# NATURE VS. NURTURE: LIMITS OF REPRESENTATION UNIVERSALITY REVEAL ARCHITECTURAL AND TRAINING DIFFERENCES IN IMAGE MODELS

**Fredo Guan**
Arizona State University
fyguan@asu.edu

**Shreyan Banerjee**
University of Michigan - Ann Arbor
shreyanb@umich.edu

**Aditya Bajoria**
University of Delaware

**Akarsh Shetty**
Foothill College
shettyakarsh@student.deanza.edu

**Cole Blondin**
Algoverse

## ABSTRACT

Recent works have hypothesized and demonstrated the general inter-compatibility and translatability of vector embeddings in learned embedding spaces. We study the limits of representation compatibility across image models encompassing a variety of architectures and training paradigms. We train adapters, comprised of linear encoders and decoders, to translate between all pairs of embedding spaces using a shared latent space. We then evaluate cross-model and cross-domain compatibility using classification, retrieval, and embedding-based benchmarks, revealing limits in compatibility across learned embedding spaces. Further analysis of pairwise translation performance reveals phylogenetic patterns that reflect the fundamental differences in model architecture and training.

## 1 INTRODUCTION AND MOTIVATING WORK

Over the past decade, improved techniques and increased scale have enabled deep neural networks to learn strong representations of their input modality. Recent work around universality and interpretability have led to two hypotheses, the Platonic Representation Hypothesis (Huh et al., 2024) and the Linear Representation Hypothesis (Park et al., 2024). These suggest that different models arrive at similar representations and that models represent concepts and relationships linearly. Growing bodies of empirical evidence provide supporting evidence and identify limitations for both hypotheses. We aim to assess the implications of the hypothesized compatibility on downstream tasks at scale by training configurable all-to-all adapters to translate between embedding spaces and evaluating on a variety of downstream tasks.

### 1.1 UNIVERSALITY OF LEARNED REPRESENTATIONS

The Platonic Representation Hypothesis (PRH) posits that different neural networks, trained on different aspects of a shared reality, converge towards a shared underlying representation (Huh et al., 2024). Many works provide evidence supporting universality between different models of the same modality, models of different modalities, and the brains of different humans (Li et al., 2015; Jha et al., 2025; Huh et al., 2024; Chen & Bonner, 2024; Ryskina et al., 2025; Marcos-Manchón & Fuentemilla, 2025; Hosseini et al., 2024; Gauthaman et al., 2025). Empirical evidence also shows that models which exhibit stronger performance and generality also exhibit stronger alignment with other models, both within and across modalities (Huh et al., 2024). Analysis of CNNs concludes that models trained on the same data and task learn similar filters (Li et al., 2015; Babaiee et al., 2025; Gavrikov & Keuper, 2022). Furthermore, controlled transformations of training images result in those same transformations appearing in the learned filters (Lenc & Vedaldi, 2015). Studies on depthwise-separable convolutions also find 8 distinct filters common across different models with varying filter sizes and architectures (Babaiee et al., 2024; 2025).

Several techniques have been introduced to translate between latent and embedding spaces, providing empirical (Moschella et al., 2022; Puri et al., 2025) and theoretical (Maystre et al., 2025) evidence that the learned geometry of different models is sufficiently similar to allow translation between different models. Additional experiments in model stitching demonstrate that stitched models still recover most of the performance of the original pretrained models, showing high alignment between distinct models (Lenc & Vedaldi, 2015; Bansal et al., 2021). A subsequent work introduces vec2vec, which further exploits this principle to align and translate between distinct models trained on the same task and modality using only an unpaired corpus of embeddings for each model, enabling text recovery from a vector database of text embeddings without access to the original embedding model (Jha et al., 2025). Using adversarial, reconstruction, and pairwise distance loss terms, the method shows that embeddings from different models can be brought into a shared latent space, suggesting that the distribution and relative geometry of embeddings is sufficient to enable translation.

Taken together, these works suggest that image and text models learn representations that include an underlying geometry common across architectures and modalities, while also exhibiting representational similarities with the human brain. They also hint at or directly isolate features and geometric structures common across different models and demonstrate translation amounts to aligning existing structure rather than constructing new representations.

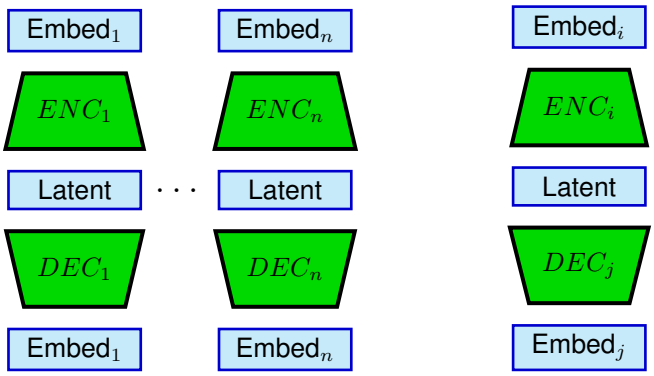
### 1.2 The Linear Representation Hypotheses

The Platonic Representation Hypothesis describes this global convergence behavior but does not specify how information is organized within representation spaces. The Linear Representation Hypothesis (LRH) addresses this, as it posits that many semantic and relational attributes are encoded along approximately linear directions in representation space, allowing them to be detected, composed, and steered using only vectors (Park et al., 2024). Evidence of this structure can be seen in language representations. In word embeddings, relations such as gender and tense correspond to approximately consistent vector offsets, indicating that these attributes align with directions in the embedding space (Mikolov et al., 2013). Structural information follows a similar pattern: a single linear transformation can map contextual representations into a space where distances reflect underlying hierarchical relationships, showing that relational structure is linearly accessible (Hewitt & Manning, 2019). Sparse autoencoders (SAEs) further exploit this by learning thousands to millions of linear features in a self-supervised manner, which can then be automatically interpreted using LLMs and used to probe or steer model internals (Cunningham et al., 2023; Paulo et al., 2025; Templeton et al., 2024; Thasarathan et al., 2025).

Additional evidence links the LRH and PRH together: most attempts to translate between embedding spaces use linear transformations (Moschella et al., 2022; Puri et al., 2025; Maystre et al., 2025; Lenc & Vedaldi, 2015; Bansal et al., 2021). Some of these works also identify that preserved dot products in different embedding spaces are sufficient for interoperability and translatability. This motivates our approach, in which we use linear mappings to align many learned representations and study their compatibility without modifying the underlying pretrained models.

### 1.3 Decoupled Representation and Task Models

Many techniques have been developed to learn image representations. These include supervised classification, various forms of language supervision, various forms of self supervision, and combinations of these techniques (see Sec. 2.1 for examples). Many recent approaches have shifted from training task-specific models end-to-end to training vision foundation models, which decouple the tasks of representing input data and using a learned representation to accomplish a downstream task (Radford et al., 2021; Fang et al., 2024; Bolya et al., 2025; Siméoni et al., 2025). The foundation model is first pretrained at scale, then used as a backbone for different downstream tasks. Often, the backbone is not trained for downstream tasks; representations are simple extracted and used to train a task-specific head without end-to-end post-training. The bulk of training compute is used to train the vision backbone, while task-specific training is done with less compute while sharing the same backbone weights across different tasks.




(a) Encoder and decoder for each model using a shared latent space

(b) Adapting from model $i$ to model $j$

Figure 1: We train our adapters to translate between the embedding spaces of any pair of supported models using a shared latent space. There are no nonlinearities, resulting in a linear transformation that translates between each pair of models.

### 1.4 Motivating Translation as a Proxy for Compatibility

Given these recent works studying the PRH and the LRH at a representation level, most existing works investigating model compatibility use similarity, alignment, or embedding-based transfer metrics. We instead aim to investigate the effect of translated embeddings on downstream tasks. This is further motivated by evidence that suggests the task heads of classification models limit universality and do not converge across different models (Klabunde et al., 2025). Additionally, works investigating downstream performance rarely extend beyond 2 models. We aim to systematically align many models using adapters to and from a shared "lingua franca" latent embedding space, then assess the effect of a using a translated non-native embedding with a pretrained task head. We view the ability to translate between many embedding spaces linearly using a shared embedding space as strong evidence of both the LRH and the PRH.

We also aim to characterize and quantify the limits of such a translation technique. Prior work suggests that models trained on different modalities using different training techniques have been observed to exhibit distinct alignment patterns when evaluated across different layers and different brain regions (Marcos-Manchón & Fuentemilla, 2025). It has also long been known that the architecture of a neural network affects its robustness, such as CNNs being more texture biased while ViTs are more shape biased (Dehghani et al., 2023). We suspect that these inherent differences, stemming from training technique, input data, and architecture, will hinder model compatibility. Thus, we fundamentally aim to analyze the impact of a pretrained model's architecture and training process, or its **nature** and **nurture**, on its compatibility with other models.

## 2 Methods

We first select several representative pretrained vision encoders. Additionally, to study cross-modality representation compatibility, we include both an image-aware CLIP text tower and an image-blind text embedding model. We then extract embeddings from each model and train adapters that use a shared latent space for each model.

### 2.1 Model Selection

In order to study the impact of architecture, training data, training technique, and training modality on embedding translatability, we first select a diverse sample of $n = 18$ models that are representative of the image representation landscape. We include 4 CNNs trained on ImageNet-1k: ResNet152 (He et al., 2016), RegNetY-8.0GF (Radosavovic et al., 2020), ConvNeXt-B (Liu et al., 2022), and MobileNetv4-Conv-L (Qin et al., 2025). We include 7 isometric vision transformers (ViT) (Dosovitskiy et al., 2021) of varying sizes trained using various techniques and datasets: a ViT-B trained us-

ing AugReg on ImageNet-21k and finetuned on ImageNet-1k (Steiner et al., 2022), a ViT-B trained using CLIP and finetuned on ImageNet-1k (Radford et al., 2021), Perception Encoder Core (PE-Core), a modified ViT-G trained using CLIP (Bolya et al., 2025), EVA-02-Base, a modified ViT-B distilled from a CLIP-trained teacher using MIM and finetuned on ImageNet-1k (Fang et al., 2024), BEiT-3, a modified ViT-G trained using MLM and MIM and finetuned on ImageNet-1k (Wang et al., 2023), AIMv2-L, a modified ViT-L trained using an autoregressive head for pixel reconstruction and captioning (Fini et al., 2025), and a modified ViT-L distilled from a ViT-7B teacher, with both trained using self-supervised learning with DINOv3 (Siméoni et al., 2025). We include Swin-L, a hierarchical transformer trained on ImageNet-21k and finetuned on ImageNet-1k (Liu et al., 2021), MaxViT-B, a hybrid model trained on ImageNet-21k and finetuned on ImageNet-1k (Tu et al., 2022), MobileNetv4-Hybrid-L, a hybrid model trained on ImageNet-1k (Qin et al., 2025), CAFormer-B36, a hybrid model, and ConvFormer-B36, a modified CAFormer-B36 with the attention layers replaced with MBConv layers, both trained on ImageNet-22k and finetuned on ImageNet-1k and from the MetaFormer family of models (Yu et al., 2024). We also include 2 text models: the image-aware text encoder paired with PE-Core (Bolya et al., 2025) and the image-blind text embedding model Qwen3-Embedding-4B (Zhang et al., 2025). We use implementations and weights from the timm library (Wightman, 2019) for all image models, the OpenCLIP library for the PE-Core text encoder (Ilharco et al., 2021), and the HuggingFace Transformers library for Qwen3-Embedding-4B (Wolf et al., 2020).

ResNet152, RegNetY-8.0GF, and ConvNeXt-B were chosen as representative supervised CNN baselines. MobileNetv4-Conv-L and MobileNetv4-Hybrid-L were chosen as representative lightweight models. All other ImageNet-trained models were chosen due to their ImageNet-1k classification performance in excess of 84%. The PE-Core models, AIMv2, and the DINOv3-trained ViT-L were chosen due to their unique training techniques. Qwen3-Embedding-4B was chosen due to its state-of-the-art performance for a text embedding model of its size as of writing.

### 2.2 Adapter Architecture

Following vec2vec, we use an encoder $ENC_i : \mathbb{R}^{d_i} \mapsto \mathbb{R}^{d_h}$ and a decoder $DEC_i : \mathbb{R}^{d_h} \mapsto \mathbb{R}^{d_i}$ for each supported model $i \in 1..n$ with dimension $d_i$ (Jha et al., 2025). We remove the shared MLP used by vec2vec and use only fully connected layers with no nonlinearity (Fig. 1a). Thus, every pairwise relationship between embedding spaces is modeled using only linear projection, allowing us to evaluate universality under both the PRH and the LRH. Each encoder takes in a vector from its corresponding model and projects into a shared latent space of a set dimension $d_h$. Each decoder then takes in a $d_h$-dimension vector in the shared latent space and projects back into the embedding space of its corresponding model. To translate an embedding from a supported model $a$ to the embedding space of another supported model $b$, the encoder of the input model projects the input embedding into the latent space and the decoder of the output model projects the latent embedding into the output space of the output model (Fig. 1b). These are trained to project between each pair of models using the training protocol and loss function described in 2.3.

### 2.3 Adapter training

We use the CC-12M dataset (Changpinyo et al., 2021), comprised of 12 million text-image pairs, to train our adapters. We first compute image embeddings for all images using each image model and text embeddings for all paired text using each text model. We then instantiate and train our adapters. Unless otherwise specified, we use a latent dimension $d_h = 4096$ for all experiments. We also use adversarial training, with an MLP discriminator $DC : \mathbb{R}^{d_h} \mapsto \Delta^{n-1}$ that takes in a 4096-dim latent, projects through 2 GELU-activated 512-dim hidden layers followed by a softmax-activated output layer with $n = 18$ outputs, where $n$ is the number of models supported by the adapter.

We use data augmentation to regularize training. Specifically, we add gaussian noise to an input $\mathbf{x}$, where $\sigma(\mathbf{x}) \in \mathbb{R}^d$ is the per-dimension standard deviation of $\mathbf{x}$, which we scale by a constant factor $\alpha$ as shown below:

$$AUG(\mathbf{x}, \alpha) = \mathbf{x} + \alpha\sigma(\mathbf{x}) \odot \mathcal{N}(0, 1)$$

During training, we start with a batch of embeds $\mathbf{x}_i$ for each model $i \in 1..n$. Noise strength was determined experimentally and disabled during inference. We compute a forward pass using the adapters as follows:

$$\tilde{\mathbf{x}}_i = AUG(\mathbf{x}_i, 0.8), \forall i \in 1..n$$
$$\mathbf{h}_i = ENC_i(\tilde{\mathbf{x}}_i), \forall i \in 1..n$$
$$\tilde{\mathbf{h}}_i = AUG(\mathbf{h}_i, 0.6), \forall i \in 1..n$$
$$\mathbf{y}_{j,i} = DEC_j(\tilde{\mathbf{h}}_i), \forall i, j \in 1..n$$
$$\mathbf{z}_{i,j} = DEC_i(ENC_j(\mathbf{y}_{j,i})), \forall i, j \in 1..n$$

For training the adapters, we adopt a modified version of the loss used to train vec2vec. We keep their self reconstruction and cycle reconstruction losses. Since we work with $n$ models that would require $n$ embedding discriminators in addition to a latent discriminator, we modify their adversarial loss to only use a latent discriminator and avoid using any embedding discriminators. Since we use paired embeddings, we additionally introduce a pairwise reconstruction and latent similarity MSE loss terms. We intentionally formulate our loss with both adversarial and similarity losses on the latent space since we want to force the latent space to act as a lingua franca between models. Our final loss is an unweighted sum of the stated loss terms as formulated below, where $MSE$ is mean squared error loss and $CE$ is cross entropy loss with a single target probability for all terms:

$$\mathcal{L}_{selfMSE} = \frac{1}{n}\sum_{i=1}^{n} MSE(\tilde{\mathbf{x}}_i, \mathbf{y}_{i,i})$$
$$\mathcal{L}_{cycleMSE} = \frac{1}{n^2-n}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\delta_{ij})MSE(\tilde{\mathbf{x}}_i, \mathbf{z}_{i,j})$$
$$\mathcal{L}_{DCpredict} = \frac{1}{n}\sum_{i=i}^{n} CE(DC(\mathbf{h}_i), \frac{1}{n})$$
$$\mathcal{L}_{pairwiseMSE} = \frac{1}{n^2-n}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\delta_{ij})MSE(\tilde{\mathbf{x}}_j, \mathbf{y}_{j,i})$$
$$\mathcal{L}_{latentMSE} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} MSE(\mathbf{h}_i, \mathbf{h}_j),$$

For training the discriminator, we use a cross-entropy loss with the target set to the source model of the input latent. Here, $CE$ is cross entropy loss with the class target set as an index:

$$\mathcal{L}_{DCtrain} = \frac{1}{n}\sum_{i=1}^{n} CE(DC(\tilde{\mathbf{h}}_i), i)$$

We train for 10 epochs using a batch size of 256 and 4 gradient accumulation epochs, for an effective batch size of 1024. We use a cosine learning rate scheduler with 1 epoch of linear warmup for both the adapters and the discriminator. We use a maximum learning rate of $3 \times 10^{-5}$ for the adapters and a maximum learning rate of $1 \times 10^{-5}$ for the discriminator. We use the Adan optimizer for the adapters and the discriminator (Xie et al., 2024).

## 3 EXPERIMENTS

We view experiments on individual tasks as isolated measures of model representation compatibility, akin to feature engineering. Our correlation, classification, and retrieval evaluations all serve as individual metrics which we then ensemble into a dendrogram.

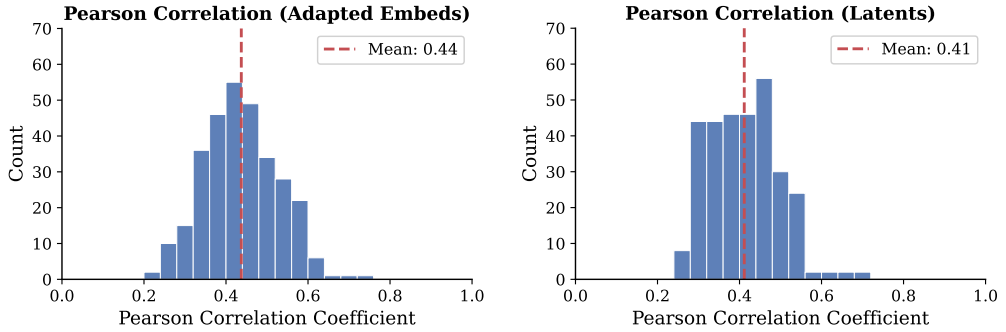



Figure 2: Histograms of off-diagonal Pearson correlation between adapted embeddings and original embeddings and between latents from different models.

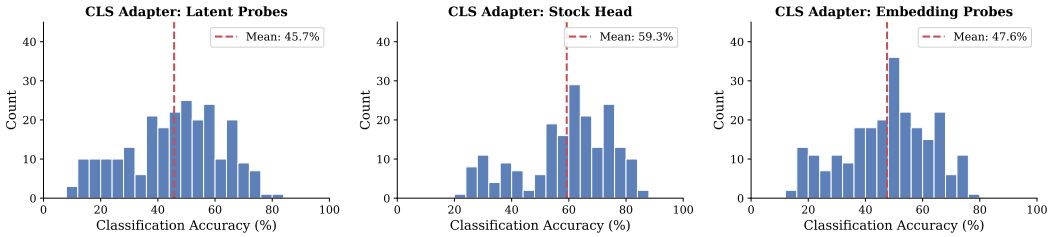



Figure 3: Histograms showing off-diagonal top-1 accuracy on ImageNet-1k using latent probes, stock classification heads, and embedding probes.

### 3.1 Correlation

We compute the Pearson correlation coefficient on the COCO Caption 2017 (Lin et al., 2015) test split between all pairs of latents $(corr(\mathbf{h}_i, \mathbf{h}_j), \forall i, j \in 1..n)$ and between all pairs of adapted embed and original embed $(corr(\mathbf{y}_{j,i}, \mathbf{x}_j), \forall i, j \in 1..n)$. We provide full histograms to visualize the off-diagonal correlation.

### 3.2 ImageNet Classification

We use the ImageNet (Deng et al., 2009) dataset to evaluate the performance of the adapters when transferring to a downstream task, specifically image classification. For this experiment, we exclude the two text models, train probes using the ILSVRC2012 training split, and evaluate using the ILSVRC2012 validation split (Russakovsky et al., 2015). For the classification head, we have explored 4 sources. First, 13 out of image models used are trained or finetuned for ImageNet-1k classification, allowing us to use the stock classification heads. Second, we can train a linear probe on the embeddings from each image model. Third, we can train a linear probe on the latents of each model by using the corresponding encoder to extract the latents. Fourth, we can train a linear probe on the latents of all models by extracting latents from all models into a combined training dataset.

Evaluating each collection of probes from the first 3 settings results in pairwise matrices, where each off-diagonal value represents transfer performance. For latent probes, the diagonal represents validation performance on the latents obtained from the training model. For the stock and trained embedding probes, the diagonal represents validation performance on the source model's embeddings adapted back into the source model's space (ie. we look at how well the adapter has learned the identity function). We provide histograms showing the off-diagonal classification performance. We include full results in the appendix (Fig. A.1).

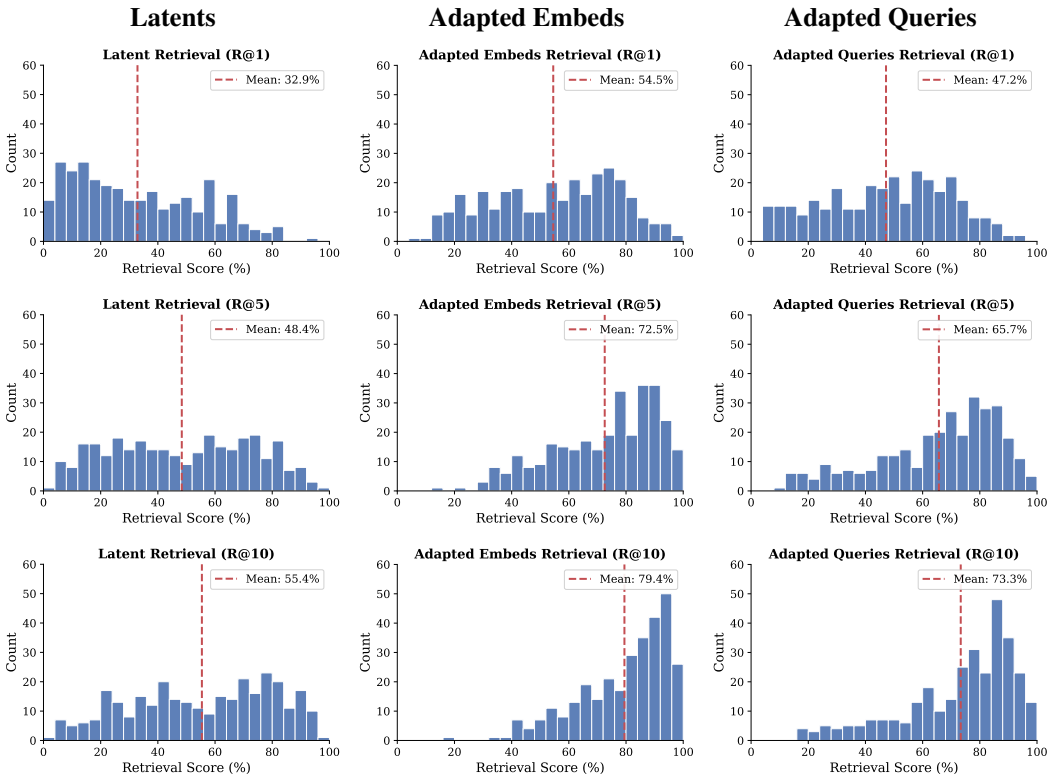



Figure 4: Histograms showing off-diagonal retrieval Recall@1, Recall@5, and Recall@10 on COCO Captions 2017 using latents, adapted embeds, and adapted queries.

### 3.3 Retrieval

We use the COCO Caption 2017 dataset (Lin et al., 2015) to evaluate the retrieval performance of different pairs of models. We first compute the image or text embeddings of the test split ($n = 5000$) for each model. We evaluate retrieval in 3 settings: latent retrieval, adapted query retrieval, and adapted embed retrieval. For latent retrieval, we compute latents for both the queries and the embeddings. For adapted query retrieval, we adapt the queries into the embedding space of the embeddings. For adapted embed retrieval, we adapt the embeddings into the embedding space of the queries. For all settings, we normalize all queries and embedding, then perform retrieval using cosine similarity across all pairs of models and report Recall@$k$ for $k \in \{1, 5, 10\}$. Retrieval can be image-to-image, text-to-image, image-to-text, or text-to-text, depending on the models used. We use histograms to visualize the off-diagonal correlation.

We observe high retrieval performance within modalities (image-to-image and text-to-text). In cross-modal retrieval, PE Core and the paired text model form the strongest pair, while other pairs follow in performance. Pairs involving Qwen3-Embedding-4B show weaker performance compared to pairs where the same image model is paired with PE Core's text model. We provide full results in the appendix (Sec. A.2).

### 3.4 Cluster Purity

We compute cluster purity using 2 metrics, kNN purity and silhouette score. We first compute latents, which are then reduced to 256-dimensional vectors using PCA to manage computational cost. For kNN purity, we run a kNN on the dimension-reduced latents of each model, then report the average proportion of latents in the identified neightborhood that came from the same model. Silhouette score measures the overlap of clusters as a whole. For both of these scores, higher values indicate more isolated latents and consequently a more distinct model representation. We report both of these scores for both COCO Caption and ImageNet-1k in Table 1.

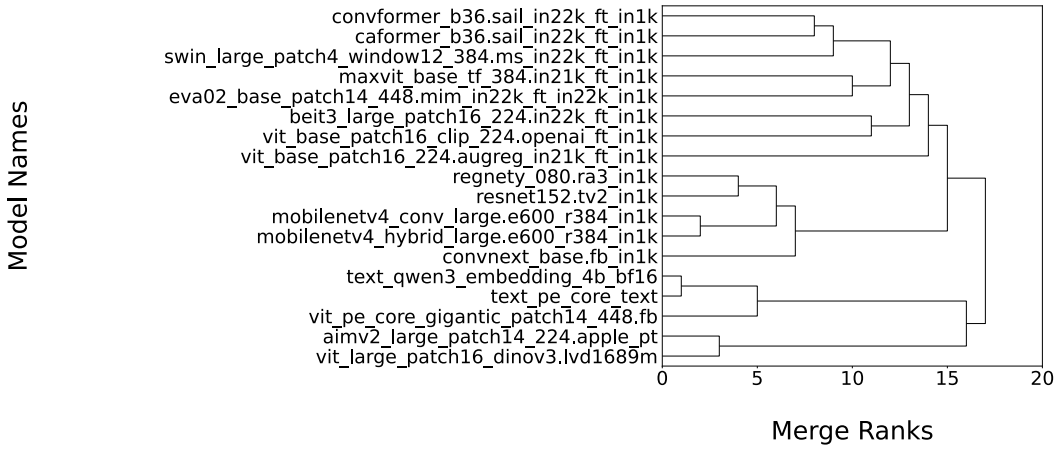



Figure 5: Dendrogram visualizing the hierarchical clustering of models derived from the SNF-fused distance matrix. Clear groupings based on architecture and training method and data emerge.

### 3.5 Aggregating Metrics into a Dendrogram

Since we have collected multiple pairwise metrics to assess model compatibility, we then aggregate all of these results into a single coherent hierarchy to identify similarities, differences, and patterns. To aggregate pairwise metrics, matrix was normalized using min-max scaling. We made an additional copy of each matrix, which we then squared to highlight strong affinities. The normalized matrices were aggregated into a single similarity matrix using similarity network fusion (SNF) with a neighborhood size (K) scaled to the dataset size (Wang et al., 2014). This fused similarity was converted to a distance matrix via a negative log transform. We used hierarchical agglomerative clustering with average linkage to cluster the models using the fused distances. For visualization, as shown in Figure 5 branch heights were replaced with merge ranks to enforce uniform spacing. The resulting tree was plotted as a dendrogram for viewing.

We observe that first 3 branches result in 4 clusters, each with distinct characteristics. The top cluster, containing 8 models, all of which are isometric transformers or based on hierarchical transformers, all of which are pretrained on concepts (language supervised or trained on ImageNet-22k), and all of which are trained on ImageNet-1k at the end of their training. The second cluster, containing 5 models, contains all of the CNN or CNN-based hybrid models trained on ImageNet-1k. These two clusters are also clustered together. The next cluster, containing 3 models, text embedding models along with PE-Core, which is closely paired with its corresponding text embedding model. The last cluster contains models where image reconstruction was part of the pretraining objective. DINOv3 is self-supervised, while AIMv2 uses both self-supervised and non-CLIP language supervision. These two clusters also cluster together. Additionally, we observe that models from the same family are consistently clustered as siblings.

## 4 Discussion, Limitations, and Conclusion

Across our experiments, we observe patterns in model representation compatibility that line up with characteristic of the models themselves. These include the architecture of the model and the training processes and data used. We observe high compatibility across many models while using a shared ”lingua franca” embedding space for translation. This enables research into the geometry, interpretability, and properties of such a shared embedding space, as demonstrated in our work. This method can likely also be extended to perform 0-shot adaptation, though this is not the focus of our study. While our setup is able to cluster models by their architectural (”nature”) and training (”nurture”) characteristics, we do not include detailed experiments into optimal training recipes for adapters. Our method sacrifices some performance compared to application-oriented methods and instead aims to explore the properties and limits of multi-model alignment. Our hyperparameters are hand-tuned for decent convergence, which warrants additional experiments if absolute peak performance is the goal.

## IMPACT STATEMENT

This paper presents work whose goal is to advance the field of machine learning. In addition to the potential impacts common to all work related to machine learning, we believe that our work may lead to more efficient, universal, and upgradable foundation models, potentially reducing energy usage spend on neural network training. Like the prior vec2vec work, our method demonstrates a degree of general compatibility between all studied text and image models. While we train using paired data with supervised alignment objectives, it is likely possible to modify our work to align embeddings in an unsupervised manner at scale and on mixed modalities. This extends the potential impacts of vec2vec's vector database attacks to image- and mixed-modality applications, making these vector databases vulnerable to attack. Attackers could potentially decode leaked image vectors into captions or use them in conjunction with an image generation model to leak an approximation of the image itself. These potential risks highlight the newfound security vulnerability of vector databases, even those that store image embeddings.

## REFERENCES

Zahra Babaiee, Peyman M Kiasari, Daniela Rus, and Radu Grosu. Unveiling the unseen: Identifiable clusters in trained depthwise convolutional kernels. *arXiv preprint arXiv:2401.14469*, 2024.

Zahra Babaiee, Peyman M. Kiassari, Daniela Rus, and Radu Grosu. The quest for universal master key filters in ds-cnns, 2025. URL https://arxiv.org/abs/2509.11711.

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

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. URL https://arxiv.org/abs/2504.13181.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567, 2021. doi: 10.1109/CVPR46437.2021.00356.

Zirui Chen and Michael F. Bonner. Universal dimensions of visual representation. *arXiv preprint arXiv:2408.12804*, 2024. URL https://arxiv.org/abs/2408.12804.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

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.

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. *arXiv preprint arXiv:2010.11929*, 2021. URL https://arxiv.org/abs/2010.11929.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2024.105171. URL https://www.sciencedirect.com/science/article/pii/S0262885624002762.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G. Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9641–9654, June 2025.

Raj Magesh Gauthaman, Brice Ménard, and Michael F. Bonner. Universal scale-free representations in human visual cortex. *arXiv preprint arXiv:2409.06843*, 2025. URL https://arxiv.org/abs/2409.06843.

Paul Gavrikov and Janis Keuper. Cnn filter db: An empirical investigation of trained convolutional filters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19066–19076, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. *arXiv preprint arXiv:1903.08859*, 2019. URL https://arxiv.org/abs/1903.08859.

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

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025. URL https://arxiv.org/abs/2505.12540.

M. Klabunde, L. Caspari, and F. Lemmerich. Revisiting the relation between robustness and universality. *arXiv preprint arXiv:2510.19427*, 2025. URL https://arxiv.org/abs/2510.19427.

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

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar (eds.), *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212, Montreal, Canada, 11 Dec 2015. PMLR. URL https://proceedings.mlr.press/v44/li15convergent.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022.

Pablo Marcos-Manchón and Lluís Fuentemilla. Convergent transformations of visual representation in brains and models. *arXiv preprint arXiv:2507.13941*, 2025. URL https://arxiv.org/abs/2507.13941.

Lucas Maystre, Alvaro Ortega Gonzalez, Charles Park, Rares Dolga, Tudor Berariu, Yu Zhao, and Kamil Ciosek. When embedding models meet: Procrustes bounds and applications. *arXiv preprint arXiv:2510.13406*, 2025.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. *arXiv preprint arXiv:1301.3781*, 2013. URL https://arxiv.org/abs/1301.3781.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2024. URL https://arxiv.org/abs/2311.03658.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2025. URL https://arxiv.org/abs/2410.13928.

Bruno Puri, Jim Berend, Sebastian Lapuschkin, and Wojciech Samek. Atlas-alignment: Making interpretability transferable across language models. *arXiv preprint arXiv:2510.27413*, 2025.

Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, Vaibhav Aggarwal, Tenghui Zhu, Daniele Moro, and Andrew Howard. Mobilenetv4: Universal models for the mobile ecosystem. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 78–96, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73661-2.

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

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Maria Ryskina, Greta Tuckute, Alexander Fung, Ashley Malkin, and Evelina Fedorenko. Language models align with brain regions that represent concepts across modalities. *arXiv preprint arXiv:2508.11536*, 2025. URL https://arxiv.org/abs/2508.11536.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2022. URL https://arxiv.org/abs/2106.10270.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*, 2025.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 459–479, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20053-3.

Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19175–19186, June 2023.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

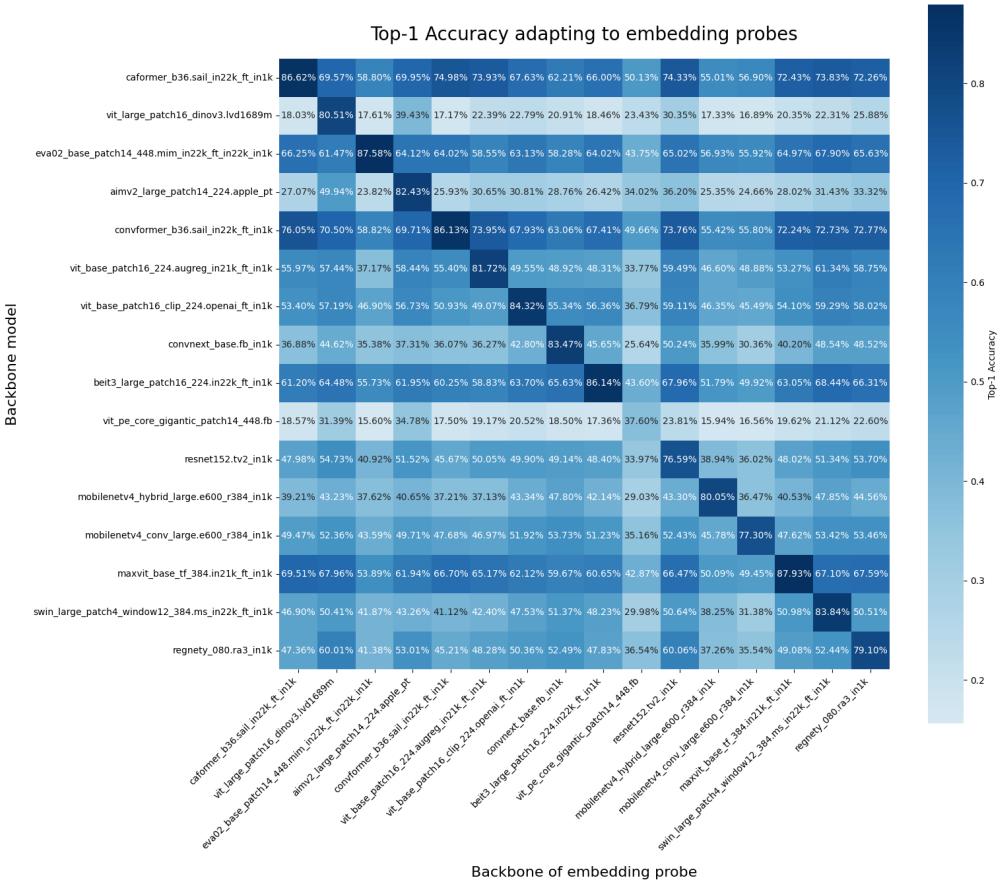
Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9508–9520, 2024.

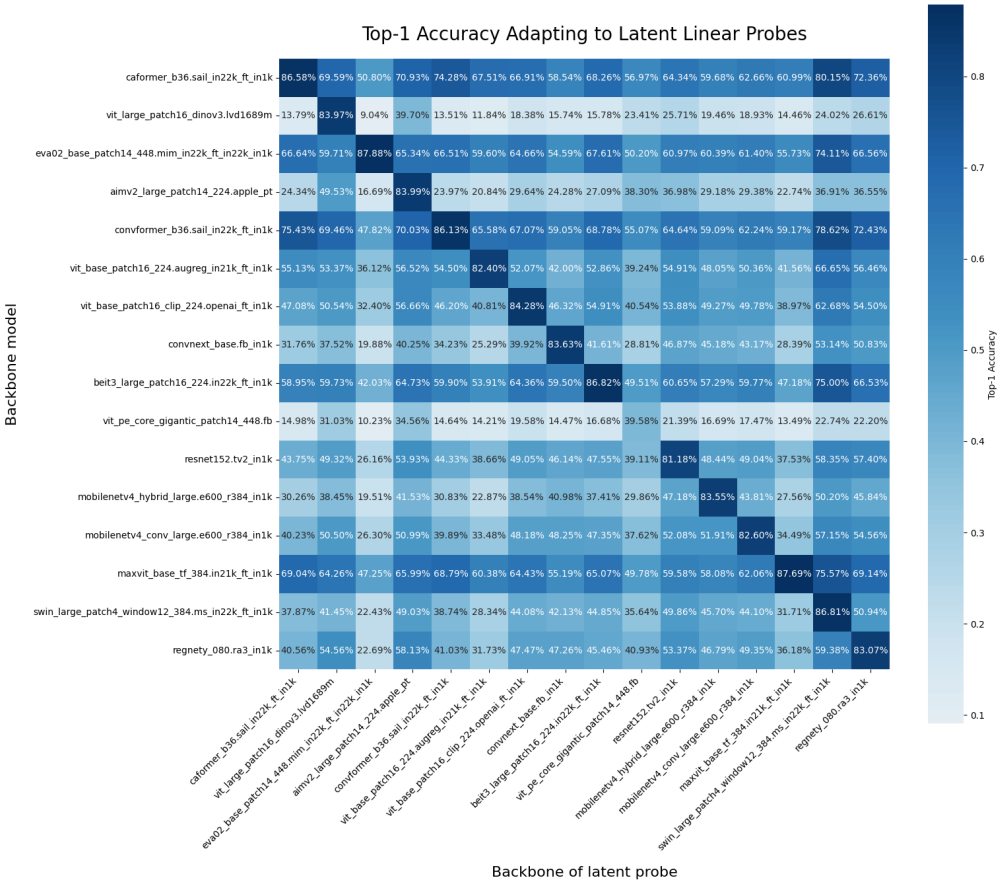
Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):896–912, 2024. doi: 10.1109/TPAMI.2023.3329173.

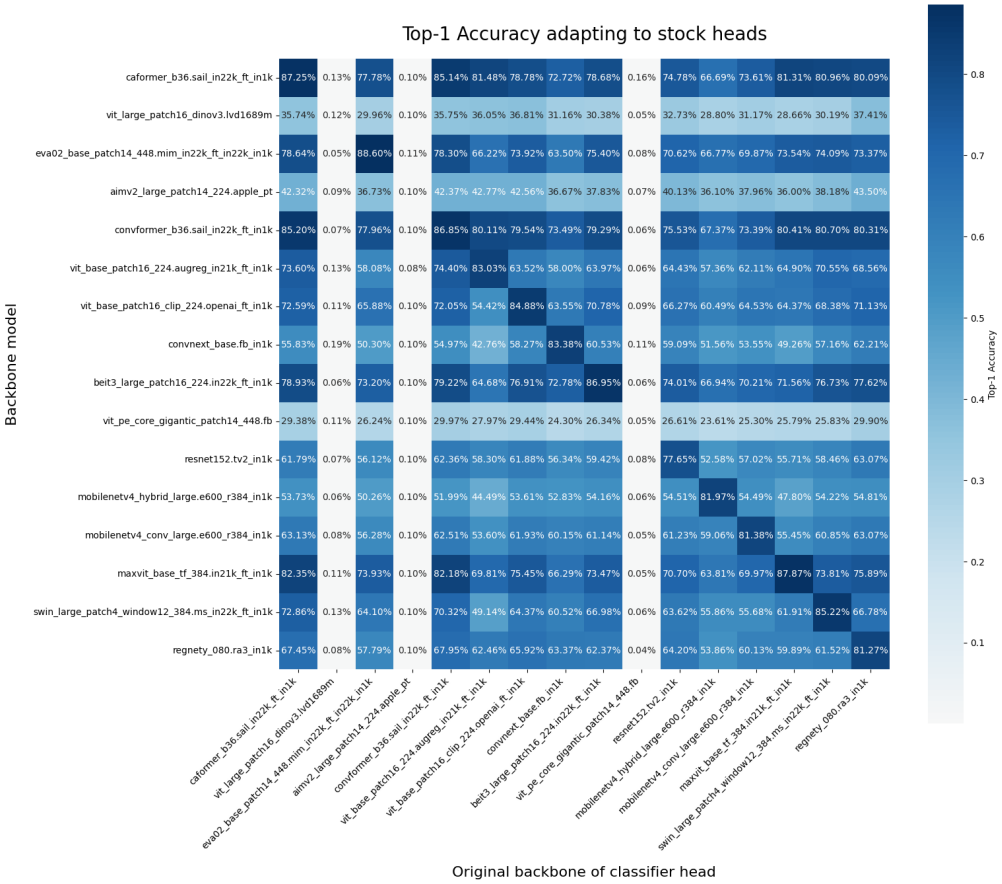
Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

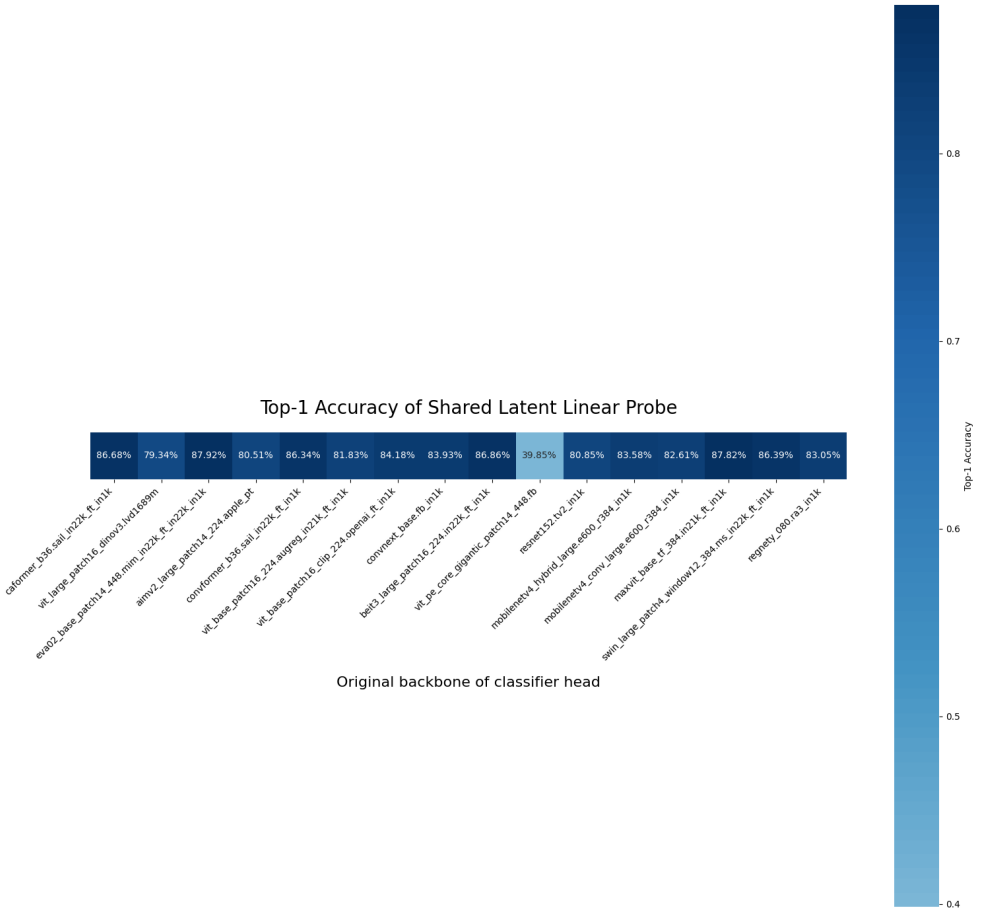
# A Additional Experimental Results

## A.1 Full Classification Heatmaps

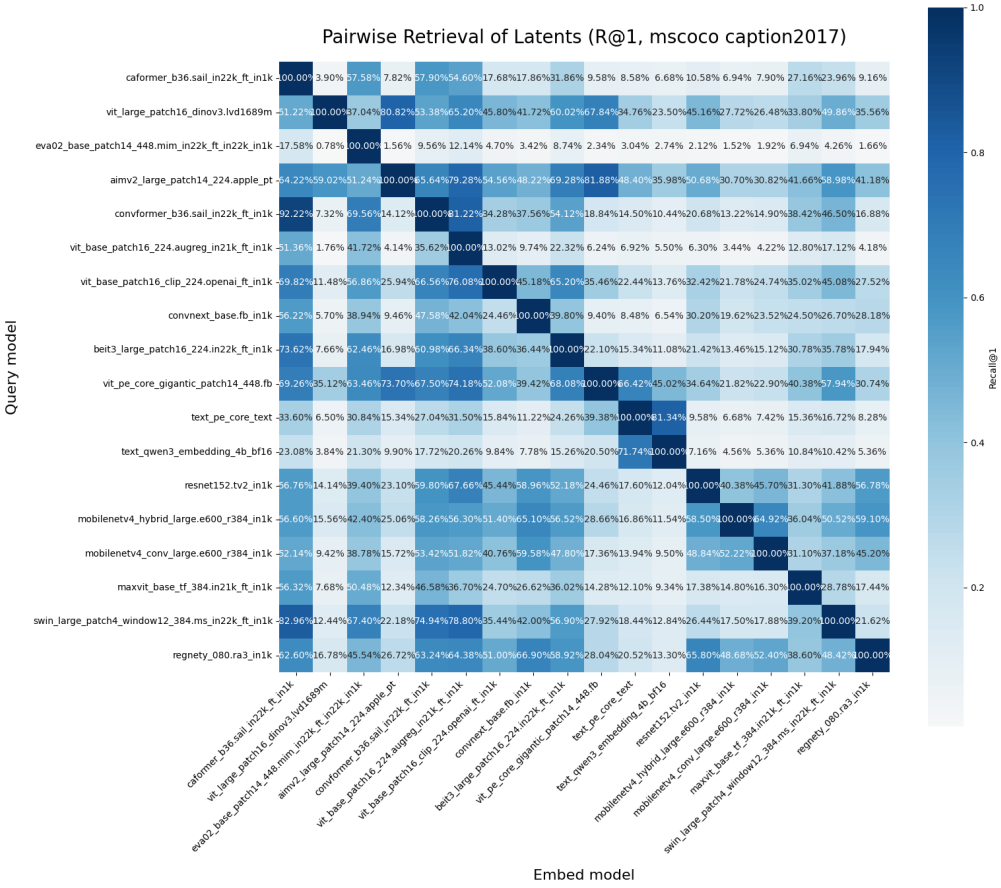


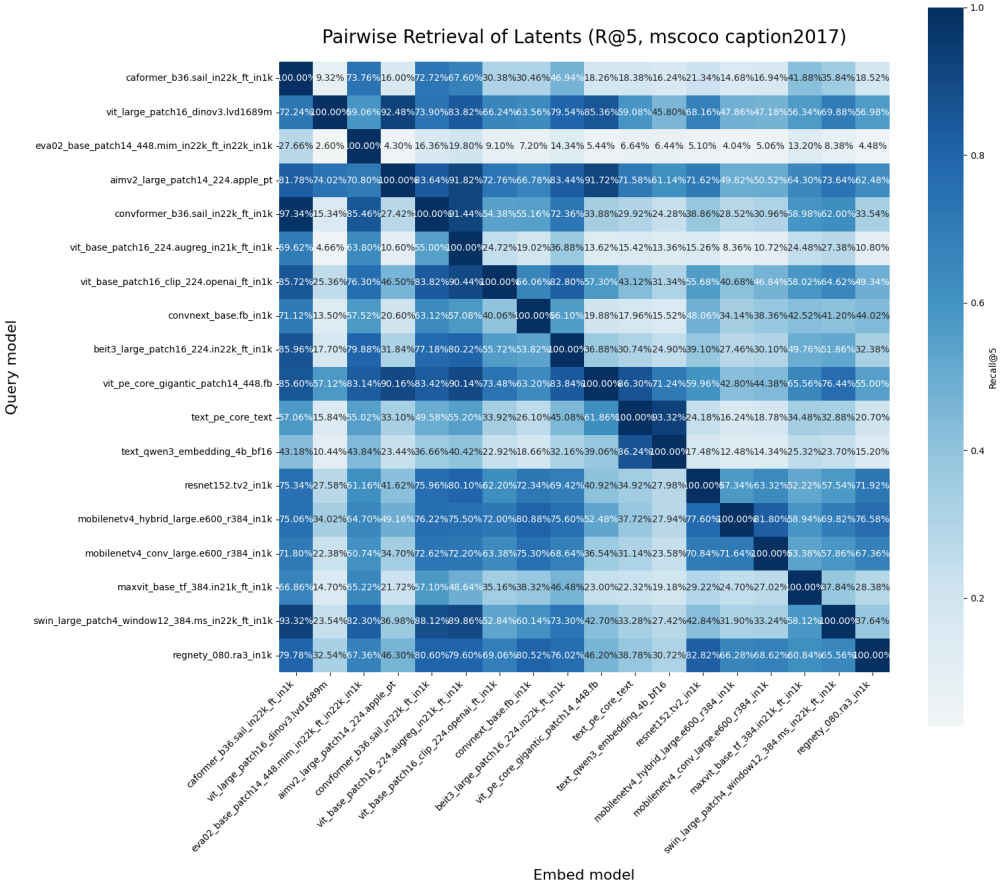

Top-1 Accuracy Adapting to Latent Linear Probes
Backbone model
caformer_b36.sail_in22k_ft_in1k 86.58% 69.59% 50.80% 70.93% 74.28% 67.51% 66.91% 58.54% 68.26% 56.97% 64.34% 59.68% 62.66% 60.99% 80.15% 72.36%
vit_large_patch16_dinov3.lvd1689m 13.79% 83.97% 9.04% 39.70% 13.51% 11.84% 18.38% 15.74% 15.78% 23.41% 25.71% 19.46% 18.93% 14.46% 24.02% 26.61%
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k 66.64% 59.71% 87.88% 65.34% 66.51% 59.60% 64.66% 54.59% 67.61% 50.20% 60.97% 60.39% 61.40% 55.73% 74.11% 66.56%
aimv2_large_patch14_224.apple_pt 24.34% 49.53% 16.69% 83.99% 23.97% 20.84% 29.64% 24.28% 27.09% 38.30% 36.98% 29.18% 29.38% 22.74% 36.91% 36.55%
convformer_b36.sail_in22k_ft_in1k 75.43% 69.46% 47.82% 70.03% 86.13% 65.58% 67.07% 59.05% 68.78% 55.07% 64.64% 59.09% 62.24% 59.17% 78.62% 72.43%
vit_base_patch16_224.augreg_in21k_ft_in1k 55.13% 53.37% 36.12% 56.52% 54.50% 82.40% 52.07% 42.00% 52.86% 39.24% 54.91% 48.05% 50.36% 41.56% 66.65% 56.46%
vit_base_patch16_clip_224.openai_ft_in1k 47.08% 50.54% 32.40% 56.66% 46.20% 40.81% 84.28% 46.32% 54.91% 40.54% 53.88% 49.27% 49.78% 38.97% 62.68% 54.50%
convnext_base.fb_in1k 31.76% 37.52% 19.88% 40.25% 34.23% 25.29% 39.92% 83.63% 41.61% 28.81% 46.87% 45.18% 43.17% 28.39% 53.14% 50.83%
beit3_large_patch16_224.in22k_ft_in1k 58.95% 59.73% 42.03% 64.73% 59.90% 53.91% 64.36% 59.50% 86.82% 49.51% 60.65% 57.29% 59.77% 47.18% 75.00% 66.53%
vit_pe_core_gigantic_patch14_448.fb 14.98% 31.03% 10.23% 34.56% 14.64% 14.21% 19.58% 14.47% 16.68% 39.58% 21.39% 16.69% 17.47% 13.49% 22.74% 22.20%
resnet152.tv2_in1k 43.75% 49.32% 26.16% 53.93% 44.33% 38.66% 49.05% 46.14% 47.55% 39.11% 81.18% 48.44% 49.04% 37.53% 58.35% 57.40%
mobilenetv4_hybrid_large.e600_r384_in1k 30.26% 38.45% 19.51% 41.53% 30.83% 22.87% 38.54% 40.98% 37.41% 29.86% 47.18% 83.55% 43.81% 27.56% 50.20% 45.84%
mobilenetv4_conv_large.e600_r384_in1k 40.23% 50.50% 26.30% 50.99% 39.89% 33.48% 48.18% 48.25% 47.35% 37.62% 52.08% 51.91% 82.60% 34.49% 57.15% 54.56%
maxvit_base_tf_384.in21k_ft_in1k 69.04% 64.26% 47.25% 65.99% 68.79% 60.38% 64.43% 55.19% 65.07% 49.78% 59.58% 58.08% 62.06% 87.69% 75.57% 69.14%
swin_large_patch4_window12_384.ms_in22k_ft_in1k 37.87% 41.45% 22.43% 49.03% 38.74% 28.34% 44.08% 42.13% 44.85% 35.64% 49.86% 45.70% 44.10% 31.71% 86.81% 50.94%
regnety_080.ra3_in1k 40.56% 54.56% 22.69% 58.13% 41.03% 31.73% 47.47% 47.26% 45.46% 40.93% 53.37% 46.79% 49.35% 36.18% 59.38% 83.07%
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k
Backbone of latent probe
Top-1 Accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

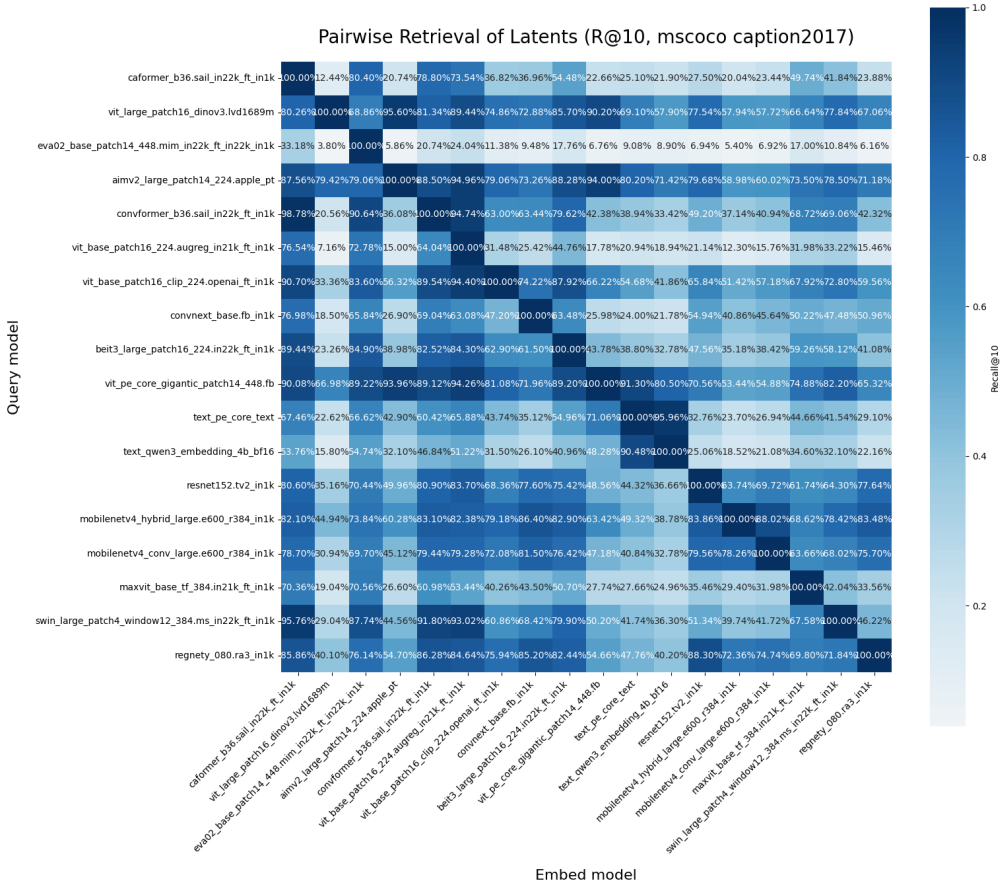

Top-1 Accuracy adapting to stock heads
Backbone model
caformer_b36.sail_in22k_ft_in1k 87.25% 0.13% 77.78% 0.10% 85.14% 81.48% 78.78% 72.72% 78.68% 0.16% 74.78% 66.69% 73.61% 81.31% 80.96% 80.09%
vit_large_patch16_dinov3.lvd1689m 35.74% 0.12% 29.96% 0.10% 35.75% 36.05% 36.81% 31.16% 30.38% 0.05% 32.73% 28.80% 31.17% 28.66% 30.19% 37.41%
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k 78.64% 0.05% 88.60% 0.11% 78.30% 66.22% 73.92% 63.50% 75.40% 0.08% 70.62% 66.77% 69.87% 73.54% 74.09% 73.37%
aimv2_large_patch14_224.apple_pt 42.32% 0.09% 36.73% 0.10% 42.37% 42.77% 42.56% 36.67% 37.83% 0.07% 40.13% 36.10% 37.96% 36.00% 38.18% 43.50%
convformer_b36.sail_in22k_ft_in1k 85.20% 0.07% 77.96% 0.10% 86.85% 80.11% 79.54% 73.49% 79.29% 0.06% 75.53% 67.37% 73.39% 80.41% 80.70% 80.31%
vit_base_patch16_224.augreg_in21k_ft_in1k 73.60% 0.13% 58.08% 0.08% 74.40% 83.03% 63.52% 58.00% 63.97% 0.06% 64.43% 57.36% 62.11% 64.90% 70.55% 68.56%
vit_base_patch16_clip_224.openai_ft_in1k 72.59% 0.11% 65.88% 0.10% 72.05% 54.42% 84.88% 63.55% 70.78% 0.09% 66.27% 60.49% 64.53% 64.37% 68.38% 71.13%
convnext_base.fb_in1k 55.83% 0.19% 50.30% 0.10% 54.97% 42.76% 58.27% 83.38% 60.53% 0.11% 59.09% 51.56% 53.55% 49.26% 57.16% 62.21%
beit3_large_patch16_224.in22k_ft_in1k 78.93% 0.06% 73.20% 0.10% 79.22% 64.68% 76.91% 72.78% 86.95% 0.06% 74.01% 66.94% 70.21% 71.56% 76.73% 77.62%
vit_pe_core_gigantic_patch14_448.fb 29.38% 0.11% 26.24% 0.10% 29.97% 27.97% 29.44% 24.30% 26.34% 0.05% 26.61% 23.61% 25.30% 25.79% 25.83% 29.90%
resnet152.tv2_in1k 61.79% 0.07% 56.12% 0.10% 62.36% 58.30% 61.88% 56.34% 59.42% 0.08% 77.65% 52.58% 57.02% 55.71% 58.46% 63.07%
mobilenetv4_hybrid_large.e600_r384_in1k 53.73% 0.06% 50.26% 0.10% 51.99% 44.49% 53.61% 52.83% 54.16% 0.06% 54.51% 81.97% 54.49% 47.80% 54.22% 54.81%
mobilenetv4_conv_large.e600_r384_in1k 63.13% 0.08% 56.28% 0.10% 62.51% 53.60% 61.93% 60.15% 61.14% 0.05% 61.23% 59.06% 81.38% 55.45% 60.85% 63.07%
maxvit_base_tf_384.in21k_ft_in1k 82.35% 0.11% 73.93% 0.10% 82.18% 69.81% 75.45% 66.29% 73.47% 0.05% 70.70% 63.81% 69.97% 87.87% 73.81% 75.89%
swin_large_patch4_window12_384.ms_in22k_ft_in1k 72.86% 0.13% 64.10% 0.10% 70.32% 49.14% 64.37% 60.52% 66.98% 0.06% 63.62% 55.86% 55.68% 61.91% 85.22% 66.78%
regnety_080.ra3_in1k 67.45% 0.08% 57.79% 0.10% 67.95% 62.46% 65.92% 63.37% 62.37% 0.04% 64.20% 53.86% 60.13% 59.89% 61.52% 81.27%
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k
Original backbone of classifier head
Top-1 Accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Top-1 Accuracy of Shared Latent Linear Probe
86.68%
79.34%
87.92%
80.51%
86.34%
81.83%
84.18%
83.93%
86.86%
39.85%
80.85%
83.58%
82.61%
87.82%
86.39%
83.05%
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
coviformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k
Original backbone of classifier head
Top-1 Accuracy
0.8
0.7
0.6
0.5
0.4

## A.2 FULL RETRIEVAL HEATMAPS

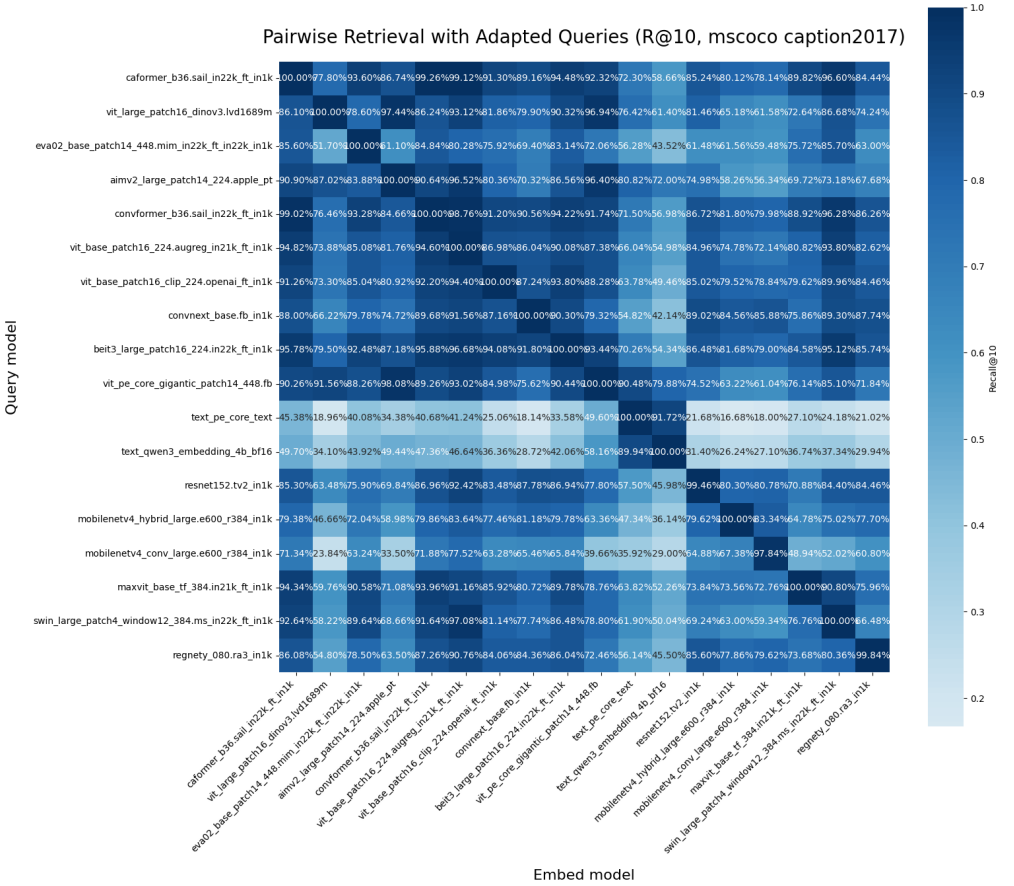



Pairwise Retrieval of Latents (R@5, mscoco caption2017)
Query model
Embed model
Recall@5
1.0
0.8
0.6
0.4
0.2
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k


Pairwise Retrieval of Latents (R@10, mscoco caption2017)
Query model
Embed model
Recall@10
1.0
0.8
0.6
0.4
0.2
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k


Pairwise Retrieval with Adapted Queries (R@1, mscoco caption2017)
Query model
Embed model
Recall@1
1.0
0.8
0.6
0.4
0.2
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k


Pairwise Retrieval with Adapted Queries (R@5, mscoco caption2017)
Query model
Embed model
Recall@5
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k


Pairwise Retrieval with Adapted Queries (R@10, mscoco caption2017)
Query model
Embed model
Recall@10
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k


Pairwise Retrieval with Adapted Embeds (R@1, mscoco caption2017)
Query model
Embed model
Recall@1
1.0
0.8
0.6
0.4
0.2
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k


Pairwise Retrieval with Adapted Embeds (R@5, mscoco caption2017)
Query model
Embed model
Recall@5
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k


Pairwise Retrieval with Adapted Embeds (R@10, mscoco caption2017)
Query model
Embed model
Recall@10
caformer_b36.sail_in22k_ft_in1k
vit_large_patch16_dinov3.lvd1689m
eva02_base_patch14_448.mim_in22k_ft_in22k_in1k
aimv2_large_patch14_224.apple_pt
convformer_b36.sail_in22k_ft_in1k
vit_base_patch16_224.augreg_in21k_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
convnext_base.fb_in1k
beit3_large_patch16_224.in22k_ft_in1k
vit_pe_core_gigantic_patch14_448.fb
text_pe_core_text
text_qwen3_embedding_4b_bf16
resnet152.tv2_in1k
mobilenetv4_hybrid_large.e600_r384_in1k
mobilenetv4_conv_large.e600_r384_in1k
maxvit_base_tf_384.in21k_ft_in1k
swin_large_patch4_window12_384.ms_in22k_ft_in1k
regnety_080.ra3_in1k

### A.3 CLUSTER PURITY TABLE

| Column 1 | COCO Caption | | ImageNet-1k | |
|---|---|---|---|---|
| | kNN Purity | Silhouette | kNN Purity | Silhouette |
| CAFormer-B36 | 0.421 | -0.148 | 0.806 | -0.23 |
| DINOv3-Large | 0.928 | 0.073 | 0.978 | 0.154 |
| EVA02-Base | 0.417 | -0.155 | 0.84 | -0.091 |
| AIMv2-Large | 0.787 | -0.005 | 0.944 | -0.086 |
| ConvFormer-B36 | 0.469 | -0.114 | 0.854 | -0.167 |
| ViT-B/16-AugReg | 0.309 | -0.198 | 0.622 | -0.234 |
| ViT-B/16-CLIP | 0.467 | -0.067 | 0.877 | -0.064 |
| ConvNeXt-Base | 0.676 | 0.062 | 0.872 | 0.086 |
| BEiT3-Large | 0.418 | -0.144 | 0.844 | -0.131 |
| PE-Core-Image | 0.681 | 0.0008 | 0.894 | -0.063 |
| PE-Core-Text | 0.770 | 0.097 | | |
| Qwen3-Embed-4B | 0.848 | -0.122 | | |
| ResNet152 | 0.412 | -0.224 | 0.768 | -0.279 |
| MobileNetV4-Hybrid | 0.792 | 0.038 | 0.972 | 0.08 |
| MobileNetV4-Conv | 0.904 | 0.053 | 0.99 | 0.091 |
| MaxViT-Base | 0.545 | -0.218 | 0.89 | -0.123 |
| Swin-Large | 0.652 | 0.018 | 0.948 | 0.201 |
| RegNetY-8.0GF | 0.581 | -0.171 | 0.864 | -0.174 |

Table 1: Cluster purity for all models.

## B GENERATIVE AI USAGE STATEMENT

We use generative AI tools to assist with writing scripts to generate figures and proofreading the paper. All figures were generated programatically; no figures were directly generated as an image using an AI tool. AI tools were used to identify relevant prior works throughout the research and writing process. No AI generated text was used directly. Some code used for experiments was written with AI assistance or entirely AI generated. All code was audited by the authors before usage.