance retrieval. We find that a small amount of finetuning on NIGHTS is best, after which performances declines (though remains above the base model). *Note that the legend applies to both the left and right figure.*

A.3 Additional RAG results

We replicate our RAG experiment on OpenFlamingo in Section 4.2 on IDEFICS2, a recently released 8B multimodal model achieving state-of-the-art results across several benchmarks [56]. As shown in Fig. 12, across the same four classification datasets, performance consistently improved when using NIGHTS-tuned models in the RAG pipeline. These results validate our RAG results on OpenFlamingo, suggesting that performance gains from perceptual alignment are not specific to a particular VLM.
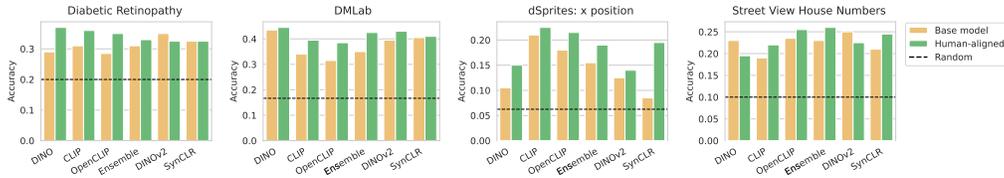
Figure 12: **Retrieval-augmented generation results on image classification for four different VTAB datasets.** We evaluate the IDEFICS2 model, an 8B-parameter multimodal model released more recently than Open-Flamingo and capable of in-context learning. We follow the same experimental setup in retrieving examples as detailed in 4.2.

A.4 Perceptual alignment for CNNs

Our main experiments assess the effects of perceptual alignment for an array of ViT models, as these are state-of-the-art pretrained vision representations. Of further interest, however, is whether the same effects are observed for pre-trained CNNs as well.

We assess the effects of perceptual alignment with NIGHTS on two popular CNNs: ResNet50 and ConvNeXt-B, using the same loss described in Section 3.2. We evaluate on counting and instance retrieval tasks; our results are shown in Tables 7a-b. Similarly to ViTs, human-aligned CNNs outperform their pretrained counterparts across both tasks. Note that we report results for both ConvNeXt and ResNet, however acknowledge that the accuracies for ResNet are likely too small to draw conclusions from.

Due to a lack of open-source LoRA implementations for CNNs, we instead train MLPs on top of the respective final layers. Previous work [18] indicated that training MLPs with NIGHTS can still affect representations, but may be less effective than full fine-tuning. Nonetheless, we still see downstream improvements with this method.

18

Model	Clevr-Count		Model	DeepFashion2		
	RMSE	MAE		Top-1	Top-3	Top-5
ConvNeXt	2.045	1.522	ConvNeXt	2.12	3.56	4.59
ConvNeXt-HA	**1.631**	**1.193**	ConvNeXt-HA	**2.81**	**4.8**	**5.98**
ResNet	3.140	2.551	ResNet	0.018	**0.12**	0.14
ResNet-HA	**1.729**	**1.282**	ResNet-HA	0.018	0.074	**0.17**

Table 7: Human-aligned vs. pretrained CNNs, evaluated on counting and instance-retrieval. In most cases, human-aligned CNNs perform better.

B   Qualitative examples

B.1   Additional visualizations

See Figures 13 and 14 for additional examples of instance retrieval on DeepFashion2 and object counting on Clevr-count.
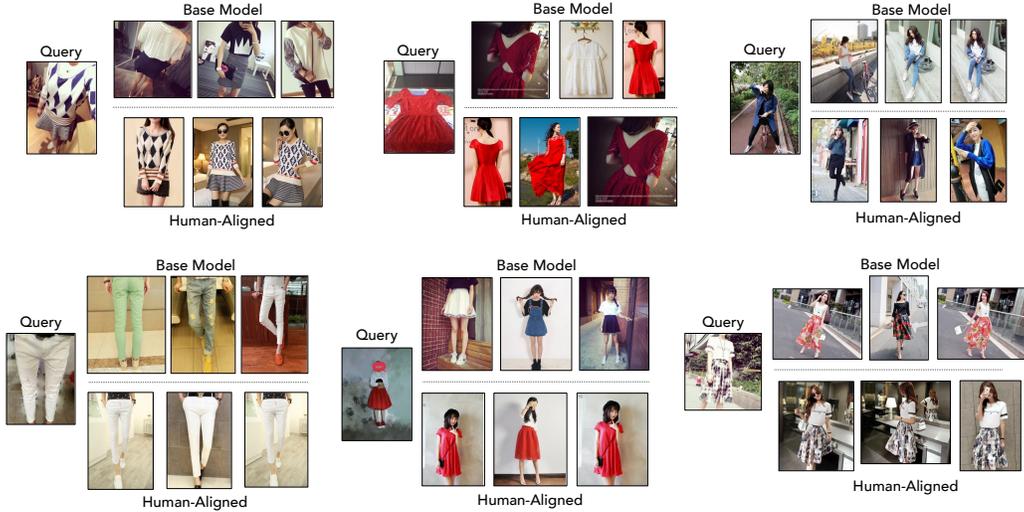
Figure 13: Additional visualizations for top-3 retrieved examples for a given query image in DeepFashion2.
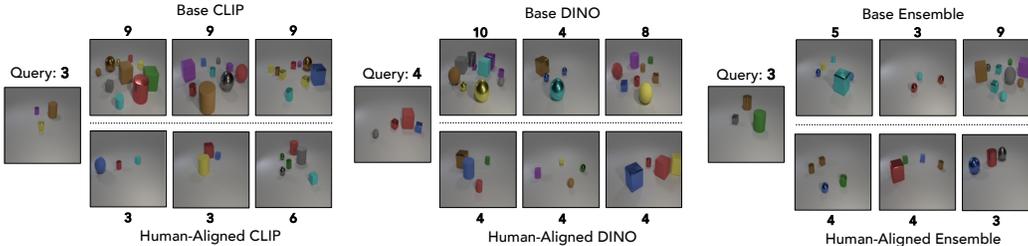
Figure 14: Additional Clevr-count examples comparing base and human-model nearest-neighbor image retrievals.

B.2   Dataset examples

See Figure 15 for examples of the triplet datasets we train on in this paper. In each triplet, the reference image (middle) and outlined image are labeled as the similar pair. Perceptually-aligned models in this paper are tuned on NIGHTS triplets (top row), whose image variations encompass a variety of mid-level perceptual attributes including object count, identity, layout, subject pose, and color. In contrast, the datasets studied in our ablation encompass differing image attributes: BAPPS

19
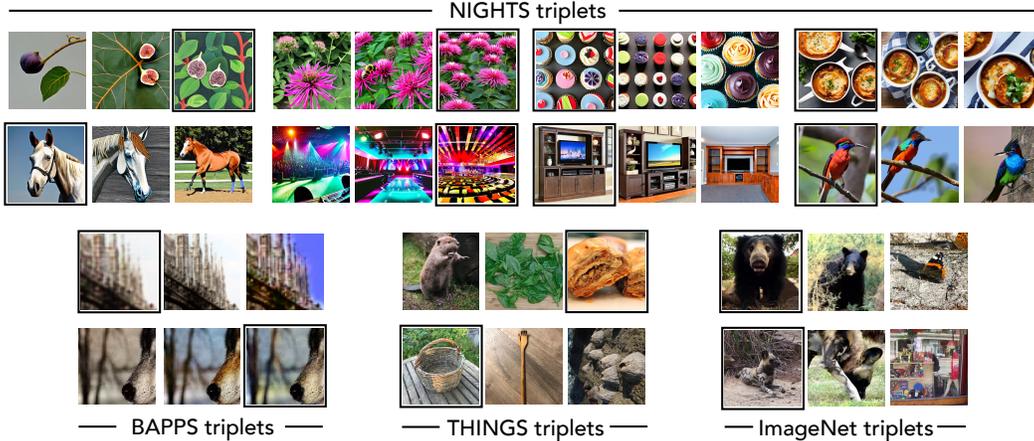
Figure 15: Examples of triplets from the NIGHTS, BAPPS, THINGS, and ImageNet datasets, with the bordered images labeled as more similar to the reference (middle image in each triplet).

focuses on image patches with low-level distortions such as CNN artifacts and color jitter, THINGS encodes similarity in concept space, and our constructed ImageNet triplets outline class boundaries.

C Implementation Details

C.1 Training human-aligned models

The human-aligned models studied in this paper were fine-tuned with LoRA parameters: $r = 16$, $\alpha = 0.5$, $p = 0.0$. All contrastive training (for image- and patch-level objectives) are trained with an Adam optimizer ($lr = 0.0003$), batch size of 16, and a hinge loss margin of $m = 0.05$ as detailed in [18]. We train all models for 8 epochs and evaluate on the checkpoint with the lowest validation loss. Train/val/test splits on NIGHTS, BAPPS, and THINGS were used as-provided in the dataset.

C.1.1 Dense prediction heads

Each semantic segmentation head is a single linear layer mapping frozen DINO/DINOv2 patch tokens to a segmentation map. We train with the Adam optimizer ($\alpha = 0.0003$) for 10 epochs with a batch size of 16 and 4 workers. The semantic segmentation head is optimized with a Jaccard Index loss (equivalently, mean Intersection-over-Union) defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$.

Dense prediction heads are defined as a single linear layer mapping frozen DINO/DINOv2 patch tokens to a spatial map, whose values are then categorized into 256 bins. We train with the Adam optimizer ($\alpha = 0.0003$) for 10 epochs with a batch size of 128 and 12 workers. The scale-invariant log loss for prediction A and target B is defined as:

$$\mathcal{L}(A, B) = \frac{1}{N} \sum_i d_i^2 + 0.15 \cdot \frac{1}{N^2} (\sum_i d_i)^2$$

where $d_i = \log(A + \epsilon) - \log(B + \epsilon)$, $\epsilon = 0.001$ to avoid gradient issues, and N is the number of valid pixels.

All training and evaluation for dense prediction tasks is done on a single NVIDIA Titan RTX GPU.

C.2 Retrieval-augmented generation

We follow the evaluation protocol in the codebase provided by [2], constructing a prompt containing three image/text examples, a query image, and a candidate class. We then average the model's output logits, and repeat for each candidate class in a dataset to form a softmaxed array of log probabilities. OpenFlamingo's classification prediction is determined as the class with the highest corresponding log probability, while classifications from IDEFICS2 are taken purely from a multiple-choice text response. Error bars were calculated as 95% confidence intervals over five trials with 1,000 randomly-sampled queries each. All evaluation for RAG tasks is done on a single NVIDIA Titan RTX GPU.

20

C.2.1 Object counting

Our k-Nearest Neighbor evaluations reported in the Counting section are run over $k \in \{1, 3, 5, 10\}$ for the training set, and the argmax k with highest classification accuracy is used to report the final performance on the test set. We use the train and test sets as provided for the FSC147 and CARPK datasets, and construct a random 80:20 data split for the Clevr-Count dataset. All evaluation for this section is done on a single NVIDIA Titan RTX GPU.

C.3 Instance retrieval

We evaluate backbones on instance retrieval by extracting features (CLIP/OpenCLIP: embedding, DINO/DINOv2/SynCLR: CLS token) from the gallery and query sets. For every query image, we compute the cosine similarity with every image in the gallery. We then calculate top-k accuracy as the number of queries for which a correct gallery pair image was in the top k retrievals.

C.4 Dataset ablations

We train models for the dataset ablation section with identical hyperparameters as the human-aligned models (details in Section C.1 and on the same amount of data as the NIGHTS training set (13,900 triplets). We randomly select these triplets without replacement from the BAPPS and THINGS training sets, and construct the ImageNet triplets by randomly sampling two images from one class and a third image from a different class.

C.5 VTAB classification

To evaluate models on VTAB, we follow standard procedure and train a linear classifier on top of the frozen representations using multinomial logistic regression, with the sci-kit learn implementation [42]. For each classifier we perform a hyperparameter search over the regularization strength using 10-fold cross validation over the training set, and report performance on the validation set. We search over the values $c = \{1e0, 1e1, 1e2, 1e3, 1e4, 1e5, 1e6\}$ where c is the inverse regularization strength.

C.6 Additional compute details

See subsections above for experiment-specific compute details. This full research project required additional compute for experiments and results that are not included in this paper; these computations were also done on single NVIDIA Titan RTX, GeForce 2080, GeForce 3090, and V100 GPUs.

C.7 Licenses for existing assets

The datasets and models we use are released under the following licenses:

Dataset	License
NIGHTS	MIT
VTAB	Apache 2.0
BAPPS	BSD 2-Clause
THINGS	CC0 1.0 Universal
ImageNet	CC BY-NC
FSC147	Apache 2.0
CARPK	CC0 1.0 Universal
Pascal VOC	CC BY 2.0
ADE20k	BSD 3-Clause
Cityscapes	CC BY-NC
COCO	CC BY 4.0
DAVIS2017	CC BY 4.0
NYUv2	MIT
4D Light Field	CC BY-NC-SA 4.0
SUN-RGBD	MIT
DeepFashion	CC BY-NC

Model	License
DINO	Apache 2.0
CLIP	MIT
OpenCLIP	MIT
DreamSim	MIT
SynCLR	Apache 2.0

21

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction are written with the paper results and scope fully in mind.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe limitations in the "Discussions" and Supplementary Material sections of our paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

22

Answer: [NA]

Justification: Our paper contains no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all experimental settings and hyperparameters in the Supplementary Material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

23

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a zip of the code that produced our experimental results, as well as our project github at: https://github.com/ssundaram21/dreamsim.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide these details in the Supplementary Material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide 95% confidence interval bars across five seeds for our experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

24

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on compute resources in the Supplementary Material.

   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics, and have determined that the research conducted conforms to the code of ethics.

   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We detail the broader impacts of our work in the Discussions section.

    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

25

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks, as we do not have new pretrained models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original papers whose datasets and models we use, and provide license information in the Supplementary Material.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

26

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

27

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

28